Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition. Volume 2 : Traitement Automatique des Langues Naturelles

Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, Stéphane Schneider

6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) (JEP-TALN-RÉCITAL) [1]

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).

Volume 2 : Traitement Automatique des Langues Naturelles

Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, Stéphane Schneider (Éds.)

Nancy, France, 08-19 juin 2020

1. https://jep-taln2020.loria.fr/

Avec le soutien de

Message des présidents de l'AFCP et de l'ATALA

En ce printemps 2020, et les circonstances exceptionnelles qui l'accompagnent, c'est avec une émotion toute particulière que nous vous convions à la 6e édition conjointe des Journées d'Études sur la Parole (JEP), de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) et des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL). Après une première édition commune en 2002 (à Nancy, déjà !), et une expérience renouvelée avec succès en 2004, c'est désormais tous les quatre ans (Avignon 2008, Grenoble 2012) que se répète cet événement commun, attendu de pied ferme par les membres des deux communautés scientifiques voisines.

Cette édition 2020 est exceptionnelle, puisque dans le cadre des mesures sanitaires liées à la pandémie mondiale de COVID-19 (confinement strict, puis déconfinement progressif), la conférence ne peut avoir lieu à Nancy comme initialement prévu, mais se déroule à distance, sous forme virtuelle, soutenue par les technologies de l'information et de la communication. Nous remercions ici chaleureusement les organisateurs, Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni et Sylvain Pogodalla, qui ont dû faire preuve de souplesse, d'inventivité, de détermination, de puissance de travail, et de tant d'autres qualités encore, afin de maintenir la conférence dans ces circonstances, en proposant un format inédit. Grâce aux différentes solutions mises en œuvre dans un délai court, la publication des communications scientifiques est assurée, structurée, et les échanges scientifiques sont favorisés, même à distance.

Bien entendu, nous regrettons tous que cette réunion JEP-TALN-RECITAL ne permette pas, comme ses prédécesseurs, de nouer ou renforcer les liens sociaux entre les différents membres de nos communautés respectives – chercheurs, jeunes et moins jeunes, académiques et industriels, professionnels et étudiants – autour d'une passionnante discussion scientifique ou d'un mémorable événement social... Notre conviction est qu'il est indispensable de maintenir à l'avenir de tels lieux d'échanges dans le domaine francophone, afin bien sûr de permettre aux jeunes diplômés de venir présenter leurs travaux et poser leurs questions sans la barrière de la langue, mais aussi de dynamiser nos communautés, de renforcer les échanges et les collaborations, et d'ouvrir la discussion autour des enjeux d'avenir, qui questionnent plus que jamais la place de la science et des scientifiques dans notre société.

Lors de la précédente édition, nous nous interrogions sur les phénomènes et tendances liés à l'apprentissage profond et sur leurs impacts sur les domaines de la Parole et du TAL. Force est de constater que l'engouement pour ces approches dans nos domaines a permis un retour sur le devant de la scène des domaines liés à l'Intelligence Artificielle, animant parfois un débat tant philosophique que technique sur la place de la machine dans la société, notamment à travers le questionnement sur la vie privée de l'utilisateur. Ces questionnements impactent tant la Parole que le TAL, d'une part sur la place de la gestion des données, d'autre part sur les modèles eux-mêmes. Malgré ces questionnements, nous constatons que les acquis et les expertises perdurent, et les nouvelles approches liées à l'apprentissage profond ont permis un rapprochement des domaines de la Parole et du TAL, sans les dénaturer, à la manière des conférences JEP-TALN-RECITAL qui créent un espace plus grand d'échange et d'enrichissement réciproques.

Nous terminons ces quelques mots d'ouverture en remerciant l'ensemble des personnes qui ont rendu possible cet événement qui restera, nous l'espérons, riche et passionnant, malgré les circonstances. L'ATALA et l'AFCP tiennent tout d'abord à réitérer leurs remerciements aux organisateurs des JEP, de TALN et de RECITAL, qui sont parvenus à maintenir le cap à travers vents et marées. Nos remerciements vont également à l'ensemble des membres des comités de programme, dont le travail et l'implication ont permis de garantir la qualité et la cohérence du programme finalement retenu. Un grand merci aux relecteurs pour le temps et le soin qu'ils ont dédiés à ce travail anonyme et indispensable. Ils se reflètent dans la qualité des soumissions que chacun pourra découvrir sur le site de la conférence.

En conclusion, cette 6e édition conjointe JEP-TALN-RECITAL est exceptionnelle parce qu'elle se tient dans un contexte de crise généralisée — crise sanitaire, économique, voire sociale et politique. Mais nous

formons le vœu qu'elle reste également dans les annales pour la qualité des échanges scientifiques qu'elle aura suscités, et pour le message envoyé à nos communautés scientifiques et à la société dans son ensemble, un message de détermination et de confiance en l'avenir, où la science et les nouvelles technologies restent au service de l'humain.

Véronique Delvaux, présidente de l'Association Francophone de la Communication Parlée
Christophe Servan, président de l'Association pour le Traitement Automatique des LAngues

Préface

En 2002, l'AFCP (Association Francophone pour la Communication Parlée) et l'ATALA (Association pour le Traitement Automatique des Langues) organisèrent conjointement leurs principales conférences afin de réunir en un seul lieu, à Nancy, les communautés du traitement automatique et de la description des langues écrites, parlées et signées.
En 2020, la sixième conférence commune revient à Nancy, après Fès (2004), Avignon (2008), Grenoble (2012) et Paris (2016). Elle est organisée par le LORIA (Laboratoire lorrain de recherche en informatique et ses applications, UMR 7503), l'ATILF (Analyse et traitement informatique de la langue française, UMR 7118) et l'INIST (Institut de l'information scientifique et technique) et regroupe :

— les 33es Journées d'Études sur la Parole (JEP),
— la 27^{e} conférence sur le Traitement Automatique des Langues Naturelles (TALN),
— la 22^{e} Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL).

Les circonstances particulières liées à l'épidémie de Covid-19 en France et dans le monde ont conduit à une virtualisation de la conférence. Ainsi, malgré un rassemblement physique qui n'a pu avoir lieu, diffusions, présentations (au gré des auteurs) et discussions des articles acceptés ont lieu sur le site internet de la conférence. Les tutoriels, certains ateliers, et le salon de l'innovation qui accompagnent la conférence ont cependant dû être annulés, mais les ateliers suivants sont maintenus :

— Défi Fouille de Textes (DEFT 2020),
— Éthique et TRaitemeNt Automatique des Langues (ÉTeRNAL).

La conférence accueille également des conférencières et conférenciers invités dont les exposés sont diffusés sur le site : Dirk Hovy (université de Bocconi, Milan, Italie, invité ÉTeRNAL) ainsi que Marie-Jean Meurs (Université du Québec à Montréal, UQAM, Canada) et Hugo Cyr (Faculté de science politique et droit à l'Université du Québec à Montréaln UQAM, Canada). En raison des circonstance particulières, un exposé conjoint de Christine Meunier (Laboratoire Parole et Langage LPL, CNRS, Aix-en-Provence, France) et Christophe Stécoli (police technique et scientifique française) a dû être annulé et reporté à une journée spéciale en septembre 2020.
Ces actes regroupent les articles des conférences JEP (volume 1), TALN (volume 2), RÉCITAL (volume 3), les articles décrivant les démonstrations (volume 4), et les articles des ateliers DEFT (volume 5) et ÉTeRNAL (volume 6). Pour la première fois, un appel spécifique à résumés en français d'articles parus dans une sélection de conférences internationales en 2019 était également proposé (volume 4). Un appel spécifique apprenti·e·s chercheur·euse·s destiné aux étudiants de licence, de master, ou en première année de thèse a également été proposé, pour leur proposer des présentations courtes ou sous forme de poster de leurs projets.
Pour les JEP, 87 articles ont été soumis, parmi lesquels 74 ont été sélectionnés, soit un taux de sélection de 85%.
Pour TALN, 58 articles ont été soumis, parmi lesquels 37 ont été sélectionnés, soit un taux de sélection de 63%, dont 10 comme article longs (17% des soumissions) et 27 comme article courts dont 20 en présentation orale (34% des soumissions) et 7 en présentation poster (12% des soumissions).
Pour RÉCITAL, 22 articles ont été soumis, parmi lesquels 16 ont été sélectionnés, soit un taux de sélection de 73%.
Nous souhaitons vivement remercier toutes les personnes qui ont participé à ce travail de relecture et de sélection :

— l'ensemble des relecteurs (voir page xi),
— le comité de programme des JEP (voir page viii),
— le comité de programme de TALN (voir page ix),
— le comité de programme de RÉCITAL (voir page x).

Nous souhaitons également remercier nos sociétés savantes : l'AFCP, assurant la continuité des éditions successives des JEP, et l'ATALA, dont le CPerm (comité permanent) assure la continuité des éditions

successives de TALN.
Nous remercions le comité d'organisation et les nombreuses personnes qui ont assuré le soutien administratif et technique pour que cette conférence se déroule dans les meilleures conditions, et en particulier Yannick Parmentier pour son travail pour la diffusion de ces actes sur HAL et les différents sites d'archives ouvertes (anthologie ACL et talnarchives.atala.org/).
Nous remercions enfin tous les partenaires institutionnels et industriels qui nous ont fait confiance, en particulier l'université de Lorraine, le CNRS, l'Inria, le LORIA, l'ATILF, l'INIST, le master TAL de l'Institut des Sciences du Digital Management & Cognition (IDMC), le projet OLKI de l'initiative Lorraine Université d'Excellence (LUE), la Région Grand Est, *The Evaluations and Language resources Distribution Agency* (ELDA), le projet ANR PARSEME-FR, la délégation générale à la langue française et aux langues de France (DGLFLF), l'Association des Professionnels des Industries de la Langue (APIL) et les entreprises Synapse, Yseop et Orange.
Bonne conférence à toutes et à tous !

Les présidentes et présidents JEP : David Langlois et Slim Ouni
TALN : Chloé Braud et Sylvain Pogodalla
RÉCITAL : Christophe Benzitoun et Laurine Huber

Comités

Comité de programme des JEP

Martine Adda-Decker (Laboratoire de Phonétique et Phonologie, CNRS)
Jean-Francois Bonastre (LIA, Université d'Avignon)
Fethi Bougares (LIUM, Le Mans Université) Philippe Boula De Mareüil (LIMSI, CNRS)
Hervé Bredin (LIMSI, CNRS)
Olivier Crouzet (LLING, Université de Nantes)
Elisabeth Delais-Roussarie (LLING, Université de Nantes)
Véronique Delvaux (Laboratoire de Phonétique, IRSTL, Université de Mons)
Camille Fauth (LiLPa, Université de Strasbourg)
Emmanuel Ferragne (CLILLAC-ARP, Université de Paris)
Cecile Fougeron (Laboratoire de Phonétique et Phonologie, CNRS)
Corinne Fredouille (LIA, Université d'Avignon)
Alain Ghio (LPL, CNRS)
Camille Guinaudeau (LIMSI, Université Paris Sud)
Anne Guyot Talbot (CLILLAC-ARP, Université de Paris 7)
Bernard Harmegnies (Laboratoire de Phonétique, IRSTL, Université de Mons)
Nathalie Henrich Bernardoni (Gipsa-lab, CNRS)
Bassam Jabaian (LIA, Université d'Avignon)
David Langlois (LORIA, Université de Lorraine)
Yves Laprie (LORIA, CNRS)
Anthony Larcher (LIUM, Université du Maine)
Gwénolé Lecorvé (IRISA, Université de Rennes)
Benjamin Lecouteux (LIG, Université Grenoble Alpes)
Georges Linarès (LIA, Université d'Avignon)
Damien Lolive (IRISA, Université Rennes)
Julie Mauclair (IRIT)
Yohann Meynadier (LPL, Aix-Marseille Université)
Slim Ouni (LORIA, Université de Lorraine)
Thomas Pellegrini (IRIT, Université de Toulouse)
François Portet (LIG, Grenoble INP)
Fabian Santiago (Structures Formelles du Langage, Université de Paris 8)
Christophe Savariaux (Gipsa-lab, CNRS)
Nathalie Vallee (Gipsa-lab, Université Grenoble Alpes)
Ioana Vasilescu (LIMSI, CNRS)

Comités de programme TALN

Maxime Amblard (LORIA, Université de Lorraine)
Chloé Braud (IRIT, CNRS)
Caroline Brun (Naver Labs Europe)
Nathalie Camelin (LIUM, Université du Maine)
Marie Candito (Universtité Paris 7)
Vincent Claveau (IRISA, CNRS)
Chloé Clavel (Telecom-ParisTech)
Mathieu Constant (ATILF, CNRS, Université de Lorraine)
Pascal Denis (Inria)
Cécile Fabre (Université Toulouse 2)
Thomas François (Université catholique de Louvain)
Núria Gala (LPL, CNRS, Aix-Marseille Université)
Natalia Grabar (STL, CNRS, Université Lille 3)
Anne-Laure Ligozat (LIMSI, CNRS, ENSIIE, Université Paris-Saclay)
Emmanuel Morin (LINA, Université de Nantes)
Sylvain Pogodalla (LORIA, Inria)
Solen Quiniou (LINA, Université de Nantes)
Corentin Ribeyre (Etermind)
Tim van de Cruys (IRIT, CNRS)
Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay)

Comité de programme RÉCITAL

Jean-Yves Antoine (Université François Rabelais de Tours)
Sonia Badene (Linagora, IRIT)
Frédéric Béchet (LIF, Aix Marseille Université)
Christophe Benzitoun (ATILF, Université de Lorraine)
Maria Boritchev (LORIA, Inria)
Léo Bouscarrat (EURA NOVA, Aix-Marseille Université)
Manon Cassier (INALCO, Paris)
Kevin Deturck (Viseo Technologies)
Emmanuelle Esperança-Rodier (GETALP, Université Grenoble Alpes)
Kim Gerdes (sorbonne nouvelle)
Nicolas Hernandez (LINA, UMR 6241, CNRS, Université de Nantes)
Lydia-Mai Ho-Dac (CLLE-ERSS, Université Toulouse Jean Jaurès)
Laurine Huber (LORIA, Université de Lorraine)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre)
Gwénolé Lecorvé (IRISA, Université de Rennes, CNRS)
Joël Legrand (LORIA, Inria, CNRS)
Anne-Laure Ligozat (LIMSI, CNRS, ENSIIE, Université Paris-Saclay)
Pierre Ludmann (LORIA, Université de Lorraine)
Yann Mathet (Université de Caen)
Anne-Lyse Minard (IRISA, CNRS)
Sandrine Ollinger (ATILF, UMR 7118, CNRS)
Yannick Parmentier (LORIA, Université de Lorraine)
Justine Reynaud (LORIA, Université de Lorraine)
Stella Zevio (LIPN, Université de Paris 13)

Relectrices et relecteurs

Gilles Adda (LIMSI, CNRS) Salah Ait-Mokhtar (Naver Labs Europe)
Charlotte Alazard (Université Toulouse 2 Jean Jaurès)
Alexandre Allauzen (LIMSI-CNRS, Université Paris-Sud)
Pascal Amsili (Université Paris Diderot)
Pierre André Hallé (Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3)
Régine André-Obrecht (Université Paul Sabatier Toulouse III)
Jean-Yves Antoine (Université François Rabelais de Tours)
Nicolas Audibert (Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3)
Nelly Barbot (IRISA, Université de Rennes 1)
Claude Barras (LIMSI, CNRS)
Loïc Barrault (University of Sheffield)
Katarina Bartkova (ATILF, Université de Lorraine)
Frédéric Béchet (LIF, Aix Marseille Université)
Nathalie Bedoin (DDL, Université Lyon 2)
Patrice Bellot (LSIS, CNRS, Aix-Marseille Université)
Asma Ben Abacha (National Library of Medicine, National Institutes of Health)
Delphine Bernhard (LiLPa, Université de Strasbourg)
Roxane Bertrand (LPL, CNRS, Aix-Marseille Université)
Laurent Besacier (Laboratoire d'Informatique de Grenoble)
Yves Bestgen (F.R.S-FNRS et Université Catholique de Louvain)
Frédéric Bimbot (IRISA, CNRS)
Caroline Bogliotti (MODYCO, UMR 7114, CNRS, Université Paris Nanterre)
Anne Bonneau (LORIA, CNRS)
Stéphanie Borel (Université de Tours)
Féthi Bougarès (LIUM, Le Mans Université)
Leila Boutora (Laboratoire Parole et Langage, Aix Marseille Université)
Paul Caillon (LORIA, Université de Lorraine)
Mélanie Canault (DDL, Université Lyon 2)
Thierry Charnois (LIPN, CNRS, Université de Paris 13)
Chloé Clavel (Telecom-ParisTech)
Maximin Coavoux (Université Grenoble Alpes, CNRS)
Vincent Colotte (LORIA, Université de Lorraine)
Juan Manuel Coria (LIMSI, Université Paris-Saclay Paris 13)
Benoît Crabbé (Université Paris 7)
Lise Crevier Buchman (Laboratoire de Phonétique et Phonologie, CNRS, Hôpital Foch)
Béatrice Daille (LINA, Université de Nantes)
Géraldine Damnati (Orange Labs)
Dan Dediu (Dynamique du Langage, UMR5596, Université Lumière Lyon 2)
Joseph Di Martino (LORIA, Université de Lorraine)
Gaël Dias (Université Caen Normandie)
Amazouz Djegdjiga (LPP, Université Sorbonne Nouvelle – Paris 3)
Benjamin Elie (IMSIA, ENSTA ParisTech)
Iris Eshkol-Taravella (Université d'Orléans)
Emmanuelle Esperança-Rodier (GETALP, Université Grenoble Alpes)
Yannick Estève (LIA, Université d'Avignon)
Dominique Estival (Western Sydney University)
Olivier Ferret (CEA LIST)
Lionel Fontan (Archean Labs)
Karën Fort (Sorbonne Université)
Claire Gardent (LORIA, CNRS)
Eric Gaussier (LIG, Université Grenoble Alpes)
Cédric Gendrot (LPP, Université Sorbonne Nouvelle – Paris 3)
James German (Laboratoire Parole et Langage, Aix Marseille Université)
Cyril Goutte (National Research Council Canada)
Cyril Grouin (LIMSI, CNRS, Université Paris-Saclay)
Pierre André Hallé (LPP, Université Sorbonne Nouvelle – Paris 3)
Olivier Hamon (Syllabs)
Thierry Hamon (LIMSI, Université Paris-Saclay, CNRS, Université Sorbonne Paris Nord)
Bernard Harmegnies (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
Nabil Hathout (CLLE, CNRS)
Amir Hazem (LS2N, Université de Nantes)
Nicolas Hernandez (LS2N, Université de Nantes)
Fabrice Hirsch (Praxiling, Université Paul Valéry Montpellier 3)
Thomas Hueber (GIPSA-lab, CNRS)
Kathy Huet (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)

Stéphane Huet (LIA, Université d'Avignon)
Mathilde Hutin (LIMSI, Université Paris-Saclay, CNRS, Université Sorbonne Paris Nord)
Irina Illina (LORIA, Université de Lorraine)
Christine Jacquin (LS2N Université de Nantes)
Adèle Jatteau (STL, UMR 8163, Université de Lille, CNRS)
Denis Jouvet (LORIA, Inria)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre)
Takeki Kamiyama (LPP, Université Paris 8 Vincennes-Saint-Denis)
Hannah King (CLILLAC-ARP, Université Paris Diderot)
Olivier Kraif (Université Grenoble Alpes)
Matthieu Labeau (Telecom Paris)
Mathieu Lafourcade (LIRMM, Université de Montpellier)
Mohamed Lahrouchi (SFL, UMR 7023, CNRS Université Paris 8)
Muriel Lalain (LPL, CNRS, Aix-Marseille Université)
Joseph Lark (Dictanova)
Thomas Lavergne (LIMSI, CNRS, Univ. Paris Sud, Université Paris Saclay)
Guillaume Le Berre (LORIA, Université de Lorraine)
Gwénolé Lecorvé (IRISA, Université de Rennes, CNRS)
Benjamin Lecouteux (Laboratoire Informatique de Grenoble)
Claire Lemaire (Université Grenoble Alpes)
Yves Lepage (Waseda University)
Joseph Le Roux (LIPN, Université de Paris 13)
Veronika Lux (ATILF, CNRS)
Paolo Mairano (STL, UMR 8163, Université de Lille)
Anna Marczyk (LPL, CNRS, Aix-Marseille Université)
Denis Maurel (Université François Rabelais de Tours)
Christine Meunier (LPL, CNRS, Aix-Marseille Université)
Alexis Michaud (LACITO, CNRS)
Richard Moot (LIRMM, CNRS)
Véronique Moriceau (LIMSI, CNRS)
Philippe Muller (IRIT, Université de Toulouse)
Alexis Nasr (LIF, Université de la Méditerranée)
Sylvain Navarro (CLLE-ERSS, CNRS)
Luka Nerima (Université de Genève)
Aurélie Névéol (LIMSI, CNRS, Université Paris-Saclay)
Jian-Yun Nie (Université de Montreal)
Damien Nouvel (INaLCO)
Nicolas Obin (IRCAM)
Yannick Parmentier (LORIA, Université de Lorraine)
Sebastian Peña Saldarriaga (Dictanova)
Marie Philippart de Foy (Université de Mons)
Myriam Piccaluga (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
Claire Pillot-Loiseau (LPP, UMR 7018, CNRS, Université Sorbonne Nouvelle – Paris 3)
Serge Pinto (LPL, CNRS, Aix-Marseille Université)
Agnès Piquard (LORIA, CNRS, Université de Lorraine)
Thierry Poibeau (LaTTiCe, CNRS)
Alain Polguère (ATILF Université de Lorraine)
Laurent Prévot (LPL, CNRS, Aix-Marseille Université)
Jean-Philippe Prost (LIRMM, Université de Montpellier)
Christian Raymond (IRISA, INSA de Rennes)
Christian Retoré (LIRMM, Université de Montpellier)
Albert Rilliard (LIMSI, CNRS, Université Paris-Saclay)
Virginie Roland (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
Sophie Rosset (LIMSI, CNRS, Université Paris-Saclay)
Véronique Sabadell (LPC, Aix Marseille Université)
Stéphane Schneider (INIST, CNRS)
Didier Schwab (Université Grenoble Alpes)
Pascale Sébillot (IRISA, INSA de Rennes)
Djamé Seddah (Almanach, Université Paris la Sorbonne)
Gilles Serasset (LIG, Université Grenoble Alpes)
Romain Serizel (LORIA, Université de Lorraine)
Kamel Smaïli (LORIA, Université de Lorraine)
Rudolph Sock (LiLPa, Université de Strasbourg)
Ludovic Tanguy (CLLE, CNRS)
Xavier Tannier (LIMICS, Sorbonne Université, INSERM)
Andon Tchechmedjiev (IMR, Mines Alès)
Juan-Manuel Torres-Moreno (LIA, Université d'Avignon)
Nicolas Turenne (LISIS, INRA)
Béatrice Vaxelaire (LiLPa, Université de Strasbourg)

Anne Vilain (GIPSA-lab, Université de Grenoble Alpes)
Coriandre Vilain (GIPSA-lab, Université de Grenoble Alpes)
Guillaume Wisniewski (LLF, Université de Paris)
Jane Wottawa (LIUM, Le Mans Université)
Yaru Wu (LPP, MoDyCo, Université Paris Nanterre)
Kossi Seto Yibokou (LiLPa, Université de Strasbourg)
François Yvon (LIMSI, CNRS, Université Paris-Sud)

Table des matières

Première partie

Articles longs

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Approche de génération de réponse à base de transformers

Imen Akermi Johannes Heinecke Frédéric Herledan
Orange / DATA-AI, Lannion, France
{imen.elakermi,johannes.heinecke,frederic.herledan}@orange.com

RÉSUMÉ

Cet article présente une approche non-supervisée basée sur les modèles Transformer pour la génération du langage naturel dans le cadre des systèmes de question-réponse. Cette approche permettrait de remédier à la problématique de génération de réponse trop courte ou trop longue sans avoir recours à des données annotées. Cette approche montre des résultats prometteurs pour l'anglais et le français.

ABSTRACT

Transformer based approach for answer generation

This paper presents an unsupervised approach for Natural Language Generation within the framework of question-answering task. This approach aims at solving the too-short or too-long answer generation problem without requiring annotated data. The proposed approach shows promising results for both English and French.

MOTS-CLÉS : systèmes de question-réponse, génération du langage naturel, analyse de dépendance.

KEYWORDS: question answering systems, natural language generation, dependency analysis.

1 Introduction

La popularité récente des assistants intelligents a accru l'intérêt pour les systèmes de question-réponse (SQR) qui sont devenus un élément central des échanges « Humain-Machine » puisqu'ils permettent aux utilisateurs d'avoir directement des réponses à leurs questions en langage naturel en utilisant leur propre terminologie sans avoir à parcourir une longue liste de documents pour trouver les réponses appropriées.

La plupart des travaux de recherche existants focalisent sur la complexité majeure de ces systèmes résidant dans le traitement et l'interprétation de la question qui exprime le besoin en information de l'utilisateur, sans pour autant donner de l'importance à la représentation de la réponse donnée en sortie. Généralement, la réponse est soit représentée par un ensemble de termes court répondant exactement à la question (cas des SQR qui extraient les réponses à partir de données structurées), soit par un passage d'un document qui indique la réponse exacte et qui peut intégrer d'autres informations inutiles ne relevant pas du contexte de la question posée.

La figure 1 présente un exemple de réponses données par deux systèmes différents.

Compte-tenu de la particularité des SQR qui extraient les réponses à partir de données structurées, les utilisateurs ne reçoivent, généralement, qu'une réponse courte et limitée à leurs questions comme c'est illustré par l'exemple de la figure 1. De plus, dans le cas où la question est posée oralement, ce type de représentation de réponse peut ne pas convenir aux attentes de l'utilisateur.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 2–11.
HAL : hal-02784751.

FIGURE 1 – Réponses affichées par deux assistants intelligents, GOOGLE ASSISTANT et ALEXA pour la question « *Qui est le directeur de thèse d'Albert Einstein ?* »

En effet, le type de réponse donné par le premier système peut être perçu comme trop bref ne rappelant pas le contexte de la question et pour le deuxième type de réponse, on a tendance à renvoyer un passage qui contient la réponse mais qui inclut également des informations qui peuvent être hors du contexte de la question et jugées par l'utilisateur comme inutile.

C'est dans ce contexte que nous présentons dans cet article une approche qui permet de générer une réponse concise en langage naturel pour le français et l'anglais. Cette approche fait partie d'un SQR que nous avons proposé dans Rojas Barahona *et al.* (2019) et que nous présenterons brièvement dans cet article.

2 État de l'art

Malgré l'abondance des travaux dans le domaine des SQR, la problématique de formulation des réponses a reçu très peu d'attention. Une première approche traitant indirectement cette tâche a été proposée dans Brill *et al.* (2001, 2002). En effet, les auteurs avaient pour but de diversifier les motifs possibles de réponses en permutant les termes de la question en vue de maximiser le nombre de documents extraits qui peuvent contenir la réponse. Une autre approche de représentation de réponse basée sur des règles de reformulation a été également proposée dans Agichtein & Gravano (2000) et Lawrence & Giles (1998) dans le contexte d'expansion de requêtes pour la recherche de documents et non pour l'extraction de la réponse exacte.

Le peu de travaux qui se sont intéressés à cette tâche dans le cadre des SQR l'ont adressée sous l'angle de la génération de résumé de texte (Ishida *et al.*, 2018; Iida *et al.*, 2019; Rush *et al.*, 2015; Chopra *et al.*, 2016; Nallapati *et al.*, 2016; Miao & Blunsom, 2016; See *et al.*, 2017; Oh *et al.*, 2016; Sharp *et al.*, 2016; Tan *et al.*, 2016; dos Santos *et al.*, 2016). La majorité de ces travaux n'ont considéré que les questions de causalité de type « *pourquoi* » où les réponses sont des paragraphes. Pour rendre ces réponses plus concises, ils procèdent à un compactage des paragraphes extraits.

D'autres approches (Kruengkrai *et al.*, 2017; Girju, 2003; Verberne *et al.*, 2011; Oh *et al.*, 2013) ont exploré cette tâche comme un problème de classification où il s'agit de prédire si un passage de texte pourra constituer une réponse à une question donnée.

Il faut noter que ces approches ont pour seul but de diversifier au maximum les formules de réponses possibles à des questions pour augmenter la probabilité d'extraire la bonne réponse et non pour générer une réponse qui soit conviviale pour l'utilisateur. Il faut également souligner que ces approches ne sont valables que pour les SQR qui génèrent les réponses sous forme d'extrait de texte et ne pourront pas être appliquées aux réponses courtes.

Notre approche de génération de réponse diffère de ces travaux dans le sens qu'elle est non supervisée, peut s'adapter à n'importe quel type de questions factuelles (à l'exception des questions de type *Pourquoi* et *Comment*) et qu'elle s'appuie que sur des données facilement accessibles et non annotées. En effet, nous partons de l'hypothèse intuitive qu'une réponse concise et facilement prononçable par un assistant intelligent n'est en fait qu'une reformulation de la question posée. Cette approche est intégrée à un système (Rojas Barahona *et al.*, 2019) qui permet d'extraire la réponse à une question à partir de données structurées.

Dans ce qui suit, nous détaillerons dans la section 3 l'approche que nous avons proposé pour la génération de réponse en Langage Naturel et nous évoquerons brièvement le SQR développé. Nous présenterons dans la section 4 les expérimentations que nous avons conduit pour évaluer cette approche. Nous conclurons dans la section 5 avec les limites relevées et les perspectives envisagées.

3 Approche de génération de réponse en langage naturel

L'approche de génération de réponse que nous décrivons dans cet article est un composant d'un SQR qui a été développé dans Rojas Barahona *et al.* (2019) et qui consiste à traduire une question en langue naturelle (français ou anglais) dans une représentation formelle pour ensuite transformer cette représentation formelle en une requête Sparql [1]. Grâce à la requête Sparql nous cherchons la réponse dans une base de connaissance RDF, dans notre cas Wikidata [2]. La réponse est toujours une liste d'URI ou de valeurs.

Bien que nous arrivons à trouver la réponse exacte à une question, sa représentation n'est pas conviviale pour l'utilisateur. De ce fait, nous proposons une approche non-supervisée qui intègre l'utilisation des modèles Transformers tels que BERT (Devlin *et al.*, 2019) et GPT (Radford *et al.*, 2018). Le choix d'une approche non supervisée émane du fait qu'il n'existe pas un corpus d'apprentissage associant une question à une réponse compacte, exhaustive et qui permettrait d'appliquer en mode supervisé une architecture neuronale End-to-End apprenant à générer une phrase répondant à une question.

Cette approche part du fait que nous avons déjà extrait la réponse exacte à une question posée. Nous supposons qu'une réponse bien formulée n'est que la reformulation de la question même associée à la réponse exacte. Cette approche est constituée de deux étapes fondamentales. La première étape consiste à effectuer une analyse de dépendance de la question en entrée et nous procédons dans une deuxième étape à la génération de la réponse.

3.1 Analyse en dépendance

Pour l'analyse en dépendance, nous utilisons une version améliorée de UdpipeFuture (Straka, 2018) qui était le système gagnant (3e avec les métriques *Labelled Attachment Score* (LAS) et *Content Word LAS* (CLAS) et 1er en utilisant la métrique *Morphology-Aware LAS* (MLAS)) de la tâche de l'analyse en dépendance (Zeman *et al.*, 2018). UdpipeFuture est un analyseur, qui fait l'étiquetage en partie de discours et la lemmatisation avec un LSTM. L'analyse en dépendance est faite avec un parser à base de graphes, inspiré de Dozat *et al.* (2017).

Notre modification consiste à intégrer les plongements contextuels à UdpipeFuture lors de l'apprentissage. Pour cela, nous nous sommes orientés vers BERT multilingue (Devlin *et al.*, 2019), XLM-R

1. https://www.w3.org/TR/sparql11-overview/
2. https://www.wikidata.org/

(Conneau *et al.*, 2019) (pour l'anglais et français), RoBERTA (Liu *et al.*, 2019) (pour l'anglais) FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2019) (pour le français) lors de l'apprentissage des treebanks French-GSD et English-EWT [3], issues du projet Universal Dependencies (UD) (Nivre *et al.*, 2016) [4]. L'ajout des plongements contextuels a significativement augmenté les résultats pour les trois métriques, LAS, CLAS et MLAS (cf. table 1).

français (Fr-GSD)			
embeddings	MLAS	CLAS	LAS
Straka (2018)	77,29	82,49	85,74
FlauBERT	79,53	84,16	87,98
BERT	81,64	86,21	89,68
CamemBERT	82,17	86,45	89,67
XLM-R	**82,62**	**86,94**	**89,82**

anglais (En-EWT)			
embeddings	MLAS	CLAS	LAS
Straka (2018)	74,71	79,14	82,51
BERT	81,16	85,89	88,63
RoBERTA	82,38	86,89	89,40
XLM-R	**82,91**	**87,24**	**89,54**

TABLE 1 – Analyse en dépendance du français et l'anglais (UD v2.2), les meilleurs résultats en gras

Néanmoins, pour l'analyse en dépendances des questions simples de type quiz, les deux treebanks UD (En-EWT et Fr-GSD) ne sont pas adaptés, car leurs corpus d'apprentissage ne contiennent pas ou très peu de questions [5]. Les résultats de l'analyse des questions avec des modèles appris sur les données standard de UD sont montrés en table 2.

français (Fr-GSD)			
embeddings	MLAS	CLAS	LAS
BERT	60,52	73,04	79,27
CamemBERT	**61,32**	**75,26**	**80,49**
FlauBERT	58,09	70,96	78,40
Word2Vec	59,83	74,43	80,14
XLM-R	59,23	73,52	79,27

anglais (En-EWT)			
embeddings	MLAS	CLAS	LAS
BERT	**80,45**	**88,02**	**90,58**
RoBERTa	80,68	89,17	91,49
XLM-R	80,68	89,42	91,88

TABLE 2 – Analyse des questions avec des modèles appris sur UD sans modifications

Afin d'améliorer l'analyse, nous avons enrichi les treebanks d'apprentissage Fr-GSD et En-EWT en annotant 309 questions anglaises des challenges QALD7 (Usbeck *et al.*, 2017) et QALD8 [6] (ainsi 91 questions pour le test) en supprimant les doublons. Pour le français, nous avons traduit des questions issues de QALD7, et formulé des questions nous-mêmes (66 test, 267 apprentissage). Les annotations ont été effectuées par deux linguistes avec le guide d'annotation du project Universal Dependencies [7] et la documentation des treebanks Fr-GSD (pour les questions françaises) et En-EWT (pour l'anglais).

Comme la table 3 le montre, la qualité de l'analyse augmente considérablement, les embeddings CamemBERT (pour le français) et BERT (anglais) ont de nouveau le meilleurs impact.

Nous nous appuyons sur la version UdpipeFuture que nous avons améliorée avec BERT/CamemBERT et qui donne les meilleurs résultats en terme d'analyse en dépendance pour procéder au découpage de la question en fragments textuels (appelés également *chunks*) : $Q = \{c_1, c_2, \ldots, c_n\}$.

3. Comme la campagne d'évaluation Conll2018, nous utilisons la version 2.2 des treebanks UD pour la comparaison
4. https://universaldependencies.org/
5. Il existe également un treebank de questions, French-FQB (Seddah & Candito, 2016), mais la plupart des questions de ce treebank sont plutôt des questions longues, et peu similaires aux questions « typiques » du quiz.
6. https://github.com/ag-sc/QALD
7. https://universaldependencies.org/guidelines.html

français (Fr-GSD)

embeddings	MLAS	CLAS	LAS
BERT	91,20	96,10	97,55
CamemBERT	**92,12**	**97,37**	**98,26**
FlauBERT	90,53	94,74	96,86
Word2Vec	90,88	95,79	97,21
XLM-R	91,23	96,14	97,56

anglais (En-EWT)

embeddings	MLAS	CLAS	LAS
BERT	**84,85**	**91,92**	**94,24**
RoBERTa	83,08	91,67	93,85
XLM-R	83,08	90,66	93,59

TABLE 3 – Analyse des questions avec des corpus d'apprentissage UD de base enrichis de questions

Si on prend l'exemple de la question *Quel est le parti politique du maire de Paris ?*, l'ensemble des fragments textuels serait *Q = {Quel, est, le parti politique du maire de Paris}*.

3.2 Génération de la réponse

Lors de cette phase, nous procédons d'abord à un premier test de l'ensemble Q pour vérifier si le fragment textuel qui contient un marqueur de question (exp : *quel*, *quand*, *qui* etc.) représente le sujet *nsubj* dans la question analysée. Si c'est le cas, nous remplaçons tout simplement ce fragment textuel par la réponse que nous avons identifié précédemment. Reprenons l'exemple de la question *Quel est le parti politique du maire de Paris ?*, le système détecte automatiquement que le fragment textuel contenant le marqueur de question *Quel* représente bien le sujet et sera donc remplacé directement par la réponse exacte *Le Parti Socialiste*. Par suite, la réponse compacte générée sera *Le Parti Socialiste est le parti politique du maire de Paris*.
Dans le cas contraire, nous procédons à la suppression du fragment textuel contenant le marqueur de question que nous avons détecté et nous ajoutons la réponse R à l'ensemble Q : $Q = \{c_1, c_2, \ldots, c_{n-1}, R\}$

À partir de l'ensemble de fragments textuels Q, nous générons par permutation toutes les structures de réponses possibles qui peuvent former la phrase répondant à la question traitée :

$S = \{s_1(R, c_1, c_2, \ldots, c_{n-1}), s_2(c_1, R, c_2, \ldots, c_{n-1}), \ldots, s_m(c_1, c_2, \ldots, c_{n-1}, R)\}$

Ces structures seront évaluées par un Modèle de Langue (LM) basé sur des modèles Transformers qui permettra d'extraire la séquence de fragments textuels la plus probable qui servira de réponse :

$$structure^* = s \in S; p(s) = \underset{s_i \in S}{\operatorname{argmax}}\, p(s_i)$$

Une fois la structure, qui représentera la réponse à la question traitée, est identifiée, l'étape de génération du terme manquant est initiée. En effet, nous supposons qu'il pourrait y avoir un terme qui ne figure pas nécessairement dans la question ou dans la réponse mais qui est par contre nécessaire à la génération d'une bonne structure grammaticale de la réponse.

Par suite, pour prédire ce terme manquant nous faisons appel à BERT en tant que modèle de langue « *masqué* ». Dans le cas ou BERT renvoie une séquence de caractère non alphabétique, nous concluons que la structure optimale, telle qu'elle était prédite par le LM, ne nécessite pas d'être complétée par un terme supplémentaire.

L'exemple suivant illustre les différentes étapes de l'approche proposée :

Question : *Dans quel pays se trouve le Lac de Garde ?*

1. Découpage de la question en fragments textuels (à partir de l'analyse en dépendance obtenu avec UDPipe) → { *Dans quel pays*, *se trouve*, *le Lac De Garde* }
2. Suppression des marqueurs de question (*Dans quel pays*) → { *se trouve*, *le Lac De Garde* }
3. Ajout de la réponse brute → { *se trouve*, *le Lac de Garde*, *Italie* }
4. Génération des structures de réponses possibles
   → — *se trouve, le Lac de Garde, Italie*
   — *Italie, se trouve, le Lac de Garde*
   — ...
5. Évaluation des structures par un modèle de langage → *structure** = {*le Lac de Garde*, *se trouve*, *Italie*}
6. Génération des termes manquants à *structure** (avec BERT pour l'anglais ou CamemBERT pour le français)
   → *le Lac de Garde se trouve [terme manquant] Italie* : « en »

Réponse : *le Lac de Garde se trouve en Italie.*

4 Évaluation

Les corpus de tests existants pour l'évaluation des SQR, sont plus adaptés aux systèmes qui génèrent la réponse exacte à la question et donc une réponse courte ou sont plus axés vers la tâche de *Machine Reading Comprehension* où la réponse consiste en un passage de texte contenant la réponse exacte. Nous avons créé un jeu de données qui consiste à associer des questions extraites du corpus QALD-7 challenge (Usbeck *et al.*, 2017) avec des réponses en langage naturel qui ont été définies manuellement par un linguiste et que nous avons revues individuellement. Ce corpus appelé QueReo consiste en 150 questions avec les réponses exactes extraites par le SQR que nous avons décrit précédemment. On note en moyenne trois réponses possibles en langage naturel pour chaque question. Ce corpus existe en versions française et anglaise.

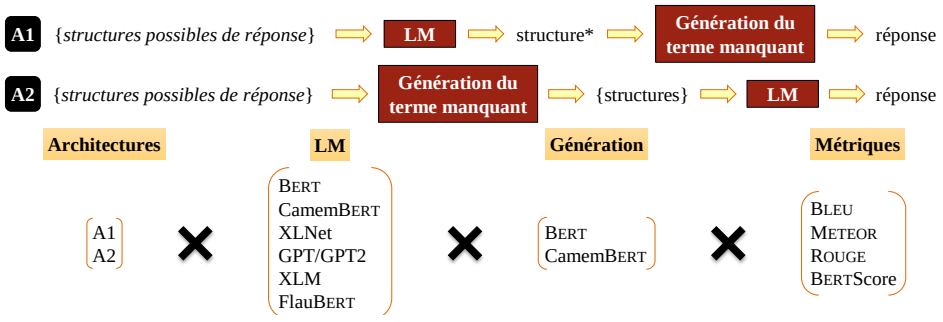


FIGURE 2 – Cadre d'expérimentation

Comme illustré dans la figure 2, deux architectures possibles de l'approche que nous proposons pour la génération de réponse ont été évaluées. La première architecture *A1* consiste à générer toutes les structures possibles de réponses pour les faire évaluer après par un LM qui permettrait la sélection de

la réponse optimale puis de générer le terme manquant à cette structure. L'architecture *A2* commence à générer le terme manquant pour chaque structure dans *Q* qui seront ensuite évaluer par le LM. Pour le moment, nous supposons qu'il y a seulement 1 terme manquant par structure.

Pour évaluer l'approche proposée, nous avons opté pour trois métriques N-gram (BLEU, METEOR et ROUGE) utilisées dans la littérature pour évaluer ce type de tâche et la métrique BERTScore qui exploite les plongements lexicaux pré-entraînés de BERT pour calculer la similarité entre la réponse générée et la réponse de référence. Pour pouvoir comparer les différentes variantes de l'approche, nous nous sommes référés au test de Friedman (Milton, 1939) qui permet de détecter les écarts de performances entre plusieurs modèles évalués par plusieurs métriques en se basant sur les rangs moyens.

La table 4 (corpus français) et la table 5 (corpus anglais) ne présentent que les résultats obtenus pour les cinq meilleurs modèles selon le classement du test de Friedman. Les modèles de langue utilisés sont adaptés selon la langue du corpus mis en test. Vanté par ses mérites en tant que modèle génératif très puissant entraîné sur un très large corpus de données, le modèle GPT a été également testé avec le corpus français pour voir s'il arrive à détecter la meilleure structure à choisir pour une question. Les valeurs mises entre parenthèses représentent le rang d'un modèle selon la métrique utilisée.

rang moyen	arch.	LM	modèle de génération	%BleuS score	%BleuS rang	%MeteorS score	%MeteorS rang	%RougeS score	%RougeS rang	%BertS score	%BertS rang
8,5	A1	BERT	CamemBERT	68,21	14	95,90	2	84,48	11	94,45	7
12	A1	GPT	CamemBERT	68,63	13	96,04	1	84,04	15	93,85	19
13,5	A1	GPT	BERT	70,39	6	94,90	38	85,70	4	94,61	6
13,5	A1	BERT	BERT	69,68	9	95,03	35	85,26	6	94,79	4
13,75	A2	BERT	BERT	75,08	1	94,78	48	85,91	3	94,98	3
16	A2	GPT	CamemBERT	69,37	10	95,73	13	83,59	21	93,82	20
18,5	A1	XLM	BERT	67,02	19	95,04	34	84,89	7	94,12	14
21,75	A2	XLM	BERT	68,67	12	94,44	59	84,67	8	94,42	8
25,25	A2	BERT	CamemBERT	63,82	52	95,82	7	83,13	26	93,98	16
26	A2	XLNET	BERT	70,87	4	94,05	74	84,65	9	93,98	17

TABLE 4 – Classement des modèles selon le test de Friedman - Corpus français

Selon la table 4, c'est CamemBERT en tant que modèle de génération du terme manquant à la réponse, associé à BERT et GPT en tant que modèles de langue, qui se classe en premières places selon la moyenne des rangs des quatre métriques d'évaluation, pour traiter les questions du corpus français. Comme attendu, le modèle GPT, malgré le fait qu'il est principalement entraîné sur un corpus de texte écrit en anglais, arrive quand même à prédire la meilleure structure de réponse pour du texte écrit en français. Ce modèle, dans sa deuxième version améliorée GPT2, arrive également à se hisser aux premiers rangs pour le corpus de questions anglais, associé à BERT comme modèle génératif du terme manquant.

La différence de performance selon l'architecture adoptée varie aussi selon la langue dans laquelle la question est écrite. En effet, nous remarquons que par rapport aux cinq premiers rangs, c'est l'architecture *A1* faisant évaluer les structures par le modèle de langue avant de générer le terme manquant, qui donne les meilleurs résultats pour le corpus français, alors que c'est la deuxième architecture *A2* qui est la plus pertinente pour les questions en anglais.

rang moyen	arch.	LM	modèle de génér.	%BleuS score	%BleuS rang	%MeteorS score	%MeteorS rang	%RougeS score	%RougeS rang	%BertS score	%BertS rang
1,25	A2	GPT2	BERT	77,82	1	94,41	1	93,48	1	97.30	2
2	A1	GPT	BERT	77,56	2	94,02	3	93,22	2	97,51	1
5	A2	XLM	BERT	76,90	5	94,14	2	92,17	7	97,16	6
8	A2	GPT	BERT	75,26	14	93,72	10	92,68	3	97,19	5
9,5	A1	XLM	BERT	76,71	7	93,92	4	91,63	12	96,79	15
12,75	A1	XLNET	BERT	74,96	15	93,77	7	91,3	17	96,8	12
16,5	A1	GPT2	BERT	75,75	10	93,49	21	91,11	24	96,87	11
17	A2	XLNET	BERT	74,19	17	93,57	15	91,2	19	96,78	17
20,5	A2	BERT	BERT	74,56	16	93,75	8	90,51	35	96,67	23
22	A1	BERT	BERT	75,37	12	93,41	25	91,12	22	96,59	29

TABLE 5 – Classement des modèles selon le test de Friedman - Corpus anglais

Il faut noter que dû à la spécificité de l'approche que nous proposons et qui fait usage des termes de la question pour générer une réponse en langage naturel, nous avons obtenu des scores relativement élevés. Nous pensons que la mesure de corrélation avec une évaluation humaine permettrait de déterminer la métrique d'évaluation la plus appropriée.

5 Conclusion

Nous avons présenté dans cet article une approche non-supervisée qui se base sur des modèles Transformer pour la génération de réponse en langage naturel dans le cadre des systèmes de question-réponse. L'évaluation que nous avons effectuée prouve que cette approche est prometteuse. Nous envisageons d'utiliser cette approche pour construire des corpus d'apprentissage de type question-réponse en langage naturel, qui permettraient d'entraîner des approches neuronales de type end-to-end.

Références

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, p. 85–94.

BRILL E., DUMAIS S. & BANKO M. (2002). An analysis of the askmsr question-answering system. In *Proceedings of the ACL conference on Empirical methods in natural language processing*, p. 257–264 : Association for Computational Linguistics.

BRILL E., LIN J., BANKO M., DUMAIS S. & NG A. (2001). Data-intensive question answering. In *In Proceedings of the Tenth Text REtrieval Conference (TREC)*, p. 393–400.

CHOPRA S., AULI M. & RUSH A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 93–98.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRACE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint : 1911.02116.

DEVLIN J., CHANG M.-W. C., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language under-standing. In *NAACL*, p. 4171–4186, Minneapolis : Association for Computational Linguistics.

DOS SANTOS C., TAN M., XIANG B. & ZHOU B. (2016). Attentive pooling networks. arXiv preprint : 1602.03609.

DOZAT T., QI P. & MANNING C. D. (2017). Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. In *Proceedings of the CoNLL 2017 Shared Task. Multilingual Parsing from Raw Text to Universal Dependencies*, p. 20–30, Vancouver, Canada : Association for Computational Linguistics.

GIRJU R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, p. 76–83 : Association for Computational Linguistics.

IIDA R., KRUENGKRAI C., ISHIDA R., TORISAWA K., OH J.-H. & KLOETZER J. (2019). Exploiting background knowledge in compact answer generation for why-questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, p. 142–151.

ISHIDA R., TORISAWA K., OH J.-H., IIDA R., KRUENGKRAI C. & KLOETZER J. (2018). Semi-distantly supervised neural model for generating compact answers to open-domain why questions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

KRUENGKRAI C., TORISAWA K., HASHIMOTO C., KLOETZER J., OH J.-H. & TANAKA M. (2017). Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*.

LAWRENCE S. & GILES C. L. (1998). Context and page analysis for improved web search. *IEEE Internet computing*, **2**(4), 38–46.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *LREC 2020*. arXiv preprint : 1912.05372.

LIU Y., OTT M., GOYAL N., DU, JINGFEI ADN JOSHI M., CHEN D., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv preprint : 1907.11692.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2019). CamemBERT : a Tasty French Language Model. arXiv preprint 1911.03894.

MIAO Y. & BLUNSOM P. (2016). Language as a latent variable : Discrete generative models for sentence compression. *EMNLP 2016*.

MILTON F. (1939). A correction : The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association. American Statistical Association*, **34**(205), 109.

NALLAPATI R., ZHOU B., DOS SANTOS C., GÜLÇEHRE Ç., XIANG B. *et al.* (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *CoNLL 2016*, p. 280–290.

NIVRE J., MARNEFFE M.-C. D., GINTER F., GOLDBERG Y., GOLDBERG Y., HAJIČ J., D. M. C., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal Dependencies v1 : A Multilingual Treebank Collection. In *the tenth international conference on Language Resources and Evaluation*, p. 23–38, Portorož, Slovenia : European Language Resources Association.

OH J.-H., TORISAWA K., HASHIMOTO C., IIDA R., TANAKA M. & KLOETZER J. (2016). A semi-supervised learning approach to why-question answering. In *Thirtieth AAAI Conference on Artificial Intelligence*.

OH J.-H., TORISAWA K., HASHIMOTO C., SANO M., DE SAEGER S. & OHTAKE K. (2013). Why-question answering using intra-and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1733–1743.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

ROJAS BARAHONA L. M., BELLEC P., BESSET B., DOS SANTOS M., HEINECKE J., ASADULLAH M., LEBLOUCH O., LANCIEN J.-Y., DAMNATI G., MORY E. & HERLÉDAN F. (2019). Spoken Conversational Search for General Knowledge. In *SIGdial Meeting on Discourse and Dialogue*, p. 110–113, Stockholm : Association for Computational Linguistics.

RUSH A. M., CHOPRA S. & WESTON J. (2015). A neural attention model for abstractive sentence summarization. arXic preprint : 1509.00685.

SEDDAH D. & CANDITO M. (2016). Hard Time Parsing Questions : Building a QuestionBank for French. In *the tenth international conference on Language Resources and Evaluation*, Portorož, Slovenia : European Language Resources Association.

SEE A., LIU P. J. & MANNING C. D. (2017). Get to the point : Summarization with pointer-generator networks. arXiv preprint : 1704.04368.

SHARP R., SURDEANU M., JANSEN P., CLARK P. & HAMMOND M. (2016). Creating causal embeddings for question answering with minimal supervision. *EMNLP 2016*.

STRAKA M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 197–207, Brussels : Association for Computational Linguistics.

TAN M., DOS SANTOS C., XIANG B. & ZHOU B. (2016). Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 464–473.

USBECK R., NGOMO A.-C. N., HAARMANN B., KRITHARA A., RÖDER M. & NAPOLITANO G. (2017). 7th Open Challenge on Question Answering over Linked Data (QALD-7). In M. DRAGONI, M. SOLANKI & E. BLOMQVIST, Éds., *Semantic Web Challenges*, p. 59–69, Cham : Springer International Publishing.

VERBERNE S., VAN HALTEREN H., THEIJSSEN D., RAAIJMAKERS S. & BOVES L. (2011). Learning to rank for why-question answering. *Information Retrieval*, **14**(2), 107–132.

ZEMAN D., HAJIČ J., POPEL M., POTTHAST M., STRAKA M., GINTER F., NIVRE J. & PETROV S. (2018). CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies. In D. ZEMAN & J. HAJIČ, Éds., *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 1–21, Brussels : Association for Computational Linguistics.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie

Maxime Amblard[1] Chloé Braud[2] Chuyuan Li[1]
Caroline Demily[3] Nicolas Franck[3] Michel Musiol[1,4]
(1) LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
{maxime.amblard, chuyuan.li}@univ-lorraine.fr
(2) IRIT, CNRS, Toulouse
chloe.braud@irit.fr
(3) Centre Hospitalier le Vinatier & UMR 5229, CNRS - Univeristé Lyon 1, Lyon, France
{caroline.demily, nicolas.franck}@ch-le-vinatier.fr
(4) ATILF, UMR 7118, Université de Lorraine, CNRS, 54000 Nancy, France
michel.musiol@univ-lorraine.fr

RÉSUMÉ

Nous présentons des expériences visant à identifier automatiquement des patients présentant des symptômes de schizophrénie dans des conversations contrôlées entre patients et psychothérapeutes. Nous fusionnons l'ensemble des tours de parole de chaque interlocuteur et entraînons des modèles de classification utilisant des informations lexicales, morphologiques et syntaxiques. Cette étude est la première du genre sur le français et obtient des résultats comparables à celles sur l'anglais. Nos premières expériences tendent à montrer que la parole des personnes avec schizophrénie se distingue de celle des témoins : le meilleur modèle obtient une exactitude de 93,66%. Des informations plus riches seront cependant nécessaires pour parvenir à un modèle robuste.

ABSTRACT

Investigating Learning Methods Applied to Language Specificity of Persons with Schizophrenia.

We present experiments to automatically identify patients with symptoms of schizophrenia in controlled conversations between patients and psychotherapists. We merge the speech turns of each interlocutor and train basic classifiers with these data, using lexical, morphological and syntactic features. This study is the first of its kind in French and obtains results comparable to the English state-of-the-art. Our first experiments highlight that the speech of patient with schizophrenia differs from control one : the best model obtains an accuracy of 93.66%. However, richer descriptions will be needed to produce an applicable and reliable model.

MOTS-CLÉS : Dialogue, schizophrénie, apprentissage automatique.

KEYWORDS: Dialog, schizophrenia, machine learning.

1 Introduction

La schizophrénie est définie comme un trouble mental sévère (Association *et al.*, 2015). Cette maladie s'accompagne de symptômes très variables, les plus manifestes étant les idées délirantes, les halluci-

Volume 2 : Traitement Automatique des Langues Naturelles, pages 12–26.
HAL : hal-02784752.

nations et le discours désorganisé. De nombreuses études convergent aujourd'hui vers l'hypothèse selon laquelle il existe une composante génétique de la pathologie schizophrénique, à tout le moins comme une condition nécessaire et non suffisante. Son étiologie demeure toutefois complexe. Pour autant, les avancées empiriques, théoriques et méthodologiques non négligeables en psychopathologie cognitive, en pragmatique linguistique et les progrès considérables, en particulier dans le domaine de l'électrophysiologie et de l'imagerie cérébrale, contribuent à une meilleure caractérisation des troubles schizophréniques (Besche-Richard *et al.*, 2018). Selon l'Organisation Mondiale de la Santé, les troubles schizophréniques touchent environ 1% de la population mondiale. En outre, les troubles cognitifs affectent 70 à 80% de la population atteinte de troubles schizophréniques (Potvin *et al.*, 2017). Pour autant les symptômes de cette pathologie, ou leurs effets, impactent largement les pratiques langagières qui apparaissent ainsi comme un bon angle pour aborder la pathologie.

L'identification automatique de patients manifestant des symptômes de la schizophrénie à partir de la production langagière, écrite ou orale, est un enjeu important dans le domaine de la santé, car cela pourrait constituer une aide décisive vers un diagnostic pour les médecins. Par ailleurs, étudier des cas de parole affectée permet d'améliorer notre compréhension du fonctionnement du langage en général, et cela devrait également permettre d'adapter des systèmes de TAL à des parties de la population qui, en plus de souffrir d'un trouble psychiatrique potentiellement désociabilisant, présentent un usage de la langue qui dévie des modèles dont nous disposons.

A terme, il serait intéressant de modéliser cet usage de la langue dans le dialogue, car c'est dans les interactions que nous attendons les déviations les plus importantes. C'est notamment les perspectives du projet SLAM [1] et son corpus en français, fondé sur des conversations contrôlées entre personnes avec schizophrénie ou témoins et psychologues qui ont été filmées et enregistrées puis retranscrites. Ces dialogues servent de base à la présente étude. Cependant, comme première approximation, nous avons choisi de modéliser le problème à partir de quasi-monologues, c'est-à-dire une fusion de l'ensemble des tours de parole de chaque interlocuteur pour chacun des dialogues considérés. À partir de cet ensemble de tour de parole, nous construisons des systèmes de classification permettant d'identifier les personnes avec schizophrénie. Cette fusion permet de construire des instances de classification contenant plus d'information qu'un simple tour de parole. Ce choix, qui exclut la parole du praticien lors de la classification, est également justifié par la spécificité de ces entretiens dans lesquels le psychologue n'est pas personnellement investi : sa mission est de maintenir l'interaction en relançant l'échange pour qu'il se poursuive le plus longtemps possible. Nous entraînons des modèles de classification sur les monologues et montrons qu'il est possible d'identifier des personnes avec schizophrénie à partir d'informations lexico-syntaxiques dans notre corpus avec une exactitude de 93,66%.

2 État de l'art

Les développements linguistiques sur le discours schizophrénique émergent avec les études de Chaika (1974); Fromkin (1975). Depuis, des indices de moins bonne maîtrise des catégories morpho-syntaxiques (*Part-Of-Speech*, *POS*) et de la syntaxe ont été étudiés (Andreasen, 1979; Fraser *et al.*, 1986; Hoffman & Sledge, 1988). Cependant, il est souvent difficile de caractériser ce qui relève de la pathologie de ce qui relève de la médication. De plus, les éléments mis en avant sont peu discriminants. Les patients ont une maîtrise moins bonne de ces niveaux linguistiques, mais ce sont leurs capacités

1. https://team.inria.fr/semagramme/fr/slam/

cognitives en général qui semblent dégradées (Docherty *et al.*, 1996). Il apparaît alors que travailler sur le discours des personnes avec schizophrénie c'est aussi travailler sur les capacités cognitives.

La détection automatique de troubles schizophréniques est un champ de recherche actif, avec des études qui se concentrent principalement sur deux types de caractéristiques : signaux biomédicaux du type électro-encéphalographie (EEG) et images de résonance magnétique (IRM) (Greenstein *et al.*, 2012; Sabeti *et al.*, 2011). Les études fondées sur les données langagières sont encore assez rares, mais un courant de recherche émerge ces dernières années dans le domaine de l'identification automatique de différents troubles comme la dépression, seule (Pestian *et al.*, 2017) ou associée à d'autres troubles comme le syndrome post-traumatique (PTSD) (Pedersen, 2015) et les pré-symptômes de la maladie d'Alzheimer (Jarrold *et al.*, 2010).

Quelques études ont été spécifiquement consacrées à la schizophrénie. Dans la première étude dédiée à ce problème, Strous *et al.* (2009) utilisent des écrits de personnes avec schizophrénie pour construire des systèmes de classification fondés sur des informations lexicales et obtiennent une exactitude de 83,3%. Ils observent des traits particuliers aux personnes avec schizophrénie comme un usage plus restreint de prépositions et une sur-représentation de la première personne. Ensuite, plusieurs études ont été menées à partir de messages écrits sur Twitter par des personnes s'auto-identifiant avec schizophrénie. Mitchell *et al.* (2015) ont collecté des données pour 174 patients (3200 *tweets*) et testé différents ensembles de traits lexicaux, comme des catégories sémantiques issues de lexique ou des clusters Brown : ils présentent des systèmes de classification (SVM) avec au mieux 82,3% d'exactitude. Cette étude est étendue dans (Birnbaum *et al.*, 2017) à partir de 1,9 millions de *tweets* collectés pour 146 patients. Ils obtiennent également des scores hauts, avec 90,0% d'exactitude, avec des traits lexicaux, notamment des catégories du LIWC (Pennebaker *et al.*, 2001). Ils observent comme précédemment une utilisation accrue des pronoms de première personne, ainsi que les termes appartenant au champ lexical de la santé.

Enfin, Kayi *et al.* (2017); Allende-Cid *et al.* (2019) ont exploré d'autres représentations notamment en se fondant sur des informations morpho-syntaxiques et syntaxiques. Allende-Cid *et al.* (2019) utilisent des textes narratifs rédigés par les sujets (13 personnes avec schizophrénie et 50 témoins) et démontrent que les catégories morpho-syntaxiques permettent déjà des performances assez hautes, 82,8% de F1. Kayi *et al.* (2017) ont eux étudié à la fois des *tweets* (174 sujets par groupe) et des textes narratifs (environ 95 sujets par groupe). Ils utilisent des informations syntaxiques, sémantiques (rôles sémantiques, LDA, clusters) et pragmatiques (sentiments), ces deux dernières se révélant particulièrement utiles pour les données issues de réseaux sociaux (81,65% en F1), tandis que, comme dans l'étude précédente, les traits morpho-syntaxiques se révèlent efficaces sur les textes narratifs (69,76% de F1).

Toutes ces études mettent en jeu des corpus différents dont les données ne sont forcément disponibles, les comparaisons entre études sont donc difficiles. Elles mettent cependant clairement en lumière le fait que les personnes atteintes de schizophrénie présentent des spécificités dans leur usage de la langue à différents niveaux, et nous testons également dans cette étude les informations lexicales, morpho-syntaxiques et syntaxiques mis en oeuvre dans des modèles de classification mais sur des données dialogiques et en français.

Différentes études ont déjà proposé des analyses du corpus du projet SLAM, sur les disfluences, les *POS* et la lexicographie (Amblard *et al.*, 2015; Amblard & Fort, 2014), ainsi que sur des aspects discursifs (Rebuschi *et al.*, 2014), notamment en s'appuyant sur une représentation de l'interaction adaptée de la S-DRT (Asher *et al.*, 2003). Ces différents travaux permettent d'avoir une modélisation fine de l'interaction, et des caractérisations sur ces niveaux linguistiques. Dans la présente étude, nous

reprenons l'analyse des usages langagiers à travers le développement d'un système de classification.

3 Origine et caractérisation des données

Dans la suite, nous appelons "entretien" le dialogue originel entre une personne avec schizophrénie (ou un témoin) et un psychologue, "document" les transcriptions de ces entretiens. Ces entretiens sont constitués des tours de parole (TDP) de chaque locuteur, et nous appelons cTDP la concaténation de tous les tours de parole d'un locuteur au sein d'un document.

Le corpus SLAM : Le corpus a été développé dans le cadre du projet SLAM[2]. Les entretiens sont réalisés en milieu hospitalier auprès de patients diagnostiqués par des médecins-psychiatres et des psychologues de l'institution d'accueil. L'entretien est associé à des tests neuropsychologiques permettant de mesurer les aptitudes des patients sur différents plans (capacité de mémoire de travail, fluence verbale, attention, vitesse motrice, fonctions exécutives, *etc*). Les interactions verbales des patients avec une psychologue sont par ailleurs enregistrées réalisent au cours d'un entretien semi dirigé. La participation des patients est libre et les éléments recueillis lors de l'expérience ne sont pas utilisés par l'équipe médicale pour le suivi du patient. Il y a donc une vraie liberté dans l'entretien. Les thématiques abordées restent simples (quotidien du patient, historique médical, anamnèse avant l'hospitalisation, *etc*). Ces entretiens sont enregistrés avec un double système d'eye-tracker permettant de travailler en parallèle sur les mouvements de regard des deux locuteurs. Les entretiens sont conduits par une psychologue qui n'est pas engagée personnellement dans le dialogue. Il ne s'agit donc pas d'une situation d'interaction symétrique du quotidien, la parole du patient se rapproche d'un monologue. Ceci explique notre choix d'extraire la production langagière du locuteur et de l'isoler comme un tout cohérent.

Description des données : Le corpus est composé de 41 documents, 18 personnes avec schizophrénie et 23 témoins pour le groupe contrôle. Une seule psychologue interroge ces 41 sujets. Chacun de ces groupes contient 15 sujets masculins, le reste (donc 3 et 8) étant féminin. Cette répartition présente donc un biais. Il est admis qu'il existe des différences significatives selon le genre (aspects cliniques et paracliniques) (Douki Dedieu *et al.*, 2012). En ce moment, la majorité des études porte surtout sur des sujets mâles et nous pensons que ces différences devront être prises en compte dans la démarche diagnostique.

De manière non surprenante, les personnes avec schizophrénie ont, en moyenne, le même nombre de TDP par document que la psychologue (199,9 vs. 200,2). Par contre, ils parlent plus (2675,5 mots par document) et leurs phrases sont plus longues (13,4 mots par phrase) par rapport à la psychologue (1814,6 mots par document, 9,1 mots par phrase). Les témoins s'expriment sensiblement plus (342 TDP et 3305 mots par document) avec des phrases plus courtes (10,5 mots par phrase). Les personnes avec schizophrénie ont par ailleurs un taux plus élevé d'utilisation de mots grammaticaux (n'appartenant pas aux catégories : nom, verbe, adverbe ou adjectif) que la psychologue ou les témoins (SCZ 56% vs. témoins 51% vs. psychologue 50%).

2. Pour Amblard *et al.* (2014), le contenu des entretiens donnait de nombreux éléments géographiques et biographiques du patient et son entourage que l'anonymisation ne suffit pas à rendre opaque, la distribution des données est difficile.

Parmi les mots les plus utilisés, que ce soit chez les personnes avec schizophrénie ou les témoins, nous observons plusieurs thématiques :

— Pour les personnes avec schizophrénie : typiquement des mots liés à la douleur comme "maladie", "hospitalisation", "hallucinations". Ce point correspond au label *Catastrophe* parmi les *top semantic features* observés par Kayi *et al.* (2017) qui présentent, dans leur étude, des traits linguistiques prédictifs des personnes avec schizophrénie à l'écrit. De ce point de vue, l'analyse empirique donne corps au contexte conversationnel au sein duquel les patients étaient amenés indirectement à évoquer les prémices de l'entrée dans la maladie. Nous constatons également une utilisation plus fréquente du mot "je" chez les personnes avec schizophrénie.

— Pour les témoins : des mots liés à l'éducation comme "master", "thèse", "licence" et à la psychologie comme "psychiatre" et "psychologue" ressortent significativement. Il se trouve que les sujets témoins sont majoritairement des étudiants de 1er ou 2me année inscrits dans une filière de sciences humaines.

Ces différences dépendent en grande partie des thématiques de conversation choisis par les sujets. Les patients sont censés parler de leur quotidien qui recoupe de fait des réalités différentes ce qui explique les différences de champ lexical. Ceci pourrait donc correspondre à un biais dans nos données.

4 Expériences

Nous présentons les résultats de classification entre locuteurs schizophrènes et non schizophrènes à partir de leurs interventions dans des dialogues transcrits.

Dans ces expériences, nous avons choisi comme première approche d'isoler les tours de parole de chaque locuteur dans les dialogues : nous extrayons et concaténons les TDP de personne avec schizophrénie (respectivement du témoin) dans le dialogue considéré et utilisons cette concaténation comme instance de classification (les cTDP du psychologue sont ici ignorées). La classification est alors traitée à partir de documents longs, contrairement à une approche à partir du dialogue (*i.e.* une instance de classification serait un tour de parole) qui se construirait sur la succession d'interventions courtes contenant trop peu d'information pour une classification précise. Cette approche nous permet d'englober tout la production d'un interlocuteur donné, mais elle a le désavantage de perdre le contexte fourni par l'autre interlocuteur, donc les éléments d'interaction. Ces aspects sont reportés à de futurs développements.

Par ailleurs, nous nous sommes concentrés sur des traits linguistiques directement extraits des transcriptions. Il est cependant certain que des informations non linguistiques seraient cruciales pour cette tâche, comme le non verbal, le genre, l'âge, ainsi que les résultats des patients obtenus aux tests neurocognitifs. Là aussi, de futurs développements intégrant le comportement oculomoteur pourrait être pris en compte.

4.1 Représentation des données

Les cTDP des personnes avec schizophrénie sont libellées comme des instances positives, et celles des témoins comme des instances négatives. Les traits sont construits à partir d'informations lexicales, morpho-syntaxiques et syntaxiques.

Type de traits	Classifieur	#Orig.	Seuil	#Sélec.	Ratio %
`bow`	NB	6504	9	6488	99,75
`bow`	SVM	6504	méd.	3254	50,03
n`-gram`	SVM	118473	8	98	0,08
`treelet`	SVM	16865	3	675	4,00
`bow + treelet`	NB	23369	8	11684	49,99
`bow + treelet`	SVM	23369	moy.	3434	14,69
`bow +` n`-gram`	SVM	124977	4	491	0,39
n`-gram + treelet`	SVM	135338	4	552	0,41
`bow +` n`-gram + treelet`	SVM	141842	5	257	0,18

TABLE 1: Nombre de traits à l'origine ("#orig.") et sélectionnés ("#selec") par les classifieurs

Traits lexicaux : Un document est représenté par les mots (tokens) qui le constituent sans tenir compte de l'ordre (sac de mots, `bow`). Cette représentation est la plus simple et sert de système de référence. Ce modèle permet par ailleurs d'identifier de potentielles préférences lexicales. Nous testons également une représentation en n-grammes sur les tokens (n`-gram`), afin de prendre en compte partiellement l'ordre des mots. Notons que les n-grammes peuvent contenir des mots à cheval sur différentes prises de parole d'un même locuteur, et donc encoder une partie du contexte dialogique. Nous testons des bi-grammes et des tri-grammes. Les données sont normalisées en utilisant le TF-IDF.

Traits morpho-syntaxiques et syntaxiques : Nous utilisons UDPipe [3] (Straka & Straková, 2017).Comme nos données sont dialogiques, les modèles classiques donnent d'assez mauvais résultats. Nous utilisons donc un modèle ré-entraîné sur un corpus oral du français (Spoken-French 2.5 [4]). Le pré-traitement supprime la ponctuation et segmente minimalement (par exemple, ajout d'un espace pour les apostrophes). Nous obtenons un étiquetage morpho-syntaxique et l'analyse syntaxique en dépendances correspondante. Afin d'encoder les traits syntaxiques, nous utilisons la méthode proposée dans (Johannsen *et al.*, 2015) qui consiste à extraire tous les sous-arbres d'au plus 3 *tokens* (`treelet`). Ces auteurs s'intéressaient, eux, aux variations syntaxiques liées au genre et à l'âge. Un `treelet` d'1 token correspond simplement au *POS* associé. Un `treelet` contenant 2 tokens est une relation typée entre une tête et un dépendant, par exemple : 'VERB→nsubj→NOUN'. Un `treelet` sur 3 tokens peut avoir deux formes, selon que l'on a une tête dominant deux dépendants ('NOUN←nsubj←VERB→dobj→NOUN') ou une chaîne de dépendances ('PRON←poss←NOUN←nsubj←VERB').

Sélection de traits : Les trois ensembles de traits construits correspondent à des vocabulaires larges (voir la table 1). Notre problème d'apprentissage est confronté à des données rares (41 instances) mais de dimension élevée, ce qui conduit généralement à des problèmes de sur-apprentissage et de manque de généralisation des modèles. Nous incluons une sélection des traits au cours de l'entraînement avec une méthode implémentée [5] dans scikit-learn (Pedregosa *et al.*, 2011). En calculant les coefficients (ou poids) attribués par un modèle à chaque trait et en ne conservant que ceux dont le poids est supérieur

3. http://ufal.mff.cuni.cz/udpipe
4. https://tinyurl.com/UniversalDependencies-French-S
5. `feature_selection.SelectFromModel` https://scikit-learn.org/

à un seuil, cette méthode permet de sélectionner les traits importants. Nous testons sans sélection (seuil `None`), puis avec un seuil correspondant à la moyenne et la médiane sur les poids obtenus, ainsi que 10 valeurs régulièrement distribuées entre $1e-5$ (la valeur par défaut dans l'implémentation utilisée) et le poids du 50^e trait le plus important. Cette valeur maximale choisie *a priori* permet de conserver au minimum 50 traits dans le modèle.

La sélection de traits nous permet de réduire drastiquement la taille du vocabulaire, comme indiqué dans la table 1. Notons que NB conduit généralement à conserver plus de traits, la distribution des coefficients étant plus continue.

4.2 Classification

Nous avons trop peu de données (41 documents) pour séparer les données entre entraînement et test, nous utilisons une validation croisée enchâssée permettant d'obtenir une estimation réaliste de l'erreur du modèle (Varma & Simon, 2006; Scheffer, 1999). Contrairement à une simple validation croisée, le modèle est choisi sur un ensemble de données différent de celui utilisé pour l'évaluation. Cette méthode repose sur deux boucles : à l'extérieur, un sous-ensemble parmi N est conservé pour l'évaluation, tandis qu'à l'intérieur, une validation croisée en M sous-ensembles permet d'optimiser le modèle (choix des hyper-paramètres et sélection de traits), le processus est répété N fois, ici $N = M = 5$.

Modèles : Nous testons différents modèles de classification implémentés dans ScikitLearn en optimisant les hyper-paramètres suivants :

— Naive Bayes : paramètre de lissage $\alpha \in V = \{0.001, 0.005, 0.01, 0.1, 0.5, 1, 5, 10, 100\}$;
— Régression logistique : norme L_2, et optimisation du coefficient de régularisation $C \in V$;
— SVM avec noyau linéaire : norme L_2, et optimisation de $C \in V \cup \{1000\}$.

Pendant le processus de validation, nous optimisons les hyper-paramètres dans les boucles intérieures. Ces hyper-paramètres restent relativement stables sur l'ensemble des expériences pour chaque modèle et chaque type de trait. Les valeurs des hyper-paramètres choisis sont, pour NB, $\alpha = 0{,}1$, avec `ngram`, 0,001 avec `bow`, `treelet`, `bow+treelet`, 0,01 dans les autres cas ; pour MaxEnt, $C = 100$ pour tous les traits ; pour SVM, on trouve $C = 1000$ pour `treelet`, $C = 100$ pour `bow`, `n-gram` et `n-gram+treelet`, sinon $C = 5$.

5 Résultats

5.1 Systèmes de référence

Nous reportons dans la table 2 les résultats de nos systèmes de référence. Le classifieur correspondant à la chance, qui attribue systématiquement la classe majoritaire (ici témoin), a une exactitude de 56,1%. Par ailleurs, nous avons examiné trois autres systèmes de référence : deux testant un indice simple de complexité des mots et le troisième la différence d'utilisation des déictiques "je" et "tu".

Avec l'hypothèse que les personnes avec schizophrénie utilisent des mots moins complexes, nous testons des systèmes fondés sur, simplement, la taille moyenne des mots ou, pour introduire une forme de normalisation, la taille moyenne des mots au dessus de la taille moyenne de tous les mots.

	Acc.	Prec.	Rec.
Majorité	56,10		
long. mot	49,51	17,21	11,11
$>$ long. moy. mot	52,43	37,43	22,78
ratio je/tu	**72,19**	**69,87**	**35,56**

TABLE 2: Résultats des systèmes de référence : classe majoritaire, longueur des mots, longueur des mots supérieure à la moyenne, ratio d'utilisation des déictiques “je" et “tu"

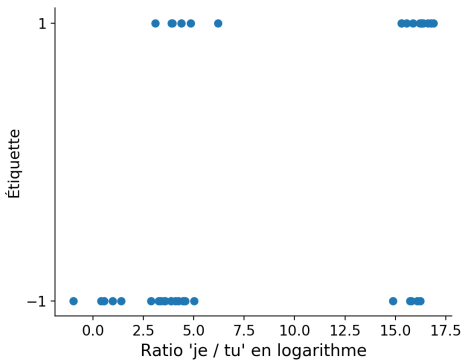


FIGURE 1: Ratio *je* / *tu* par document

Les deux classifieurs donnent des résultats inférieurs à la chance (exactitude de, resp., 49,51% et 52,43%), la longueur des mots n'est donc pas une information pertinente pour cette tâche.

Comme de précédentes études ont noté l'importance des pronoms de première personne pour la tâche, nous testons également un système fondé sur le ratio des déictiques "je" et "tu" (normalisé en logarithme) dans chacune des deux classes. Le système obtient de meilleurs résultats, largement au dessus de la chance (72,19% d'exactitude). La figure 1 montre que, dans nos données, les témoins (étiquette -1) ont un taux quasiment similaire d'utilisation de ces déictiques (majorité des documents autour de 0) tandis que les personnes avec schizophrénie (resp. 1) favorisent la première personne.

5.2 Meilleurs systèmes

La table 3 présente les résultats obtenus par les différents classifieurs pour les différents ensembles de traits testés. Le meilleur système obtient 93,66% d'exactitude (F1 92,21%), il est fondé sur l'algorithme bayésien naïf (NB) et les traits de type sac de mots (`bow`). Avec SVM, nous obtenons 90,98% d'exactitude (89,79 en F1). Ces résultats sont supérieurs à ceux présentés dans (Allende-Cid *et al.*, 2019) (87,50% en F1) qui utilisent également une représentation sac de mots et SVM mais un corpus plus large, et également supérieurs à ceux présentés dans (Birnbaum *et al.*, 2017), ceux-ci obtenant 90% d'exactitude sur des données Twitter également plus nombreuses, avec des traits de type n-grammes ($n = 1, 2$ et 3, donc correspondant ici à `bow+n-gram`) ainsi que des catégories sémantiques issues du lexique LIWC (Pennebaker *et al.*, 2001) et un classifieur de type Random Forest. Ceci pourrait indiquer que nous avons un vocabulaire plus restreint et peut-être biaisé dans nos données. Le second meilleur système, 92,20% d'exactitude (F1 90,38%), est obtenu avec ces mêmes traits lexicaux combinés aux traits syntaxiques (`bow+treelet`) et également NB.

De manière assez classique, les meilleurs scores sont obtenus avec le classifieur SVM, sauf lorsque les traits `bow` sont pris seuls ou du moins dominent (dans la combinaison avec les `treelet`, la phase de sélection a tendance a plus largement supprimé des traits de cette catégorie que des traits sac de mots, cf. table 1), auxquels cas c'est NB qui permet les meilleures performances.

Comme les différences observées entre les scores sont assez faibles et que la taille du corpus est restreinte, nous avons vérifié la significativité statistique de certains résultats : notamment, nous cherchons à vérifier si NB est vraiment supérieur à SVM ($\pm$3,66% avec `bow`, $\pm$3,42% avec `bow+treelet`), et si ces deux ensembles de traits correspondent à des performances différentes ($\pm$1,46 avec NB et $\pm$2,2 avec SVM). Nous utilisons le *test de Student* qui est interprétable avec un échantillon de très petite

Algorithme Sélection	SVM non	SVM oui	MaxEnt oui	NB oui
`bow`	90,00	90,98	87,07	**93,66**
n-`gram`	68,78	**81,71**	79,76	65,61
`treelet`	61,46	**66,83**	58,29	58,05
`bow+`n-`gram`	80,49	**88,54**	86,59	70,49
`bow+treelet`	87,07	88,78	84,88	**92,20**
n-`gram+treelet`	68,54	**80,73**	77,56	62,20
`bow+`n-`gram+treelet`	80,98	**85,85**	84,15	77,07

TABLE 3: Exactitude moyenne ("Avg Acc.") sans ou avec sélection ("SVM", "MaxEnt" et "NB") pour chaque ensemble de traits

Groupe d'échantillons		t-statistique	p-value	d de Cohen	Taille d'effet
`bow_nb`	`bow_svm`	2,74	0,01	1,23	fort
`bow+treelet_nb`	`bow+treelet_svm`	2,10	0,05	0,94	fort
`bow_nb`	`bow+treelet_nb`	1,21	0,24	0,54	moyen
`bow_svm`	`bow+treelet_svm`	1,49	0,15	0,67	moyen

TABLE 4: Résultats des Tests de Student pour la comparaison de classifieurs

taille, en particulier si la taille d'effet (*effect size*, calculée en utilisant le coefficient d de Cohen [6]) et la corrélation entre les échantillons (t-statistiques) sont suffisamment importantes (De Winter, 2013).

Les résultats sont présentés dans la table 4. Ces tests démontrent que l'algorithme NB permet effectivement des performances significativement supérieures à celles obtenues avec SVM (p-value $\leq$ 0,05). Par contre, l'algorithme étant fixé, la perte en performance observée en ajoutant les informations syntaxiques (`treelet`) n'est pas significative, cette combinaison n'apporte rien.

Le volume de données étant limité, les erreurs de classement des documents entraînent une variation importante des résultats. Ainsi, SVM a en fait systématiquement un meilleur rappel (pour `bow` : 90,56 vs 86,11 ; pour `bow+treelet` : 86,67 vs 84,44) et, pour `bow`, il classe correctement 16,3 instances de personnes avec schizophrénie contre 15,5 pour NB (sur 18) tandis que NB classe généralement correctement toutes les instances de témoins, donc la classe majoritaire.

5.3 Différents jeux de traits

Les informations lexicales semblent clairement les plus pertinentes pour la tâche : après `bow`, ce sont les n-`gram` qui conduisent aux meilleures performances (au mieux, 81,71% d'exactitude) tandis que les performances sont bien plus basses (66,83% d'exactitude, 57,30% en F1) avec les traits syntaxiques (`treelet`) pris seuls. Ce score est supérieur à la chance (56,10%) ce qui semble indiquer qu'un signal est présent, mais largement moins discriminant que l'information lexicale et apparemment non complémentaire, comme le montre l'absence d'amélioration lorsque les traits sont

6. Traditionnellement, un d autour de 0,2 est décrit comme un effet faible, 0,5 moyen et 0,8 comme fort.

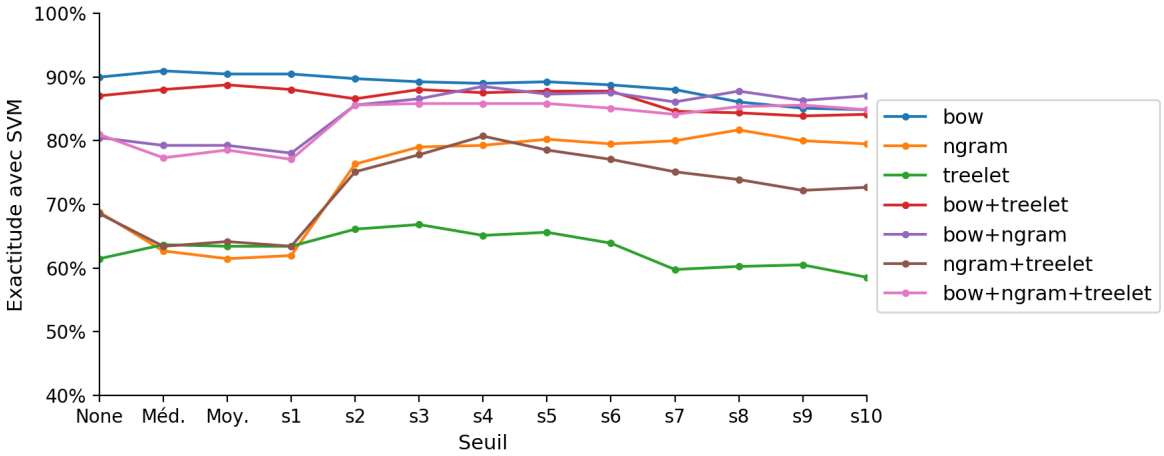


FIGURE 2: Score d'exactitude en fonction des seuils de sélection pour les traits

combinés (seconde partie de la table 3). Par contre, le système fondé sur les `treelet` obtient un score d'exactitude inférieur au système de référence utilisant le ratio de déictique, ce qui démontre l'importance de cette information pour la tâche. Kayi *et al.* (2017) rapportent des scores supérieurs (F1 de 68,48% pour les textes narratifs et 63,19% pour les tweets)) avec le même algorithme (SVM) et en ne considérant que les POS, pourtant incluses dans `treelet`. Cette baisse est attribuable soit au bruit amené par les autres traits pris en compte dans notre système (`treelet` de 2 ou 3 tokens, cf. Section 4.1) soit à la taille des données, ces auteurs disposant de plus d'instances de classification (348 tweets ou 190 textes narratifs contre 41 ici).

Dans nos expériences, combiner les traits ne conduit à aucune amélioration, voire dégrade les performances. L'information contenue dans les traits lexicaux semblent redondantes, puisque combiner `bow` et n`-gram` laissent les performances inchangées avec sélection. Sans sélection, on peut constater (cf. figure 2, seuil "None") que cette combinaison surpasse n`-gram` seul mais pas `bow` seul. La représentation par n`-gram` ne semble ici pas adapté, peut-être à cause des chevauchements entre tours de parole. Par ailleurs, comme le montre la figure 2, l'étape de sélection est cruciale, elle permet une amélioration pour tous les traits, et les meilleurs scores sont très généralement obtenus avec des seuils hauts.

6 Analyse des traits

Traits lexicaux : Afin d'évaluer la pré-dominance de certains champs lexicaux, nous utilisons le test de corrélation de Spearman pour classer les tokens et regardons ceux avec les plus hautes valeurs (p-valeur $< 0,05$ et coefficient $|\rho| > 0,3$). Nous obtenons 210 mots, une sélection est présentée dans la table 5. Les termes relatifs à la pathologie comme "maladie, traitement, médecin, diagnostic" sont positivement corrélés à la classe schizophrène, tandis que des termes relatifs à la scolarité comme "licence, thèse" et à la vie comme "vacances, bio, monde" reçoivent des poids négatifs, ce qui rejoint les observations de la Section 3. Par ailleurs, les sujets avec schizophrénie utilisent plus de références à la première personne, ce que l'on voit avec des déictiques ("j'" ("je"), "mon", "ma", "mes") ainsi que des formes d'auxiliaires ("suis" et "ai") tandis que les témoins utilisent plus de références de seconde personne ("tu", "es" et "as"). Nous évaluons également l'impact de ces traits dans les modèles : en ignorant "je" et "tu" (et les formes élidées "j'" et "t'"), nous constatons une chute faible de l'exactitude

Vocabulaire	ρ	p-value
Douleur		
maladie	0,540	$< 1e-3$
hospitalisé	0,509	$< 1e-3$
hallucinations	0,420	0,006
Éducation		
master	$-0,505$	$< 1e-3$
concours	$-0,496$	$< 1e-3$
fac	$-0,490$	0,001

Vocabulaire	ρ	p-value
Psycho		
psychologie	$-0,536$	$< 1e-3$
psychologue	$-0,453$	0,002
Déictique		
j' / je	0,635	$< 1e-5$
mon	0,613	$< 1e-5$
t' / tu	$-0,467$	0,002
nous	$-0,342$	0,028

TABLE 5: Valeurs ρ et p du test de Spearman pour les mots traits

avec NB ($-0,49\%$) mais importante avec SVM ($-6,59\%$).

Ces observations rejoignent les conclusions d'études précédentes : par exemple, Strous *et al.* (2009) argumentent qu'une utilisation plus importante des déictiques à la première personne et moins de références aux sujets à la troisième personne, accompagnée par des répétitions lexicales sont des caractéristiques de sujets renfermés sur eux-même. D'autres études ont également affirmé que l'utilisation de la première personne du singulier est associée à des états affectifs négatifs tels que la dépression (Rude *et al.*, 2004; Chung & Pennebaker, 2007). Évidemment, ce type de résultat est à apprécier relativement aux conditions contextuelles et conversationnelles dans lesquelles les données sont recueillies.

Traits syntaxiques : Pour les traits syntaxiques (voir la table 6), les verbes semblent un marqueur fort des personnes avec schizophrénie tandis que les noms apparaissent plus souvent dans le corpus des témoins. Les statistiques des *2-token* `treelet` tendent à indiquer que les personnes avec schizophrénie utiliseraient plus de groupes verbaux et moins de groupes nominaux. Ainsi le *2-token* `treelet` "VERB→aux→AUX" (par exemple : "(j')ai fait", "(c')est (pas) gagné") et "VERB→nsubj→PRON" (par exemple : "ça va", "(je) sais pas") sont les traits les plus discriminants des personnes avec schizophrénie. Une forte utilisation des auxiliaires montre aussi que les personnes avec schizophrénie parlent souvent du passé, de ce qu'ils ont fait. Côté témoins, on trouve plus de relations qui capturent des nominaux : "expl" capture des nominaux explicatifs ou pléonastiques ; les cas ("case") sont traités comme des dépendants du nom auquel ils s'attachent souvent avec des adpositions [7] (par exemple : "(fait partie) de l'expérience").

7 Conclusion

Nous avons proposé les premiers systèmes d'identification automatique de personnes atteintes de schizophrénie dans des données dialogiques et en français. Nous avons testé différentes représentations, incluant des informations lexicales, morpho-syntaxiques et syntaxiques, ainsi que différents classifieurs. Notre meilleur système utilise des informations lexicales uniquement et obtient une exactitude de 93,66 avec le classifieur NB. Cependant, l'étude des données et des modèles nous a permis

7. Dans le format conllu, l'adposition recouvre les prépositions et postpositions.

`treelet`	SCZ	ρ	Témoin	ρ
1-*token*	verb	0,21	noun	−0,17
2-*token*	verb→aux→aux	0,41	pron→nsubj→pron	−0,64
	verb→nsubj→pron	0,37	cconj→nsubj→pron	−0,46
	aux→advcl→verb	0,34	propn→conj→pron	−0,46
3-*token*	pron→nsubj→verb←iobj←pron	0,51	pron→obj→verb←mark←sconj	−0,66
	aux→aux→verb←obl←pron	0,49	adp→mark→verb←det←det	−0,39
	adj→advcl→verb←nsubj←pron	0,47	verb→expl→noun→case→adp	−0,36

TABLE 6: Traits typiques des classes SCZ et témoins (p-value $<0,05$ pour les *2-tokens* et *3-tokens*)

d'identifier de potentiels biais lexicaux dans notre corpus, notamment à travers un groupe témoins dont le vocabulaire est centré sur les études, et des patients habitués à décrire leur environnement médical, ce qui rend probablement nos modèles peu robustes.

Par ailleurs, nous nous limitons au contenu linguistique de l'échange sans considérer l'évolution de la phonologie, de la phonétique, ni ce qui appartient au non-verbal (position, regards, *etc.*). Cependant, ce groupe reste selon nous pertinent pour aller vers le développement d'applications ciblées, comme la détection de changement d'états chez les patients ou l'adaptation automatique d'un *chatbot* par exemple.

Dans une extension de l'étude nous souhaitons tester d'autres classifieurs comme random forest ou des perceptrons. Par ailleurs, nous souhaitons tester plus de traits certains toujours linguistiques comme les déictiques, et surtout intégrer des informations sémantiques comme des connecteurs, d'autres extra linguistiques, en particulier les résultats aux tests neuro-cognitifs. Il nous apparaît aussi primordial de nous intéresser à la production langagière dans sa dynamique en l'analysant comme un dialogue et pas seulement comme un quasi-monologue.

Les résultats obtenus montrent que si les classifications ne sont pas parfaitement opérantes, elles ouvrent vers des indices intéressants. Des analyses récentes tendent à prouver qu'il ne faut pas limiter l'analyse à la seule production des patients. De manière surprenante, il semble que les interlocuteurs des personnes avec schizophrénie adaptent leur manière de parler à leur interlocuteur. Ainsi, tout un chacun identifierait inconsciemment des défaillances chez l'autre, les spécialistes interprétant ces indices de façon diagnostique. Ainsi, un prolongement de cette étude doit s'intéresser à la production langagière du psycho-thérapeute soit en face d'un patient, soit en face d'un témoin.

Remerciements

Nous remercions les relecteurs pour leurs commentaires pertinents. Ce travail a obtenu le soutien du projet PIA "Lorraine Université d'Excellence", ANR-15-IDEX-04-LUE, ainsi que les infrastrustures du CPER LCHN (Contrat de Plan État-Région - Langues, Connaissances et Humanités Numériques). Nous remercions le Centre Hospitalier Le Vinartier pour avoir contribué de manière décisive à la mise en place de l'expérimentation.

Références

ALLENDE-CID H., ZAMORA J., ALFARON-FACCIO P. & ALONSO M. (2019). A machine learning approach for the automatic classification of schizophrenic discourse. *IEEE Access*, p. 45544–45554. DOI : 10.1109/ACCESS.2019.2908620.

AMBLARD M. & FORT K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. In *TALN - Traitement Automatique des Langues Naturelles*, p. 292–303, Marseille, France. HAL : hal-01054391.

AMBLARD M., FORT K., DEMILY C., FRANCK N. & MUSIOL M. (2015). Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Langues*, **55**(3), 91 – 115. HAL : hal-01188677.

AMBLARD M., FORT K., MUSIOL M. & REBUSCHI M. (2014). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France. HAL : hal-01079308.

ANDREASEN N. C. (1979). Thought, language, and communication disorders : I. clinical assessment, definition of terms, and evaluation of their reliability. *Archives of general Psychiatry*, **36**(12), 1315–1321.

ASHER N., ASHER N. M. & LASCARIDES A. (2003). *Logics of conversation*. Cambridge University Press.

ASSOCIATION A. P. *et al.* (2015). *DSM-5-Manuel diagnostique et statistique des troubles mentaux*. Elsevier Masson.

BESCHE-RICHARD C., TERRIEN S., RINALDI R., VERHAEGEN F., LEFEBVRE L. & MUSIOL M. (2018). Les troubles du spectre de la schizophrénie. In C. BESCHE-RICHARD, Éd., *Psychopathologie cognitive. Enfant, adolescent, adulte*, Univers Psy, chapitre 6, p. 153–179. Dunod.

BIRNBAUM M. L., ERNALA S. K., RIZVI A. F., DE CHOUDHURY M. & KANE J. M. (2017). A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research*, **19**(8), e289. DOI : 10.2196/jmir.7956.

CHAIKA E. (1974). A linguist looks at "schizophrenic" language. *Brain and language*, **1**(3), 257–276.

CHUNG C. & PENNEBAKER J. W. (2007). The psychological functions of function words. In K. FIEDLER, Éd., *Social Communication*, volume 1, chapitre 12, p. 343–359. Psychology Press. DOI : 10.4324/9780203837702.

DE WINTER J. C. (2013). Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, **18**. DOI : 10.7275/e4r6-dj05.

DOCHERTY N. M., HAWKINS K. A., HOFFMAN R. E., QUINLAN D. M., RAKFELDT J. & SLEDGE W. H. (1996). Working memory, attention, and communication disturbances in schizophrenia. *Journal of Abnormal Psychology*, **105**(2), 212–219. DOI : 10.1037/0021-843X.105.2.212.

DOUKI DEDIEU S., OUALI U. & NACEF F. (2012). Schizophrénie et genre. In J. DALÉRY, T. D'AMATO & M. SAOUD, Éds., *Pathologies schizophréniques*, Psychatrie, p. 199–205. Lavoisier.

FRASER W. I., KING K. M., THOMAS P. & KENDELL R. E. (1986). The diagnosis of schizophrenia by language analysis. *The British Journal of Psychiatry*, **148**(3), 275–278.

FROMKIN V. A. (1975). A linguist looks at "linguist looks at 'schizophenic language' ". *Brain and Language*, **2**, 498–503. DOI : 10.1016/S0093-934X(75)80087-3.

GREENSTEIN D., WEISINGER B., MALLEY J. D., CLASEN L. & GOGTAY N. (2012). Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. *Frontiers in psychiatry*, **3**. DOI : 10.3389/fpsyt.2012.00053.

HOFFMAN R. E. & SLEDGE W. (1988). An analysis of grammatical deviance occurring in spontaneous schizophrenic speech. *Journal of neurolinguistics*, **3**(1), 89–101.

JARROLD W. L., PEINTNER B., YEH E., KRASNOW R., JAVITZ H. S. & SWAN G. E. (2010). Language analytics for assessing brain health : Cognitive impairment, depression and pre-symptomatic alzheimer's disease. In *International Conference on Brain Informatics*, p. 299–307 : Springer.

JOHANNSEN A., HOVY D. & SØGAARD A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, p. 103–112.

KAYI E. S., DIAB M., PAUSELLI L., COMPTON M. & COPPERSMITH G. (2017). Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, p. 241–250.

MITCHELL M., HOLLINGSHEAD K. & COPPERSMITH G. (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology : From linguistic signal to clinical reality*, p. 11–20.

PEDERSEN T. (2015). Screening twitter users for depression and ptsd with lexical decision lists. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology : from linguistic signal to clinical reality*, p. 46–53.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PENNEBAKER J., FRANCIS M. & BOOTH R. (2001). *Linguistic inquiry and word count (LIWC)*.

PESTIAN J. P., SORTER M., CONNOLLY B., BRETONNEL COHEN K., MCCULLUMSMITH C., GEE J. T., MORENCY L.-P., SCHERER S., ROHLFS L. & GROUP S. R. (2017). A machine learning approach to identifying the thought markers of suicidal subjects : a prospective multicenter trial. *Suicide and Life-Threatening Behavior*, **47**(1), 112–121.

POTVIN S., AUBIN G. & STIP E. (2017). L'insight neurocognitif dans la schizophrénie. *L'Encéphale*, **43**(1), 15–20.

REBUSCHI M., AMBLARD M. & MUSIOL M. (2014). Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances. In M. REBUSCHI, M. BATT, G. HEINZMANN, F. LIHOREAU, M. MUSIOL & A. TROGNON, Éds., *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics : Dialogue, Rationality, and Formalism*, volume 3 de *Logic, Argumentation & Reasoning*, p. 343 – 368. Springer. DOI : 10.1007/978-3-319-03044-9_15, HAL : hal-00910725.

RUDE S., GORTNER E.-M. & PENNEBAKER J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, **18**(8), 1121–1133.

SABETI M., KATEBI S., BOOSTANI R. & PRICE G. (2011). A new approach for eeg signal classification of schizophrenic and control participants. *Expert Systems with Applications*, **38**(3), 2063–2071.

SCHEFFER T. (1999). *Error Estimation and Model Selection*. Thèse de doctorat, Technischen Universitet Berlin, School of Computer Science.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

STROUS R. D., KOPPEL M., FINE J., NACHLIEL S., SHAKED G. & ZIVOTOFSKY A. Z. (2009). Automated characterization and identification of schizophrenia in writing. *The Journal of nervous and mental disease*, **197**(8), 585–588.

VARMA S. & SIMON R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, **7**.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Classification de relations pour l'intelligence économique et concurrentielle

Hadjer Khaldi[1, 2] Amine Abdaoui[1] Farah Benamara[2] Grégoire Sigel[1]
Nathalie Aussenac-Gilles[3]
(1) Geotrend, Toulouse, France
(2) IRIT, Université de Toulouse, France
(3) IRIT, CNRS, Toulouse, France
prenom.nom@irit.fr, prenom@geotrend.fr

RÉSUMÉ

L'extraction de relations reliant des entités par des liens sémantiques à partir de texte a fait l'objet de nombreux travaux visant à extraire des relations génériques comme l'hyperonymie ou spécifiques comme des relations entre gènes et protéines. Dans cet article, nous nous intéressons aux relations économiques entre deux entités nommées de type organisation à partir de textes issus du web. Ce type de relation, encore peu étudié dans la littérature, a pour but l'identification des liens entre les acteurs d'un secteur d'activité afin d'analyser leurs écosystèmes économiques. Nous présentons BIZREL, le premier corpus français annoté en relations économiques, ainsi qu'une approche supervisée à base de différentes architectures neuronales pour la classification de ces relations. L'évaluation de ces modèles montre des résultats très encourageants, ce qui est un premier pas vers l'intelligence économique et concurrentielle à partir de textes pour le français.

ABSTRACT

Relation Classification for Competitive and Economic Intelligence

Relation extraction aims at identifying semantic relations that may hold between entities in raw text. This task has been widely studied in the litterature focusing either on extracting generic relations like hyperonymy or domain-dependent relations like those linking genes and proteins. In this paper, we aim at extracting business relations between two organizations from web textual contents. In particular, we propose BIZREL, the first French annotated dataset for business relations as well as a supervised approach based on several neural architectures to classify these relations. Our results are encouraging and constitute a first step towards economic and competitive intelligence from French texts.

MOTS-CLÉS : Classification de relation, Relation économique, Ressource linguistique.

KEYWORDS: Relation classification, Business relation, Linguistic resources.

1 Motivations

L'extraction de relations sémantiques est une tâche d'extraction d'information qui permet de détecter les liens sémantiques reliant des entités à partir d'expressions en langage naturel, dans le but de produire de l'information structurée à partir de textes bruts (Aussenac-Gilles *et al.*, 2013). Les entités ainsi reliées peuvent être exprimées par des syntagmes nominaux désignant des classes (Hendrickx

Volume 2 : Traitement Automatique des Langues Naturelles, pages 27–39.
HAL : hal-02784753.

et al., 2010), comme *entreprise, produit*, ou des entités nommées (Mitchell *et al.*, 2005), comme *Google*, *Paris*.

Cette tâche, qui s'apparente à un problème de classification supervisée de fragments de phrases, a fait l'objet de nombreux travaux basés sur des approches à base de patrons (Aussenac-Gilles & Jacques, 2008), à base de traits ou de noyaux (Kambhatla, 2004; Culotta & Sorensen, 2004), ou neuronales (Wang *et al.*, 2016; Lee *et al.*, 2019). Nous distinguons ceux qui classifient des relations génériques telles que les relations *cause-effet*, *message-sujet* (Hendrickx *et al.*, 2010) ou encore des relations d'*hypernonymie* ou de *synonymie* (Hearst, 1992; Lee *et al.*, 2017), et ceux se focalisant sur des relations spécifiques telles que les relations biomédicales reliant des protéines ou des gènes (Zhou *et al.*, 2014; Zhang *et al.*, 2018; Fan *et al.*, 2018). Les approches s'appuyant sur de l'apprentissage supervisé utilisent des corpus manuellement annotés par des types de relations prédéfinis dont la plupart sont en anglais (tels que les corpus de SemEval-2010 Tâche 8 (Hendrickx *et al.*, 2010), TACRED (Zhang *et al.*, 2017) et BioNLP-OST 2019 (Bossy *et al.*, 2019)). Nous citons néanmoins quelques corpus dans d'autres langues comme ACE 2004 (Mitchell *et al.*, 2005) pour le chinois et l'arabe, ou ReRelEM (Freitas *et al.*, 2009) pour le portugais.

Dans cet article, nous nous intéressons à la classification de relations économiques prédéfinies entre deux entités nommées de type organisation (ORG) à partir de textes issus du web. Ces relations permettent de créer des réseaux commerciaux, valoriser des entreprises et suivre leurs activités. C'est donc un moyen crucial pour l'identification des liens entre les acteurs d'un secteur d'activité et pour l'analyse de leurs écosystèmes économiques. Contrairement aux relations génériques, peu de travaux se sont attaqués à la classification des relations économiques. Ces derniers se focalisent principalement sur deux types de relations, la *compétition* et la *coopération* (Zuo *et al.*, 2017; Lau & Zhang, 2011). Les techniques adoptées visent à adapter les travaux sur la classification de relations génériques pour prendre en compte la spécificité des relations économiques. Les approches varient de méthodes à base de patrons (Braun *et al.*, 2018) à des approches supervisées (Yan *et al.*, 2019) ou semi-supervisées (Lau & Zhang, 2011; Zuo *et al.*, 2017). Là encore, les données utilisées pour entraîner ces différents systèmes sont majoritairement en anglais (Lau & Zhang, 2011) avec quelques initiatives en allemand (Braun *et al.*, 2018) et en chinois (Yan *et al.*, 2019). À notre connaissance, il n'existe aucun corpus annoté pour la classification de relations économiques pour la langue française. Dans ce contexte, nous proposons :

1. *Une taxonomie* composée de cinq relations jugées pertinentes dans le cadre d'intelligence économique : *investissement, compétition, coopération, vente-achat* et *poursuite judiciaire*.
2. BIZREL, *le premier corpus français* de 10k instances de relations annotées manuellement selon cette taxonomie via une plateforme d'annotation collaborative. Ce corpus est mis à disposition de la communauté. [1]
3. *Une approche supervisée* pour la classification de relations économiques à base de différentes architectures neuronales. Nous évaluons en particulier le modèle de langue multilingue M-BERT (Devlin *et al.*, 2019) mais aussi ses variantes françaises CamemBERT (Martin *et al.*, 2019) et FlauBERT (Le *et al.*, 2019), les modèles de langues récemment développés pour le français. Ces deux derniers n'ont jamais été utilisés pour une tâche de classification de relations. Nous comparons les performances de ces modèles avec celles obtenues par d'autres modèles de classification de relations génériques à la Tâche 8 de SemEval-2010.
4. Une analyse d'erreurs montrant les limites des approches proposées.

1. BizRel_corpus.

Dans la suite de cet article, la section 2 présente un bref état de l'art des approches de classification de relations. La section 3 définit les relations économiques d'intérêt et rapporte la procédure suivie pour construire le corpus BIZREL. La section 4 décrit les modèles utilisés, ainsi que les résultats obtenus. Une analyse d'erreur est présentée en section 5. Nous présentons les perspectives envisagées des travaux futurs en section 6.

2 Classification de relations : état de l'art

2.1 Classification de relations génériques

Les approches classiques visent à identifier une combinaison de traits lexicaux, syntaxiques et sémantiques (Kambhatla, 2004; Zhou *et al.*, 2005; Nguyen *et al.*, 2007) à partir desquels un classifieur apprend à distinguer les différentes relations prédéfinies ou à exploiter les liens de dépendances entre les deux entités, en utilisant des représentations plus riches des instances de relations telles que des arbres syntaxiques ou de dépendances (Collins & Duffy, 2001; Culotta & Sorensen, 2004). Plus récemment, les approches neuronales, telles que des CNN, arrivent à extraire des traits aux niveaux du mot et de la phrase à l'aide de modèles de plongements de mots, tout en prenant en compte leurs positions relatives par rapport aux deux entités concernées par la relation (Zeng *et al.*, 2014; Nguyen & Grishman, 2015; dos Santos *et al.*, 2015). D'autres y rajoutent un mécanisme d'attention qui capture, pour chaque mot d'une instance de relation, les éléments de contexte qui lui sont pertinents, améliorant ainsi les performances (Wang *et al.*, 2016; Shen & Huang, 2016). Des modèles basés sur une architecture RNN Bi-LSTM [2] avec attention ont été proposés par (Lee *et al.*, 2019; Xiao & Liu, 2016) pour capter des traits à longue distance et à distance variable. Plus récemment, l'utilisation des modèles de langue à base de transformers a encore amélioré les performances. Par exemple, Wu & He (2019) ont utilisé BERT (Devlin *et al.*, 2019) en concaténant les représentations contextuelles des entités avec la représentation de la phrase produite par le transformer. Ce modèle a pu atteindre un score F1 de 89.25% sur le corpus SemEval-2010 Tâche 8, en dépassant les modèles neuronaux. Baldini Soares *et al.* (2019) a exploré l'efficacité de différentes représentations possibles d'une instance de relation où l'utilisation des représentations des marqueurs d'entités s'est avérée la plus efficace. Enfin, Tao *et al.* (2019) ont combiné la représentation fournie par BERT avec des traits syntaxiques de la phrase, améliorant ainsi les performances.

2.2 Classification de relations économiques

Selon la nature de l'activité reliant deux organisations, Zhao *et al.* (2010) répartit les relations économiques en quatre types : coopération, investissement, vente et approvisionnement. D'autres travaux les classent en deux groupes uniquement (Zuo *et al.*, 2017; Yamamoto *et al.*, 2017) : co-opération et concurrence. Globalement, les relations économiques sont peu présentes dans les bases de connaissances telles que Freebase (Bollacker *et al.*, 2008) et DBpedia (Auer *et al.*, 2007) où figurent par exemple les relations *subsidiary* et *ownership_of* (Zuo *et al.*, 2017). Certaines relations économiques sont néanmoins annotées dans des corpus de relations génériques, comme par exemple *Employment/Membership/Subsidiary* dans le corpus ACE 2004 (Mitchell *et al.*, 2005), ou la relation *Component-Whole* dans le corpus SemEval-2010 Tâche 8 (Hendrickx *et al.*, 2010) mais avec des

2. Bidirectionnel à mémoire court terme étendue.

fréquences assez faibles (de l'ordre de 1k). Pour extraire ces relations, la plupart des travaux existants utilisent des approches semi-supervisées, soit à l'aide de patrons lexico-syntaxiques générés à partir d'arbres de dépendance (Braun *et al.*, 2018), soit grâce à une liste de mots clés représentatifs de chaque relation (Lau & Zhang, 2011). Les réseaux de neurones ont été utilisés dans (Yan *et al.*, 2019) qui propose un modèle entraîné sur un corpus chinois de 1k instances de relation économiques, afin d'aider les institutions financières à gérer le risque des crédits aux entreprises.

3 BIZREL : corpus français de relations économiques

BIZREL est le premier corpus annoté en relations économiques pour le français. Il décrit cinq types de relations binaires de nature économique, reliant deux entités nommées de type ORG, notés *EO*. D'abord, nous reprenons les quatre types de relations économiques proposés par Zhao *et al.* (2010) : *investissement*, *coopération*, *vente* et *approvisionnement*. Cependant, nous combinons les deux relations de *vente* et d'*approvisionnement* en une seule relation *vente-achat*, puisque nous ne nous intéressons pas pour le moment à l'orientation des relations entre les entités (i.e., $R(a, b) = R(b, a)$). Puis, nous rajoutons à la liste deux nouveaux types de relation, à savoir *compétition* et *poursuite judiciaire*.

La construction de ce corpus est passée par trois étapes principales : la collecte des phrases potentielles à annoter, l'annotation des entités nommées et enfin l'annotation des relations. Initialement, un ensemble de documents est collecté à partir du web en interrogeant les deux moteurs de recherche Google et Bing et ce en utilisant une liste de mots clés relatifs aux différents domaines d'activités d'entreprise comme *voitures autonomes, impression 3D, etc*. Les documents collectés passent par une étape de pré-traitement où les en-têtes, les pieds de page et les menus de navigation sont supprimés. Ensuite, le contenu de la page est segmenté en phrases et chaque entité nommée présente est identifiée automatiquement à l'aide des systèmes de reconnaissance d'entités nommées de SpaCy et StanfordNLP. Pour augmenter la précision, ne sont retenues que les entités organisations (*EO*) reconnues par les deux systèmes à la fois. Seules les phrases contenant au moins deux *EO* sont conservées. En revanche, celles dont les mots sont à 95 % des *EO* (comme c'est le cas dans des énumérations d'entreprises) sont rejetées.

Ces phrases sont annotées en relations en fonction du lien qu'entretiennent les *EO* annotées. Soulignons qu'entre deux mêmes *EO*, des relations de types différents peuvent exister dans des phrases différentes. De plus, une seule relation est annotée par phrase même si cette dernière peut exprimer plusieurs relations entre différents paires d'*EO* qui ne se chevauchent pas. Enfin, une relation est annotée si et seulement s'il existe dans la phrase un ou des indices explicites de la relation sans avoir recours à des connaissances externes, c'est l'un des principes d'annotation des corpus ACE. Ceci est illustré dans l'exemple suivant :

"Présents dans la ville de Wuhan, Faurecian, ***PSA****,* ***Renault*** *ou encore Valeo ont dû fermer leurs sites situés dans la zone de confinement en attendant le feu vert des autorités chinoises pour reprendre leurs activités."*.

Ici, aucune relation de type économique ne relie les deux *EO* ***PSA*** et ***Renault***, même si ces deux entités peuvent également être reliées par la relation ***Compétition*** car elles partagent le même marché de construction d'automobiles. Cependant, étant donnée que cette seconde relation n'est pas exprimée linguistiquement dans la phrase, elle ne sera pas annotée.

Les types de relations à annoter sont définis comme suit (tous les exemples sont extraits de notre corpus, les *EO* concernées par la relation sont en gras) :

— **Investissements** : une *EO* est filiale d'une autre *EO*, ou *EO* détient (toutes ou une partie) des actions d'une autre *EO*, i.e. : *Le missilier européen **MBDA** (filiale commune de Airbus, Leonardo et **BAE**) espère que l'accord signé à Helsinski lui donnera à terme accès à des financements pour développer de nouvelles versions de son missile antichar de moyenne portée (MMP).*

— **Compétition** : traduit une compétition/rivalité entre deux *EO* fournissant les mêmes biens ou services, ou voulant accéder à un même marché relativement restreint, par exemple : *Boeing et l'avionneur brésilien **Embraer**, rival de **Bombardier** sur les avions régionaux, ont annoncé discuter sur un éventuel rapprochement de leurs activités.*

— **Coopération** : ce type de relation apparaît lorsqu'il existe une coopération contractuelle entre deux *EO*, que deux *EO* travaillent ensemble pour le même projet, par exemple : *Depuis le 25 novembre 2017, 32 associations et startups , 400.000 citoyen.nes, la Fondation **Kering**, **Facebook** et la Région Île-de-France ont travaillé ensemble avec Make.org pour élaborer le premier plan de actions de la société civile contre les violences faites aux femmes.*

— **Poursuite judiciaire** : c'est lorsqu'une *EO* lance une poursuite judiciaire contre une autre *EO*, comme dans : *Grégoire Triet a représenté **Shionogi** dans une action en contrefaçon de brevet portant sur un médicament contre le VIH, qui l'a opposé à **Merck** et ses filiales.*

— **Vente-achat** : une *EO* est cliente de l'autre, ou elle lui fournit des biens ou services, par exemple : *Le capot d'un réacteur d'un **Airbus** A320 de la compagnie **Frontier Airlines** s'est rompu en plein décollage.*

Une classe **Autres** est rajoutée, qui regroupe tous les autres types de relations possibles qui peuvent exister entre deux *EO* et qui ne sont pas des relations économiques.

Un corpus de 10k instances a été construit et annoté par six annotateurs francophones non-experts via la plateforme d'annotation collaborative *Isahit*[3]. L'annotation a été faite par lots contenant chacun 2k instances de relations. Pour chaque lot, 10 % des données annotées sont ré-annotées par des experts. Ceci permet d'évaluer la qualité des annotations et la qualité du guide d'annotation fourni aux annotateurs, et de les améliorer à travers une démarche rétroactive itérative. Sur un total de 1k instances de relations, la moyenne des accords Kappa de Cohen (1960), calculés entre les annotateurs et les experts, est de 0.685. Nous considérons cette valeur comme un bon accord étant donné la diversité des relations économiques à annoter. Le tableau 1 présente le nombre total des relations annotées ainsi que la répartition des instances en corpus de développement, d'entraînement et de test.

	Invest.	**Compét.**	**Coopérat.**	**Poursuit.**	**Vente.**	**Autres.**	**Total**
Entrain.	220	1 229	598	41	188	4 747	**7 023**
Dev.	48	263	128	9	40	1 017	**1 505**
Test.	47	263	129	8	40	1 018	**1 505**
Total	**315**	**1825**	**855**	**58**	**268**	**6 782**	**10 033**

TABLE 1 – Répartition des données annotées par type de relation et type de corpus.

3. https://isahit.com/en/

Le tableau 2 présente quelques statistiques sur notre corpus, avec notamment le nombre total d'*EO* uniques dans le corpus, le nombre de paires d'*EO* ainsi que le nombre maximum, minimum et la moyenne des *EO* par instance de relation. Nous observons que le nombre moyen de *EO* par phrase est de 5 *EO*. Ceci dit, un maximun de 10 relations, en moyenne, pourraient exister dans une seule phrase entre différentes paires de *EO*. De plus, des instances de relations ayant 41 mots en moyenne reflètent la complexité des contextes dans lesquels ces relations économiques sont exprimées dans le web.

	Entrain.	**Dev.**	**Test.**	**Total**
Nb. total EO uniques	1 096	650	688	**1 182**
Nb. paires EO uniques	4 315	1 302	1 305	**5 280**
EO par instance	max = 47 min = 2 moy. = 5	max=40 min = 2 moy.= 5	max = 32 min = 2 moy. = 5	**max = 47** **min = 2** **moy. = 5**
Mots par instance	max = 203 min = 5 moy. = 41	max = 253 min = 7 moy. = 42	max = 189 min = 7 moy. = 42	**max = 253** **min = 5** **moy. = 41**

TABLE 2 – Statistiques sur les EOs dans les relations annotées par type de corpus.

4 Modèles proposés et résultats

Le corpus BIZREL, composé de phrases préalablement annotées en *EO* (deux *EO* par phrase), a été utilisé pour entraîner différents modèles neuronaux afin qu'ils identifient le type de relation économique qui relie les deux *EO*. Notre objectif est de répondre aux deux questions suivantes : (a) Les modèles de la littérature proposés pour la classification de relations génériques sont-ils adaptés pour la classification de relations économiques ? (b) Comment adapter ces modèles pour le français ?

Pour répondre à ces questions, nous proposons cinq modèles. Les trois premiers (M_1, M_2 et M_3) ont obtenu les meilleures scores à la tâche 8 de SemEval 2010 de classification de relations génériques, en anglais. Ces modèles étant indépendants de ressources lexicales et d'outils externes, nous souhaitions tester leurs performances pour la classification de relations économiques. Les deux derniers modèles (M_4 et M_5) sont des adaptations de plongements de mots contextuels issus de modèle de langue basé sur les transformers bidirectionnels pour le français, qui sont utilisés ici pour la première fois pour une tâche de classification de relations. Nous utilisons les plongements de mot français pré-entraînés sur Wikipedia et Common Crawl *FastText* (Joulin *et al.*, 2016, 2017) ayant une dimension de 300 pour les modèles (M_1) et (M_2). Les hyper-paramètres des modèles ont été ajustés sur le corpus de développement. Nous préseontons dans ce qui suit les modèles, puis nous détaillons les résultats obtenus par les expérimentations.

4.1 Description des modèles

Nos modèles sont comme suit :

(M_1) : CNN (Zeng *et al.*, 2014). L'architecture de ce modèle repose sur 3 types de couches :

— Une couche d'entrée qui utilise les plongements de mots pré-entraînés *FastTest* pour associer un vecteur à chaque mot de l'instance. Des plongements de position de dimension 5 sont également calculés par mot pour encoder sa position relative par rapport aux deux EO.

— Trois couches cachées de convolution constituées de 100 filtres chacune, de tailles multiples (des fenêtres de 3,4,5 vecteurs), qui permettent d'extraire des traits au niveau de la phrase à partir de plongements de mot et de position. Au niveau du mot, les plongements des deux *EO* ainsi que de leurs contextes sont utilisés comme traits.

— Une couche de sortie entièrement connectée *(fully connected)* dotée d'un *softmax* qui exploite la concaténation des traits aux niveaux du mot et de la phrase pour calculer la distribution de probabilités des différents types de relation et d'en déduire la relation candidate en prenant le type ayant la probabilité maximale.

L'apprentissage des paramètres du modèle est fait sur 200 *epochs*[4] sur des lots de données *(batch)* de 100 instances, en utilisant la méthode d'optimisation *Adam* (Kingma & Ba, 2014) avec un taux d'apprentissage de 10^{-3}. Une régularisation de type *dropout* (Srivastava *et al.*, 2014) à un taux de 50% est appliquée entre les différentes couches afin de prévenir le sur-apprentissage lors de l'entraînement.

(M_2) : Bi-LSTM avec mécanisme d'attention (Zhou *et al.*, 2016). L'architecture de ce modèle est composée comme suit :

— Une couche d'entrée qui utilise les plongements de mots pré-entraînés *FastTest* pour associer un vecteur à chaque mot de l'instance, avec un taux de dropout de 70% lors de l'entraînement.

— Deux couches cachées composées de cellules LSTM (Hochreiter & Schmidhuber, 1997) bidirectionnelles, de 100 unités par direction avec un taux de dropout de 70% lors de l'entraînement. Ces cellules récurrentes dites "à mémoire court et long terme", permettent de calculer la sortie d'une cellule en prenant en considération l'élément courant de la séquence, mais aussi l'historique passé des cellules précédentes. De plus, la bidirectionnalité permet de considérer le contexte passé et futur pour tout élément de la séquence.

— Une couche d'attention qui fusionne les traits extraits au niveau du mot par le BiLSTM en un vecteur de niveau phrase en les multipliant par un vecteur de poids calculé à partir des sorties des couches LSTM. Ceci permet de capturer les informations sémantiques les plus importantes dans une phrase.

— Une couche de sortie entièrement connectée dotée d'un *softmax* qui exploite les traits extraits par le BiLSTM pour calculer la distribution de probabilités des différents types de relation et d'en déduire la relation candidate en prenant le type ayant la probabilité maximale.

L'apprentissage est fait sur 100 *epochs* sur des lots de données *(batch)* de 10 instances, en utilisant la méthode d'optimisation *Adam* (Kingma & Ba, 2014) avec un taux d'apprentissage de 1, , un taux de décroissance du taux d'apprentissage de 0.9 et un *facteur de régularisation* L_2 de 10^{-5}. Pour prévenir le sur-apprentissage, un *dropout* global à un taux de 50% est appliqué entre les couches cachées et la couche de sortie.

(M_3) : R-mBERT (Wu & He, 2019). Ce modèle est une réadaptation du modèle de langue pré-entraîné BERT (Devlin *et al.*, 2019) pour la tâche de classification de relation en prenant en compte, en plus de la représentation de la phrase, les représentations des deux *EO*.

Un marqueur [CLS] désignant la sortie de classification est rajouté au début de chaque instance de relation, et un marqueur [SEP] est utilisé pour séparer les phrases dans une même instance. Comme

4. Une passe complète sur les données d'entraînement.

réadaptation du modèle, des marqueurs spéciaux sont utilisés pour identifier le début et la fin des *EO* afin de capturer des traits de position les concernant : [E_{11}],[E_{12}] pour *EO1* et [E_{21}],[E_{22}] pour *EO2*. Les instances ainsi modifiées sont introduites au modèle de langue pré-entraîné pour du *fine-tuning* où tous les paramètres pré-entraînés du modèle sont réajustés. À la sortie du modèle de langue, seul l'état caché final est exploité. Les moyennes des vecteurs d'états cachés finaux pour *EO1* et *EO2* sont concaténées avec le vecteur de l'état caché final du marqueur [CLS]. La représentation obtenue passe par une couche entièrement connectée doté d'un classifieur *softmax* qui identifie le type de la relation.

Un modèle pré-entainé de BERT multilingue `bert-base-multilingual-cased`[5], disponible dans la librairie python *HuggingFace*, est utilisé. Le modèle, constitué d'un vocabulaire de $110k$ sous-mots, couvre 104 langues les plus utilisées sur Wikipedia, dont le français. Il compte 12 couches de 768 unités chacune et ayant $110M$ paramètres. Le modèle a été pré-entraîné de façon non-supervisée sur les deux tâches qui sont : (1) *Modèle de langue masqué* (MLM) qui apprend à prédire les sous-mots masqués de façon aléatoire ; et (2) *Prédiction de phrase suivante* (NSP) dans laquelle le modèle apprend à prédire si B est la phrase suivante réelle qui suit A, étant donné une paire de phrases en entrée A, B. Pour l'ajustement des paramètres de ce modèle sur la tâche de classification de relation, la méthode d'optimisation *Adam* est utilisée. La table 3 présente les principaux hyper-paramètres utilisés pour cela.

Nb. epochs	5
Taille de lots par gpu	16
Longueur max. phrase	260
Taux d'apprentissage Adam	2^{-5}
L2 régul. lambda	5^{-3}

TABLE 3 – Réglages des hyper-paramètres.

(M_4) **: R-CamemBERT**. L'architecture du modèle de langue CamemBERT (Martin *et al.*, 2019) est basée sur celle de RoBERTa (Liu *et al.*, 2019), une version optimisée de BERT. Ce modèle est monolingue pré-entraîné sur du textes français issus du corpus multilingue OSCAR (Suárez *et al.*, 2019). R-CamemBERT est adapté à la tâche de classification de relation selon l'architecture proposée par Wu & He (2019), décrite dans M_3. La seule différence du modèle R-mBERT réside dans les marqueurs de classification et de séparation des phrases d'une instance qui sont $< s >$ et $< /s >$, respectivement, pour le modèle R-CamemBERT.

Un modèle pré-entraîné de CamemBERT `camembert-base`, disponible dans la librairie python *HuggingFace*, est utilisé. Ce modèle, constitué d'un vocabulaire de $32k$ sous-mots pour le français, compte 12 couches de de 768 unités chacune et ayant $110M$ paramètres. Il a été pré-entraîné sur la tâche de *Modèle de langue masqué* (MLM) qui consiste ici à prédire des mots masqués aléatoirement au lieu de sous-mots. La méthode d'optimisation Adam est utilisée pour ajuster les paramètres de CamemBERT sur notre tâche. Les mêmes valeurs d'hyper-paramètres qui sont présentées dans la table 3 sont utilisées pour cela.

(M_5) **: R-FlauBERT**. L'architecture du modèle de langue FlauBERT (Le *et al.*, 2019) est basée sur celle de BERT (Devlin *et al.*, 2019). Ce modèle est monolingue pré-entraîné sur du textes français provenant de différentes sources, couvrant différents sujets et des styles d'écriture allant du texte formel (par exemple Wikipédia et livres)[6] au texte aléatoire extrait d'Internet (par exemple Common

5. https://github.com/google-research/bert/blob/master/multilingual.md
6. http://www.gutenberg.org

Crawl [7]). R-FlauBERT reprend également la même architecture proposée par Wu & He (2019), décrite dans M_3, pour prendre en considération les représentations des deux *EO* lors de la classification des relations. Cependant, R-FlauBERT utilise un seul marqueur qui est $< /s >$ pour marquer la classification et la séparation des phrases.

Un modèle pré-entraîné de FlauBERT `flaubert-base-cased`, disponible dans la librairie python *HuggingFace*, est utilisé. Ce modèle, constitué d'un vocabulaire de $50k$ sous-mots pour le français, compte 12 couches de 768 unités chacune et a $138M$ paramètres. Il a été pré-entraîné sur les deux tâche de *Modèle de langue masqué* (MLM) et de *Prédiction de phrase suivante*(NSP). La méthode d'optimisation Adam est utilisée pour ajuster les paramètres de ce modèle sur notre tâche. Les mêmes valeurs d'hyper-paramètres qui sont présentées dans la table 3 sont utilisées pour cela.

4.2 Résultats

Modéle	**Exactitude**	**Précision**	**Rappel**	**F1-score**
M_1 : CNN (Zeng *et al.*, 2014)	76.2	66.8	51.3	57.0
M_2 : Att-Bi-LSTM (Zhou *et al.*, 2016)	74.2	56.6	55.0	53.3
M_3 : R-mBERT (Wu & He, 2019)	80.3	71.2	**64.1**	**67.1**
M_4 : R-CamemBERT	**81.2**	74.6	53.8	59.5
M_5 : R-FlauBERT	80.7	**77.2**	59.7	66.3

TABLE 4 – Résultats de classification de relations économiques sur le corpus de test.

La table 4 présente les résultats obtenus. Nous observons que les modèles basés sur les transformers dépassent ceux basés sur des architectures neuronales de type *CNN* ou *RNN*. Nous constatons également que le modèle (M_3), qui donne aujourd'hui les meilleurs résultats sur le corpus SemEval-2010 Tâche 8 pour la classification de relations génériques, est aussi bon pour la classification de relations économiques, offrant un bon compromis entre *Précision* et *Rappel* (*F1-score*). D'un autre côté, l'adaptation des transformers français à notre tâche (modèles (M_4) et (M_5)) donne des résultats assez bons avec R-FlauBERT qui dépasse légèrement R-CamemBERT. Ceci est semblable aux résultats rapportés par Le *et al.* (2019) sur une tâche de classification de texte du benchmark FLUE.

L'analyse des matrices de confusion telles que obtenues par (M_3), (M_4) et (M_5) montre que ces modèles arrivent à bien distinguer les relations économiques entre elles. Les principales sources d'erreurs proviennent de la relation *Autres*, ce qui montre la difficulté de distinguer les relations économiques des non économiques. Par exemple, pour les relations *Compétition* et *Coopération*, respectivement 163 et 85 instances sont bien classées, 97 et 36 sont classées comme *Autres* et 3 et 5 sont classées comme *autres relations économiques* par le modèle M_3.

5 Analyse d'erreurs

Une analyse manuelle plus poussée des erreurs de classification montre deux autres sources d'erreur. La première concerne les phrases contenant plusieurs relations mais seule celle qui relie les

7. http://data.statmt.org/ngrams/deduped2017

deux *EO* marquées est à identifier, comme dans l'exemple (1). Ici la relation prédite par (M_3) est *Coopération*(EO_2, EO_3) alors que l'annotation de référence indique *Autre*(EO_2, EO_3). En revanche, il existe bien une relation de coopération dans cette phrase mais entre EO_1 et EO_2.

(1) [Sikorsky]$_1$ et ses partenaires [**Hensoldt**]$_2$, Liebherr-Aerospace , MTU , [**Rheinmetall**]$_3$ et ZF lancent le CH-53K qu'il construit pour l'US Marine Corps et qui est le digne successeur de l'actuel CH-53 .

La seconde source d'erreur provient des relations exprimées implicitement, comme dans l'exemple (2) où c'est l'expression métaphorique *se ranger derrière* qui déclenche la relation de compétition entre EO_1 et EO_2. Dans ce cas, le modèle *(M3)* prédit souvent la relation *Autre*.

(2) Plusieurs entreprises de la Silicon Valley dont Google Facebook, [**Dell**]$_1$ et [**HP**]$_2$ se sont rangées derrière Samsung.

6 Conclusions et perspectives

Cet article présente le premier corpus annoté en relations économiques entre organisations pour le français ainsi qu'une approche supervisée pour la classification de ces relations dans des textes issus du web. L'évaluation des modèles de l'état de l'art proposés pour la classification de relations génériques sur notre corpus a donné des résultats prometteurs, ce qui est un premier pas vers l'intelligence économique et concurrentielle à partir de textes pour le français. La prochaine étape est l'adaptation de ces modèles pour prendre en compte les spécificités des relations économiques.

Références

AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R. & IVES Z. G. (2007). Dbpedia : A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 de *LNAI*, p. 722–735 : Springer. DOI : 10.1007/978-3-540-76298-0_52.

AUSSENAC-GILLES N. & JACQUES M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with caméléon. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, **14**(1), 45–73.

AUSSENAC-GILLES N., KAMEL M., BUSCALDI D. & COMPAROT C. (2013). Construction d'ontologies à partir de pages web structurées. In *Journées Francophones d'Ingénierie des Connaissances (IC 2013), Lille*, p. 1–17 : AFIA.

BALDINI SOARES L., FITZGERALD N., LING J. & KWIATKOWSKI T. (2019). Matching the blanks : Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2895–2905, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1279.

BOLLACKER K., EVANS C., PARITOSH P., STURGE T. & TAYLOR J. (2008). Freebase : A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, p. 1247–1250, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1376616.1376746.

BOSSY R., DELÉGER L., CHAIX E., BA M. & NÉDELLEC C. (2019). Bacteria biotope at bionlp open shared tasks 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, p. 121–131.

BRAUN D., FABER A., HERNANDEZ-MENDEZ A. & MATTHES F. (2018). Automatic relation extraction for building smart city ecosystems using dependency parsing. In P. BASILE, V. BASILE, D. CROCE, F. DELL'ORLETTA & M. GUERINI, Éds., *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018), Trento, Italy, November 22nd to 23rd, 2018*, volume 2244 de *CEUR Workshop Proceedings*, p. 29–39 : CEUR-WS.org.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

COLLINS M. & DUFFY N. (2001). Convolution kernels for natural language. In *Proceedings of the 14th International Conference on Neural Information Processing Systems : Natural and Synthetic*, NIPS'01, p. 625–632, Cambridge, MA, USA : MIT Press.

CULOTTA A. & SORENSEN J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 423–429, Barcelona, Spain. DOI : 10.3115/1218955.1219009.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DOS SANTOS C., XIANG B. & ZHOU B. (2015). Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 626–634, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1061.

FAN Z., SOLDAINI L., COHAN A. & GOHARIAN N. (2018). Relation extraction for protein-protein interactions affected by mutations. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, p. 506–507, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3233547.3233617.

FREITAS C., SANTOS D., MOTA C., GONÇALO OLIVEIRA H. & CARVALHO P. (2009). Relation detection between named entities : report of a shared task. In *Proceedings of the Workshop on Semantic Evaluations : Recent Achievements and Future Directions (SEW-2009)*, p. 129–137, Boulder, Colorado : Association for Computational Linguistics.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2 : The 15th International Conference on Computational Linguistics*, p. 539–545 : Association for Computational Linguistics.

HENDRICKX I., KIM S. N., KOZAREVA Z., NAKOV P., SÉAGHDHA D. O., PADÓ S., PENNACCHIOTTI M., ROMANO L. & SZPAKOWICZ S. (2010). Semeval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, p. 33–38, USA : Association for Computational Linguistics.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

JOULIN A., GRAVE E., BOJANOWSKI P., DOUZE M., JÉGOU H. & MIKOLOV T. (2016). Fasttext.zip : Compressing text classification models. arXiv preprint : 1612.03651.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431, Valencia, Spain : Association for Computational Linguistics. .

KAMBHATLA N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, p. 22–es, USA : Association for Computational Linguistics. DOI : 10.3115/1219044.1219066.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. arXiv preprint : 1412.6980.

LAU R. & ZHANG W. (2011). Semi-supervised statistical inference for business entities extraction and business relations discovery. In *SIGIR 2011 workshop, Beijing, China*, p. 41–46.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of 12th Language Resources and Evaluation Conference (LREC 2020)*. arXiv : 1912.05372.

LEE J., SEO S. & CHOI Y. (2019). Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. *Symmetry*, **11**(6). DOI : 10.3390/sym11060785.

LEE J. Y., DERNONCOURT F. & SZOLOVITS P. (2017). MIT at SemEval-2017 task 10 : Relation extraction with convolutional neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 978–984, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/S17-2171.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. arXiv preprint : 1907.11692.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model. arXiv preprint : 1911.03894.

MITCHELL A., STRASSEL S., HUANG S. & ZAKHARY R. (2005). Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*.

NGUYEN D. P. T., MATSUO Y. & ISHIZUKA M. (2007). Relation extraction from wikipedia using subtree mining. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, p. 1414–1420 : AAAI Press.

NGUYEN T. H. & GRISHMAN R. (2015). Relation extraction : Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, p. 39–48, Denver, Colorado : Association for Computational Linguistics. DOI : 10.3115/v1/W15-1506.

SHEN Y. & HUANG X. (2016). Attention-based convolutional neural network for semantic relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 2526–2536, Osaka, Japan : The COLING 2016 Organizing Committee.

SRIVASTAVA N., HINTON G., KRIZHEVSKY A., SUTSKEVER I. & SALAKHUTDINOV R. (2014). Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, **15**(1), 1929–1958.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, p.9̃.

TAO Q., LUO X., WANG H. & XU R. (2019). Enhancing relation extraction using syntactic indicators and sentential contexts. In *IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 1574–1580. DOI : 10.1109/ICTAI.2019.00227.

WANG L., CAO Z., DE MELO G. & LIU Z. (2016). Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1298–1307, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1123.

WU S. & HE Y. (2019). Enriching pre-trained language model with entity information for relation classification. arXiv preprint : 1905.08284.

XIAO M. & LIU C. (2016). Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1254–1263, Osaka, Japan : The COLING 2016 Organizing Committee.

YAMAMOTO A., MIYAMURA Y., NAKATA K. & OKAMOTO M. (2017). Company relation extraction from web news articles for analyzing industry structure. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, p. 89–92 : IEEE.

YAN C., FU X., WU W., LU S. & WU J. (2019). Neural network based relation extraction of enterprises in credit risk management. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, p. 1–6 : IEEE.

ZENG D., LIU K., LAI S., ZHOU G. & ZHAO J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 2335–2344, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

ZHANG Y., LIN H., YANG Z., WANG J., ZHANG S., SUN Y. & YANG L. (2018). A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, **81**, 83–92.

ZHANG Y., ZHONG V., CHEN D., ANGELI G. & MANNING C. D. (2017). Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 35–45, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1004.

ZHAO J., JIN P. & LIU Y. (2010). Business relations in the web : Semantics and a case study. *Journal of Software*, **5**(8), 826–833.

ZHOU D., ZHONG D. & HE Y. (2014). Biomedical relation extraction : from binary to complex. *Computational and mathematical methods in medicine*, **2014**.

ZHOU G., SU J., ZHANG J. & ZHANG M. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 427–434, Ann Arbor, Michigan : Association for Computational Linguistics. DOI : 10.3115/1219840.1219893.

ZHOU P., SHI W., TIAN J., QI Z., LI B., HAO H. & XU B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 207–212, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-2034.

ZUO Z., LOSTER M., KRESTEL R. & NAUMANN F. (2017). Uncovering business relationships : Context-sensitive relationship extraction for difficult relationship types. In M. LEYER, Éd., *Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Rostock, Germany, September 11-13, 2017*, volume 1917 de *CEUR Workshop Proceedings*, p. 271 : CEUR-WS.org.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Représentation dynamique et spécifique du contexte textuel pour l'extraction d'événements

Dorian Kodelja Romaric Besançon Olivier Ferret
CEA, LIST, Laboratoire Analyse Sémantique Texte et Image, Gif-sur-Yvette, F91191 France
dorian.kodelja, romaric.besancon, olivier.ferret@cea.fr

RÉSUMÉ

Dans cet article, focalisé sur l'extraction supervisée de mentions d'événements dans les textes, nous proposons d'étendre un modèle opérant au niveau phrastique et reposant sur une architecture neuronale de convolution de graphe exploitant les dépendances syntaxiques. Nous y intégrons pour ce faire un contexte plus large au travers de la représentation de phrases distantes sélectionnées sur la base de relations de coréférence entre entités. En outre, nous montrons l'intérêt d'une telle intégration au travers d'évaluations menées sur le corpus de référence TAC Event 2015.

ABSTRACT

Dynamic and specific textual context representation for event extraction.

In this paper, which focuses on the supervised detection of event mentions in texts, we propose to extend a neural sentence level model based on graph convolution exploiting syntactic dependencies. To do so, we integrate a larger context through the representation of distant sentences selected on the basis of co-reference relations between entities. We show the interest of such an integration through evaluations carried out on the TAC Event 2015 reference corpus.

MOTS-CLÉS : Extraction d'information événementielle, convolution de graphe, contexte.

KEYWORDS: Event information extraction, graph convolution, context.

1 Introduction

Le travail présenté dans cet article se focalise sur l'extraction d'événements supervisée à partir de textes (Grishman, 2019; Xiang & Wang, 2019; Kodelja *et al.*, 2019b). Cette tâche, incarnée en particulier par les évaluations ACE 2005 (Doddington *et al.*, 2004) et TAC Event (Getman *et al.*, 2018), consiste à identifier dans des textes les mots ou séquences de mots, appelés mentions d'événements, marquant la présence d'un type d'événement défini a priori. Par exemple, le mot « pow-wow » de la phrase suivante :

Putin had invited Tony Blair to the ***pow-wow*** *in Saint Petersburg's Grand Hotel Europe.*

est à extraire pour marquer la présence d'un événement de type *Meet*. Diverses méthodes ont été élaborées au fil du temps pour réaliser cette tâche mais les meilleures performances sont obtenues actuellement par des méthodes présentant deux caractéristiques principales : elles sont fondées sur des architectures neuronales et opèrent principalement à l'échelle phrastique, à l'image de (Nguyen & Grishman, 2018). Néanmoins, se limiter à cette échelle ne permet pas toujours de disposer de

Volume 2 : Traitement Automatique des Langues Naturelles, pages 40–53.
HAL : hal-02784754.

tous les éléments nécessaires à une bonne décision. C'est pourquoi un certain nombre de travaux se sont attachés à exploiter des informations au-delà de la phrase pour extraire de celle-ci des mentions d'événements. Ces travaux peuvent schématiquement se répartir en deux grandes catégories : ceux exploitant des informations au niveau du document pour réaliser une extraction à un niveau local et ceux opérant une extraction plus collective à l'échelle du document.

Les premiers sont représentés historiquement par (Liao & Grishman, 2010) et plus récemment par (Kodelja *et al.*, 2019a), avec une approche par amorçage consolidant au niveau du document les prédictions réalisées à un niveau local afin d'améliorer ces dernières. Sont ainsi prises en compte les dépendances entre types d'événements au niveau du document. (Hong *et al.*, 2011) représente de ce point de vue une vision plus centrée sur les relations entre les types d'entités et les types d'événements, ces relations étant là aussi envisagées à l'échelle du document. Une autre perspective est d'intégrer à un niveau local une représentation globale du document. Duan *et al.* (2017) le font ainsi en s'appuyant sur une méthode de représentation générique, en l'occurrence *Doc2Vec* (Le & Mikolov, 2014). Zhao *et al.* (2018) définissent pour leur part un modèle hiérarchique neuronal de représentation des documents en relation avec la tâche d'extraction d'événements.

Le travail de Chen *et al.* (2018) peut être vu comme une forme d'association entre les deux catégories distinguées ci-dessus : il repose pour une bonne part sur l'exploitation au niveau phrastique des informations présentes à l'échelle du document, exploitation réalisée dans ce cas de façon plus sélective grâce à des mécanismes d'attention, qu'intégratrice, mais l'extraction des événements au niveau phrastique s'effectue de façon collective. Reichart & Barzilay (2012), Liu *et al.* (2016) ainsi que Yang & Mitchell (2016) implémentent des approches collectives opérant quant à elles au niveau du document. La première articule l'utilisation de modèles de type Conditional Random Field (Lafferty *et al.*, 2001) à un niveau local et de contraintes déclaratives à l'échelle du document par le biais d'une procédure de décomposition duale (Rush *et al.*, 2010). La deuxième repose sur le paradigme *Probabilistic Soft Logic* (Kimmig *et al.*, 2012) pour exprimer des contraintes globales sous la forme de formules logiques. Enfin, la dernière s'appuie sur un modèle graphique sous forme de graphe de facteurs (*factor graph*).

Nous proposons dans cet article une nouvelle méthode de prise en compte du contexte textuel pour l'extraction d'événements en étendant un modèle opérant au niveau phrastique fondé sur la convolution de graphe. Cette nouvelle méthode prend en compte de façon sélective le contexte textuel d'une mention d'événement candidate en intégrant la représentation de phrases distantes sur la base des relations de coréférence entretenues avec les entités entourant cette mention. Nous évaluons cette nouvelle méthode sur le corpus de référence TAC Event 2015 et montrons son intérêt en termes de résultats par rapport aux résultats de référence sur ce corpus.

2 Description de l'approche

Comme nous l'avons vu en introduction, l'extraction d'événements consiste à identifier dans un texte les mentions d'événements et leur associer un type selon une taxonomie préalablement établie. Dans cet article, nous nous appuyons sur les 38 types d'événements de la taxonomie DEFT Rich ERE (Linguistic Data Consortium, 2015) utilisée dans le cadre des campagnes TAC Event (Getman *et al.*, 2018). Ces 38 types sont organisés en 9 grands types pour lesquels nous donnons un exemple entre parenthèses : Business (Merge.Org), Conflict (Attack), Contact (Meet), Justice (Sentence), Life (Divorce), Manufacture (Artifact), Movement (Transport.Person), Personnel (Nominate) et Transaction

(Transfert.Money). Dans les annotations liées à cette taxonomie, les mentions d'événements sont en grande majorité des mots simples. Comme nous pouvons le constater dans le tableau 1, la proportion de mentions d'événements multi-mots pour les corpus d'évaluation que nous avons adoptés se situe aux alentours de 3 %, ce qui est faible. Dans le prolongement de la plupart des modèles neuronaux développés pour l'extraction d'événements, nous choisissons donc d'aborder le problème non pas comme une tâche d'annotation de séquences mais comme une tâche de classification multi-classe de mots, chacun des 38 types d'événements constituant l'une de ces classes, auxquelles s'ajoute une classe NULLE (absence d'événement, classe très majoritaire). Ce choix est d'un impact négatif négligeable mais simplifie la modélisation et permet l'introduction d'un vecteur de positions contribuant grandement aux performances (Nguyen *et al.*, 2016). Enfin, toujours dans la continuité de la plupart des approches neuronales récentes, nous nous plaçons à l'échelle intra-phrastique. Notre modèle est composé de deux sous-modèles : le premier est un modèle hybride opérant au niveau phrastique tandis que le second, un modèle récurrent, est appliqué au niveau inter-phrastique pour produire une représentation intégrée au premier modèle.

2.1 Modèle phrastique d'extraction d'événements

L'exploitation des informations contenues dans le contexte intra-phrastique joue un rôle capital dans la résolution de la tâche d'extraction d'événements. La compréhension du sens d'un mot en contexte dépend bien évidemment de certains a priori sur son sens en général tels qu'ils sont capturés par des représentations distribuées de mots. Cependant, ce sens contextualisé du mot se sélectionne, voire se construit, par interaction avec le contexte. C'est pourquoi il est nécessaire d'utiliser un modèle exploitant les autres mots de la phrase et la manière dont ils interagissent avec le mot cible. Bien que les modèles exploitant la forme de surface de la phrase (réseaux convolutifs (CNN) ou récurrents (RNN)) permettent d'obtenir des performances satisfaisantes en extraction, la prise en compte de l'arbre de dépendance de la phrase est également bénéfique, comme l'ont notamment montré (Orr *et al.*, 2018) et (Nguyen & Grishman, 2018).

Le modèle de convolution de graphe proposé par (Nguyen & Grishman, 2018) pour l'extraction d'événements ayant obtenu de bonnes performances, nous nous focaliserons par la suite sur cette architecture. Ce modèle est constitué de quatre composantes principales : une couche de plongement associant une représentation vectorielle à chaque mot en entrée spécifiant les différents traits qui lui sont associés ; un BiLSTM (Hochreiter & Schmidhuber, 1997) appliqué à ces représentations d'entrée permettant leur contextualisation dans l'espace de la phrase ; plusieurs couches de convolution de graphe opérant sur ces représentations contextualisées afin de prendre en compte l'environnement des mots par le biais des relations de dépendance syntaxique ; enfin, une couche de pooling permettant d'agréger les représentations ainsi produites pour chaque mot de la phrase avant la couche de classification pour le mot cible. Nous présentons ici plus en détail ces composantes.

2.1.1 Représentation et contextualisation des entrées

Comme nous l'avons vu ci-dessus, la tâche d'extraction d'événements prend la forme d'une classification multiclasse pour chaque mot de la phrase. Pour chaque candidat w_t, nous créons une représentation qui lui est spécifique de la phrase $S = (w_1, w_2, \ldots, w_n)$ dans laquelle il apparaît. Pour ce faire, la représentation x_i de chaque mot w_i est obtenue en concaténant les représentations vectorielles réelles suivantes correspondant aux différents traits associés aux mots en entrée.

— **Plongement de mot :** cette représentation encode les propriétés distributionnelles du mot w_i. Elle est initialisée à partir de représentations produites sur la base d'un grand corpus non annoté.

— **Plongement de position :** pour chaque mot w_i de la phrase, sa distance $i - t$ au mot cible w_t est calculée. Un dictionnaire de vecteurs initialisés aléatoirement associe les différentes distances possibles à des vecteurs.

— **Plongement d'entités :** les entités nommées jouent un rôle important en extraction d'information car elles sont généralement les arguments des relations à extraire. Dans le cas présent, elles sont plus précisément les participants des événements à extraire et sont donc partie prenante, à un niveau plus local, de relations entre les mentions d'événements et les participants de ces événements. Même si toutes les entités nommées d'une phrase ne sont pas nécessairement liées à une mention d'événements présente dans cette phrase, elles constituent néanmoins des indices potentiellement intéressants de sa présence. Du point de vue des représentations, comme pour les plongements de position, chaque type d'entité (dont le type NUL correspondant à l'absence d'entité pour un mot) est associé à un vecteur initialisé aléatoirement. Le type d'entité e_i du mot w_i permet ainsi d'obtenir son plongement d'entité.

La concaténation des représentations x_i de chaque mot w_i d'une phrase forme la séquence X :

$$X = (x_0, x_1, \ldots, x_n) \tag{1}$$

Celle-ci constitue une représentation de la phrase, centrée sur le mot cible w_t grâce aux plongements de position. Cette représentation est ensuite mise en entrée d'un modèle BiLSTM permettant de produire une représentation contextualisée de chaque mot de la phrase. Plus précisément, la partie *forward* de ce BiLSTM intègre la partie de la phrase allant de son début jusqu'au mot considéré tandis que la partie *backward* intègre la partie complémentaire, allant de la fin de la phrase jusqu'au mot considéré. Les réseaux de type LSTM ont vocation à prendre en compte un contexte large, ce qui pourrait sembler a priori suffisant pour couvrir la totalité d'un contexte phrastique. En réalité, l'expérience montre que l'horizon effectif de ce type de réseaux est beaucoup plus limité, ce qui explique d'ailleurs que les performances d'un réseau LSTM en extraction d'événements soient très comparables à celles d'un réseau convolutif, dont l'horizon est supposé plus réduit.

2.1.2 Convolution de graphe pour l'exploitation des dépendances syntaxiques

Afin de dépasser les limites des modèles convolutifs et récurrents, l'idée est de s'appuyer sur les relations de dépendance syntaxique pour élargir l'horizon des modèles de façon sélective. Les participants des événements, qui peuvent constituer des traits particulièrement discriminants pour les identifier, sont en effet fréquemment liés aux mentions d'événements par le biais de relations syntaxiques. Pour mettre en œuvre cette idée, Nguyen & Grishman (2018) ont proposé d'utiliser la notion de convolution de graphe (Kipf & Welling, 2017). Dans ce cadre, le voisinage d'un mot n'est plus constitué de son environnement séquentiel – les mots qui le précèdent et le suivent – mais des mots qui lui sont liés par le biais de relations de dépendance. La représentation d'un mot est alors produite en appliquant une convolution aux représentations des mots constituant ce voisinage.

De façon plus formelle, à partir d'une phrase S de n mots (w_1, w_2,...w_n), est construit le graphe $G = \{V, E\}$ tel que l'ensemble $V = \{w_1, w_2, \ldots, w_n\}$ de ses nœuds correspond aux mots de la phrase considérée et l'ensemble de ses arêtes E, à l'ensemble des relations de dépendance syntaxique

qui les lient. Pour chaque paire de mots (w_i,w_j) pour lesquels w_i et w_j sont respectivement gouverneur et gouverné d'une relation, L(w_i,w_j) indique le type de la dépendance en question.

Afin que la représentation d'un mot par le graphe tienne compte à la fois de la représentation d'entrée du mot, de ses gouvernants et de son gouverneur, l'ensemble des arêtes E est constitué de trois sous-ensembles d'arêtes dirigées et étiquetées.

— **Direct :** pour chaque dépendance syntaxique (w_i,w_j) de type L(w_i,w_j), nous ajoutons une arête de w_i vers w_j étiquetée par le type de la dépendance (p. ex. *nmod*).

— **Inverse :** nous produisons également une arête inverse, de w_j vers w_i, étiquetée par le type de la dépendance et suffixée par l'intitulé *inverse* (p. ex. *nmod_inverse*).

— **Self-loop :** pour chaque nœud du graphe, une arête vers lui-même de type *self-loop* est ajoutée au graphe. Contrairement aux précédents types d'arêtes, celui-ci ne traduit pas une relation syntaxique au sein de la phrase mais permet au modèle de prendre en compte la représentation propre du nœud à la couche précédente lors de la convolution.

Un modèle de convolution de graphe est constitué de K couches de convolutions appliquées à un graphe dont les arêtes peuvent être de différents types. Pour un nœud u de voisinage $N(u)$, sa représentation h_u^{k+1} à la couche $k+1$ est alors :

$$h_u^{k+1} = \sigma\left(\sum_{v \in N(u)} W_{L(u,v)}^k h_v^k + b_{L(u,v)}^k\right) \tag{2}$$

où $W_{L(u,v)}^k$ et $b_{L(u,v)}^k$ sont respectivement la matrice de poids et les biais associés au type de dépendance $L(u,v)$ entre u et v. σ est une fonction d'activation.

Afin de distinguer l'influence de différents voisins, une pondération des nœuds voisins est obtenue ainsi :

$$s_{(u,v)}^k = \sigma\left(h_v^k \overline{W}_{L(u,v)}^k + \bar{b}_{L(u,v)}^k\right) \tag{3}$$

En introduisant la pondération des voisins (3), l'équation de convolution de graphe (2) devient :

$$h_u^{k+1} = \sigma\left(\sum_{v \in N(u)} s_{(u,v)}^k \left(W_{L(u,v)}^k h_v^k + b_{L(u,v)}^k\right)\right) \tag{4}$$

Le nombre de paramètres des matrices $W_{L(u,v)}^k$, $\overline{W}_{L(u,v)}^k$ et des biais $b_{L(u,v)}^k$ et $\bar{b}_{L(u,v)}^k$ est proportionnel au nombre de types de dépendances syntaxiques. Or, les *Universal Dependencies* utilisées sont constituées de 37 dépendances différentes. Le modèle produisant également des dépendances inverses et des arêtes self-loop, il serait nécessaire d'utiliser 75 étiquettes différentes. Compte tenu de la taille relativement petite des jeux de données en extraction d'événements, il est préférable de restreindre le nombre de types de dépendances. Pour ce faire, et comme proposé par (Marcheggiani & Titov, 2017), nous limitons le nombre de relations syntaxiques à trois : lien direct, inverse et self-loop.

Pour la première couche du graphe, la représentation h_u^0 est la représentation contextualisée du mot x_u produite par le BiLSTM. Il est à noter que cette représentation est en pratique complémentaire de l'approche par graphe de convolution : elle permet de prendre en compte d'une façon que l'on sait efficace le contexte local du mot considéré tandis que la vocation du graphe de convolution est surtout de regarder au-delà de ce contexte local.

2.1.3 Pooling

Une fois produite la séquence des représentations vectorielles $h^k_{w_1}, h^k_{w_2}, \dots h^k_{w_n}$ de chaque mot par la dernière (K-ième) couche de convolution de graphe, il est nécessaire d'agréger cette séquence en une représentation p_t du mot cible w_t à fournir en entrée d'une couche linéaire dotée d'un softmax afin de réaliser la classification. Nguyen & Grishman (2018) comparent les méthodes existantes, *pooling cible* (extraction de la représentation du mot cible uniquement), *pooling global* (*max-pooling* sur l'ensemble des mots de la phrase), *multipooling dynamique* (concaténation des poolings globaux des contextes gauche et droit du mot cible) et proposent une nouvelle méthode d'agrégation tenant compte des entités de la phrase : le pooling d'entités.

Cette proposition est motivée par les limites des autres méthodes à tirer spécifiquement profit des représentations vectorielles produites par le graphe des entités. Comme évoqué précédemment, le nombre K de couches de convolution de graphe peut être insuffisant pour que l'information des entités distantes soit propagée jusqu'à la représentation finale h^k_t extraite par la méthode de pooling cible. De plus, la présence d'informations plus spécifiques aux entités, présentes dans leurs représentations propres, est ignorée par cette méthode. Pour ce qui est des deux autres méthodes, leur traitement indifférencié de l'ensemble des mots du contexte peut mener au rejet d'informations pertinentes des entités dans le cas où certaines représentations de mots non informatifs obtiendraient des valeurs plus élevées. Pour éviter ces écueils, la méthode de pooling d'entités consiste à appliquer un *max-pooling* uniquement aux mots cibles et aux entités de la phrase :

$$p_t = maxpool(\{h^K_{w_t}\} \cup \{h^K_{w_i} : 1 \leq i \leq n, e_i \neq \text{NUL}\}) \quad (5)$$

Cette méthode reposant sur une annotation fiable des entités, elle pourrait s'avérer moins efficace dans des cas où l'annotation des entités serait réalisée automatiquement. C'est pourquoi nous proposons un pooling intermédiaire entre le pooling d'entités et le pooling global, ne dépendant pas des entités mais se focalisant également sur les mots les plus porteurs de sens. Cette stratégie, que nous appelons pooling syntaxique, consiste à appliquer un max-pooling au mot cible et à l'ensemble des noms, verbes et adjectifs de la phrase.

2.2 Prise en compte du contexte inter-phrastique

Le modèle de convolution de graphe étant à même d'exploiter un contexte intra-phrastique distant, nous nous intéressons à présent à la prise en compte du contexte inter-phrastique, c'est-à-dire l'exploitation des autres phrases du document pour l'enrichissement de la représentation locale.

2.2.1 Problématique

À notre connaissance, seulement deux autres études se sont portées sur l'intégration d'un contexte distant pour l'extraction d'événements. Ces deux modèles produisent une représentation unique du document, utilisée de manière indifférenciée pour tous les exemples d'apprentissage. Nous faisons au contraire l'hypothèse qu'il est souhaitable de déterminer un contexte spécifique pour chaque exemple afin de produire une représentation du contexte inter-phrastique plus à même de résoudre les ambiguïtés locales spécifiques de la phrase d'exemple.

Dans le cadre de l'extraction d'événements, la présence d'entités communes, en tant qu'arguments potentiels d'événements similaires, nous semble un indice fort du lien contextuel entre deux phrases.

Nous supposons en effet que de telles phrases font référence à des événements proches (tels que différents événements judiciaires partageant un même accusé), successifs (succession d'une blessure puis de la mort), voire contiennent deux mentions d'un même événement.

Nous distinguons ici la notion d'entité, c'est-à-dire une instance unique et spécifique telle qu'une personne, un lieu ou une organisation, de celle de mention d'entité, c'est-à-dire la mention d'une entité au sein d'une phrase. Ainsi, dans les deux phrases suivantes provenant d'un même document du jeu de données ACE 2005 :

— « Putin had invited Tony Blair to the **pow-wow** in *Saint Petersburg*'s Grand Hotel Europe. »

— « But the *Saint Petersburg* **summit** ended without any formal declaration on Iraq. »

le mot « pow-wow » de la première phrase est déclencheur d'un événement *Meet*. Mais ce mot est particulièrement atypique, donc peu susceptible d'avoir été rencontré dans un ensemble d'entraînement. L'identification de la coréférence entre les deux mentions « Saint Petersburg » permet néanmoins de relier cette phrase à la seconde dans laquelle la présence de l'événement est plus évidente. Dans cet esprit, nous proposons de restreindre le contexte d'une phrase cible à l'ensemble des phrases contenant des mentions associées à des entités communes avec la phrase cible. Notre méthode consiste alors à extraire des représentations de ces phrases de contexte puis à les combiner aux représentations locales des mentions d'entités correspondantes de la phrase cible afin de les enrichir. Cette méthode peut donc se diviser en trois étapes que nous présentons dans la suite de cette section : l'identification des phrases de contexte, l'extraction des représentations des entités du contexte puis l'intégration de ces représentations au modèle local.

2.2.2 Sélection des phrases de contexte

À partir d'une phrase $S^j = (w_1^j, w_2^j, \ldots, w_n^j)$, pour chaque mot w_i^j tel que $e_i^j \neq \text{NUL}$, $E(e_i^j)$ désigne l'entité correspondant à la mention e_i^j. Pour deux phrases S^j et S^k, $links(S^j, S^k)$ est l'ensemble des liens d'entités entre les deux phrases.

$$links(S^j, S^k)) = \{(l, m) : E(e_l^j) = E(e_m^k)\} \tag{6}$$

Le contexte d'une phrase S^j est alors donné par $Links(S^j)$. Cette fonction produit l'ensemble des tuples associant une phrase de contexte et les liens qu'elle entretient avec la phrase cible :

$$Links(S^j) = \left\{ \left(S^k, links(S^j, S^k) \right) \ : \ j \neq k \right\} \tag{7}$$

2.2.3 Extraction des représentations du contexte

Nous souhaitons produire des représentations spécifiques pour chaque entité e_i^jde la phrase cible S^j. Nous définissons donc $Ent\text{-}Links(S^j, i)$, qui associe à une mention d'entité de la phrase cible l'ensemble des mentions d'entités en coréférence dans des phrases de contexte :

$$Ent\text{-}Links(S^j, i) = \left\{ (S^k, l) \ : \ S^k \in Links(S^j) \text{ et } E(e_i^j) = E(e_l^k) \right\} \tag{8}$$

Pour chacune de ces paires (S^k, l) du contexte, nous produisons une représentation d'entrée $X^{k,l} = x_1^{k,l}, x_2^{k,l}, \ldots, x_n^{k,l}$ similaire à celle présentée en section 2.1.1, à la différence du vecteur de position.

Ici, pour chaque mot, le vecteur de position exprime cette fois la distance par rapport à la position l de la mention d'entité e_l^k.

Deux méthodes d'extraction sont possibles : le mode *Finale* (éq. 9) consiste à concaténer les représentations finales des deux modèles récurrents tandis que le mode *Mention* (éq. 10) extrait les représentations à l'emplacement de l'entité.

$$\textbf{Finale} : h_{\textbf{contexte}}(w^{k,l}) = [h_{\text{forward}}(x_n^{k,l}); h_{\text{backward}}(x_1^{k,l})] \tag{9}$$

$$\textbf{Mention} : h_{\textbf{contexte}}(w^{k,l}) = [h_{\text{forward}}(x_l^{k,l}); h_{\text{backward}}(x_l^{k,l})] \tag{10}$$

$$h_{\textbf{contexte}}(w^{k,l}) \in \mathbb{R}^{2d_c}$$

où d_c est la dimension de la couche cachée des modèles *forward* et *backward*.

2.2.4 Intégration du contexte

Il est à présent nécessaire d'intégrer les représentations du contexte global de la mention d'entité e_i dans le contexte local. Cette représentation peut être intégrée à deux niveaux : soit sous la forme de plongements supplémentaires lors de la production de la représentation d'entrée, soit sous la forme d'un nœud supplémentaire dans le graphe, voisin de la mention d'entité. Ces deux modes d'intégration sont présentés à la figure 1. Que ce soit pour une intégration au niveau des plongements

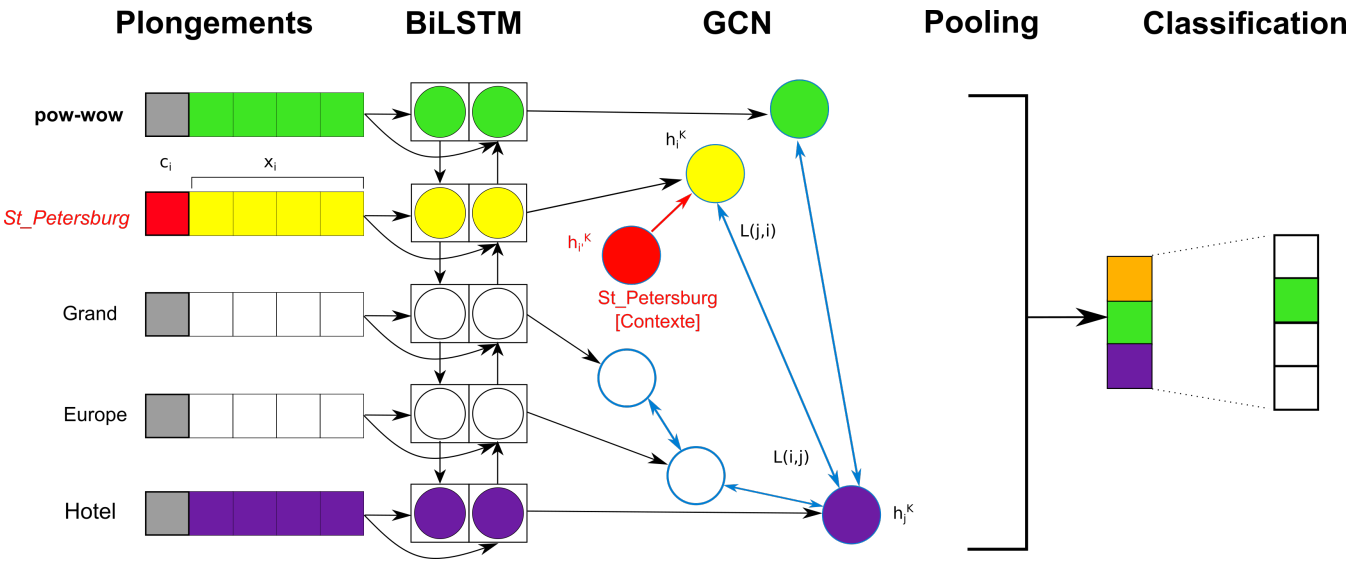


FIGURE 1 – Possibilités d'intégration d'une représentation contextuelle d'entité au sein d'un modèle de convolution de graphe. Cette représentation (en rouge) peut être intégrée en entrée du modèle ou dans le graphe par l'introduction d'un nouveau nœud connecté à la mention locale par un lien de type *contexte*. Le changement de couleur du jaune à l'orange explicite l'influence du contexte sur la représentation finale

ou du graphe, la représentation attendue est un vecteur. Nous agrégeons donc l'ensemble des vecteurs obtenus via une étape de max-pooling permettant d'obtenir le vecteur de contexte [1] :

$$contexte_i = maxpool(\{h_{\textbf{contexte}}(w^{k,l}) \ : \ (S^k, l) \in Ent\text{-}Links(S^j, i)\} \tag{11}$$

1. Par souci de concision et de cohérence avec la notation utilisée en section 2.1.1, nous n'utilisons pas l'index de phrase cible j en exposant de $contexte_i$.

Jeux de données	N^{bre} docs	N^{bre} phrases	N^{bre} mots	N^{bre} évts	N^{bre} évts/doc	% évts multi-mots
train	346	14 568	252 355	7 865	22,7	4,5
dév.	100	6 255	94 064	4 436	44,2	3,7
test	202	4 756	97 956	6 438	31,9	1,8

TABLE 1 – Tailles des jeux de données utilisés : nombre de documents (N^{bre} docs), de phrases (N^{bre} phrases), de mots (N^{bre} mots), de mentions d'événements (N^{bre} évts), nombre de mentions d'événements par document (N^{bre} évts/doc) et pourcentage des mentions d'événements de type multi-mot (% évts multi-mot)

Pour l'intégration au niveau des nœuds, nous modifions donc le graphe $G = \{V, E\}$ en ajoutant un nœud i' dans V associé au mot $w_{i'}$ avec la représentation initiale $h^0_{i'} = contexte_i$ et la représentation finale $h^K_{i'}$. Nous ne créons donc qu'un seul nœud agrégeant l'ensemble des représentations vectorielles distantes. Nous définissons alors un nouveau type d'arête *contexte* pour relier les mentions locales à leur représentation de contexte puis nous introduisons une arête supplémentaire $(w_i, w_{i'})$ de ce type dans E. Pour l'intégration au niveau des plongements, nous concaténons cette représentation aux autres plongements utilisés. Cependant, il est également nécessaire de créer une représentation par défaut pour les mots n'ayant pas de représentation de contexte. Cette représentation $c_{défaut}$ sera modifiée durant l'apprentissage. Le vecteur de contexte (11) est alors généralisé à l'ensemble des mots de la phrase en introduisant :

$$c_i = \begin{cases} contexte_i & \text{si } |Ent\text{-}Links(S^j, i)| > 0 \\ c_{défaut} & \text{sinon} \end{cases} \tag{12}$$

En y introduisant le vecteur (12), nous redéfinissons la séquence d'entrée (1) ainsi :

$$X = ([x_0, c_0], [x_1, c_1], \ldots, [x_n, c_n]) \tag{13}$$

3 Expériences

3.1 Données et prétraitements

Afin de nous comparer aux performances originelles du modèle de graphe présenté par (Nguyen & Grishman, 2018), nous nous évaluons sur le jeu de test TAC 2015. Nous partageons les données d'entraînement de TAC 2015 entre une partie pour l'apprentissage (58 documents) et une partie pour le développement (100 documents) en complétant les données d'apprentissage avec les jeux de données DEFT Rich ERE (R2 V2 et V2). Le tableau 1 présente un certain nombre de statistiques pour ces trois jeux de données : apprentissage (*test*), développement (*dév.*) et test (*test*). Les annotations en entités n'étant pas fournies, nous appliquons un modèle de reconnaissance d'entités nommées pour identifier les mentions. Il est alors nécessaire d'identifier les entités auxquelles ces mentions font référence. Pour ce faire, des outils de désambiguïsation d'entités (*entity linking*) pourraient sembler adaptés. Cependant, ces outils ayant pour objectif de rattacher les mentions d'entités à des entités spécifiques d'une ontologie, ils ne sont pas à même de traiter des mentions telles que « les trois

touristes » ou « l'agresseur », qui sont propres au document considéré. Nous appliquons donc plutôt un outil de résolution de coréférences. Nous redéfinissons alors la notion d'entité comme le groupe de coréférence auquel appartient une mention. Ce processus n'étant pas parfait, certaines mentions sont ignorées. Afin d'élargir la couverture du processus de coréférence, nous fusionnons donc les entités dont les mentions sont identiques.

3.1.1 Génération des exemples

Comme nous l'avons vu lors de la présentation de la convolution de graphe en section 2.1.2, le nombre de couches de convolution K correspond à la distance maximale séparant deux nœuds pouvant mutuellement s'influencer. Afin de faciliter l'accès aux mots supposés porteurs de sens dans la phrase, nous réalisons un filtrage préalable des mots de la phrase en fonction de leur étiquette morphosyntaxique. Nous supprimons ainsi les mots appartenant aux catégories suivantes : ponctuation, symbole, chiffre, déterminant, préposition, conjonction, interjection. Afin de préserver la connexité de l'arbre de dépendances syntaxiques, lorsqu'un mot supprimé est gouverneur d'autres mots, nous remplaçons le gouverneur de ces dépendances par le gouverneur du mot supprimé.

De plus, nous introduisons un masque de prédiction pour ne prédire que pour les noms, les verbes et les adjectifs. Les autres mots sont automatiquement associés à l'étiquette de la classe NULLE. Ce masque permet ainsi de réduire sensiblement le nombre d'exemples négatifs dans le jeu de données en ne perdant que très peu d'exemples positifs. Cette étape présente un double intérêt. D'une part la réduction importante de la taille des jeux de données se traduit par des temps d'apprentissage et de prédiction plus rapides. D'autre part, ce filtrage permet de réduire l'important déséquilibre entre la classe NULLE et les autres classes.

3.1.2 Hyperparamètres

Nous nous appuyons sur la suite d'outils linguistiques Stanford CoreNLP (Manning *et al.*, 2014) pour réaliser l'extraction d'entités nommées, la résolution de coréférences et l'analyse en dépendances utilisée pour produire les graphes. Concernant les entités nommées, nous exploitons l'ensemble étendu de 24 types couvrant des types de base tels que PERSON, LOCATION ou ORGANIZATION mais également des types relevant moins directement de la notion d'entité nommée comme IDEOLOGY ou CAUSE_OF_DEATH. Du point de vue syntaxique, les dépendances considérées sont les dépendances *Basic dependencies*. La matrice de poids des plongements de mots, à 300 dimensions, est initialisée à partir des plongements pré-entraînés GloVe (Pennington *et al.*, 2014). Les plongements de position et de type d'entités sont de taille 50 et les dimensions du BiLSTM du modèle local et des couches de convolution de graphe sont respectivement de 400 et 300. Les plongements de mots, d'entités et de distances sont les mêmes pour les phrases cibles et de contexte. Nous avons repris de (Nguyen & Grishman, 2018) un nombre K de couches de convolution égal à 2, les performances se dégradant à partir de 3, sans doute parce qu'au-delà d'un pas de deux dans le graphe de dépendances, le ratio entre les informations pertinentes observées par rapport à ce que l'on peut considérer comme du bruit vis-à-vis de la tâche devient trop faible. De même, la fonction d'activation σ est aussi la fonction ReLU. Le modèle est entraîné via SGD avec momentum avec des lots de 10 exemples. Les autres paramètres sont déterminés par optimisation bayésienne d'hyperparamètres grâce à hyperopt [2], les

2. https://github.com/hyperopt/hyperopt

	P	R	$F_{max.}$	$F_{moy.}$	F_{σ}
C-GCN	63,39	**57,34**	**60,51**	**60,19**	0,20
Intégration - Nœud	**65,53**	55,40	60,39	59,96	0,38
Pooling - Entité	63,35	57,07	60,49	59,96	0,30
Extraction - Mention	63,11	57,07	60,40	59,86	0,44
Pooling - Cible	62,14	56,92	59,99	59,37	0,36

TABLE 2 – Performances en développement suivant les choix de modélisation (P : précision, R : rappel, $F_{max.}$: F-mesure maximale sur 10 reproductions, $F_{moy.}$: F-mesure moyenne, F_{σ} : écart-type de la F-mesure)

configurations présentées étant sélectionnées à l'aide des performances en développement. Toutes les performances moyennes fournies sont calculées pour 10 reproductions avec les mêmes paramètres.

3.2 Étude des hyperparamètres du modèle

Nous étudions en premier lieu l'influence des différents choix de modélisation présentés :

— **Extraction :** *Finale*/*Mention* (cf. section 2.2.3).

— **Intégration :** *Plongement*/*Nœud* (cf. section 2.2.4).

— **Pooling :** *Cible*/*Syntaxique*/*Entité* (cf. section 2.1.3).

Nous avons réalisé sur le jeu de développement une recherche de valeur optimale pour ces différents paramètres ainsi que pour les paramètres d'optimisation (learning rate, régularisation l2, dropout, momentum). Le meilleur modèle obtenu utilise l'extraction *Finale*, l'intégration *Plongements* et le pooling *Syntaxique*. Cette tendance se confirme également en observant les autres configurations explorées lors de la recherche des meilleures valeurs pour les hyperparamètres. Il n'est cependant pas aisé de résumer directement cet ensemble d'expériences. C'est pourquoi nous présentons dans le tableau 2 les performances du meilleur modèle, C-GCN, et des versions obtenues en modifiant chacun des paramètres présentés. Le pooling cible est très significativement inférieur ($p < 0,001$) au pooling syntaxique utilisé par le modèle C-GCN, ce qui indique que la représentation du nœud à prédire n'est pas suffisante pour en prédire le type. Nous observons également que le pooling des entités obtient des performances légèrement inférieures au pooling syntaxique bien que cette différence soit peu significative ($p = 0,058$). L'exploitation d'un ensemble plus large de mots étant bénéfique au modèle, nous en déduisons que les représentations des entités de la phrase ne suffisent pas à enrichir la représentation du nœud cible. Notre pooling syntaxique est relativement proche du pooling *overall* proposé par Nguyen & Grishman (2018) qui obtient dans l'article d'origine des performances plus basses que le pooling des entités. Puisque nous utilisons un système d'extraction d'entités différent de celui utilisé par Nguyen & Grishman (2018), nous supposons que cette différence de performance est liée à la qualité des entités détectées.

Les moindres performances de l'extraction au niveau des mentions peuvent également s'expliquer par l'imprécision des entités ou simplement par le fait que les représentations finales des phrases de contexte sont plus informatives que les représentations spécifiques des mentions d'entités du contexte. Enfin, concernant l'intégration de la représentation du contexte au modèle, l'intégration au niveau des nœuds ne dégrade pas de manière significative les performances mais produit un profil plus déséquilibré entre précision et rappel.

	P	R	$F_{max.}$	$F_{moy.}$	F_{σ}
GCN_{repro}	78,48	46,96	59,1	58,73	0,82
$C\text{-}GCN_{générique}$	74,50	48,35	59,04	58,64	0,57
C-GCN	75,57	50,42	60,35	**60,47**	0,64
GCN_{nguyen}	70,3	50,6	58,8	-	-
RPI_BLENDER	75,23	47,74	58,41	-	-

TABLE 3 – Performances sur TAC 2015 test

3.3 Comparaison avec l'état de l'art

Nous comparons maintenant notre réimplémentation du modèle de graphe $\mathbf{GCN_{repro}}$ et notre proposition d'extension **C-GCN** à l'implémentation originale $\mathbf{GCN_{nguyen}}$ ainsi qu'au meilleur modèle de la campagne TAC 2015, **RPI_BLENDER** (Hong *et al.*, 2015), fondé sur un classifieur d'entropie maximale utilisant un large ensemble de traits. Afin de confirmer l'intérêt d'un contexte spécifique pour chaque exemple, nous entraînons également $\mathbf{C\text{-}GCN_{générique}}$, exploitant l'ensemble des phrases du document en tant que contexte. Dans ce cas, le vecteur de position n'intervient pas pour les phrases de contexte et la représentation produite sert plongement pour l'ensemble des mots de la phrase cible. Les résultats présentés dans le tableau 3 confirment l'intérêt de notre proposition. En effet, l'introduction de notre représentation de contexte apporte un gain de 1,74 point en F-mesure et permet de dépasser le modèle GCN_{nguyen} détenant jusqu'alors la meilleure performance sur TAC 2015 test, ainsi que le meilleur modèle de la campagne ayant recours à un ensemble large d'attributs définis manuellement. Nous constatons également que l'intégration de la représentation de contexte dans $C\text{-}GCN_{générique}$ ne permet pas d'améliorer les performances par rapport au modèle local, confirmant la pertinence de notre motivation première concernant l'intérêt de fournir un contexte spécifique.

4 Conclusion et perspectives

Nous avons proposé dans cet article une extension d'un modèle de convolution de graphe permettant la prise en compte du contexte inter-phrastique. Cette méthode consiste, à partir d'une phrase cible, à générer une représentation de phrases distantes dans lesquelles apparaissent des mentions d'entités également mentionnées dans la phrase cible. Cette représentation est ensuite utilisée pour enrichir la représentation d'entrée des mentions correspondantes de la phrase cible. L'évaluation de cette méthode sur le jeu de test TAC 2015 permet d'obtenir un gain significatif par rapport au modèle initial, les performances obtenues étant par ailleurs les meilleures sur ce jeu de données.

Notre modèle global étant tributaire de la qualité des étapes d'extraction des mentions d'entités et de leurs liens, une première piste d'extension évidente consiste à étudier leur impact exact sur les résultats en considérant d'autres outils pour l'extraction d'entités nommées et la résolution de coréférences. Par ailleurs, l'utilisation de modèles de désambiguïsation d'entités (*entity linking*) pourrait également étendre les liens entre entités et donc les contextes considérés. De façon complémentaire, substituer au max-pooling utilisé pour agréger les différentes mentions d'entités un mécanisme d'attention permettrait d'apprendre à discriminer et filtrer les phrases de contexte de façon plus fine. Dans le même esprit, un tel mécanisme pourrait aussi être utilisé pour trouver un équilibre plus optimal entre le nombre de couches de convolution et la prise en compte du contexte phrastique par le BiLSTM.

Références

CHEN Y., YANG H., LIU K., ZHAO J. & JIA Y. (2018). Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 1267–1276, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1158.

DODDINGTON G., MITCHELL A., PRZYBOCKI M., RAMSHAW L., STRASSEL S. & WEISCHEDEL R. (2004). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *4^{th} Conference on Language Resources and Evaluation (LREC 2004)*, p. 837–840, Lisbon, Portugal : European Language Resources Association (ELRA).

DUAN S., HE R. & ZHAO W. (2017). Exploiting Document Level Information to Improve Event Detection via Recurrent Neural Networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 1, p. 352–361.

GETMAN J., ELLIS J., STRASSEL S., SONG Z. & TRACEY J. (2018). Laying the Groundwork for Knowledge Base Population : Nine Years of Linguistic Resources for TAC KBP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Languages Resources Association (ELRA).

GRISHMAN R. (2019). Twenty-five years of information extraction. *Natural Language Engineering*, **25**(6), 677—692. DOI : 10.1017/S1351324919000512.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(9), 1735–1780. DOI : 10.1162/neco.1997.9.8.1735.

HONG Y., LU D., YU D., PAN X., WANG X., CHEN Y., HUANG L. & JI H. (2015). RPI_BLENDER TAC-KBP2015 System Description. In *Proceedings of the 2015 Text Analysis Conference*.

HONG Y., ZHANG J., MA B., YAO J., ZHOU G. & ZHU Q. (2011). Using Cross-Entity Inference to Improve Event Extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, volume 1, p. 1127–1136 : ACL.

KIMMIG A., BACH S., BROECHELER M., HUANG B. & GETOOR L. (2012). A short introduction to probabilistic soft logic. In *NIPS Workshop on Probabilistic Programming : Foundations and Applications*, p. 1–4.

KIPF T. N. & WELLING M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

KODELJA D., BESANÇON R. & FERRET O. (2019a). Exploiting a More Global Context for Event Detection Through Bootstrapping. In *Proceedings of the 41st European Conference on Information Retrieval* , p. 763–770.

KODELJA D., BESANÇON R. & FERRET O. (2019b). Modèles neuronaux pour l'extraction supervisée d'événements : état de l'art. *Traitement Automatique des Langues (TAL), numéro varia*, **60**(1), 13–37.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *Eighteenth International Conference on Machine Learning (ICML'01)*, p. 282–289, Williamstown, MA, USA.

LE Q. & MIKOLOV T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, volume 2 de *ICML'14*, p. 1188–1196, Beijing, China : JMLR.org.

LIAO S. & GRISHMAN R. (2010). Using Document Level Cross-Event Inference to Improve Event Extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 789–797, Uppsala, Sweden : ACL.

LINGUISTIC DATA CONSORTIUM (2015). *DEFT Rich ERE Annotation Guidelines : Events v.2.6*. Rapport technique.

LIU S., LIU K., HE S. & ZHAO J. (2016). A Probabilistic Soft Logic Based Approach to Exploiting Latent and Global Information in Event Classification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, AZ, USA : AAAI Press.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), System Demonstrations*, p. 55–60.

MARCHEGGIANI D. & TITOV I. (2017). Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1506–1515.

NGUYEN T. H. & GRISHMAN R. (2018). Graph Convolutional Networks with Argument-Aware Pooling for Event Detection. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA : AAAI Press.

NGUYEN T. H., GRISHMAN R. & MEYERS A. (2016). New York University 2016 System for KBP Event Nugget : A Deep Learning Approach. In *Proceedings of the 2016 Text Analysis Conference*, Gaithersburg, MD, USA : NIST.

ORR J. W., TADEPALLI P. & FERN X. (2018). Event Detection with Neural Networks : A Rigorous Empirical Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods on Natural Language Processing*, Brussels, Belgium : ACL.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, p. 1532–1543, Doha, Qatar : ACL.

REICHART R. & BARZILAY R. (2012). Multi-event extraction guided by global constraints. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2012)*, p. 70–79, Montréal, Canada.

RUSH A. M., SONTAG D., COLLINS M. & JAAKKOLA T. (2010). On dual decomposition and linear programming relaxations for natural language processing. In *2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, p. 1–11, Cambridge, MA.

XIANG W. & WANG B. (2019). A survey of event extraction from text. *IEEE Access*, **7**, 173111–173137.

YANG B. & MITCHELL T. M. (2016). Joint Extraction of Events and Entities within a Document Context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 289–299, San Diego, California : ACL.

ZHAO Y., JIN X., WANG Y. & CHENG X. (2018). Document Embedding Enhanced Event Detection with Hierarchical and Supervised Attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, p. 414–419, Melbourne, Australia : ACL.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Les modèles de langue contextuels CamemBERT pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement

Louis Martin*[1,2,3] Benjamin Muller*[2,3] Pedro Javier Ortiz Suárez*[2,3]
Yoann Dupont[3] Laurent Romary[2] Éric Villemonte de la Clergerie[2]
Benoît Sagot[2] Djamé Seddah[2]
[1]Facebook AI Research, Paris, France
[2]Inria, Paris, France
[3]Sorbonne Université, Paris, France
louismartin@fb.com,yoa.dupont@gmail.com,prenom.nom@inria.fr.

Résumé

Les modèles de langue neuronaux contextuels sont désormais omniprésents en traitement automatique des langues. Jusqu'à récemment, la plupart des modèles disponibles ont été entraînés soit sur des données en anglais, soit sur la concaténation de données dans plusieurs langues. L'utilisation pratique de ces modèles — dans toutes les langues sauf l'anglais — était donc limitée. La sortie récente de plusieurs modèles monolingues fondés sur BERT (Devlin *et al.*, 2019), notamment pour le français, a démontré l'intérêt de ces modèles en améliorant l'état de l'art pour toutes les tâches évaluées. Dans cet article, à partir d'expériences menées sur CamemBERT (Martin *et al.*, 2019), nous montrons que l'utilisation de données à haute variabilité est préférable à des données plus uniformes. De façon plus surprenante, nous montrons que l'utilisation d'un ensemble relativement petit de données issues du web (4Go) donne des résultats aussi bons que ceux obtenus à partir d'ensembles de données plus grands de deux ordres de grandeurs (138Go).

Abstract

CamemBERT Contextual Language Models for French: Impact of Training Data Size and Heterogeneity

Contextual word embeddings have become ubiquitous in Natural Language Processing. Until recently, most available models were trained on English data or on the concatenation of corpora in multiple languages. This made the practical use of models in all languages except English very limited. The recent release of monolingual versions of BERT (Devlin *et al.*, 2019) for French established a new state-of-the-art for all evaluated tasks. In this paper, based on experiments on CamemBERT (Martin *et al.*, 2019), we show that pretraining such models on highly variable datasets leads to better downstream performance compared to models trained on more uniform data. Moreover, we show that a relatively small amount of web crawled data (4GB) leads to downstream performances as good as a model pretrained on a corpus two orders of magnitude larger (138GB).

Mots-clés : Modèles de langue contextuels, BERT, CamemBERT, impact jeu de données.

Keywords: Contextual language models, BERT, CamemBERT, dataset impact.

∗. Les trois premiers auteurs ont contribué à parts égales à ce travail

Volume 2 : Traitement Automatique des Langues Naturelles, pages 54–65.
HAL : hal-02784755.

1 Introduction

En préface à son *Introduction to Deep Learning*, Charniak (2019) évoque son scepticisme initial face à la révolution apportée par l'apprentissage profond de réseaux neuronaux au traitement automatique des langues :

> « (..) I can rationalize this since this is the third time neural networks have threatened a revolution but only the first time they have delivered. *(Charniak, 2019, page* XI*)* »

En effet, la surprise apportée par l'avènement des plongements lexicaux et le gain de performance qu'ils ont permis en peu de temps (Mikolov *et al.*, 2013; Pennington *et al.*, 2014; Mikolov *et al.*, 2018) n'a eu pour équivalent que le saut qualitatif apporté par la prise en compte du contexte dans les représentations vectorielles, permettant *de facto* une prise en charge effective de la polysémie et donc l'obtention de modèles plus efficaces et plus fins (Peters *et al.*, 2018; Akbik *et al.*, 2018). Ces avancées ont ouvert la voie à des modèles contextuels plus larges, entraînés sur des objectifs de modèles de langue (Dai & Le, 2015). Ces approches, qui reposaient au départ sur des architectures LSTM (Howard & Ruder, 2018), ont évolué vers des architecturs de type *Transformer*, avec notamment GPT2 (Radford *et al.*, 2019), BERT (Devlin *et al.*, 2019), RoBERTa (Liu *et al.*, 2019) et plus récemment ALBERT (Lan *et al.*, 2019) et T5 (Raffel *et al.*, 2019).
Bien que plusieurs modèles développés pour d'autres langues aient été publiés (modèles ELMo [1] pour le japonais, le portugais, l'allemand et le basque ; modèles BERT pour le chinois simplifié et classique (Devlin *et al.*, 2018) ou pour l'allemand (Chan *et al.*, 2019)), le différentiel quant à la taille de leurs données de pré-entraînement n'a pas permis l'émergence de travaux les comparant au modèle original. Cependant, des modèles multilingues reposant sur la concaténation de larges jeux de données (principalement basés sur Wikipedia) sont apparus (Devlin *et al.*, 2018; Conneau *et al.*, 2019) et ont permis des avancées notables via l'apprentissage par transfert (Pires *et al.*, 2019). Ce n'est toutefois que très récemment que des modèle monolingues à grande échelle ont été développés (Martin *et al.*, 2019; Le *et al.*, 2019; Virtanen *et al.*, 2019; Delobelle *et al.*, 2020) et ont permis de confirmer l'intérêt des modèles monolingues sur d'autres langues.

En ce qui concerne le français, Le *et al.* (2019) ont montré sur diverses tâches que leur modèle, FlauBERT, offrait un panel de performances équivalentes à celles de CamemBERT (Martin *et al.*, 2019), soulignant qui plus est la complémentarité des deux modèles sur des tâches d'analyse syntaxique. Sachant que ces modèles ont été entraînes sur des données *in fine* différentes bien que d'origine similaire (avec un filtrage plus intense et l'utilisation d'un équivalent francophone du *Bookcorpus* dans un cas, un filtrage principalement sur le bruit et l'identification de la langue cible dans l'autre), il est pertinent de s'interroger sur l'impact qu'ont les données de pré-entraînement, tant en termes de taille que de type de données, sur les performances des modèles de langue neuronaux contextuels. D'autres paramètres sont d'importance, en particulier la stratégie de *masking* utilisée (*subword* ou *whole-word* ?) et le nombre de couches et de têtes d'attention (modèle *Base* ou *Large* ?).

Nous présentons ici une série d'expériences construites autour de CamemBERT visant à répondre à ces questions. Nos résultats montrent que, contrairement à l'idée qui prévalait, il est possible d'obtenir des résultats étonnement bons, au niveau de l'état de l'art pour toutes les tâches ou presque, avec des modèles entraînés sur seulement 4Go de données. Le point essentiel est qu'il semble préférable d'utiliser des données à haute variabilité, éventuellement bruitées, plutôt que des données proprement éditées et stylistiquement homogènes telles qu'on peut en trouver dans des jeux de données tirés de Wikipedia. Ce résultat permet d'envisager l'entraînement de ce type de modèles pour

1. https://allennlp.org/elmo

des langues relativement peu dotées voire pour des domaines spécialisés, dans les cas où une stratégie de *fine-tuning* ne serait pas efficace.

2 Protocole expérimental

Nous reprenons le même environnement expérimental (paramètres, outils, métriques, etc.) que celui utilisé par Martin *et al.* (2019).

2.1 Modèles et architectures

BERT, RoBERTa et CamemBERT CAMEMBERT est basée sur ROBERTA (Liu *et al.*, 2019), une évolution de BERT (Devlin *et al.*, 2019) sur plusieurs plans, notamment par l'utilisation du *masked language model* comme seul objectif de pré-entrainement. Outre le modèle CAMEMBERT$_{\text{BASE}}$ originel entraîné avec 12 couches, 768 dimensions cachées et 12 têtes d'attention, soit 110M de paramètres, nous utilisons un CAMEMBERT$_{\text{LARGE}}$ entrainé avec 24 couches, 1024 dimensions cachées et 16 têtes d'attention, soit 340M paramètres.
Selon les expériences, nous évaluons nos modèles en fonction de plusieurs hyper-paramètres : (i) la stratégie de *masking* (*subword* ou *whole word*), (ii) l'architecture du modèle (*BASE* ou *LARGE*), (iii) le nombre d'étapes d'entraînement (steps) et (iv) l'utilisation du modèle via *fine-tuning* ou via l'extraction de plongements lexicaux.

Données d'entraînement Pour étudier l'impact des données d'entraînement sur les performances de CAMEMBERT, nous utilisons alternativement le sous-corpus français du corpus multilingue OSCAR extrait de Common Crawl (Ortiz Suárez *et al.*, 2019), un autre corpus extrait de Common Crawl nommé CCNET (Wenzek *et al.*, 2019) et un snapshot récent de la Wikipedia française.

— **OSCAR** (Ortiz Suárez *et al.*, 2019) est un ensemble de corpus monolingues extraits de Common Crawl (*dump* de novembre 2018). Les corpus ont été sélectionnés par un modèle de classification par langues en suivant l'approche de (Grave *et al.*, 2018) s'appuyant sur le classifieur linéaire FASTTEXT (Grave *et al.*, 2017; Joulin *et al.*, 2016) pré-entraîné sur les corpus Wikipedia, Tatoeba et SETimes, et couvrant 176 langues.
— **CCNet** (Wenzek *et al.*, 2019), un jeu de données extrait lui aussi de Common Crawl mais avec un filtrage différent de celui d'OSCAR. Il a été construit avec un modèle de langue entraîné sur Wikipedia, lui permettant ainsi de filtrer le bruit (code, tables, etc.). CCNET contient ainsi des documents plus longs en moyenne qu'OSCAR. Ce filtrage a pour effet de biaiser les données en leur donnant un aspect « Wikipedia » et nous permet de considérer CCNET comme se positionnant entre OSCAR, peu filtré voire bruité, et WIKIPEDIA, totalement édité.
— **Wikipedia**, un corpus homogène en termes de genre et de style. Nous utilisons le *dump* français officiel de Wikipedia (avril 2019). Le corpus est prétraité à l'aide de *WikiExtractor*[2].

Afin de pouvoir comparer équitablement l'impact du type de données de pré-entraînement, nous créons des échantillons aléatoires à partir de OSCAR et CCNET, et ce au niveau du document, de la même taille que celle de notre WIKIPEDIA, soit 4Go de texte brut non compressé. Ceci nous permet d'étudier également les effets de la taille des données d'entraînement sur les performances des modèles.

2. https://github.com/attardi/wikiextractor

Jeux de données et tâches d'évaluation Nous évaluons nos différents modèles en étiquetage morphosyntaxique, en analyse sytaxique, en reconnaissance d'entités nommées (NER) et en reconnaissance d'implication textuelle (*Natural Language Inference*, NLI), qui consiste à prédire la relation entre une phrase hypothèse et phrase prémisse (implication, contradiction, neutralité). Pour les évaluations en étiquetage morphosyntaxique (POS tagging) et analyse en dépendances (parsing), nous utilisons dans leurs versions *Universal Dependencies 2.2* (Nivre *et al.*, 2018) les corpus Sequoia (Candito & Seddah, 2012), UD French GSD, UD French Spoken et UD French ParTut. L'évaluation de la NER se fait sur l'instance du *French treebank* (Abeillé *et al.*, 2003) annotée en entités nommées par Sagot *et al.* (2012). Pour la tâche de NLI, nous utilisons la partie française du jeu de données XNLI (Conneau *et al.*, 2018) qui étend le corpus *Multi-Genre NLI)* (Williams *et al.*, 2018) [3].

Toutes nos expériences suivent les *splits* usuels et utilisent les métriques classiques associées à ces tâches (UPOS, LAS, F1 et exactitude) . La Table 1 présente des statistiques sur ces jeux de données.

Corpus	Taille (texte brut non compr.)	#tokens	#docs	tokens/doc quantiles : 5%	50%	95%
Wikipedia	4Go	990M	1.4M	102	363	2530
CCNet	135Go	31.9B	33.1M	128	414	2869
OSCAR	138Go	32.7B	59.4M	28	201	1946

TABLE 1 – Statistiques sur les jeux de données de pré-entrainement.

Corpus	#tokens	#phrases	Genres
GSD	389,363	16,342	Blogs, News Reviews, Wiki
Sequoia	68,615	3,099	Medical, News Non-fiction, Wiki
Spoken	34,972	2,786	Spoken
ParTUT	27,658	1,020	Legal, News, Wikis
FTB	350,930	27,658	News

TABLE 2 – Statistiques des corpus arborés utilisés en étiquetage morphosyntaxique, analyse en dépendance et NER.

2.2 Utilisation de CAMEMBERT pour des tâches en aval

Nous utilisons CAMEMBERT de deux façons. Dans la première, *fine-tuning*, nous affinons le modèle sur une tâche spécifique de bout en bout. Dans la seconde, nous extrayons de CAMEMBERT des plongements lexicaux contextuels figés. Les performances de ces deux approches complémentaires illustrent la qualité des représentations cachées que capture CAMEMBERT.

Fine-tuning Pour chaque tâche, nous ajoutons la couche prédictive pertinente au-dessus du modèle de CAMEMBERT. Suite au travail effectué sur BERT (Devlin *et al.*, 2019) en étiquetage de séquence, nous ajoutons une couche linéaire qui prend respectivement en entrée la dernière représentation cachée du token spécial <s> et la dernière représentation cachée du premier token de sous-mot de chaque mot. Pour l'analyse de dépendance, nous branchons une tête de prédiction de graphes *bi-affine* inspirée de Dozat & Manning (2017). Nous renvoyons le lecteur à cet article pour plus de détails sur ce module. Nous affinons CAMEMBERT sur XNLI en ajoutant une tête de classification composée d'une couche cachée avec une non-linéarité et une couche de projection linéaire, avec un dropout d'entrée pour chaque couche.

Nous affinons CAMEMBERT indépendamment pour chaque tâche et chaque ensemble de données. Nous optimisons le modèle en utilisant l'optimiseur Adam (Kingma & Ba, 2014) avec un taux d'apprentissage fixe. Nous effectuons une *grid-search* sur une combinaison de taux d'apprentissage

3. Seules les parties de validation et de test ont été manuellement traduites de l'anglais, la partie d'entraînement l'a été automatiquement (122k exemples d'entraînement, 2490 de développement et 5010 de test).

et de tailles de lots. Nous sélectionnons le meilleur modèle sur l'ensemble de validation parmi les 30 premières *epoch*. Pour la tache de NLI, nous utilisons les hyper-paramètres par défaut fournis par les auteurs de RoBERTa sur la tâche MNLI. [4] Bien que cela aurait pu encore accroître les performances, nous n'appliquons aucune technique de régularisation telle que le *weight decay*, *learning rate warm-up* ou un affinage discriminant, sauf dans le cas de NLI. En effet, les expériences de Martin *et al.* (2019) ont montré que ce n'était pas nécessaire étant donné qu'un affinage simple de CamemBERT a contribué à établir l'état de l'art sur toute les taches et surpasse les modèles BERT multilingues. [5] Les expériences d'étiquetage morpho-syntaxique, d'analyse syntaxique en dépendance et de reconnaissance d'entités nommées sont exécutées à l'aide de la bibliothèque Transformer d'HuggingFace étendue pour prendre en charge CamemBERT et l'analyse de dépendance (Wolf *et al.*, 2019). Les expériences NLI utilisent la bibliothèque FairSeq reposant sur l'implémentation de RoBERTa.

Plongements lexicaux Suivant en celà Straková *et al.* (2019) et Straka *et al.* (2019) pour mBERT et le BERT originel, nous utilisons aussi CamemBERT dans scénario d'extraction de plongements lexicaux. Afin d'obtenir une représentation pour un token donné, nous calculons d'abord la moyenne des représentations de chaque sous-mot dans les quatre dernières couches du Transformer, puis faisons la moyenne des vecteurs des sous-mot résultants.

Nous évaluons CamemBERT dans cette utilisation sous forme de plongements lexicaux dans des taches d'étiquetage morpho-syntaxique, d'analyse de dépendance et en NER, avec les implémentation open-source de Straková *et al.* (2019) et Straka *et al.* (2019) entraînés sur les jeux de données décrit auparavant. [6]

3 Facteurs influençant les performances des modèles

Dans cette section, nous étudions l'influence de plusieurs facteurs sur les performances des tâches aval. Dans ce but, nous produisons plusieurs versions de CamemBERT en faisant varier les données de pré-entraînement. Sauf indication contraire, nous utilisons l'architecture BASE et fixons le nombre d'étapes de pré-entraînement à 100k et permettons alors au nombre d'*epochs* de varier en conséquence (plus d'*epochs* pour des tailles de jeu de données plus petites).

3.1 Common Crawl vs. Wikipedia

Les résultats présentés à la Table 3 montrent que les modèles entraînés sur les versions réduites (4Go) d'OSCAR et de CCNet (issus tous deux de Common Crawl) obtiennent des performances constamment supérieures à celles du modèle entraîné sur Wikipedia, que l'on utilise les modèles en configuration *fine-tuning* ou comme sources de plongements lexicaux. Sans surprise, l'écart est plus grand sur les tâches impliquant des textes dont le genre et le style sont plus éloignés de Wikipédia, notamment pour l'étiquetage et l'analyse syntaxique du corpus French Spoken (transcriptions de

4. Voir https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.glue.md pour plus détails.
5. Résultat confirmé ensuite dans plusieurs travaux décrivant des modèles BERT monolingues, eg. (Le *et al.*, 2019).
6. UDPipe Future est disponible sur https://github.com/CoNLL-UD-2018/UDPipe-Future, et le code pour le *nested NER* est disponible sur https://github.com/ufal/acl2019_nested_ner.

Dataset	Size	GSD		Sequoia		Spoken		ParTUT		**Average**		NER	NLI
		UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1	Acc.
Fine-tuning													
Wiki	4GB	98.28	93.04	98.74	92.71	96.61	79.61	96.20	89.67	97.45	88.75	89.86	78.32
CCNet	4GB	98.34	93.43	98.95	93.67	96.92	**82.09**	96.50	**90.98**	97.67	**90.04**	90.46	**82.06**
OSCAR	4GB	98.35	93.55	98.97	93.70	96.94	81.97	96.58	90.28	97.71	89.87	90.65	81.88
OSCAR	138GB	**98.39**	**93.80**	**98.99**	**94.00**	**97.17**	81.18	**96.63**	90.56	**97.79**	89.88	**91.55**	81.55
Plongements lexicaux (avec UDPipe Future (tagging, parsing) ou LSTM+CRF (NER))													
Wiki	4GB	98.09	92.31	98.74	93.55	96.24	78.91	95.78	89.79	97.21	88.64	91.23	-
CCNet	4GB	**98.22**	**92.93**	99.12	94.65	97.17	**82.61**	**96.74**	89.95	97.81	90.04	**92.30**	-
OSCAR	4GB	98.21	92.77	99.12	**94.92**	97.20	82.47	**96.74**	**90.05**	**97.82**	**90.05**	91.90	-
OSCAR	138GB	98.18	92.77	**99.14**	94.24	**97.26**	82.44	96.52	89.89	97.77	89.84	91.83	-

TABLE 3 – Résultats sur quatre tâches aval de modèles de langues entraînés avec des jeux de données d'homogénéité et de taille variable. Nous rapportons les scores sur les ensemble de validation de chaque tâche (moyenne de 4 expériences de *fine-tuning* en POS tagging, en parsing et en NER, moyenne de 10 expériences de fine-tuning en NLI).

l'oral, sans ponctuation). L'écart de performance est également important en NLI, probablement en raison de la plus grande diversité thématique et en genre dans les corpus issus de Common Crawl, que l'on retrouve probablement dans les données XNLI, lui même divers thématiquement et en genre, et combinant données orales et écrites.

3.2 De combien de données avons-nous besoin ?

Un résultat inattendu de nos expériences est que le modèle CamemBERT standard, entraîné sur l'ensemble des 138Go de texte d'OSCAR, ne surpasse pas massivement le modèle entraîné « uniquement » sur l'échantillon de 4Go. Dans les configurations où le modèle de langue est utilisé comme plongements, le modèle entraîné sur 4Go conduit plus souvent à de meilleurs résultats que le CamemBERT standard entraîné sur 138Go, bien que les différences de scores soient rarement frappantes. Dans les configurations *fine-tuning*, le CamemBERT standard fonctionne généralement mieux que celui entraîné sur 4Go, mais là encore les différences sont toujours faibles.

En d'autres termes, lorsque les modèles sont entraînés sur des corpus tels que OSCAR et CCNet, hétérogènes en termes de genre et de style, 4Go de texte non compressé constitue un corpus de pré-entraînement suffisamment volumineux pour atteindre l'état de l'art avec l'architecture *BASE*, et notamment supérieurs dans tout les cas à ceux obtenus avec mBERT (pré-entraîné sur 60 Go de texte dans une centaine de langues). Cela remet en question la nécessité d'utiliser la totalité de très larges corpus tel qu'OSCAR ou CCNet lors du pré-entrainement de modèles tels que CamemBERT, sauf peut-être lorsque l'on utilise une architecture LARGE.

Cela signifie que des modèles de type CamemBERT peuvent être entraînés pour toutes les langues pour lesquelles un corpus varié d'au moins 4 Go peut être construit. OSCAR est disponible en 176 langues et fournit un tel corpus pour 38 langues. De plus, il est possible que des corpus légèrement plus petits (par exemple jusqu'à 1 Go) soient également suffisants pour entraîner des modèles de langue très performants.
Cependant, même avec une architecture BASE et 4 Go de données d'entraînement, la *validation loss* continue de diminuer au-delà de 100 000 *steps* (et 400 *epochs*). Cela suggère que nous sous-entraînons toujours sur le jeu de données de pré-entraînement de 4 Go, et qu'un entraînement plus long pourrait conduire à de meilleures performances. Quoiqu'il en soit, nos résultats ont été obtenus sur des modèles *BASE*, des recherches supplémentaires sont donc nécessaires pour confirmer la validité de nos résultats sur des architectures plus grandes et sur d'autres tâches plus complexes de

compréhension de la langue.

CORPUS	MASKING	ARCH.	#PARAM.	#STEPS	UPOS	LAS	NER	XNLI
	Stratégie de masking							
CCNET	*subword*	BASE	110M	100K	97.78	89.80	**91.55**	81.04
CCNET	*whole word*	BASE	110M	100K	**97.79**	**89.88**	91.44	**81.55**
	Taille du modèle							
CCNet	*whole word*	BASE	110M	100K	97.67	89.46	90.13	82.22
CCNet	*whole word*	LARGE	335M	100k	**97.74**	**89.82**	**92.47**	**85.73**
	Données d'entraînement							
CCNET	*whole word*	BASE	110M	100K	97.67	89.46	90.13	**82.22**
OSCAR	*whole word*	BASE	110M	100K	**97.79**	**89.88**	**91.44**	81.55
	Nombre de steps							
CCNet	*whole word*	BASE	110M	100k	**98.04**	89.85	90.13	82.20
CCNet	*whole word*	BASE	110M	500k	97.95	**90.12**	91.30	**83.04**

TABLE 4 – Comparaison des scores sur les ensemble de **Validation** des différents choix de conception. Les scores d'étiquetage morphosyntaxique et d'analyse syntaxique sont moyennés sur les 4 jeux de données.

3.3 Impact de la stratégie de *masking*

Dans le tableau 4, nous comparons les modèles entraînés avec une stratégie de *subword masking* à ceux en *whole word masking*. Le *whole word masking* a un impact positif sur les performances en NLI (mais seulement de 0,5 point de précision). À notre grande surprise et contrairement à l'anglais, cette stratégie de *masking* ne profite pas à des tâches de plus bas niveau (NER, étiquetage morphosyntaxique et analyse syntaxique).

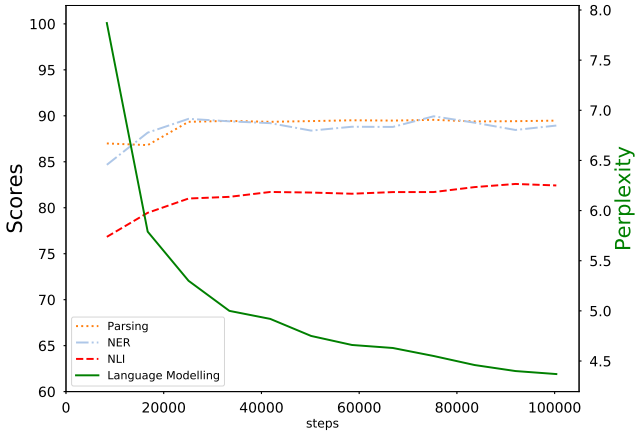
3.4 Impact de la taille du modèle

Le tableau 4 compare les modèles entraînés avec les architectures *BASE* et *LARGE*. Pour des raisons pratiques, ces modèles ont été entraînés avec le corpus CCNET (135 Go). Les résultats confirment l'impact positif de modèles plus grands sur les tâches NLI et NER. L'architecture *LARGE* conduit à une réduction d'erreur respectivement de 19,7% et 23,7% sur ces tâches. Étonnamment, sur les tâches d'étiquetage morphosyntaxique et d'analyse en dépendances, le fait d'avoir trois fois plus de paramètres ne conduit pas à des résultats significativement meilleurs qu'avec le modèle *BASE*.

Tenney *et al.* (2019) et Jawahar *et al.* (2019) ont montré que les informations morphosyntaxiques et syntaxiques sont apprises dans les couches inférieures de BERT tandis que les représentations sémantiques plus profondes se retrouvent dans les couches supérieures. Les couches inférieures de l'architecture *BASE* suffisent probablement à capturer ce qui est nécessaire aux tâches d'étiquetage morphosyntaxique et d'analyse syntaxique.

3.5 Impact du nombre de *steps*

La Figure ci-contre indique la perplexité du modèle de langue CAMEMBERT original ainsi que de ses performances sur nos tâches d'évaluation en fonction du nombre d'*epochs*, et ce à chaque *epoch* (8360 *steps*). Les résultats ci-contre suggèrent que plus la tâche est complexe, plus le nombre de *steps* a d'impact. Ainsi, alors qu'on peut observer un plateau pour les tâches bas-niveaux autour de 22000 *steps*, il semble que les performances continuent marginalement d'augmenter pour le NLI.



La comparaison entre deux modèles CCNET entraînés sur 100k et 500k *steps* respectivement (cf. Table 4) montre une légère augmentation des scores en NLI (+0,84) alors que ceux-ci stagnent en étiquetage et en analyse syntaxique. Ces résultats suggèrent que les représentations syntaxiques de bas niveau sont capturées bien plus tôt au cours de l'apprentissage que ne sont extraites les informations sémantiques complexes nécessaires au NLI.

4 Conclusion

Nous avons étudié l'impact de la taille et du niveau d'hétérogénéité des données de pré-entraînement sur la performance des modèles de langue neuronaux contextuels CAMEMBERT du français, ainsi qu'entre autres, l'impact de la taille du modèle et du nombre de *steps* de pré-entraînement. Nos résultats montrent que la taille des données d'entraînement n'a finalement que peu d'impact sur les performances globales et ouvrent donc la voie à des modèles de langages neuronaux contextuels spécialisés, liés à des domaines précis ou à des langues très peu dotées. La question de leur éventuelle complémentarité avec des modèles *fine-tuné* sur des modèles de langage générique est reste évidemment à explorer.

Entrainés sur des corpus *open-source* et disponibles sous une licence MIT, tous les modèles discutés dans cet article sont accessibles librement sur `https://camembert-model.fr`.

Remerciements

Nous tenons à remercier Clémentine Fourrier pour ses relectures et ses commentaires précieux, ainsi qu'Alix Chagué pour son fantastique logo. Ce travail a été en partie financé par trois projets de l'Agence Nationale de la Recherche accordés à Inria, les projets PARSITI (ANR-16-CE33-0021), SoSweet (ANR-15-CE38-0011) et BASNUM (ANR-18-CE38-0003), ainsi que par la chaire du dernier auteur dans l'Institut Prairie financée par l'ANR via le programme "Investissements d avenir" (ANR-19-P3IA-0001).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks*, p. 165–187. Kluwer : Dordrecht.

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, p. 1638–1649 : Association for Computational Linguistics.

CANDITO M. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, Grenoble, France, June 4-8, 2012*, p. 321–334.

CHAN B., MÖLLER T., PIETSCH M., SONI T. & YEUNG C. M. (2019). German bert. https://deepset.ai/german-bert.

CHARNIAK E. (2019). *Introduction to deep learning*. The MIT Press.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint : 1911.02116.

CONNEAU A., RINOTT R., LAMPLE G., WILLIAMS A., BOWMAN S. R., SCHWENK H. & STOYANOV V. (2018). XNLI : evaluating cross-lingual sentence representations. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, p. 2475–2485 : Association for Computational Linguistics.

DAI A. M. & LE Q. V. (2015). Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, p. 3079–3087.

DELOBELLE P., WINTERS T. & BERENDT B. (2020). RobBERT : a Dutch RoBERTa-based Language Model. arXiv preprint : 2001.06286.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). Multilingual bert. https://github.com/google-research/bert/blob/master/multilingual.md.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 4171–4186 : Association for Computational Linguistics.

DOZAT T. & MANNING C. D. (2017). Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* : OpenReview.net.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. : European Language Resources Association (ELRA).

GRAVE E., MIKOLOV T., JOULIN A. & BOJANOWSKI P. (2017). Bag of tricks for efficient text classification. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2 : Short Papers*, p. 427–431 : Association for Computational Linguistics.

HOWARD J. & RUDER S. (2018). Universal language model fine-tuning for text classification. In I. GUREVYCH & Y. MIYAO, Éds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1 : Long Papers*, p. 328–339 : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1031.

JAWAHAR G., SAGOT B., SEDDAH D., UNICOMB S., IÑIGUEZ G., KARSAI M., LÉO Y., KARSAI M., SARRAUTE C., FLEURY É. *et al.* (2019). What does bert learn about the structure of language ? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

JOULIN A., GRAVE E., BOJANOWSKI P., DOUZE M., JÉGOU H. & MIKOLOV T. (2016). Fasttext.zip : Compressing text classification models. arXiv preprint : 1612.03651.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. arXiv preprint : 1412.6980.

LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2019). ALBERT : A lite BERT for self-supervised learning of language representations. arXiv preprint : 1909.11942.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. arXiv preprint : 1912.05372.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. arXiv preprint : 1907.11692.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2019). CamemBERT : a Tasty French Language Model. arXiv preprint : 1911.03894.

MIKOLOV T., GRAVE E., BOJANOWSKI P., PUHRSCH C. & JOULIN A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, p. 3111–3119.

NIVRE J., ABRAMS M., AGIĆ Ž., AHRENBERG L., ANTONSEN L., ARANZABE M. J., ARUTIE G., ASAHARA M., ATEYAH L., ATTIA M., ATUTXA A., AUGUSTINUS L., BADMAEVA E., BALLESTEROS M., BANERJEE E., BANK S., BARBU MITITELU V., BAUER J., BELLATO S., BENGOETXEA K., BHAT R. A., BIAGETTI E., BICK E., BLOKLAND R., BOBICEV V., BÖRSTELL C., BOSCO C., BOUMA G., BOWMAN S., BOYD A., BURCHARDT A., CANDITO M., CARON B., CARON G., CEBIROĞLU ERYIĞIT G., CELANO G. G. A., CETIN S., CHALUB F., CHOI J., CHO Y., CHUN J., CINKOVÁ S., COLLOMB A., ÇÖLTEKIN Ç., CONNOR M., COURTIN M., DAVIDSON

E., DE MARNEFFE M.-C., DE PAIVA V., DIAZ DE ILARRAZA A., DICKERSON C., DIRIX P., DOBROVOLJC K., DOZAT T., DROGANOVA K., DWIVEDI P., ELI M., ELKAHKY A., EPHREM B., ERJAVEC T., ETIENNE A., FARKAS R., FERNANDEZ ALCALDE H., FOSTER J., FREITAS C., GAJDOŠOVÁ K., GALBRAITH D., GARCIA M., GÄRDENFORS M., GERDES K., GINTER F., GOENAGA I., GOJENOLA K., GÖKIRMAK M., GOLDBERG Y., GÓMEZ GUINOVART X., GONZÁLES SAAVEDRA B., GRIONI M., GRŪZĪTIS N., GUILLAUME B., GUILLOT-BARBANCE C., HABASH N., HAJIČ J., HAJIČ JR. J., HÀ MỸ L., HAN N.-R., HARRIS K., HAUG D., HLADKÁ B., HLAVÁČOVÁ J., HOCIUNG F., HOHLE P., HWANG J., ION R., IRIMIA E., JELÍNEK T., JOHANNSEN A., JØRGENSEN F., KAŞIKARA H., KAHANE S., KANAYAMA H., KANERVA J., KAYADELEN T., KETTNEROVÁ V., KIRCHNER J., KOTSYBA N., KREK S., KWAK S., LAIPPALA V., LAMBERTINO L., LANDO T., LARASATI S. D., LAVRENTIEV A., LEE J., LÊ HỒNG P., LENCI A., LERTPRADIT S., LEUNG H., LI C. Y., LI J., LI K., LIM K., LJUBEŠIĆ N., LOGINOVA O., LYASHEVSKAYA O., LYNN T., MACKETANZ V., MAKAZHANOV A., MANDL M., MANNING C., MANURUNG R., MĂRĂNDUC C., MAREČEK D., MARHEINECKE K., MARTÍNEZ ALONSO H., MARTINS A., MAŠEK J., MATSUMOTO Y., MCDONALD R., MENDONÇA G., MIEKKA N., MISSILÄ A., MITITELU C., MIYAO Y., MONTEMAGNI S., MORE A., MORENO ROMERO L., MORI S., MORTENSEN B., MOSKALEVSKYI B., MUISCHNEK K., MURAWAKI Y., MÜÜRISEP K., NAINWANI P., NAVARRO HORÑIACEK J. I., NEDOLUZHKO A., NEŠPORE-BĒRZKALNE G., NGUYỄN THỊ L., NGUYỄN THỊ MINH H., NIKOLAEV V., NITISAROJ R., NURMI H., OJALA S., OLÚÒKUN A., OMURA M., OSENOVA P., ÖSTLING R., ØVRELID L., PARTANEN N., PASCUAL E., PASSAROTTI M., PATEJUK A., PENG S., PEREZ C.-A., PERRIER G., PETROV S., PIITULAINEN J., PITLER E., PLANK B., POIBEAU T., POPEL M., PRETKALNIŅA L., PRÉVOST S., PROKOPIDIS P., PRZEPIÓRKOWSKI A., PUOLAKAINEN T., PYYSALO S., RÄÄBIS A., RADEMAKER A., RAMASAMY L., RAMA T., RAMISCH C., RAVISHANKAR V., REAL L., REDDY S., REHM G., RIESSLER M., RINALDI L., RITUMA L., ROCHA L., ROMANENKO M., ROSA R., ROVATI D., ROȘCA V., RUDINA O., SADDE S., SALEH S., SAMARDŽIĆ T., SAMSON S., SANGUINETTI M., SAULĪTE B., SAWANAKUNANON Y., SCHNEIDER N., SCHUSTER S., SEDDAH D., SEEKER W., SERAJI M., SHEN M., SHIMADA A., SHOHIBUSSIRRI M., SICHINAVA D., SILVEIRA N., SIMI M., SIMIONESCU R., SIMKÓ K., ŠIMKOVÁ M., SIMOV K., SMITH A., SOARES-BASTOS I., STELLA A., STRAKA M., STRNADOVÁ J., SUHR A., SULUBACAK U., SZÁNTÓ Z., TAJI D., TAKAHASHI Y., TANAKA T., TELLIER I., TROSTERUD T., TRUKHINA A., TSARFATY R., TYERS F., UEMATSU S., UREŠOVÁ Z., URIA L., USZKOREIT H., VAJJALA S., VAN NIEKERK D., VAN NOORD G., VARGA V., VINCZE V., WALLIN L., WASHINGTON J. N., WILLIAMS S., WIRÉN M., WOLDEMARIAM T., WONG T.-S., YAN C., YAVRUMYAN M. M., YU Z., ŽABOKRTSKÝ Z., ZELDES A., ZEMAN D., ZHANG M. & ZHU H. (2018). Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

ORTIZ SUÁREZ P. J., SAGOT B. & ROMARY L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In P. BAŃSKI, A. BARBARESI, H. BIBER, E. BREITENEDER, S. CLEMATIDE, M. KUPIETZ, H. LÜNGEN & C. ILIADI, Éds., *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom : Leibniz-Institut für Deutsche Sprache. HAL : hal-02148693.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In A. MOSCHITTI, B. PANG & W. DAELEMANS, Éds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 1532–1543 : ACL.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. A. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, p. 2227–2237 : Association for Computational Linguistics.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How multilingual is multilingual bert ? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*. DOI : 10.18653/v1/P19-1493.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners. preprint, https://paperswithcode.com/paper/language-models-are-unsupervised-multitask.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint : 1910.10683.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de Paris 7 en entités nommées (referential named entity annotation of the paris 7 french treebank) [in french]. In G. ANTONIADIS, H. BLANCHON & G. SÉRASSET, Éds., *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, Grenoble, France, June 4-8, 2012*, p. 535–542 : ATALA/AFCP.

STRAKA M., STRAKOVÁ J. & HAJIC J. (2019). Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing. arXiv preprint : 1908.07448.

STRAKOVÁ J., STRAKA M. & HAJIC J. (2019). Neural architectures for nested NER through linearization. In A. KORHONEN, D. R. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1 : Long Papers*, p. 5326–5331 : Association for Computational Linguistics.

TENNEY I., DAS D. & PAVLICK E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4593–4601, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1452.

VIRTANEN A., KANERVA J., ILO R., LUOMA J., LUOTOLAHTI J., SALAKOSKI T., GINTER F. & PYYSALO S. (2019). Multilingual is not enough : Bert for finnish. arXiv preprint : 1912.07076.

WENZEK G., LACHAUX M.-A., CONNEAU A., CHAUDHARY V., GUZMÁN F., JOULIN A. & GRAVE E. (2019). CCNet : Extracting High Quality Monolingual Datasets from Web Crawl Data. arXiv preprint : 1911.00359.

WILLIAMS A., NANGIA N. & BOWMAN S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, p. 1112–1122.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M. & BREW J. (2019). Huggingface's transformers : State-of-the-art natural language processing. arXiv preprint : 1910.03771.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Génération automatique de définitions pour le français

Timothee Mickus[1] Mathieu Constant[1] Denis Paperno[2]
(1) ATILF, 44 Avenue de la Libération, 54000 Nancy, France
(2) Universiteit Utrecht, Domplein 29, 3512 JE Utrecht, Netherlands
tmickus@atilf.fr, mconstant@atilf.fr, d.paperno@uu.nl

RÉSUMÉ

La génération de définitions est une tâche récente qui vise à produire des définitions lexicographiques à partir de plongements lexicaux. Nous remarquons deux lacunes : (i) l'état de l'art actuel ne s'est penché que sur l'anglais et le chinois, et (ii) l'utilisation escomptée en tant que méthode d'évaluation des plongements lexicaux doit encore être vérifiée. Pour y remédier, nous proposons un jeu de données pour la génération de définitions en français, ainsi qu'une évaluation des performances d'un modèle de génération de définitions simple selon les plongements lexicaux fournis en entrée.

ABSTRACT

Definition Modeling in French

Definition modeling is a recent task that aims at producing dictionary definitions based on word embeddings. We observe two gaps : (i) the current state of the art has yet to tackle languages other than English or Chinese and (ii) the purported usability as an evaluation method for word embeddings has yet to be verified. Hence we propose a dataset for French definition modeling and evaluate how using different input embeddings impacts the performances of a simple definition modeling system.

MOTS-CLÉS : Génération de définitions – plongements lexicaux – sémantique distributionnelle.

KEYWORDS: Definition modeling – word embeddings – distributional semantics.

1 Introduction

La *génération de définitions* (Noraset *et al.*, 2017, ou '*definition modeling*') vise à convertir un jeu de plongements lexicaux en un jeu de définitions équivalentes, telles qu'elles pourraient apparaître dans un dictionnaire. Fournir à un modèle de génération de définitions le vecteur du mot "*répétitivité*" devrait produire une sortie telle que "*Qualité, caractère de ce qui est répétitif (événements, gestes, etc.).*". [1] Suivant les conventions lexicographiques, on appellera ici le mot à définir le *definiendum* (pluriel : *definienda*).

Un ensemble conséquent de travaux a établi que les plongements lexicaux correspondent à des représentations de sémantique distributionnelle (Lenci, 2018). Ces dernières sont par conséquent explicitement mises en équivalence, dans les modèles de génération de définitions, avec les descriptions traditionnelles du sens. Le linguiste peut donc voir dans la génération de définitions une expérience montrant que la représentation distributionnelle d'un mot est conforme aux intuitions

1. Définition du wiktionnaire (fr.wiktionary.org)

Volume 2 : Traitement Automatique des Langues Naturelles, pages 66–80.
HAL : hal-02784756.

des locuteurs. [2] Pour l'informaticien, cette tâche permet de vérifier qu'un plongement lexical a été correctement pré-entraîné : si un vecteur de mot est une représentation correcte d'un mot, on s'attend à ce qu'il contienne toute l'information nécessaire afin de reconstruire une description du sens plus traditionnelle, en particulier la définition du mot telle qu'elle apparaît dans un dictionnaire.

Cet usage escompté de la génération de définitions comme outil d'évaluation faisait partie intégrante de la proposition initiale de Noraset *et al.* (2017) ; or aucune vérification n'en a été réalisée. La littérature subséquente sur le sujet a considéré comme acquise cette utilisation : aucun des travaux portés à notre connaissance ne compare deux jeux de plongements lexicaux, ni n'étudie l'importance des plongements pour la production de définitions. Pour y remédier, nous proposons de comparer les performances d'une architecture selon les représentations vectorielles fournies en entrée.

L'emploi suggéré par Noraset *et al.* (2017) n'est pas le seul attrait que présente cette tâche. On peut évoquer plusieurs applications pratiques à la génération de définitions : par exemple produire des ébauches de dictionnaires pour des langues peu dotées ou bien servir d'aide à la lecture pour les apprenants d'une langue étrangère. Ces usages, cependant, requièrent d'étudier les capacités des modèles sur plusieurs langues. Toute la littérature concernant la génération de définitions s'est pourtant focalisée sur l'anglais, à l'exception notoire de Yang *et al.* (2019) qui étudient le chinois. Pour répondre en partie à cette seconde lacune, nous produisons un jeu de données pour le français.

Le présent article est structuré comme il suit : nous mentionnerons d'abord en section 2 l'état de l'art dans le domaine de la géneration de définitions, avant de détailler le jeu de données que nous proposons pour le français (section 3), les modèles étudiés (section 4), les résultats préliminaires dont nous disposons (section 5) et enfin les erreurs typiquement commises par nos modèles (section 6). Nous terminerons par quelques perspectives en section 7.

2 État de l'art

L'utilisation de dictionnaires est une idée féconde et ancienne en traitement automatique des langues. Dans le cadre de cet exposé, on distinguera trois types de travaux. Une première catégorie contiendrait les travaux qui cherchent à modéliser ces ressources lexicales, que ce soit sous forme d'ontologie (Chodorow *et al.*, 1985), de graphes (Gaume *et al.*, 2014) ou autres. Une seconde s'intéresse davantage à comment exploiter les définitions qu'ils contiennent : par exemple, Gaume *et al.* (2004), qui indiquent qu'un dictionnaire permet de désambiguïser un mot ou encore Hill *et al.* (2016b), qui se penchent notamment sur la tâche de dictionnaire inversé, correspondant à trouver un mot à partir d'une définition ; on pensera aussi à Hill *et al.* (2016a) qui cherchent à produire un modèle de sémantique compositionnelle à partir de dictionnaires. Une troisième catégorie cherche à étoffer ces ressources, parmi lesquels on peut citer Sierra *et al.* (2015), qui proposent une méthodologie pour extraire automatiquement des définitions dans du texte 'brut'. De nombreuses entreprises de recherches s'inscrivent dans plusieurs de ces registres à la fois : par exemple, Bosc & Vincent (2018) proposent un système d'auto-encodage des définitions qui peut être vu à la fois comme une modélisation du dictionnaire, ainsi qu'une manière d'utiliser ceux-ci afin de produire des plongements lexicaux. Le programme de recherche qu'introduit la tâche de génération de définitions peut être considéré comme

2. Nous notons que l'établissement de dictionnaires (ou documents similaires) est attesté à travers les civilisations et les époques : à titre d'exemple, on peut citer le Er ya de la Chine antique, le glossaire de Cormac, texte irlandais du X^{ème} siècle, ou encore le projet moderne Wiktionary (wiktionary.org). Contrairement à d'autres ressources "expertes" telles que WordNet (Fellbaum, 1998), ces différentes entreprises ne sont pas guidées par les intuitions des linguistes spécialistes du domaine.

relevant de ces trois catégories de travaux : il propose de développer un modèle neuronal convertissant des *definienda* en définitions, ce qui servirait à la fois à évaluer des représentations vectorielles de mots et à compléter des dictionnaires existants.

La génération de définitions a été introduite par Noraset *et al.* (2017), qui la conçoivent avant tout comme une tâche d'évaluation extrinsèque de jeux de plongements lexicaux ; en cela, elle est à rapprocher de la vaste littérature dédiée à l'analyse des jeux de plongements lexicaux (Levy & Goldberg, 2014a,b; Arora *et al.*, 2018; Batchkarov *et al.*, 2016; Swinger *et al.*, 2018, entre autres), qui plus traditionnellement se concentre sur la structure de l'espace vectoriel et de l'application des vecteurs à la tâche d'analogie formelle (Mikolov *et al.*, 2013b; Gladkova *et al.*, 2016; Grave *et al.*, 2018).

La formulation initiale de la tâche par Noraset *et al.* (2017) considère le *definiendum* isolé de son contexte : Gadetsky *et al.* (2018) remarquent que ceci est problématique pour les mots ambigus ou polysémiques, pour lesquels établir le sens requiert d'avoir accès au contexte ; ils utilisent par conséquent des plongements contextualisés AdaGram (Bartunov *et al.*, 2016). Mickus *et al.* (2019) soulignent qu'une architecture de type "encodeur-décodeur" permet de traiter de manière uniforme les *embeddings* contextualisés et non-contextualisés, ainsi que les unités polylexicales et les mots simples. Zhu *et al.* (2019) et Chang *et al.* (2018) proposent de décomposer les sens d'un mot en représentations vectorielles creuses (Arora *et al.*, 2018). Chang *et al.* (2018) critiquent de plus que ces modèles sont relativement difficiles à interpréter, ce qui nuit à l'utilité de la tâche en tant qu'outil d'évaluation (d'où leur emploi de représentations creuses). Zhang *et al.* (2019) explorent différentes architectures qui permettent de générer non seulement la définition, mais aussi un exemple d'utilisation du *definiendum*.

Tous les travaux précédents étudient spécifiquement le cas de l'anglais. Yang *et al.* (2019) proposent un jeu de données pour le chinois, qui inclut des caractères sémantiquement liés au *definiendum* manuellement annotés. Tout comme le jeu proposé par Noraset *et al.* (2017), le jeu de Yang *et al.* (2019) ne contient pas d'exemples d'usage des *definienda*.

3 Jeu de données

Le jeu de données que nous proposons est tiré de GLAWI (Hathout & Sajous, 2016), une version du wiktionnaire français au format XML. GLAWI recense pour chaque mot-forme les parties du discours applicables, et pour chacune de celles-ci les sens possibles, accompagnés d'une définition (balise `<gloss>`) ainsi qu'éventuellement d'un exemple d'usage (balise `<example>`). Nous retirons les exemples d'usage où le *definiendum* n'est pas présent dans le contexte. [3] Nous convertissons le jeu de données entier en caractères minuscules. L'emploi du wiktionnaire comme source de données implique que notre jeu contient aussi des expressions polylexicales. [4] L'utilisation de *definienda* polylexicaux n'est pas standard en génération de définitions : comme le but premier de cette tâche est l'évaluation extrinsèque de plongements lexicaux, la capacité à prendre en compte une entrée séquentielle est souvent sacrifiée en faveur d'une architecture liant directement une représentation vectorielle à une production. [5]

3. Comme Mickus *et al.* (2019), nous conservons les exemples d'usages où une variante fléchie du *definiendum* est présente ; pour ce faire nous utilisons la librairie python `spacy` (https://spacy.io/).

4. Seuls les *definienda* polylexicaux réalisés de manière continue ont étés conservés.

5. Laissant un instant de côté la formalisation initiale de la tâche en tant qu'outil d'évaluation, il est notoire que l'avènement de plongements contextualisés est allé de pair avec l'accroissement de la demande en ressources computationnelles pour l'entraînement. Un nombre important de ces modèles, pour y répondre, font usage de tokenisation en sous-mots afin de limiter

# d'exemples	*Definienda* distincts	Taille moy. des exemples	Taille moy. des définitions
232037	100288	25,36	12,01

TABLE 1: Distribution des définitions utilisables de GLAWI.

La table 1 présente quelques statistiques concernant notre jeu de données. L'exemple d'usage est généralement deux fois plus long que la définition à produire. De plus, 12% des exemples correspondent à un *definiendum* polylexical ; la longueur moyenne des exemples d'utilisations et des définitions est relativement similaire dans les deux cas des *definienda* polylexicaux et monolexicaux. À titre de comparaison, les jeux de données proposés par Yang *et al.* (2019) contiennent un peu plus de 100000 exemples pour le chinois. Quant à l'anglais, le jeu de données de Gadetsky *et al.* (2018) contient un peu plus de 122000 exemples (dont 20% de cas où le *definiendum* ne peut pas être extrait de l'exemple d'utilisation associé), Zhu *et al.* (2019) proposent un jeu de données d'un peu plus de 120000 exemples et enfin Zhang *et al.* (2019) proposent près de 300000 exemples.

Nous séparons le jeu de données en trois sous-ensembles, un d'entraînement (80 %), un de validation (10 %) et un de test (10 %) ; nous utilisons la même partition dans toutes les expériences mentionnées plus bas. La partition est réalisée de manière à ce que les *definienda* soient uniques au sous-ensemble où ils apparaissent. Les sous-ensembles totalisent respectivement 185363, 23178 et 23496 exemples.

4 Modèle et jeux de plongements lexicaux

4.1 Architecture de génération de définitions

Nous reprenons le modèle de Mickus *et al.* (2019), qui permet de traiter uniformément les unités polylexicales et monolexicales. Ce modèle est entraîné à générer des définitions à partir des exemples d'usages associés aux définitions ; nous y renvoyons le lecteur pour tout détail complémentaire. À un niveau formel, il correspond à une architecture Transformer (Vaswani *et al.*, 2017), dotée de vecteurs "marqueurs", ici $\vec{D}$ et $\vec{C}$ distinguant respectivement le *definiendum* de son contexte. Dans le cadre du présent travail, nous ajoutons systématiquement le *definiendum* au début de l'exemple d'usage, précédé d'un token spécial `[DDUM]`, et suivi d'un autre token spécial `[DENS]`.[6]

Une illustration de l'architecture encodeur-décodeur utilisée par ce modèle est présentée dans la Figure 1 : la séquence composée du token spécial `[DDUM]`, suivi du *definiendum* suivi du token spécial `[DENS]`, suivi de l'exemple d'usage est passée à l'encodeur (en rouge dans la figure), qui la convertit en une séquence de représentations intermédiaires $\vec{r_1}, \ldots, \vec{r_n}$, désignées collectivement comme "banque d'attention" (en vert dans la figure). Le décodeur (en jaune dans la figure) est prompté

la taille du vocabulaire. Pour prendre un exemple récent, CamemBERT (Martin *et al.*, 2019) et FlauBERT (Le *et al.*, 2019), qui pré-entraînent tous deux l'architecture BERT (Devlin *et al.*, 2019) pour le français, font respectivement usage des algorithmes SentencePiece (Kudo & Richardson, 2018) et BPE (Sennrich *et al.*, 2016). Or l'éclatement d'un mot en sous-mots conduit à remplacer une entrée composée d'une seule représentation vectorielle par une série ordonnée de vecteurs : c'est-à-dire conduit à considérer certaines unités monolexicales comme polylexicales.

6. L'architecture Transformer faisant un usage systématique de connections résiduelles entre chacune des couches de l'encodeur, interrompues seulement par des opérations de normalisation, on peut garantir que la représentation intermédiaire $\vec{r_k}$ correspondant au k-ième token de la séquence gardera trace de la composition linéaire du vecteur fourni en entrée. Par conséquent les représentations intermédiaires correspondant aux éléments de contexte résideront dans un sous-espace vectoriel distinct de celui correspondant aux représentations de *definienda*.

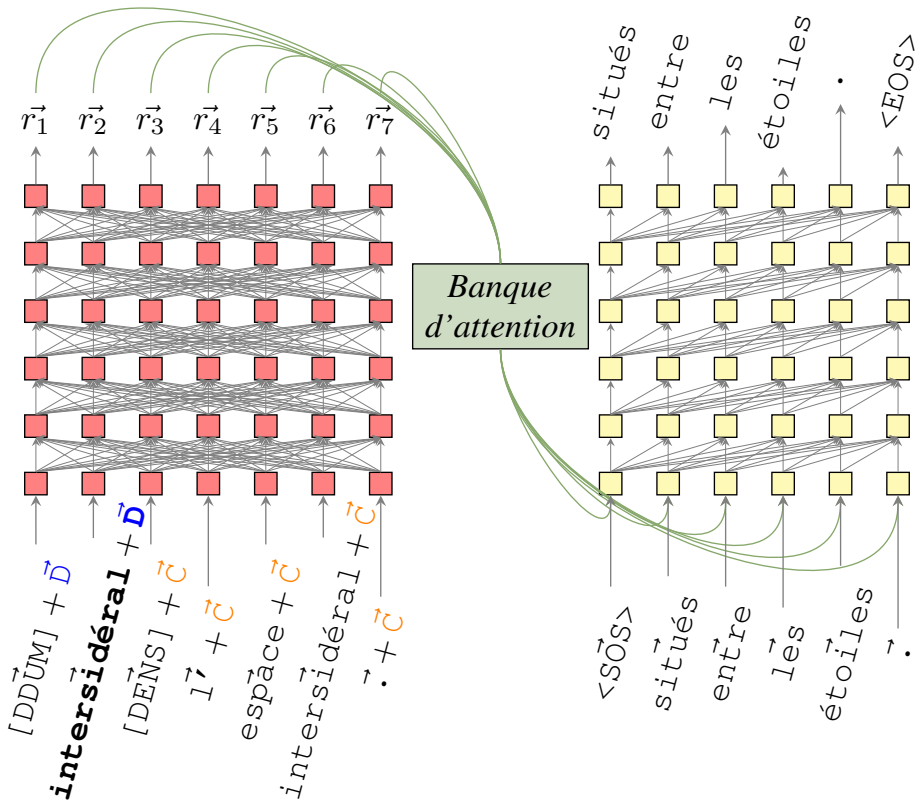


FIGURE 1: Vue d'ensemble de l'architecture du modèle de génération de définitions.

avec un symbole spécial indiquant le début de séquence <SOS> ; la génération se termine lorsqu'il produit un symbole de fin de séquence <EOS>. Lors de l'apprentissage, la définition tirée du jeu de données est fournie en entrée du décodeur ; pendant la génération, l'entrée à l'étape t correspond au symbole généré à l'étape précédente $t-1$. Le décodeur a toujours accès à la banque d'attention, qu'il utilise à travers des mécanismes d'attention multi-tête. Comme les modèles de génération de définitions sont censés servir à l'évaluation des plongements lexicaux, nous gelons les poids des représentations vectorielles.

Nous utilisons 12 couches dans l'encodeur et le décodeur, un *warmup* de 10000 étapes, un taux d'apprentissage de 1 et un *label smoothing* de 0,15 ; autrement les hyperparamètres correspondent à ceux initialement suggérés dans Mickus *et al.* (2019). Ces paramètres correspondent à la configuration optimale trouvée lors d'expériences préliminaires.

4.2 Plongements lexicaux

Nous comparons plusieurs architectures de plongements lexicaux : word2vec (Mikolov *et al.*, 2013a, CBOW), GloVe (Pennington *et al.*, 2014) et FastText (Bojanowski *et al.*, 2017) ; toutes nos représentations sont de dimension 300. Pour ce qui est de word2vec et GloVe, nous utilisons la concaténation de FRCOW (Schäfer, 2015) et d'un dump de Wikipedia français parsé par Coavoux (2017) [7] comme corpus d'entraînement. Ces deux corpus sont mis en minuscules, pour un total de 7,25 milliards de tokens. Pour word2vec, nous utilisons 20 contre-exemples et une fenêtre de 10 mots et parcourons

7. Disponible à l'adresse suivante : http://www.llf.cnrs.fr/wikiparse/

aléatoires	word2vec	GloVe	FastText MC	FastText FB
0,00	36,46	58,32	57,38	68,63

TABLE 2: Analogie formelle : précision des différentes représentations vectorielles (pourcentages).

le corpus sur 10 itérations ; les vecteurs GloVe sont entraînés pendant 10 itérations avec les hyperparamètres proposés dans le script d'exemple fourni par Pennington *et al.* (2014). Pour FastText, nous utilisons deux jeux de plongements : un jeu entraîné sur ce même corpus pendant 5 itérations ("FastText MC"), et celui fourni par Grave *et al.* (2018) ("FastText FB"). Ce jeu de plongements FastText FB est décrit en davantage de détails par Grave *et al.* (2018) ; nous soulignerons seulement ici que le corpus d'apprentissage utilisé contient un dump de Wikipedia (1,11 milliard de tokens) ainsi qu'un sous ensemble de Common Crawl (68,36 milliards de tokens), c'est-à-dire un ensemble de données d'apprentissage environ 10 fois plus grand. Comparer ces deux jeux devrait nous permettre d'appréhender la sensibilité de la génération de définitions aux choix d'hyperparamètres et à l'accès aux données. Enfin, nous évaluons aussi une matrice initialisée aléatoirement $M_{ij} \sim \mathcal{N}(0, 1)$ avec la même dimension que nos plongements lexicaux.

La précision atteinte par ces différentes représentations sur le jeu d'analogie formelle de Grave *et al.* (2018) est indiquée dans la table 2. [8] Évaluer les performances des représentations vectorielles à l'aide d'une méthode répandue comme l'analogie formelle nous permet d'étudier si la génération de définitions peut servir en tant que méthode d'évaluation des plongements lexicaux. Ceci nous permet d'estimer *a priori* un degré raisonnable de variation des performances des différents jeux de plongements : si les résultats obtenus sur l'analogie formelle et ceux obtenus en génération de définitions divergent radicalement, il nous faudra soit remettre en cause la validité de la tâche générative en tant que méthode d'évaluation, soit avancer que les aspects mesurés par ces deux méthodes diffèrent radicalement. Un tel scénario est effectivement envisageable : rien ne garantit que la régularité linéaire de l'espace sémantique des plongements soit utile à la génération de définitions. Soulignons toutefois que les deux tâches devraient être toutes deux sensibles à la qualité des représentations vectorielles testées ; aussi est-il raisonnable d'escompter une certaine congruence de leurs résultats respectifs.

5 Résultats

Nous évaluons nos résultats à l'aide de deux métriques : perplexité et score BLEU (Papineni *et al.*, 2002). Ces métriques sont couramment utilisées en génération automatique. La perplexité est censée dépeindre l'incertitude du modèle de pouvoir engendrer la cible. Elle se calcule comme l'exponentiation de l'entropie croisée : par conséquent on préférera les modèles qui minimisent la perplexité. Les scores BLEU calculent la similarité du vocabulaire employé dans la cible et dans la production du modèle : les modèles où le score BLEU est maximal sont *a priori* à privilégier.

Les performances des modèles que nous étudions sont rapportées dans la table 3. Pour chaque jeu de représentations vectorielles, nous rapportons les scores de perplexité et les scores BLEU sur l'ensemble de test. Deux remarques ressortent immédiatement de l'étude des scores de perplexité. Premièrement,

8. Nous supprimons les doublons du jeu de Grave *et al.* (2018), ainsi que les exemples contenant la paire "*son*" — "*sa*", qui correspond à un contraste de genre grammatical et non de genre social. 646 exemples sont concernés. Nous mettons le jeu entier en minuscules.

Plongements	Perplexité	BLEU
aléatoires	83,70	19,90
word2vec	52,13	30,60
GloVe	48,55	29,00
FastText MC	45,04	30,50
FastText FB	47,84	32,80

TABLE 3: Génération de définitions : résultats généraux (ensemble de test).

le modèle entraîné sur des représentations aléatoires produit des résultats bien moins bons que les modèles entraînés sur des plongements lexicaux. Ceci suggère que la génération de définitions répond aux critères minimum pour être utilisée afin d'évaluer les plongements lexicaux. Deuxièmement, les différences de performance entre jeux de plongements lexicaux sur la tâche d'analogie formelle (cf. table 2) ne sont pas celles observées pour la génération de définitions : si l'ordre des trois architectures varie peu, la différence entre les vecteurs GloVe et FastText est moindre que celle observée pour l'analogie formelle ; [9] de plus, le modèle FastText que nous avons entraîné s'avère comparable à celui distribué par Grave *et al.* (2018), alors qu'on observait une différence claire en analogie formelle. Les scores BLEU donnent un élément de contraste à ce que nous indique la perplexité. Ici, les représentations aléatoires sont toujours nettement inférieures ; cependant les vecteurs GloVe sont cette fois-ci perçus comme moins bons que les vecteurs word2vec. Nous remarquons de plus un compromis entre le score BLEU et la perplexité des deux modèles FastText, ce qui n'aide pas à départager les deux métriques.

Notons toutefois deux points importants. D'une part, les hyperparamètres de nos modèles n'ont pas étés spécifiquement réglés : une recherche plus extensive pourrait avoir un impact significatif sur ces résultats. D'autre part, une autre interprétation de ces résultats serait que les modèles disposent de suffisamment de ressources pour gommer les différences entre chaque jeu de plongements, ce qui gênerait l'utilisation de la génération de définitions en tant que tâche d'évaluation.

Afin de compléter le tableau que dressent les deux métriques automatiques, nous pouvons comparer également les scores BLEU obtenus en récupérant simplement la définition correspondant à l'entrée la plus proche dans l'ensemble d'entraînement. Ceci nous permet d'établir un point de comparaison utile, en indiquant le score qui serait obtenu par un modèle qui copierait simplement des exemples déjà rencontrés. Nous étudions deux variantes pour cette ligne de référence. La première consiste à récupérer l'entrée globalement la plus similaire, ce que l'on peut mesurer approximativement en calculant la similarité entre exemples d'utilisation fournis lors de l'inférence et exemples d'utilisation vus lors d'apprentissage ; nous mesurons la similarité des exemples d'utilisation à l'aide du cosinus des vecteurs moyens des mots qu'ils contiennent. La seconde repose sur l'idée que des mots aux sens similaires auront des définitions semblables ; nous comparons donc aussi les définitions dont les *definienda* sont les plus similaires. Nous calculons ces deux lignes de référence à partir de l'ensemble de validation.

Les résultats correspondant à ces deux lignes sont décrits en table 4 ; nous incluons aussi les scores BLEU obtenus sur l'ensemble de validation par les modèles de génération correspondants à fins

9. Rappelons cependant que l'algorithme FastText encode linéairement les régularités orthographiques, ce qui améliore ses résultats sur la tâche d'analogie formelle : le jeu d'analogies de Grave *et al.* (2018) inclut notamment des alternances flexionnelles, qui utilisent souvent des affixes orthographiquement réguliers.

Plongements	Validation	Meilleur ex. d'usage	Meilleur *definiendum*
aléatoires	19,80	17,10	16,20
word2vec	31,60	17,50	17,80
GloVe	28,40	17,60	18,00
FastText MC	30,30	17,80	18,70
FastText FB	32,30	17,30	18,80

TABLE 4: Génération de définitions : lignes de référence (scores BLEU).

de comparaisons. Premièrement, nous voyons que la marge de différence entre les représentations aléatoires et les jeux de plongements lexicaux est bien moindre que ce qu'on observe pour les modèles entraînés. Ceci s'explique en partie du fait qu'aucune des définitions pour ces lignes de référence n'est apprise : toute représentation même aléatoire est associée à une production tirée de l'ensemble de validation, et par conséquent toutes sont stylistiquement parfaites en ce qu'elles réemploient les tournures exactes du jeu d'entraînement. Par conséquent on peut supposer que tout écart à la référence s'explique par une différence d'ordre sémantique, plutôt que stylistique. Admettons donc que l'écart des scores BLEU issus de ces lignes de référence et de nos modèles reflète l'importance de la similarité sémantique : alors, bien qu'on puisse supposer que cette similarité sémantique est encodée dans nos plongements (ce que l'on peut voir en comparant les gains importants des modèles appris sur des plongements aux gains pauvres du modèle appris sur des représentations aléatoires), elle ne l'est pas aussi directement que la dépendance linéaire mesurée en analogie formelle (ce qui s'observe par l'écart limité entre les vecteurs aléatoires et des plongements dans ces lignes de référence).

De plus, si l'on observe les scores BLEU obtenus en récupérant le *definiendum* le plus similaire dans l'ensemble d'entraînement, on voit que les jeux de plongements distributionnels où la similarité entre *definienda* induit une plus grande similarité entre définitions sont ceux qui produisent les meilleurs résultats sur la tâche de génération de définitions. Ceci tend à confirmer que la génération de définitions requiert que le vecteur passé en entrée encode la similarité sémantique ; par extension, ceci confirmerait de plus l'utilité de cette tâche pour l'évaluation des jeux de plongements lexicaux.

Notons cependant que d'autres explications peuvent répondre à ces faits. En particulier les plongements lexicaux non-contextualisés représentent généralement tous les sens associés à un mot-type par le même vecteur : [10] par conséquent les phénomènes de poysémie et d'homonymie perturbent les mesures de similarité sémantique, et à son tour ceci implique que le *definiendum* le plus 'similaire' peut correspondre à un sens autre que celui visé par la définition auquel on l'associe. Enfin, la métrique BLEU employée ici peut n'être pas adaptée à l'évaluation que nous conduisons.

6 Analyse d'erreurs

Jusqu'ici, nous avons analysé les résultats produits par nos modèles à l'aide des scores BLEU et de la perplexité. Ces métriques peinent cependant à capturer la composante sémantique des productions d'un modèle ; notre emploi ici se justifie essentiellement par l'utilisation de ces métriques dans la littérature existante. Nous remarquons que la difficulté inhérente à la génération de définitions ne réside non pas dans des questions de variation ou cohérence stylistique mais bien plutôt dans

10. En particulier, on s'attend à ce que le sens le plus fréquent domine dans la représentation vectorielle, cf. Arora *et al.* (2018); Bartunov *et al.* (2016).

l'ancrage sémantique et la véracité du contenu des définitions : une définition stylistiquement parfaite et textuellement très proche de la référence peut être entièrement erronnée et inutilisable.

À titre d'illustration, envisageons une définition du mot "*chimique*" qui serait "*Relatif à ou issu de la Martinique*" plutôt que "*Relatif à ou issu de la chimie*", comme ce mot est définit dans le wikitionnaire : les métriques textuelles automatiques telles que BLEU ne pénalisent pas particulièrement ce type d'exemples — du moins, cet exemple erronné sera préféré presque systématiquement à une définition utilisant une autre formulation, mais cependant valide, telle que par exemple celle du TLFi : [11] "*Qui par nature appartient à la chimie, qui relève de la chimie.*".

De fait, cette illustration est loin d'être invraisemblable. Ayant sélectionné aléatoirement un échantillon E_{GLAWI} de 10000 définitions de GLAWI, nous avons calculé pour chaque définition d_i la distance d'édition minimale au reste de l'échantillon, c'est-à-dire :

$$d_i = \min\{D_{edit}(d_i, d_j) \quad \textit{tel que} \quad d_j \in E_{\text{GLAWI}} \ \wedge \ d_j \neq d_i\}$$

où D_{edit} correspond à la distance d'édition (ou de Levenshtein), définie sur les mots pluôt que les caractères, et d_i et d_j sont des définitions de l'échantillon E_{GLAWI}. Nous avons pu observer que 77,19% des définitions de notre échantillon ne différaient que par un mot ajouté, supprimé ou remplacé d'une autre définition de l'échantillon. De manière générale, les dictionnaires utilisent avec une grande fréquence un inventaire limité de tournures spécifiques, rendant la question de l'évaluation par métriques automatiques d'autant plus prégnante.

En bref, si le problème de la paraphrase est commun à toute tâche de génération automatique, il se présente de manière encore plus épineuse dans le cadre de la génération de définitions de par l'effet conjoint de la relative pauvreté des tournures stylistiques employées et de la proéminence de l'ancrage sémantique dans ce qui constitue une bonne ou une mauvaise définition. C'est aussi cette importance primordiale de l'ancrage sémantique, paradoxalement, qui fait l'attrait de la tâche générative pour l'évaluation des plongements lexicaux.

D'autres métriques similaires à BLEU pourraient être envisagées, notamment METEOR (Banerjee & Lavie, 2005; Elloumi *et al.*, 2015, pour l'adaptation au français) ; [12] nous notons toutefois que ces métriques ne ciblent pas spécifiquement les éléments lexicaux clefs ("*Martinique*" ou "*chimie*" dans notre exemple), mais prennent également en compte les éléments purement stylistiques dans le calcul d'un score : aussi le problème que nous soulignons n'est pas résolu par de telles métriques.

Afin de pallier ce défaut et d'obtenir une compréhension plus détaillée des productions qu'ils génèrent, nous sélectionnons aléatoirement 100 exemples et comptabilisons (i) le nombre de productions valides pour une partie du discours autre que celle du *definiendum*, (ii) le nombre de productions où le *definiendum* est présent dans sa propre définition et (iii) le nombre de productions contenant un mot ou un syntagme répété. Ces informations supplémentaires nous permettent d'établir une vision plus fine du genre d'erreurs de génération commises par nos modèles ; comme souligné plus haut, une définition peut évidemment être entièrement erronnée tout en n'enfreignant aucun de ces critères.

Les résultats correspondants sont consignés dans la table 5. Si, à première vue, les représentations aléatoires semblent autant erronées que les plongements GloVe, nous notons que les erreurs des

11. Trésor de la Langue Française informatisé. Disponible à l'adresse suivante : http://atilf.atilf.fr/.

12. Dans le cas de Elloumi *et al.* (2015), la ressource lexicale exploitée afin d'établir des relations de synonymies recouvre une partie significative de notre jeu de données, car les deux sont dérivés du wiktionnaire. Si nous employions cette métrique, des exemples tirés du jeu de test pourraient par conséquent être pris en compte dans le calcul de scores sur les jeux d'entraînement et de validation.

Plongements	Mauvaise POS	Auto-référence	Répétitions
aléatoires	25	1	6
word2vec	19	7	4
GloVe	24	2	5
FastText MC	16	7	5
FastText FB	22	4	0

TABLE 5: Erreurs typiques de génération.

vecteurs aléatoires sont souvent de plus grande ampleur : par exemple le nombre de répétitions dans une seule définition y est plus important. Un autre défi majeur rencontré par tous nos modèles est de distinguer les différentes parties du discours, ce que l'on peut supposer être dû à l'usage d'un unique encodeur pour le *definiendum* et l'exemple d'usage. De manière intéressante, les modèles basés sur des représentations non-aléatoires produisent davantage d'auto-références, et ce en l'absence de tout mécanisme de copie. Ceci peut être directement imputé à ce que l'espace sémantique des vecteurs distributionnels est effectivement structuré d'une manière exploitable par le réseau de neurones artificiel : bien qu'il s'agisse à proprement parler d'une erreur, l'auto-référence nous indique que la représentation contextualisée du *definiendum* influence effectivement le processus de génération, en ce que le décodeur choisit d'émettre le symbole au sens le plus similaire à cette entrée.

Penchons-nous à présent davantage sur les facteurs sémantiques qui peuvent conduire à une définition erronnée. Ceux-ci sont difficiles à étudier de manière systématique, en partie du fait que les erreurs de génération peuvent impacter ces facteurs sémantiques (notamment dans le cas fréquent où la partie du discours de la définition générée ne correspond pas à celle du *definiendum*), en partie aussi du fait du faible ancrage sémantique de nos modèles, qui conduit souvent à des productions difficiles à juger sur certains critères. Nous étudions par conséquent trois critères pour lesquels on peut espérer une certaine rigueur : (i) si le champ sémantique des éléments lexicaux présents dans la production est en lien avec le sens ciblé par l'exemple d'usage ; (ii) si le champ sémantique des éléments lexicaux présents dans la production est en lien avec un quelconque sens du *definiendum* et (iii) la proportion de productions suivant une forme *genus–differentia* [13] où le *genus* est un hyperonyme du *definiendum*.

Plongements	Champ sémantique inapproprié		*Genus–differentia*		
	à la définition visée	à tout sens	# défs. concernées	# avec genus incorrect	*(Pourcentage)*
aléatoires	93	91	58	50	86,2 %
word2vec	63	56	62	35	56,5 %
GloVe	67	57	65	40	61,5 %
FastText MC	61	45	69	38	55,1 %
FastText FB	70	57	65	41	63,1 %

TABLE 6: Erreurs à caractère sémantique.

Les résultats correspondants sont consignés dans la table 6. Nous donnons les résultats pour chacun des trois aspects sémantiques discutés précédemment, ainsi que le nombre de définitions produites

13. Les définitions de forme *genus–differentium* sont composées d'un mot ou d'une expression donnant la classe générale sémantique, le *genus*, et d'une expression restreignant cette classe au sens visé par le *definiendum*. Noraset *et al.* (2017) remarquent que cette forme de définition est très fréquente : 85 % des définitions de WordNet (Fellbaum, 1998) et 50 % des définitions du GCIDE y correspondraient.

ayant une forme *genus–differentia*. Nous soulignons que les représentations aléatoires produisent fréquemment des définitions métalinguistiques ("*variante orthographique de ...*", "*synonyme de...*", etc.), ce que nous suggérons être dû à l'incapacité de ces modèles à lier cohéremment un *definiendum* à un possible hyperonyme. De manière plus générale, la vision d'ensemble qui se dégage de cette évaluation manuelle indique clairement que l'adéquation sémantique des définitions demeure un défi majeur. Un point intéressant est que les scores attribués aux représentations aléatoires sont encore une fois nettement supérieurs à ceux attribués aux autres jeux de plongements lexicaux. En cela les résultats suggèrent encore une fois que la génération de définitions peut dans une certaine mesure servir de crible pour les représentations distributionnelles, puisque les définitions issues d'entrées sémantiquement incohérentes paraissent clairement moins acceptables. Il est intéressant de remarquer qu'ils indiquent aussi que les plongements FastText MC seraient les plus performants.

1.	Entrée	la **structuration** de la pensée.
	Réf.	*Action de structurer, de donner une structure.*
	aléatoire	`qualité de ce qui est net.`
	word2vec	`action de composer.`
	GloVe	`action de grouper.`
	FastText MC	`action de grouper.`
	FastText FB	`action de structurer.`
2.	Entrée	**chercheur** d'or, **chercheur** d'aventures, **chercheur** de querelles.
	Réf.	*Personne qui cherche.*
	aléatoire	`étoffe de soie, de coton, etc.`
	word2vec	`spécialiste de l'étude des connaissances.`
	GloVe	`celui qui s'occupe de statistique.`
	FastText MC	`celui, celle qui connaît.`
	FastText FB	`celui, celle qui connaît.`
3.	Entrée	le grand **mélinet** est une petite plante de la famille des bourraches qui se rencontre dans les terrains sablonneux relativement humides .
	Réf.	*Nom vulgaire de borraginées (cérinthe, tournefort), dont quelques espèces sont indigènes dans le midi.*
	aléatoire	`espèce d'insecte lépidoptère (papillon) de la famille des géométridés (noctuidae), dont les ailes sont caduques alternes.`
	word2vec	`plante monocotylédone de la famille des saxifragacées.`
	GloVe	`famille de plantes dicotylédones monopétales.`
	FastText MC	`qui croît sur les écorces.`
	FastText FB	`synonyme de préfloraison.`

TABLE 7: Productions choisies du jeu de validation (*definiendum* en gras dans l'entrée ; échantillon non-représentatif).

Parmi les erreurs à caractère sémantiques, d'autres que celles mentionnées dans la table 6 sont plus difficiles à annoter de manière systématique. La table 7 contient quelques exemples choisis associant une entrée à la définition attendue et aux productions de chacun des modèles. Plusieurs modèles produisent des définitions identiques pour des entrées identiques, ce qui suggère que les résultats en génération de définitions dépendent en grande partie des données lexicographiques utilisées. Certains des faits que nous mentionnons plus haut sont aussi visibles ici : par exemple, les productions sémantiquement inadéquates des représentations aléatoires ou bien la tendance générale des différents modèles à ignorer l'exemple d'usage pour se focaliser sur un sens particulier du *definiendum* (exemple 2). Enfin, un problème épineux concerne le cas de définitions en domaine de spécialité : sans une bonne connaissance de la botanique, il est difficile de juger de l'adéquation de la production du modèle word2vec pour l'exemple 3.

7 Conclusions

Cet article vise à adapter la tâche de génération de définitions à la langue française. À ce titre, nous proposons un jeu de données, des jeux de plongements normalisés et plusieurs éléments de comparaison pour les recherches ultérieures sur ce sujet. [14] Nos expériences suggèrent que la tâche de génération de définitions permet d'évaluer des représentations vectorielles car elle distingue proprement des représentations aléatoires de plongements lexicaux préentraînés. Les aspects qu'elle mettrait en exergue diffèrent de ceux capturés par l'analogie formelle : elle permet d'évaluer si des composantes sémantiques sont encodées de manière non linéaire dans les vecteurs, et favorise les jeux de plongements lexicaux où ces composantes structurent le plus clairement l'espace vectoriel.

Nous notons toutefois que la capacité à distinguer des représentations aléatoires de représentations sémantiquement cohérentes n'est qu'un des éléments nécessaires pour établir que cette tâche puisse effectivement jouer le rôle d'outil d'évaluation que la littérature suggère. D'autres expériences sont évidemment requises. En particulier, il serait utile de mettre en correspondance les productions générées par des modèles avec les jugements de locuteurs natifs quant à leur adéquation sémantique. Un autre point qu'il sera crucial d'étudier dans des travaux futurs consiste à établir des métriques automatiques et critères d'évaluations fiables adaptés à la tâche de génération de définitions.

L'évaluation manuelle que nous avons conduite laisse penser que ces modèles sont encore loin d'être effectivement utilisables. En vue d'améliorer ces résultats, plusieurs points méthodologiques demanderaient un examen plus minutieux : notamment les effets des choix d'hyperparamètres, tant pour les jeux de plongements que pour les modèles de génération. Le point central de la présente étude concernant l'adéquation de la génération de définitions à l'évaluation des plongements lexicaux, nous avons étudié en priorité comment différentes architectures de plongements interagissaient avec la tâche en question. Il est par conséquent plus que vraisemblable que les hyperparamètres optimaux pour la génération de définitions varient pour chaque jeu de plongements ; aussi l'établissement d'un protocole de séléction des hyperparamètres est un point méthodologique important que nous comptons aborder dans nos travaux futurs.

Ces travaux suggèrent aussi plusieurs pistes de recherches futures, en particulier sur le traitement des définitions en domaine de spécialité ou bien sur l'adaptation de la tâche à d'autres langues. Étendre la tâche de la génération de définitions dans un contexte multilingue nous paraît particulièrement crucial : non seulement parce qu'enrichir les jeux de données relatives à cette tâche permettra sans doute d'avancer nos connaissances en sémantique distributionnelles et représentations neuronales, mais aussi, et surtout, parce qu'une des applications pratiques majeures de la génération de définitions consiste en la documentation de langues peu dotées.

Remerciements

Nous remercions trois relecteurs anonymes dont les commentaires ont permis une amélioration significative du présent manuscrit. Ce travail a bénéficié d'une aide de l'État, gérée par l'Agence Nationale de la Recherche, au titre du projet Investissements d'Avenir Lorraine Université d'Excellence, portant la référence ANR-15-IDEX-04-LUE.

14. Ces éléments seront disponibles à l'adresse suivante : https://github.com/TimotheeMickus/dm-french

Références

ARORA S., LI Y., LIANG Y., MA T. & RISTESKI A. (2018). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, **6**, 483–495. DOI : 10.1162/tacl_a_00034.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

BARTUNOV S., KONDRASHKIN D., OSOKIN A. & VETROV D. P. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, p. 130–138.

BATCHKAROV M., KOBER T., REFFIN J., WEEDS J. & WEIR D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, p. 7–12, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W16-2502.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BOSC T. & VINCENT P. (2018). Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 1522–1532 : Association for Computational Linguistics.

CHANG T., CHI T., TSAI S. & CHEN Y. (2018). xSense : Learning Sense-Separated Sparse Representations and Textual Definitions for Explainable Word Sense Networks. arXiv : 1809.03348.

CHODOROW M. S., BYRD R. J. & HEIDORN G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *23rd Annual Meeting of the Association for Computational Linguistics*, p. 299–304, Chicago, Illinois, USA : Association for Computational Linguistics. DOI : 10.3115/981210.981247.

COAVOUX M. (2017). *Discontinuous Constituency Parsing of Morphologically Rich Languages*. Thèse de doctorat, Univ Paris Diderot, Sorbonne Paris Cité.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

ELLOUMI Z., BLANCHON H., SERASSET G. & BESACIER L. (2015). METEOR For Multiple Target Languages Using DBnary. In *MT Summit 2015*, Miami, United States. HAL : hal-01350109.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database*. Bradford Books.

GADETSKY A., YAKUBOVSKIY I. & VETROV D. (2018). Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 266–271 : Association for Computational Linguistics.

GAUME B., HATHOUT N. & MULLER P. (2004). Word sense disambiguation using a dictionary for sense similarity measure. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.3115/1220355.1220528.

GAUME B., NAVARRO E., DESALLE Y. & GAILLARD B. (2014). Mesurer la similarité structurelle entre réseaux lexicaux. In *TALN-20 2014*, Proceedings of TALN-20 2014 : Atelier RLTLN, Réseaux Lexicaux et Traitement des Langues Naturelles, Marseille, France. HAL : hal-01321990.

GLADKOVA A., DROZD A. & MATSUOKA S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings : what works and what doesn't. In *SRWHLT-NAACL*.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

HATHOUT N. & SAJOUS F. (2016). Wiktionnaire's Wikicode GLAWIfied : a Workable French Machine-Readable Dictionary. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. GROBELNIK, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia : European Language Resources Association (ELRA).

HILL F., CHO K. & KORHONEN A. (2016a). Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1367–1377, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1162.

HILL F., CHO K., KORHONEN A. & BENGIO Y. (2016b). Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, **4**, 17–30.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). FlauBERT : Unsupervised Language Model Pre-training for French. arXiv preprint : 1912.05372.

LENCI A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, **4**, 151–171.

LEVY O. & GOLDBERG Y. (2014a). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, p. 171–180 : Association for Computational Linguistics. DOI : 10.3115/v1/W14-1618.

LEVY O. & GOLDBERG Y. (2014b). Neural word embedding as implicit matrix factorization. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. D. LAWRENCE & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 27*, p. 2177–2185. Curran Associates, Inc.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2019). CamemBERT : a Tasty French Language Model. arXiv preprint : 1911.03894.

MICKUS T., PAPERNO D. & CONSTANT M. (2019). Mark my Word : A Sequence-to-Sequence Approach to Definition Modeling. *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*. HAL : hal-02362397.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. arXiv : 1301.3781.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013b). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, p. 746–751.

NORASET T., LIANG C., BIRNBAUM L. & DOWNEY D. (2017). Definition modeling : Learning to define word embeddings in natural language. In *AAAI*.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 311–318, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

SCHÄFER R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. BAŃSKI, H. BIBER, E. BREITENEDER, M. KUPIETZ, H. LÜNGEN & A. WITT, Éds., *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster : UCREL IDS.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

SIERRA G., TORRES-MORENO J.-M. & MOLINA A. R. (2015). Regroupement sémantique de définitions en espagnol. arXiv : 1501.04920.

SWINGER N., DE-ARTEAGA M., IV N. T. H., LEISERSON M. D. M. & KALAI A. T. (2018). What are the biases in my word embedding ? arXiv preprint : 1812.08769.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30*, p. 5998–6008. Curran Associates, Inc.

YANG L., KONG C., CHEN Y., LIU Y., FAN Q. & YANG E. (2019). Incorporating sememes into chinese definition modeling. arXiv preprint : 1905.06512.

ZHANG H., DU Y., SUN J. & LI Q. (2019). Improving Interpretability of Word Embeddings by Generating Definition and Usage. arXiv preprint : 1912.05898.

ZHU R., NORASET T., LIU A., JIANG W. & DOWNEY D. (2019). Multi-sense Definition Modeling using Word Sense Decompositions. arXiv preprint : 1909.09483.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Du bon usage d'ingrédients linguistiques spéciaux pour classer des recettes exceptionnelles

Elham Mohammadi [1, 2, 4] Louis Marceau [2] Eric Charton [2] Leila Kosseim [1]
Luka Nerima [3] Marie-Jean Meurs [4]
(1) Université Concordia, Montréal, Québec Canada
(2) Banque Nationale du Canada (BNC), Montréal, Québec Canada
(3) Université de Genève (UNIGE), Genève, Suisse
(4) Université du Québec à Montréal (UQAM), Montréal, Québec Canada
{elham.mohammadi,leila.kosseim}@concordia.ca, louis.marceau@bnc.ca,
eric.charton@bnc.ca, luka.nerima@unige.ch, meurs.marie-jean@uqam.ca

RÉSUMÉ

Nous présentons un modèle d'apprentissage automatique qui combine modèles neuronaux et linguistiques pour traiter les tâches de classification dans lesquelles la distribution des étiquettes des instances est déséquilibrée. Les performances de ce modèle sont mesurées à l'aide d'expériences menées sur les tâches de classification de recettes de cuisine de la campagne DEFT 2013 (Grouin *et al.*, 2013). Nous montrons que les plongements lexicaux (*word embeddings*) associés à des méthodes d'apprentissage profond obtiennent de meilleures performances que tous les algorithmes déployés lors de la campagne DEFT. Nous montrons aussi que ces mêmes classifieurs avec plongements lexicaux peuvent gagner en performance lorsqu'un modèle linguistique est ajouté au modèle neuronal. Nous observons que l'ajout d'un modèle linguistique au modèle neuronal améliore les performances de classification sur les classes rares.

ABSTRACT

Using Special Linguistic Ingredients to Classify Exceptional Recipes

We propose a joint model composed of neural and linguistic sub-models, to address classification tasks in which the distribution of labels over samples is imbalanced. Different experiments are performed on tasks 1 and 2 of the DEFT 2013 shared task (Grouin *et al.*, 2013), which focused on classification of cooking recipes based on difficulty or meal type. In one set of experiments, the joint model is used for both classification tasks, whereas the second set of experiments involves using the neural sub-model, independently. This allows us to measure the impact of using linguistic features in the joint model. The results for both tasks show that adding a linguistic model to the neural model improves classification performance on the rare classes.

MOTS-CLÉS : Classification de textes, apprentissage profond, caractéristiques linguistiques.

KEYWORDS: Text classification, deep learning, linguistic features, cooking recipes.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 81–94.
HAL : hal-02784757.

1 Introduction

Différentes techniques issues du traitement automatique des langues (TAL) ont été utilisées au fil des ans pour traiter la tâche de classification de textes. Avec l'émergence de corpus de taille plus importante et l'avènement de l'apprentissage profond, les architectures de réseaux neuronaux sont devenues de plus en plus populaires dans l'exécution de différentes tâches de TAL (Amini *et al.*, 2019). Toutefois, malgré les progrès réalisés par l'apprentissage profond et l'obtention de résultats de pointe dans de nombreuses tâches, peu d'études ont abordé le défi que représentent les tâches de classification sur des ensembles de données déséquilibrés. Ces tâches de classification correspondent pourtant à de nombreux scénarios et applications de la vie réelle (Johnson & Khoshgoftaar, 2019), et représentent toujours un défi majeur.

Dans cet article, nous nous intéressons à deux tâches de classification multi-classes avec une distribution très déséquilibrée des étiquettes sur les données et nous mesurons l'efficacité de différentes méthodes pour relever un tel défi. Les tâches considérées sont la classification des recettes de cuisine selon 4 niveaux de difficulté (*Très facile*, *Facile*, *Assez difficile*, *Difficile* – Tâche 1) et la classification en fonction du type de repas (*Entrée*, *Plat principal*, *Dessert* – Tâche 2). Ces tâches et les ensembles de données, tous deux très déséquilibrés, sont issus de la campagne d'évaluation DEFT (DÉfi Fouille de Textes) 2013 (Grouin *et al.*, 2013). Nous expérimentons un modèle neuronal qui utilise des plongements lexicaux pré-entraînés comme caractéristiques d'entrée puis un modèle hybride composé de sous-modèles neuronaux et linguistiques et mesurons leur efficacité pour gérer une distribution de classe déséquilibrée. Par cette double expérimentation, nous cherchons à évaluer dans quelle mesure les plongements lexicaux pré-entraînés sont réellement en mesure de supprimer, dans une tâche de classification, tout recours au prétraitement linguistique. Nous souhaitons en particulier évaluer si l'influence du pré-traitement linguistique des données textuelles sur le caractère discriminant du classifieur demeure, en particulier lorsque ce classifieur est déjà renforcé par des plongements lexicaux.

Cet article est organisée comme suit : dans la section 2 nous passons brièvement en revue les travaux antérieurs connexes. La section 3 présente les ensembles de données utilisés et détaille les particularités de la campagne d'évaluation DEFT 2013. L'architecture globale du modèle, les sous-modèles et les différentes configurations utilisées sont décrits dans la section 5. La section 6 analyse les résultats obtenus et la section 8 conclut ce travail.

2 État de l'art

Selon Johnson & Khoshgoftaar (2019), en apprentissage automatique, trois types d'approches pour traiter les données déséquilibrées ont été proposées : (1) les approches au niveau des données qui consistent à modifier la distribution des classes par un sous- ou sur-échantillonnage des données limitant le déséquilibre ; (2) les approches algorithmiques qui adaptent l'apprentissage pour prendre en compte le déséquilibre en utilisant par exemple des poids sur les classes pour attribuer une pénalité plus importante à une erreur commise dans une classe rare par rapport à une classe fréquente ; (3) les approches hybrides qui combinent les approches (1) et (2) pour mieux gérer une distribution déséquilibrée.

Tâche 1 - Niveau de difficulté	Entraînement		Développement		Test	
	Nb instances	Proportion	Nb instances	Proportion	Nb instances	Proportion
Très facile	5569	50,2%	1393	50,2%	1132	49,0%
Facile	4601	41,5%	1151	41,5%	968	41,9%
Assez difficile	855	7,7%	213	7,7%	189	8,2%
Difficile	64	0,6%	16	0,6%	20	0,9%
Total	11089	100%	2773	100%	2309	100%
Tâche 2 - Type de plat	**Entraînement**		**Développement**		**Test**	
	Nb instances	**Proportion**	**Nb instances**	**Proportion**	**Nb instances**	**Proportion**
Entrée	2599	23,4%	647	23,3%	562	24,4%
Plat	5167	46,6%	1280	46,1%	1084	47,0%
Dessert	3323	30,0%	846	30,5%	661	28,6%
Total	11089	100%	2773	100%	2307	100%

TABLE 1 – Composition des ensembles de données pour les tâches 1 et 2

Cependant, ces méthodes peuvent être d'une efficacité limitée dans le cas de classes extrêmement déséquilibrées. Krawczyk (2016) montre qu'il est alors intéressant d'extraire les caractéristiques discriminantes, en tenant compte du déséquilibre entre classes. Utiliser ces approches parallèlement aux représentations distribuées dans une architecture profonde améliore les résultats dans plusieurs types de tâches (Bogdanova *et al.*, 2017). De même, le modèle d'étiquetage morpho-syntaxique (POS) proposé par Bach *et al.* (2019) – un réseau bidirectionnel à mémoire (BiLSTM) suivi d'une couche de champs aléatoires conditionnels (CRF) prédisant les étiquettes – obtient des performances améliorées quand il est enrichi avec des caractéristiques conçues manuellement. Pour la classification des textes courts, les travaux de Wang *et al.* (2017) utilisent un modèle faisant appel à des représentations implicites (plongements lexicaux et plongements de caractères pré-entrainés) et explicites (concepts, extraits d'une base de connaissances). En alimentant un modèle convolutif à branches avec ces représentations, les auteurs obtiennent des résultats à l'état de l'art. Cela suggère que l'enrichissement des caractéristiques par des éléments issus d'une base de connaissances améliore les capacités de classification. Dans ce travail, nous ajoutons des caractéristiques linguistiques aux modèles neuronaux et mesurons le gain de performance obtenu, en mettant l'accent sur les classes minoritaires.

3 Ensembles de données

Nous avons choisi de retenir pour nos expériences le corpus de données proposé dans le cadre de la campagne d'évaluation Défi Fouille de Texte 2013 (DEFT 2013) [1]. DEFT est l'une des rares séries de campagnes d'évaluation annuelles à proposer à la communauté scientifique du TALN des tâches de classification en français, sur des corpus volumineux, avec des annotations originales. Elle offre également l'avantage de susciter l'intérêt de nombreux laboratoires et donc d'impliquer la contribution de plusieurs équipes qui appliquent, pour répondre à la tâche, des méthodes de classification très différenciées.

La tâche 2013 n'a pas échappé à cette tradition et a éveillé l'intérêt de 7 équipes qui représentent leurs laboratoires en France et au Canada. L'intérêt particulier de la campagne 2013 pour l'expérimentation que nous souhaitions mener est précisément sa date. Les méthodes proposées par ces 7 équipes établissent un état de l'art sur une tâche de classification de texte dé-balancée précisément quelques mois avant la généralisation, dans la littérature, de l'usage de l'apprentissage profond, puis des

1. https ://deft.limsi.fr/2013/

plongements lexicaux. Elle nous permet donc de comparer avec précision les performances de ces méthodes classiques avec les approches récentes, en bénéficiant d'un protocole expérimental rigoureux et, comme il est d'usage dans une campagne d'évaluation, fiable puisque géré par une tierce partie tant pour ce qui est de l'annotation des corpus que de la mesure de performance des systèmes proposés.

Le corpus utilisé pour le défi 2013 est composé de recettes de cuisine extraites du site Marmiton. Marmiton.org est l'un des sites culinaire francophone de large audience [2]. Les recettes rassemblées dans la base de données de Marmiton.org depuis 1999 sont proposées par les internautes via un formulaire validé pour publication par l'équipe de Marmiton. Lors de la soumission d'une recette, les internautes doivent indiquer le type de plat, le niveau de difficulté, le coût et le type de cuisson en sélectionnant des valeurs parmi des listes de choix pré-établies. Les paramètres numériques de la recette tels les temps de préparation et de cuisson, et le nombre de convives, sont à renseigner dans des champs contraints. Les ingrédients, les consignes de préparation et la boisson conseillée sont recueillis dans des champs en texte libre.

Le corpus d'entraînement contient 13.864 recettes pour un volume de données de 19,2 MB. Les corpus de test pour les tâches 1, 2 et 4 sont composés respectivement de 2309, 2307 et 2306 recettes pour des volumes de données de 3MB, 2,9MB et 2,2MB. Les recettes sont fournies au format XML. Chaque fichier du corpus d'entraînement contient le titre de la recette, son type, son niveau, son coût, la liste non normalisée de ses ingrédients et leur quantité d'usage, ainsi que les indications de préparation en texte libre.

Les données utilisées pour nos expériences sont celles des deux premières tâches de la campagne d'évaluation DEFT 2013 (Grouin *et al.*, 2013). La composition de ces ensembles de données est détaillée dans le tableau 1. Les données de la tâche 1 sont des recettes de cuisine en français, étiquetées avec leur niveau de difficulté respectifs sur une échelle de 1 à 4 (1 pour *Très facile*, 4 pour *Difficile*). La répartition des étiquettes dans cet ensemble de données est très déséquilibrée, avec plus de 90% des échantillons portant une étiquette *Très facile* ou *Facile* et un nombre nettement inférieur d'échantillons portant une étiquette *Assez difficile* ou *Difficile*. Les données de la tâche 2 sont des recettes de cuisine en français, étiquetées avec le type de plat de la recette, soit *Entrée*, *Plat principal* ou *Dessert*. Bien que la distribution des étiquettes ne soit pas aussi déséquilibrée que celle de la tâche 1, près de la moitié des échantillons appartiennent à la classe *Plat principal*, ce qui pose le problème d'un ensemble de données déséquilibré dans la tâche 2 également.

Lors de DEFT 2013, les données ont été publiées en deux parties, pour l'entraînement et le test. Pour les expériences rapportées dans ce document, 20% des données d'entraînement ont été utilisées pour la mise au point du modèle (développement) et le corpus de test original de la campagne a été utilisé pour comparer les résultats finaux. Les proportions des différentes parties du corpus sont présentées dans le tableau 1.

4 Méthodes de classification utilisées pour la comparaison

Dans la perspective de comparer efficacement les performances des algorithmes de classification avant et après l'apparition des méthodes à base d'apprentissage profond renforcé par des plongements lexicaux, il est important que les familles d'algorithmes utilisées pour les deux tâches principales de DEFT 2013 puissent être considérées comme représentatives de l'état de l'art de cette période.

2. Avec plus de 300.000 visiteurs par jour (chiffres Smart AdServer mars 2010, source : http ://www.marmiton.org/)

Dans nos analyses, nous utiliserons en tant que référence de comparaison les résultats obtenus par les systèmes proposés par les trois premières équipes, tant pour la tâche 1 que pour la tâche 2 de DEFT 2013. Ces trois équipes ont utilisées des méthodes de pré-traitement linguistique plus ou moins sophistiquées.

Pour la tâche 1, (Collin *et al.*, 2013) applique diverses méthodes de pré-traitement lexical et de sélection de variables. La classification est ensuite conduite avec l'algorithme d'entropie maximale. En ce qui concerne (Bost *et al.*, 2013), un algorithme de boosting est d'abord utilisé pour la tâche 1, associé à une préparation statistique des paramètres (polygrammes et données numériques issues d'observations sur le corpus). Des algorithmes tels que le SVM ainsi qu'une méthode issue de la de recherche d'information (similarité cosinus) sont également testés. Un algorithme de fusion de résultat de classifieur par combinaison linéaire est également utilisé. L'équipe (Charton *et al.*, 2013) qui obtient les meilleurs résultats sur la tâche 1, utilise un modèle d'arbre de décision dont les feuilles sont des fonctions de régression logistique (LMT). Pour la tâche 2, (Collin *et al.*, 2013) applique à nouveau diverses méthodes de pré-traitement lexical et de sélection de variables, accompagnées de variables numériques construites par calcul. La classification est ensuite conduite avec un algorithme propriétaire non décrit. Pour ce qui est de (Bost *et al.*, 2013) la tâche 2 est conduite avec des modèles classiques (boosting, SVM, méthode Electre), dont les résultats sont combinés par fusion. L'équipe (Charton *et al.*, 2013) obtient ses meilleurs résultats avec un classifieur SVM.

De manière générale, et au delà du classement final, on observe que ces trois équipes obtiennent avec leurs systèmes, appliqués sur le corpus de DEFT 2013, des résultats très proches avec les algorithmes et les techniques de pré-processing à l'état de l'art de l'époque : SVM, boosting, combinés avec une sélection fine des paramètres retenus pour l'apprentissage. On notera que des études subséquentes de (Charton *et al.*, 2014) sur cette campagne d'évaluation, entreprennent une comparaison systématique de classifieurs (Régression Logistique, Réseau Bayésien, SVM, arbres, Naîves Bayes) sur le corpus DEFT 2013, sans remettre en cause les résultats de la campagne.

Dans notre cadre expérimental, on peut donc considérer que les résultats de systèmes de classification de texte décrits dans la littérature et appliqués sur le corpus DEFT 2013 offrent une bonne représentation de l'état de l'art en matière de classification supervisée de documents textuels avant 2014. Et qu'il est en conséquence possible d'utiliser le corpus de recherche concerné comme base de comparaison avec des techniques plus récentes.

Par commodité, dans la suite de cette communication, nous désignerons par l'expression *modèle classique* les classifieurs non neuronaux utilisés sur les corpus DEFT 2013. Nous désignerons par l'expression *modèle linguistique* l'ensemble de pré-traitements appliqués aux corpus d'apprentissage textuels.

5 Conception des modèles

Notre but est de conduire deux séries de comparaisons. En premier lieu, nous voulons donc déterminer si un modèle neuronal associé à des plongements lexicaux est capable de performances supérieures à un *modèle classique* associé à un *modèle linguistique*. Nous désignerons dans nos résultats cette série d'expérience par le qualificatif de *modèle neuronal*. En second lieu, nous voulons déterminer si ce même modèle neuronal associé à des plongements lexicaux et complété du *modèle linguistique* améliore ses performances. Nous désignerons, dans nos résultats, cette série d'expérience par le

qualificatif de *modèle combiné*.

Dans cette section, nous présentons tout d'abord l'architecture des modèles neuronaux. En utilisant trois modèles de plongements (BERT, FLAUBERT, CAMEMBERT) que nous exploitons avec une architecture neuronale récurrente ou convolutive, nous obtenons 6 modèles (voir Table 2). Nous présentons ensuite le *modèle linguistique* et les détails de sa conception.

5.1 Modèle neuronal

Plongement lexical. Un plongement lexical est utilisé pour transformer la concaténation du titre d'une recette et du texte de préparation en vecteurs denses. Dans ce travail, trois plongements lexicaux pré-entraînés à base de transformeurs sont utilisés : la version multilingue de BERT (Devlin *et al.*, 2019) ainsi que CamemBERT (Martin *et al.*, 2019) et FlauBERT (Le *et al.*, 2020), construits sur des données françaises en utilisant un modèle BERT. Seules les caractéristiques de la dernière couche des modèles sont extraites, donnant une représentation dense de taille 768 pour chaque token.

Architecture récurrente. Pour la couche cachée de l'architecture récurrente, des réseaux de neurones récurrents à portes (GRUs) (Cho *et al.*, 2014) ont été utilisés, car ayant moins de paramètres, ils sont moins sujets au sur-apprentissage (Chung *et al.*, 2014) que les LSTM (Hochreiter & Schmidhuber, 1997). Un GRU bidirectionnel traite les plongements lexicaux consécutivement vers l'avant et vers l'arrière. La sortie de la couche GRU est ensuite transmise à une couche d'attention qui calcule sa moyenne pondérée à l'aide de l'équation $Attention = \sum_{t=1}^{n} y_t \omega_t$ où y_t représente la sortie de la couche GRU au pas de temps t et où ω_t est le poids attribué à y_t par le mécanisme d'attention.

Architecture convolutive. Par soucis de comparaison, nous avons également testé une architecture convolutive. Un réseau de neurones convolutif (CNN) (LeCun *et al.*, 1999) traite des n-grammes d'entrée (n représentations consécutives de tokens) en utilisant des filtres de convolution de taille n. La couche cachée est suivie d'une couche de regroupement moyen, maximum ou combinant les deux.

5.2 Modèle linguistique

Pour chacune des tâches de DEFT 2013 (tâche 1, identification d'une classe de difficulté de recette, et tâche 2, identification du type de plat préparé), un extracteur de caractéristiques est conçu. Les corpus d'apprentissage de la campagne DEFT 2013 sont composés de titres de recettes et de textes libres décrivant ces recettes. Le principe des extracteurs de caractéristiques est d'identifier des paramètres discriminants qui pourraient être dérivés de ces contenus textuels. Par une analyse quantitative systématique, on peut par exemple découvrir que le nombre de mots dans le titre d'une recette, ou encore le nombre d'ingrédients qu'elle contient, sont plus caractéristiques de sa difficulté que les mots (en tant qu'entités lexicales) qui composent ce titre. On espère ainsi, en utilisant ces paramètres dérivés du corpus textuel, plutôt que des sacs de mots, maximiser les performances du classifieur.

Aidé des observations faites sur les corpus, on bâtit des extracteurs dont le rôle est de construire des vecteurs avec les caractéristiques identifiées. Ce sont ces vecteurs qui seront utilisés pour entraîner et faire fonctionner le classifieur. Cette démarche de construction de vecteurs d'apprentissage constitués

de paramètres finement sélectionnés, voir fabriqués de toutes pièces (comme dans le cas des comptages de mots), est classique des systèmes de classification tels qu'on les observe dans la littérature avant la généralisation de l'utilisation des réseaux de neurones et l'avènement des plongements. C'est cette même méthode d'ingénierie des paramètres d'apprentissage que la littérature récentes sur les plongements considère comme moins utile voire superflue (Liu *et al.* (2015); Tymoshenko *et al.* (2016)). L'un des buts de nos expériences est de déterminer si le caractère discriminant des plongements est effectivement suffisant pour rendre l'ingénierie de paramètres inutiles.

Ces extracteurs de caractéristiques sont détaillés dans Charton *et al.* (2013), et leurs grands lignes sont présentées ci-après.

Caractéristiques pour la tâche 1 (niveau de difficulté). Sont utilisés le nombre de tokens dans le titre de la recette et dans la partie préparation, le nombre d'ingrédients, le coût du repas sur une échelle de 3 points, la présence de 22 mots et de 48 trigrammes discriminants, et enfin, le nombre de verbes dans la recette qui appartiennent à 3 familles de verbes discriminants. L'extraction de ces caractéristiques donne un vecteur de taille 77 pour chaque recette.

Caractéristiques pour la tâche 2 (type de plats). Comme pour la tâche 1, le nombre de tokens dans le titre de la recette et la partie préparation, le nombre d'ingrédients et le coût associé au repas sur une échelle de 3 points sont les quatre premières caractéristiques. S'y ajoutent 1231 noms d'ingrédients, 48 trigrammes discriminants et le nombre de verbes dans la recette appartenant à chacune des trois familles prédéfinies. Un vecteur de caractéristiques de taille 1286 est extrait pour chaque recette.

Pour être utilisé par le classifieur dans les expériences menées avec les *modèles combinés*, les vecteurs de caractéristiques ainsi extraits sont ensuite transmis à un réseau neuronal à simple couche à action anticipée, en faisant correspondre chaque vecteur de caractéristiques à un vecteur de même taille pour obtenir les représentations de sortie du modèle linguistique.

5.3 Composante de fusion

La composante de fusion concatène la sortie des deux modèles puis applique une couche entièrement connectée sur le vecteur résultant, lui faisant correspondre un vecteur de taille 4 dans le cas de la tâche 1, et de taille 3 dans le cas de la tâche 2. Cette couche est suivie d'une fonction d'activation softmax qui produit les probabilités des classes. Afin de mesurer l'impact du modèle linguistique, certaines expériences n'utilisent que le modèle neuronal. La fusion est alors remplacée par une couche entièrement connectée faisant correspondre la sortie de la couche d'attention avec le nombre de classes, suivie d'une fonction d'activation produisant la distribution des probabilités sur les classes.

5.4 Entraînement

Les modèles ont été entraînés avec des lots de taille 32 et 20 passes (*epochs*). Les paramètres ont été choisis lors de la passe qui a obtenu le meilleur micro score sur les données de développement. Le tableau 2 présente les hyperparamètres des modèles et des détails sont expliqués ci-après.

Optimiseur. AdamW (Loshchilov & Hutter, 2019) a été utilisé pour optimiser l'apprentissage. Pour tous les modèles, le taux d'apprentissage initial est de 10^{-3}. Il a été adapté pour les modèles CNN à 10^{-4} après deux passes dans la tâche 1, et après cinq passes dans la tâche 2.

	Modèle	Tâche 1			Tâche 2		
		#HL / #KH	#HN / #K	Regroupement	#HL / #KH	#HN / #K	Regroupement
Neuronal	CNN-BERT	1, 2, 3, 4	300, 200, 100, 100	max	2	200	max
	GRU-BERT	1	64	attention	2	32	attention
	CNN-FlauBERT	2	200	max, moyen	1, 2	400, 200	max
	GRU-FlauBERT	1	64	attention	2	32	attention
	CNN-CamemBERT	2, 3	250, 50	max	2	200	max, moyen
	GRU-CamemBERT	2	32	attention	2	64	attention
Combiné	CNN-BERT	2	250	max	2	200	max, moyen
	GRU-BERT	1	64	attention	2	32	attention
	CNN-FlauBERT	1, 2	300, 200	max, moyen	1, 2	300, 200	max, moyen
	GRU-FlauBERT	1	64	attention	2	32	attention
	CNN-CamemBERT	1	400	max, moyen	2	400	max, moyen
	GRU-CamemBERT	2	32	attention	2	32	attention

TABLE 2 – Hyperparamètres pour chaque modèle. *#HL / #KH* : Nombre de couches cachées dans les modèles récurrents ou hauteur du noyau dans les CNN. *#HN / #K* : Nombre de noeuds cachés dans chaque couche récurrente ou nombre de noyaux dans les CNN.

Poids des classes. Des poids de classe ont été ajoutés dans la fonction de perte d'entropie croisée. Pour les expériences n'utilisant que le modèle neuronal, les poids ont été calculés automatiquement, en tenant compte de la proportion des échantillons de chaque classe. Dans les expériences utilisant le modèle combiné, les poids ont été manuellement réglés à 0,1, 0,1, 0,2 et 0,6 (correspondant respectivement aux classes *Très facile*, *Facile*, *Assez difficile* et *Difficile* pour la tâche 1, et à 0,6, 0,3 et 0,1 (correspondant aux classes *Entrée*, *Plat principal* et *Dessert*, respectivement) pour la tâche 2.

Régularisation. Afin de régulariser le réseau, l'optimiseur a été utilisé avec un taux de décroissance du poids de 0,02 . De plus, une couche de décroissance (*dropout*) avec une probabilité de 0,2 a été appliquée sur la concaténation de la sortie des deux modèles dans la composante de fusion et sur la sortie de la couche attention/regroupement dans les modèles neuronaux.

Réglage fin des modèles BERT et CamemBERT. Comme dans les expériences qui ont fait appel à des modèles combinés et neuronaux, la couche de plongement lexical a été figée, les modèles BERT et CamemBERT ont été affinés sur les deux tâches en tant qu'expériences supplémentaires.

6 Résultats

Les résultats des différentes expériences pour la tâche 1, ainsi que ceux des équipes DEFT 2013 ayant obtenu les meilleurs micro scores sont présentés dans le tableau 3. Dans tous les cas, le modèle combiné a permis d'obtenir des performances (souvent très) supérieures en termes de micro et macro F1 par rapport à un modèle uniquement neuronal. L'amélioration des résultats peut être également observée en termes de macro précision et de macro rappel.

Les scores micro et macro les plus élevés (sauf pour la macro précision (Charton *et al.*, 2014)) sont obtenus par des modèles combinés qui utilisent les plongements CamemBERT, illustrant leur efficacité sur des données en français. En outre, le tableau 3 montre que le modèle combiné CNN-CamemBERT dépasse largement tous les autres en termes de micro score, de macro F1 et de rappel dans la tâche 1. Dans 3 des 4 classes, le meilleur score F1 est également obtenu par des modèles combinés, en particulier ceux qui utilisent CamemBERT. Les résultats des modèles qui utilisent CamemBERT montrent que l'ajout de caractéristiques linguistiques améliore les performances par classe et que ces caractéristiques ont complété plus efficacement les plongements CamemBERT que BERT.

	Modèle	Développement				Test			
		Micro Score	Macro F1	Macro P	Macro R	Micro Score	Macro F1	Macro P	Macro R
Neuronal	CNN-BERT	61,5	39,3	41,1	37,6	58,8	37,7	39,4	36,1
	GRU-BERT	59,9	38,5	38,5	38,4	58,0	36,8	37,0	36,7
	BERT affiné	56,9	39,3	42,9	36,2	55,9	36,0	36,8	35,3
	CNN-FlauBERT	59,4	36,2	41,9	31,8	58,1	30,0	28,9	31,3
	GRU-FlauBERT	56,2	37,4	39,1	35,8	54,4	33,6	33,9	33,3
	CNN-CamemBERT	60,9	42,7	43,4	42,0	59,3	38,6	39,9	37,3
	GRU-CamemBERT	62,4	36,1	38,1	34,3	60,1	36,9	40,1	34,1
	CamemBERT affiné	61,2	37,3	38,5	36,2	59,3	37,6	38,9	36,4
Combiné	CNN-BERT	64,5	49,1	60,0	41,6	62,0	47,3	59,3	39,3
	GRU-BERT	65,8	41,7	45,5	38,5	63,1	39,3	42,1	36,8
	CNN-FlauBERT	63,7	45,1	56,7	37,4	60,6	43,3	55,1	35,7
	GRU-FlauBERT	62,6	49,0	49,8	**48,3**	61,2	44,3	45,2	**43,4**
	CNN-CamemBERT	**66,4**	50,3	58,5	44,2	**63,8**	**50,0**	62,0	42,0
	GRU-CamemBERT	65,3	**51,1**	**68,5**	40,8	63,1	40,5	42,5	38,7
DEFT 2013	Première équipe (Charton *et al.*, 2014)	-	-	-	-	62,5	48,4	**68,2**	37,5
	Seconde équipe (Collin *et al.*, 2013)	-	-	-	-	61,2	45,1	52,4	39,5
	Troisième équipe (Bost *et al.*, 2013)	-	-	-	-	59,2	45,3	63,3	35,3

	Modèle	Développement				Test			
		Très facile	Facile	Assez difficile	Difficile	Très facile	Facile	Assez difficile	Difficile
Neuronal	CNN-BERT	66,3	61,6	25,1	0,0	63,1	59,7	23,5	0,0
	GRU-BERT	68,0	56,0	**29,9**	0,0	65,8	55,2	26,1	0,0
	CNN-FlauBERT	68,1	53,0	0,9	0,0	67,8	50,9	0,0	0,0
	GRU-FlauBERT	66,7	47,5	22,7	9,1	65,9	44,4	22,0	0,0
	CNN-CamemBERT	71,0	51,9	22,6	19,5	69,9	51,2	19,2	9,8
	GRU-CamemBERT	71,1	56,7	7,3	0,0	68,7	55,0	12,6	0,0
Combiné	CNN-BERT	72,1	60,8	24,9	21,1	70,0	58,1	22,5	17,4
	GRU-BERT	73,9	61,1	22,6	0,0	72,0	58,1	18,9	0,0
	CNN-FlauBERT	73,6	54,9	4,5	20,0	71,5	51,1	5,8	17,4
	GRU-FlauBERT	67,8	**62,2**	22,8	**33,3**	67,1	**61,5**	16,9	22,6
	CNN-CamemBERT	74,0	61,8	27,0	27,3	72,2	59,0	25,2	**25,0**
	GRU-CamemBERT	**74,4**	58,3	27,9	11,8	**72,5**	56,3	**29,4**	0,0
DEFT 2013	Première équipe (Charton *et al.*, 2014)	-	-	-	-	71,7	56,2	18,8	9,5
	Seconde équipe (Collin *et al.*, 2013)	-	-	-	-	69,2	57,0	26,1	16,0
	Troisième équipe (Bost *et al.*, 2013)	-	-	-	-	68,6	52,5	15,6	9,5

TABLE 3 – Résultats généraux et par classe (score F1) pour la tâche 1.

Sur l'ensemble des tests, le modèle combiné CNN-CamemBERT obtient des scores F1 supérieurs au meilleur modèle de référence. Ce modèle combiné obtient également le meilleur score F1, soit 25%, dans la classe *Difficile*, qui est la plus rare. Seuls 3 des 8 modèles obtiennent un score F1 non nul dans la classe *Difficile* et deux sont des modèles combinés. Les résultats par classe montrent l'efficacité des caractéristiques linguistiques lorsque la tâche implique un ensemble de données très déséquilibré. Le tableau 4 montre les résultats obtenus pour la tâche 2. Le modèle CamemBERT affiné (miam !) a obtenu la meilleure performance globale. Toutefois, le modèle combiné CNN-CamemBERT, déjà le meilleur dans la tâche 1, fait mieux dans la phase de test. Cela montre que le modèle combiné CNN-CamemBERT semble mieux généraliser en présence de nouvelles instances.

D'après le tableau 4, tous les modèles combinés surpassent leurs homologues neuronaux en termes de micro et macro scores. Toutefois, contrairement à la tâche 1, cette amélioration n'est pas assez importante pour faire la différence en score micro moyen. Cette performance peut s'expliquer par les caractéristiques linguistiques utilisées pour la tâche 2, qui sont peut-être moins représentatives des classes que dans la tâche 1.

En examinant les résultats de la tâche 1 dans le tableau 3, on peut observer qu'après ajout du modèle linguistique, lorsqu'il y a une amélioration, celle-ci est significativement plus importante pour les classes rares. On peut donc émettre l'hypothèse que le modèle combiné gère mieux une distribution déséquilibrée des étiquettes. Sachant que cette distribution est nettement plus déséquilibrée dans la tâche 1 que dans la tâche 2, le modèle combiné est donc plus efficace dans le premier cas que dans le second. Enfin, le tableau 4 montre également les scores F1 par classe obtenus par les trois équipes les plus performantes de la campagne DEFT 2013.

	Modèle	Développement				Test			
		Micro Score	Macro F1	Macro P	Macro R	Micro Score	Macro F1	Macro P	Macro R
Neuronal	CNN-BERT	86,4	84,9	85,7	84,2	85,9	84,9	85,6	84,2
	GRU-BERT	84,4	83,2	83,2	83,2	84,8	84,0	84,0	83,9
	BERT affiné	86,3	85,8	85,1	86,5	86,4	86,2	85,6	86,8
	CNN-FlauBERT	86,4	85,2	85,3	85,1	86,7	85,8	86,2	85,5
	GRU-FlauBERT	84,2	83,3	82,6	84,0	85,1	84,6	83,8	85,4
	CNN-CamemBERT	87,6	86,8	86,5	**87,2**	88,1	87,6	87,6	87,7
	GRU-CamemBERT	86,5	85,6	85,5	85,6	87,1	86,5	86,6	86,4
	CamemBERT affiné	**88,2**	**87,1**	**87,3**	86,9	88,1	87,4	87,5	87,3
Combiné	CNN-BERT	86,0	85,2	84,9	85,4	87,0	86,5	86,3	86,7
	GRU-BERT	85,0	84,2	83,9	84,6	85,5	85,0	84,8	85,2
	CNN-FlauBERT	86,6	85,3	85,8	84,8	87,6	86,8	87,5	86,1
	GRU-FlauBERT	85,3	83,5	84,3	82,8	86,1	85,0	86,1	84,0
	CNN-CamemBERT	87,5	86,8	86,4	87,1	88,6	**88,2**	88,0	**88,3**
	GRU-CamemBERT	86,9	86,1	85,8	86,5	87,8	87,3	87,2	87,4
DEFT 2013	Première équipe (Bost *et al.*, 2013)	-	-	-	-	**88,9**	**88,2**	**88,4**	88,1
	Seconde équipe (Charton *et al.*, 2014)	-	-	-	-	85,6	84,7	85,0	84,3
	Troisième équipe (Hamon *et al.*, 2013)	-	-	-	-	84,9	84,1	84,2	84,1

	Modèle	Développement			Test		
		Entrée	Plat	Dessert	Entrée	Plat	Dessert
Neuronal	CNN-BERT	70,2	86,5	97,6	71,0	86,4	96,8
	GRU-BERT	67,9	83,8	97,8	70,5	85,0	96,4
	CNN-FlauBERT	71,4	85,9	98,1	73,0	87,0	97,2
	GRU-FlauBERT	69,3	83,4	96,9	72,6	85,1	95,7
	CNN-CamemBERT	**75,1**	**87,1**	98,2	77,1	88,0	97,7
	GRU-CamemBERT	72,8	86,4	97,4	75,4	87,8	96,4
Combiné	CNN-BERT	72,0	85,6	97,7	75,0	87,2	97,3
	GRU-BERT	70,7	84,5	97,2	72,6	85,8	96,5
	CNN-FlauBERT	70,9	86,7	98,0	74,4	87,8	97,9
	GRU-FlauBERT	67,0	85,7	97,4	69,9	86,8	97,5
	CNN-CamemBERT	74,7	86,9	**98,6**	**78,1**	88,5	98,0
	GRU-CamemBERT	73,4	86,3	98,5	76,7	87,8	97,5
DEFT 2013	Première équipe (Bost *et al.*, 2013)	-	-	-	77,3	**88,8**	**98,6**
	Seconde équipe (Charton *et al.*, 2014)	-	-	-	70,3	85,6	97,9
	Troisième équipe (Hamon *et al.*, 2013)	-	-	-	69,4	84,8	98,2

TABLE 4 – Résultats généraux et par classe (score F1) pour la tâche 2.

Parmi les 8 modèles que nous avons développés, les résultats montrent que le modèle conjoint CNN-CamemBERT obtient les scores F les plus élevés pour les trois classes de l'ensemble des tests. Il obtient également les meilleurs résultats sur la classe *Entrée* qui est la classe la plus rare. Ceci est en accord avec l'hypothèse selon laquelle la force du modèle combiné réside dans le traitement des classes rares.

7 Discussion

Il est intéressant d'observer que sur les corpus de tests, aucun des modèles neuronaux dont les plongements sont non personnalisés, n'obtient de meilleurs performances que les *modèles classiques* associés aux modèles linguistiques déployés lors de la campagne DEFT 2013. Il apparaît également que les modèles neuronaux dont les plongements sont affinés sur les corpus d'entraînement, produisent eux aussi des résultats inférieurs à ceux des *modèles classiques* associés aux modèles linguistiques.

On peut envisager ces résultats comme une indication que les modèles neuronaux à apprentissage profond, y compris lorsqu'ils sont finement entraînés ne produisent pas automatiquement de performances supérieures à celles des *modèles classiques* quand ces derniers sont finement paramétrés. Ceci semble particulièrement vrai sur des corpus textuels dé-balancés : les modèles neuronaux sont

incapables, dans toutes les configurations de nos expériences qui ne font pas appel aux modèles linguistiques, de modéliser la classe *difficile* de la tâche 1 alors que les *modèles classiques* y parviennent. Cette difficulté qu'ont les modèles neuronaux à modéliser les classes sous représentées apparaît dans d'autres expériences (Marceau *et al.*, 2019; Johnson & Khoshgoftaar, 2019). Il conviendrait de mener d'autres études dans d'autres contextes pour confirmer ce point. Si l'incapacité des modèles à apprentissage profond à modéliser une classe fortement dé-balancée venait à être confirmée plus largement, elle les disqualifierait pour de nombreuses applications, au profit d'autres classifieurs beaucoup plus performants dans ce contexte, tels que XGBoost (Chen & Guestrin, 2016; Nielsen, 2016).

On remarque néanmoins que les modèles neuronaux, avec très peu d'efforts de préparation, atteignent des scores très proches (moins de 1 point d'écart pour Camembert) de ceux obtenus par les *modèles classiques*, lorsque les corpus sont mieux balancés. Nous observons pour finir que tous les modèles neuronaux, lorsqu'ils voient leur données d'apprentissage enrichies par les vecteurs de paramètres issus du modèle linguistique, obtiennent les meilleures performances, y compris sur les classes dé-balancés. On peut déduire de cette observation qu'un réseau de neurone à apprentissage profond — même associé à des plongements censés lui fournir des caractères linguistiques discriminants — gagne à voir ses données d'entraînement complétées par des paramètres complémentaires et sélectionnés pour leurs propriétés discriminantes, obtenues via une phase d'ingénierie des paramètres.

8 Conclusion

Ces travaux dédiés à la classification de textes présentent un modèle combiné, composé d'un modèle neuronal et d'un modèle linguistique. Ce modèle combiné est évalué sur les tâches 1 et 2 de la campagne DEFT 2013 (Grouin *et al.*, 2013), qui consistent à classer les recettes de cuisine en français en fonction du niveau de difficulté ou du type de repas. Les résultats de ces expériences montrent que, dans les deux tâches, les modèles combinés sont plus performants que leurs homologues uniquement neuronaux. Dans la tâche 1, le modèle combiné a pu obtenir les meilleurs résultats en termes de micro et macro F1 moyens, ce qui montre son efficacité dans les contextes de classes très déséquilibrées.

Dans notre contexte expérimental, la conception rapide d'un système performant est possible avec les modèles neuronaux combinés avec des plongements, si les classes ne sont pas exagérément débalancées. Cependant, le système le plus performant et robuste est obtenu en soumettant aux modèles neuronaux des caractères discriminants patiemment sélectionnés.

Reproductibilité

Pour faciliter la reproduction de nos travaux et permettre les comparaisons, nos systèmes sont disponibles en code source libre dans le dépôt suivant :
`https://github.com/cooking-classification/TALN2020`.
Les données peuvent être obtenues en contactant le comité d'organisation de la campagne DEFT 2013 (voir `https://deft.limsi.fr/2013/index.php`)

Références

AMINI H., FARAHNAK F. & KOSSEIM L. (2019). Natural Language Processing : An Overview. In M. BLOM, N. NOBILE & C. Y. SUEN, Éds., *Frontiers in Pattern Recognition and Artificial Intelligence*, volume 5, chapitre 3, p. 35–55. World Scientific. DOI : 10.1142/9789811203527_0003.

BACH N. X., DUY T. K. & PHUONG T. M. (2019). A POS Tagging Model for Vietnamese Social Media Text Using BiLSTM-CRF with Rich Features. In *Pacific Rim International Conference on Artificial Intelligence*, p. 206–219 : Springer.

BOGDANOVA D., FOSTER J., DZENDZIK D. & LIU Q. (2017). If You Can't Beat them Join them : Handcrafted Features Complement Neural Nets for Non-factoid Answer Reranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 121–131.

BOST X., BRUNETTI I., CABRERA-DIEGO L. A., COSSU J.-V., LINHARES A., MORCHID M., TORRES-MORENO J.-M., EL-BÈZE M. & DUFOUR R. (2013). Systèmes du LIA à DEFT 13. In *Actes du neuvième DÉfi Fouille de Textes (DEFT2013), atelier de la 20e conférence de la conférence sur le Traitement Automatique du Langage Naturel 2013 (TALN 2013) et de la 15ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2013)*, Les Sables-d'Olonne, France. HAL : hal-01313065.

CHARTON E., JEAN-LOUIS L., MEURS M.-J. & GAGNON M. (2013). Trois recettes d'apprentissage automatique pour un systeme d'extraction d'information et de classification de recettes de cuisines. In *Actes du neuvième DÉfi Fouille de Textes (DEFT2013), atelier de la 20e conférence de la conférence sur le Traitement Automatique du Langage Naturel 2013 (TALN 2013) et de la 15ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2013)*, p. 17–21, Les Sables-d'Olonne, France. https://deft.limsi.fr/actes/2013/pdf/deft13_submission_6.pdf.

CHARTON E., MEURS M.-J., JEAN-LOUIS L. & GAGNON M. (2014). Using Collaborative Tagging for Text Classification : From Text Classification to Opinion Mining. *Informatics*, **1**(1), 32–51. DOI : 10.3390/informatics1010032.

CHEN T. & GUESTRIN C. (2016). XGBoost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, p. 785–794.

CHO K., VAN MERRIENBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, p. 1724–1734, Doha, Qatar.

CHUNG J., GULCEHRE C., CHO K. & BENGIO Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS 2014 Deep Learning and Representation Learning Workshop*, Montreal, Canada.

COLLIN O., GUERRAZ A., HIOU Y. & VOISINE N. (2013). Participation de Orange Labs à DEFT 2013. In *Actes du neuvième DÉfi Fouille de Textes (DEFT2013), atelier de la 20e conférence de la conférence sur le Traitement Automatique du Langage Naturel 2013 (TALN 2013) et de la 15ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2013)*, p. 67–79, Les Sables-d'Olonne, France. https://deft.limsi.fr/actes/2013/pdf/deft13_submission_5.pdf.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (ACL/HLT 2019), p. 4171–4186, Minneapolis, Minnesota.

GROUIN C., PAROUBEK P. & ZWEIGENBAUM P. (2013). DEFT2013 se met à table : présentation du défi et résultats. In *Actes du neuvième DÉfi Fouille de Textes (DEFT2013), atelier de la 20e conférence de la conférence sur le Traitement Automatique du Langage Naturel 2013 (TALN 2013) et de la 15ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2013)*, Les Sables-d'Olonne, France. https://deft.limsi.fr/actes/2013/pdf/deft13_submission_0.pdf.

HAMON T., PÉRINET A. & GRABAR N. (2013). Efficacité combinée du flou et de l'exact des recettes de cuisine. In *Actes du neuvième DÉfi Fouille de Textes (DEFT2013), atelier de la 20e conférence de la conférence sur le Traitement Automatique du Langage Naturel 2013 (TALN 2013) et de la 15ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2013)*, Les Sables-d'Olonne, France. https://deft.limsi.fr/actes/2013/pdf/deft13_submission_1.pdf.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780.

JOHNSON J. M. & KHOSHGOFTAAR T. M. (2019). Survey on Deep Learning with Class Imbalance. *Journal of Big Data*, **6**(1), 27.

KRAWCZYK B. (2016). Learning from Imbalanced Data : Open Challenges and Future Directions. *Progress in Artificial Intelligence*, **5**(4), 221–232.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference (LREC 2020)*. arXiv : 1912.05372.

LECUN Y., HAFFNER P., BOTTOU L. & BENGIO Y. (1999). Object Recognition with Gradient-based Learning. In D. A. FORSYTH, J. L. MUNDY, V. DI GESÚ & R. CIPOLLA, Éds., *Shape, contour and grouping in computer vision*, volume 1681 de *Lecture Notes in Computer Science*, p. 319–345. Springer. DOI : 10.1007/3-540-46805-6_19.

LIU P., JOTY S. & MENG H. (2015). Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1433–1443.

LOSHCHILOV I. & HUTTER F. (2019). Decoupled Weight Decay Regularization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2019)*, New Orleans, Louisiana, USA.

MARCEAU L., QIU L., VANDEWIELE N. & CHARTON E. (2019). A comparison of deep learning performances with other machine learning algorithms on credit scoring unbalanced data. arXiv preprint : 1907.12363.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). CamemBERT : A Tasty French Language Model. arXiv preprint : 1911.03894.

NIELSEN D. (2016). Tree boosting with XGBoost- Why does XGBoost win "every" machine learning competition ? Mémoire de master, NTNU.

TYMOSHENKO K., BONADIMAN D. & MOSCHITTI A. (2016). Convolutional neural networks vs. convolution kernels : Feature engineering for answer sentence reranking. In *Proceedings of the*

2016 conference of the North American chapter of the association for computational linguistics : human language technologies, p. 1268–1278.

WANG J., WANG Z., ZHANG D. & YAN J. (2017). Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, p. 2915–2921.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude sur le résumé comparatif grâce aux plongements de mots

Valentin Nyzam Aurélien Bossard
Laboratoire d'Informatique Avancée de Saint-Denis, Université Paris 8
2 rue de la Liberté, 93526 Saint-Denis
valentin.nyzam@iut.univ-paris8.fr, aurelien.bossard@iut.univ-paris8.fr

RÉSUMÉ

Dans cet article, nous présentons une nouvelle méthode de résumé automatique comparatif. Ce type de résumé a pour objectif de permettre de saisir rapidement les différences d'information entre deux jeux de documents. En raison de l'absence de ressources disponibles pour cette tâche, nous avons composé un corpus d'évaluation. Nous présentons à la fois la méthodologie de son élaboration ainsi que le corpus lui-même. Notre méthode utilise les avancées récentes dans le calcul de similarité entre phrases afin de détecter les informations comparatives. Nous montrons que sur ce corpus, notre méthode est comparable en termes de qualité de résultats à une méthode de l'état de l'art, tout en réduisant d'un facteur dix le temps de calcul, la rendant donc exploitable dans le cadre de l'aide à l'analyse de documents.

ABSTRACT

Comparative summarization study using word embeddings.

This paper introduces a new method for comparative automatic summarization. This kind of summary allows to quickly identify the differences between two sets of documents. Because of the lack of evaluation resources, we built an evaluation corpus. We present both the methodology used to build this corpus, and the corpus itself. Our method makes use of recent advances in sentence similarity computation in order to detect comparative information. We show that on that corpus, our method compares to a state-of-the-art method in terms of quality and divides the computation time by a factor ten. This increase in computation speed makes automatic comparative summarization usable for support for document analysis.

MOTS-CLÉS : Résumé automatique, résumé comparatif, plongements de mots.

KEYWORDS: Automatic summarization, comparative summarization, word embeddings.

1 Introduction

Étant donné le nombre d'actualités disponibles en ligne, l'analyse automatique des articles de presse est devenue un enjeu important. Dans ce cadre, nous nous intéressons au résumé comparatif qui permet de regrouper les informations connexes au sein d'un groupe de documents qui traitent d'un même type de sujet pour les présenter à l'utilisateur sous la forme de courts résumés. Une telle présentation analytique est une réelle plus-value pour le lecteur du résumé. Elle met en effet en avant les informations comparables au sein de plusieurs documents, ce qui peut s'apparenter au résumé guidé par les aspects, mais de manière non supervisée. Cependant, cette tâche représente un défi pour le traitement automatique de la langue.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 95–107.
HAL : hal-02784758.

Il est en effet nécessaire d'avoir une connaissance approfondie des sujets traités par les documents afin de trouver de bonnes comparaisons. C'est pourquoi l'analyse manuelle prend généralement beaucoup de temps et de travail. La génération automatique d'un résumé qui mette automatiquement en évidence les informations connexes entre les sujets d'actualité serait ainsi une aide primordiale afin de pouvoir effectuer des analyses de manière plus simple et efficace. Dans la suite de cet article, nous appelons "informations comparatives" deux informations qui peuvent être rapprochées l'une de l'autre d'après le thème qu'elles portent.

Le résumé comparatif d'actualités cherche ainsi à comparer une paire de documents ou une paire d'ensembles de documents. Le résumé généré est donc composé de deux "blocs", chacun résumant un seul document (ou un seul groupe de documents). De plus, les résumés doivent être composés de phrases qui transmettent des informations à la fois comparatives et représentatives de chaque document ou ensemble de documents.

Nous présentons ici une nouvelle méthode de résumé comparatif fondée sur les récentes avancées en sémantique computationnelle : les plongements de mots et la *Word Mover Distance* (WMD) (Kusner *et al.*, 2015) afin de détecter les phrases sémantiquement proches pour la comparaison. Les résultats expérimentaux démontrent l'efficacité de ce modèle en termes de temps d'exécution ainsi que de qualité d'identification et de résumé des comparaisons.

Dans un premier temps, nous présentons les travaux associés à la tâche du résumé comparatif. Nous expliquons ensuite notre méthode de résumé comparatif, ainsi que les méthodes d'évaluation de la comparativité et de l'informativité. Nous présentons alors le contexte de notre expérience ainsi que les résultats obtenus, puis les conclusions et suites que nous envisageons à cette étude.

2 Définition

Wang *et al.* (2009) sont les premiers à avoir défini le résumé comparatif. Ils définissent le résumé comparatif comme suit : *"Étant donnée une collection de groupes de documents, le résumé de comparaison consiste à générer un court résumé composé des différences entre ces documents en extrayant les phrases les plus discriminantes dans chaque groupe de documents."*. En n'extrayant que des phrases discriminantes, le résumé comparatif s'écarte de la définition d'une comparaison car il y manquera sûrement les aspects comparatifs. Il se rapproche en revanche de la tâche traditionnelle de résumé automatique, si ce n'est qu'au lieu d'extraire les informations les plus discriminantes d'un document ou groupe de documents, on extrait les informations les plus discriminantes d'un groupe de documents vis-à-vis des autres. En effet, l'extraction des seules phrases discriminantes sans tenir compte d'une thématique commune implique que le résumé comparatif perd alors sa cohérence et son sens. Wang *et al.* (2009) fournissent un exemple de phrases discriminantes en utilisant leur méthode *Discriminative Sentence Selection*. Le tableau 1 présente un résumé automatique tiré de l'article qui présente ce travail. Bien que l'on ne dispose pas des articles d'origine, on peut constater que cet exemple constitue difficilement un résumé. Il est même compliqué de comprendre quelle en est l'information principale. De plus, les phrases ainsi extraites de leur contexte sont difficilement interprétables, et l'intérêt de cette méthode vis-à-vis d'un utilisateur final est donc discutable.

Huang *et al.* (2014) étendent la définition précédente en déclarant que : *"Un résumé comparatif doit contenir une combinaison de deux composants (provenant de différents ensembles de documents), chacun lié au même sujet."*. Huang *et al.* (2014) fournissent un exemple de résumé de référence (écrit

ID	Discriminative sentences
1	There is no cold war, there is no Saddam. Lebanon has also changed.
2	If hiring rap sheet-free intelligent people means they won't hire a black applicant for another five years.
3	He should drop the case against the lacrosse players but not the sexual assault case itself.
4	Rahman, who is about 41 years old, converted from islam to christianity over 16 years ago.
5	To be totally honest with you, we believed that there may have been a classified annex.
6	I suspect that his position reflects conventional wisdom among the Chinese military establishment.
7	In both the short and long term what those displaced by hurricane Katrina need most is money.

TABLE 1 – Extraction de phrases discriminantes par Wang *et al.* (2009).

par un humain) présenté dans le tableau 2, qui compare la coupe du monde de football 2006 à celle de 2010. Les phrases qui traitent du même thème ne sont pas mises en vis-à-vis. Toutefois, les thèmes sont ordonnés de la même manière au sein des deux résumés, et on peut dès lors identifier les thèmes principaux mis en avant par le rédacteur. Le rapprochement entre ce type de résumé comparatif et le résumé guidé par les aspects est évident. La lecture de ces résumés apporte des éléments concrets de compréhension et fait bien ressortir les spécificités des jeux de documents source.

Nous voyons donc le résumé comparatif de la même manière que Huang *et al.* (2014), à savoir la mise en parallèle de phrases issues de deux ensembles différents afin de faire apparaître les similarités thématiques et les différences des informations à l'intérieur de chaque thème qui font la spécificité de chaque jeu de documents.

Ensemble de documents A	Ensemble de documents B
Italy claimed a fourth world title in a penalty shoot-out victory over France after the two sides finished a goal apiece following extra-time in Berlin's Olympic Stadium on Sunday	Spain have won the 2010 FIFA World Cup South Africa final, defeating Netherlands 1-0 with a wonderful goal from Andres Iniesta deep into extra-time.
France captain Zinedine Zidane won the Golden Ball award for the tournament's best player	Uruguay star striker Diego Forlan won the Golden Ball Award as he was named the best player of the tournament at the FIFA World Cup 2010 in South Africa
Lukas Podolski was named the inaugural Gillette Best Young Player by FIFA's TSG after scoring three goals and contributing boundless energy to Germany's enthralling FIFA World Cup campaign	German youngster Thomas Mueller got double delight after his side finished third in the tournament as he was named Young Player of the World Cup by the FIFA Technical Study Group (TSG) and he also won the Golden Boot Award for the tournament's top-scorer.
Germany striker Miroslav Klose was the Golden Shoe winner for the tournament's leading scorer	
Germany's minister of economics and technology, Michael Glos, says he is confident the World Cup will boost the economy.	The net economic benefit from hosting the World Cup for South Africa, in terms of current and future tourism impact, is unclear
An average of 52,500 fans packed into the 12 stadiums for the 64 matches	South Africa will have five brand new state of the art football stadiums that seat an average of 50,400 spectators and five newly renovated stadiums that seat an average of 53,300
In Berlin, for example, police estimated that up to one million fans converged on the official Fan Fest public viewing venue in front of the Brandenburger Tor on Saturday to watch the host nation beat Sweden for a quarterfinal berth	In Berlin, about 3,50,000 people watched Germany at the FIFA fan fest on Wednesday night, while 56,836 people attended the fan fest in Durban
Television audiences for the 2006 FIFA World cupTM in Germany are being collated as the tournament progresses and it already looks as if they are heading for the record books	A global TV audience of more than 700 million watched Sunday's World Cup final, according to the tournament's organizers

TABLE 2 – Exemple d'un résumé de référence issu de Huang *et al.* (2014).

3 Travaux associés

Le résumé comparatif est étudié depuis de nombreuses années en linguistique et de nombreux travaux ont étudié la connotation, l'extension, les formes et les usages des comparaisons (Kennedy, 2007; Lerner & Pinkal, 2003). L'analyse comparative a été appliquée dans de nombreux domaines, et plusieurs sujets académiques connexes ont émergé, tels que la littérature linguistique (Anttila, 1972), la littérature (Prawer, 1973), l'histoire et la politique comparées. L'analyse comparative est également largement utilisée dans les applications web. De nombreux systèmes de commerce électronique fournissent des comparaisons de produits de base sur les prix et les fonctionnalités en se basant sur les données structurelles sous-jacentes.

Plus récemment, l'extraction d'informations comparatives à partir de données non structurées a attiré beaucoup d'attention. Plusieurs chercheurs proposent d'étudier la comparaison d'entités en utilisant des modèles linguistiques (Bao *et al.*, 2008) ou des mesures de similarités entre distributions de probabilités (Jain & Pantel, 2011; Liu *et al.*, 2007). Certains travaux tentent d'identifier des comparaisons linguistiques explicites et d'en extraire les éléments de comparaison (Jindal & Liu, 2006), d'autres cherchent à extraire des caractéristiques individuelles dans les phrases et à les faire correspondre (Kim & Zhai, 2009; Paul *et al.*, 2010; Zhai *et al.*, 2004).

D'autres encore utilisent des méthodes de résumé automatique conventionnelles telles que LSA, LDA, méthodes à base de graphe, ILP... L'intérêt des méthodes à base de LSA et LDA (Campr & Ježek, 2013) est la représentation des documents sous la forme de distributions de probabilités sur des thématiques. Les méthodes à base de graphes (Wan *et al.*, 2011) et de programmation linéaire en nombres entiers (Huang *et al.*, 2014) permettent de conserver simplement l'informativité présente dans les ensembles de documents, comme dans le cadre du résumé automatique classique.

Si la plupart de ces études se concentrent sur la comparaison des aspects communs au sein de phrases, il existe également certaines recherches axées sur la détection des informations uniques des sujets qui composent les documents (Wang *et al.*, 2012) ou de la nouveauté de ces informations.

La méthode utilisée par (Huang *et al.*, 2014) étant la plus proche de nos travaux, nous l'utiliserons comme baseline de comparaison dans nos expériences. Afin de générer les résumés, les auteurs utilisent à la fois la comparabilité, la centralité et la non-redondance dans une fonction objectif dont ils cherchent le maximum grâce à la programmation linéaire en nombres entiers (ILP). Leur méthode fait appel à une ressource sémantique exogène : WordNet (Pedersen *et al.*, 2004) afin de capturer les similarités entre des paires de concepts (unigrammes ou bigrammes) et ainsi déterminer le poids "comparatif" d'un ensemble de paires de concepts. Une telle méthode possède deux inconvénients majeurs :

— son coût de calcul : elle nécessite le calcul de similarité entre concepts dans un arbre et ce pour tous les concepts identifiés dans le corpus à résumer ;
— sa portabilité vers d'autres langues, bien que des ressources telles que BabelNet (Navigli & Ponzetto, 2012) répondent en partie à ce problème.

4 Méthode proposée

Dans notre étude, nous nous intéressons au résumé comparatif tel que défini dans (Huang *et al.*, 2011, 2014). Les auteurs tentent d'identifier les phrases comparables entre deux ensembles documents, i.e.

qui partagent un thème commun, tout en véhiculant une information différente. Il faut donc extraire des paires de phrases comparables depuis les deux ensembles de documents mais également identifier parmi ces paires comparables quelles sont celles qui véhiculent un thème important. Il faut en effet éviter de présenter au lecteur du résumé des informations non essentielles. La difficulté inhérente à la tâche consiste donc à trouver un compromis entre la comparabilité des phrases à extraire dans le résumé et l'aspect central des informatios qu'elles contiennent.

Nous construisons un résumé comparatif en extrayant des phrases sémantiquement similaires. Afin d'identifier les phrases similaires, nous utilisons une mesure de distance sémantique destinée aux plongements de mots : la *Word Mover Distance (WMD)* (Kusner *et al.*, 2015). La distance WMD mesure la dissimilarité entre deux documents (dans notre cas, des phrases) comme la distance minimale que les plongements de mots d'un document doivent "parcourir" pour atteindre les plongements de mots d'un autre document. Le calcul de cette distance consiste donc en une résolution d'un problème de transport.

Nous déterminons la comparabilité au niveau de la phrase afin d'améliorer le temps de traitement, contrairement à (Huang *et al.*, 2014) qui travaillent au niveau des concepts. En effet, utiliser les concepts est problématique : Les jeux de documents sont constitués d'entre trois et cinq milles concepts différents. Pour chaque résumé, il serait donc nécessaire de calculer en moyenne 15 millions de similarités sémantiques entre paires de concepts. En travaillant au niveau des phrases, nous limitons le nombre de paires de phrases à 100 000, ce qui semble plus réaliste d'un point de vue calculatoire.

Il nous faut également une mesure de centralité des phrases, c'est-à-dire de l'importance de l'information véhiculée par les phrases vis-à-vis du jeu de documents auquel elles appartiennent. Pour cette première approche du problème, et bien que d'autres mesures plus performantes existent (Gambhir & Gupta, 2017) nous avons choisi d'utiliser, la somme du tf.idf des concepts d'une phrase afin d'évaluer sa centralité.

Ces deux scores de comparativité (entre des paires de phrases) et de centralité (pour une phrase) sont normalisés puis combinés. Cette combinaison des deux scores constitue alors une fonction objectif destinée à représenter le compromis entre comparativité et centralité, et qui peut être intégrée à un algorithme d'optimisation afin de déterminer le résumé qui la maximise.

La fonction objectif utilisée est la suivante :

$$obj(sum) = (1 - \lambda)Score_{info}(sum) + \lambda Score_{comp}(sum) \tag{1}$$

avec $Score_{info}$ la fonction représentant l'informativité du résumé et $Score_{comp}$ la fonction représentant la comparativité du résumé.

Le score d'informativité du résumé est calculé simplement à l'aide d'une somme sur les TF-IDF des concepts présents :

$$Score_{info}(R) = \sum_{i=1}^{2} \sum_{k=1}^{|C_i|} w_{ij} \cdot présent(c_{ij}, R) \tag{2}$$

avec $C_i = \{c_{ij}\}$ l'ensemble des concepts présents dans l'ensemble de documents D_i. Un concept est défini comme un unigramme ou un bigramme. $présent(c_{ij}, R)$ représente la présence ou non du concept c_{ij} dans le résumé R. Chaque concept c_{ij} a un poids $w_{ij} \in \mathbb{R}$ calculé comme le score TF-IDF du concept sur l'ensemble de documents D_i :

$$w_{ij} = TF(c_{ij}, D_i) \cdot IDF(c_{ij}) \tag{3}$$

Le score de comparativité est calculé comme la somme des poids comparatifs présents dans le résumé. Un poids comparatif est le poids associé à la comparaison entre deux phrases issues des deux différents jeux de documents. Il se calcule comme la somme des similarités normalisées entre les phrases appartenant au résumé A et les phrases appartenant au résumé B, calculées avec la distance WMD.

$$Score_{comp}(R) = \sum_{s_{1j} \in R} \sum_{s_{2k} \in R} \frac{sim_{WMD}(s_{1j}, s_{2k})}{\max_{\forall m,n}(sim_{WMD}(s_{1m}, s_2 n))} \tag{4}$$

avec s_{ij} la phrase j de l'ensemble de documents i, R le résumé (composé du résumé de chacun des deux jeux de documents) et sim_{WMD} la similarité WMD calculée à partir de la *Word Mover Distance*. Une paire de phrases $< s_{1j}, s_{2k} >$ a ainsi un poids considéré comme égal à la distance WMD normalisée qui les sépare. Ce poids indique si la paire de phrases est porteuse d'une comparaison importante.

Nous appliquons en dernier lieu un algorithme d'extraction destiné à la résolution du problème du sac à dos (Pisinger *et al.*, 2007), qui a déjà été utilisé avec succès pour le résumé automatique (McDonald, 2007) avec la fonction objectif définie en équation 1.

5 Expérience

Dans cette section, nous décrivons l'expérience mise en place afin d'évaluer notre méthode. Nous y présentons notre corpus, la chaîne de traitement de notre méthode à des fins de reproductibilité, la *baseline* utilisée ainsi que la mesure utilisée pour l'évaluation. En dernier lieu, nous présentons les paramètres expérimentaux et les résultats de l'expérience.

5.1 Corpus

En raison de la nouveauté de la tâche du résumé comparatif, il n'existe pas de jeu de données disponible pour son évaluation. Huang *et al.* (2014); Campr & Ježek (2013) ont chacun créé un jeu de données de dix corpus, chaque corpus étant composé d'une paire d'ensembles de dix documents sur des sujets comparables. Sur le même principe illustré en figure 1, nous avons donc créé notre propre ensemble de corpus en anglais.

Nous avons d'abord choisi dix paires de sujets comparables, présentées en figure 3. Puis nous avons récupéré les dix premiers articles de presse liés à chaque sujet en utilisant un moteur de recherche populaire. Enfin, nous avons créé manuellement le résumé humain pour chaque paire de sujets. Le résumé humain a été créé en fonction d'instructions souples : "les deux résumés doivent contenir uniquement les informations comparables issues des deux jeux de documents, et sont chacun limités à 100 mots. Informations comparables : informations participant à un même thème à l'intérieur des jeux de documents.". Il est important de noter que chaque résumé de référence contient également deux "blocs" limités à 100 mots, chacun d'entre eux comportant les phrases de l'ensemble de documents A ou B. Un résumé se compose donc d'un maximum de 200 mots.

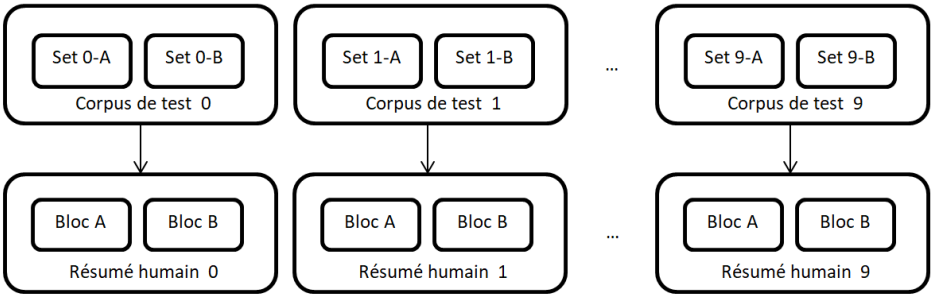


FIGURE 1 – Schéma illustrant la création du corpus.

5.2 Chaîne de traitement

La chaîne de traitement de notre méthode pour cette expérience est présentée dans la figure 2. Elle consiste tout d'abord dans les traitements classiques : tokenisation et segmentation en phrases, puis en une mise à jour, sur le corpus de résumé, des plongements de mots calculés au préalable sur un corpus issu de Wikipedia. Cette mise à jour spécifique au corpus d'évaluation vise à ajouter aux plongements de mots préexistants le vocabulaire spécifique qui n'y serait pas présent. Avoir une liste exhaustive des plongements de mots est en effet un prérequis nécessaire pour que la *Word Mover Distance* puisse délivrer des résultats satisfaisants.

Ces pré-traitements sont suivis des différents traitements visant à mettre en place la fonction objectif, suite à quoi l'algorithme d'optimisation de type résolution du problème du sac à dos décrit dans (Pisinger *et al.*, 2007) est lancé. Notre méthode, décrite en §4, est nommée SenWE-KP.

ID	Thématique	Sujets
1	Catastrophe naturelle	Tremblement de terre Haïti / Chili
2	Terrorisme	Attaque Paris / Nice
3	Politique	Élection présidentielle US / France
4	Politique	Mariage homosexuel US / France
5	Politique	Candidature Paris JO 2024 / 2012
6	Catastrophe naturelle	Feu Forêt Portugal / USA
7	Catastrophe naturelle	Inondation France / USA
8	Catastrophe naturelle	Tsunami Fukushima / 2004 Océan Indien
9	Catastrophe naturelle	Ouragan Irma / Katrina
10	Scandale	Affaire DSK / Affaire Weinstein

TABLE 3 – Paires de sujets comparables et leur thématique constituant l'ensemble de corpus d'évaluation.

5.3 Baseline

Nous utilisons notre implémentation de la méthode de Huang *et al.* (2014) comme méthode de référence (WordNet-ILP). Malheureusement, les auteurs n'ont pas fourni le code correspondant. Nous espérons notre implémentation la plus fidèle possible à leur méthode décrite en §3 ; elle associe une similarité sémantique entre concepts issue de WordNet avec une mesure de centralité pour obtenir une

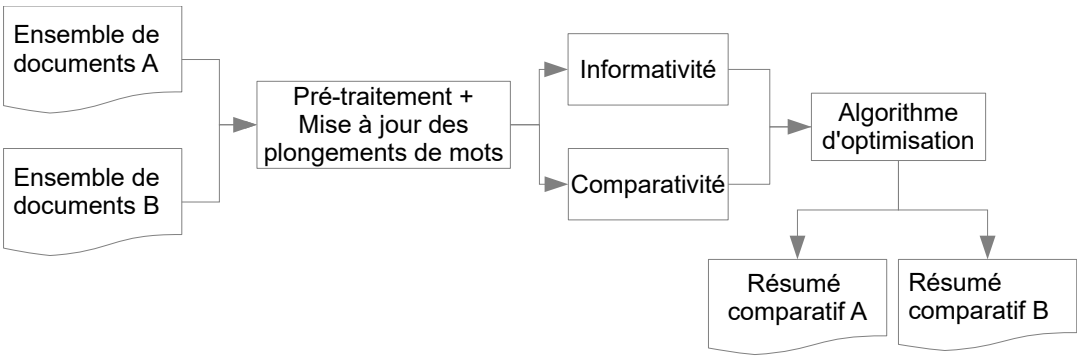


FIGURE 2 – Chaîne de traitement pour le résumé comparatif.

fonction objectif. Cette fonction objectif est maximisée grâce à un solveur de programmation linéaire en nombres entiers, guidé par des contraintes similaires à celles décrites dans (Gillick & Favre, 2009) qui interdisent toute redondance de concepts au sein du résumé généré.

5.4 Évaluation

L'évaluation d'un résumé reste toujours problématique. On peut catégoriser les méthodes d'évaluation en quatre catégories :

— automatiques sans référence (Louis & Nenkova, 2013) ;
— automatiques avec référence (ROUGE, BLEU a aussi été utilisé dans certaines études) (Lin, 2004) ;
— manuelles avec score automatique : Pyramid (Nenkova & Passonneau, 2004) ;
— manuelles entièrement subjectives.

Huang *et al.* (2014) ont proposé une méthode d'évaluation semblable à Pyramid mais adaptée au résumé comparatif. Il s'agit d'évaluer si les résumés automatiques traitent ou non des thèmes relevés dans les résumés manuels. Cependant, les évaluations manuelles étant coûteuses en temps humain, nous souhaitons les réserver pour les étapes futures de notre étude. Les évaluations automatiques sans références, quant à elles, sont inexploitables dans notre contexte. En effet, elles s'appuient sur la similarité entre le jeu de documents à résumer et les résumés à évaluer, par le biais des distributions de probabilité. L'objectif du résumé comparatif s'éloigne de celui du résumé "traditionnel" pour lequel la centralité, bien reflétée par les mesures de divergence entre distributions de probabilité, est la notion la plus importante. Nous avons donc opté pour les mesures ROUGE.

Nous évaluons les deux résumés créés A et B pour chaque ensemble de documents vis-à-vis respectivement des résumés humains A et B. Dans cette expérience, nous utilisons les paramètres ROUGE possédant la meilleur corrélation avec les scores humains d'après Graham (2015) :

— ROUGE-2 : Précision ROUGE-2 moyenne avec racinisation et suppression des mots vides [1]
— ROUGE-3 : F-Mesure ROUGE-3 moyenne sans racinisation et sans suppression des mots vides [2].

1. Arguments ROUGE-2 : *-n 2 -x -m -s -c 95 -r 1000 -f A -p 1 -t 0*
2. Arguments ROUGE-3 : *-n 3 -x -c 95 -r 1000 -f A -p 0.5 -t 0*

Résumé A (Chili) : Reuters now reports that 47 people have died in the quake. The earthquake struck at 3 :34 a.m. in central Chile, centered roughly 200 miles southwest of Santiago at a depth of 22 miles, the United States Geological Survey reported. More than 300 people were killed, according to Chile's Office of Emergency Management, and 15 are missing. A tsunami watch has been issued for Ecuador, Colombia, Panama, Costa Rica and Antarctica. No damage was expected from possibly stronger waves to follow, Ryan said. Chilean President Michele Bachelet said that altogether two million people had been affected.
Résumé B (Haïti) : The headquarters of the U. N. peacekeeping mission in Port-au-Prince collapsed, a U. N. official told CNN. "We stand ready to assist the people of Haiti," Mr. Obama said. Some said that they had been able to get through immediately after the earthquake. The World Bank forgave the country's $36 million balance in May. "I saw people under the rubble, and people killed. A "large number" of UN personnel were reported missing by the organisation. The quake was felt in the Dominican Republic, sending people in the capital Santo Domingo running on to the streets in panic.

FIGURE 3 – Résumé comparatif issu de la méthode WordNet-KP pour le sujet "Tremblement de terre Chili / Haiti 2010".

5.5 Paramètres expérimentaux

Tous les systèmes génèrent un résumé comparatif composé de deux résumés de 100 mots chacun. Toutes les expériences sont effectuées sur un processeur Intel Xeon à 2,20 GHz composé de 40 coeurs. Les résultats de l'évaluation sont présentés dans le tableau 4. Les scores présentés correspondent à la moyenne des scores ROUGE obtenus sur les deux "blocs" qui composent un résumé.

5.6 Résultats

La méthode SenWE-KP permet d'obtenir des résultats légèrement meilleurs en terme de scores Rouge-2 et Rouge-3. Étant donné la taille de notre corpus et la faible différence de score entre les deux méthodes, on peut considérer que leurs résultats ROUGE sont comparables. Notre méthode semble donc bien capturer les aspects comparatifs entre deux jeux de documents. Les figures 3 et 4 présentent les résumés obtenus pour le sujet "Tremblement de terre Chili / Haïti 2010" respectivement avec la *baseline* WordNet-ILP et notre méthode SenWE-KP.

Néanmoins, en étudiant manuellement les résumés générés, on peut observer que ceux-ci sont en général pollués par une certaine redondance. La méthode WordNet-ILP pallie ce problème en utilisant une optimisation sur une somme de concepts et de paires de concepts. Étant donné que chaque concept et chaque paire n'est compté qu'une seule fois, le meilleur résumé doit limiter la redondance. Dans notre méthode, même si nous ne comptons qu'une seule fois chaque score tf-idf dans la partie représentativité de la fonction objectif, nous ne pouvons pas réduire les similitudes aussi facilement.

Cependant, l'utilisation d'une granularité au niveau de la phrase améliore considérablement le temps de traitement, de plus d'un facteur 10, la rendant utilisable en application réelle ; la table 4 montre que le calcul des résumés pour l'ensemble des 10 sujets de notre corpus d'évaluation prend plus de 5h pour WordNet-ILP contre 22 minutes pour notre méthode. Le temps de traitement relativement long pour la méthode WordNet provient des traitements réalisés au niveau des concepts. D'une part, le calcul des distances de similarité entre concepts dans la taxonomie WordNet est coûteux. D'autre part, le solveur de programmation linéaire en nombre entiers possède un très grand nombre de contraintes,

Résumé A (Chili) : The Geological Survey said that another earthquake on Saturday, a 6.3-magnitude quake in northern Argentina, was unrelated. The magnitude-8.8 earthquake struck at 3 :34 AM at a depth of 35km. It also recorded at least eight aftershocks, the largest of 6.9 magnitude at 0801 GMT. The earthquake struck at 0634 GMT, 115km (70 miles) north-east of the city of Concepcion and 325km south-west of the capital Santiago. This was in direct contrast to Haiti, which was unprepared for the Jan. 12 earthquake. The quake was followed by 76 aftershocks of 4.9 magnitude or greater. The damage from Chile's earthquake was widespread.
Résumé B (Haïti) : The tremor hit at 1653 (2153 GMT) on Tuesday, the US Geological Survey said. A massive 7.0-magnitude earthquake has struck the Caribbean nation of Haiti. The earthquake hit at 4 :53 PM some 15 miles (25 km) southwest of the Haitian capital of Port-au-Prince. Most severely affected was Haiti, occupying the western third of the island. At least 10 aftershocks followed, including two in the magnitude 5 range, the USGS reported. The hospital in Petionville, a well to do neighbourhood, home to diplomats and expatriates, was wrecked. In addition, less than one-third of the population was steadily employed.

FIGURE 4 – Résumé comparatif issu de la méthode SenWE-KP pour le sujet "Tremblement de terre Chili / Haiti 2010".

dû à la nécessité d'encoder les similarités entre les concepts présents dans les jeux de documents différents.

Méthodes	R2	R3	Temps d'exécution
WordNet-ILP	0.08704	0.05207	19335s
SenWE-KP	0.09006	0.05410	1352s

TABLE 4 – Scores ROUGE et temps d'exécution pour chaque méthode.

6 Conclusion et perspectives

Dans cet article, nous présentons une nouvelle méthode pour le résumé comparatif de deux jeux de documents traitant de sujets comparables. Nous utilisons les plongements de mots et la mesure de distance sémantique *Word Mover Distance* au niveau de la phrase afin de découvrir celles les plus susceptibles d'être comparatives. Nous comparons notre méthode à une méthode de l'état de l'art décrite par Huang *et al.* (2014), qui utilise WordNet pour repérer les phrases comparatives. Nos résultats confirment que l'utilisation des plongements de mots et de la *Word Mover Distance* au niveau de la phrase permet d'obtenir des résultats comparables à l'utilisation de similarités dérivées de WordNet. Nous améliorons en revanche considérablement le temps de traitement par rapport à la méthode de Huang *et al.* (2014).

Nous travaillons à l'extension de nos baselines, en y ajoutant la méthode de Campr & Ježek (2013). Une évaluation manuelle est également prévue afin de confirmer les résultats obtenus avec la méthode ROUGE.

Dans la suite de nos recherches, nous devons travailler à l'élimination de la redondance au sein du résumé comparatif, par exemple en supprimant les paires de phrases trop similaires à des phrases déjà présentes dans le résumé ou en intégrant une mesure de la redondance dans notre fonction objectif. Nous avons également constaté l'importance des entités nommées dans la détection des phrases

comparables. En effet, elles sont souvent l'élément de différenciation au sein de phrases qui traitent du même thème. Nous souhaitons donc incorporer des traitements sur les entités nommées afin de mieux identifier les phrases comparables.

Les avancées dans la recherche sur les plongements de mots permettent aujourd'hui d'obtenir des plongements de mots multilingues (Conneau *et al.*, 2017). Nous souhaitons appliquer notre méthode au résumé comparatif multilingue ou translingue, donc en travaillant sur des jeux de documents dans des langues différentes. Cela permettrait d'étendre la portée de notre travail. L'utilisation de plongements de mots multilingues nous permettrait d'appliquer directement notre méthode au résumé comparatif multilingue ou crosslingue.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-16-CE38-0008 (projet ANR JCJC ASADERA).

Références

ANTTILA R. (1972). *An introduction to historical and comparative linguistics*. Macmillan New York.

BAO S., LI R., YU Y. & CAO Y. (2008). Competitor mining with the web. *IEEE Transactions on Knowledge and Data Engineering*, **20**(10), 1297–1310.

CAMPR M. & JEŽEK K. (2013). Topic models for comparative summarization. In *International Conference on Text, Speech and Dialogue*, p. 568–574 : Springer.

CONNEAU A., LAMPLE G., RANZATO M., DENOYER L. & JÉGOU H. (2017). Word translation without parallel data. In *International Conference on Learning Representations*. arXiv preprint : 1710.04087.

GAMBHIR M. & GUPTA V. (2017). Recent automatic text summarization techniques : A survey. *Artif. Intell. Rev.*, **47**(1), 1–66. DOI : 10.1007/s10462-016-9475-9.

GILLICK D. & FAVRE B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, p. 10–18, Stroudsburg, PA, USA : Association for Computational Linguistics.

GRAHAM Y. (2015). Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, p. 128–137.

HUANG X., WAN X. & XIAO J. (2011). Comparative news summarization using linear programming. In *Proceedings of the 49th Annual Meeting of the ACL*, p. 648–653.

HUANG X., WAN X. & XIAO J. (2014). Comparative news summarization using concept-based optimization. *Knowledge and information systems*, **38**, 691–716.

JAIN A. & PANTEL P. (2011). How do they compare ? automatic identification of comparable entities on the web. In *2011 IEEE International Conference on Information Reuse & Integration*, p. 228–233 : IEEE.

JINDAL N. & LIU B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 244–251 : ACM.

KENNEDY C. (2007). Modes of comparison. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, volume 43.1, p. 141–165 : Chicago Linguistic Society.

KIM H. D. & ZHAI C. (2009). Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, p. 385–394 : ACM.

KUSNER M., SUN Y., KOLKIN N. & WEINBERGER K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning*, p. 957–966.

LERNER J.-Y. & PINKAL M. (2003). Comparatives and nested quantification. In J. GUTIÉRREZ-REXACH, Éd., *Semantics : critical concepts in linguistics. Vol. V : Operators and sentence types*, chapitre 68, p. 70–87. Routledge.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics. Anthologie ACL : W04-1013.

LIU J., WAGNER E. & BIRNBAUM L. (2007). Compare&contrast : using the web to discover comparable cases for news stories. In *Proceedings of the 16th international conference on World Wide Web*, p. 541–550.

LOUIS A. & NENKOVA A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, **39**(2), 267–300. DOI : 10.1162/COLI_a_00123.

MCDONALD R. (2007). A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, p. 557–564 : Springer.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

NENKOVA A. & PASSONNEAU R. (2004). Evaluating content selection in summarization : The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics : HLT-NAACL 2004*, p. 145–152, Boston, Massachusetts, USA : Association for Computational Linguistics.

PAUL M. J., ZHAI C. & GIRJU R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, p. 66–76 : Association for Computational Linguistics.

PEDERSEN T., PATWARDHAN S. & MICHELIZZI J. (2004). Wordnet : : Similarity : measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, p. 38–41 : Association for Computational Linguistics.

PISINGER D., RASMUSSEN A. & SANDVIK R. (2007). Solution of large-sized quadratic knapsack problems through aggressive reduction. *I N F O R M S Journal on Computing*, **19**(2), 280–290. DOI : 10.1287/ijoc.1050.0172.

PRAWER S. S. (1973). *Comparative literary studies : an introduction*. Duckworth London.

WAN X., JIA H., HUANG S. & XIAO J. (2011). Summarizing the differences in multilingual news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, p. 735–744 : ACM.

WANG D., ZHU S., LI T. & GONG Y. (2009). Comparative document summarization via discriminative sentence selection. In *Proceedings of the 18th ACM conference on Information and knowledge management*, p. 1963–1966.

WANG D., ZHU S., LI T. & GONG Y. (2012). Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **6**, 12.

ZHAI C., VELIVELLI A. & YU B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 743–748 : ACM.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Réseaux de neurones pour la résolution d'analogies entre phrases en traduction automatique par l'exemple

Valentin Taillandier[1] Liyan Wang[2] Yves Lepage[2]
(1) École Normale Supérieure de Lyon, 46 allée d'Italie, 69007 Lyon, France
(2) Université Waseda, 2-7 Hibikino, Wakamatu, 808-0135 Kitakyûsyû, Japon
valentin.taillandier@ens-lyon.fr, wangliyan0905@toki.waseda.jp,
yves.lepage@waseda.jp

RÉSUMÉ

Cet article propose un modèle de réseau de neurones pour la résolution d'équations analogiques au niveau sémantique et entre phrases dans le cadre de la traduction automatique par l'exemple. Son originalité réside dans le fait qu'il fusionne les deux approches, directe et indirecte, de la traduction par l'exemple.

ABSTRACT

Neural networks for the resolution of analogies between sentences in EBMT

We introduce a neural network architecture for the resolution of semantic analogies between sentences for the purpose of example-based machine translation. Our proposal merges the direct and indirect approaches in example-based machine translation.

MOTS-CLÉS : Analogie, traduction par l'exemple, réseaux de neurones.

KEYWORDS: Analogy, example-based machine translation, neural networks.

1 Introduction

1.1 Approche directe en traduction par l'exemple

Dans l'approche directe de traduction automatique par l'exemple (Nagao, 1984), étant donnée une phrase à traduire et un couple constitué d'une phrase exemple et de sa traduction, les similarités et les différences entre la phrase à traduire et la phrase exemple sont identifiées, puis transférées en langue cible en se fondant sur des connaissances symboliques ou autres.

Formellement, soit une phrase D dans une langue source $\mathcal{L}$ à traduire dans une langue cible $\mathcal{L}'$, l'approche directe consiste à chercher un couple de phrases (A, A') dans un bi-corpus donné afin de produire une phrase D', proposée comme traduction candidate, qui soit solution de l'équation analogique bilingue $A : A' :: D : D'$ (voir figure 1).

Volume 2 : Traitement Automatique des Langues Naturelles, pages 108–121.
HAL : hal-02784759.

he 's coming .	:	*il est en train d' arriver .*	::	*i am eating an apple .*	:	??

FIGURE 1 – Approche directe en traduction automatique par analogie. Une phrase en langue source correspond à une phrase en langue cible. De la même manière, à quelle phrase en langue cible correspond une nouvelle phrase en langue source ?

he 's coming .	:	*i am coming .*	::	*he 's eating an apple .*	:	*i am eating an apple .*
il est en train d' arriver .	:	*j' arrive .*	::	*il est en train de manger une pomme .*	:	??

FIGURE 2 – Approche indirecte en traduction automatique par analogie. Une phrase en langue source (en haut à droite) est en relation d'analogie avec trois autres phrases en langue source (en haut). La traduction en langue cible de ces trois phrases étant connue (en bas), quelle est la phrase en langue cible en relation d'analogie avec elles, que l'on peut supposer traduction de la première phrase ?

1.2 Approche indirecte en traduction par l'exemple

Le problème de l'approche directe est la nécessaire expression explicite du transfert des différences par traduction. Afin de remédier à ce problème, l'approche indirecte en traduction automatique par l'exemple (Lepage & Denoual, 2005; Langlais *et coll.*, 2008; Dandapat *et coll.*, 2010) ne cherche pas un seul couple de phrases en traduction, mais trois. La traduction s'effectue en deux étapes : si l'analogie entre les quatre phrases en langue source tient, alors l'analogie en langue cible est tentée.

Formellement, soit la phrase D à traduire. On explore des triplets de couples de phrases en relation de traduction $((A, A'), (B, B'), (C, C'))$. Si l'analogie monolingue en langue source $A : B :: C : D$ est vérifiée, on peut essayer la résolution monolingue en langue cible de l'analogie $A' : B' :: C' : D'$ d'inconnue D' (voir figure 2). Différents algorithmes et méthodes de résolution d'analogies monolingues ont été proposés, soit par approche symbolique (Lepage, 1998; Langlais *et coll.*, 2009; Lepage, 2017; Rhouma & Langlais, 2018; Rhouma, 2018), soit par apprentissage automatique (Kaveeta & Lepage, 2016).

L'approche indirecte suppose d'avoir accès à trois phrases dont la traduction est connue et formant une analogie avec la phrase à traduire. La partie en haut à gauche de la figure 3 montre un exemple de phrase à traduire. La figure est séparée en deux : à gauche, les phrases en langue source ; à droite, les phrases en langue cible. À gauche, trois phrases ont été trouvées pour former une analogie monolingue en langue source avec la phrase à traduire. Les traductions de ces trois phrases sont supposées connues. Elles peuvent provenir d'une mémoire de traduction, d'un corpus bilingue ou de précédentes épreuves de traduction. Elles sont représentées à droite de la figure. Résoudre l'analogie en langue cible (celle qui est représentée à droite du plan) permet d'obtenir la traduction souhaitée.

1.3 Fusion des approches directe et indirecte

L'approche directe peut être fusionnée avec l'approche indirecte. En effet la phase d'extraction de l'approche indirecte fait apparaître quatre nouvelles analogies bilingues dont deux pourraient permettre la résolution, par approche directe, de la traduction souhaitée (voir la figure 3, en haut à droite).

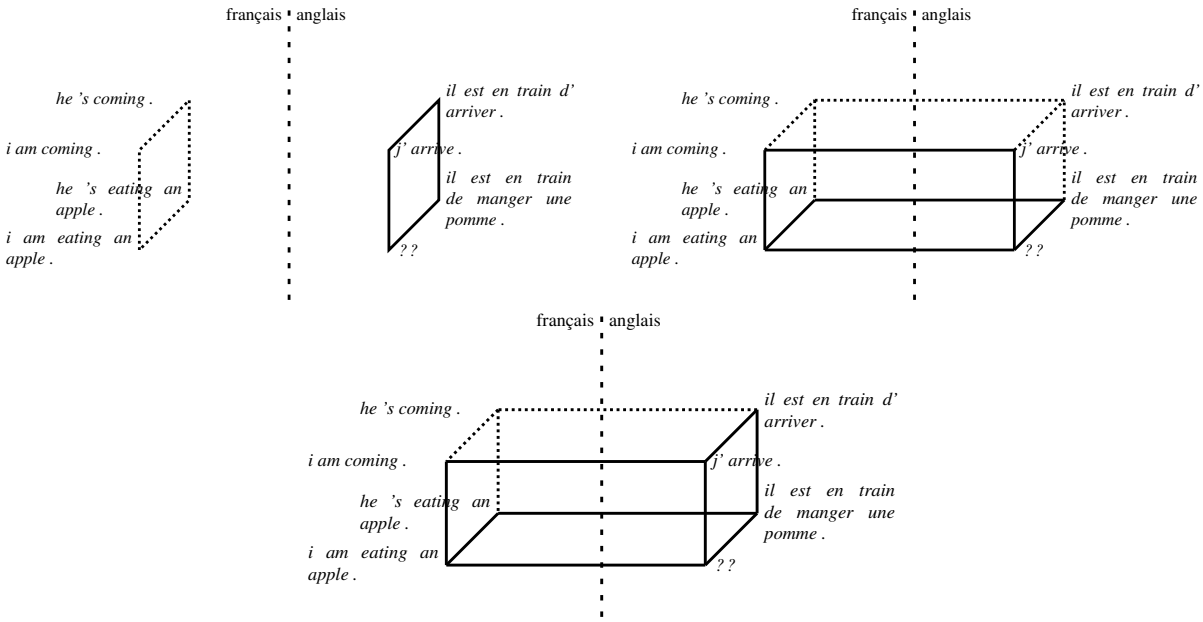


FIGURE 3 – En haut à gauche, l'approche indirecte consiste à résoudre la seule analogie monolingue, visualisée par des traits pleins. Mais on voit en haut à droite que deux analogies bilingues, données en traits pleins, peuvent aussi être utilisées. La fusion des deux approches directe et indirecte, visualisée en bas de la figure, consiste à utiliser ces trois analogies pour produire la traduction souhaitée.

1.4 Contribution

La contribution essentielle de cet article est de montrer comment utiliser les deux analogies bilingues de l'approche directe en complément de la seule analogie monolingue de l'approche indirecte. Autrement dit, il s'agit de se fonder sur trois analogies pour obtenir une traduction.

Le premier pas dans cette direction consiste à utiliser la notion d'appariement aussi bien monolingue que bilingue. Cette notion est en effet unificatrice : elle permet de prendre en considération les trois analogies mentionnées plus haut, grâce aux techniques d'appariement issues par exemple de la traduction automatique statistique.

Le second pas consiste à ne pas se limiter à l'appariement mais à élargir à la notion de matrices de correspondances entre phrases. Il s'agit de passer d'une vue binaire, celle qui existait dans les approches en traduction automatique statistique, à une vue plus souple, grâce, par exemple, aux représentations vectorielles des mots. Ces représentations permettent en effet de calculer des valeurs réelles comme mesure de similarité ou de distance entre les mots de deux phrases, cela aussi bien de façon monolingue que bilingue, grâce, par exemple, aux représentations vectorielles multilingues de mots.

2 Matrices d'appariement mot à mot

Nous reformulons maintenant le problème en redonnant les notations nécessaires. Nous donnons aussi des informations plus précises sur les notions de similarité utilisées dans notre proposition.

Soit une phrase D dans une langue source $\mathcal{L}$. Le problème de la traduction de D dans une langue

cible $\mathcal{L}'$ consiste à *rechercher* dans un corpus bilingue trois phrases A, B et C en langue source $\mathcal{L}$ qui soient telles que $A : B :: C : D$, (à ce sujet, voir p. ex., (Langlais, 2016)) puis à *réutiliser* leurs traductions en langue $\mathcal{L}'$, A', B' et C', extraits de ce corpus bilingue, en les *adaptant* pour trouver une phrase D' qui soit telle que :

$$\begin{cases} A' : B' :: C' : D' \\ B : B' :: D : D' \\ C : C' :: D : D' \end{cases} \tag{1}$$

La phrase D' est alors proposée comme traduction de D.

Si on ajoute le fait que la phrase D et sa traduction D' peuvent être *ajoutées* au corpus bilingue, on aura alors reconnu, à la suite de (Collins & Somers, 2003), les principales étapes du raisonnement à partir de cas (Aamodt & Plaza, 1994) : la recherche de cas connus (angl. : *retrieve*), leur réutilisation (angl. : *reuse*), leur adaptation (angl. : *revise*) et la mémorisation du nouveau cas créé (angl. : *retain*).

Nous ne nous intéressons pas ici à la recherche de cas dans le corpus bilingue. Nous partons de quadruplets de phrases donnés (A, A'), (B, B'), (C, C') et (D, D'). En retirant l'une des phrases en langue cible, nous créons une instance du problème. La phrase retirée servira de référence lors de l'évaluation de la traduction obtenue.

Notre structure de données de base est celle de matrices d'appariement. Pour un couple de phrases, chacune vue comme une chaîne de mots, une matrice d'appariement mot à mot est composée de cases contenant chacune une valeur réelle reflétant la proximité entre les mots des deux phrases. Pour chacun des deux cas, monolingue et bilingue, nous utilisons une mesure de proximité différente.

Pour le cas monolingue, nous exploiterons des plongements lexicaux. Nous utiliserons la similarité couramment utilisée qui repose sur le calcul du cosinus entre les vecteurs représentant les mots. Pour deux mots m_1 et m_2, on posera donc :

$$\text{sim}(m_1, m_2) = \cos(\overrightarrow{m_1}, \overrightarrow{m_2}) \tag{2}$$

Pour le cas bilingue, nous exploiterons les probabilités de traduction entre mots. On peut obtenir de telles probabilités, conditionnelles, à partir d'une table de traduction. Nous utiliserons la moyenne géométrique de telles probabilités conditionnelles. Pour deux mots m et m', on posera donc :

$$\text{sim}(m, m') = \sqrt{p(m|m') \times p(m'|m)} \tag{3}$$

Dans le cas où les deux mots n'apparaissent pas comme traduction possibles l'un de l'autre dans la table de traduction, la valeur est évidemment zéro.

Les deux mesures de proximité données ci-dessus ont évidemment la propriété de symétrie. Dans le cas où les deux mots sont égaux (cas monolingue) ou traduction l'un de l'autre sans variante possible (cas bilingue), la valeur de la proximité est 1.

On remarque bien sûr que l'utilisation de plongements lexicaux bilingues permet de se dispenser de table de traduction. De tels plongements bilingues permettent une vue unifiée. La formule (2) peut alors être utilisée dans le cas bilingue comme monolingue.

En toute généralité, étant données deux phrases X et Y n'appartenant pas nécessairement à la même langue, il est toujours possible de construire la matrice $\mathcal{M}_{X,Y}$ dont les cases portent les valeurs des similarité entre les mots correspondant aux indices dans les phrases. La figure 4 donnent deux exemples de telles matrices. À gauche, entre deux phrases appartenant à la même langue ; à droite,

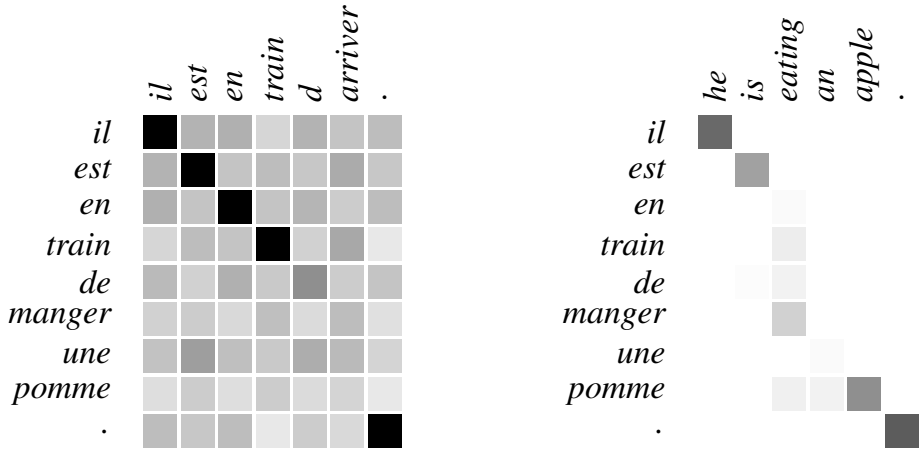


FIGURE 4 – Exemples de matrices d'appariement. À gauche, entre deux phrases françaises du corpus Tatoeba, avec des proximités calculées à l'aide de la formule 3. À droite, entre deux phrases traduction l'une de l'autre, issue de la partie anglais-français du même corpus, avec des proximités calculées à l'aide de la formule (3). La noirceur des cases reflète la proximité.

entre deux phrases traduction l'une de l'autre. Pour notre problème, nous aurons affaire aux trois types de matrices suivantes :

— $\mathcal{M}_{X,Y}$ entre deux phrases X et Y de la langue source $\mathcal{L}$;
— $\mathcal{M}_{X',Y'}$ entre deux phrases X' et Y' de la langue cible $\mathcal{L}'$;
— $\mathcal{M}_{X,X'}$ entre deux phrases traduction l'une de l'autre, et appartenant respectivement à $\mathcal{L}$ et à $\mathcal{L}'$.

3 Réseaux de neurones

3.1 Architectures proposées

Comme *baseline,* nous utiliserons une architecture de réseaux de neurones composée d'une unique couche dépourvue de fonction d'activation. Il s'agit donc d'une simple application affine. Elle permet de justifier l'usage des architectures plus complexes suivantes.

La première architecture *Architecture* 1 est un perceptron multicouche à deux ou trois couches cachées de type *ReLu*. La couche de sortie est équipée de l'activation $\tanh$ pour imposer une sortie dans $[-1, 1]$. Cette architecture prend en entrée le vecteur composée de toutes les cellules de chacune des matrices d'entrée et retourne un vecteur composée de toutes les cellules des matrices de sorties. Dans cette architecture chacune des cellules d'entrée peut influer sur chacune des cellules de sorties.

La deuxième architecture *Architecture* 2 exploite la remarque suivante. Chacune des matrices de sortie est impliquée dans deux analogies. Par exemple, la matrice $\mathcal{M}_{B',D'}$ est impliquée dans les analogies $B : B' :: D : D'$ et $A' : B' :: C' : D'$. Cette matrice peut donc être calculée en utilisant les informations contenues dans les matrices $\mathcal{M}_{B,D}$, $\mathcal{M}_{B,B'}$, $\mathcal{M}_{A',B'}$ et $\mathcal{M}_{A',C'}$ (les matrices $\mathcal{M}_{C,D}$ et $\mathcal{M}_{C,C'}$ sont ici exclues du calcul de la matrice $\mathcal{M}_{B',D'}$). On utilise donc un réseau de neurones unique pour chacune des matrices de sortie à calculer. Chacun des réseaux prend en entrée les quatre

matrices impliquées dans la résolution de la matrice de sortie et adopte la même architecture que celle donnée dans le paragraphe précédent (*Architecture* 1). Chacun des trois réseaux est donc un canal isolé d'un réseau plus grand contraignant les trois matrices de sortie à partir des six matrices d'entrée.

La dernière architecture *Architecture* 3 utilise un canal séparé pour chacune des cases des matrices de sortie. Chaque canal est conçu selon le modèle précédent (*Architecture* 2) et prend en entrée les quatre matrices impliquées dans la résolution de chaque case.

Bien que le nombre de neurones sur les couches externes soit fixé par les dimensions des entrées et des sorties des réseaux, le nombre de neurones des couches cachées peut varier. Pour une architecture donnée, il est possible de moduler le nombre de neurones de chaque couche cachée et donc de faire varier le nombre de paramètres entrainables. Un nombre de paramètre élevé tend à améliorer la précision d'un réseau mais requiert plus de puissance de calcul lors de l'apprentissage et l'utilisation. Une bonne architecture devant pouvoir fournir un bon compromis entre précision et coût, la figure 6 permet de comparer la capacité des différentes architectures selon le nombre de paramètres alloués.

Toutes les architectures précédentes produisent en sortie des estimations des valeurs contenues dans les trois matrices d'appariement $\mathcal{M}_{D,D'}$, $\mathcal{M}_{B',D'}$ et $\mathcal{M}_{C',D'}$. La figure 7 donne un exemple de sortie réelle. On y distingue clairement trois parties qui correspondent aux trois matrices : en haut la matrice bilingue $\mathcal{M}_{D,D'}$ avec au dessous les deux matrices monolingues $\mathcal{M}_{B',D'}$ et $\mathcal{M}_{C',D'}$.

3.2 Décodage des matrices en phrase de la langue cible

Le décodage des matrices en sortie en une phrase en langue cible est effectué à partir des valeurs de proximité données par les matrices d'appariement en sortie. Le mot $\widehat{m'_j}$ à la position j dans D' est choisi par minimisation classique de l'erreur quadratique.

$$\widehat{m'_j} = \arg \min_{m' \in \mathcal{L}'} \sum_{X \in \{ D,\, B',\, C' \}} \sum_{i=1}^{|X|} \left(\mathcal{M}_{X,D'}[i,j] - \mathrm{sim}(X[i], m') \right)^2 \tag{4}$$

Dans la formule (4), les matrices d'appariement en sortie $\mathcal{M}_{D,D'}$, $\mathcal{M}_{B',D'}$ et $\mathcal{M}_{C',D'}$ apparaissent sour la forme $\mathcal{M}_{X,D'}$ avec X parcourant l'ensemble $\{ D,\, B',\, C' \}$. Les valeurs trouvées dans ces matrices de sortie, sont interprétées comme des valeurs de similarité. Pour un mot candidat m', elles sont comparées avec les similarités du mot m' aux mots des phrases D, B' et C' donnés dans la formule (4) par $X[i]$, avec X parcourant encore l'ensemble $\{ D,\, B',\, C' \}$. Pour un problème similaire, (Kaveeta & Lepage, 2016) utilisent le même type de minimisation de l'erreur quadratique.

Dans l'expérience rapportée plus bas, les analogies utilisées, décrites dans la section 4, sont des analogies formelles. Pour de telles analogies, on sait que les mots de D' apparaissent nécessairement soit dans B', soit dans C', soit dans les deux à la fois. Notre propos est de tester l'approche en toute généralité. C'est pourquoi la formule (4) ne fait pas une telle hypothèse : tout le vocabulaire de la langue cible est théoriquement exploré pour choisir chacun des mots $\widehat{m'_j}$.

En pratique, cependant, avec une table de traduction, nous restreignons l'espace de recherche à un ensemble de mots candidats prédéfinis, construit en trois étapes. Nous construisons d'abord un premier ensemble de mots candidats qui correspond à l'ensemble des mots de la table de traduction qui sont traduction des mots de A, B et C. Nous élargissons ensuite à un deuxième ensemble de mots candidats en ajoutant tous les mots de la langue cible qui font partie des vingt mots les plus proches

d'au moins un mot du premier ensemble. Enfin, pour construire le troisième et dernier ensemble de mots candidats, nous ajoutons au deuxième ensemble les cent mots les plus fréquents de la langue cible. Cet ajout se justifie par des expériences, non rapportées ici, qui ont montré que cela améliorait les scores de traduction.

4 Jeu de données

4.1 Corpus

Nous utilisons la partie français-anglais du corpus *Tatoeba.org* [1]. Chaque phrase a au plus dix mots. Toutes les analogie formelles de commutation (Lepage, 2003) entre phrases françaises d'une part et entre phrases anglaises d'autre part ont été extraites automatiquement grâce à une librairie dédiée au calcul des analogies (Fam & Lepage, 2018) [2]. L'intersection par traduction permet d'obtenir un jeu de quadruplets de bi-phrases $((A, A'), (B, B'), (C, C'), (D, D'))$ tels que la partie française, comme la partie anglaise, constitue une analogie. Nous avons obtenu de cette façon $327\ 461$ quadruplets de bi-phrases.

En réordonnant les termes d'un quadruplet, on augmente facilement les données : trois autres quadruplets, vérifiant aussi les analogies dans les deux langues, sont obtenues grâce aux axiomes de l'analogie [3]. À partir de $327\ 461$ quadruplets de bi-phrases, l'énumération des quatre formes équivalentes pour chaque analogie nous permet d'obtenir un jeu de données contenant $1\ 309\ 844$ quadruplets de bi-phrases ($327\ 461 \times 4 = 1\ 309\ 844$).

A des fins d'entraînement et de test, notre jeu de données a été divisé aléatoirement en trois parties distinctes :

— 60 % constitue le jeu d'entraînement ;
— 20 % sert de jeu de validation : l'apprentissage est arrêté quand les performances sur le jeu de validation n'augmentent plus ;
— les 20 % restants constituent le jeu de test.

Les statistiques sur les données sont présentées dans le tableau 1. Le jeu de test contient $261\ 969$ analogies, soit autant de phrases à traduire. Beaucoup de phrases à traduire sont répétées dans le jeu de test : il n'y a en fait que $15\ 470$ phrases anglaises distinctes. Il faut observer que certaines de ces phrases sont traduites différemment selon les analogies dans lesquelles elles interviennent. C'est ce que reflète le nombre supérieur de phrases françaises distinctes, $18\ 089$, dans le tableau 1. Un exemple de phrase anglaise à traduire de deux façons différentes est donné dans la figure 5.

Enfin, on peut constater, chose caractéristique de la ressource Tatoeba utilisée, que le vocabulaire utilisé dans notre ressource n'est pas très riche : $1\ 450$ mots différents en anglais et $2\ 533$ en français sur l'ensemble de notre jeu de données. Ces chiffres sont donnés en dernière ligne dans le tableau 2.

1. https://tatoeba.org/
2. http://lepage-lab.ips.waseda.ac.jp > Projects > Kakenhi 15K00317 > Tools – Nlg Module
3. Voir (Lepage, 2003, p. 116). Il existe huit formes équivalentes de l'analogie qui sont en fait le groupe de transformations des coins du carré connu en algèbre sous le nom de D_8. Ici, pour notre problème, quatre formes redondantes sont éliminées par échange des moyens qui affirme que $A : B :: C : D \Leftrightarrow A : C :: B : D$.

Jeu de données	Nombre d'analogies	de phrases angl.	de phrases fr.	de mots / phrase angl.	de mots / phrase fr.	de carac. / phrase angl.	de carac. / phrase fr.
entraînement	785 906	17 208	20 465	5.8±1.5	5.8±1.7	22,6±6,3	26,1±7,7
validation	261 969	15 461	18 129	5,7±1,5	5,8±1,7	22,4±6,2	25,9±7,6
test	261 969	15 470	18 089	5,7±1,5	5,8±1,7	22,4±6,2	25,9±7,6
total	1 309 844						

TABLE 1 – Statistique des données utilisées. Les nombres de phrases sont les nombres de phrases distinctes.

Jeu de données	Nombre d'analogies	Taille du vocabulaire angl.	Taille du vocabulaire fr.
entraînement	785 906	1 449	2 532
validation	261 969	1 416	2 459
test	261 969	1 428	2 463
total	1 309 844	1 450	2 533

TABLE 2 – Taille du vocabulaire pour les données utilisées. En comparant les tailles sur l'ensemble des données et celles pour chacune des parties, on observe qu'il existe un important recouvrement des vocabulaires des trois parties du jeu de données. On fait aussi la remarque classique qu'à contenu équivalent le nombre de mots distincts en français est plus important que celui de l'anglais.

4.2 Matrices d'appariement

Les matrices d'appariement sont produites automatiquement pour chaque quadruplet de bi-phrases. Elles se répartissent en deux groupes : celles qui n'impliquent pas D', utilisées en entrée du système, et celles qui impliquent D', qui servent de comparaison pour l'évaluation des sorties du système.

Pour les matrices monolingues, des modèles de plongement lexicaux pré-entrainés pour chacune des langues ont été utilisés (Bojanowski *et coll.*, 2017) [4]. Pour les matrices bilingues, les valeurs des cases ont été calculées de deux manières différentes : d'une part, avec les probabilités d'une table de traduction obtenue à partir du corpus avec l'outil *Hieralign* (Wang & Lepage, 2017) [5] ; d'autre part, à partir de l'alignement automatique des deux plongements monolingues précédents, au préalable toilettés [6], avec l'outil MUSE (Artetxe *et coll.*, 2018) [7].

Les matrices d'appariement étant de dimensions variables selon les longueurs des phrases, nous leur donnons une taille fixe. Chaque phrase ayant moins de dix mots, nous re-dimensionnons les matrices à 10×10. Chaque mot d'une phrase est répété un même nombre de fois afin d'approcher au plus la longueur de dix mots. L'espace restant est rempli à l'aide d'un mot réservé de fin de phrase. Par exemple la phrase *Bonjour à tous .* est re-dimensionnée en *Bonjour Bonjour à à tous tous . .*

4. https://fasttext.cc/docs/en/crawl-vectors.html.
5. https://github.com/wang-h/Hieralign
6. Nous éliminons les mots contenant un signe de ponctuation, les séquences de longueur supérieure à 21, et les mots contenant plus d'un tiret ou mélangeant les casses. Ce toilettage réduit la taille des plongements par deux environ.
7. https://github.com/facebookresearch/MUSE

he ’s my best friend .	:	*he ’s a liar .*	::	*you ’re my best friend .*	:	*you ’re a liar .*
c’ est mon meilleur ami .	:	*c’ est un menteur .*	::	*tu es mon meilleur ami .*	:	*tu es un menteur .*
i ’m not crazy .	:	*you ’re crazy .*	::	*i ’m not a liar .*	:	*you ’re a liar .*
je ne suis pas fou .	:	*tu es fou .*	::	*je ne suis pas une menteuse .*	:	*tu es une menteuse .*
he ’s coming .	:	*i am coming .*	::	*he ’s eating an apple .*	:	*i am eating an apple .*
il est en train d’ arriver .	:	*j’ arrive .*	::	*il est en train de manger une pomme .*	:	*je mange une pomme .*

FIGURE 5 – Exemples d’analogies en deux langues extraites de la partie anglais-français de Tatoeba. On observera que la même phrase anglaise à droite dans les deux premiers quadruplets de bi-phrases se traduit par deux phrases différentes en français. Cette ressource contient beaucoup de phrases ne différant que par un adjectif au masculin ou au féminin.

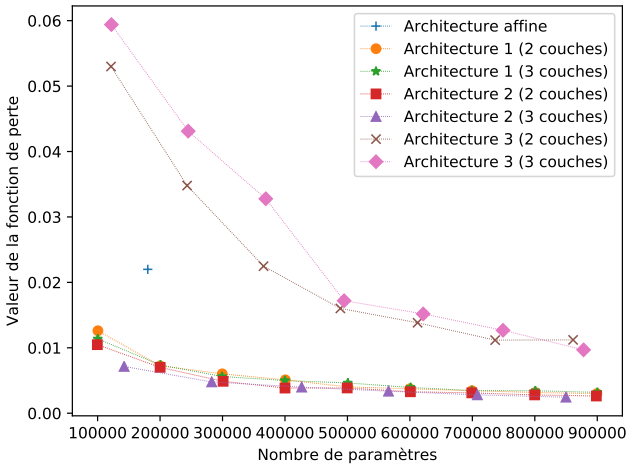


FIGURE 6 – Perte moyenne sur le jeu de test après entraînement dans les expériences pour le choix de l’architecture. Pour rendre les architectures comparables entre elles, les pertes moyennes sont exprimées en fonction du nombre de paramètres entraînables alloué. L’architecture 2 à trois couches offre le meilleur rapport perte / nombre de paramètres.

<*fin*> <*fin*>. Ce re-dimensionnement conserve l’analogie formelle de commutation s’il est appliqué pareillement aux quatre phrases de l’analogie (Lepage, 2018).

5 Résultats et conclusion

5.1 Évaluation des modèles neuronaux

Les différentes architectures de réseaux de neurones présentées précédemment ont été entraînées puis testées sur le jeu de données décrit ci-dessus.

La fonction de perte utilisée est simplement l’erreur quadratique moyenne entre la sortie du réseau et la sortie attendue. La figure 6 présente la perte moyenne obtenue sur le jeu de test après entraînement des différentes architectures dans une expérience préliminaire. L’architecture permettant d’obtenir la plus petite perte moyenne est l’*architecture* 2 à trois couches cachées qui correspond à un canal par matrice de sortie. Comme le montre la figure 6, cette architecture offre le meilleur compromis entre

Type de couche	Dimension de l'entrée	Dimension de la sortie	Activation	Nombre de paramètres
linéaire	$4 \times 10 \times 10$	352	ReLU	141 152
linéaire	352	352	ReLU	124 256
linéaire	352	352	ReLU	124 256
linéaire	352	$1 \times 10 \times 10$	tanh	35 300
total				424 964

TABLE 3 – Caractéristiques et nombre de paramètres pour chacun des trois canaux effectuant la prédiction de $\mathcal{M}_{DD'}$, $\mathcal{M}_{B'D'}$ ou $\mathcal{M}_{C'D'}$. Un canal prend quatre matrices de taille 10×10 en entrée et retourne une matrice de taille 10×10 en sortie. Il aurait été possible de partager les paramètres entre les canaux pour $\mathcal{M}_{B'D'}$ et $\mathcal{M}_{C'D'}$ puisqu'ils effectuent un travail de même nature.

Paramètres	Valeurs
taille des lots	64
taux d'apprentissage initial	0,001
décroissance	0,9
patience	20
optimiseur	Adam
critère de convergence	erreur quadratique en moyenne

TABLE 4 – Hyper-paramètres utilisés lors de l'entrainement.

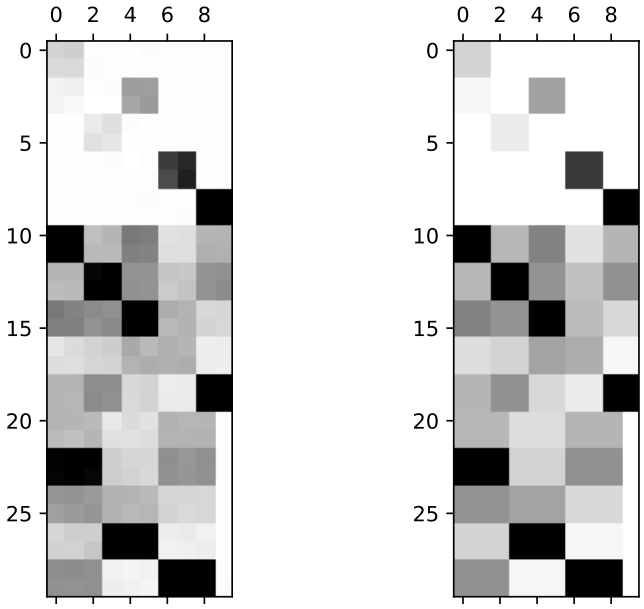


FIGURE 7 – Exemple de sortie du réseau (à gauche) comparée à sa sortie attendue (à droite). Une sortie est constituée par les trois matrices $\mathcal{M}_{DD'}$, bilingue, en haut et $\mathcal{M}_{B'D'}$ et $\mathcal{M}_{C'D'}$, toutes deux monolingues, dessous.

le coût de calcul, lié au nombre de paramètres, et sa capacité d'apprentissage, donnée par la perte moyenne après convergence.

La version finale du réseau utilisée dans les expériences décrites au paragraphe 5.2 utilise 352 neurones par couches cachées pour chaque canal prédisant une matrice (voir tableau 3), ce qui fait un nombre total de $424\,964$ paramètres. Au total on a donc $3 \times 424\,964 = 1\,274\,892$ paramètres. Les hyper-paramètres utilisés sont listés dans le tableau 4.

Dans nos expériences, la perte moyenne atteint des valeurs autour de $0{,}0012$, soit environ un millième. Ce résultat est assez positif pour des valeurs dans l'intervalle $[-1, 1]$. La figure 7 présente un exemple de sortie comparée à la sortie attendue : les deux sont visuellement très proches.

5.2 Évaluation de la traduction

Les résultats d'évaluation de la traduction sont donnés dans le tableau 5, page suivante. Dans une expérience avec des tables de traduction pour le calcul des matrices d'appariement bilingues, le décodage des phrases cibles à partir des matrices sorties par le réseau de neurones, par application de la formule (4), permet d'obtenir un score BLEU de $94{,}7$ sur les phrases du jeu de test [8].

Quatre-vingt-dix pour cent des phrases ont été exactement traduites par notre méthode. En moyenne, les phrase traduites diffèrent de la phrase de référence par un cinquième de mot ou un peu plus de la moitié d'un caractère. Rappelons qu'une phrase contient $5{,}8$ mots ou $25{,}9$ caractères en moyenne (voir tableau 1). On observe en parcourant les résultats que cette différence consiste assez souvent dans le « e » du féminin.

Il est indéniable que la tâche est facile, et le système de traduction automatique neuronal OpenNMT [9], dans sa configuration la plus simple, obtient un score BLEU de $90{,}3$ sur le même jeu de données. Ce score est cependant en retrait du nôtre et la différence statistique est significative comme le montrent les intervalles de confiance donnés dans le tableau 5.

Ces très bons résultats contrastent avec ceux obtenus avec des plongements lexicaux bilingues. Les scores sont extrêmement décevants. Ils sont certainement à expliquer d'une part par la quantité considérable de bruit provenant des plongements monolingues, même toilettés, et d'autre part par le manque de fiabilité de l'alignement automatique des plongements monolingues.

5.3 Remarques finales

Le point le plus critiquable de l'approche de traduction automatique mentionnée ici est la supposition que la recherche de trois couples de bi-phrases peut toujours être couronnée de succès. Cette première étude a laissé ce point important de côté. Il est cependant à noter que l'utilisation de plongements lexicaux, et donc de mesures de similarité sémantique, ouvre des portes qui étaient fermées par l'utilisation d'analogies formelles reposant sur des égalités entre mots. Les résultats de l'étape de recherche pourront être beaucoup moins rigides.

Les données utilisées dans cette première étude étaient très particulières : ce sont en fait des analogies

8. En référence à la construction de l'ensembles des mots candidats d'un mot donné décrite au paragraphe 3.2, mentionnons que l'ajout des 100 mots les plus fréquents du français permet un gain de 2 ou 3 points BLEU.

9. https://opennmt.net/

Méthode	BLEU	Distance en mots	Distance en caractères	Exactitude (%)
OpenNMT	90,3 ± 0,1	0,5	1,0	82,7
méthode proposée :				
table de traduction	94,7 ± 0,1	0,2	0,6	90,2
plongement bilingue	14,4 ± 0,1	6,3	20,6	2,5

TABLE 5 – Résultats de traduction sur les phrases du jeu de test. La méthode proposée est testée pour deux configurations pour l'appariement bilingue : l'une utilise les scores données par une table de traduction ; l'autre utilise un espace de représentations vectorielles de mots partagé par les langues source et cible. La système de traduction neuronal OpenNMT est utilisé à titre de comparaison.

formelles de commutation. Nous désirons étendre nos travaux à des cas plus souples, comme les analogies sémantico-formelles introduites dans (Lepage, 2019), ou généraliser encore en exploitant directement des représentations vectorielles de phrases comme dans (Diallo *et coll.*, 2019).

Remerciements

Les résultats de cette étude ont été en partie obtenus dans le cadre d'un projet subventionné par la Société japonaise pour la promotion de la science, JSPS, Kakenhi Kiban C, n° 18K11447 intitulé « *Self-explainable and fast-to-train example-based machine translation using neural networks.* »

Références

AAMODT A. & PLAZA E. (1994). Case-based reasoning : Foundational issues, methodological variations, and system approaches. *AI Communications*, **7**(1), 39–59. DOI : 10.3233/AIC-1994-7104.

ARTETXE M., LABAKA G. & AGIRRE E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 789–798, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1073.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

COLLINS B. & SOMERS H. (2003). *EBMT seen as case-based reasoning*, In *Recent Advances in Example-Based Machine Translation*, p. 115–153. Springer Netherlands : Dordrecht. DOI : 10.1007/978-94-010-0181-6_4.

DANDAPAT S., MORRIESSY S., NASKAR S. K. & SOMERS H. (2010). Mitigating problems in analogy-based EBMT with SMT and vice versa : a case study with named entity transliteration. In *Proceedings of the 24th Pacific Asia Conference on Language Information and Computation (PACLIC 2010)*, p. 365–372, Sendai, Japan. ACL anthology : Y10-1041.

DIALLO A., ZOPF M. & FÜRNKRANZ J. (2019). Learning analogy-preserving sentence embeddings for answer selection. In *Proceedings of the 23rd Conference on Computational Natural Language*

Learning (CoNLL), p. 910–919, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/K19-1085.

FAM R. & LEPAGE Y. (2018). Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, p. 1060–1066, Miyazaki, Japan : ELRA. ACL anthology : L18-1171.

KAVEETA V. & LEPAGE Y. (2016). Solving analogical equations between strings of symbols using neural networks. In *Proceedings of the Computational Analogy Workshop at the 24th International Conference on Case-Based Reasoning (ICCBR-16)*, volume 1815, p. 67–76, Atlanta, Georgia. CEUR-WS : Vol-1815/paper7.

LANGLAIS P. (2016). Efficient identification of formal analogies. In *Proceedings of the Computational Analogy Workshop at the 24th International Conference on Case-Based Reasoning (ICCBR-16)*, p. 77–86, Atlanta, Georgia. CEUR-WS : Vol-1815/paper8.

LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2008). Analogical translation of medical words in different languages. In A. RANTA & N. NORDSTRÖM, Éds., *Gotal'08 : Proceedings of the 6th international conference on Advances in Natural Language Processing*, volume 5221 de *Lecture Notes in Artificial Intelligence*, p. 284–295, Berlin, Heidelberg : Springer Verlag. DOI : 10.1007/978-3-540-85287-2_27.

LANGLAIS P., ZWEIGENBAUM P. & YVON F. (2009). Improvements in analogical learning : application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, p. 487–495, Athens, Greece : Association for Computational Linguistics. ACL anthology : E09-1056.

LEPAGE Y. (1998). Solving analogies on words : an algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98) and the 17th International Conference on Computational Linguistics (COLING'98)*, volume I, p. 728–735, Montreal. DOI : 10.3115/980845.980967.

LEPAGE Y. (2003). *De l'analogie rendant compte de la commutation en linguistique*. Mémoire d'habilitation à diriger les recherches, Université de Grenoble. HAL Id : tel-00004372.

LEPAGE Y. (2017). Character–position arithmetic for analogy questions between word forms. In *Proceedings of the Computational Analogy Workshop at the 24th International Conference on Case-Based Reasoning (ICCBR-17)*, p. 17–26, Trondheim, Norway. CEUR-WS : Vol-2028/paper2.

LEPAGE Y. (2018). String transformations preserving analogies. In *Proceedings of the 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2018)*, Yogyakarta. DOI : 10.1109/ICACSIS.2018.8618162.

LEPAGE Y. (2019). Semantico-formal resolution of analogies between sentences. In Z. VETULANI & P. PAROUBEK, Éds., *Proceedings of the 9th Language & Technology Conference (LTC 2019) – Human Language Technologies as a Challenge for Computer Sicience and Linguistics*, p. 57–61. En ligne.

LEPAGE Y. & DENOUAL E. (2005). Purest ever example-based machine translation : detailed presentation and assessment. *Machine Translation*, **19**, 251–282. DOI : 10.1007/s10590-006-9010-x.

NAGAO M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. ELITHORN & R. BANERJI, Éds., *Proceedings of the international NATO symposium on Artificial and human intelligence*, p. 173–180 : Elsevier Science Publishers, NATO. En ligne.

RHOUMA R. (2018). *Apprendre à résoudre des analogies de forme*. Thèse de doctorat, université de Montréal. Permalien : 1866/21742.

RHOUMA R. & LANGLAIS P. (2018). Experiments in learning to solve formal analogical equations. In M. T. COX, P. FUNK & S. BEGUM, Éds., *Proceedings of the 26th International Conference on Case-Based Reasoning (ICCBR-18)*, p. 438–453, Stockholm, Sweden : Springer. DOI : 10.1007/978-3-030-01081-2_40.

WANG H. & LEPAGE Y. (2017). Hierarchical sub-sentential alignment with IBM models for statistical phrase-based machine translation. *Journal of Natural Language Processing*, **24**(4), 619–646. DOI : 10.5715/jnlp.24.619.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Impact de la structure logique des documents sur les modèles distributionnels : expérimentations sur le corpus TALN

Ludovic Tanguy, Cécile Fabre, Yoann Bard
CLLE : Université de Toulouse & CNRS, France
ludovic.tanguy@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr,
yoann.bard@hotmail.fr

RÉSUMÉ

Nous présentons une expérience visant à mesurer en quoi la structure logique d'un document impacte les représentations lexicales dans les modèles de sémantique distributionnelle. En nous basant sur des documents structurés (articles de recherche en TAL) nous comparons des modèles construits sur des corpus obtenus par suppression de certaines parties des textes du corpus : titres de section, résumés, introductions et conclusions. Nous montrons que malgré des différences selon les parties et le lexique pris en compte, ces zones réputées particulièrement informatives du contenu d'un article ont un impact globalement moins significatif que le reste du texte sur la construction du modèle.

ABSTRACT

Impact of document structure on distributional semantics models: a case study on NLP research articles

We present an experiment which aims at measuring the impact of document structure on distributional semantics models. Based on a structured corpus of French research articles in NLP, we have built different models by removing specific parts of the corpus: section headers, abstracts, introductions and conclusions. We show that removing these parts has different effects, depending on the target words and the nature of the removed part. Most importantly, we show that these different parts of a research article that are considered particularly informative of its content have a significantly lesser effect than the rest of the text on the resulting distributional models.

MOTS-CLÉS : structure de document, analyse distributionnelle, corpus spécialisé.

KEYWORDS: Document structure, distributional semantics, specialised corpus.

1 Introduction

Les modèles de sémantique distributionnelle sont généralement construits à partir de très grands corpus rassemblant des textes hétérogènes, afin de tirer bénéfice de la masse des données, dont l'impact sur la qualité du modèle généré a été maintes fois démontré, par exemple par Sahlgren & Lenci (2016). Or, de nombreuses applications nécessitent de construire des représentations sémantiques propres à un domaine de spécialité dans le cadre d'un travail terminologique qui vise en particulier la construction de vocabulaires contrôlés et d'ontologies. C'est le cas par exemple de Bernier-Colborne (2014) pour le domaine de l'environnement, ou de Cohen & Widdows (2009) dans le domaine bio-médical. Dans le cadre de ces travaux à visée applicative où la nature précise des données importe,

Volume 2 : Traitement Automatique des Langues Naturelles, pages 122–135.
HAL : hal-02784760.

ces modèles génériques ne sont pas directement utilisables (El Boukkouri *et al.*, 2019). Or, construire des word embeddings à partir de corpus spécialisés pose des problèmes spécifiques : ces corpus sont généralement de taille plus modeste que les grands corpus qui servent à entraîner les modèles de référence –même s'il faut nuancer cette affirmation pour certains domaines comme la médecine, dans le cas de l'anglais. En outre, les unités de sens sont souvent des termes complexes, qui, du fait de leur spécificité, réduisent encore le volume des contextes exploitables. Enfin, les protocoles d'évaluation sont plus difficiles à établir, sauf à disposer de ressources termino-ontologiques. En revanche, ces corpus spécialisés peuvent posséder des caractéristiques potentiellement intéressantes : le lexique est réduit, moins ambigu, et les documents sont généralement très structurés. C'est cette dernière caractéristique dont nous cherchons à tirer parti dans ce travail.

L'expérience que nous décrivons dans cet article cherche à évaluer la possibilité de prendre en compte la structure des documents dans la construction de modèles sémantiques distributionnels à partir de corpus spécialisés. Ce niveau d'information structurel a été exploité dans différentes tâches, comme le résumé ou l'extraction de connaissances (Hofmann *et al.*, 2009; Teufel & Moens, 2002), en partant du principe que certaines zones textuelles facilitent la détection de contenus plus saillants et donc plus utiles. Dans le même ordre d'idées, notre objectif ici est de déterminer si certaines zones de texte ont un impact particulier sur le modèle sémantique et pourraient donc être privilégiées. Notre hypothèse est en effet que certaines parties des textes comme le résumé ou l'introduction sont particulièrement denses en information et susceptibles de fournir une information distributionnelle de meilleure qualité.

Dans cet objectif, nous avons choisi de privilégier un corpus structuré homogène, dont les parties peuvent être identifiées de façon systématique. Le corpus sur lequel est fondé notre expérience est constitué d'articles scientifiques dans le domaine du traitement automatique des langues. Le caractère standardisé des articles scientifiques a été largement étudié. La spécificité des différentes sections ou étapes qui jalonnent ces textes a permis de dégager des modèles argumentatifs génériques valables à travers la diversité des disciplines (Swales, 1990; Teufel *et al.*, 2009). Dans les disciplines expérimentales, le format de type IMRaD (Introduction, Method, Results and Discussion) s'est imposé (Sollaci & Pereira, 2004). Cette codification facilite l'exploitation de la structure des textes, accessible par les titres de section, qui constituent des indices de présentation externe relativement stables. La spécificité argumentative de chaque section se traduit également sur le plan lexical, ce qui permet de faire l'hypothèse de différences de contribution dans la construction de modèles distributionnels. Ainsi, (Bertin & Atanassova, 2014), s'intéressant au vocabulaire verbal lié aux contextes de citation, font état de différences lexicales importantes d'une section à l'autre. Plus près des objectifs qui sont les nôtres dans cet article, (Badenes-Olmedo *et al.*, 2017) ont étudié la capacité de différents fragments spécifiques d'un article à fournir un équivalent informationnel représentatif de l'ensemble. Leur étude compare l'apport du résumé et celui d'autres zones du texte présentant le contexte, l'approche, ou les résultats de l'article. Ils montrent des différences nettes de représentativité selon les sections considérées, mettant en cause l'utilisation par défaut du résumé comme source privilégiée d'accès au texte.

Nous examinons le rôle spécifique de certaines zones des documents, à savoir l'introduction, le résumé, les titres de section et les conclusions, qui constituent les sections les plus systématiquement représentées dans les articles du corpus TALN, dont le degré de codification est limité. Nous mesurons leur impact en utilisant comme critère le score de variation entre des modèles construits en intégrant ou en supprimant les zones concernées. Nous présentons tout d'abord le dispositif déployé : le corpus constitué pour l'occasion (section 2), les modèles distributionnels construits et la façon de les comparer (section 3). La section 4 présente les analyses quantitatives (mesures globales de la variation entre les modèles) et quantitatives (étude des éléments du lexique plus ou moins impactés).

2 Le corpus TALN

Le corpus TALN utilisé dans cette expérience a été construit à partir des archives qui rassemblent les articles des conférences TALN et RÉCITAL [1] des années 1997 à 2019. Ce corpus a été constitué en plusieurs étapes :

— constitution des archives PDF et récupération des métadonnées des articles de 1997 à 2015, *cf.* (Boudin, 2013) [2] ;
— collecte additionnelle des actes de 2016 à 2019 sur le site des éditions de la conférence ;
— sélection des articles rédigés en français ;
— extraction du contenu textuel en balisant les éléments de la structure logique du document présentée dans la figure 1.

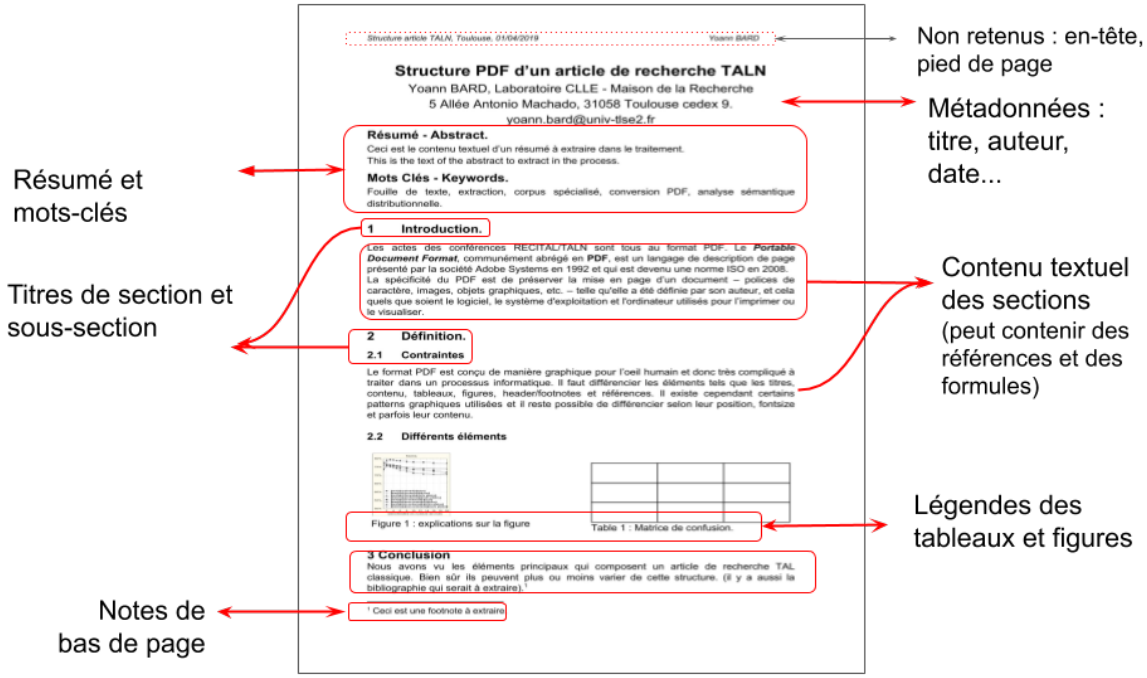


FIGURE 1 – Éléments structurels extraits et annotés dans un article TALN

Cette dernière étape a été réalisée en convertissant les fichiers PDF au format XML grâce à des outils de conversion comme la bibliothèque *pdfminer* de python [3], puis en récupérant le contenu textuel pertinent et en segmentant chaque partie des articles grâce à l'outil ParsCit (Councill *et al.*, 2008).

Un premier travail de restructuration a été effectué sur la sortie de ParsCit pour conserver la structure interne des articles en la représentant dans un format XML *ad hoc*. Nous avons ensuite procédé à un nettoyage automatique pour corriger certains problèmes liés aux caractères spéciaux, traiter les césures et éliminer des éléments résiduels comme les numéros de page et les noms d'auteur figurant dans les en-têtes et les pieds de page, ainsi que différents symboles et idéogrammes. Plusieurs légendes de figures et de tableaux, titres de section et notes de bas de page n'avaient pas été segmentés correctement ou se trouvaient affectés à d'autres éléments du document. La plupart de ces erreurs de segmentation suivant un schéma bien précis, il a été possible de les corriger automatiquement. Les erreurs qui subsistaient après cette phase correspondaient à des ambiguïtés d'interprétation de la

1. Disponibles sur https://www.atala.org/-Conference-TALN-RECITAL
2. Disponibles sur https://github.com/boudinfl/taln-archives
3. https://pypi.org/project/pdfminer/

forme des documents, parmi lesquelles des titres de section identifiés comme des notes de bas de page ou encore des parties de légende sectionnés qui se confondaient avec le corps d'un paragraphe. Nous avons opté pour une approche semi-automatique consistant à développer des règles pour identifier les cas ambigus et à faire ensuite appel à une vérification manuelle des zones correspondantes.

Le marquage obtenu indique donc les méta-données (auteurs, date, type d'article), les informations de la première page (titre, résumés et mots-clés en français et en anglais), les titres des sections et sous-sections, les paragraphes, les légendes de figures et de tableaux, les notes de bas de page ainsi que la bibliographie (mais cette dernière reste non analysée à ce stade). Les sections ont été dotées d'un attribut supplémentaire qui indique leur fonction sur la base de règles ad hoc, pour identifier les introductions et les conclusions. Nous n'avons considéré comme introductions que les sections dont le titre est "Introduction". Pour identifier les conclusions nous avons intégré un ensemble de variantes ("Conclusion", "Perspectives", "Conclusion et Perspectives", etc.). Notons que certains articles ne disposent pas de telles sections, soit parce que les auteurs ont fait un autre choix dans la structuration ou le titrage, soit parce qu'ils relèvent d'un type de communication spécifique (démonstration de logiciel, prise de position, charte, etc.).

Le corpus final est constitué de 1602 articles pour un total de 5,8 millions de mots (tokens). Parmi eux, 1321 articles contiennent un corps de texte. Pour les autres, seuls les métadonnées, résumés et mots-clés ont pu être identifiés. Ce cas de figure s'explique par le codage PDF spécifique de certains fichiers empêchant leur conversion en texte, et par la présence d'articles en anglais dont les métadonnées ont été conservées lorsqu'on disposait d'au moins un contenu exploitable en français (résumé, traduction du titre, mots-clés). L'ensemble du corpus (métadonnées et contenu) a été mis sous format XML en respectant la norme TEI P5. Le corpus TALN est utilisable librement pour les besoins de la recherche et à des fins non commerciales grâce à une licence spécifique accordée par l'ATALA. Il est disponible sur la plateforme de diffusion *Ortolang* [4].

3 Construction et comparaison des modèles distributionnels

Nous présentons ici les opérations qui mènent du corpus TALN aux modèles sémantiques et à leur comparaison.

3.1 Construction des versions de corpus

Nous avons créé pour notre expérience plusieurs corpus dérivés du corpus TALN décrit dans la section précédente. Nous avons commencé par en supprimer les éléments suivants : métadonnées, titre de l'article, auteurs et bibliographie. Nous avons ensuite construit un corpus de référence et 4 sous-corpus correspondant chacun à la suppression d'une zone particulière des documents :

- un corpus de référence comportant l'ensemble du contenu des zones de document suivantes : titre, résumé en français, titres des sections et sous-sections, paragraphes, notes de bas de page et légendes de figure ou de tableau (pour un total de 4 915 365 mots).
- un corpus sans titres de section et sous-section (4 865 324 mots, 99% du corpus de référence).
- un corpus sans les résumés (4 734 393 mots, 96 % du corpus de référence).
- un corpus sans les introductions (4 420 989 mots, 90 % du corpus de référence).
- un corpus sans les conclusions (4 640 218 mots, 94 % du corpus de référence).

La taille et le nombre des éléments supprimés du corpus de référence peuvent varier selon les articles.

4. https://www.ortolang.fr/market/corpora/corpus-taln

La table 1 fournit les caractéristiques des segments pris en compte.

Zone	Nombre	Longueur moy.	Longueur min.	Longueur max.	Total (mots)
Titres de section	14 027	4 mots	1 mot	28 mots	50 041
Résumés	1534	118 mots	16 mots	907 mots	180 972
Introductions	1202	422 mots	23 mots	1798 mots	494 376
Conclusions	1165	241 mots	21 mots	1171 mots	275 147

TABLE 1 – Nombre et taille des zones de documents étudiées dans le corpus

3.2 Modèles distributionnels

Ces corpus ont été utilisés pour construire des modèles distributionnels en utilisant une technique fréquentielle classique, basée sur l'outil DISSECT (Dinu *et al.*, 2013). Ces modèles sont plus précisément construits en extrayant les cooccurrences entre mots dans une fenêtre symétrique de trois mots (en respectant les limites des phrases et des éléments de structure). Les mots sont identifiés par leur lemme et leur catégorie grammaticale. Nous retenons ceux qui appartiennent à une classe ouverte (nom, adjectif, verbe ou adverbe) et qui ont une fréquence minimale de 50 occurrences. La valeur de la cooccurrence entre deux mots est mesurée par le score d'information mutuelle positive (PPMI) et la similarité entre les mots par la mesure de cosinus entre les vecteurs correspondants, sans étape préalable de réduction du nombre de dimensions de la matrice. La simplicité de ces modèles est justifiée par notre volonté de préserver leur interprétabilité et notamment de pouvoir identifier le rôle des contextes individuels sur les similarités calculées.

Nous avons ainsi construit cinq modèles distributionnels à partir des cinq corpus présentés en 3.1, que nous appelons : (1) modèle complet, (2) modèle sans titres, (3) modèle sans résumés, (4) modèle sans introductions, (5) modèle sans conclusions. La comparaison entre le modèle complet et un modèle calculé sur un sous-corpus nous permet d'observer les variations induites par le retrait d'une des quatre zones de document et d'identifier les mots dont la représentation est modifiée de façon importante, ou est au contraire inchangée.

3.3 Modèles de contrôle

Afin de vérifier si ces éléments structurels jouent un rôle spécifique, nous avons également construit des modèles basés sur des corpus construits en retirant de manière aléatoire une quantité de mots égale à celle des structures étudiées, sans scinder les phrases. Pour chacun des quatre sous-modèles nous avons construit dix modèles aléatoires équivalents. L'analyse de ces derniers par rapport aux sous-modèles originaux nous aide à vérifier si la variation est seulement causée par la diminution de la taille du corpus ou si le contenu des zones ciblées a un impact spécifique.

Nous avons également testé des versions plus récentes des modèles distributionnels, en construisant des modèles prédictifs. Nous nous sommes limités à l'utilisation de la version de base de *Word2vec* (Mikolov *et al.*, 2013) et construit pour chacun des cinq corpus un modèle *SGNS* (*skip-gram with negative sampling*) en choisissant pour les hyperparamètres les valeurs par défaut de l'outil (plus précisément de son implémentation dans la bibliothèque GenSim (Řehůřek & Sojka, 2010)) sauf pour ceux qui font écho aux caractéristiques des modèles fréquentiels précédents : fréquence minimale de 50 occurrences et fenêtre de taille 3 pour les contextes.

3.4 Mesure de la variation entre deux modèles

Pour évaluer la variation nous utilisons le *Ranked Biased Overlap* (ou RBO) (Webber *et al.*, 2010) qui calcule la similarité entre deux listes ordonnées L_1 etL_2 jusqu'au rang n suivant la formule suivante :

$$RBO(L_1, L_2, n) = (1-p)\sum_{k=1}^{n} p^{k-1}\frac{|L_1(1:k) \cap L_2(1:k)|}{k}$$

Où $L(1:k)$ représente les k premiers éléments de la liste L. Cette mesure considère donc le recouvrement partiel des deux listes comparées à chaque rang, en accordant plus d'importance aux débuts des listes. Le paramètre p (fixé ici à 0,9) est le coefficient d'atténuation de la prise en compte des différences lorsque l'on avance dans les listes. Les listes comparées ici sont les $n = 20$ plus proches voisins d'un mot donné dans deux modèles distributionnels. Le score RBO est normalisé pour varier de 0 (aucun voisin commun) à 1 (listes identiques).

La table 2 illustre cette mesure dans le cas du nom *significativité* dont les 5 premiers voisins sont donnés pour 3 modèles. On peut voir que le RBO est très élevé (0,89) dans le cas d'une suppression des titres de section, avec une différence très faible entre les deux listes (identiques pour les 5 premiers voisins). Le voisinage est de plus en plus perturbé par les autres suppressions, avec un score RBO qui tombe à 0,1 lorsqu'on enlève les conclusions.

Rang	**Modèle complet**	**Modèle sans titres**	**Modèle sans résumés**	**Modèle sans introductions**	**Modèle sans conclusions**
1	corrélation	corrélation	corrélation	corrélation	performance
2	statistiquement	statistiquement	performance	statistiquement	proximité
3	performance	performance	statistiquement	proximité	statistiquement
4	proximité	proximité	proximité	significatif	empiriquement
5	cohésion	cohésion	empiriquement	performance	confiance
RBO	-	0,89	0,46	0,34	0,10

TABLE 2 – Comparaison du score RBO et des voisins du mot *significativité* entre le modèle de référence et chacun des quatre modèles obtenus par suppression d'une zone de texte dans le corpus

Nous avons également considéré deux autres mesures qui permettent de comparer le voisinage distributionnel d'un mot entre deux modèles. La première est le coefficient de Jaccard, autrement dit le ratio de voisins en commun qu'a un mot entre les deux modèles. Cette mesure a été couramment utilisée pour comparer des modèles distributionnels, notamment dans (Pierrejean & Tanguy, 2018b). Nous l'avons calculé en prenant en compte, comme pour le RBO, les 20 premiers voisins de chaque mot dans l'ordre de similarité. La différence principale avec le RBO est que l'ordre des voisins n'est pas pris en compte.

La dernière mesure que nous avons calculée se concentre exclusivement sur le plus proche voisin du mot dans chacun des deux modèles comparés, et consiste en la moyenne des différences de rangs de ces plus proches voisins, comme indiqué dans la formule ci-dessous :

$$diffrang_{M_1,M_2}(m) = \frac{|rang^m_{M_2}(m_1)| + |rang^m_{M_1}(m_2)|}{2}$$

où m_i est le plus proche voisin du mot-cible m suivant le modèle M_i et $rang^m_M(n)$ est le rang du mot n dans la liste des voisins du mot m suivant le modèle M, ordonnés par similarité décroissante

(en comptant à partir de 0). Suivant cette formule, une différence de zéro indique que les deux plus proches voisins sont identiques et une différence importante indique que l'un ou l'autre de ces premiers voisins (ou les deux) se trouve relégué plus loin dans la liste.

3.5 Mesure de la spécificité des mots

Dans un corpus spécialisé, la question de la qualité de la représentation des termes du domaine est cruciale. Afin de distinguer l'impact des modèles sur les mots spécifiques au domaine du TAL, nous avons utilisé une mesure de spécificité en comparant les fréquences dans notre corpus avec celles du corpus FrWac telles que fournies par le lexique GLAFF (Sajous *et al.*, 2013). Nous avons utilisé le score de χ^2 pour identifier les mots dont la fréquence relative est significativement plus élevée dans le corpus TALN. Les mots relatifs au domaine de spécialité comme *corpus*, *sémantique*, *annotation*, *supervisé* ou encore *syntaxique* se démarquent ainsi avec un χ^2 élevé. On y retrouve aussi des mots du discours scientifique tels que *afin* et *ci-dessous*. Certains mots sont, à l'inverse, sous-représentés (*avoir*, *numéro* ou *national*). Pour les calculs statistiques utilisant cette mesure, nous avons attribué un signe négatif au χ^2 des mots dont la fréquence relative est inférieure dans le corpus TALN.

Nous avons repris le même procédé pour identifier le vocabulaire spécifique aux zones de texte étudiées par rapport au corpus TALN entier. La table 3 donne les cinq mots les plus spécifiques de chaque partie envisagée (par rapport au reste du corpus).

Titres de section	**Résumés**	**Introductions**	**Conclusions**
introduction	article	introduction	conclusion
conclusion	présenter	section	perspective
référence	automatique	automatique	envisager
perspective	montrer	travail	améliorer
discussion	proposer	langue	montrer

TABLE 3 – Mots avec le score de spécificité (χ^2) le plus élevé pour chaque partie d'article

On voit que les mots les plus spécifiques aux titres sont ceux des sections génériques (*introduction*, *conclusion*, *références* etc.), que l'on retrouve également dans les parties correspondantes. Les mots des autres colonnes sont aisément interprétables comme des éléments de formulations canoniques pour les résumés ("Dans cet article nous présentons/montrons/proposons [...]") ou les conclusions ("Nous avons montré que [...] ; dans la suite nous envisageons de [...]"). Les introductions contiennent quant à elles des termes très génériques du domaine (*langue*, *automatique*) qu'on retrouvera moins dans le développement du texte, conformément au rôle de cette section de créer une niche au sein d'un espace de recherche englobant (Swales, 1990). Quant à *section*, il correspond à la présentation du plan de l'article qui est une des fonctions de l'introduction d'un article scientifique.

4 Analyse

Toutes les analyses présentées ici portent sur les mots dont la fréquence est de 50 ou plus dans chacun des quatre sous-corpus considérés, soit 3107 mots représentés par leur catégorie et leur lemme (uniquement les classes ouvertes).

4.1 Variation entre les modèles distributionnels

Dans un premier temps, nous avons calculé le score RBO de chaque lemme en comparant chaque modèle au modèle de référence. La figure 2 montre la distribution de ce score dans les quatre cas.

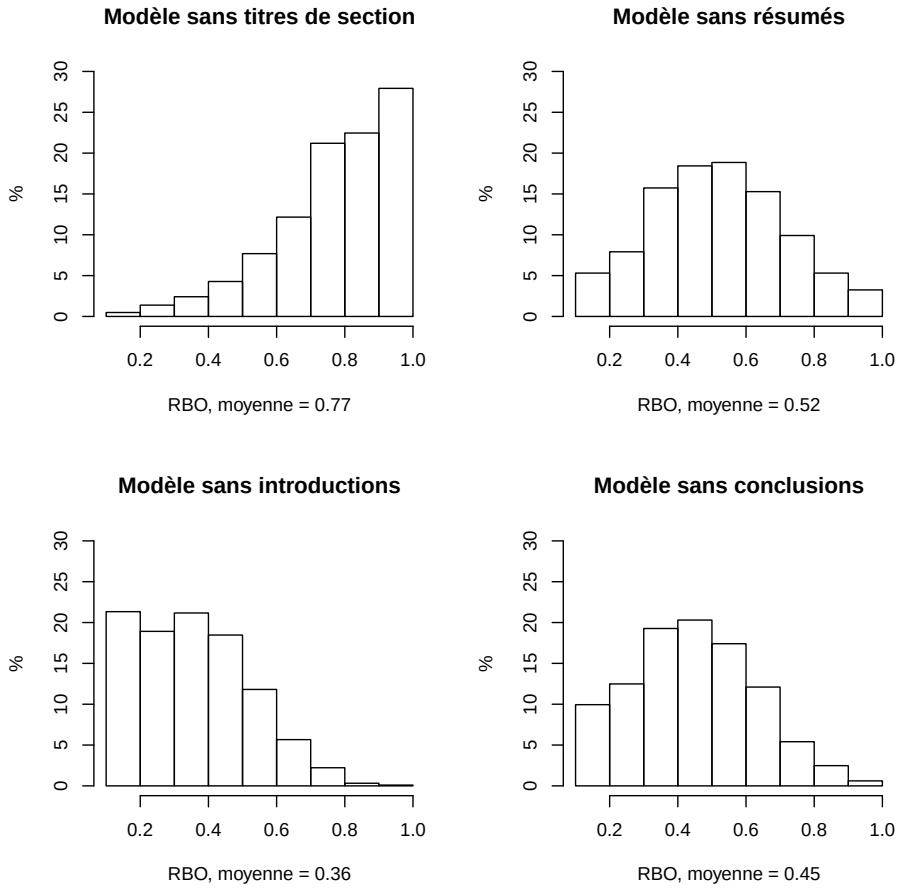


FIGURE 2 – Distribution du score de RBO pour les quatre modèles comparés à celui du corpus complet

Dans le modèle sans titres de section, le RBO moyen est à 0,77, avec une importante proportion de valeurs très élevées, correspondant à des voisinages distributionnels inchangés (RBO de 1). Ce score est nettement plus bas pour les autres modèles : il tombe à 0,36 lorsqu'on supprime les introductions, ce qui correspond au modèle le plus éloigné du corpus entier, avec un nombre important de voisinages radicalement différents (RBO proche de 0). Les modèles correspondant à la suppression des résumés et des conclusions présentent des profils intermédiaires (RBO moyen de 0,52 et 0,45 respectivement) avec une distribution plus symétrique et plus homogène, indiquant que la plupart des mots y voient leur voisins sémantiques partiellement modifiés, mais que les perturbations majeures sont minoritaires.

Cette différence de comportement s'explique en grande partie par le fait que l'amplitude de la variation est proportionnelle à la taille de la zone supprimée. Néanmoins, il apparaît au vu des distributions de la figure 2 que les scores de RBO varient de façon importante au sein du lexique, l'ensemble du spectre étant occupé dans les quatre cas. Ces observations soulèvent deux questions : l'instabilité d'un modèle dépend-elle uniquement de la quantité de texte supprimé ou bien la nature des zones de document ciblées joue-t-elle aussi un rôle déterminant ? Peut-on identifier les principaux facteurs expliquant les variations internes au sein du lexique ?

La première question nous amène à l'analyse des modèles aléatoires. Nous avons calculé la moyenne des RBO (par rapport au modèle complet) pour chacune des quatre séries de 10 modèles aléatoires puis nous l'avons comparée au RBO moyen de chacun des quatre modèles correspondant aux suppressions de parties spécifiques, les résultats étant présentés en table 4.

Zones supprimées	RBO du modèle	RBO moyen des 10 modèles aléatoires (IC 95%)
Titres de section	$\mathbf{0,772}$	$0,757 \pm 0,002$
Résumés	$\mathbf{0,519}$	$0,497 \pm 0,003$
Introductions	$0,360$	$0,358 \pm 0,002$
Conclusions	$0,447$	$0,443 \pm 0,007$

TABLE 4 – Comparaison des RBO moyens pour les quatre modèles et de la moyenne sur les dix modèles aléatoires équivalents en taille, par rapport au modèle complet.

On peut y voir une différence entre le modèle sans titres de section et les dix modèles aléatoires équivalents : le RBO est plus élevé (*i.e.* la variation est moindre) quand on enlève ces titres que lorsqu'on retire une quantité de texte identique choisie aléatoirement (0,772 contre 0,757), la différence étant significative pour les 10 modèles ($p < 0,05$) suivant le test des rangs signés de Wilcoxon sur l'ensemble du vocabulaire. L'analyse du modèle sans les résumés donne des résultats identiques avec un RBO moyen de 0,519 contre 0,497 sur les modèles aléatoires (différence significative également). Par contre, le RBO moyen d'un modèle sans introductions est identique à celui des modèles aléatoires (0,360 et 0,358, aucune différence significative pour les 10). La situation des conclusions est plus mitigée mais la différence avec les modèles aléatoires n'est significative que dans 5 cas sur 10, ce qui nous conduit à conclure que la différence n'est pas suffisamment marquée pour être prise en considération.

Les résumés et les titres de section ont donc bien un effet singulier, mais qui va à l'encontre de notre hypothèse de départ : ces zones influencent moins les modèles distributionnels que d'autres parties de texte de taille équivalente. Quant aux introductions et aux conclusions, elles n'apportent aucune différence nette : elles se comportent sur ce plan comme le reste du texte.

Nous avons enfin calculé le coefficient de corrélation (ρ de Spearman) entre le RBO et le χ^2 du vocabulaire distinctif de chaque zone. La relation linéaire est négative pour les quatre modèles au même niveau : titres de section ρ = -0,39, résumés ρ = -0,37, introductions ρ = -0,38 et conclusions ρ =-0,35. En toute logique, plus un mot est spécifique à une zone du texte (quelle qu'elle soit), moins sa représentation a de chances d'être stable lorsqu'on la retire. Ce n'est cependant pas systématique car la relation linéaire négative n'est pas si forte. La variation dépend donc également d'autres facteurs.

4.2 Variation sur le vocabulaire spécifique

Nous avons cherché à mesurer si la représentation du vocabulaire spécifique du corpus TALN était plus affectée que le vocabulaire courant par le retrait d'une des différentes zones de document. Nous avons mesuré son impact sur le vocabulaire spécialisé du corpus (3.5) en calculant le coefficient de corrélation de Spearman entre le χ^2 (obtenu en comparant les fréquences dans notre corpus et dans FrWac) et le RBO dans la table 5.

Pour les modèles sans résumés, sans introductions et sans conclusions le ρ de Spearman est positif (de 0,15 à 0,24). Les mots spécifiques du corpus TALN (avec un χ^2 élevé) tendent donc à être moins perturbés dans leur représentation (RBO plus haut) que les autres. La corrélation est pratiquement

Modèles	ρ de Spearman
Sans titres de section	-0,06
Aléatoires	0,10
Sans résumés	0,15
Aléatoires	0,18
Sans introductions	0,24
Aléatoires	0,19
Sans conclusions	0,15
Aléatoires	0,18

TABLE 5 – Corrélation de Spearman entre spécificité du mot dans le corpus (χ^2) et variation (RBO) par rapport au modèle complet

nulle (-0,06) pour les titres de section qui n'influent pas plus sur le vocabulaire du domaine que sur les autres mots. Si l'on regarde les modèles aléatoires utilisés comme point de comparaison, la corrélation entre le χ^2 et la moyenne du RBO de ceux-ci est plus élevée, sauf pour les introductions. L'effet de perturbation des mots du domaine semble donc être amoindri à mesure que l'on ôte des portions de texte plus larges.

4.3 Autres mesures

Comme indiqué en section 3.4, nous avons mesuré la variation du voisinage distributionnel d'un mot en utilisant deux autres techniques : le recouvrement des 20 premiers voisins avec le coefficient de Jaccard et la différence de rang des premiers voisins.

Concernant le Jaccard, il est fortement corrélé avec le RBO (ρ de 0,6 à 0,8 suivant les modèles) si bien que l'ensemble des conclusions obtenues précédemment avec le RBO sont confirmées par cette mesure.

La variation du rang des premiers voisins, quant à elle, présente un profil de distribution différent : pour les quatre paires de modèles ce score est nul dans 87% des cas (i.e. les deux modèles comparés ont le même premier voisin pour un mot donné) ce qui rend son usage très délicat pour les analyses statistiques. Sa corrélation est de ce fait faible avec les deux autres coefficients (ρ de 0,1 à 0,4 en valeur absolue [5]). Par contre, ce score est très utile pour isoler les cas de perturbation importante du voisinage distributionnel et donc effectuer une analyse qualitative permettant de mieux comprendre les mécanismes à la base de ces variations.

Nous avons enfin voulu comparer ces résultats à ceux que l'on obtient en utilisant des méthodes prédictives de construction des représentations distributionnelles. En mesurant les scores RBO pour comparer les modèles construits sur les sous-corpus à celui du corpus complet, nous n'observons que très peu de différence entre les modèles. Le RBO moyen va en effet de 0,62 à 0,68 en suivant le même ordre que celui observé en 4.1, mais avec un resserrement qui ne permet pas de distinguer clairement le rôle des différentes parties de corpus supprimées. On observe d'ailleurs une très forte corrélation entre les 4 séries de score RBO (0,7 en moyenne, contre seulement 0,4 sur les modèles fréquentiels). Il nous apparaît donc clairement que ces modèles prédictifs sont des instruments trop grossiers pour mettre au jour des variations fines, et que les différences entre les corpus d'apprentissage ne sont pas suffisantes au vu des techniques qui tendent à lisser le matériau (notamment les processus aléatoires qui ont une influence très importante sur les modèles, *cf* (Pierrejean & Tanguy, 2018a)).

5. La corrélation est négative, deux listes identiques entraînent un RBO de 1 et une variation de rang de 0.

4.4 Analyse qualitative

Dans cette dernière partie nous effectuons des observations plus fines en nous basant sur les scores (RBO et différence de rang), les listes des plus proches voisins, et les contextes d'apparition. Nous nous concentrons tout d'abord sur les variations des représentations des termes dans le modèle **sans résumés**, en observant les mots présentant des variations importantes. On y trouve deux cas de figure distincts :

— Des mots spécifiques aux résumés (fort χ^2 par rapport aux autres parties du corpus), qui subissent donc en toute logique une forte variation de voisins relative à la diminution de leur fréquence. C'est le cas de *démonstration*, *assistance*, *arboré* et *informatisé*. Ces termes sont employés régulièrement dans le domaine.

— Des mots non spécifiques aux résumés mais présentant contre toute attente une forte variation. Ces mots sont pour la plupart polysémiques, tels que *synthétique*, *couper*, *interrompre* et ont des emplois multiples dans le corpus. Cela se voit notamment dans la grande variété de leurs voisins, que l'on peut rattacher à plusieurs sens. Leurs contextes sont donc très variés. Logiquement, une légère modification des données d'entrée du calcul distributionnel suffit à provoquer leur instabilité, sans qu'on puisse pour autant identifier clairement une variation de sens.

À l'inverse, si l'on regarde du côté des mots très stables (RBO élevé) lorsqu'on supprime les résumés, on trouve deux cas de figure :

— Des mots spécifiques aux résumés dont les voisins sont stables malgré la baisse importante de leur fréquence. Parmi eux on retrouve les verbes d'exposition tels que *présenter*, *finaliser* ou encore des évaluatifs comme *vaste*, *important*, *faiblement*. Leur stabilité viendrait de leur ancrage dans des contextes très réguliers, ce qui les rendrait moins vulnérable à la suppression d'un passage spécifique. La suppression de ces zones n'affecte donc pas leur représentation.

— Des mots non spécifiques aux résumés et dont le voisinage n'est pas affecté tels que *nœud*, *pattern* ou *possessif*. Ils sont aussi inscrits dans des zones de texte stables indépendantes de la structure du document.

Cette étude qualitative confirme certains résultats des analyses précédentes. Tout d'abord on retrouve peu de mots du domaine du TAL affectés par l'absence de résumés. La plupart du vocabulaire instable est en fait spécifique à cette zone du texte et pas au corpus. Cependant on remarque que certaines classes (ou clusters) de mots sont très stables et leur représentation le restera quels que soient la quantité ou le type de texte supprimé. L'instabilité d'un mot n'est donc pas seulement provoquée par la suppression de ses occurrences mais dépend surtout des contextes dont il est entouré.

L'examen des variations entraînées par la suppression d'une autre zone nous permet de dégager des phénomènes sémantiques intéressants. C'est le cas lorsqu'on observe les voisinages fortement impactés par la suppression des **conclusions**. Parmi les mots spécifiques à cette partie des articles, on trouve le champ lexical habituel à la présentation des perspectives d'un travail de recherche : *futur*, *prévoir*, *affiner*, *approfondir*, *poursuivre*. Les perturbations des voisinages sont importantes mais limitées à un réordonnancement local des mots les plus proches sans tendance globale précise, comme on l'a vu pour les résumés. En revanche, pour certains de ces mots spécifiques la baisse de fréquence s'accompagne plus clairement d'un changement de sens. C'est le cas notamment de l'adjectif *idéal* qui voit ses voisins passer d'un sens technique (*optimal*, *final*, *contrôlé*) à un évaluatif (*trivial*, *véritable*, *suffisant*). Une autre distinction polysémique est observable pour *engager* dont les emplois correspondent au sens de /participer/, /initier une action/ (avec des voisins comme *participant*, *organisateur*, *entreprendre*), ou au sens plus abstrait et dialectique : *défendre*, *exprimer*, *confronter*, *inciter*.

Ces observations nous conduisent à formuler l'hypothèse que la stabilité d'un mot quand on compare deux modèles distributionnels est liée à l'existence d'un groupe de "bons" voisins, dont l'identification résiste à des modifications partielles des contextes. L'absence de tels points d'ancrage semble entraîner une forte instabilité, quelle que soit la modification subie par le corpus. Les cas intéressants mais rares (qu'illustrent *idéal* et *engager*) que nous recherchons correspondraient à des situations où la représentation d'un mot passe d'un cluster de proches voisins à un autre.

5 Conclusion et perspectives

Les études de variation entre un modèle distributionnel de référence et d'autres modèles obtenus en retirant une partie du document des mêmes données d'apprentissage nous permettent de tirer certaines conclusions concernant l'impact de la structure des articles scientifiques sur la représentation distributionnelle d'un corpus spécialisé. De manière globale les résumés et titres de section impactent moins le voisinage des mots que les autres parties de texte. Ce résultat, qui semble contre-intuitif, peut s'expliquer en faisant l'hypothèse que ces passages sont redondants par rapport au reste du texte, ou qu'y figurent des mots qui s'y comportent différemment, comme nous le suggère l'analyse qualitative. Quant aux introductions et aux conclusions, elles n'ont pas d'influence significative, les enlever a le même effet que retirer une part aléatoire équivalente du corpus.

La représentation des mots spécifiques au domaine et au genre n'est pas affectée par le retrait des titres et se trouve même plus stable que les autres lorsque les résumés et introductions sont retirés. Nous pouvons en déduire que le cœur du vocabulaire technique n'est pas particulièrement employé à l'intérieur de ces zones. Seul le vocabulaire propre à la zone est logiquement touché. Ce dispositif nous pousse à penser que ces parties de document scientifique ne jouent pas de rôle prépondérant vis-à-vis de l'analyse distributionnelle. Une optimisation/pondération des contextes au regard de la zone dans laquelle ils se trouvent serait donc négligeable voir déconseillée selon l'objectif de modélisation recherché.

Pour mettre au jour de véritables variations entre les emplois des mots dans les différentes parties d'un document, l'idéal aurait été de comparer directement des espaces sémantiques construits sur des parties différentes du corpus, à la manière des travaux initiés par (Hamilton *et al.*, 2016) pour la diachronie. Cependant, notre corpus et les sous-parties visées représentent clairement un trop petit volume pour ce type de méthode. L'approche indirecte que nous avons suivie ici pourrait néanmoins être poursuivie à plus large échelle, mais devrait alors se concentrer sur le discours scientifique général, puisque la restriction à un domaine particulier ne permettrait pas d'atteindre la masse critique nécessaire.

Remerciements

Le travail présenté dans cet article a été réalisé dans le cadre du projet ADDICTE [6] (Analyse distributionnelle en domaine spécialisé), financé par l'Agence Nationale de la Recherche (ANR-17-CE23-0001). Il a également reçu le soutien du consortium CORLI pour la finalisation du corpus TALN. Les auteurs tiennent à remercier l'ATALA et son président Patrick Paroubek pour avoir autorisé l'utilisation et la diffusion du corpus constitué des actes des conférences TALN et RECITAL. Ils remercient enfin Alice Adnot-Albinet, Charline Fabre et Clémentine Mailly pour leur aide dans la correction du corpus.

6. https://anr-addicte.ls2n.fr/

Références

BADENES-OLMEDO C., GARCÍA J. L. R. & CORCHO O. (2017). An initial analysis of topic-based similarity among scientific documents based on their rhetorical discourse parts. In *Workshop on Enabling Open Semantic Science co-located with the 16th International Semantic Web Conference (ISWC)*, p. 15–22, Vienna, Austria.

BERNIER-COLBORNE G. (2014). Identifying semantic relations in a specialized corpus through distributional analysis of a cooccurrence tensor. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, p. 57–62.

BERTIN M. & ATANASSOVA I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. In *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with the 36th European Conference on Information Retrieval (ECIR 2014)*, Amsterdam, Netherlands.

BOUDIN F. (2013). TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. In *Actes de TALN'2013*, p. 507–514, Les Sables d'Olonne.

COHEN T. & WIDDOWS D. (2009). Empirical distributional semantics : methods and biomedical applications. *Journal of biomedical informatics*, **42**(2), 390–405.

COUNCILL I. G., LEE GILES C. & KAN M.-Y. (2008). An open-source CRF reference string and logical document structure parsing package. In *Proceedings of the Language Resources and Evaluation Conference (LREC 08), Marrakesh, Morrocco, May.*

DINU G., THE PHAM N. & BARONI M. (2013). Dissect - distributional semantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 31–36.

EL BOUKKOURI H., FERRET O., LAVERGNE T. & ZWEIGENBAUM P. (2019). Embedding strategies for specialized domains : Application to clinical entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 295–301.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 1489–1501, Berlin.

HOFMANN K., TSAGKIAS M., MEIJ E. & DE RIJKE M. (2009). The impact of document structure on keyphrase extraction. In *Proceedings of the 18th ACM conference on Information and knowledge management*, p. 1725–1728.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *ICLR 2013, workshop track*.

PIERREJEAN B. & TANGUY L. (2018a). Predicting word embeddings variability. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, New Orleans.

PIERREJEAN B. & TANGUY L. (2018b). Towards qualitative word embeddings evaluation : Measuring neighbors variation. In *Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, p. 32–39.

ŘEHŮŘEK R. & SOJKA P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45–50, Valletta, Malta : ELRA.

SAHLGREN M. & LENCI A. (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 975–980, Austin, Texas : Association for Computational Linguistics.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.

SOLLACI L. B. & PEREIRA M. G. (2004). The introduction, methods, results, and discussion (imrad) structure : a fifty-year survey. *Journal of the medical library association*, **92**(3), 364.

SWALES J. (1990). *Genre analysis : English in academic and research settings*. Cambridge University Press.

TEUFEL S. & MOENS M. (2002). Summarizing scientific articles : experiments with relevance and rhetorical status. *Computational linguistics*, **28**(4), 409–445.

TEUFEL S., SIDDHARTHAN A. & BATCHELOR C. (2009). Towards discipline-independent argumentative zoning : evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3-Volume 3*, p. 1493–1502 : Association for Computational Linguistics.

WEBBER W., MOFFAT A. & ZOBEL J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, **20**.

Deuxième partie

Articles courts

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Prédire automatiquement les intentions du locuteur dans des questions issues du discours oral spontané

Angèle Barbedette[1] Iris Eshkol-Taravella[1]
(1) Université Paris Nanterre, MoDyCo UMR 7114, 200 avenue de la République, 92000 Nanterre, France
angele.barbedette@gmail.com, ieshkolt@parisnanterre.fr

RÉSUMÉ

Cette étude porte sur la classification automatique des intentions exprimées dans des questions issues d'un corpus d'échanges oraux spontanés. Nous proposons une typologie dans laquelle nous distinguons trois classes d'intentions (AVIS, VOLONTÉ et DOUTE). Après plusieurs prétraitements et ajouts de traits lexicaux aux données (lexiques, nombre de mots et de caractères), nous implémentons un algorithme de classification automatique et nous en présentons et évaluons les résultats qui atteignent une F-mesure de 0,62. Nous proposons ensuite une interprétation de ceux-ci, basée sur une comparaison entre les expériences menées et des mesures liées aux traits linguistiques intégrés avant la tâche de classification.

ABSTRACT

Automatically predicting the speaker's intentions in questions from spontaneous oral speech

This study focuses on the automatic classification of intentions expressed in questions from a corpus of spontaneous oral interactions. We suggest a typology in which we distinguish three categories of intentions (OPINION, WILL and DOUBT). After several preprocessings and additions of lexical features to the data (lexicons, number of words and characters), we implement an automatic classification algorithm and we present and evaluate the results which reach a F-measure of 0.62. We then provide an interpretation of these, based on a comparison between our experiments and some measures related to the linguistic features we integrated before the classification task.

MOTS-CLÉS : intentions, implicite, actes de dialogue, discours oral spontané.

KEYWORDS: intentions, implicit, dialog acts, spontaneous oral speech.

1 Introduction

L'objectif de ce travail est d'étudier les spécificités linguistiques de l'aspect non-littéral, c'est-à-dire de ce qui n'est pas dit de façon explicite, dans des questions issues de transcriptions d'échanges oraux et spontanés. Il s'agit d'une tâche primordiale pour tendre vers des systèmes de dialogue homme-machine plus performants et plus proches de conversations dites « naturelles ». Pour rendre compte de sa complexité, nous proposons un bref état de l'art des travaux en lien avec cette étude réalisés dans les domaines de la linguistique et du TAL.

L'énoncé est porteur d'une valeur illocutoire correspondant à la force ou l'intention qui lui est attribuée (Austin, 1962; Kerbrat-Orecchioni, 1986). Searle introduit de son côté la notion d'acte de langage indirect qui se compose d'un acte illocutoire primaire (non littéral), et d'un acte illocutoire

Volume 2 : Traitement Automatique des Langues Naturelles, pages 137–145.
HAL : hal-02784761.

secondaire (littéral) (Searle, 1975). Dans (Allen & Perrault, 1980), les auteurs parlent également d'action intentionnelle pour définir l'acte de langage. Ces différents travaux introduisent la notion d'intention. Du point de vue du TAL, les travaux sur les intentions s'inscrivent dans les domaines de l'analyse d'opinion et des actes de dialogue. D'après (Karoui *et al.*, 2019, 2014), une opinion est une « expression subjective du langage » qui est soit explicite (repérée à l'aide d'indices textuels), soit implicite (s'appuyant sur des connaissances culturelles ou pragmatiques communes). Les actes de dialogue ont quant à eux pour objectif d'aider à l'analyse fine de la conversation et de tous les types d'énoncés qui la composent. Ils correspondent, selon Bunt, à la combinaison entre le contenu sémantique de l'énoncé et sa fonction communicative (Bunt, 2005). Plusieurs taxonomies d'actes de dialogue ont été proposées pour la tâche de classification automatique en actes de langage (Moldovan *et al.*, 2011) tel que le schéma d'annotation DIT ++ (Bunt, 2009) qui s'inspire d'autres taxonomies telles que DAMSL (Allen & Core, 1997), SWBD-DAMSL (Jurafsky *et al.*, 1997), HCRC Map Task (Anderson *et al.*, 1991) ou encore VERBMOBIL (Alexandersson *et al.*, 2011). Le dialogue est souvent étudié en TAL du point de vue de la sémantique formelle, avec des approches telles que celle de Ginzburg (Ginzburg, 1996a,b) et des théories telles que la SDRT par exemple (Segmented Discourse Representation Theory) (Maudet *et al.*, 2004), où l'identification de l'incompréhension entre les locuteurs s'appuie sur la recherche d'incohérences logiques dans les combinaisons entre les actes de parole, calculées à partir des algorithmes compositionnels fondés sur une représentation logique et dynamique d'une question et d'une réponse (Boritchev & Amblard, 2018, 2019). Selon ces théories, les nouveaux énoncés produits réalisent souvent des actes qui doivent être en lien avec d'autres éléments du contexte et s'inscrivent donc dans une structure complexe et non pas linéaire. Elles mettent en évidence l'importance de la prise en compte du contexte pour permettre la cohérence du dialogue.

Nous considérons, d'après la définition de l'énoncé performatif donnée par Austin (Austin, 1962) qu'en posant une question, le locuteur veut toujours dire quelque chose de plus que ce que la valeur locutoire de la question exprime en réalité. Les questions accomplissent toujours, selon nous et tels que les définit Searle, un acte illocutoire primaire et un acte illocutoire secondaire. L'association des caractéristiques des opinions et des actes de dialogue nous amène à définir ce que nous appelons dans cette étude l'*intention*. Il s'agit de l'activité illocutoire exprimée par un énoncé (« l'ensemble des actes qui s'accomplissent, immédiatement et spécifiquement, par l'exercice de la parole » (Ducrot, 1972)), qui permet de le caractériser selon son but, explicite ou implicite. L'activité illocutoire ne se restreint pas seulement à l'expression des opinions mais à tous les types d'objectifs impliqués par la production d'un énoncé. Dans (Chen *et al.*, 2013), les auteurs différencient les intentions explicites des intentions implicites. Des exemples illustrent ces deux cas : « I am looking for a brand new car to replace my old Ford Focus » qui correspond à une intention explicite d'achat et « Anyone knows the battery life of iPhone ? » qui est une intention implicite d'achat.

Ce travail s'intéresse particulièrement aux intentions implicites exprimées par les locuteurs lorsqu'ils posent une question et tente d'en effectuer la classification automatique ainsi que d'en dégager des spécificités linguistiques en se concentrant uniquement sur l'écrit. Il s'agit de tenir compte du contexte de chaque question à la fois pour l'étape des annotations manuelles et pour celle des annotations automatiques, sans se placer du point de vue de la linguistique formelle. Nous utilisons des techniques existantes ayant déjà fait leurs preuves dans la fouille d'opinion, un des domaines se rapprochant le plus de notre tâche.

2 Corpus, typologie et annotations

Les données utilisées proviennent des corpus oraux ESLO1 et ESLO2 (Enquêtes SocioLinguistiques à Orléans) (Baude & Dugua, 2011; Eshkol-Taravella *et al.*, 2011), comportant respectivement environ 300 et 400 heures d'enregistrement. Les transcriptions des enregistrements sont disponibles sur le site web d'ESLO [1] et ouvertes au public. Nous utiliserons ici toutes les transcriptions d'enregistrements effectués au cours de repas dans ESLO1 et ESLO2, soit un total de 28 fichiers au format *.xml* (sept issus du corpus ESLO1 et 21 issus d'ESLO2). Ils forment un tout d'environ 19 heures d'enregistrement. Le choix de cette catégorie précise est lié au fait de vouloir utiliser les données les plus spontanées possibles pour répondre à l'objectif principal de ce travail consistant à prédire l'intention du locuteur à travers les questions.

Après de premières observations pour déterminer des points communs et divergences entre les questions du corpus (ou cibles), nous avons fait plusieurs essais d'annotation, en nous demandant à chaque fois ce que le locuteur cherchait à exprimer à travers sa question. Ces essais d'annotation étaient entrecoupés de phases de discussion pour affiner les étiquettes, durant lesquelles se sont posées les questions de leur généricité et de leur applicabilité à l'ensemble du corpus. Un des objectifs était de parvenir à dégager des critères distinctifs pour chacune d'elles. Ces étapes nous ont permis de construire une typologie des intentions dans les questions se divisant en deux parties : la première s'intéressant au type de réponse attendu pour chacune des questions et donc à leur aspect explicite, et la seconde portant sur l'intention exprimée par le locuteur à travers sa question et donc plutôt sur son aspect implicite.

Pour l'explicite, nous avons défini deux classes permettant de catégoriser les questions en fonction du type de réponse attendue. Une question peut donc être :

— une *demande d'accord* (une interrogation totale), par exemple « je peux mettre ça là ? » ;
— une *demande d'information* (une interrogation partielle), par exemple « c'est où Saint Raphaël ? ».

L'implicite suppose des classes plus complexes à déterminer puisqu'elles nécessitent une interprétation : il s'agit d'un message non littéral. Nous avons dégagé trois classes représentant l'intention exprimée par le locuteur produisant la question (*avis*, *volonté* et *doute*), sur lesquelles nous nous concentrerons pour la tâche de classification automatique qui constitue la suite de ce travail :

— l'expression d'un *avis* correspond à un jugement positif ou négatif qui n'implique pas nécessairement une action de la part d'un des locuteurs (« ils sont vraiment bêtes hein ? ») et dont les marqueurs possibles peuvent être des adjectifs, des adverbes ou des locutions verbales qui aident à l'expression d'opinions (« je trouve que », « j'adore », « ennuyeux », « honnêtement », etc. ;
— l'expression d'une *volonté* correspond au désir d'une action ou d'un comportement de la part du locuteur lui-même ou de son interlocuteur, implique une réponse correspondant à une action (« tu nous sers à boire mon chéri ? ») et peut avoir des marqueurs tels que des verbes d'action (« aller », « manger », « dormir », « regarder », etc.) ou des verbes exprimant une volonté (« vouloir », « souhaiter », « désirer », etc.) ;
— l'expression d'un *doute* correspond à une mise en doute de ce qui est dit, du caractère vrai ou faux d'une chose, qui n'implique pas nécessairement une action, qui s'apparente à une répétition, à de la surprise, à une demande de confirmation ou de précisions (« sûrement

1. http://eslo.huma-num.fr

non ? ») et dont les marqueurs possibles peuvent être la répétition d'une partie du contexte précédent, la présence de mots interrogatifs (« qui », « quel », « quoi », etc.) ou encore la présence d'adverbes d'affirmation/modaux (« sûrement », « peut-être », « probablement », etc.).

Nous avons utilisé les définitions de cette typologie pour annoter les 3647 questions de notre corpus, la prise de décision finale pour chacune d'entre elles s'appuyant également sur son contexte, c'est-à-dire sur les dix tours de parole la précédant et la suivant, celui-ci étant peu compréhensible avec une longueur inférieure. Les questions dites « phatiques » ont été considérées comme peu pertinentes pour cette tâche et n'ont donc pour la plupart pas été annotées.

Pour s'assurer de la fiabilité des définitions de nos classes et de nos annotations, nous avons mis en place un formulaire Google Forms pour l'annotation de quinze questions issues de notre corpus, associé à des consignes qui prévoyaient la prise en considération de leurs contextes. Ce formulaire a récolté vingt-six participations de locuteurs natifs du français, dont douze ayant une formation en linguistique. Nous avons ainsi calculé un accord inter-annotateur avec un Kappa de Cohen entre nos propres annotations et celles de chacun des participants. Nous avons obtenu pour 50% des participations un accord inter-annotateur supérieur à 0,73 pour l'explicite et supérieur à 0,6 pour l'implicite, et pour 25% des participants un accord supérieur à 0,86 pour l'explicite et supérieur à 0,8 pour l'implicite (Landis & Koch, 1977). Ces scores sont plutôt satisfaisants compte tenu de la nature de la tâche demandée qui nécessitait de s'appuyer à la fois sur des éléments concrets tels que l'analyse du contexte des questions et la prise en considération des indices lexicaux présents, mais aussi sur sa propre intuition.

3 Prétraitements et intégration de traits linguistiques

Pour préparer l'étape de classification automatique, nous avons procédé à la lemmatisation et à l'étiquetage morphosyntaxique des données à l'aide de *TreeTagger* (Schmid, 1994), et plus particulièrement à l'aide des fichiers de paramètres du projet *PERCEO, un Projet d'Étiqueteur Robuste pour l'Écrit et pour l'Oral* (Benzitoun *et al.*, 2012), plus adaptés aux données orales et donc aux transcriptions d'enregistrements. Nous avons ensuite vectorisé avec un TF-IDF les questions et leurs contextes pour obtenir une représentation du texte utilisable en entrée de l'algorithme de classification. En comparaison d'autres types de vectorisations tels que *word2vec* avec les modèles *CBOW* et *Skip-Gram* ou les vecteurs préentraînés de *Flair*, le TD-IDF nous a permis d'obtenir des résultats de classification légèrement meilleurs.

La mise en place de notre typologie et les phases d'annotation nous ont fait remarquer des informations lexicales permettant de distinguer nos différentes classes. Elles constituent donc des indices que nous avons regroupés sous la forme de six lexiques, dont cinq construits spécifiquement pour ce travail :

— les *verbes de parole* : *dire, demander, proposer, suggérer, expliquer,* etc. ;
— les *verbes de mouvement*, issus de la ressource lexicale DinaVmouv, (Stosic & Aurnague, 2017) : *accrocher, suivre, s'asseoir, remplir, parcourir,* etc. ;
— les *mots interrogatifs* : *qui, combien, lequel, pourquoi, quand,* etc. ;
— les *interjections*, correspondant à la liste d'interjections présentées dans le guide de transcription des corpus ESLO : *mouais, hein, miam, bof, bah,* etc. ;
— les *sentiments* : *apprécier, ravi, haine, nul, inquiéter,* etc. ;

— les *adverbes et adjectifs modaux* : *vraiment, impossible, certainement, peut-être, vrai,* etc..

En plus des informations liées aux lexiques (la fréquence d'apparition des mots de chacun des lexiques dans chaque cible et contexte gauche ou droit), nous avons intégré à nos données le nombre de mots et le nombre de caractères par cible et par contexte.

4 Classification automatique : expériences et résultats

Après avoir équilibré les classes pour éviter un biais dans les résultats (286 occurrences par classe soit un total de 858 occurrences), nous avons implémenté l'algorithme *Random Forest* et utilisé la méthode de validation croisée *k-fold cross-validation* ainsi que la fonction *GridSearchCV* de *sklearn* pour choisir les valeurs optimales pour des hyperparamètres donnés, dans notre cas *n_estimators*, *criterion* et *bootstrap*. La performance du modèle a ensuite été évaluée avec une moyenne des mesures de précision, rappel et F-mesure obtenues pour chaque groupe.

		EXPÉRIENCES								
		1	2	3	4	5	6	7	8	9
TRAITS	vecteur c.	×	×	×	×	×	×	×	×	×
	vecteur g.		×							
	vecteur d.		×							
	POS tagging c.			×	×	×	×	×	×	
	POS tagging g.									
	POS tagging d.									
	sentiments c.				×	×	×	×	×	×
	sentiments g.					×				
	sentiments d.					×				
	interjections c.				×	×	×	×	×	×
	interjections g.					×				
	interjections d.					×				
	interrogatifs c.				×	×	×	×	×	×
	interrogatifs g.					×				
	interrogatifs d.					×				
	mouvement c.				×	×	×	×	×	×
	mouvement g.					×				
	mouvement d.					×				
	parole c.				×	×	×	×	×	×
	parole g.					×				
	parole d.					×				
	modaux c.				×	×	×	×	×	×
	modaux g.					×				
	modaux d.					×				
	explicite						×	×	×	×
	nb. mots c.							×	×	×
	nb. mots g.								×	×
	nb. mots d.								×	×
	nb. caractères c.							×	×	×
	nb. caractères g.								×	×
	nb. caractères d.								×	×
RÉSULTATS	Précision	0,622	0,492	0,618	0,632	0,6	0,63	0,621	0,631	0,613
	Rappel	0,612	0,493	0,612	0,623	0,592	0,622	0,617	0,624	0,606
	F-mesure	0,611	0,489	0,61	0,622	0,59	0,62	0,616	0,622	0,603

TABLE 1 – Récapitulatif des expériences et résultats

Nous avons testé l'algorithme avec plusieurs combinaisons de traits présentées dans le tableau 1, un récapitulatif global de l'ensemble des résultats. Chaque expérience correspond à un ensemble choisi de traits (cochés dans le tableau). Nous voyons par exemple que toutes les expériences incluent la vectorisation de la cible mais que seules les expériences 8 et 9 incluent le nombre de caractères pour les contextes gauche et droit. Les résultats sont très proches dans l'ensemble comme nous le voyons pour les expériences 4, 6 et 8 pour lesquelles nous obtenons des scores de F-mesure avoisinant 0,62. Certaines mesures semblent cependant se détacher, telles que celles de l'expérience 2 qui sont inférieures à 0,5 ou celles de l'expérience 5 inférieures à 0,6, ces deux expériences prenant en compte des traits liés aux contextes de la question cible (la vectorisation dans le premier cas et la présence de mots des lexiques dans le second).

5 Discussion

Une première observation concerne la baisse des performances lorsque des traits liés aux contextes sont ajoutés, comme dans les expériences 2 et 5. Pour vérifier l'importance du contexte des cibles, nous avons reproduit les expériences 2 et 8 avec deux puis cinq tours de parole avant et après la question. Les scores obtenus pour l'expérience 2 (qui ne comprenait que la vectorisation des cibles et de leurs contextes) montrent une amélioration des performances lorsqu'il y a moins de tours de parole dans chaque contexte, ce qui peut s'expliquer par la trop grande quantité d'informations non pertinentes rapportées par la vectorisation des contextes lorsque ceux-ci sont plus larges. En revanche, nous observons pour l'expérience 8 (qui comprenait la vectorisation des cibles et des informations lexicales) que les scores sont meilleurs lorsque la fenêtre de contexte est plus grande : lorsque la vectorisation des contextes est utilisée en entrée du classifieur, elle semble rapporter beaucoup de bruits et faire baisser les performances, tandis que lorsque l'algorithme s'appuie sur des informations lexicales plutôt que sur la vectorisation, il est plus performant avec plus de contextes.

	Expérience 2			Expérience 8		
Nb tours de parole par contexte	2	5	10	2	5	10
Précision	0,575	0,54	0,492	0,588	0,591	0,631
Rappel	0,572	0,536	0,493	0,584	0,587	0,624
F-mesure	0,571	0,535	0,489	0,582	0,585	0,622

TABLE 2 – Comparaison des expériences 2 et 8 en fonction du nombre de tours de parole par contexte

Pour interpréter les résultats, nous nous sommes concentrés sur l'expérience 8, une des expériences ayant eu les meilleurs scores et prenant en compte le plus de traits. Nous avons calculé la médiane du nombre de mots et du nombre de caractères pour chaque cible, contexte gauche et contexte droit. Nous avons également calculé le nombre de mots appartenant aux différents lexiques présents en moyenne dans chaque cible (tableau 3).

	A→A	A→D	A→V	V→V	V→A	V→D	D→D	D→A	D→V
méd. mots c.	6	5	5	6	5	5	3	5	6
méd. mots g.	96	92	93.5	87.5	95	88	92	85	101
méd. mots d.	97	97	81	88.5	97	88	84	94	92
méd. caractères c.	21	19	21	21	19	17	11	17	21
méd. caractères g.	331	326	330,5	312,5	322	302	327	312	341
méd. caractères d.	344	338	301	310	345	317	299	331	309,5
vb. de parole	0,03	0,04	0,07	0,04	0,04	0,02	0,05	0,08	0,07
vb. de mouvement	0,06	0,07	0,12	0,19	0,12	0,1	0,08	0,11	0,16
mots interrogatifs	0,34	0,53	0,47	0,33	0,35	0,37	0,58	0,46	0,59
interjections	0,54	0,27	0,25	0,18	0,37	0,3	0,08	0,3	0,24
sentiments	0,41	0,09	0,12	0,03	0,09	0,13	0,04	0,16	0,1
modaux	0,31	0,07	0,09	0,13	0,26	0,13	0,09	0,11	0,07

TABLE 3 – Mesures pour les prédictions des classes avis (A), volonté (V) et doute (D)

Pour la classe *doute*, la médiane pour le nombre de caractères des questions est de 11. Il s'agit d'une longueur courte par rapport aux questions de la classe *doute* ayant été mal classées en *avis* ou en *volonté* pour lesquelles nous trouvons respectivement des médianes de 17 et de 21. Pour cette classe, la présence de mots interrogatifs est également significative puisqu'elle est plus élevée en moyenne qu'ailleurs (0,58). Les questions de cette classe ayant été mal classées en *volonté* contiennent en moyenne 0,16 verbes de mouvement, un chiffre qui se rapproche de la moyenne du nombre de verbes de mouvement présents dans les bonnes prédictions de *volonté* (0,19).

Pour la classe *avis*, les cibles bien classées ont en moyenne une forte présence de mots appartenant aux lexiques des sentiments (0,41), des interjections (0,55) et des modaux (0,31) comparé aux autres classes et aux cibles mal classées de la classe *avis*. Ces dernières, lorsqu'elles sont mal classées dans *volonté* ont en moyenne 0,12 mots appartenant au lexique des sentiments, 0,25 mots appartenant aux interjections et 0,09 mots étant des modaux. Ceci se confirme également pour les cibles *avis* mal classées en *doute*, pour lesquelles nous voyons par ailleurs que le nombre de mots interrogatifs présents en moyenne est de 0,53, un chiffre qui se rapproche de celui obtenu pour les cibles *doute* bien classées qui est de 0,58.

Enfin, pour la classe *volonté*, nous remarquons une plus forte présence de verbes de mouvement qu'ailleurs, avec une moyenne de 0,19 pour les cibles bien classées en *volonté*. Ce chiffre est plus bas pour les cibles mal classées en *doute* (0,1) et se rapproche notamment de la moyenne du nombre de verbes de mouvement pour les bonnes prédictions de *doute* qui est de 0,08. Pour les cibles de *volonté* mal classées en *avis*, nous observons une présence plus forte de mots des lexiques d'interjections (0,37) et de modaux (0,26), qui sont des caractéristiques de la classe *avis*.

6 Perspectives et conclusions

Le but de cette étude était de parvenir à définir et à identifier les intentions implicites exprimées par les locuteurs. Il s'agissait plus précisément d'étudier les spécificités linguistiques permettant de caractériser l'aspect non-littéral de questions issues du discours oral spontané, en prenant en considération leurs contextes. Après avoir constitué un corpus de référence dont nous avons évalué les étiquettes créées à partir de nos recherches et de notre typologie, nous avons procédé à une tâche de classification automatique avec Random Forest pour laquelle nous avons également intégré des traits linguistiques à nos données originales. Les scores obtenus pour ces expériences sont supérieurs au hasard mais restent tout de même améliorables puisqu'ils tournent pour la plupart autour de 0,6. Ces résultats semblent constituer un plafond et rendent ainsi compte de la complexité de la tâche effectuée. Ils restent cependant encourageants lorsqu'ils sont mis en regard avec les accords inter-annotateurs obtenus lors de l'évaluation des annotations manuelles, bons ou excellents pour la moitié d'entre eux.

Plusieurs idées et perspectives pourraient améliorer ces résultats :

— l'augmentation de la taille du corpus de référence pour permettre un meilleur apprentissage ;
— l'identification de nouveaux traits discriminants ;
— l'enrichissement de traits existants (par exemple l'ajout de nouvelles entrées à nos lexiques) ;
— l'utilisation de techniques d'apprentissage profond.

Une perspective à plus long terme de ce travail serait d'adapter cette étude sur les intentions implicites à d'autres types d'énoncés que les questions ou encore à d'autres types de données que les transcriptions de discours oral et spontané.

Références

ALEXANDERSSON J., BUSCHBECK-WOLF B., FUJINAMI T., KIPP M., KOCH S., MAIER E., REITHINGER N., SCHMITZ B. & SIEGEL M. (2011). *Dialogue acts in VERBMOBIL-2*, volume 1998. 2. ed. édition. Report 226.

ALLEN J. & CORE M. (1997). Draft of DAMSL : Dialog Act Markup in Several Layers. https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/.

ALLEN J. F. & PERRAULT C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, **15**(3), 143–178.

ANDERSON A. H., BADER M., BARD E. G., BOYLE E., DOHERTY G., GARROD S., ISARD S., KOWTKO J., MCALLISTER J., MILLER J. *et al.* (1991). The HCRC map task corpus. *Language and speech*, **34**(4), 351–366.

AUSTIN J. L. (1962). *How to do things with words*. William James Lectures. Oxford University Press.

BAUDE O. & DUGUA C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*, **10**, 99–118.

BENZITOUN C., FORT K. & SAGOT B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*. HAL : hal-00709187.

BORITCHEV M. & AMBLARD M. (2018). Coffee or tea ? Yes. In L. PRÉVOT, M. OCHS & B. FAVRE, Éds., *Proceedings of the 22nd workshop on the Semantics and Pragmatics of Dialogue*. HAL : hal-01922137.

BORITCHEV M. & AMBLARD M. (2019). A compositional view of questions. In *Proceedings of the 2019 Workshop on Widening NLP*. HAL : hal-02269603.

BUNT H. (2005). A framework for dialogue act specification. *Proceedings of SIGSEM WG on Representation of Multimodal Semantic Information*.

BUNT H. (2009). The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts (EDAML 2009)*, p. 13–24.

CHEN Z., LIU B., HSU M., CASTELLANOS M. & GHOSH R. (2013). Identifying intention posts in discussion forums. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics : human language technologies*, p. 1041–1050.

DUCROT O. (1972). *Dire et ne pas dire : principes de sémantique linguistique*. Savoir. Hermann.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral « disponible » : le corpus d'Orléans 1 1968-2012. *Traitement Automatique des Langues*, **52**(3). HAL : halshs-01163053.

GINZBURG J. (1996a). Dynamics and the semantics of dialogue. In J. SELIGMAN & D. WESTERSTÅHL, Éds., *Logic, Language and Computation, Volume 1*, volume 58 de *CSLI Lecture Notes*, chapitre 15. CSLI Publications.

GINZBURG J. (1996b). Interrogatives : Questions, facts and dialogue. In *The handbook of contemporary semantic theory. Blackwell, Oxford*, p. 359–423. Citeseer.

JURAFSKY D., SHRIBERG E. & BIASCA D. (1997). *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13*. Rapport interne 97-02, Institute of Cognitive Science University of Colorado, Boulder. https://www.colorado.edu/ics/technical-reports-1990-1999, Part 1, Part 2.

KAROUI J., BENAMARA F. & MORICEAU V. (2019). *Détection automatique de l'ironie : Application à la fouille d'opinion dans les microblogs et les médias sociaux*. ISTE Group.

KAROUI J., GILLES N. A., BENAMARA ZITOUNE F. & BELGUITH L. (2014). Le langage figuratif dans le web social : cas de l'ironie et du sarcasme. In *Workshop Fouille d'opinion dans le Web social*, Lyon, France. HAL : hal-01686491.

KERBRAT-ORECCHIONI C. (1986). *L'implicite*. Paris : A. Colin.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

MAUDET N., MULLER P. & PRÉVOT L. (2004). Tableaux conversationnels en SDRT. *Actes de TALN*.

MOLDOVAN C., RUS V. & GRAESSER A. (2011). Automated speech act classification for online chat. In S. VISA, A. INOUE & A. RALESCU, Éds., *Proceedings of The 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011*, p. 23–29. http://ceur-ws.org/Vol-710/paper22.pdf.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

SEARLE J. (1975). Indirect Speech Acts. In P. COLE & J. L. MORGAN, Éds., *Syntax and Semantics, Volume 3 : Speech Acts*. Academic Press.

STOSIC D. & AURNAGUE M. (2017). DinaVmouv : Description, INventaire, Analyse des Verbes de mouvement. An annotated lexicon of motion verbs in French. Ressource lexicale, HAL : hal-01979613.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Réduire l'effort humain d'amélioration des ressources lexicales grâce aux inférences

Nadia Bebeshina[1, 2] Mathieu Lafourcade[1]
(1) LIRMM, 860 rue de St Priest, 34095 Montpellier, France
(2) Praxiling, Université Paul Valéry, Route de Mende 34199, Montpellier, France
nadia.clairet@gmail.fr, mathieu.lafourcade@lirmm.fr

RÉSUMÉ

Les inférences translingues représentent une piste intéressante pour la construction des ressources lexico-sémantiques multilingues. Cependant, la validation des éléments candidats nécessite un effort humain considérable. Nous décrivons une façon de construire des ressources lexico-sémantiques via des inférences monolingue et translingue. Son intérêt principal consiste à implémenter dans le contexte d'une ressource lexico-sémantique multilingue une approche où le processus de construction est un processus auto-apprenant car l'évaluation participe à la construction de celle-ci.

ABSTRACT

Reducing the Knowledge Resource Enhancement Human Effort through Inferences.

The inference based knowledge resource enhancement mechanism allows building lexical semantic resource incrementally and, presumably, minimises the human effort necessary for the resource building. Unfortunately, the knowledge parts obtained by inference have to be semi-automatically double checked to avoid the propagation of errors. The overall human effort in the inference process may be very important. In the present paper, we describe a resource building pipeline based on monolingual inference and cross-lingual inference. Its main contribution is proposing a method where the resource building process appears as a self learning process.

MOTS-CLÉS : inférence, ressource multilingue, relations sémantiques, validation.

KEYWORDS: inference, multilingual knowledge resource, semantic relations, validation.

1 Introduction

Les ressources lexico-sémantiques sous forme de graphe sont fréquemment utilisées pour l'analyse sémantique et la désambguïsation. Dans le contexte de construction de ces ressources, inférer de nouveaux éléments à partir des éléments existants dans la ressource semble permettre de réduire l'effort humain nécessaire à celle-ci. Cependant, les éléments candidats doivent être validés de façon supervisée afin d'éviter la propagation des erreurs. Ainsi, la part de l'effort humain nécessaire à la mise en place et au fonctionnement des méthodes à base d'inférence endogène peut être très importante (élaboration des règles d'inférence, validation, mise à jour des éléments existants). Après avoir introduit l'état de l'art des techniques de construction des ressources lexico-sémantiques, nous détaillons les ressources qui ont servi pour l'expérience. Puis, nous décrivons notre méthode d'inférence et d'évaluation par inférence des éléments candidats et ses résultats.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 146–154.
HAL : hal-02784762.

2 État de l'art

Construire ou améliorer le ressources de connaissance multilingues peut s'appuyer sur un mécanisme d'inférence qui permet d'exploiter les différentes partitions monolingues de ces ressources et les liens entre ces partitions. Ainsi, dans le cadre des bases de connaissance fondées sur les entités et les faits sur les entités telles que NELL (Carlson *et al.*, 2010), plusieurs auteurs ont proposé d'établir les équivalences entre les entités et les relations des différentes partitions. Les auteurs dans (Hernández-González *et al.*, 2017) ont proposé de fusionner le différentes éditions monolingues de NELL. (Nickel *et al.*, 2015) détaillent l'apprentissage statistique dans le contexte multilingue en mettant l'accent sur la transitivité et les contraintes de type de relation. De même que les auteurs dans (Wang *et al.*, 2015), ils fondent leur méthode sur des bases à large couverture telles que NELL, KnowItAll (Etzioni *et al.*, 2005), YAGO (Suchanek *et al.*, 2007) ou DeepDive (Shin *et al.*, 2015).

Des travaux ont également été menés pour enrichir les modèles des ressources telles que Princeton WordNet (Fellbaum, 1998) et construire des ressources ayant une expressivité comparable dans d'autres langues, plus pauvres en ressources telles que le basque (Agirre *et al.*, 2002). Ces auteurs mettent l'accent sur l'importance de l'alignement concept à concept et mot à mot. Le projet BabelNet (Navigli & Ponzetto, 2012) a été la première expérience à grande échelle de construction des ressources non supervisée et intégrant plusieurs ressources pré-existantes et construites manuellement. Les processus d'inférence endogènes à base de règles a été étudié notamment par (Zarrouk, 2015) et (Ramadier, 2016) dans le cadre des travaux sur la construction endogène du réseau lexico-sémantique pour le français RezoJDM. Leurs méthodes reposent sur les relations sémantiques et annotations (méta-informations) de ces relations déjà présentes dans cette ressource pour proposer de nouveaux éléments grâce aux inférences déductives, inductives, abductives (exploitant la similarité sémantique) et fondées sur les raffinements de sens. (Gelbukh, 2018) a introduit un mécanisme d'inférence comparable pour enrichir une base de connaissances collocationnelle et proposer de nouvelles collocations par abduction à partir de celles déjà présentes dans la base. Cette méthode utilise WordNet pour calculer la similarité sémantique.

3 Ressources

RezoJDM (Lafourcade, 2007) est un réseau lexico-sémantique du français construit par peuplonomie dont, en particulier, les jeux avec un but tels que JeuxDeMots [1] ainsi que d'autres jeux [2]. Il s'agit d'un graphe orienté, typé et pondéré. Au moment où nous écrivons RezoJDM contient 4,2 millions de termes (noeuds) et 310 millions de relations (arcs). Le modèle de RezoJDM offre trois caractéristiques importantes pour notre expérience. Premièrement, il contient des relations à poids négatif (relations sémantiquement fausses) ce qui rend possible la détection des relations fausses par inférence. Deuxièmement, les termes de ce réseau peuvent être raffinés afin de représenter les distinctions de sens. Troisièmement, les relations peuvent être annotées et comporter plusieurs annotations. Nous nous référerons à RezoJDM comme *ressource de référence*.
RLSM (Bebeshina-Clairet, 2019) est un réseau lexico-sémantique multilingue (français, anglais, russe, espagnol) avec un pivot interlingue. Au moment où nous écrivons RLSM contient 821 781 noeuds (termes) and 2 231 197 arcs (relations). Il a été construit pour les domaines de la cuisine et

1. http://www.jeuxdemots.org
2. http://imaginat.name/JDM/Page_Liens_JDMv2.html

de la nutrition, mais contient également de la connaissance générale conformément à l'hypothèse sur la non-séparation de la connaissance générale et de spécialité vérifiée par (Ramadier, 2016). Inter-opérable avec RezoJDM (types de relations, raffinements de sens), le RLSM est un graphe orienté, typé et valué. Il contient k sous-graphes correspondant à chacune des k langues du réseau ainsi qu'un sous-graphe spécifique, le pivot interlingue. *Valué* se réfère au fait que le RLSM n'a pas été construit par peuplonomie et a nécessité d'autres moyens de valuation complémentaires à la pondération. L'ajout des méta-informations et l'attribution des scores de confiance ont été exploités pour permettre de renforcer ou non la relation entre les termes.

4 Expérience

4.1 Observations générales

L'expérience s'est appuyée sur les observations suivantes : (1) dans certaines langues les informations sémantiques peuvent être découvertes à partir des traits morpho-syntaxiques tandis que dans d'autres langues la même information est plus difficile à obtenir ; (2) dans le contexte des ressources lexico-sémantiques multilingues, une partition lexicalisée (sous-graphe d'une langue) peut fournir l'information sémantique manquante à ses autres partitions ; (3) dans le cas de deux ressources lexico-sémantiques interopérables, les composants (par exemple, termes, relations sémantiques) d'une ressource peuvent être validés en s'appuyant sur le contenu réel ou inférable de l'autre ressource. Compte tenu de ces observations, nous supposons que les relations sémantiques extraites depuis le corpus de langue riche en traits sémantico-syntaxiques (comme la langue russe) peuvent être exploitées pour enrichir une ressource sémantique dans une autre langue. La condition serait alors l'existence d'un lien (traduction, pivot) entre la ressource à évaluer et la ressource de référence. L'interopérabilité du RLSM avec la ressource plus riche et plus stable RezoJDM a permis la mise en place de la procédure d'évaluation automatique. L'exploitation des résultats de la procédure d'évaluation automatique rapproche l'expérience du processus auto-apprenant. L'expérience comprend trois étapes : (1) **identification et extraction des relations sémantiques** à partir du corpus de textes en russe pré-étiqueté en partie de discours en utilisant un ensemble de règles ; (2) **inférence translingue** des relations extraites et création de nouvelles relations dans le sous-graphe français du RLSM à partir des relations intégrées dans le sous-graphe russe suite à l'extraction ; (3) **validation translingue des relations inférées** grâce à la ressource de référence RezoJDM.

4.2 Extraction

Corpus et portée de l'extraction. L'extraction a été effectuée à partir d'un corpus des instructions de cuisine (2 473 654 mots) collecté sur le Web [3]. Ce corpus a été étiqueté en utilisant le parseur *Russian Malt parser* (Sharoff & Nivre, 2012). L'extraction a concerné les types de relations sémantiques suivants : (a) *caractéristique* avec quelques distinctions entre les caractéristiques de composition, et des caractéristiques qualitatives pouvant être observées à travers les traits morpho-syntaxiques en russe ; (b) *manière* avec distinctions observées entre les expressions adverbiales liées à l'instrument (décorer *avec des fruits tranchés*), manière partie-tout (*couper en cubes*) et d'autres réalisations ; (c) *lieu* avec distinction entre les lieux de type « surface plane » et les lieux de type « contenant ».

3. Principalement, à partir du portail https://www.gotovim.ru/

Règles. Les règles pour l'extraction des caractéristiques reposent sur une série de traits propres aux adjectifs en russe. Comme remarqué par (Corbett, 2004), les adjectifs du russe couvrent les aspects typologiques inhabituels. Un sous-ensemble de règles (*carac-r1*) correspond aux adjectifs canoniques qui donnent des caractéristiques de dimension, âge, valeur, et couleur. Un autre sous-ensemble de règles (*carac-r2*) sur les modèles de dérivation saillants pour les adjectifs du russe tels que décrits par (Zemskaja *et al.*, 1981). Plus spécifiquement, nous avons exploré les adjectifs dérivés des noms et verbes. La dérivation à partir de noms peut révéler la relation partie-tout (composant ou membre) tandis que la dérivation à partir des verbes peut révéler les changements d'état. Il est à remarquer que le rôle des adjectifs dans la représentation des caractéristiques typiques nécessaires catégorisation ont fait objet de nombreux travaux, notamment dans le cadre de la terminologie fondée sur les cadres, *Frame-Based Terminology* (Faber & Cabezas-García, 2019). Dans (Altmanova *et al.*, 2018), les adjectifs renfermant des relations sémantiques sont utilisés pour extraire d'autres adjectifs porteurs des mêmes types de relation (fonction, composition, localisation, cause et forme).

Sur la base des appariements entre les structures de surface (lexicalisation du mouvement) et les composants du sens nécessaires pour représenter le mouvement (objet mouvant, point de référence, chemin) la typologie de (Talmy, 1985) distingue deux groupes essentiels de langues : les langues S (langues qui exploitent des mots-satellites)[4] et les langues V où les verbes de mouvement sont directionnels et l'indication de manière est optionnelle comme détaillé dans (Strömqvist & Verhoeven, 2004). Dans notre expérience il s'agit de l'acquisition des relations de manière et trajectoire à partir d'une *langue S* pour les transférer vers la *langue V* sous forme de relations sémantiques explicites.

Les prémisses des règles sont des traits syntaxiques, grammaticaux et morphologiques tels que les cas des noms, « satellites » (préfixes, suffixes, mots-outils). La conclusion est la création d'une relation candidate dans le sous-graphe russe du RLSM.

Exemples des règles :
(*)`if X Y` (*contexte*) `and X Verb` (*prémisse 1*) `and Y Nom:CaseAccusative` (*prémisse 2*)
`X r_object Y` (*conclusion*)
(**)`if X` "na" `Y and X Verb and Y Noun:CaseLocative`
`X r_has_part::essential component Y` (*conclusion - une relation annotée*)

Pour donner un exemple d'une règle fondée sur les traits morphologiques, nous pouvons citer le cas des adjectifs avec une suffixe `[-yann]` qui permettent d'extraire les informations sur la matière (bois, verre) qui constitue l'objet.

L'ensemble des règles comporte moins de 30 règles pour les types de relations abordés par l'expérience. L'étape d'extraction reste dépendante de la sortie du parseur.

Les types d'information sémantique extraits sont présentés dans le tableau 1. Une mesure de confiance est attribuée à chaque instance de règle. Initialement, le score c est fixé à $c = 1$. Au fur et à mesure que les relations inférées sont évaluées grâce à la ressource de référence, ce score évolue et contribue à refléter la fiabilité de la règle *pour la langue cible*. Il est affecté par le nombre de relations justes inférées grâce à cette règle, le nombre de sens des termes polysémiques (dans le cas d'une relation source dont l'inférence ascendante produit n relations interlingues, la confiance pour chacune de ces relations est de $c_{int} = \frac{c_s}{n}$ *etc.*) et le statut suite à l'évaluation.

4. Finno–ougriennes, germaniques, sino–tibétaines et slaves.

type-règle	**#ext**	**%ext**
manner-r1 (adverbes)	8 229	25,4%
manner-r2 (instrument, forme : *en cubes*, *avec les mains*)	1 879	5,8%
manner-r3 (avec peu d'intensité)	295	0,9%
manner-r4 (action récurrente)	154	0,4%
carac-r1	12 488	38%
carac-r2 (dérivation nominale, *partie-tout*)	543	1,6%
carac-r3 (dérivation verbale, *état*)	756	2%
place-r1	5 679	17,5%
place-r2 (lieu plan ou contenant)	2 298	7%
Total	32 321	100%

TABLE 1 – Détails concernant les règles d'extraction et le nombre des instances obtenues **#ext**.

4.3 Inférence et évaluation

Pivot « interlingue ». Les schémas d'inférence dans le RLSM sont fondées sur l'utilisation du pivot interlingue. Ce pivot a été amorcé en tant que pivot naturel à partir des liens de traduction présents dans DBnary (Sérasset, 2014) (édition anglaise et russe) et évolue actuellement vers le pivot interlingue. La ressource DBNary est orientée vers l'alignement par sens (Tchechmedjiev, 2016). Le pivot a été enrichi grâce à des processus d'acquisition exogènes (à partir des ressources monolingues et multilingues pré-existantes) ainsi que d'extractions à partir des textes de spécialité (Bebeshina-Clairet, 2019). Le pivot peut être vu comme l'union de tous les sens disponibles dans le RLSM. Les termes interlingues qui appartiennent au pivot sont reliés aux termes lexicalisés grâce à la relation typée *r_covers*. Un terme lexicalisé peut avoir plusieurs termes couvrants. Un lien translingue correspond à un chemin qui traverse le pivot, processus également détaillé dans (Bebeshina-Clairet & Lafourcade, 2019). Ainsi, les relations extraites à partir des textes en russe sont créées dans le sous-graphe du russe. Lorsque le terme source t_s et le terme cible t_c d'une relation r_type_{ru} ainsi créée sont couverts par le pivot interlingue, la relation r_type_{inter} est inférée dans le pivot et relie le terme couvrant du t_s à celui du t_c. il est ensuite possible de produire des relations candidates en français afin de permettre la validation.

Inférence. Le processus d'inférence actuel détaillé dans (Bebeshina-Clairet & Lafourcade, 2019) combine la phase ascendante (langue source → pivot) et la phase descendante (pivot → langue cible). Pendant la phase ascendante, les relations sémantiques du sous-graphe de la langue source sont inférées dans le pivot grâce à un ensemble de contraintes qui dépendent du type de relation à inférer. Pour former une relation typée « caractéristique », il est nécessaire que le terme source soit un nom tandis que le terme-cible soit un adjectif. Vérifier si les termes sont sémantiquement liés (triangulation) ainsi que le calcul de similarité sémantique sont également utilisés. Pendant la phase descendante, le processus d'inférence est complété par le processus d'évaluation fondé sur la recherche et l'inférence dans la ressource de référence (RezoJDM).

Outil d'évaluation. Premièrement, les relations inférées sont ajoutées dans le sous-graphe français du RLSM. Deuxièmement, l'évaluation est effectuée grâce à l'outil Helix [5] destiné à vérifier la

5. http://www.jeuxdemots.org/rezo-ask.php

présence des relations et à compléter RezoJDM. Troisièmement, le retour de l'évaluation de la relation candidate affecte le poids et le statut des relations ayant contribué à son inférence (relations des sous-graphes russe, interlingue et français).

Les résultats d'inférence ascendante reflètent la couverture de la langue source par le pivot. L'inférence descendante montre la couverture de la langue cible par le pivot en ce qui concerne un type de relation particulier. Grâce à l'évaluation, cette phase descendante permet l'acquisition des relations sémantiquement fausses qui sont maintenues dans le RLSM afin de prévenir leur acquisition ultérieure par d'autres moyens (processus exogènes, inférences). Les résultats d'inférences en deux phases sont listés dans la table 2a. Les processus d'inférence génèrent de nombreuses relations candidates à cause de la présence des raffinements de sens dans le RLSM, (en particulier, les distinctions de sens acquises à partir de WordNet (Fellbaum, 1998), RezoJDM, ConceptNet (Speer & Havasi, 2012) qui sont reliées à des termes polysémiques mais non raffinés (notamment, dans le sous graphe russe). Lorsque les relations candidates sont automatiquement validées, les termes en langue source sont désambiguïsés car un poids négatif est attribué aux relations qui concernent le sens erroné $t > s_1$ du terme t à un terme p lorsqu'un autre sens de t, $t > s_2$, a une relation évaluée comme *vrai* avec p.

type	**#ext**	**#asc**	**#desc**
carac-r1	12 488	22 631	40 762
carac-r2	543	3 371	2 983
carac-r3	756	1 318	440
manner-1	8 229	6 000	8 929
manner-2	1 879	1 413	1 655
manner-3	295	283	192
manner-4	154	177	169
place-r1	5 679	2 548	2 969
place-r2	2 298	5 551	6 267
Total	32 321	43 292	66 366

(a) Nombre des relations extraites (**#ext**), inférences ascendante (**#asc**) et descendante (**#desc**).

type	**#inf**$_d$	**#true**$_r$	**#true**$_{inf}$	**#und**
carac-r1	40 762	815	2 454	37 493
carac-r2	2 983	64	283	2 636
carac-r3	440	44	264	132
manner-r1	8 929	35	2 419	6 475
manner-r2	1 655	17	83	1 555
manner-r3	192	13	27	152
manner-r4	169	6	13	150
place-r1	2 969	74	453	2 442
place-r2	6 267	49	3 181	2 777
Total	66 366	1 117	9 177	53 812

(b) Évaluation des relations inférées. Par souci de simplicité, seulement les relations acceptées par recherche (**#true**$_r$) ou inférence (**#true**$_{inf}$) et les relations « indécidables » (**#und**) soit « ne sait pas », « terme absent », « faux par inférence » sont listées.

TABLE 2 – Inférence et évaluation des relations

Évaluation et ajustement. Le processus d'évaluation est un processus fondé sur l'inférence monolingue dans le réseau lexico-sémantique RezoJDM destiné à améliorer automatiquement le RLSM et le processus d'extraction des relations sémantiques. L'évaluation avec l'outil Helix retourne les valeurs suivantes pour les relations testées : (1) « vrai » (la relation est présente dans RezoJDM) ; (2) « vrai par inférence » (la relation est inférable dans le RezoJDM) ; (3) « ne sait pas » (la relation est absente et le processus d'inférence ne parvient pas à valider la relation, cette réponse est quasi-équivalente à « faux ») ; (4) « faux » (la relation a un poids négatif) ; (5) « faux par inférence » (la relation est inférable et fausse dans RezoJDM) ; (6) « terme inconnu » (terme source ou cible de la relation à tester est absent de RezoJDM). L'expérience s'intéresse principalement aux relations validées par recherche, par inférence et aux relations invalidées (classification à 3 termes). Une analyse ultérieure plus approfondie des relations candidates invalidées se situera dans un cadre plus large qui concerne la ressource de référence.

Les résultats d'évaluation montrent le rôle significatif d'inférence monolingue. Les relations sémantiques qui exploitent les propriétés telles que *plan*, *contenant* (place-r2), action faite « un par un » (manner-r4) peuvent ne pas être explicitement présentes dans la ressource de référence mais elles sont inférables. Le pourcentage des relations validées automatiquement comme « vraies » correspond au pourcentage généralement observé dans le contexte d'évaluation humaine : **#true**$_{inf}$ indiqué dans la table 2b est d'environ 5% à 10% des relations inférées par inférence translingue ce qui est similaire aux résultats de validation manuelle dans (Zarrouk, 2015) et (Bebeshina-Clairet, 2019).

5 Discussion

L'expérience a été menée sur la base des réseaux lexico-sémantiques et s'est appuyée sur des processus d'inférence conçus pour ce type de ressource. Est-elle facile à reproduire ? Pour une paire de langues donnée, la réponse à cette question dépend de la disponibilité de la ressource de référence (RezoJDM dans le cadre de l'expérience) dans une des langues. La ressource de référence est celle où la proportion des termes polysémiques désambiguïsés (raffinement de sens), la couverture et le nombre de relations sémantiques sont importants. Cette ressource doit également disposer d'un nombre suffisant de relations typées entre les termes et les raffinements pour permettre des inférences. Une des affirmations défendues à travers l'expérience présentée est que, pour un réseau lexico-sémantique multilingue, il suffit de disposer d'une ressource de référence dans une seule langue pour améliorer et raffiner l'ensemble de la ressource de façon faiblement supervisée. Cette expérience est ainsi un encouragement à la conception de ressources raisonnablement interopérables avec une ressource de référence afin de pouvoir de réduire l'effort humain de leur construction.

Un intérêt primordial des réseaux lexico-sémantiques pour le TAL est de contenir des termes et des relations entre ces termes représentés de façon explicite aussi bien pour un usage humain que pour un usage machine. La ressource que nous avons exploitée en tant que référence est le RLS le plus important construit pour le français. L'état actuel du RezoJDM permet d'apprécier son modèle polyvalent. Par ailleurs, l'expérience de construction d'un RLSM avec un pivot interlingue a démontré que, dans le contexte d'un RLS multilingue, les différences de granularité de sens (le terme anglais « stew » correspond à la fois à *ragoût* et *pot-au-feu* en français) peuvent aisément être représentés.

6 Conclusion

L'impact d'extraction à base de critères morpho-syntaxiques peut être limité car de nombreux traits sont « polysémiques » (correspondent à plusieurs relations sémantiques potentielles). Les inférences translingues et monolingues apparaissent comme un moyen de résoudre ces ambiguïtés et réduire l'effort humain nécessaire à la construction et l'amélioration des ressources de façon significative. Nos expériences montrent également qu'il suffit de disposer de la ressource de référence (pouvant supporter les processus d'inférence) dans une seule langue couverte par la ressource multilingue pour construire et améliorer les partitions des autres langues de façon moins supervisée. L'expérience d'évaluation a montré que le recours à une base de connaissance importante semble être le bon chemin vers la construction non supervisée des ressources lexico-sémantiques multilingues.

Références

AGIRRE E., ANSA O., ARREGI X., ARRIOLA J. M., DE ILARRAZA A. D., POCIELLO E. & URIA L. (2002). Methodological issues in the building of the Basque WordNet : quantitative and qualitative analysis. In *Proceedings of First International WordNet Conference*, p. 32–40.

ALTMANOVA J., GRIMALDI C. & ZOLLO S. (2018). Le rôle de l'adjectif dans la catégorisation des déchets. *SHS Web of Conferences*, **46**, 05004. DOI : 10.1051/shsconf/20184605004.

BEBESHINA-CLAIRET N. (2019). *Construction d'une ressource termino-ontologique multilingue pour les domaines de la cuisine et de la nutrition*. Theses, Université Paris 13.

BEBESHINA-CLAIRET N. & LAFOURCADE M. (2019). Inférence des relations sémantiques dans un réseau lexico-sémantique multilingue. In *TALN : Traitement Automatique des Langues Naturelles*, volume PFIA, Toulouse, France. HAL : hal-02269113.

CARLSON A., BETTERIDGE J., KISIEL B., SETTLES B., JR. E. R. H. & MITCHELL T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, p. 1306–1313.

CORBETT G. G. (2004). *The Russian Adjective : A Pervasive Yet Elusive Category*, In *Adjective classes : a cross-linguistic typology*, p. 199–222. Oxford University Press.

ETZIONI O., CAFARELLA M., DOWNEY D., POPESCU A.-M., SHAKED T., SODERLAND S., WELD D. S. & YATES A. (2005). Unsupervised named-entity extraction from the Web : An experimental study. *Artificial Intelligence*, **165**(1), 91 – 134. DOI : http ://dx.doi.org/10.1016/j.artint.2005.03.001.

FABER P. & CABEZAS-GARCÍA M. (2019). Specialized knowledge representation : From terms to frames. *Research in Language*, **17**, 197–211. DOI : 10.2478/rela-2019-0012.

FELLBAUM C. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London : The MIT Press.

GELBUKH A. F. (2018). Inferences for enrichment of collocation databases by means of semantic relations. *Computación y Sistemas*, **22**(1), 103–117. DOI : http ://dx.doi.org/10.13053/cys-22-1-2923.

HERNÁNDEZ-GONZÁLEZ J., HRUSCHKA JR. E. R. & MITCHELL T. M. (2017). Merging knowledge bases in different languages. In *Proceedings of TextGraphs-11 : the Workshop on Graph-based Methods for Natural Language Processing*, p. 21–29, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/W17-2403.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing*, Pattaya, Chonburi, Thailand. HAL : lirmm-00200883.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

NICKEL M., MURPHY K., TRESP V. & GABRILOVICH E. (2015). A review of relational machine learning for knowledge graphs : From multi-relational link prediction to automated knowledge graph construction. In *CBMM Memos*. CBBM.

RAMADIER L. (2016). *Indexation and learning of terms and relations from reports of radiology*. Theses, Université de Montpellier. HAL : tel-01479769.

SÉRASSET G. (2014). DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, p. 355–361. DOI : 10.3233/SW-140147.

SHAROFF S. & NIVRE J. (2012). The proper place of men and machines in language technology. processing russian without any linguistic knowledge. In *Dialogue 2011, Russian Conference on Computational Linguistics*, p. 657–670.

SHIN J., WU S., WANG F., DE SA C., ZHANG C. & RÉ C. (2015). Incremental knowledge base construction using deepdive. *Proc. VLDB Endow.*, **8**(11), 1310–1321. DOI : 10.14778/2809974.2809991.

SPEER R. & HAVASI C. (2012). Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, p. 3679–3686, Istanbul, Turkey : European Languages Resources Association (ELRA).

STRÖMQVIST S. & VERHOEVEN L. (2004). *Relating events in narrative : Typological and contextual perspectives*. Mahwah, NJ : Lawrence Erlbaum.

SUCHANEK F., KASNECI G. & WEIKUM G. (2007). Yago : A Core of Semantic Knowledge. Unifying WordNet and Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, p. 697 – 697, Banff, Canada. DOI : 10.1145/1242572.1242667.

TALMY L. (1985). *Lexicalization patterns : Semantic structure in lexical forms.*, In *Language typology and syntactic description.*, volume 3, p. 57–149. Cambridge : Cambridge University Press.

TCHECHMEDJIEV A. (2016). *Semantic Interoperability of Multilingual Lexical Resources in Lexical Linked Data*. Theses, Université Grenoble Alpes. HAL : tel-01681358.

WANG Q., WANG B. & GUO L. (2015). Knowledge base completion using embeddings and rules. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, p. 1859–1865 : AAAI Press.

ZARROUK M. (2015). *Consolidation endogène de réseaux lexico-sémantiques : Inférence et annotation de relations, règles d'inférence et langage dédié*. Theses, Université de Montpellier. HAL : tel-01300285.

ZEMSKAJA E., KITAJGORODSKAJA M. & SIRJAEV E. (1981). *Russkaja razgovornaja rec' : Obcie voprosy, slovoobrazovanie, sintaksis*. Moskva : Nauka.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Extraction de thèmes d'un corpus de demandes de support pour un logiciel de relation citoyen

Mokhtar Boumedyen Billami Christophe Bortolaso Mustapha Derras
Berger-Levrault, 64 rue Jean Rostand, 31670 Labège
mb.billami@berger-levrault.com,
christophe.bortolaso@berger-levrault.com,
mustapha.derras@berger-levrault.com

RÉSUMÉ

Nous nous intéressons dans cet article à l'extraction de thèmes (*topics*) à partir de commentaires textuels provenant des demandes de support de l'éditeur de logiciel Berger-Levrault. Le corpus de demandes analysé est celui d'un outil de gestion de la relation citoyen. Ce corpus n'est pas formaté et est peu structuré avec plusieurs locuteurs qui interviennent (le citoyen et un ou plusieurs techniciens support). Nous décrivons une étude expérimentale qui repose sur l'utilisation de deux systèmes. Le premier système applique une LDA (Allocation Dirichlet Latente), tandis que le second combine l'application d'une LDA avec l'algorithme k-Moyennes (*k-Means*). Nous comparons nos résultats avec un échantillon de ce corpus, annoté par un expert du domaine. Nos résultats montrent que nous obtenons une classification de meilleure qualité comparable avec celle effectuée manuellement par un expert en utilisant une combinaison LDA/k-Moyennes.

ABSTRACT

Topic extraction from a corpus of support requests for citizen relations software.
In this paper, we are interested in topic modeling from textual comments provided by support requests from Berger-Levrault software editor. The selected corpus relates to a citizen relationship management platform. This corpus is not formatted and not well-structured involving multiple speakers (the citizen and one or more technicians). We describe an experimental study based on the use of two systems. The first system applies an LDA (Latent Dirichlet Allocation), while the second combines the application of an LDA with the k-Means algorithm. We compare our results with a datatest derived from this corpus, annotated by an expert in the field. Our results show that we obtain a better classification like the one carried out manually by an expert using a combination of LDA/k-Means.

MOTS-CLÉS : Modélisation de thèmes, Allocation Dirichlet Latente, k-Moyennes.
KEYWORDS: Topic Modeling, Latent Dirichlet Allocation, k-Means.

1 Introduction

La modélisation et l'identification du thème auquel appartiennent les documents d'une collection donnée de textes sont essentielles pour de nombreuses applications. Par exemple, la recherche d'information (Yi & Allan, 2009), les systèmes de recommandation (Al-Ghossein et al., 2018), la classification de textes (Guha Neogi et al., 2019), les sciences cognitives (Pirnay-Dummer & Walter, 2009) et l'analyse de sentiments dans les réseaux sociaux (Naskar et al., 2016), pour n'en

Volume 2 : Traitement Automatique des Langues Naturelles, pages 155–163.
HAL : hal-02784763.

citer que quelques-unes. Dans ce contexte, un topic est un groupe de mots clés qui est considéré intuitivement comme représentant un thème sémantique latent présent dans un document. L'extraction des thèmes consiste à effectuer une analyse distributionnelle sur un corpus de données pour en mesurer les probabilités de distribution des thèmes sur les termes du vocabulaire utilisé. La modélisation des thèmes (en anglais *topic modeling*) est un processus sémantique qui applique un regroupement de niveau supérieur sur les mots clés et repose sur des modèles statistiques qui capturent la probabilité d'apparition des mots sémantiquement liés dans un contexte défini. Par exemple, pour capturer l'idée que *demande* et *inscription* se rapportent au même thème, tandis que *demande* et *FAQ* (i.e. Foire Aux Questions) se rapportent à un thème différent.

Plusieurs travaux se sont concentrés sur la modélisation et l'identification du thème auquel appartiennent les documents d'un corpus de données (Lin, 1995 ; Medelyan et al., 2008 ; Varga et al., 2014 ; Venkatesaramani et al., 2019). Ces approches reposent sur l'analyse de termes (ou parfois d'entités nommées), extraits à partir des textes, sur lesquels un regroupement peut être effectué afin de les lier par thème (par exemple, dans notre cas d'étude, *inscription, gestion de doublons* et *renumérotation*). Par la suite, un étiquetage du thème (Sorodoc et al., 2017 ; Gourru et al., 2018) peut être (facultativement) appliqué pour attribuer à ce groupe de termes un identifiant approprié se référant au thème en question (par exemple, *doublons d'inscription*).

Nous nous intéressons dans cet article à l'extraction de thèmes à partir de commentaires textuels évoqués par des clients et techniciens support pour le traitement des données du logiciel e.élections appartenant à l'éditeur Berger-Levrault. Il s'agit d'un logiciel de gestion des inscriptions électorales et de la préparation des scrutins. Plusieurs tâches sont nécessaires à la gestion des élections, par exemple, la modification des adresses des électeurs, l'optimisation d'édition des cartes électorales, la mise à jour du nom du maire, voire la gestion des transmissions dans un répertoire national induisant la gestion des doublons citée ci-dessus. Ce travail présente une étude expérimentale de méthodes d'extraction de thèmes à partir d'un corpus de demandes de support. Plus précisément, nous traitons la tâche d'identification des thèmes. Pour ce qui concerne l'étiquetage des thèmes, cette tâche n'est pas traitée dans ce travail mais constitue un complément essentiel.

Après avoir présenté dans la section 2 les travaux état-de-l'art d'extraction de thèmes, nous décrivons dans la section 3 le corpus de travail et d'évaluation de nos méthodes. Nous discutons du format de ces corpus et des conséquences sur le processus d'extraction de thèmes. Dans la section 4, nous présentons la méthodologie que nous avons suivie pour créer des modèles d'apprentissage permettant l'identification des thèmes. Enfin, les résultats d'expérimentation sont discutés dans la section 5 avant de terminer par une conclusion et quelques perspectives (section 6).

2 Travaux antérieurs d'extraction de thèmes

L'identification des thèmes peut être envisagée en appliquant directement les méthodes traditionnelles proposées dans la littérature telles que l'analyse sémantique latente–LSA (Deerwester et al., 1990), la LSA probabiliste–pLSA (Hofmann, 2001) ou l'allocation Dirichlet latente–LDA (Blei et al., 2003). Cependant, ces approches ont certaines limites. Premièrement, elles ne fonctionnent généralement que sur des mots individuels et non sur des expressions polylexicales, bien que des extensions aient été proposées pour tenir compte des termes multi-mots (Nokel & Loukachevitch, 2016; Blei & Lafferty, 2009). Deuxièmement, les thèmes sont considérés comme des variables latentes associées à une probabilité de générer des mots, et ne sont donc pas directement « étiquetés », ce qui les rend difficile à externaliser, même si des extensions pour étiqueter les thèmes sont disponibles (Gourru et al., 2018). Enfin, les mots sont difficiles à

interpréter sémantiquement et sont généralement considérés comme des références symboliques sur lesquelles une inférence statistique/probabiliste peut être appliquée. Toutefois, il existe une alternative qui permet de surmonter cette dernière limite lorsque le modèle LDA est appliqué. Il s'agit de l'outil de visualisation pyLDAvis (Sievert & Shirley, 2014). Cet outil permet de faciliter l'interprétation des thèmes à l'aide de la représentation graphique proposée. En effet, sélectionner un ou plusieurs termes sur le graphe permet non seulement de mieux expliquer le thème en question mais aussi de tirer des conclusions sur l'importance de certains termes pour certains thèmes. Par ailleurs, des approches ont émergé et proposent d'utiliser les informations structurées disponibles dans les bases de connaissances ontologiques pour améliorer l'identification des thèmes dans les textes (Jain & Pareek, 2010). Toutefois, ces ressources ne sont pas toujours disponibles principalement lorsque nous sommes confrontés au traitement d'un corpus de spécialité.

D'autres travaux se sont concentrés sur l'utilisation des méthodes traditionnelles en conjonction avec les informations extraites à partir de ressources standards, par exemple, la ressource DBpedia (Auer et al., 2007). Hulpuş et al. (2013) ont appliqué une LDA pour regrouper les mots en thèmes et ensuite, ils ont associé les mots groupés avec la ressource DBpedia pour étiqueter les thèmes. Todor et al. (2016), quant à eux, ont proposé un travail en sens inverse, c'est-à-dire, d'abord associer les termes avec la base DBpedia avant d'appliquer un algorithme d'identification de thèmes dans le but d'enrichir les textes avec des annotations en catégories de Wikipédia, hyperonymes, entités liées aux entités extraites, etc. Medelyan et al. (2008) ont proposé un travail qui consiste à indexer les thèmes sur tout le corpus Wikipédia par application d'un contrôleur de vocabulaire. Ce contrôleur repose sur l'utilisation des noms des articles de Wikipédia.

Pour l'étiquetage des thèmes, des approches à base de graphes de connaissances ont été proposées en utilisant des ressources lexico-sémantiques ou ontologiques (Hulpuş et al., 2013 ; Allahyari & Kochut, 2015). Ce type d'approches a été aussi proposé pour l'amélioration de l'identification des thèmes en utilisant des plongements d'entités (*entity embeddings*), par exemple (Yao et al., 2017 ; Brambilla & Altinel, 2019). Yao et al. (2017) ont entraîné les représentations d'entités au moyen d'une utilisation des réseaux sémantiques comme WordNet (Miller, 1995) et des bases de connaissances en ligne comme FreeBase (Bollacker et al., 2008).

3 Données de travail

Nous utilisons un corpus français de demandes de support à propos du logiciel e.élections. Nous appellerons ce corpus par la suite le corpus support Élections. Ce dernier a plusieurs particularités (par exemple, mots non structurés, formatage, ponctuation faible, etc.). Cela s'explique par le fait que pour une demande donnée, plusieurs échanges ont eu lieu soit par écrit, soit par téléphone. Toutefois, le contenu textuel que nous avons récupéré pour chaque demande regroupe la question du citoyen et les différentes réponses ayant été rapportées à sa demande par suite des échanges effectués avec les techniciens support. Par ailleurs, le vocabulaire utilisé par les citoyens est très varié et parfois mal structuré. Il dépasse la plupart du temps les termes (mots) utiles pour apporter une réponse à une demande donnée. L'utilisation de l'ensemble des mots du vocabulaire du corpus peut facilement engendrer un biais. L'aspect de réduction du vocabulaire est traité dans ce travail et est discuté dans la section qui suit.

Le corpus de travail ayant servi à l'entraînement de nos modèles d'apprentissage contient 94 478 échanges représentant 31 036 demandes différentes. Ce corpus a été collecté durant le dernier trimestre de l'année 2019. Une première analyse de ce corpus avec Spacy (Honnibal & Johnson, 2015), chaîne du traitement automatique de la langue, nous a montré la diversité du vocabulaire

utilisé : 12 412 mots pleins différents, c'est-à-dire, mots portant du sens (noms, adjectifs, adverbes et verbes). Le corpus d'évaluation, quant à lui, représente un échantillon de données collecté durant le troisième trimestre de l'année 2019. Cet échantillon décrit 535 demandes différentes dont près de 74 % des demandes (394 cas) ont été annotées avec un seul thème par un expert du domaine.

4 Méthodologie

Cette section décrit l'architecture de notre approche et les modèles d'apprentissage que nous avons développés. Cette approche répond au problème d'identification des thèmes en trois parties différentes, à savoir : (1) quels sont les outils et les ressources nécessaires pour prétraiter les données et réduire le vocabulaire utilisé dans le corpus support Élections ? (2) quel est le modèle de représentation vectorielle à utiliser pour décrire les demandes ? et (3) quel est le modèle d'apprentissage non supervisé à utiliser pour apprendre les thèmes discutés dans le corpus ? La figure 1 présente l'architecture de notre approche.

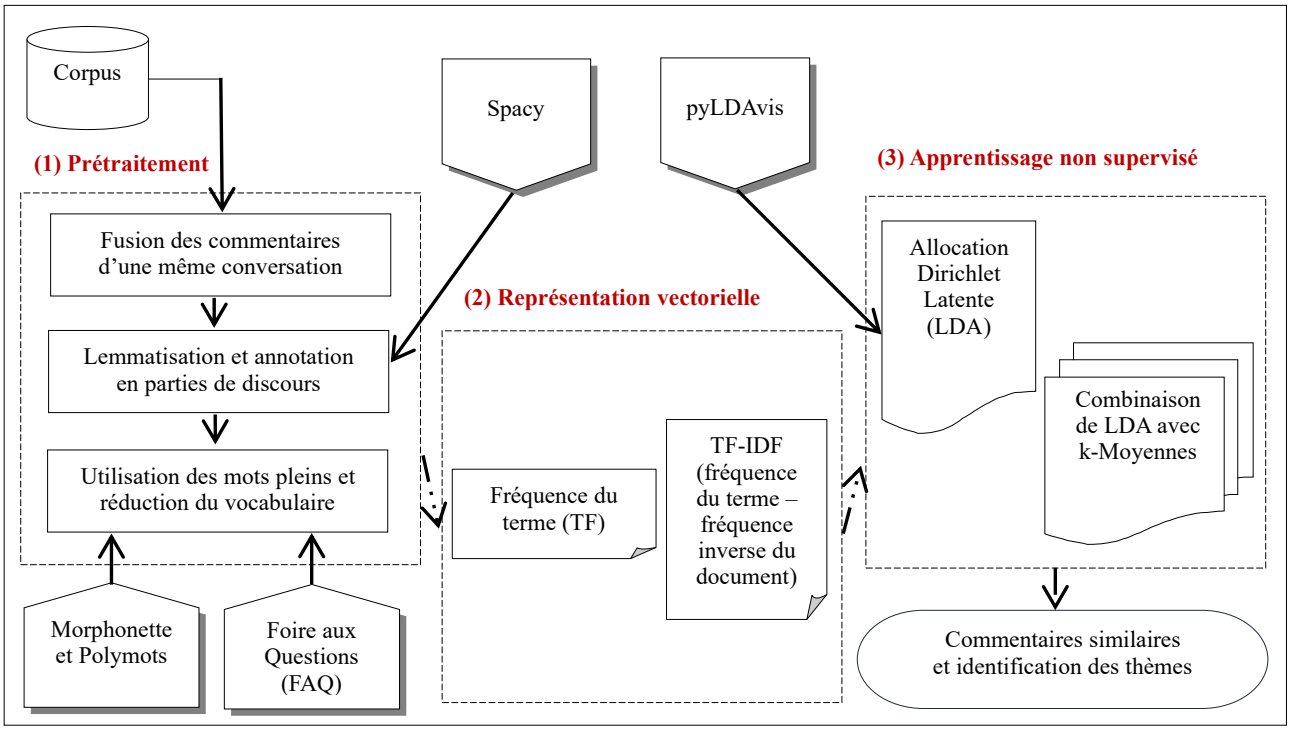


FIGURE 1: Approche d'identification des thèmes à partir des demandes de support

Pour la première partie, tout d'abord un travail de fusion des différents échanges liés à une même demande est effectué, en se référant au numéro de référence de la demande. Ensuite, nous utilisons Spacy (Honnibal & Johnson, 2015) pour extraire les formes lemmatisées des mots portant du sens.

L'étape suivante consiste à contrôler le vocabulaire exprimé dans les différents commentaires en utilisant un vocabulaire spécialisé et en effectuant des substitutions lexicales par l'utilisation de familles morphologiques. Plus précisément, nous avons utilisé un ensemble de documents FAQ liés aux corpus support Élections (22 documents avec une moyenne de 2 pages de contenu textuel, hors figures, par document). Ces documents FAQ permettent d'apporter des réponses aux questions des utilisateurs sur le produit e.élections. Un vocabulaire de 804 mots pleins a ainsi été obtenu.

Après avoir obtenu le vocabulaire lié aux documents FAQ, nous effectuons une substitution lexicale des mots du corpus par d'autres mots appartenant à une même famille morphologique. L'idée consiste à remplacer chaque mot plein du corpus par le mot le plus fréquent (dans le corpus). Pour cela, nous utilisons deux ressources, à savoir : Morphonette (Hathout, 2008) et Polymots (Gala & Rey, 2008). Morphonette est un réseau lexical morphologique construit à partir de la liste des mots du Trésor de la langue Française (Dendien & Pierrel, 2003), par utilisation des mesures de similarité morphologique. Ce réseau contient plus de 8 600 familles morphologiques. Polymots (Gala & Rey, 2008), quant à lui, représente une base de données lexicale permettant d'avoir des familles morpho-phonologiques. Cette base contient 8 000 mots communs en français et a été construite à partir de la continuité de sens et des formes phoniques comparables. Nous utilisons à la fois ces deux ressources puisqu'elles sont complémentaires. Par exemple, Polymots contient la famille morphologique *doublon*, *double*, *doublé*, *doubleur*, *doublage* et *doublement* alors que Morphonette ne propose pas de famille pour le mot *doublon*. D'un autre côté, Morphonette contient la famille *procuration*, *procurer*, *procureur* alors que Polymots ne propose pas de famille pour le mot *procuration*. La substitution que nous effectuons nous permet non seulement d'avoir une réduction supplémentaire du vocabulaire (nous passons de 804 à 513 mots) mais aussi d'ajouter une importance aux mots les plus fréquents. Par exemple, *doublon* est représenté avec 2 494 occurrences dans le corpus de travail contre 697 pour *double* et 3 pour *doublé*. La substitution lexicale donne ainsi un total d'occurrences de 3 194 pour *doublon*.

Pour la représentation des commentaires, nous utilisons Scikit-learn (Buitinck et al., 2013). Plus précisément, les modèles de représentation *CountVectorizer* et *TfidfVectorizer*. *CountVectorizer* permet de créer des représentations tenant compte seulement du nombre d'occurrences de mots dans un texte. *TfidfVectorizer*, quant à lui, permet de créer des représentations à base de TF-IDF, fréquence du terme-fréquence inverse du document (Jones, 1972). Cette technique donne des poids les plus élevés aux termes (i.e. mots) qui apparaissent plus fréquemment dans un texte par rapport aux autres textes du corpus. Pour l'expérimentation, nous avons utilisé ces deux techniques.

Pour le modèle d'apprentissage, nous utilisons l'allocation Dirichlet latente (Blei et al., 2003) et k-Moyennes (McQueen, 1967). L'allocation Dirichlet latente permet de représenter les textes comme des distributions de thèmes. En comparaison avec l'analyse sémantique latente probabiliste, la LDA suppose que les thèmes sont répartis entre les textes et les mots répartis entre les thèmes. Par ailleurs, k-Moyennes permet d'aboutir pour notre cas à une répartition des commentaires du corpus en *k* groupes (*clusters*). Les commentaires de chaque groupe partagent une certaine sémantique qui peut être identifiée en analysant la distribution des termes pour le groupe (i.e. fréquence des mots). Nous avons fait le choix de combiner LDA avec k-Moyennes pour voir le comportement de ce dernier sur des demandes évoquant au moins deux thèmes proches sémantiquement. Pour cette combinaison, l'algorithme LDA travaille sur la distribution des mots alors que l'algorithme k-Moyennes prend en entrée la distribution de probabilités des thèmes retournés par LDA.

5 Résultats d'expérimentation

Nous avons tout d'abord entrainé un modèle LDA sur les données du corpus de travail. Afin de choisir les meilleures valeurs pour les deux paramètres à fournir pour LDA, à savoir : le nombre de composants (thèmes) et la décomposition de l'apprentissage (*learning decay, ld*), nous avons appliqué la méthode *GridSearchCV* qui est implémentée dans Scikit-learn. Le nombre de thèmes a été varié entre 2 et 20 et nous avons pris une valeur pour la variable *ld* soit de 0,5, 0,7 ou 0,9. Les résultats obtenus ont montré que le paramétrage optimal pour notre modèle LDA serait de prendre 10 thèmes avec une valeur *ld* égale à 0,5, cela en utilisant les deux types de représentation des

commentaires, à savoir TF (fréquence du terme) et TF-IDF. Avec cette configuration, nous avons obtenu un rapport de vraisemblance de -1 204 185 et une perplexité de 123 par simple application du TF au type de représentation, ce qui est relativement bien pour une telle valeur de vraisemblance.

Nous avons comparé les résultats obtenus par l'utilisation des deux types de représentation des demandes de support. Cette comparaison nous a montré que le simple TF permet de mieux distinguer au moins quatre thèmes. La figure 2 montre les cartes retournées par pyLDAvis (Sievert & Shirley, 2014) sur les deux types de représentation. Nous remarquons que l'application de TF-IDF (voir la carte 'b)' de la figure 2) renvoie une projection très dense de la plupart des thèmes où la distance entre eux est très petite. Cela peut s'expliquer par le fait que la pondération TF-IDF accapare une très large part de la variance en raison de la présence des deux thèmes 9 et 10.

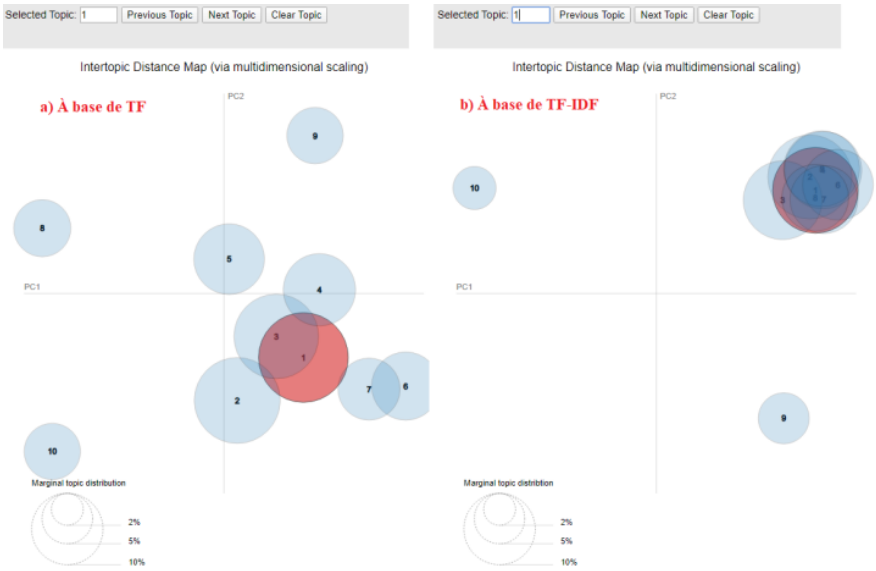


FIGURE 2: Visualisation des thèmes obtenus à l'aide de l'outil pyLDAvis

Pour ce qui suit, nous présentons nos résultats par application du type de représentation TF. Le tableau 1 décrit les mots les plus fréquents pour chaque thème du corpus de travail support Élections.

Thème	Liste de mots les plus fréquents
1	inscription ; notification ; traiter ; radiation ; pas ; date ; office ; avoir ; client ; compte.
2	logiciel ; mise ; jour ; rappeler ; nom ; synchronisation ; version ; avance ; usage ; pas.
3	demande ;insee ; transmettre ;pas; instruit ;message ; viser ;traiter ; notification ; erreur.
4	liste ; numéro ; avoir ; pas ; renumérotation ; apparaitre ; client ; fois ; carte ; électeur.
5	client ;connexion ; erreur ;pas ; message ;portail ; service ;période ;application ; besoin.
6	adresse ; doublon ; naissance ; changement ; ville ; commune ; gestion ; supprimer.
7	bureau ; vote ; synchronisation ; faire ; lancer ; pas ; base ; exister ; erreur ; sauvegarde.
8	faq ;client ;demande ;site ;lien ;recevoir ; cas ;disponible ;assistance ;réponse ; pouvoir.
9	carte ; retour ; procuration ; commission ; premium ; liste ; européen ; imprimer ;saisir.
10	tableau ; commander ; relatif ; espace ; solliciter ; récupérer ; transmettre ; condition.

TABLE 1 : Mots les plus fréquents de chaque thème par application du LDA

Nous remarquons que les thèmes numérotés en 5, 8, 9 et 10 se distinguent des autres puisque chacun est lié à une tâche de traitement différente, par exemple, 'problème de connexion et gestion des

messages d'erreurs' pour le thème 5 et 'demande des FAQ' pour le thème 8. Par ailleurs, certaines demandes n'ont pas pu être représentées (i.e. vecteurs nuls) en raison du texte très court utilisé par le citoyen (par exemple, *problème*) et l'absence de la réponse textuelle (fournie en général par téléphone pour ces cas). Il s'agit de 386 demandes (*Out-Of-Vocabulary, OOV*) sur 31 036.

Nous avons appliqué par la suite l'algorithme k-Moyennes en utilisant comme entrées les probabilités de distribution des thèmes renvoyées par LDA. Nous avons effectué une analyse de la silhouette en variant de la même façon le nombre de composants (thèmes dans notre cas) de 2 à 20. Le meilleur coefficient de la silhouette obtenu est 0,41 et revient à l'utilisation de 10 thèmes. Cela confirme d'une certaine manière le nombre de thèmes à prendre en compte pour classifier les demandes. La comparaison de nos deux systèmes sur la distribution des demandes du corpus de travail pour le thème le plus probable (majoritaire) nous montre des différences : 1 050 demandes évoquent un thème différent selon le système utilisé (3,38 % de cas). Le tableau 2 présente la distribution des demandes sur le thème majoritaire.

Modèle	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**OOV**
LDA	5010	4223	4077	4322	4060	2407	2013	1105	2071	1362	386
LDA/k-Means	4870	4429	3962	4380	3924	2639	2035	1009	2078	1324	386

TABLE 2 : Nombre de demandes associées au thème majoritaire

Nous avons évalué la qualité de nos deux systèmes sur 394 nouvelles demandes de support ayant été annotées par un expert du domaine. Sur l'ensemble de ces cas, 8 thèmes ont été évoqués, à savoir : (a) changement d'adresse, (b) demande d'inscription/radiation, (c) doublon d'inscription, (d) inscrits d'Office, (e) erreur du numéro d'émargement, (f) synchronisation, (g) suppression de la radiation et (h) divers. Nous avons remarqué plusieurs équivalences de thèmes entre ceux retournés par nos systèmes et ceux annotés par l'expert, voir même de nouveaux thèmes que l'expert a pris comme divers. Par exemple, l'édition des cartes électorales (thème 9). Il est à noter que l'expert était libre de choisir les thèmes jugés pertinents et n'avait pas l'obligation de spécifier exactement 8 thèmes. Par ailleurs et afin de mesurer la performance de nos systèmes, nous avons mis en correspondance les thèmes générés automatiquement et ceux proposés par l'expert. L'application du système LDA/k-Moyennes s'est vu meilleure que l'utilisation seule de LDA lorsque les demandes évoquent au moins deux thèmes importants. Nous avons obtenu un taux d'exactitude de 72 % pour LDA contre 77,69 % pour LDA/k-Moyennes.

6 Conclusion et perspectives

Dans cet article, nous avons présenté deux méthodes d'identification des thèmes. La première repose sur l'utilisation de LDA, tandis que la seconde est en deux temps : elle applique d'abord une LDA sur la distribution des mots pour ensuite appliquer l'algorithme k-Moyennes sur la distribution des thèmes retournés par LDA. Nous avons montré que l'utilisation de la seconde méthode à deux niveaux est plus performante lors du traitement des commentaires ayant au moins deux thèmes proches sémantiquement. Par ailleurs, en raison de la variété du vocabulaire exprimé par les locuteurs (principalement les citoyens), une réduction de la liste des mots du corpus était nécessaire en utilisant non seulement un vocabulaire spécialisé (documents FAQ liés au corpus support Élections) mais aussi certaines ressources de familles morphologiques. Deux perspectives s'ouvrent à ce travail : (1) une intégration des expressions polylexicales par l'utilisation des N-grammes (Nokel & Loukachevitch, 2016) est possible afin d'améliorer nos modèles; et (2) un étiquetage des thèmes (Gourru et al., 2018) est envisageable afin de faciliter leur intégration dans l'amélioration de plusieurs applications du traitement automatique du langage naturel.

Références

ALLAHYARI M. & KOCHUT K. (2015). Automatic Topic Labeling Using Ontology-Based Topic Models. *International Conference on Machine Learning and Applications (ICMLA)*, p. 259–264.

AL-GHOSSEIN M., MURENA P. A., ABDESSALEM T., BARRE A. & CORNUEJOLS A. (2018). Adaptive collaborative topic modeling for online recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, p. 338–346. DOI : 10.1145/3240323.3240363.

AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R. & IVES Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *International Semantic Web Conference (ISWC-ASWC)*, p. 722–735.

BLEI D. M. & LAFFERTY J. D. (2009). Visualizing Topics with Multi-Word Expressions. *arXiv preprint* arXiv:0907.1013.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, **3**(Jan), p. 993–1022.

BOLLACKER K., EVANS C., PARITOSH P., STURGE T. & TAYLOR J. (2008). Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In *Proceedings of the ACM SIGMOD international conference on Management of data*, p. 1247–1250.

BRAMBILLA M. & ALTINEL B. (2019). Improving Topic Modeling for Textual Content with Knowledge Graph Embeddings. *AAAI-MAKE Spring Symposium*.

BUITINCK L., LOUPPE G., BLONDEL M., PEDREGOSA F., MUELLER A., GRISEL O., NICULAE V., PRETTENHOFER P., GRAMFORT A., GROBLER J., LAYTON R., VANDERPLAS J., JOLY A., HOLT B. & VAROQUAUX G. (2013). API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, p. 108–122.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**, p. 391–407.

DENDIEN J. & PIERREL J.-M. (2003). Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues*. Sous la direction de M. ZOCK & J. CAROLL, **44**(2), p. 11–37.

GALA N. & REY V. (2008). POLYMOTS: une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. *Actes de TALN 08: Traitement Automatique des Langues Naturelles*.

GOURRU A., VELCIN J., ROCHE M., GRAVIER C. & PONCELET P. (2018). United we stand: Using multiple strategies for topic labeling. *NLDB: Natural Language Processing and Information Systems*, p. 352–363. DOI : 10.1007/978-3-319-91947-8_37, HAL : lirmm-01910614.

HATHOUT N. (2008) Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the COLING Workshop Textgraphs-3*, p. 1–8.

HOFMANN T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, **42**(1), p. 177–196.

HONNIBAL M. & JOHNSON M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1373–1378. DOI : 10.18653/v1/D15-1162.

HULPUŞ I., HAYES C., KARNSTEDT M. & GREENE D. (2013). Unsupervised graph-based topic labelling using DBpedia. In *Web Search and Web Data Mining (WSDM)*, p. 465–474. ACM.

JAIN S. & PAREEK J. (2010). Automatic Topic(s) Identification from Learning material: An Ontological Approach. In *Computer Engineering and Applications (ICCEA)*, volume 2, p. 358–362. IEEE.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation,* **28**, p. 11–21.

LIN C.-Y. (1995). Knowledge-based automatic topic identification. In *Annual meeting of the Association for Computational Linguistics (ACL)*, p. 308–310.

MACQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume **1**, p. 281–297.

MEDELYAN O., WITTEN I. H. & MILNE D. (2008). Topic indexing with Wikipedia. In *2008 AAAI workshop "Wikipedia and Artificial Intelligence: An Evolving Synergy"*.

MILLER G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, **38**(11), p. 39–41.

NASKAR D., MOKADDEM S., REBOLLO M. & ONAINDIA E. (2016). Sentiment Analysis in Social Networks through Topic modeling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 46–53.

NOKEL M. & LOUKACHEVITCH N. (2016). Accounting ngrams and multi-word terms can improve topic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, p. 44–49. DOI : 10.18653/v1/W16-1806.

GUHA NEOGI P. P., DAS A. K., GOSWAMI S. & MUSTAFI J. (2019). Topic Modeling for Text Classification. *Emerging Technology in Modelling and Graphics*, p. 395–407, Springer. DOI : 10.1007/978-981-13-7403-6_36.

PIRNAY-DUMMER P. & WALTER S. (2009). Bridging the world's knowledge to individual knowledge using latent semantic analysis and web ontologies to complement classical and new knowledge assessment technologies. *Technology, Instruction, Cognition & Learning*, **7**(1).

SIEVERT C. & SHIRLEY K. E. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, p. 63–70. DOI : 10.3115/v1/W14-3110.

SORODOC L., LAU J. H., ALETRAS N. & BALDWIN T. (2017). Multimodal Topic Labelling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, p. 701–706.

TODOR A., LUKASIEWICZ W., ATHAN T. & PASCHKE A. (2016). Enriching topic models with DBpedia. *On the Move to Meaningful Internet Systems*, p. 735–751, Springer.

VARGA A., CANO BASAVE A. E., ROWE M., CIRAVEGNA F. & HE Y. (2014). Linked knowledge sources for topic classification of microposts: A semantic graph-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, **26**, p. 36–57. DOI : 10.1016/j.websem.2014.04.001.

VENKATESARAMANI R., DOWNEY, D., MALIN, B. & VOROBEYCHIK, Y. (2019). A Semantic Cover Approach for Topic Modeling. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, p. 92–102. DOI : 10.18653/v1/S19-1011.

YAO L., ZHANG Y., WEI B., JIN Z., ZHANG R., ZHANG Y. & CHEN Q. (2017). Incorporating Knowledge Graph Embeddings into Topic Modeling. In *Thirty-First AAAI Conference on Artificial Intelligence*, p. 3119–3126.

YI X. & ALLAN J. (2009). A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Boughanem M., Berrut C., Mothe J. & Soule-Dupuy C. (eds) Advances in Information Retrieval. ECIR 2009*. Lecture Notes in Computer Science, vol 5478, Springer.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Recommandation d'âge pour des textes

Alexis Blandin[1] Gwénolé Lecorvé[1] Delphine Battistelli[2] Aline Étienne[2]
(1) Univ Rennes, CNRS, IRISA, 6, rue de Kerampont, 22300 Lannion, France
(2) Univ. Paris-Nanterre, CNRS, MoDyCo, 200, avenue de la République, 92001 Nanterre, France
{alexis.blandin, gwenole.lecorve}@irisa.fr,
{delphine.battistelli, aline.etienne}@parisnanterre.fr

RÉSUMÉ

Cet article étudie une première tentative pour prédire une recommandation d'âge estimant à partir de quand un enfant pourrait comprendre un texte donné. À ce titre, nous présentons d'abord des descripteurs issus de divers domaines scientifiques, puis proposons différentes architectures de réseaux de neurones et les comparons sur un ensemble de données textuelles en français, dédiées à des publics jeune ou adulte. Pour contourner la faible quantité de données de ce type, nous étudions l'idée de prédire les âges au niveau de la phrase. Les expériences montrent que cette hypothèse, quoique forte, conduit d'ores et déjà à de bons résultats, meilleurs que ceux fournis par des experts psycholinguistes, y compris lorsque les phrases isolées sont remplacées par textes complets.

ABSTRACT

Age recommendation for texts.

This paper studies a first attempt to predict an age recommendation from which a text can be understood by a child. As such, we first exhibits features derived from various scientific domains, then propose different architectures of neural network and compare them on a dataset of French texts dedicated to young or adult audiences. To circumvent the lack of data, we study the idea to predict ages at the sentence level. The experiments show that this strong assumption leads to good results yet, better than those provided by psycholinguists, even when shifting isolated sentences to full texts.

MOTS-CLÉS : recommandation d'âge, enfants, psycholinguistique, réseaux de neurones.

KEYWORDS: age recommendation, children, psycholinguistics, neural networks.

1 Introduction

La façon dont un individu comprend un texte dépend à la fois des caractéristiques du texte et des capacités de l'individu. Par exemple, les capacités à se souvenir d'informations, à positionner un événement dans un scénario, analyser la structure d'une phrase, comprendre un mot ou simplement à lire sont autant de facteurs qui peuvent évoluer d'un individu à l'autre. C'est particulièrement vrai en fonction de l'âge puisque, pendant l'enfance, les capacités cognitives, linguistiques et culturelles évoluent beaucoup. Dans ce contexte, cet article vise à prédire automatiquement des recommandations d'âge pour des textes en français dans le but de maximiser leur compréhension par des enfants.

En tant que tâche, la recommandation d'âge peut être globalement affiliée à celle d'analyse de la lisibilité d'un texte, au sens de la prédiction de la difficulté de lecture d'un texte pour une population spécifique (François, 2015), par exemple, la lecture d'un texte par une personne non-native, ou celle d'un formulaire par des clients. Cependant, la lisibilité du texte est centrée sur l'activité de lecture,

Volume 2 : Traitement Automatique des Langues Naturelles, pages 164–171.
HAL : hal-02784764.

alors que notre travail cherche à intégrer une dimension lié à la compréhension du langage. Cela signifie que nous ne nous limitons pas aux personnes en capacité de lire et considérons également des textes transmis oralement (par exemple, une histoire racontée). En tant que tâche nouvelle, ce document porte les contributions suivantes : (*i*) nous listons un ensemble de descripteurs issus de divers domaines et potentiellement pertinents pour notre tâche ; (*ii*) nous étudions différentes façons de formaliser la prédiction de l'âge comme un problème de régression ; (*iii*) nous répondons au besoin en données annotées en collectant des textes dédiés aux enfants et en exploitant les recommandations fournies par les auteurs ou éditeurs ; (*iv*) nous testons l'hypothèse selon laquelle toutes les phrases d'un même texte partagent la même recommandation d'âge. Bien que linguistiquement contestable, les expériences montrent qu'il s'agit d'une hypothèse de travail concluante ; (*v*) enfin, nous étudions l'acceptabilité des erreurs produites par nos modèles en les comparant avec des prédictions faites par des experts psycholinguistes.

Dans cet article, la section 2 dresse un panorama des travaux connexes en psycholinguistique et en traitement automatique des langues (TAL), puis la section 3 présente les descripteurs retenus dans notre travail. La section 4 présente les différentes approches proposées et la section 5 l'ensemble de données sur lesquelles elles sont étudiées. Enfin, la section 6 présente les résultats.

2 États de l'art

Les études en psycholinguistique mettent en avant différents facteurs majeurs de l'évolution de la compréhension du langage. Tout d'abord, la mémoire phonologique à court terme, qui se développe fortement entre 2 et 8 ans (Gathercole, 1999), joue un rôle essentiel dans le stockage et la restitution de l'information. L'acquisition des notions temporelles est également cruciale car elle permet à un enfant de se situer dans le temps, ainsi que d'ordonner chronologiquement des événements (Tartas, 2010; Hickmann, 2012). Ainsi, la compréhension de notions comme les jours de la semaine, la vitesse, une date très ancienne ou certains connecteurs ou adverbes temporels plus ou moins complexes varie en fonction de l'âge (Vion & Colas, 1999). Les émotions sont également rapportées comme contributives à l'établissement et au maintien de la cohérence des faits dans un texte (Mouw *et al.*, 2019). Cependant, leur repérage progresse avec le temps, s'appuyant initialement sur le lexique, puis progressivement sur des suggestions d'ordre culturel (Blanc, 2010). De même, celles reconnues le plus tôt sont centrées sur l'enfant (la peur, la joie, etc.), puis intègrent progressivement une dimension sociale (la culpabilité, l'empathie, etc.) (Davidson, 2006).

Dès 5-6 ans, les études en apprentissage de la lecture apportent d'autres éléments d'analyse. En particulier, le modèle historique de Frith (Frith, 1985) fait valoir que la lecture est acquise en trois étapes principales à travers lesquelles l'enfant passe d'un décodage des mots par la reconnaissance de symboles globaux, puis de graphèmes et, enfin, de morphèmes. Par ailleurs, d'autres travaux notent que l'intonation lors de la lecture d'un texte – induite par le lexique, la ponctuation, la syntaxe, etc. – influe sur la perception d'un texte et que cette intonation évolue avec l'âge (Aguert *et al.*, 2009).

Enfin, des approches calculatoires existent depuis longtemps pour lier la lisibilité d'un texte à un niveau d'étude. Historiquement, celles-ci se fondent sur les complexités lexicale et syntaxique, à l'image de l'indice Flesch-Kincaid (Flesch, 1948), ou de la formule Dale-Chall qui considère en outre la notion de mots « difficiles » (Dale & Chall, 1948). Plus récemment et plus généralement en TAL, des travaux sur la simplification du texte pour les enfants (De Belder & Moens, 2010; Gala *et al.*, 2018) ou sur l'acquisition du français comme langue étrangère (François & Fairon, 2012) se rapprochent des nôtres. En particulier, (François & Fairon, 2012) propose de prédire des niveaux de lisibilité en utilisant des approches par apprentissage automatique et 46 critères linguistiques

mêlant lexique, syntaxe et sémantique. Notre travail s'en distingue de multiples manières. Outre des méthodes d'apprentissage revisitées, la différence principale tient dans le fait que nous intègrons des informations liées à certaines dimensions développementales et cognitives (temporalité, émotions) pour prédire un âge, là où (François & Fairon, 2012) se limite à des éléments linguistiques pour prédire un niveau de compétence dans une langue (A1, C2...).

3 Descripteurs

Nous considérons 10 dimensions linguistiques qui reflètent l'état de l'art précédent. Pour chaque dimension, nous avons cherché à maximiser le nombre d'informations pouvant être extraites automatiquement. Lorsque les informations linguistiques portent au niveau des mots, les descripteurs sont calculés comme la moyenne et l'écart-type des valeurs par mot. Finalement, pour chaque énoncé, le vecteur de descripteurs globaux est composé de 606 valeurs réelles. Le détail des informations extraites est donné ci-dessous.

Plongements (1 descripteur de dimension 500) : plongement moyen des mots du texte [1].
Lexique (5 descripteurs) : log-prob. des mots en français (estimé sur un vaste corpus d'articles de journaux, romans, transcriptions...) ; nombre de mots/lemmes[2] différents p/r à la longueur du texte.
Graphie/typographie (6) : Score de confusion graphique des mots [3] ; longueur des mots ; ratio de caractères alphanumériques par mot ; ratio de ponctuations par mot.
Morphosyntaxe (7) : classes grammaticales (verbes, verbes d'état, noms, adjectifs, clitiques, adverbes)[2] ; mots-outils.
Temps verbaux (24) : diversité des temps verbaux ; proportions de 14 temps (simples et composés) ; modes ; systèmes temporels (passé, présent, futur).
Personne et forme verbale (5) : proportion des première/deuxième/troisième personnes ; proportion des formes singulier/pluriel.
Syntaxe (8) : mots par phrase ; distances moy. et max. entre un mot et ses dépendances syntaxiques[2] ; nombre de dépendances entrantes/sortantes par mot ; profondeur de l'arbre de dépendances.
Connecteurs logiques (16) : addition ; temps ; but ; cause ; comparaison ; concession ; conclusion ; condition ; conséquence ; énumération ; explicat. ; illustrat. ; justificat. ; opposit. ; restrict. ; exclusion [4].
Phonétique (9) : longueur de la phrase en phonèmes [5] ; nombre de phonèmes par mot ; diversité des phonèmes dans le texte / dans les mots ; scores d'ordinarité phonétique [6].
Sentiments/émotions (26) : Scores de subjectivité et de polarité [7] ; mots identifiés comme déclencheurs d'une parmi 24 émotions [8].

4 Modèles

Notre objectif est de prédire à partir de quel âge un texte d'entrée peut être compris. Précisons que nous ne tenons pas compte d'éventuels retards d'apprentissage. Dans ce problème de régression, l'une des questions majeures est de savoir si l'âge peut être considéré comme une valeur réelle unique

1. Skip-grammes (Fauconnier, 2015) appris sur le corpus FrWaC (Baroni *et al.*, 2009).
2. En utilisant Bonsai (Candito *et al.*, 2010).
3. Inspiré de (Geyer, 1977).
4. Les catégories et les connecteurs ont été établis par consensus de diverses sources.
5. En utilisant eSpeak : `http://espeak.sourceforge.net/index.html`
6. Calculé comme la probabilité moyenne de chaque phonème en français, comme indiqué dans (Gromer & Weiss, 1990).
7. Utilisation du classifieur de sentiments TextBlob
8. Les mots et les émotions sont issus d'un raffinement du dictionnaire EMOTAIX (Piolat & Bannour, 2009).

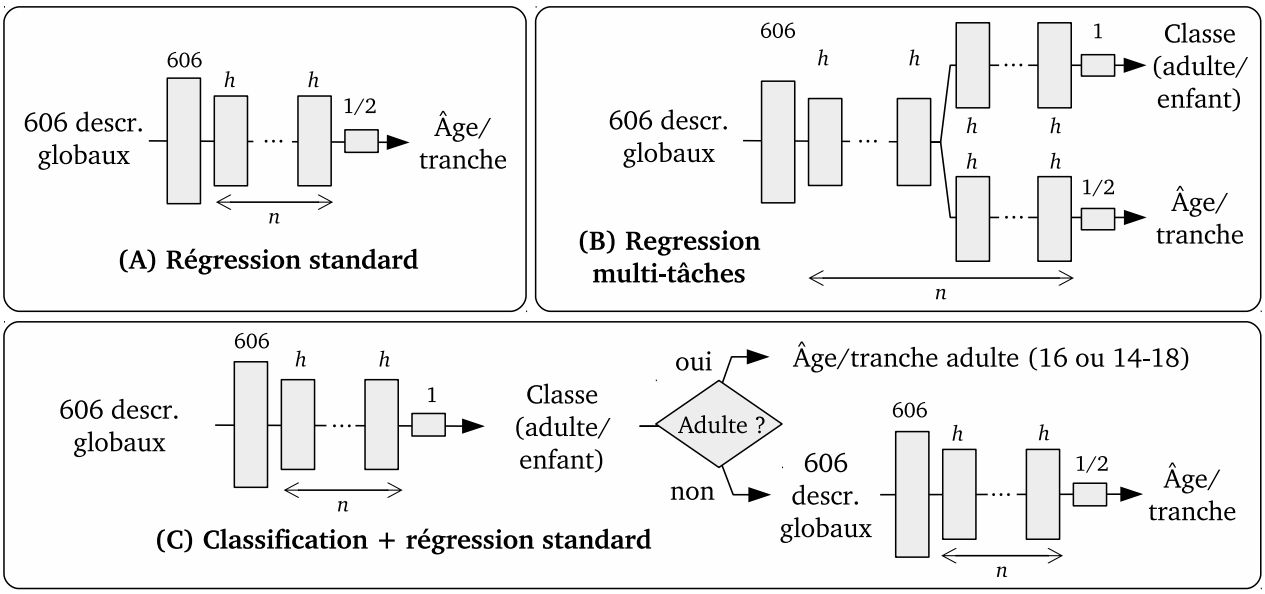


FIGURE 1 – Architectures des approches étudiées.

ou, comme le font souvent les auteurs et éditeurs, comme une tranche d'âges (un âge bas et un âge haut). Cette seconde modélisation reflète le fait que les enfants ne développent pas leurs compétences exactement au même âge et que les textes ont des irrégularités en terme de complexité. Dans ce travail, nous testons les deux modélisations. Par ailleurs, nous fixons une borne supérieure à nos recommandations, borne intuitivement associée à ce que nous qualifions de niveau « adulte ». Elle est fixée à 16 ans, ou 14-18 ans en terme de tranche. Cela coïncide avec la période du lycée en France. Pour un texte, l'évaluation de la tâche se fait en termes de différence absolue entre l'âge prévu et l'âge attendu (vérité terrain). Ainsi, à l'échelle d'un corpus, la métrique utilisée dans notre travail est l'erreur absolue moyenne, notée MAE pour *Mean Absolute Error*. Dans le cas des tranches d'âges, les deux bornes sont prédites, puis ramenés à leur moyenne. Les MAE sont ainsi reportés sur l'âge bas, l'âge haut et la moyenne des deux.

La figure 1 présente les 3 architectures de modèles neuronaux non-récurrents que nous étudions, toutes fondées sur un vecteur de 606 descripteurs globaux et produisant soit un âge ou une tranche d'âges recommandés. Le modèle A est un modèle de régression standard. Le modèle B est un modèle multi-tâches où la prédiction de l'âge est augmentée d'une classification binaire adulte/enfant. Enfin, le modèle C enchaîne un classificateur et un modèle de régression si la classe prédite est "enfants". L'idée est que la régression est inutile pour les textes considérés comme "adultes" car les âges associés sont fixes (16 et 14-18). Le modèle de régression est le même que A mais estimé sur l'ensemble d'apprentissage restreint aux seuls textes pour enfants.

La taille h et le nombre n des couches, ainsi que les fonctions d'activation sont réglées sur le modèle A et dupliqués pour les autres modèles. Le réglage de ces paramètres, ainsi que le nombre de couches spécifiques dans le modèle B, sont détaillés en section 6.

5 Données

Comme détaillé par la figure 2, nous avons collecté un ensemble de 631 textes, dont 541 sont destinés aux enfants de 0 à 14 ans, les 90 autres étant pour les adultes. Les textes pour enfants proviennent de

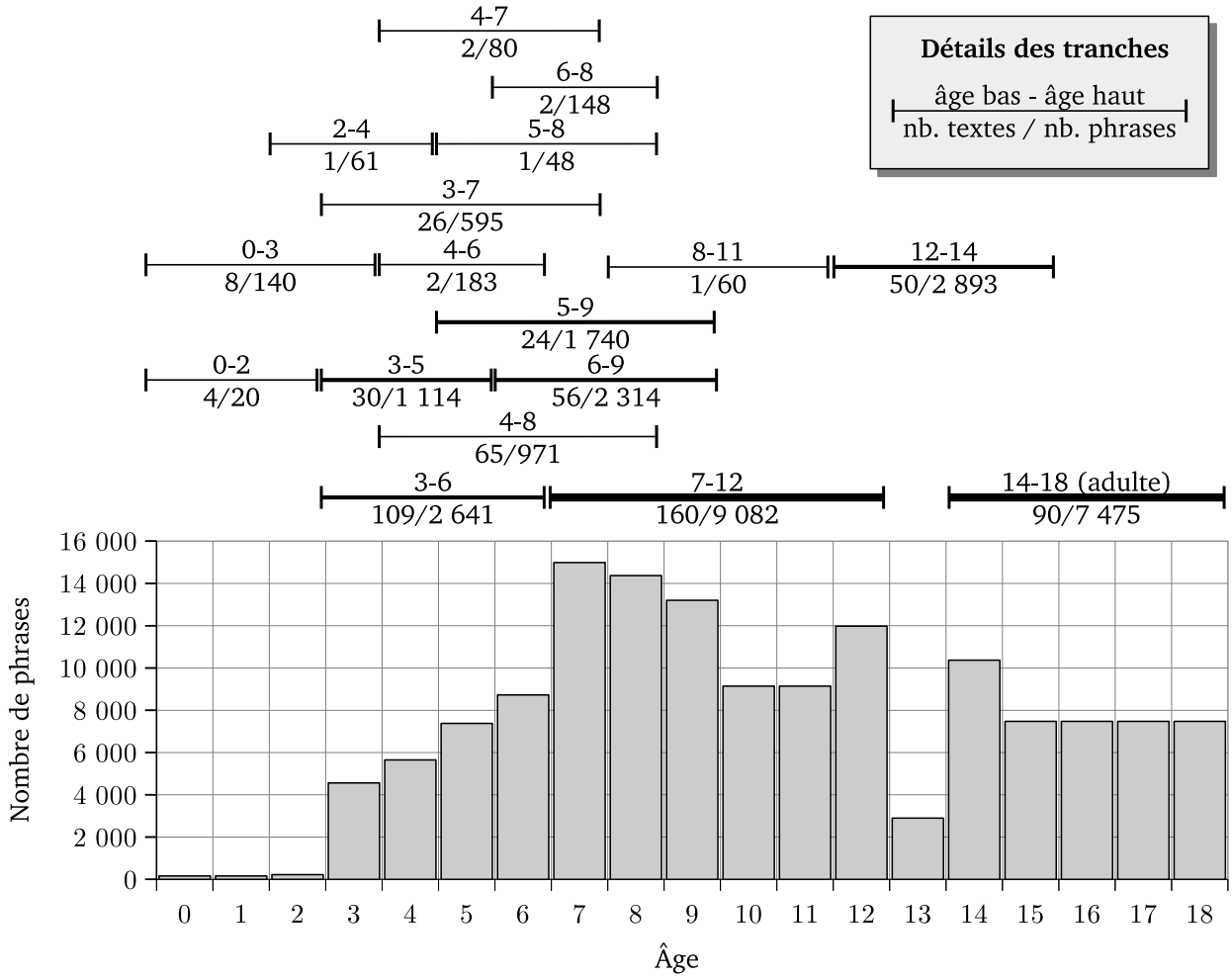


FIGURE 2 – Répartition des phrases et textes en tranches d'âges.

contes, romans, magazines et journaux [9]. Ces textes sont annotés avec les indications des éditeurs ou des auteurs sous la forme d'une tranche d'âge A-B et d'un âge $\frac{A+B}{2}$. Les textes pour adultes sont de genres similaires et sont d'un niveau difficile pour des enfants, par exemple des romans avec un langage soutenu, des articles Wikipedia et de journaux sur des sujets avancés (capitalisme, génétique, diplomatie…). Dans l'ensemble, la validité et l'homogénéité des annotations en tranche d'âge sont à considérer avec prudence car les recommandations des éditeurs et auteurs peuvent refléter des motivations autres que psycholinguistiques, par exemple commerciales ou thématiques.

Étant donné la faible quantité de données compte tenu du nombre de paramètres à régler dans nos modèles, nous avons décidé de décomposer les textes en phrases, chacune partageant la même annotation que son texte d'origine. Il s'agit d'une hypothèse forte et manifestement erronée car, par exemple, toutes les phrases d'un texte pour adulte ne sont pas nécessairement inaccessibles à des enfants. La complexité peut venir de certaines phrases spécifiques ou du raisonnement déroulé par leurs articulations réciproques.

Comme le montre la section 6, elle conduit cependant à de bons résultats en pratique. Le corpus ainsi obtenu est composé de 30K phrases et d'environ 446K mots, répartis en ensembles d'apprentissage, de développement et de test à raison de 60, 20 et 20 %. Une partie de l'ensemble test est composé de phrases provenant de 20 textes qui ne sont pas du tout vus dans les autres ensembles. Cette portion

9. Nous n'utilisons pas les encyclopédies dédiées aux enfants comme Wikimini (fr.wikimini.org) ou Vikidia (fr.vikidia.org) car les articles peuvent être écrits par des enfants. Cela introduirait un biais important car les capacités à écrire et à comprendre sont des capacités différentes.

	MAE Âge	Exac.	MAE Âge bas	MAE Âge haut	MAE Âge moyen	Exac.
Naïve	3,44	74,7%	3,59	3,30	3,44	74,7%
A	2,06	–	2,12	2,08	2,09	–
B	**2,01**	**84,7%**	**2,02**	**1,99**	**2,00**	**85,3%**
C	2,12	84,0%	2,11	2,09	2,09	84,2%

(a) Ensemble de développement

	MAE Âge	MAE Âge bas	MAE Âge haut	MAE Âge moyen
Naïve	3,67	3,74	3,61	3,67
A	**2,24**	**2,26**	**2,29**	**2,27**
B	2,26	2,27	2,30	**2,27**
C	2,31	2,29	2,33	2,30

(b) Ensemble de test

TABLE 1 – MAE et exactitude pour les approches naïve, A, B et C.

« vierge » a une distribution différente en terme de tranches d'âges. Sur l'ensemble d'apprentissage, l'âge moyen est d'environ $10,26$, tandis que la tranche d'âge moyenne est de $8,33$-$12,19$. Sur la portion vierge, ces valeurs sont différentes : $9,01$ et $7,54$-$10,48$.

6 Résultats

Tous les modèles sont entraînés sur l'ensemble d'entraînement en utilisant l'ensemble de développement pour éviter un surapprentissage. La fonction de coût est l'erreur quadratique moyenne pour la régression et l'entropie croisée binaire pour la classification. L'algorithme d'optimisation est Adam, avec 500 époques et une taille de lot de 256 phrases. Après avoir comparé les MAE sur l'ensemble de développement, il apparaît que les meilleurs résultats sont rapportés avec ReLU et $n = 6$ couches cachées de $h = 200$ unités, sans *dropout*. Pour le modèle B, le meilleur nombre de couches spécifiques est de 3 lorsque l'on considère des âges, de 4 pour des tranches d'âge.

Le tableau 1 présente les résultats des modèles A, B et C sur les ensembles de développement (a) et de test (b). Les colonnes de gauche montrent les résultats lorsqu'un âge unique est directement prédit, celle de droite quand une tranche d'âge est prédite. Dans ce second cas, les MAE pour chaque borne sont également signalés, ainsi que celle avec le barycentre de la tranche. Sur l'ensemble de développement, l'exactitude de la classification adultes/enfants est également fournie pour les modèles B et C. Tous les modèles sont comparés à l'approche naïve qui prédit constamment les valeurs moyennes observées dans l'ensemble d'apprentissage (*cf.* section 5). Dans l'ensemble, tous les modèles surpassent clairement cette approche naïve. Ensuite, bien que le modèle B fonctionne un peu mieux sur l'ensemble de développement, la différence avec le modèle A disparaît sur l'ensemble de test. Enfin, il semble que prédire des âges ou des tranches d'âges ne change pas grand chose.

Pour affiner ces résultats, nous avons fait annoter la portion vierge de l'ensemble de test par trois psycholinguistes spécialisés dans le développement des enfants. Ces annotations ont été effectuées soit sur l'ensemble des 20 textes concernés, soit sur 80 phrases isolées tirées aléatoirement. Le tableau 2.a donne les résultats, incluant ceux des experts pris individuellement ou après moyennage de leurs prévisions. Sur les phrases, il apparaît que les prédictions de notre modèle sont meilleures que celles des experts, même en essayant de trouver un consensus entre elles (dernière ligne). Lorsque plus de contexte est donné aux experts à travers des textes complets (b), leurs recommandations tendent à battre celles de notre modèle lorsque ses prédictions sont comptabilisées phrase par phrase (sauf pour un expert). Cependant, un simple calcul de moyenne de ces prédictions phrase par phrase amène à une amélioration substantielle (ligne "par texte") et à des prédictions meilleures que celles des experts. Outre les conclusions positives concernant notre approche de recommandation automatique,

	Âge	Âge bas	Âge haut	Âge moyen
Naïve	4,46	4,29	4,63	4,46
Modèle B	**2,70**	**2,53**	**2,65**	**2,57**
Expert 1	3,14	2,95	3,45	3,14
Expert 2	3,38	3,48	3,39	3,38
Expert 3	3,07	2,93	3,54	3,07
Moy. experts	2,88	2,86	3,05	2,88

(a) 80 phrases isolées

		Âge	Âge bas	Âge haut	Âge moyen
Naïve		4,57	4,32	4,83	4,57
Modèle B	Par phrase	3,13	3,07	3,35	3,18
	Par texte	**2,39**	**2,51**	**2,57**	**2,53**
Expert 1		2,60	2,60	2,80	2,60
Expert 2		3,50	3,80	3,30	3,50
Expert 3		2,70	2,90	2,60	2,70
Moy. experts		2,95	3,19	2,81	2,95

(b) 20 textes

TABLE 2 – Comparaisons entre notre approche et les experts sur une portion de l'ensemble de test.

ces résultats reflètent sans doute aussi un décalage quant à la notion de tranche d'âge entre les experts psycholinguistiques et les éditeurs ou auteurs. L'analyse des prédicitions montrent en effet que les experts tendent à utiliser une plage d'âges restreinte par rapport à celle autorisée (et utilisée par nos modèles), à savoir 4-13 ans *versus* 0-18 ans.

7 Conclusion et perspectives

Dans cet article, nous avons étudié la tâche originale de recommandation d'âge pour des textes. Plusieurs modélisations ont été proposées et les résultats ont été comparés aux recommandations de psycholinguistes. Ces résultats montrent que les prévisions de nos modèles sont meilleures que celles des experts. En outre, tout en s'appuyant sur une hypothèse forte selon laquelle toutes les phrases d'un texte peuvent être considérées comme toutes associées à une tranche d'âge unique, nos résultats sur l'agrégation des résultats au niveau des phrases sont clairement encourageants. Cela démontre la viabilité de l'approche et appelle des investigations supplémentaires.

Cependant, nous sommes conscients que ces résultats doivent être pris avec prudence car différents aspects portent de l'incertitude. En particulier, la précision et le bien-fondé scientifique des recommandations fournies par les éditeurs et les auteurs sont sans doute parfois discutables. Ensuite, considérer les erreurs absolues moyennes avec un âge cible est probablement trop dur. Il serait intéressant de comparer des tranches d'âges et non des âges dans l'évaluation. Enfin, il serait intéressant de corréler les résultats avec une campagne d'évaluation *in situ* auprès des enfants. Ceci est prévu dans les prochains mois.

Remerciements

Ce travail a bénéficié du soutien financiers des projets ANR TREMoLo et TextToKids.

Références

AGUERT M., BERNICOT J. & LAVAL V. (2009). Prosodie et compréhension des énoncés chez les enfants de 5 à 9 ans. *Enfance*, **3**.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, **43**(3).

BLANC N. (2010). La compréhension des contes entre 5 et 7 ans : Quelle représentation des informations émotionnelles ? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **64**(4).

CANDITO M., NIVRE J., DENIS P. & ANGUIANO E. H. (2010). Benchmarking of statistical dependency parsers for french. In *Proceedings of the International Conference on Computational Linguistics : Posters*, COLING '10 : Association for Computational Linguistics.

DALE E. & CHALL J. S. (1948). A formula for predicting readability : Instructions. *Educational research bulletin*, **27**.

DAVIDSON D. (2006). The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, **30**(3).

DE BELDER J. & MOENS M.-F. (2010). Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*.

FAUCONNIER J.-P. (2015). French word embeddings. http://fauconnier.github.io.

FLESCH R. (1948). A new readability yardstick. *Journal of applied psychology*, **32**(3).

FRANÇOIS T. (2015). When readability meets computational linguistics : a new paradigm in readability. *Revue française de linguistique appliquée*, **20**(2).

FRANÇOIS T. & FAIRON C. (2012). An "AI readability" formula for french as a foreign language. In *Proceedings of EMNLP-CoNLL*.

FRITH U. (1985). Beneath the Surface of Developmental Dyslexia. In K. PATTERSON, J. C. MARSHALL & M. COLTHEART, Éds., *Surface Dyslexia. Neuropsychological and Cognitive Studies of Phonological Reading*, Psychology Library Editions : Psychology of Reading, chapitre 13. Routledge. DOI : 10.4324/9781315108346-18.

GALA N., FRANCOIS T., JAVOUREY-DREVET L. & ZIEGLER J. C. (2018). La simplification de textes, une aide à l'apprentissage de la lecture. *Langue française*, **199**(3/2018).

GATHERCOLE S. (1999). Cognitive approaches to the development of short-term memory. *Trends in cognitive sciences*, **3**.

GEYER L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, **22**(5).

GROMER D. & WEISS M. (1990). *Lire, tome 1 : apprendre à lire*. Armand Colin.

HICKMANN M. (2012). Diversité des langues et acquisition du langage : espace et temporalité chez l'enfant. *Langages*, **188**(4/2012).

MOUW J. M., VAN LEIJENHORST L., SAAB N., DANEL M. S. & VAN DEN BROEK P. (2019). Contributions of emotion understanding to narrative comprehension in children and adults. *European Journal of Developmental Psychology*, **16**(1).

PIOLAT A. & BANNOUR R. (2009). An example of text analysis software (emotaix-tropes) use : The influence of anxiety on expressive writing. *Current psychology letters. Behaviour, brain & cognition*, **25**(2).

TARTAS V. (2010). Le développement de notions temporelles par l'enfant. *Développements*, **4**.

VION M. & COLAS A. (1999). L'emploi des connecteurs en français : contraintes cognitives et développement des compétences narratives (le cas de la narration de séquences arbitraires d'événements). In *Proceedings of the Conference of the International Association for the Study of Child Language*.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Traduire des corpus pour construire des modèles de traduction neuronaux : une solution pour toutes les langues peu dotées ?

Raoul Blin
CNRS-CRLAO, 105 Bd Raspail, 75006 Paris, France
blin@ehess.fr

RÉSUMÉ

Nous comparons deux usages des langues pivots en traduction automatique neuronale pour des langues peu dotées. Nous nous intéressons au cas où il existe une langue pivot telle que les paires source-pivot et pivot-cible sont bien ou très bien dotées. Nous comparons la traduction séquentielle traditionnelle (source→pivot→cible) et la traduction à l'aide d'un modèle entraîné sur des corpus traduits à l'aide des langues pivot et cible. Les expériences sont menées sur trois langues sources (espagnol, allemand et japonais), une langue pivot (anglais) et une langue cible (français). Nous constatons que quelle que soit la proximité linguistique entre les langues source et pivot, le modèle entraîné sur corpus traduit a de meilleurs résultats que la traduction séquentielle, et bien sûr que la traduction directe.

ABSTRACT

Corpus Translation to Build Translation Models : a Solution for all Low-Resource Languages ?

We compare two uses of pivot languages in neural machine translation when a language pair is low or not resourced, but there is a pivot language such that the source-pivot and pivot-target pairs are well or very well resourced. We compare traditional sequential translation (source→pivot→target) and translation using a model trained on corpora translated with the pivot and target languages. Experiments conducted with three source languages (Spanish, German and Japanese), a pivot language (English) and a target language (French), show that the trained model on translated corpus has better results than sequential translation, and of course than direct translation.

MOTS-CLÉS : langues peu dotées, traduction automatique neuronale, langue pivot.

KEYWORDS: low-resource languages, neural machine translation, pivot language .

1 Introduction

Pour profiter des avantages de la traduction automatique neuronale, il est nécessaire de disposer de très volumineux corpus (Koehn & Knowles, 2017). Malheureusement, peu de paires de langues disposent de telles ressources. Une première solution très étudiée actuellement est le recours à des modèles multilingues de traduction (Rikters *et al.*, 2018), etc.). L'inconvénient est que l'entraînement nécessite plus de données (grand corpus) et plus de puissance de calcul. Cette solution s'inscrit d'ailleurs dans une tendance générale observable tant dans le domaine académique que dans l'industrie, d'un intérêt croissant pour des systèmes toujours plus puissants (Aharoni *et al.*, 2019; Arivazhagan *et al.*, 2019). Pourtant, il existe aussi un besoin pour des systèmes «frugaux» avec de faibles capacités de calcul. Il est alors important de limiter la taille des corpus.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 172–180.
HAL : hal-02784765.

Il existe quelques alternatives à cette course à la puissance pour traiter les paires de langues peu dotées. Une première piste récemment proposée est l'apprentissage par transfert de modèles existants (Kocmi & Bojar, 2019). Une autre solution est le recours aux langues pivots, technique ancienne en traduction mais qui revient sous de nouvelles formes grâce à la traduction automatique neuronale. La pratique traditionnelle consiste à traduire séquentiellement, de la langue source vers la langue pivot puis vers la langue cible. C'est une méthode encore appliquée même dans les systèmes commerciaux (Blin, 2018b). L'inconvénient majeur de la traduction séquentielle automatique avec des systèmes qui restent encore imparfaits, c'est que des informations peuvent être perdues lors de la traduction de la source vers le pivot.

Pour résoudre ce problème, (Currey & Heafield, 2019) ont proposé d'utiliser le pivot pour traduire les corpus eux-mêmes et produire des corpus synthétiques sur lesquels des modèles de traduction directe sont entraînés. Avant de décrire la procédure, nous convenons des abréviations suivantes. a, a' et a^i désignent des corpus monolingues pour une langue donnée «a». a^n et a^m désignent des corpus distincts d'une même langue. $m(a, b)$ désigne un modèle entraîné sur un corpus aligné a-b pour traduire de a vers b. «src», «piv» et «tgt» désignent respectivement les langues source, pivot et cible. La procédure est la suivante. Elle n'est envisageable que si il existe deux corpus «de base» (non traduits) alignés piv^1-src et piv^2-tgt, et un corpus monolingue piv^0. Supposons que l'on veuille traduire d'une langue source src vers une langue cible tgt. On génère un corpus src' en langue source en traduisant piv^0 grâce au modèle $m(piv^1, src)$ (voir Fig.1). En parallèle, on génère un corpus tgt' en langue cible en traduisant ce même corpus piv^0 à l'aide du modèle $m(piv^2, tgt)$. On dispose ainsi d'un corpus synthétique aligné bilingue src'-tgt'. Il est utilisé pour entraîner un modèle de traduction directe de la source vers la cible (par abus de langage et pour faire simple, nous parlerons de «modèle traduit»). Les auteurs ajoutent au corpus synthétique les corpus originaux piv^1-src et piv^2-tgt et obtiennent ainsi un corpus synthétique et multilingue. Cette technique a été utilisée pour traduire la paire russe-allemand avec l'anglais comme pivot et produit des résultats prometteurs (BLEU ≈ 23). Le premier intérêt est que cette technique fonctionne pour une paire de langue sans corpus aligné. Il devrait donc a fortiori fonctionner pour une paire de langues peu dotée. L'autre intérêt est que le corpus synthétique peut être aussi grand que le corpus monolingue.

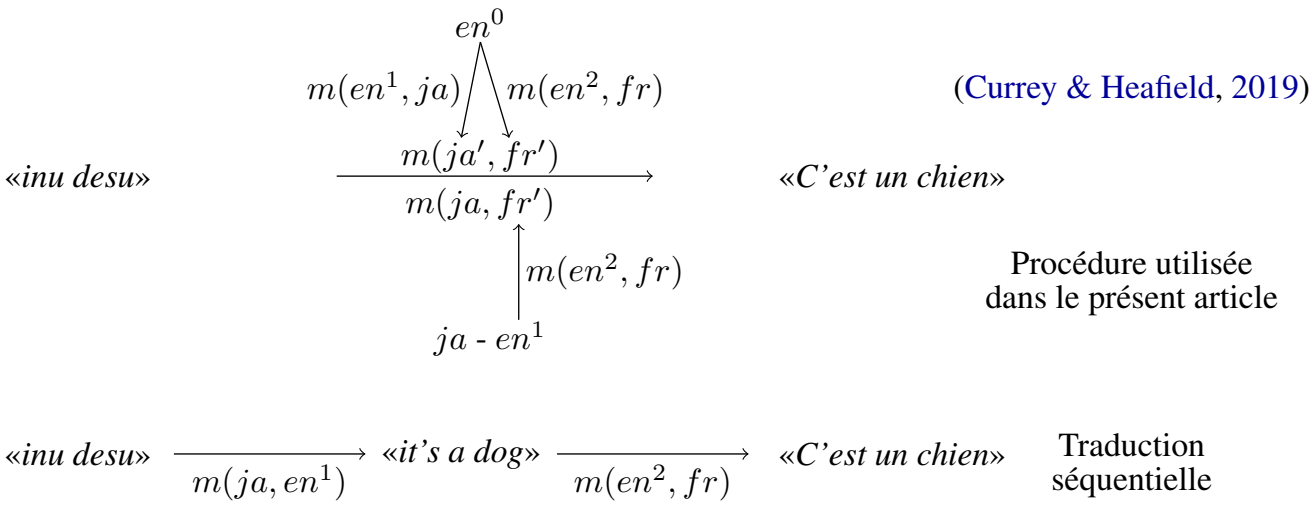


FIGURE 1: Trois procédés de traduction avec langue pivot ; ici, les langues source, pivot et cible sont repectivement le japonais, l'anglais et le français ; $a \xrightarrow{m(x,y)} b$ désigne une traduction de a vers b à l'aide du modèle $m(x, y)$.

Nous nous sommes demandé si cette technique produisait une meilleure traduction que la méthode séquentielle, et dans quelle mesure la proximité entre les langues manipulées influençait les résultats. Pour répondre à cette question, nous avons effectué des comparaisons en jouant sur plusieurs variables : la proximité linguistique des langues source et pivot (les langues pivot et cible restant toujours les mêmes), la part de vocabulaire commun entre les corpus piv^1 et piv^2, et la qualité des traductions. Contrairement à (Currey & Heafield, 2019), nous n'avons pas intégré le multilinguisme car il pouvait affecter différemment la traduction séquentielle et la traduction par modèle traduit.

L'article est organisé en deux parties. Dans un premier temps, nous décrivons les ressources utilisées et la méthode. Puis nous présentons les résultats et les discutons.

2 Expériences

2.1 Langues observées et corpus

Pour prendre en compte la proximité linguistique et la part de vocabulaire commun tout en limitant le nombre d'expériences à mener, nous construisons un ensemble de corpus pour chaque cas extrême : un premier ensemble pour des langues sources proches de la langue pivot et un maximum de vocabulaire en commun, et un second ensemble avec une langue source éloignée et un moindre volume de vocabulaire en commun.

Nous menons les observations pour trois langues sources : espagnol (es), allemand (al) et japonais (ja). La langue cible est le français. La langue pivot est l'anglais. L'anglais a été choisi pour des questions pratiques : il existe un corpus moyen ou grand pour toutes les paires de langues incluant l'anglais et les langues ci-dessus. Les langues sources sont choisies de sorte que la proximité linguistique avec la langue pivot soit progressive. La plus éloignée est le japonais (de type SOV avec marqueurs casuels, pro-drop, non indo-européen, langue agglutinante, écriture différente). L'espagnol est la plus proche (langue SVO, indo-européenne, non agglutinante). L'allemand se situe entre les deux (partiellement SOV, indo-européenne, non agglutinante).

Nous constituons tout d'abord un corpus pour entraîner un modèle de traduction directe, de la source vers la cible. La paire de langues ja-fr est réellement peu dotée ($\approx$ 50K phrases.) (Blin, 2018a). {al,es}-fr sont moyennement ou même bien dotées. Pour simuler des paires faiblement dotées avec l'espagnol et l'allemand, nous utilisons des corpus d'approximativement la même taille que le corpus ja-fr. Nous extrayons ces corpus d'Europarl (Koehn, 2005).

Nous nous dotons ensuite de corpus de taille moyenne ($\approx$ 400-900K phrases) pour entraîner des modèles de la langue source vers la langue pivot. Pour le japonais-anglais, nous avons rassemblé l'ensemble des corpus librement disponibles. Les thématiques sont très variées. Les corpus $\{al, es\}$-en sont à nouveau extraits d'Europarl. Enfin nous nous dotons d'un grand corpus (14 millions de phrases) pour la traduction de la langue pivot vers la langue cible. Ce corpus intègre Europarl, ce qui permet d'augmenter le vocabulaire commun avec les corpus $\{al, es\}$-en de langues européennes. Au final presque 100% du vocabulaire des corpus $\{al, es\}$-en est présent dans le corpus en-fr contre seulement 40% pour le corpus ja-en.

Pour éviter les biais causés par des structures phrastiques trop différentes d'un corpus à l'autre, nous limitons les observations à des phrases structurellement «standards». L'autre intérêt de cette restriction, c'est que les segments du corpus test (PUD, voir plus loin) sont majoritairement des

Langues	Corpus	nb phrases.	content	voc. src	voc. tgt
ja-fr	rbjafr-191204	51 659	varia	32K	27K
de-fr	rbeurdefr-1.0	52 175	Europarl	28 424	43 332
es-fr	rbeuresfr-1.0	52 807	Europarl	29 706	33 149
ja-en	rbjaen-191023	752 518	varia	150 004	137 415
de-en	rbeurdeen-1.0	973 557	Europarl	150002	56447
es-en	rbeuresen-1.0	491 387	Europarl	49144	86038
en-fr	rbenfr-192710	14 048 795	varia*	170004	170002
*ja-en***	rbjaen-191023	idem	idem	150004	137415
*es-en***	rbeuresen-1.0	idem	idem	49144	86038
*en-fr***	rbenfr-191124	14 048 795	varia*	150004	150002

TABLE 1: Corpus utilisés (* inclu Europarl ; ** pour entraîner les modèles améliorés)

.

phrases «standards». Cette sélection permet donc d'harmoniser les corpus d'entraînement et le corpus d'évaluation. Nous n'utilisons que des ressources librement disponibles et de bonne qualité, sans bruit. En plus, nous retenons les phrases qui finissent par un signe de ponctuation («. ? !») et contiennent moins de 10 chiffres et moins de 20 majuscules. Ce critère simple suffit à éliminer les phrases non standards sans trop d'erreurs. Ceci a réduit considérablement le nombre de phrases exploitées par rapport au nombre initialement disponibles, en particulier pour l'anglais-français.

Les techniques de pré et post-traitement des corpus varient d'une paire de langues à l'autre et peuvent biaiser les comparaisons. Pour uniformiser le traitement, nous utilisons un pré et post traitement a minima, avant tout pensés pour réduire la taille du vocabulaire. Les corpus ont été prétraités simplement en ajoutant une marque de début de phrase à chaque phrase, et en transformant les majuscules en minuscule en tête de phrase. Pour limiter la taille du vocabulaire et en particulier le nombre de mots inconnus, nous avons décomposé et balisé les numéraux et abréviations ("123" → "<num> 1 2 3 </num>"). Bien qu'elle permette de sensiblement réduire la taille du vocabulaire, nous n'avons pas utilisé de segmentation de type BPE (Sennrich *et al.*, 2016). Nous avons en effet estimé qu'elle aurait biaisé les comparaisons en favorisant trop les paires de langues utilisant un même système d'écriture.

Pour faciliter la comparaison avec d'autres études incluant le japonais (par exemple (Sekizawa *et al.*, 2017)), nous avons segmenté le corpus japonais en utilisant l'analyseur morphologique MeCab [1] et le dictionnaire IPADIC (Asahara & Matsumoto, 2003) [2]. Les deux sont très utilisés pour ce type de travail.

Pour l'évaluation, nous avons utilisé le corpus multilingue PUD [3] (1000 phrases). Le contenu est thématiquement très varié. Pour évaluer les modèles de base (c'est-à-dire non traduits) et les comparer à l'état de l'art, nous avons aussi extrait des corpus test de 1000 phrases, par échantillonnage des corpus d'origine.

La description des corpus de base est donnée dans le tableau 1.

1. Kudou, Taku. "MeCab : Yet Another Part-of-Speech and Morphological Analyzer". taku910.github.io (in Japanese). 23/01/2018.
2. https ://ja.osdn.net/projects/ipadic/docs/ipadic-2.7.0-manual-en.pdf/en/1/ipadic-2.7.0-manual-en.pdf.pdf
3. https ://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2184

2.2 Méthode

La procédure est la suivante (voir aussi le Fig.1) . On construit les modèles de base $m(src, piv^1)$ et $m(piv^2, tgt)$ pour procéder à la traduction séquentielle. Puis on traduit le corpus piv^1 en tgt' à l'aide de $m(piv^2, tgt)$. On évalue alors la qualité de la traduction directe effectuée à l'aide du modèle traduit $m(src, tgt')$.

Pour tenir compte de l'influence de la qualité de la traduction sur les performances des deux techniques de traduction par pivot, nous avons réalisé les expériences à deux reprises avec un même système de traduction, mais des réglages différents. Avec le premier réglage, l'entraînement réclamait moins de temps de calcul mais la qualité des traductions de base était globalement plus modeste. Nous avons refait l'expérience en adoptant des réglages plus gourmants en puissance de calcul, mais produisant des traductions de base de meilleure qualité. Puisque l'allemand a toujours produit des résultats intermédiaires avec la première configuration, nous sommes parti du principe qu'il en serait de même avec la seconde. En conséquence, nous ne l'avons pas observé pour les traductions de qualité améliorée.

Le système de traduction neuronale utilisé est OpenNmt-py (Klein *et al.*, 2017). Pour la configuration «non améliorés», nous avons utilisé les réglages par défaut : réseau neuronal bi-récurrent, 2 couches, RNN de taille 500, word embedding de taille 500, 10 époques, beam de taille 5 pour la traduction. Cette configuration s'est révélée inadaptée pour traiter les corpus de grandes tailles. Pour améliorer les performances des modèles, la configuration «améliorée» différait sur les points suivant : 4 couches, RNN de taille 1000, word embedding de taille 600. En plus, les paramètres du traducteur ont été modifiés : batch de taille 20, beam de taille 8, longueur maximum 150. Les mots inconnus sont remplacés par les mots source qui ont un poids d'attention plus grand. Cette stratégie avantage les langues qui ont un même système d'écriture, et a fortiori un vocabulaire commun.

Nous avons utilisé BLEU (Papineni *et al.*, 2002) pour l'évaluation [4]. Nous avons aussi calculé le score Meteor (Denkowski & Lavie, 2014) mais il n'a pas produit d'information complémentaire. En conséquence, nous ne l'évoquons pas ici. On a aussi observé la fréquence des mots inconnus.

3 Résultats et discussion

La qualité des traductions fournies par les systèmes de base (voir Table 2) sont conformes aux attentes. La traduction directe avec les petits corpus obtient un score très bas. Malgré tout, BLEU est corrélé à la proximité linguistique. Plus les langues sont linguistiquement proches, meilleure est la traduction. Le fait que la qualité des traductions de PUD soit moins bonne que la traduction des corpus tests s'explique par le fait que ce corpus n'a pas de rapport thématique avec le corpus d'entraînement (et donc un faible volume de vocabulaire commun). Pour les traductions impliquant l'anglais, les scores sont inférieurs à l'état de l'art (BLEU de plus de 41.5 pour l'anglais-français dans (Shaw *et al.*, 2018), plus de 27 pour le japonais-anglais (Cromieres *et al.*, 2017) etc). Ces résultats s'expliquent par l'absence d'optimisation tant du côté de la préparation de nos corpus que du côté du réglage du système de traduction (entraînement et traduction). Nous avons certes amélioré la qualité des traductions avec la seconde configuration, mais sans pour autant chercher à rivaliser avec les meilleurs systèmes. En effet, il ne s'agissait pas dans ce travail de surpasser l'état de l'art mais de comparer

4. Nous avons utilisé multi-bleu.perl avec les réglages par défaut (https ://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl)

	Test		PUD		Etat de l'art	
Langues	BLEU	% unk.	BLEU	% unk.	BLEU	source
ja-fr	8,63	0	2,22	0	10,03	(Blin, 2018b)
de-fr	12,86	0	4,47	0	28,63	(Bougares *et al.*, 2019)
es-fr	19,77	0	8,36	0	32,7	(Klein *et al.*, 2017)
ja-en	21,68	0.61	10,16	2.73	27,66	(Cromieres *et al.*, 2017)
de-en	29,64	0	16,15	0	42,5	(Popovic, 2017)
es-en	37,00	0	23,53	0	36,05	(Duma & Menzel, 2018)
en-fr	24,90	0.91	22,15	2.65	41,5	(Shaw *et al.*, 2018)
*ja-en***	23,18	0	10,92	0		
*es-en***	38,00	0	24,97	0		
*en-fr***	33,24	0	25,86	0		

TABLE 2: Evaluation des modèles de base (** configurarion améliorée) ; Comparaison à l'état de l'art tous corpus confondus (et à l'exclusion du corpus PUD).

deux techniques pour des niveaux égaux de préparation des corpus et de performances des modèles.

Sans surprise, quelle que soit la qualité des traductions, la traduction par pivot (voir Tables 3 et 4) obtient de meilleurs scores que la traduction directe avec corpus de base. Nous attribuons ce résultat à la différence de taille des corpus puisque le corpus pour la traduction directe est au moins 8 fois plus petit que tous les autres corpus). Le gain est plus important pour les langues européennes (proximité et corpus en commun plus grand). Mais les résultats restent toutefois faibles. (voir Fig.2)

Source	BLEU	%<unk>	gain
ja	6,74	3.33	4.52
de	10,26	0.38	5.79
es	13,49	1.38	5.13

TABLE 3: Traduction séquentielle du corpus PUD (voir Table 1 pour les traductions source→pivot correspondantes) ; gain (nombre de points) par rapport à la traduction directe ; configuration non améliorée.

Source	BLEU	%<unk>	gain
ja	6,29	5.18	4,07
de	10,05	4.42	5,58
es	14,00	5.32	5,64

TABLE 4: Evaluation des «modèles traduits» ; gain (en nombre de points) par rapport à la traduction directe ; configuration non améliorée.

Comparons les deux méthodes de traduction avec langue pivot (voir Tables 3 et 4, Fig.2). Avec une traduction de qualité modeste, la méthode séquentielle et la méthode par modèle traduit sont très proches. Par contre, on observe une différence au niveau des mots inconnus, plus nombreux avec la traduction par modèle traduit.

Si l'on augmente la qualité des traductions, les deux techniques n'ont plus la même efficacité (Table 5 et Fig.2). La méthode par modèle traduit est meilleure ($\approx +3$). Ce constat est le même quelle que soit la proximité linguistique entre les langues source et la langue pivot, et quelle que soit la quantité de vocabulaire commun entre les deux corpus piv^1 et piv^2. Par extrapolation, nous pouvons avancer que ce résultat vaut quelle que soit la distance thématique entre le corpus d'entraînement et le corpus d'évaluation. La qualité des traductions est malheureusement trop basse pour faire une comparaison qualitative pertinente. On observe des erreurs dans tous les domaines, lexico, syntaxiques et logiques.

	Source	en	fr
Traduction séquentielle	ja	10,92	8,38
	es	24.79	16.76
Model traduit	ja	_	10,11
	es	_	19,79

TABLE 5: Score BLEU pour la traduction séquentielle et par «modèle traduit», configuration améliorée

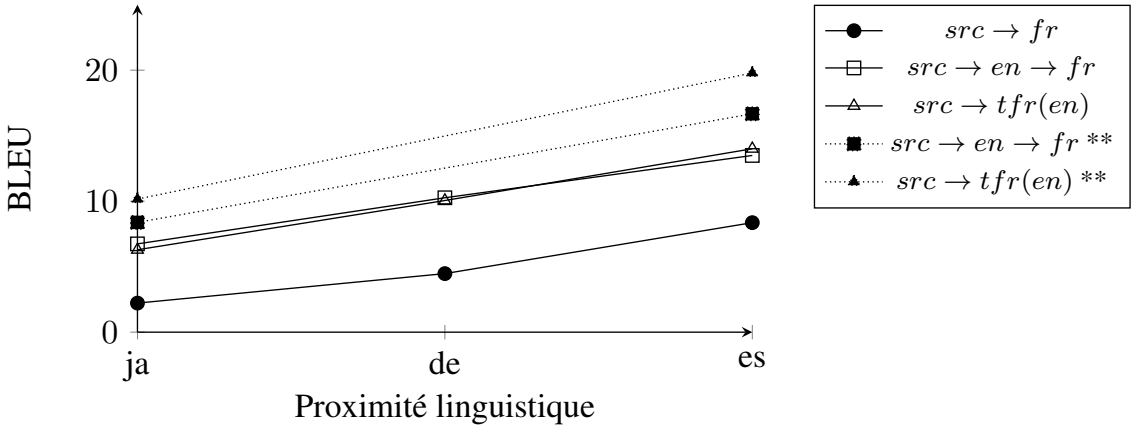


FIGURE 2: Traduction séquentielle et par modèle traduit (** Configuration améliorée)

4 Conclusion

Nos résultats montrent qu'utiliser un «modèle traduit» produit de meilleurs résultats qu'une traduction séquentielle par pivot, sous réserve que les modèles de traduction source→pivot et pivot→cible soient de bonne qualité. Sinon, la traduction séquentielle reste compétitive. Ces résultats confirment les observations de (Currey & Heafield, 2019). Ils montrent en plus que les résultats vont dans le même sens quelles que soient la proximité entre les langues source et cible et la quantité de vocabulaire commun entre les corpus. Enfin, ils montrent que la supériorité de la traduction par «modèle traduit» est vraie même sans recourir à un corpus multilingue.

Remerciements

Nous remercions chaleureusement le Centre de Calcul CNRS / IN2P3 (Lyon - France) pour la mise à disposition des moyens informatiques nécessaires à ces travaux.

Références

AHARONI R., JOHNSON M. & FIRAT O. (2019). Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Pa-*

pers), p. 3874–3884, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1388.

ARIVAZHAGAN N., BAPNA A., FIRAT O., LEPIKHIN D., JOHNSON M., KRIKUN M., CHEN M. X., CAO Y., FOSTER G., CHERRY C., MACHEREY W., CHEN Z. & WU Y. (2019). Massively Multilingual Neural Machine Translation in the Wild : Findings and Challenges. arXiv : 1907.05019.

ASAHARA M. & MATSUMOTO Y. (2003). Ipadic user manual. ipadic-2.7.0-manual-en.pdf.pdf.

BLIN R. (2018a). Automatic evaluation of alignments without using a gold-corpus - example with french-japanese aligned corpora. In S. KIYOAKI, Éd., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France : European Language Resources Association (ELRA). 23_W29.pdf.

BLIN R. (2018b). Traduction automatique du japonais vers le français : Bilan et perspectives. In *Traitement Automatique du Langage Naturel*, Rennes, France. HAL : hal-01796313.

BOUGARES F., WOTTAWA J., BAILLOT A., BARRAULT L. & BARDET A. (2019). LIUM's Contributions to the WMT2019 News Translation Task : Data and Systems for German-French Language Pairs. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, p. 129–133, Florence, Italy : Association for Computational Linguistics.

CROMIERES F., DABRE R., NAKAZAWA T. & KUROHASHI S. (2017). Kyoto university participation to wat 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, p. 146–153, Taipei, Taiwan : Asian Federation of Natural Language Processing.

CURREY A. & HEAFIELD K. (2019). Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, p. 99–107, Hong Kong : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5610.

DENKOWSKI M. & LAVIE A. (2014). Meteor Universal : Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

DUMA M.-S. & MENZEL W. (2018). Translation of Biomedical Documents with Focus on Spanish-English. In *Proceedings of the Third Conference on Machine Translation : Shared Task Papers*, p. 637–643, Belgium, Brussels : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6444.

KLEIN G., KIM Y., DENG Y., SENELLART J. & RUSH A. (2017). OpenNMT : Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, p. 67–72, Vancouver, Canada : Association for Computational Linguistics.

KOCMI T. & BOJAR O. (2019). Transfer Learning across Languages from Someone Else's NMT Model. arXiv : 1909.10955.

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X*, p. 79–86.

KOEHN P. & KNOWLES R. (2017). Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, p. 28–39, Vancouver : Association for Computational Linguistics. DOI : 10.18653/v1/W17-3204.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*, p. 311–318.

POPOVIC M. (2017). Comparing Language Related Issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, **108**. DOI : 10.1515/pralin-2017-0021.

RIKTERS M., PINNIS M. & KRIŠLAUKS R. (2018). Training and Adapting Multilingual NMT for Less-resourced and Morphologically Rich Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

SEKIZAWA Y., KAJIWARA T. & KOMACHI M. (2017). Improving Japanese-to-English Neural Machine Translation by Paraphrasing the Target Language. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, p. 64–69, Taipei, Taiwan : Asian Federation of Natural Language Processing.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

SHAW P., USZKOREIT J. & VASWANI A. (2018). Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 464–468, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2074.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Construction de plongements de concepts médicaux sans textes

Vincent Claveau
Univ. de Rennes, CNRS, IRISA
Campus de Beaulieu
35042 Rennes, France
vincent.claveau@irisa.fr

RÉSUMÉ

Dans le domaine médical, beaucoup d'outils du TAL reposent désormais sur des plongements de concepts issus de l'UMLS. Les approches existantes pour générer ces plongements nécessitent de grandes quantités de documents médicaux. Au contraire des ces approches, nous proposons dans cet article de nous appuyer sur les traductions en japonais, plus précisément en kanjis, disponibles dans l'UMLS pour générer ces plongements. Testée sur différents jeux d'évaluation proposés dans la littérature, notre approche, qui ne requiert donc aucun texte, donne de bons résultats comparativement à l'état-de-l'art. De plus, nous montrons qu'il est intéressant de les combiner avec les plongements – contextuels – existants.

ABSTRACT

Embedding medical concepts without texts.

In the medical field, many TAL tools are now based on embeddings of concepts from the UMLS. Existing approaches to generate these embeddings require large amounts of medical data. Contrary to these approaches, we propose in this article to rely on Japanese translations of the concepts, more precisely in Kanjis, available in the UMLS to generate these embeddings. Tested on different evaluation tasks proposed in the literature, our approach, which therefore requires no text, yields good results compared to the state of the art. Moreover, we show that it is interesting to combine them with existing — contextual-based — embeddings.

MOTS-CLÉS : TAL médical, plongements de concepts, CUI, UMLS, kanjis.

KEYWORDS: Biomedical NLP, concept embedding, CUI, UMLS, kanjis.

1 Introduction

De nombreuses techniques de TAL reposent désormais sur les réseaux de neurones ; ceux-ci nécessitent en entrée des représentations numériques des mots – les plongements de mots. Le domaine médical ne fait pas exception et la plupart des travaux exploitent donc désormais des approches neuronales. La question de la construction de plongements adaptés au vocabulaire médical a donc été posée depuis quelques années.

Plus particulièrement, plusieurs travaux se sont intéressés à produire des plongements des concepts médicaux identifiés dans l'UMLS (CUI - *Concept Unique Identifiers*). Le MetaTheusaurus de l'UMLS (Tuttle *et al.*, 1990) recense en effet des entités et des termes médicaux (termes simples ou complexes, voire portion de phrases) dans de nombreuses langues, et rattachés à ces identifiants uniques de

Volume 2 : Traitement Automatique des Langues Naturelles, pages 181–188.
HAL : hal-02784766.

concepts.

Dans cet article, nous revenons sur la question de ces représentations de CUI en proposant une nouvelle approche pour leur construction. Au contraire des travaux existants qui exploitaient de très grosses quantités de textes du domaine pour produire les plongements de CUI, notre approche ne nécessite aucun texte d'entraînement et repose uniquement sur l'UMLS et en particulier les traductions en kanjis. Deux versions de cette approche, l'une produisant une représentation creuse et l'autre dense, sont ici décrites et comparées aux plongements existants. Enfin, nous proposons une combinaison de nos plongements et de ceux de l'état-de-l'art, reposant sur une approche distributionnelle standard.

Nous revenons dans la section suivante sur les travaux existants cherchant à produire des plongements de CUI et sur ceux exploitant les propriété des kanjis. Nous présentons ensuite successivement nos deux représentations, dont les performances sont comparées à l'état-de-l'art en section 4. Nous terminons en donnant quels pistes de travaux qui nous semblent prometteuses.

2 Travaux connexes

L'UMLS est une ressource largement utilisée pour des tâches de TAL biomédical. Son MetaThesaurus rassemble de nombreuses terminologies dans plusieurs langues qui sont agrégées grâce à des identifiants de concepts : chaque terme est rattaché à un identifiant de concept ou CUI. Les CUI permettent au sein d'une langue de trouver les variantes d'un terme (plusieurs termes de la même langue ayant un même CUI), de même qu'ils permettent de trouver des traductions (un CUI partagé par des termes de langues différentes). Par ailleurs, les CUI sont rattachés à des types sémantiques organisés hiérarchiquement.

La construction de plongements de CUI a fait l'objet de plusieurs travaux (De Vine *et al.*, 2014; Choi *et al.*, 2016; Beam *et al.*, 2020). Tous adoptent une approche similaire et les outils usuels de cette tâche, tels que Glove (Pennington *et al.*, 2014) ou Word2vec (Mikolov *et al.*, 2013). Ainsi, ils sont tous construits à partir de textes médicaux anglais. La seule particularité, par rapport à un plongement lexical standard, est la phase de pré-traitement permettant de repérer les termes de l'UMLS dans les phrases et de les associer à leur CUI. Cette étape est faite avec l'outil Metamap (Aronson, 2001).

Il est important de noter plusieurs points communs à ces travaux. Ils reposent tous sur l'anglais, du fait de l'outil MetaMap qui n'existe que pour l'anglais, et de la disponibilité de quantités de textes cliniques.

Ainsi, De Vine *et al.* (2014) ont utilisé 350 000 résumés de journaux du domaine médical pour générer avec word2vec (skip-gram) des plongements pour 60 000 concepts (différentes dimensionnalités ont été testées, 200 étant celle obtenant les meilleures performances).

Choi *et al.* (2016) utilisent également word2vec et une SVD sur 4 millions de donnés d'assurances santé et 20 millions de notes cliniques pour générer 28 000 plongements de dimension 200.

Enfin, les plongements les plus performants et les plus utilisés sont CUI2vec (Beam *et al.*, 2020). Il s'agit de plongements dans $\mathbb{R}^{500}$ de 110 000 CUI qui ont nécessité une quantité gigantesque de textes (plus de 80 millions de documents, incluant ceux des travaux précemmet cités) pour être générés.

Utiliser des représentations par kanjis de termes médicaux a déjà été fait dans un contexte d'analyse morphologique (Claveau & Kijak, 2013). Dans ce travail, les kanjis servaient d'indices pour aider à

décomposer les termes morphologiquement complexes (eg. photo|chimio|thérapie). Bien que pour des buts différents, notre approche repose sur la même idée d'utiliser les kanjis comme des éléments atomiques de sens.

3 Plongements par kanjis

Comme nous l'avons expliqué, l'idée directrice de notre travail est d'exploiter les termes en kanjis pour s'en servir comme des représentations sémantiques des concepts (CUI). De telles traductions sont disponibles directement dans le MetaThesaurus de l'UMLS dans lequel des termes en plusieurs langues sont rattachés aux identifiants de concept CUI. L'intérêt des kanjis est qu'ils offrent une représentation ayant une portée sémantique (chaque kanji pris isolément a un – ou plusieurs – sens), ne sont pas soumis à de la variation morphologique, et sont en nombre limités (moins de 2 000 dans l'UMLS) comparativement aux mots en anglais par exemple. Dans l'UMLS version 2019AB que nous utilisons pour les expériences, il y a 72 000 CUI ayant au moins un terme en japonais. Par exemple, le concept identifié par le CUI C0031740, qui concerne la photochimiothérapie, a plusieurs réalisations en kanjis dans l'UMLS : 光化学法, 光力学的法, 光力学治, 光力学的治, 光力学的治, 光力学的治法.

3.1 Plongements creux

Le premier plongement que nous proposons est simplement la projection des CUI dans l'espace des kanjis. C'est-à-dire qu'un CUI est décrit par le vecteur pondéré construit à partir du sac de kanjis des termes relevant de ce CUI. Le concept C0031740 est donc représenté par un vecteur nul sauf aux dimensions correspondant aux kanjis 光, 化, 学, , 法, 力, 治, ... L'espace de représentation, après suppression des kanjis n'apparaissant qu'une fois, est $\mathbb{R}^{1462}$.

Plusieurs pondérations ont été testées : binaire (ou *one-hot*), Hellinger, TF, TF-IDF, Okapi-BM25 (Robertson *et al.*, 1998). Pour des raisons de place, seule la plus performante, Okapi-BM25, est rapportée ici. Dans les expériences rapportées dans la section suivante, cette approche est appelée approche par kanjis.

3.2 Plongements denses

Les vecteurs obtenus par l'approche précédente sont de dimension importante, mais creux (beaucoup de dimensions à 0). Il peut être intéressant d'en proposer une version dense et de dimension inférieure, notamment si la perspective est d'utiliser ces plongements dans des réseaux de neurones pour une application donnée.

Pour ce faire, on recourt à des techniques de réductions de dimensions (Sarveniazi, 2014). Plusieurs techniques standard on été testées. Pour des raisons de place nous ne rapportons les résultats que pour l'Analyse en Composantes Principales (ACP) qui a obtenu parmi les meilleurs résultats avec des temps de calcul courts. Sauf indication contraire, dans les expériences rapportées ci-dessous, nous avons choisi de réduire en 500 dimensions, à des fins de comparaison avec CUI2vec. Dans les expériences rapportées dans la section suivante, cette approche est appelée approche par ACP.

3.3 Fusion de plongements

Les plongements que nous proposons reposent sur des indices très différents des indices contextuelles habituellement exploités par Glove, word2vec ou autre. Notre approche ne repose donc pas sur l'hypothèse distributionnelle pour capturer le sens des CUI. On peut ainsi espérer que les notions encodées par les plongements par kanjis et les plongements contextuels de l'état-de-l'art soient complémentaires. Pour tester cette hypothèse, nous proposons de fusionner le plongement CUI2vec (Beam *et al.*, 2020) et nos plongements denses. Cela est fait simplement par concaténation des vecteurs ; le plongement obtenu est donc dans $\mathbb{R}^{1000}$.

Toujours dans un soucis de dimensionnalité moindre, nous proposons aussi de tester une deuxième version de plus petite dimensionnalité. Le plongement précédent, obtenu par concaténation, est donc ramené par ACP d'une dimension 1 000 à une dimension de 500. Dans les expériences rapportées dans la section suivante, cette approche est appelée approche par fusion.

4 Expérimentations

Dans cette section, nous évaluons les performances des plongements proposés dans la section précédente. À des fins de comparaison avec les plongements CUI2vec (Beam *et al.*, 2020), largement employés, nous choisissons de tester des plongements obtenus par ACP avec la même dimensionnalité (500).

4.1 Jeux d'évaluation

Nous reprenons plusieurs jeux d'évaluation proposés dans la littérature. Nous les présentons brièvement ci-dessous, le lecteur intéressé peut se reporter à Beam *et al.* (2020) pour plus de détails.

Corrélation avec des jugements humains. Pakhomov *et al.* (2010) ont développé un jeu de données dans lequel des médecins ont indiqué leur perception de proximité entre 566 paires de concepts UMLS. Chaque paire de concepts a ainsi une mesure moyenne de la façon dont ils sont jugés similaires (*similarity*) ou liés (*relatedness*). Nous rapportons la corrélation ρ de Spearman entre ces deux scores d'évaluation humaine (noté ρ *sim.* et ρ *rel.*) et la similarité (cosinus) des plongements.

Types sémantiques. Les types sémantiques sont des méta-informations sur la catégorie à laquelle appartient un concept, et ces catégories sont organisées hiérarchie. Comme proposé par Beam *et al.* (2020), nous avons extrait le type sémantique le plus spécifique disponible pour chaque concept (à partir du fichier MRSTY fourni par UMLS). Nous évaluons la capacité de nos plongements à retrouver, dans les plus proches voisins (au sens du cosinus) d'un concept donné, des concepts partageant le même type sémantique. Cela est évalué avec des mesures de précision sur la liste ordonnée des plus proches voisins des concepts. Par exemple, le concept C0031740 (photochiomiothérapie) vu précédemment est de type *Therapeutic or Preventive Procedure*, tout comme C010009 (lutéolyse), C0043308 (radiothérapie X), etc.

Relations UMLS. Le dernier jeu d'évaluation exploite les relations encodées entre les concepts dans l'UMLS. Ces relations sont générales (e.g. *is-a*) ou spécifiques au domaine médical (e.g. *diagnoses*), la liste complète est disponible à `https://www.nlm.nih.gov/research/umls/META3_`

Plongements	ρ *sim.*	ρ *rel.*	types sémantiques P@5	types sémantiques P@10	relations UMLS P@5	relations UMLS P@10
De Vine *et al.* (2014)	0,455	0,423	0,3940	0,3751	0,1631	0,1275
Choi *et al.* (2016) (claims)	**0,552**	0,384	0,5784	0,5559	0,2444	0,1906
Beam *et al.* (2020)	0,522	0,430	0,5095	0.4781	0,2645	0,2069
kanjis (sec. 3.1)	0,296	0,317	0,6378	0,6117	0,3991	0,3110
ACP (sec. 3.2)	0,228	0,163	0,6213	0,6051	0.3557	0,2814
fusion (sec. 3.3)	0,538	**0,481**	0,6507	0,6138	**0,4265**	**0,3299**
fusion ACP (sec. 3.3)	0,525	0,474	**0,6518**	**0,6158**	0,4180	0.3238

TABLE 1 – Résultats des plongements proposés et de ceux de l'état-de-l'art sur les jeux d'évaluation. Les meilleurs résultats sont indiqués en gras.

`current_relations.html`. Nous évaluons donc la capacités de nos plongements à retrouver, dans les plus proches voisins (au sens du cosinus) d'un concept donné, des concepts liés par n'importe quelle relation. Comme précédemment, cela est mesuré par des précisions a différents seuils sur la liste ordonnée des voisins des concepts. Par exemple, le concept C0031740 (Photochimiothérapie) est en relation *is-a* avec C0087111 (Traitement), et en ralation *has-clinical-form* avec C0034172 (Photothérapie UVA), etc.

4.2 Résultats

Le tableau 1 recense les résultats des différents plongements de la littérature et ceux proposés en section 3. Nous rapportons également les résultats des plongements disponibles de l'état-de-l'art, à savoir les plongements de Choi *et al.* (2016) appris sur les données d'assurance (*claims*), de De Vine *et al.* (2014) et CUI2vec (Beam *et al.*, 2020).

Approches par kanjis. De ces résultats, on peut noter que les approches fondées sur les kanjis se comportent différemment des plongements de l'état-de-l'art. Sur les tâches de comparaison avec l'impression de proximité donnée par des médecins (deux premières colonnes), les approches par kanjis sont largement moins performantes que les autres. Il semble que pour cette tâche, les indices contextuels soient très importants. En revanche, sur les deux autres tâches, fondées sur des proximités encodées dans l'UMLS (types sémantiques ou relations sémantiques), les approches par kanjis apportent des gains, et plus particulièrement sur les relations UMLS. Concernant la différence entre la représentation creuse et celle dense obtenue par ACP, on observe une baisse très légère sur l'ensemble des tâche, du fait de la perte d'information dans le processus de réduction de dimension.

Intérêt de la fusion. Les plongements construits classiquement sur l'hypothèse distributionnelle et ceux construits sur la représentation par kanjis semblent avoir des performances complémentaires. Leur fusion, par concaténation suivie ou non d'une ACP, tire le meilleur parti de chacun et permettent donc d'obtenir de bons résultats sur l'ensemble des jeux d'évaluation. Comme précédemment, la réduction de dimension dégrade très peu les résultats.

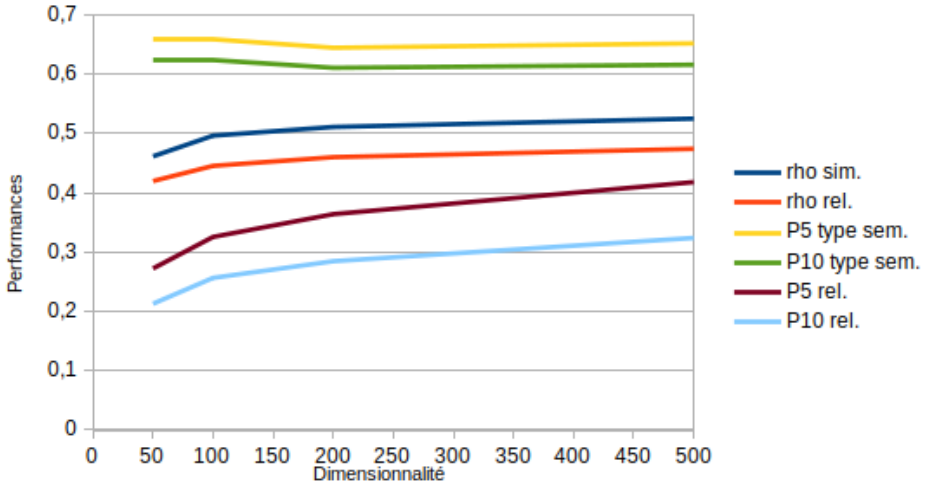


FIGURE 1 – Performances de la fusion de plongements selon la dimensionnalité obtenue par ACP.

Influence de la dimensionnalité. La figure 1 présente l'évolution des performances du plongement obtenu par ACP de la concaténation de cui2vec avec notre représentation ACP/kanjis. Comme attendu, on observe une diminution des performances lorsque le nombre de dimensions retenues à l'ACP finale diminue. Cette baisse de performances reste relativement faible, sauf pour l'évaluation via l'ensemble des relations UMLS. La variété des relations prises en compte dans cette évaluation semble difficile à concilier avec de plus faibles nombres de dimensions.

5 Conclusion et travaux futurs

L'approche que nous proposons est conceptuellement très simple, n'exploite que des données facilement accessibles (issues de l'UMLS) et ne nécessite que peu de temps de calcul. Elle offre pourtant de bons résultats par rapport aux approches existantes qui reposaient sur d'énormes quantités de textes médicaux. Les indices exploités (kanjis vs. contextes distributionnels) étant de nature différente, les performances sont variables selon les jeux d'évaluation. Sur la base de ce constat, nous avons aussi montré l'intérêt de les combiner au sein d'un plongement.

Les plongements existants sont dépendants de l'apparition des termes UMLS dans les corpus sur lesquels ils sont appris. Bien que jouissant d'une bonne couverture, comparable aux plongements de l'état-de-l'art, notre approche est quant à elle complètement dépendante de la disponibilité des termes japonais dans l'UMLS. L'anglais étant plus largement présent dans l'UMLS, des systèmes de traductions peuvent permettre de générer une ou plusieurs traductions candidates en kanjis et ainsi générer les plongements correspondants. Pour résoudre ce problème, nous sommes en train de développer une approche neuronale de génération de représentation kanjis. Les performances des plongements obtenus tendant à être sensiblement inférieurs à ceux obtenus directement des termes en kanjis, mais cela permet d'accroître considérablement la couverture à l'ensemble des CUI de l'UMLS.

D'autres indices, comme la structure même de l'UMLS (graphe de liens typés entre les concepts) peuvent également être utilisés en plus des liens de traductions. Ils peuvent notamment efficacement exploités par des approches de plongements de graphes ou de bases de connaissances (Wang *et al.*, 2018, *inter alia*). Mais cela pose alors la question de l'évaluation, puisque certains jeux d'évaluation

reposent actuellement sur ce graphe.

Récemment, le développement de plongements dynamiques tels que Bert (Devlin *et al.*, 2019) et leurs pendants dans le domaine biomédical (BioBert (Lee *et al.*, 2019) ou ClinicalBert (Alsentzer *et al.*, 2019)...) ont ouvert de nouvelles voies de recherche sur les plongements. Ils reposent cependant sur la même hypothèse distributionnelle et le même besoin de grandes quantités de textes. Par ailleurs, les jeux d'évaluation actuels ne sont pas très adaptés à leur fonctionnement (puisque l'évaluation des plongements est fait hors contexte dans les jeux de données utilisés dans cet article). Cependant, même pour ces approches, l'adjonction d'indices autres tels que nos représentations par kanjis pourraient potentiellement améliorer les représentations - uniquement contextuelles - apprises par ces approches.

Enfin, les différents plongements testés (incluant ceux de l'état de l'art), le code pour les générer (requiert un accès à l'UMLS), et le code pour les évaluer (requiert un accès à l'UMLS) sont disponibles pour la recherche sur demande auprès de l'auteur.

Références

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JIN D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

ARONSON A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program. In *Actes de AMIA 2001*, p. 17–21.

BEAM A. L., KOMPA B., SCHMALTZ A., FRIED I., WEBER G., PALMER N. P., SHI X., CAI T. & KOHANE I. S. (2020). Clinical concept embeddings learned from massive sources of multimodal medical data. In *Proceedings of the Pacific Symposium on BioComputing*, p. 295–306, Hawaï, USA.

CHOI Y., CHIU C. Y.-I. & SONTAG D. A. (2016). Learning low-dimensional representations of medical concepts. In *Clinical Research Informatics* : AMIA.

CLAVEAU V. & KIJAK E. (2013). Analyse morphologique non supervisée en domaine biomédical. Application à la recherche d'information. *Traitement Automatique des Langues*, **54**(1), 13–45. HAL : hal-00912301.

DE VINE L., ZUCCON G., KOOPMAN B., SITBON L. & BRUZA P. (2014). Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, p. 1819–1822, New York, NY, USA : ACM. DOI : 10.1145/2661829.2661974.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**. DOI : 10.1093/bioinformatics/btz682.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU,

Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, p. 3111–3119.

PAKHOMOV S., MCINNES B., ADAM T., LIU Y., PEDERSEN T. & MELTON G. (2010). Semantic similarity and relatedness between clinical terms : An experimental study. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, **2010**, 572–576.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

ROBERTSON S. E., WALKER S. & HANCOCK-BEAULIEU M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proceedings of the 7th Text Retrieval Conference, TREC-7*, p. 199–210.

SARVENIAZI A. (2014). An actual survey of dimensionality reduction. *American Journal of Computational Mathematics*, **04**, 55–72. DOI : 10.4236/ajcm.2014.42006.

TUTTLE M., SHERERTZ D., OLSON N., ERLBAUM M., SPERZEL D., FULLER L. & NESLON S. (1990). Using meta-1 – the 1st version of the UMLS metathesaurus. In *Proc. of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, p. 131–135, Washington, USA.

WANG Z., LV Q., LAN X. & ZHANG Y. (2018). Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 349–357, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1032.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Qu'apporte BERT à l'analyse syntaxique en constituants discontinus ? Une suite de tests pour évaluer les prédictions de structures syntaxiques discontinues en anglais.

Maximin Coavoux
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG,
38000 Grenoble, France
maximin.coavoux@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article propose d'analyser les apports d'un modèle de langue pré-entraîné de type BERT (*bidirectional encoder representations from transformers*) à l'analyse syntaxique en constituants discontinus en anglais (PTB, Penn Treebank). Pour cela, nous réalisons une comparaison des erreurs d'un analyseur syntaxique dans deux configurations (i) avec un accès à BERT affiné lors de l'apprentissage (ii) sans accès à BERT (modèle n'utilisant que les données d'entraînement). Cette comparaison s'appuie sur la construction d'une suite de tests que nous rendons publique. Nous annotons les phrases de la section de validation du Penn Treebank avec des informations sur les phénomènes syntaxiques à l'origine des discontinuités. Ces annotations nous permettent de réaliser une évaluation fine des capacités syntaxiques de l'analyseur pour chaque phénomène cible. Nous montrons que malgré l'apport de BERT à la qualité des analyses (jusqu'à 95 en F_1), certains phénomènes complexes ne sont toujours pas analysés de manière satisfaisante.

ABSTRACT

What does BERT contribute to discontinuous constituency parsing ? A test suite to evaluate discontinuous constituency structure predictions in English.

We propose to analyse the contributions of a pretrained language model such as BERT to discontinuous constituency parsing of English (Penn Treebank). To do so, we perform a comparison of a parsing model in two experimental configuration (i) BERT fine-tuning (ii) without BERT. The comparison relies on the construction of a test suite that we release publicly. We manually annotate the sentences from the development section of the Penn Treebank with information about syntactic phenomena causing discontinuities. We use these annotations to evaluate the syntactic capabilities of a parser for each target phenomenon. Our experiments show that despite the contributions of BERT to very high scores (approaching 95 F_1), certain complex syntactic phenomena are still not identified reliably.

MOTS-CLÉS : Analyse syntaxique en constituants discontinus, analyse d'erreur, discontinuités syntaxiques, suite de tests.

KEYWORDS: Discontinuous constituency parsing, error analysis, syntactic discontinuities, test suite.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 189–196.
HAL : hal-02784767.

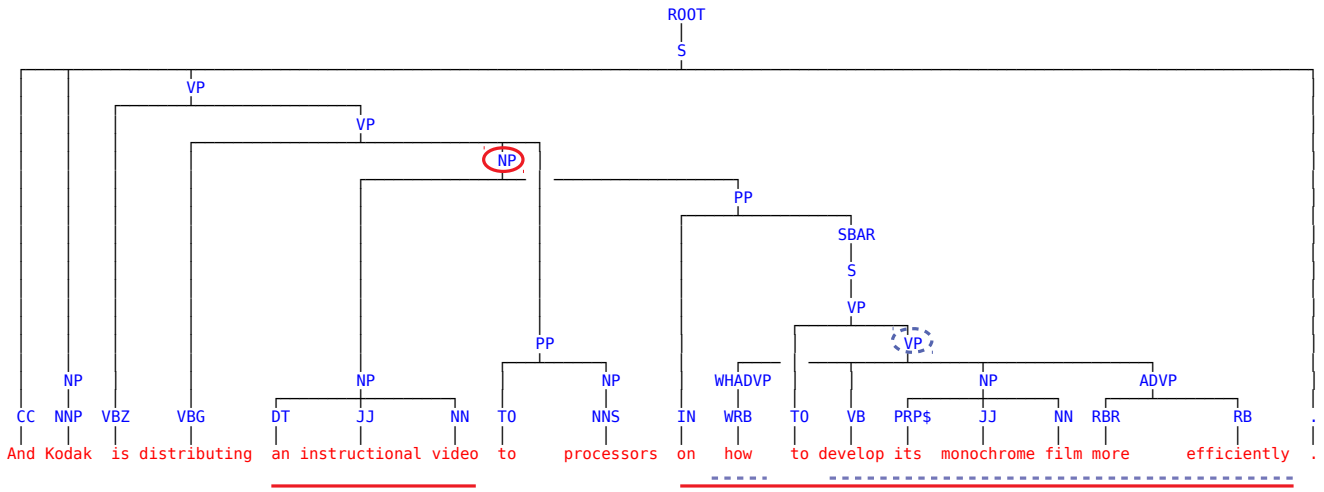


FIGURE 1 – Un arbre syntaxique discontinu issu du Discontinuous Penn Treebank (Evang, 2011). Les discontinuités sont dues à un syntagme prépositionnel (PP) extraposé et à une extraction-*Qu*. Les deux constituants discontinus correspondant ainsi que leurs paires d'empans respectives sont indiquées en rouge / trait plein (NP) et indigo / trait pointillé (VP).

1 Introduction

Dans le cas des langues pour lesquelles on dispose de beaucoup de ressources linguistiques, les analyseurs syntaxiques en constituants qui combinent des systèmes de scorage basés sur les réseaux de neurones et des modèles de langue pré-entraînés tels que BERT (Devlin *et al.*, 2019) obtiennent d'excellents résultats (Kitaev & Klein, 2018; Kitaev *et al.*, 2019; Zhou & Zhao, 2019), atteignant fréquemment plus de 95 F_1 sur le Penn Treebank. Cependant, l'évaluation exclusivement quantitative des outils de traitement automatique des langues (TAL) peut être trompeuse. En effet, un unique score F_1 est difficile à interpréter, et pénalise de la même façon des erreurs sans importance ou des erreurs qui ont un impact sur l'interprétation de la phrase. Ainsi, pour comprendre les capacités des analyseurs actuels, il est nécessaire d'utiliser des méthodes d'évaluation alternatives qui ciblent des phénomènes spécifiques.

Dans cet article, nous construisons une suite de tests pour automatiser l'analyse des erreurs commises sur les phrases présentant des discontinuités syntaxiques du Penn Treebank. Dans un arbre en constituants discontinus (figure 1), il est possible que les tokens dominés par un nœud représentent plusieurs empans (*spans*) non contigus de la phrase. Par exemple, le groupe nominal (NP) discontinu de la figure 1 matérialise l'extraposition en fin de phrase d'un syntagme prépositionel qui est un modifieur du NP. Ce nœud domine « *an instructional video* » et « *on how [...] efficiently* », mais pas le syntagme prépositionnel (PP) « *to processors* ». De telles structures relèvent de grammaires formelles légèrement sensibles au contexte, telles que les LCFRS (Vijay-Shanker *et al.*, 1987, *linear context-free rewriting system*), c'est-à-dire des grammaires d'une classe plus expressive que les grammaires hors-contexte (CFG). Ce type de structure est particulièrement adéquat pour représenter les phénomènes syntaxiques relatifs aux variations d'ordres des mots, et constitue donc des donneés intéressantes pour évaluer les analyseurs syntaxiques.

Une suite de tests similaire à celle que nous proposons existe pour l'allemand (Maier *et al.*, 2014, Discosuite). Les auteur·es de Discosuite ont sélectionné les 1500 premières phrases présentant des discontuités syntaxiques issues du corpus arboré de l'allemand Tiger (Brants *et al.*, 2004), et ont

annoté pour chacune quel phénomène syntaxique est à l'origine de la discontinuité. Ce jeu de tests permet une évaluation fine des analyseurs syntaxiques, sur des phénomènes cibles réputés difficiles à analyser (extrapositions, *scrambling*, ...). À notre connaissance, un tel jeu de tests n'existe pas pour d'autres langues. Certains travaux présentent une analyse d'erreurs sur le Penn Treebank (Evang, 2011; Coavoux *et al.*, 2019), mais d'une part effectuent une analyse manuelle, arbre par arbre, et d'autres part ne rendent pas publiques leurs annotations. À l'inverse, nous rendons publique notre suite de tests ainsi que des scripts d'évaluation qui permettent d'obtenir de manière automatique une évaluation par phénomène d'un analyseur syntaxique.

En effet, nous utilisons la suite de tests pour analyser et comparer les erreurs commises par un analyseur à l'état de l'art dans deux configurations expérimentales distinctes : (i) entraînement utilisant uniquement les données d'entraînement du Penn Treebank (ii) entraînement utilisant (et affinant) les vecteurs de BERT comme source complémentaire d'information. Nous montrons que même si BERT permet d'atteindre des scores approchant 95% de F_1, il ne permet toujours pas d'analyser certains phénomènes cibles de manière satisfaisante.

2 Processus d'annotations

Nous extrayons tous les arbres présentant au moins un constituant discontinu dans le corpus de validation [1] de la version discontinue du Penn Treebank (Evang & Kallmeyer, 2011; Evang, 2011). Nous ignorons les phrases pour lesquelles la discontinuité provient uniquement du rattachement de la ponctuation. Cela correspond en tout à 266 phrases, c'est-à-dire environ 16% des phrases du corpus. Ensuite, pour chaque phrase, nous annotons manuellement quels phénomomènes syntaxiques sont à l'origine des discontinuités, en suivant la classification de Evang (2011) reprise par Coavoux *et al.* (2019) : (i) extractions à longue distance, (ii) citations avec proposition principale en incise, (iii) extraposition à gauche d'une citation (iv) extraposition-*it* (v) autres extrapositions (vi) inversion sujet-verbe. Nous exemplifions ces phénomènes ci-dessous, en matérialisant en gras le constituant discontinu principal :

i *[...] the worst thing* ***that*** *anyone* ***can do*** *[...]* (phrase 53 du corpus de développement) ;

ii ***The stock market has lost some precursory power ,*** *analysts at the Columbia center claim* ***, because of the growing impact of international developments .*** (1190) ;

iii ***Currently , average pay for machinists is $ 13.39 an hour ,*** *Boeing* ***said .*** (238) ;

iv *But* ***it*** *remains to be seen* ***whether their ads will be any more effective .*** (64) ;

v *[...] as* ***the news*** *spread* ***that Wall Street was moving up*** *[...]* (300) ;

vi ***Says*** *James Norman , the mayor of Ava , Mo.* ***: " I do n't invest in stocks .*** (1575).

Cette classification couvre tous les cas présents dans le corpus [2].

Pour certaines phrases, les discontinuités sont dues à plusieurs occurrences de phénomènes distincts dans la même phrase, par exemple la phrase de la figure 1 comporte une extraposition et une extraction. Il s'agit de 21 phrases avec 2 occurrences et d'une seule phrase avec 3 occurrences. Nous rendons disponibles ces annotations [3] en ligne.

1. Nous avons choisi de travailler sur le corpus de validation pour que les phrases annotées soient distinctes du corpus d'entraînement. Cela permettra d'évaluer n'importe quel analyseur syntaxique entraîné sur le split standard de ce corpus.

2. En pratique, il serait possible d'utiliser une granularité encore plus fine, par exemple distinguer parmi les extractions, celles qui sont dues à une question directe et celles qui sont dues à des propositions relatives.

3. https://gitlab.com/mcoavoux/disco-eval-ptb/-/releases/v1.0_taln2020

	Corpus de validation						Corpus de test					
	P	R	F_1	Disc. P	Disc. R	Disc. F_1	P	R	F_1	Disc. P	Disc. R	Disc. F_1
-BERT (Coavoux & Cohen, 2019)	91.5	91.3	91.4	76.1	66.4	70.9	91.3	90.6	90.9	73.3	62.1	67.3
+BERT	94.9	94.9	94.9	80.5	77.6	79.0	95.0	94.5	**94.8**	76.5	70.9	**73.6**
Δ	+3.4	+3.6	+3.5	+4.4	+11.2	+8.1	+3.7	+3.9	+3.9	+3.2	+8.8	+6.3
Corro (2020)									**94.8**	90.8	49.7	64.2

TABLE 1 – Le modèle sans BERT est le modèle pré-entraîné fourni avec le code de l'analyseur et décrit par Coavoux & Cohen (2019).

3 Expériences

Analyseur syntaxique Nous utilisons l'analyseur discontinu décrit par Coavoux & Cohen (2019) et librement disponible [4]. Il s'agit d'un analyseur statistique par transitions, paramétré par un réseau de neurones modulaire :

— un mot-forme est représenté par la concaténation d'un plongement lexical standard ($\mathbf{w}$) et de la sortie ($\mathbf{c}$) d'un bi-LSTM basé sur la séquence de ses caractères ;

— la séquence de vecteurs pour une phrase $([\mathbf{w}_1;\mathbf{c}_1],[\mathbf{w}_2;\mathbf{c}_2],\ldots,[\mathbf{w}_n;\mathbf{c}_n])$ est ensuite donnée en entrée à un bi-LSTM qui calcule des représentations contextualisées de chaque token ;

— enfin, à chaque étape de l'analyse, un réseau à propagation avant (*feedforward*) prédit une action de parsing à partir des vecteurs contextualisés de tokens extraits d'une configuration de l'analyseur (se reporter à Coavoux & Cohen, 2019, pour plus de détails).

Nous avons augmenté cet analyseur d'une option qui permet d'utiliser les vecteurs contextualisés issus de BERT comme représentation de chaque token, concaténés aux représentations lexicales originelles de l'analyseur. En d'autres termes, nous donnons la séquence $([\mathbf{w}_1;\mathbf{c}_1,\mathbf{b}_1],[\mathbf{w}_2;\mathbf{c}_2,\mathbf{b}_2],\ldots,[\mathbf{w}_n;\mathbf{c}_n,\mathbf{b}_n])$ au bi-LSTM qui cacule les représentations contextualisées, où $\mathbf{b}_i$ est la sortie de BERT pour le token i.

Nous affinons les paramètres de BERT lors de l'entraînement. Nous utilisons le modèle de base de BERT qui prend en compte la casse (Devlin *et al.*, 2019, `bert-base-cased`) via l'interface de la bibliothèque `transformers` [5] (Wolf *et al.*, 2019).

Évaluation générale Nous présentons les résultats des deux modèles dans la table 1. Nous utilisons l'évaluateur `discodop` [6] (van Cranenburgh *et al.*, 2016) avec ses paramètres standards. Ce paramétrage ignore la ponctuation et les racines des arbres. Il neutralise également la distinction entre ADVP (syntagmes adverbiaux) et PRT (particule). Nous présentons les précisions (P), rappels (R) et F mesures (F_1) calculés soit sur l'ensemble des constituants, soit uniquement sur les constituants discontinus (Disc. P/R/F_1). Par exemple, pour l'arbre de la figure 1, ces dernières mesures ne considèreraient que les deux constituants discontinus. Toutes ces mesures prennent en compte les étiquettes des constituants, c'est-à-dire que pour qu'un constituant prédit soit considéré juste, il faut qu'il ait la bonne étiquette et qu'il couvre le(s) bon(s) empan(s).

Le modèle entraîné avec BERT (+BERT) obtient une mesure F_1 de 94.8 sur le corpus de test. Ce résultat est identique à celui publié récemment par Corro (2020), qui est le seul autre résultat utilisant

4. https://gitlab.com/mcoavoux/discoparset
5. https://github.com/huggingface/transformers
6. https://github.com/andreasvc/disco-dop

Phénomène	Effectif	Reconnus		Partiellement reconnus		Précision		Rappel		F_1	
		-BERT	+BERT	-BERT	+BERT	-BERT	+BERT	-BERT	+BERT	-BERT	+BERT
Évaluation avec étiquettes											
Extraction	93	65.6	83.9	81.7	91.4	81.6	87.3	77.2	86.2	79.4	86.7
Citation extraposée	73	89.0	93.2	91.8	94.5	94.4	93.4	90.7	94.7	92.5	94.0
Autre extraposition	39	23.1	59.0	25.6	64.1	90.9	93.1	20.0	54.0	32.8	68.4
Citation avec incise	16	0.0	0.0	93.8	100.0	48.9	53.7	46.0	58.0	47.4	55.8
Extraposition-*it*	12	41.7	75.0	50.0	83.3	85.7	83.3	50.0	83.3	63.2	83.3
Inversion sujet-verbe	6	50.0	66.7	66.7	83.3	100.0	71.4	50.0	62.5	66.7	66.7
Extraction et citation extraposée	6	83.3	100.0	100.0	100.0	100.0	100.0	94.1	100.0	97.0	100.0
Évaluation sans étiquettes											
Extraction	93	65.6	84.9	81.7	91.4	81.3	89	78.7	89	80	89
Citation extraposée	73	89	93.2	91.8	94.5	94.4	93.4	90.7	94.7	92.5	94
Autre extraposition	39	23.1	61.5	25.6	66.7	90.9	96.4	20.4	55.1	33.3	70.1
Citation avec incise	16	81.2	87.5	93.8	100	94.4	97.4	89.5	97.4	91.9	97.4
Extraposition-*it*	12	41.7	75	50	83.3	85.7	83.3	50	83.3	63.2	83.3
Inversion sujet-verbe	6	50	66.7	66.7	83.3	100	85.7	50	75	66.7	80
Extraction et citation extraposée	6	83.3	100	100	100	100	100	93.8	100	96.8	100

TABLE 2 – Évaluation par phénomène pour les deux modèles d'analyse (avec ou sans BERT), sur le corpus de validation. La précision, le rappel et le score F_1 sont calculés uniquement sur les constituants discontinus et correspondent aux mesures Disc. P/R/F_1 de la table 1, décomposées par phénomène.

BERT sur ce jeu de données. Le modèle +BERT obtient une amélioration de 4 points par rapport au modèle sans BERT (-BERT). L'apport de BERT est particulièrement fort sur les constituants discontinus (+8.1 en Disc. F_1 sur le corpus de développement). En particulier, nous observons un très fort effet sur le rappel : le modèle avec BERT est bien plus compétent pour détecter les phénomènes syntaxiques qui produisent des discontinuités. Cet effet est d'autant plus important que les travaux en analyse syntaxique discontinue obtiennent constamment une précision bien plus elevée que le rappel (Maier, 2015; Stanojević & G. Alhama, 2017; Coavoux *et al.*, 2019), ce qui est lié à la rareté des constituants discontinus dans les données d'entraînement. Malgré ces améliorations, le modèle +BERT obtient une mesure F_1 de seulement 73.6 sur le corpus de test, signe que les phénomènes syntaxiques à l'origine des discontinuités ne sont toujours pas identifiés de manière satisfaisante. Pour cette raison, nous proposons une analyse plus fine des résultats de ces analyseurs à l'aide de la suite de tests que nous avons construite.

Processus d'évaluation par phénomène Nous procédons comme suit pour évaluer automatiquement les prédictions des analyseurs sur le corpus de développement. À l'aide de l'évaluateur `discodoop`, nous récupérons une évaluation individuelle pour chaque phrase contenant une discontinuité. En particulier, nous récupérons son score Disc. F_1. Nous considérons que le phénomène annoté pour une phrase donnée a été (i) reconnu si la phrase obtient 100% en Disc. F_1 (ii) partiellement reconnu si son Disc. F_1 est strictement supérieur à 0 (c'est-à-dire si au moins un constituant discontinu a été bien prédit). Dans la table 2 nous rapportons ces résultats par phénomène pour chaque modèle, ainsi que la précision, le rappel et le F_1 (micro-moyenne sur l'ensemble des constituants discontinus) pour ce phénomène. Nous rapportons ces valeurs dans deux cas : le cas standard où on prend en compte les étiquettes des constituants (partie supérieure de la table), et le cas non étiqueté, où il suffit de prédire les bons empans pour qu'on considère que la prédiction d'un constituant est correcte (partie inférieure de la table). Cela nous permet d'isoler les cas où le système fait simplement des erreurs d'étiquetage des constituants.

Lorsqu'il y a plusieurs occurrences de phénomènes cibles pour une phrase, nous ne pouvons pas

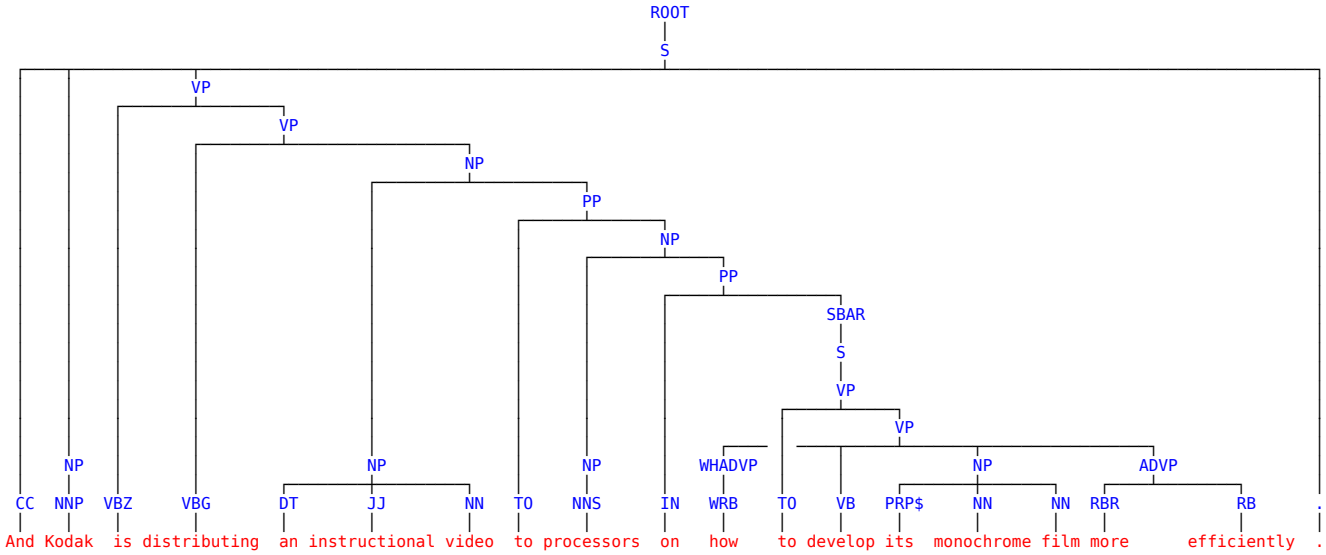


FIGURE 2 – Arbre prédit par le modèle -BERT pour la phrase de la figure 1, présentant des erreurs de rattachement prépositionnel. Le modèle +BERT ne fait aucune erreur sur cette phrase.

conclure de manière automatique lequel des phénomènes a été reconnu ou non (hormis quand le F_1 est à 0 ou 100). Par ailleurs, comme il y a très peu d'occurrences de phrases ayant une combinaison spécifique de 2 phénomènes cibles, nous rapportons des statistiques seulement pour la combinaison la plus fréquente (c'est-à-dire les phrases qui contiennent à la fois une extraction à longue distance et une citation extraposée), et nous ignorons les autres (moins de 6 occurrences chacune).

Résultats par phénomène Nous rapportons l'ensemble des résultats dans la table 2. Les phénomènes produisant des discontinuités les mieux identifiés sont les extractions, les citations extraposées, ainsi que, dans le cas +BERT, les extrapositions-*it*. Notons qu'il s'agit de phénomènes pour lesquels l'analyseur peut s'appuyer sur des indices lexicaux forts : les mots-*Qu* (tels que *what*, *that*, *when*, ...) pour les extractions, les verbes de discours (*say*) pour les citations extraposées, et la présence de *it*. Ce sont également des phénomènes plutôt fréquents.

En revanche, les citations avec incise sont plutôt mal prédites (0% reconnues par les deux modèles). Leur structure est plutôt bien identifiée ($> 80\%$ de reconnaissance non étiquetée), mais les analyseurs font systèmatiquement au moins une erreur sur les étiquettes des constituants discontinus. Ces erreurs d'étiquetage sont sans doute liées à un manque de traits pour classifier les constituants discontinus (Coavoux & Cohen, 2019), ce que nous prévoyons de tester.

Enfin, les extrapositions et les inversions sujet-verbe sont plutôt mal identifiées ($\approx 50\%$ de reconnaissance pour le modèle -BERT). Les extrapositions sont particulièrement difficiles à prédire dans la mesure où elles nécessitent souvent des connaissances syntaxiques et sémantiques fines pour être identifiées. Par exemple, pour désambiguïser un des deux rattachements de syntagmes prépositionnels dans l'arbre de la figure 1, il faut savoir que le sujet de la video (*on how to [...]*) est un modifieur possible pour *video* mais pas pour *processors*. Le modèle +BERT ne fait aucune erreur sur cette phrase, alors que le modèle -BERT attachent les deux prépositions localement (figure 2) [7].

7. Reviewer 2 nous fait également remarquer que *distributing* et *video* ont plusieurs occurrences dans le corpus d'entraînement mais jamais dans des structures syntaxiques similaires à celles que l'on trouve dans cet exemple, suggérant que BERT apporte ici avant tout des informations sur la valence et les cadres de sous-catégorisation de ces lexèmes.

De manière générale, l'entraînement avec BERT améliore tous les résultats en terme de reconnaissance et de score F_1. En particulier, son effet sur la reconnaissance des extrapositions (de 23.1 à 59%) et sur les extrapositions-*it* (41.7 à 75%) est remarquable.

4 Conclusions

Cet article présente une suite de tests permettant d'évaluer automatiquement les prédictions d'un analyseur en constituants discontinus sur un ensemble de phénomènes cibles réputés difficiles à analyser. Nous rendons publique cette suite de tests. Enfin, nous présentons une analyse des erreurs d'un analyseur à l'état de l'art dans deux configurations expérimentales, selon qu'il est entraîné par affinage des paramètres de BERT ou sans BERT. L'analyse met en lumière un résultat prometteur : BERT permet d'obtenir le meilleur résultat publié à ce jour en analyse syntaxique discontinue sur le Penn Treebank, et améliore significativement la prédiction des discontuités syntaxiques présentes dans l'anglais journalistique de l'époque du PTB. Cependant, elle montre également qu'il reste une marge d'amélioration pour ces phénomènes. À l'avenir, nous prévoyons d'expérimenter avec d'autres méthodes d'apprentissage semi-supervisées pour traiter cette limitation des analyseurs actuels.

Remerciements

Je remercie Jibril Frej, Caio Corro, ainsi que 3 relecteurices anonymes pour leurs remarques et suggestions sur cet article.

Références

BRANTS S., DIPPER S., EISENBERG P., HANSEN-SCHIRRA S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKOREIT H. (2004). Tiger : Linguistic interpretation of a german corpus. *Research on language and computation*, **2**(4), 597–620.

COAVOUX M. & COHEN S. B. (2019). Discontinuous constituency parsing with a stack-free transition system and a dynamic oracle. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 204–217, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1018.

COAVOUX M., CRABBÉ B. & COHEN S. B. (2019). Unlexicalized transition-based discontinuous constituency parsing. *Transactions of the Association for Computational Linguistics*, **7**, 73–89. DOI : 10.1162/tacl_a_00255.

CORRO C. (2020). Span-based discontinuous constituency parsing : a family of exact chart-based algorithms with time complexities from $\mathcal{O}(n^6)$ down to $\mathcal{O}(n^3)$. arXiv preprint : 2003.13785.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EVANG K. (2011). Parsing discontinuous constituents in English. Mémoire de master, University of Tübingen.

EVANG K. & KALLMEYER L. (2011). PLCFRS parsing of English discontinuous constituents. In *Proceedings of the 12th International Conference on Parsing Technologies*, p. 104–116, Dublin, Ireland : Association for Computational Linguistics.

KITAEV N., CAO S. & KLEIN D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3499–3505, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1340.

KITAEV N. & KLEIN D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2676–2686, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1249.

MAIER W. (2015). Discontinuous incremental shift-reduce parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1202–1212, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1116.

MAIER W., KAESHAMMER M., BAUMANN P. & KÜBLER S. (2014). Discosuite - a parser test suite for German discontinuous structures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2905–2912, Reykjavik, Iceland : European Language Resources Association (ELRA).

STANOJEVIĆ M. & G. ALHAMA R. (2017). Neural discontinuous constituency parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1666–1676, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1174.

VAN CRANENBURGH A., SCHA R. & BOD R. (2016). Data-oriented parsing with discontinuous constituents and function tags. *Journal of Language Modelling*, **4**(1), 57–111.

VIJAY-SHANKER K., WEIR D. J. & JOSHI A. K. (1987). Characterizing structural descriptions produced by various grammatical formalisms. In *25th Annual Meeting of the Association for Computational Linguistics*, p. 104–111, Stanford, California, USA : Association for Computational Linguistics. DOI : 10.3115/981175.981190.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M. & BREW J. (2019). Huggingface's transformers : State-of-the-art natural language processing. arXiv preprint : 1910.03771.

ZHOU J. & ZHAO H. (2019). Head-driven phrase structure grammar parsing on Penn treebank. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2396–2408, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1230.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Sur l'impact des contraintes structurelles pour l'analyse en dépendances profondes fondée sur les graphes

Caio Corro
Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
caio.corro@limsi.fr

RÉSUMÉ

Les algorithmes existants pour l'analyse en dépendances profondes fondée sur les graphes capables de garantir la connexité des structures produites ne couvrent pas les corpus du français. Nous proposons un nouvel algorithme qui couvre l'ensemble des structures possibles. Nous nous évaluons sur les corpus français FTB et Sequoia et observons un compromis entre la production de structures valides et la qualité des analyses.

ABSTRACT

On the impact of structural constraints for graph-based deep dependency parsing

Existing approaches for graph-based deep dependency parsing that force connexity in predicted graphs do no cover the structures observed in French treebanks. We propose a novel algorithm that covers the full set of possible structures. We evaluate our approach on the French corpora FTB and Sequoia and observe a trade-off between the validity of predicted structures and the quality of predictions.

MOTS-CLÉS : analyse syntaxique, analyse en dépendances profondes, optimisation combinatoire.

KEYWORDS: syntactic analysis, deep dependency parsing, combinatorial optimization.

1 Introduction

La structure en dépendances *surfaciques* d'une phrase est usuellement représentée sous forme d'un graphe dirigé où chaque mot est représenté par un nœud et chaque dépendance bi-lexicale par un arc (figure 1). Un nœud supplémentaire est introduit pour connecter le mot racine de la phrase. Le graphe ainsi construit est une arborescence couvrante : chaque nœud sauf la racine possède exactement un arc entrant et il est existe un chemin à partir de la racine jusque chaque nœud. Il en découle qu'une arborescence ne peut pas contenir de cycle. L'analyse en dépendances fondée sur les graphes d'une phrase comporte deux étapes (McDonald *et al.*, 2005) :

1. la construction d'un graphe dirigé et pondéré complet où le poids associé à chaque arc correspond à la vraisemblance d'une relation bi-lexicale calculée par un modèle statistique ;
2. la calcul de l'arborescence couvrante de poids maximal de ce graphe, c'est-à-dire la sélection d'un sous-graphe contraint de poids maximal.

Le sous-graphe calculé peut alors être directement transposé à la structure syntaxique correspondante. Le problème à résoudre lors de la seconde étape a une complexité quadratique par rapport au nombre de mots dans la phrase d'entrée et peut être résolu en utilisant la variante de Fredman & Tarjan (1987)

Volume 2 : Traitement Automatique des Langues Naturelles, pages 197–204.
HAL : hal-02784768.

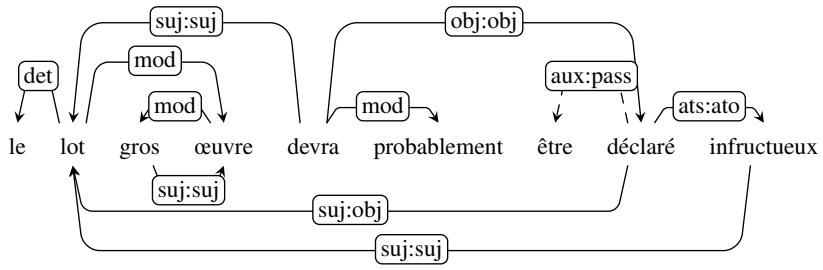


FIGURE 1 – Exemple d'analyse en dépendances surfaciques (arcs au-dessus de la phrase) et en dépendances profondes (combinaison des arcs pleins du dessus et des arcs en dessous) repris de Candito *et al.* (2014). Le mot "être" n'est pas présent dans l'analyse en dépendances profondes et il existe un cycle contenant "gros" et "œuvre".

Complexité	FTB	Sequoia
$O(n^3)$	57.88%	65.34%
$O(n^4)$	83.63%	86.99%
$O(n^5)$	84.34%	87.45%
NP-diff.	100%	100%

TABLE 1 – Couverture en terme de graphes complets du FTB et de Sequoia en fonction de la complexité temporelle de l'analyseur. Le parser en $\mathcal{O}(n^3)$ est celui proposé par Kuhlmann & Jonsson (2015), ceux en $O(n^4)$ et $O(n^5)$ sont ceux proposés par Cao *et al.* (2017), le NP-difficile celui proposé dans cet article.

de l'algorithme de Chu-Liu-Edmonds (Edmonds, 1967). D'autres contraintes sur le sous-graphe peuvent être ajoutées comme la projectivité (Eisner & Satta, 1999) ou le degré de bloc limité et la bonne imbrication (Gómez-Rodríguez *et al.*, 2009; Corro *et al.*, 2016) au prix d'une complexité temporelle plus élevée, voire d'une complexité rédhibitoire.

Nous nous intéressons ici à l'analyse en dépendances *profondes* fondée sur les graphes (figure 1). À la différence des dépendances surfaciques, une structure en dépendances profondes s'abstrait des variations syntaxiques pour représenter plus explicitement les relations prédicats-arguments, entre autres. Cette représentation ne contient que les mots sémantiquement pleins. Nous renvoyons le lecteur au schéma d'annotation du corpus Sequoia (Perrier *et al.*, 2014) pour de plus amples informations. Notons qu'une analyse en dépendances profondes n'est pas une extension des dépendances surfaciques : par exemple sur la figure 1, la dépendance entre « être » et « déclaré » n'est pas présente. Pour l'analyse fondée sur les graphes, le calcul du sous-graphe de poids maximal doit donc prendre en compte les spécificités de cette représentation :

1. le sous-graphe n'est pas contraint d'être couvrant car les mots sémantiquement vides n'apparaissent pas dans la structure syntaxique ;
2. un nœud peut avoir 0 ou plusieurs arcs entrants ;
3. le sous-graphe peut contenir des cycles.

En d'autres termes, la seule propriété qui doit être contrainte sur le sous-graphe pour qu'il soit valide est sa faible connexité, c'est à dire qu'il doit exister un chemin à partir de la racine jusque tous les autres mots nœud du graphe sans prendre en considération la direction des arcs. Malheureusement, ce problème est NP-difficile (Ideker *et al.*, 2002). Bien qu'il existe des algorithmes tractables permettant de calculer des sous-graphes faiblement connexes plus contraints (Kuhlmann & Jonsson, 2015; Cao *et al.*, 2017), ceux-ci ne permettent pas de couvrir l'ensemble des structures présentes dans les jeux de données (voir table 1).

Dans cet article, nous proposons un algorithme qui garantit la connexité des analyses produites tout en couvrant l'ensemble des corpus et nous examinons empiriquement si cette contrainte est importante pour l'analyse en dépendances profondes fondée sur les graphes. En effet, dans le cadre de l'analyse en dépendances surfaciques, Zhang *et al.* (2017) ont montré qu'il était possible de supprimer les contraintes d'arborescence lors de la prédiction, c'est-à-dire de choisir localement un arc entrant pour chaque nœud sans garantir explicitement la contrainte globale d'arborescence. De plus, empiriquement la plupart des graphes produits par leur analyseur non contraint sont des

arborescences. Dans le cas de l'analyse en dépendances profondes, la suppression de la contrainte de connexité transforme le calcul du sous-graphe de poids maximal en la simple sélection de tous les arcs de poids positifs. Cette approche non contrainte est celle choisie par les analyseurs à l'état de l'art pour les dépendances profondes du français. En effet, si l'analyseur de Ribeyre *et al.* (2016) ne se limite pas à une simple sélection des arcs de poids positifs, c'est seulement dû à l'utilisation d'un modèle d'ordre supérieur [1]. Cependant, la connexité du graphe produit n'est pas garantie.

Nos contributions peuvent être résumées comme suit :

1. Nous proposons un nouvel algorithme pour l'analyse en syntaxe profonde qui permet de couvrir l'ensemble des structures des jeux de données tout en garantissant la connexité des graphes produits ;
2. Nous comparons expérimentalement les résultats produits par notre analyseur contraint et la simple sélection des arcs de poids positifs.

L'implémentation pour reproduire les expériences de cet article est disponible en ligne [2].

2 Analyse en dépendances profondes fondée sur les graphes

Nous proposons de réduire le problème de l'analyse syntaxique profonde au problème du sous-graphe connexe de poids maximal (SGCPM), un problème de graphe NP-difficile ayant des applications en chimie et en biologie moléculaire (Dittrich *et al.*, 2008; Ideker *et al.*, 2002; Loboda *et al.*, 2016). À la différence de l'approche proposée par Ribeyre *et al.* (2016), nous contraignons la bonne formation de la structure syntaxique. À la différence des approches proposées par Kuhlmann & Jonsson (2015) et Cao *et al.* (2017), nous pouvons couvrir l'ensemble des jeux de données (voir table 1).

2.1 Réduction à un problème de graphe non dirigé

Soit $\boldsymbol{s} = s_0 ... s_n$ une phrase de n mots où s_0 est un faux mot supplémentaire. Nous construisons un graphe dirigé complet pondéré $G = \langle V, A, \boldsymbol{w} \rangle$ avec $V = \{0...n\}$ l'ensemble des nœuds, $A = V \times V$ l'ensemble des arcs et $\boldsymbol{w} \in \mathbb{R}^{|A|}$ un vecteur de poids indexé par les arcs. Le problème du SGCPM sur un graphe dirigé peut se réduire à sa variante non dirigée en construisant un graphe non dirigé pondéré $G' = \langle V, E, \boldsymbol{w}' \rangle$ avec $E \subset V \times V$ l'ensemble des arêtes et $\boldsymbol{w}' \in \mathbb{R}^{|E|}$ de la façon suivante : pour chaque couple de nœuds $u, v \in V$ tels que $u < v$, nous ajoutons une arêtes (u, v) au graphe G' de poids : $w'_{uv} = max(w_{uv}, w_{vu}, w_{uv} + w_{vu})$. Une solution du SGCPM sur G peut alors être reconstruite à partir d'une solution du SGCPM sur G' en tenant compte des arcs de G qui ont contribué au poids des arêtes de G'.

2.2 Programme mathématique

Nous proposons ici une formulation sous forme de programme mathématique du problème du SGCPM, fondée sur les formulations de Haouari *et al.* (2013) et Loboda *et al.* (2016). Le programme

1. Dans un modèle du première ordre, le poids d'un graphe est la somme des poids des arcs qu'il contient. Les modèles d'ordre supérieur incluent aussi des poids pour les couples d'arcs.
2. https://github.com/FilippoC/deep-syntactic-dependency-parsing-release

que nous proposons est (légèrement) plus simple car il existe un nœud obligatoirement présent dans le sous-graphe : la racine. Le sous-graphe recherché étant connexe, tous les nœuds doivent être accessibles en suivant un chemin à partir de celle-ci : nous devons garantir l'existence d'au moins une traversée possible à partir de la racine. Pour cela, nous représentons l'ensemble des chemins de cette traversée par une arborescence où un arc indique que le nœud de départ est visité avant le nœud d'arrivée dans la traversée [3]. Pour forcer la structure d'arborescence, nous utilisons une variable entière qui représente la distance de la racine à chaque nœud du sous-graphe plutôt qu'une modélisation par flot.

Nous introduisons 3 ensembles de variables binaires. Les variables y_e, $e \in E$, et x_v, $v \in V \setminus \{0\}$, représentent la sélection (valeur 1) ou non (valeur 0) des différentes arêtes et nœuds, respectivement. Les variables z_a, $a \in A$, représentent la sélection des arcs de la traversée utilisée pour garantir la connexité du sous-graphe produit. De plus, les variables entières $d_v, v \in V$, sont utilisées pour décrire les solutions faisables de l'arborescence. Nous notons $\delta(v)$ l'ensemble des arêtes incidentes au nœud v. Le programme mathématique pour résoudre le problème du SGCPM sur G' est défini de la façon suivante :

$$\max \quad \sum_{e \in E} w'_e y_e, \tag{1}$$

$$\text{t.q.} \quad y_e \leq x_v, \qquad \forall v \in V, e \in \delta(v), \tag{2}$$

$$z_{uv} + z_{vu} \leq y_{uv}, \qquad \forall (u,v) \in E, \tag{3}$$

$$\sum_{(u,v) \in A} z_{uv} = x_v, \qquad \forall v \in V \setminus \{0\}, \tag{4}$$

$$d_0 = 1, \tag{5}$$

$$d_v z_{uv} = (d_u + 1) z_{uv}, \qquad \forall (u,v) \in A, \tag{6}$$

$$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \{0,1\}, \boldsymbol{d} \in [1...n+1]. \tag{7}$$

La contrainte (2) garantit qu'une arête ne peut être sélectionnée que si ses deux nœuds incidents sont sélectionnés. La contrainte (3) s'assure que les arcs qui forment l'arborescence ne peuvent être présents que si l'arête équivalente dans le sous-graphe non-dirigé est présente. Ensuite, les contraintes (4)-(6) assurent la bonne formation de l'arborescence : chaque nœud doit avoir exactement un arc entrant s'il est sélectionné (4) ; la distance de la racine à elle même est égale à un (5) ; si un arc est présent dans l'arborescence, alors la distance de la racine de son nœud de destination est égale à la distance de la racine du nœud de départ plus un (6) [4]. Pour une démonstration de l'exactitude de cette formulation, nous reportons le lecteur à Haouari *et al.* (2013). Ce programme mathématique n'est pas un programme linéaire en nombre entiers à cause de la contrainte (6). Cependant, celle-ci peut être linéarisée en suivant la procédure décrite par Loboda *et al.* (2016). En pratique, nous pouvons donc résoudre ce programme mathématique avec l'outil CPLEX [5].

3. Précisons bien que les arcs de cette arborescence ne représentent pas forcément des dépendances syntaxiques. Les dépendances sélectionnées sont représentées par les arêtes (non-dirigées) du sous-graphe connexe. De plus ils ne décrivent qu'une traversée possible, il peut en exister d'autres.

4. Cette contrainte interdit les cycles dans l'arborescence.

5. https://www.ibm.com/fr-fr/analytics/cplex-optimizer

	% de struct. con.	LP	LR	LF	Exacte
	FULLYSUP				
POSITIVEARCS	64.84 / 66.83	85.90 / 86.15	82.70 / 83.00	**84.27 / 84.55**	17.95 / 18.83
→ FTB	63.16 / 65.39	85.63 / 85.89	82.43 / 82.75	84.00 / 84.29	15.39 / 16.26
→ Sequoia	75.50	88.21	85.02	86.58	34.25
SGCPM	100 / 100	84.70 / 85.02	83.48 / 83.78	84.09 / 84.40	**18.40 / 19.29**
→ FTB	100 / 100	84.43 / 84.77	83.22 / 83.53	83.82 / 84.14	15.82 / 16.72
→ Sequoia	100	87.05	85.77	86.41	34.75
	PRETRAINED				
POSITIVEARCS	69.16 / 70.65	89.10 / 89.46	85.38 / 85.77	**87.20 / 87.57**	21.32 / 22.36
→ FTB	67.41 / 69.05	88.83 / 89.20	85.07 / 85.48	86.91 / 87.30	18.93 / 20.01
→ Sequoia	80.25	91.48	88.02	89.72	36.50
SGCPM	100 / 100	88.19 / 88.59	86.04 / 86.41	87.10 / 87.49	**22.34 / 23.43**
→ FTB	100 / 100	87.89 / 88.32	85.74 / 86.14	86.80 / 87.21	19.83 / 20.97
→ Sequoia	100	90.78	88.61	89.68	38.25

TABLE 2 – Résultats de nos deux analyseurs avec les deux réseaux de neurones en terme de pourcentage de structures connexes, de précision étiquetée, rappel étiqueté, f-mesure étiquetée et de correspondance exacte. Le FTB contenant des erreurs de conversion induisant des structures non connexes, nous reportons deux scores : score sur tout le jeu de test / score sur les phrases ayant des structures de référence connexes seulement.

3 Pondération neuronale

Nous utilisons l'architecture neuronale de Dozat & Manning (2017) avec les mêmes hyper-paramètres. Nous expérimentons avec deux variantes qui diffèrent par la façon dont elles construisent les plongements lexicaux.

L'architecture FULLYSUP est entièrement initialisée aléatoirement et est entrainée de bout en bout. Elle est fondée sur des plongements lexicaux et des plongements de caractères.

L'architecture PRETRAINED utilise un *self-attentive encoder* (Vaswani *et al.*, 2017) pré-entrainé pour extraire les plongements sensibles au contexte (Martin *et al.*, 2019). Étant donné que ce modèle pré-entrainé utilise son propre tokenizer de mots, nous utilisons la sortie qui correspond au premier sous-token de chaque mot. De plus, plutôt que d'utiliser la représentation en sortie de la dernière couche d'attention, nous apprenons une combinaison convexe des couches 4 à 7, de la même façon que proposée pour l'architecture Elmo (Peters *et al.*, 2018). Ensuite, nous concaténons cette sortie à un plongement lexical appris de bout en bout, suivi de la même architecture que Dozat & Manning (2017).

Le calcul de la fonction de perte et de sa dérivée est une étape coûteuse en prédiction structurée. Nous décomposons la fonction de perte par partie, c'est-à-dire par arc, d'une façon similaire à ce qui a été proposé par Dozat & Manning (2017) et Zhang *et al.* (2017) mais adapté aux dépendances profondes : pour chaque arc dans le graphe complet nous utilisons une perte de type entropie croisée binaire.

4 Expériences

Nous nous évaluons sur les corpus French Treebank (FTB, Abeillé *et al.*, 2003) et Sequoia (Candito & Seddah, 2012) convertis en dépendances profondes (Perrier *et al.*, 2014). La division du corpus

pour l'entrainement / la validation / le test comporte 14759 / 1235 / 2541 et 2202 / 497 / 400 phrases pour le FTB et Sequoia, respectivement. Pour l'entrainement, nous concaténons les deux corpus. Le corpus FTB contient des structures non connexes (137 phrases sur 2541 dans le test) qui semblent dues à des erreurs de conversion automatique. Nous reportons donc les résultats à la fois sur le corpus de test complet et sur la sous-partie des phrases dont la structure syntaxique profonde de référence est connexe. De plus, nous utilisons deux analyseurs différents : POSITIVEARCS est un analyseur non structuré qui sélectionne tous les arcs de poids positifs, SGCPM est notre analyseur qui ne produit que des analyses valides, c'est-à-dire que des graphes faiblement connexes. Notons que nos résultats sont dans les mêmes ordres de grandeur que ceux de Ribeyre *et al.* (2016) qui obtiennent une f-mesure de 80.79 sans extraction de charactéristiques supplémentaires et 85.18 en utilisant des charactéristiques extérieurs (analyse en constituants, ...) [6].

La première question qui nous intéresse est de savoir si les réseaux de neurones peuvent apprendre implicitement à ne produire que des structures connexes ou s'il est important de forcer cette contrainte lors de la prédiction. Nous reportons le nombre de structures connexes produites par POSITIVEARCS dans la table 2 (partie de gauche). Notons que les structures non connexes peuvent être problématiques lors de l'utilisation des prédictions pour des tâches cibles.

La seconde interrogation porte sur la qualité des prédictions : est ce que garantir des structures connexes lors de la prédiction permet de corriger des erreurs ? Les scores de nos deux réseaux de neurones sont reportés sur le Tableau 2 (partie de droite). Comme nous pouvons l'observer, forcer les structures à être connexe diminue le score en f-mesure, mais augmente le nombre de phrases dont la structure syntaxique est exactement prédite.

5 Conclusion

Dans cet article, nous proposons un nouvel algorithme pour l'analyse en dépendances profondes fondée sur les graphes qui permet de couvrir l'ensemble des jeux de données du français. Nous avons évalué notre approche sur le français. Nous observons que garantir ces structures impacte négativement les résultats des prédictions. Cependant, sans cet algorithme une partie des structures produites sont non connexes, ce qui peut poser problème pour des tâches cibles. De plus, le nombre de structures parfaitement prédites augmente lorsque l'on force la connexité.

Remerciements

Nous remercions les 3 relecteurices anonymes pour leurs remarques et suggestions. Nous remercions vivement Marie Candito et Djamé Seddah pour nous avoir aidé avec la mise en place du corpus Sequoia et FTB en dépendances profondes ainsi que pour avoir répondu à nos interrogations. Nous remercions Corentin Ribeyre et Djamé Seddah pour nous avoir aidé à mettre en place l'évaluateur. Nous remercions Maximin Coavoux et Matthieu Labeau pour les relectures.

6. En l'absence d'un archivage par identifiant unique et découpage commun publiquement disponible pour le jeu de données, comme réalisé par exemple pour le corpus Abstract Meaning Representation qui identifie chaque parution de façon unique (LDC2016E25, LDC2014T12, ...), nous ne savons pas à quel point ces résultats sont comparables. Nous supposons que le découpage du FTB de Ribeyre *et al.* (2016) suit celui de la shared task SPMRL comme nous, mais nous ne savons pas exactement si les versions sont comparables (correction des dépendances non-locales, expressions multi-mots...).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In *Treebanks*, p. 165–187. Springer.

CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & DE LA CLERGERIE É. (2014). Deep syntax annotation of the sequoia French treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2298–2305, Reykjavik, Iceland : European Language Resources Association (ELRA).

CANDITO M. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of TALN 2012*.

CAO J., HUANG S., SUN W. & WAN X. (2017). Parsing to 1-endpoint-crossing, pagenumber-2 graphs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2110–2120, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1193.

CORRO C., LE ROUX J., LACROIX M., ROZENKNOP A. & WOLFLER CALVO R. (2016). Dependency parsing with bounded block degree and well-nestedness via Lagrangian relaxation and branch-and-bound. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 355–366, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1034.

DITTRICH M. T., KLAU G. W., ROSENWALD A., DANDEKAR T. & MÜLLER T. (2008). Identifying functional modules in protein-protein interaction networks : an integrated exact approach. *Bioinformatics*, **24**(13), i223–i231.

DOZAT T. & MANNING C. D. (2017). Deep biaffine attention for neural dependency parsing. In *Proceedings of the 5th International Conference on Learning Representations*.

EDMONDS J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards*, **71**(4), 233–240.

EISNER J. & SATTA G. (1999). Efficient parsing for bilexical context-free grammars and head automaton grammars. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 457–464 : Association for Computational Linguistics.

FREDMAN M. L. & TARJAN R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM (JACM)*, **34**(3), 596–615.

GÓMEZ-RODRÍGUEZ C., WEIR D. & CARROLL J. (2009). Parsing mildly non-projective dependency structures. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, p. 291–299, Athens, Greece : Association for Computational Linguistics.

HAOUARI M., MACULAN N. & MRAD M. (2013). Enhanced compact models for the connected subgraph problem and for the shortest path problem in digraphs with negative cycles. *Computers & operations research*, **40**(10), 2485–2492.

IDEKER T., OZIER O., SCHWIKOWSKI B. & SIEGEL A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**(suppl_1), S233–S240.

KUHLMANN M. & JONSSON P. (2015). Parsing to noncrossing dependency graphs. *Transactions of the Association for Computational Linguistics*, **3**, 559–570. DOI : 10.1162/tacl_a_00158.

LOBODA A. A., ARTYOMOV M. N. & SERGUSHICHEV A. A. (2016). Solving generalized maximum-weight connected subgraph problem for network enrichment analysis. In *International Workshop on Algorithms in Bioinformatics*, p. 210–221 : Springer.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model. arXiv preprint : 1911.03894.

MCDONALD R., PEREIRA F., RIBAROV K. & HAJIČ J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 523–530, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

PERRIER G., CANDITO M., GUILLAUME B., RIBEYRE C., FORT K. & SEDDAH D. (2014). Un schéma d'annotation en dépendances syntaxiques profondes pour le français. In *Proc. of TALN 2014*, Marseille, France.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1202.

RIBEYRE C., DE LA CLERGERIE E. V. & SEDDAH D. (2016). Accurate deep syntactic parsing of graphs : The case of French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 3563–3568, Portorož, Slovenia : European Language Resources Association (ELRA).

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.

ZHANG X., CHENG J. & LAPATA M. (2017). Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 665–676, Valencia, Spain : Association for Computational Linguistics.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

L'expression des émotions dans les textes pour enfants : constitution d'un corpus annoté

Aline Étienne [1, 2] Delphine Battistelli [1] Gwénolé Lecorvé[2]
(1) Univ. Paris-Nanterre, CNRS, MoDyCo, 200 av. de la République, 92001 Nanterre, France
(2) Univ Rennes, CNRS, IRISA, 6, rue de Kerampont, 22300 Lannion, France
{aline.etienne,delphine.battistelli}@parisnanterre.fr,
gwenole.lecorve@irisa.fr

Résumé

Cet article présente une typologie de divers modes d'expression linguistique des émotions, le schéma d'annotation sous Glozz qui implémente cette typologie et un corpus de textes journalistiques pour enfants annoté à l'aide de ce schéma. Ces travaux préliminaires s'insèrent dans le contexte d'une étude relative au développement des capacités langagières des enfants, en particulier de leur capacité à comprendre un texte selon des critères émotionnels.

Abstract

Expressing emotions in texts for children: constitution of an annotated corpus.
This paper presents a classification of various types of linguistic expression of emotions, the Glozz annotation schemata implementing this classification and a corpus of journalistic texts for children manually annotated with this schemata. This preliminary work fits in the broader context of a psycholinguistic study of children's language abilities development, and more precisely of their ability to understand a text depending on emotional criteria.

Mots-clés : émotions, schéma d'annotation, textes pour enfants.
Keywords: emotions, annotation schemata, texts for children.

1 Introduction

Nous nous intéressons ici aux modes d'expression des émotions dans les textes pour enfants. Ce travail s'inscrit dans l'objectif à plus long terme du développement d'outils à même de repérer automatiquement des portions de textes difficiles à comprendre pour des enfants jeunes lecteurs. Dans la lignée de travaux en psycholinguistique comme ceux de (Zwaan et Radvansky, 1998), nous abordons la compréhension d'un texte comme la construction d'une représentation mentale de la situation décrite par le texte, intégrant plusieurs dimensions situationnelles (le temps, l'espace, la cause par exemple). À l'instar de travaux comme (Blanc, 2010), nous intégrons les émotions ressenties par les personnages (à distinguer des émotions ressenties par le lecteur (Dijkstra et al., 1995)) aux dimensions situationnelles d'un texte, et nous nous intéressons en particulier ici à leurs divers modes d'expression linguistique. Ces émotions sont en effet primordiales pour la compréhension de texte puisqu'elles permettent de saisir pourquoi les personnages agissent de telle ou telle manière et ainsi de (re)créer des liens de causalité entre des situations évoquées dans un texte. De fait, il a été montré que, rendus saillants par des émotions, certains évènements (ceux présentant un enjeu pour les personnages), sont mémorisés puis restitués plus aisément, chez les

Volume 2 : Traitement Automatique des Langues Naturelles, pages 205–212.
HAL : hal-02784769.

enfants (Blanc, 2010 ; Davidson et al., 2001), comme chez les adultes (Dijkstra et al., 1995). Le présent article décrit le corpus annoté en émotions constitué afin de mieux cerner l'expression linguistique des émotions dans des textes journalistiques pour enfants. Après un état de l'art du traitement des émotions en psycholinguistique, en linguistique et en TAL (section 2), nous détaillons les étapes de l'annotation manuelle du corpus (section 3). Les premiers résultats d'exploration de ce corpus annoté sont abordés en section 4.

2 Les émotions en psycholinguistique, en linguistique et en TAL

En psycholinguistique, l'impact favorable des émotions sur la compréhension des textes chez les enfants a été démontré, voir notamment (Blanc, 2010) et (Davidson et al., 2001). En tant qu'individus en construction (développement linguistique, cognitif etc.), les enfants ne saisissent cependant pas toujours complètement l'émotion exprimée, ce qui pourrait amoindrir l'effet positif escompté de la présence d'émotions sur l'accès au texte. Il a ainsi été montré qu'avant 10 ans, les émotions de base (colère, dégoût, joie, peur, surprise et tristesse, cf. (Ekman, 1992)) sont mieux retenues que les émotions complexes (culpabilité, embarras, fierté, jalousie, cf. (Lewis, 2008)) (Davidson, 2006 ; Blanc et Quenette, 2017). La façon dont les émotions sont exprimées influence également leur compréhension (Blanc, 2010 ; Creissen et Blanc, 2017). De 6 à 10 ans, les émotions explicitement mentionnées dans un texte, que ce soit *via* des descriptions de comportements associés à une émotion - par exemple "il éclate en sanglots" - ou *via* l'emploi d'unités lexicales désignant des émotions (heureux, effrayer, ...), sont mieux comprises que les émotions simplement suggérées par la situation exposée dans un texte (ex. "Le loup arrive"). Dans l'étude de la compréhension de textes par les enfants, l'expression linguistique des émotions occupe ainsi une place déterminante. Cette question représente par ailleurs un enjeu de taille en linguistique. Le rôle crucial que les émotions jouent dans la communication verbale (ex. persuader son auditoire lors d'une argumentation rhétorique, (Micheli, 2010)) les érige en objet d'étude de choix mais leur omniprésence dans le discours rend leur modélisation linguistique particulièrement complexe (Micheli, 2014). De nombreux travaux se focalisent alors sur l'étude du lexique émotionnel (ex. EmoBase (Diwersky et al., 2014)), délaissant les marqueurs ne relevant pas du lexique des émotions (ex. structures syntaxiques ou discursives). Lorsque ces marqueurs sont pris en compte, les travaux n'abordent le plus souvent que des catégories sémantiques restreintes (ex. analyse des expressions de la peur (Bresson et Dobrovol'skij, 1995) ; analyse des structures (morpho-)syntaxiques exprimant la joie et la rage (Gross, 1995)) ou des structures précises (ex. structures des verbes de sentiments (Mathieu, 2006)). La typologie des "modes de sémiotisation" des émotions de (Micheli, 2014) fait alors figure d'exception, puisque ce travail propose une caractérisation générale (absence de focus sur une catégorie émotionnelle spécifique) des différentes manières d'exprimer les émotions dans les textes destinés aux adultes. Cette typologie distingue les émotions dénotées directement par le lexique (appelées émotions "désignées"), celles que l'on devine grâce à la structure syntaxique d'un énoncé et aux choix des mots opérés par l'énonciateur (appelées émotions "montrées") et celles suggérées par la situation décrite par le texte (appelées émotions "étayées").

Les émotions constituent une catégorie sémantique dont les marqueurs linguistiques, impliquant plusieurs niveaux de la langue (lexique, syntaxe etc.), sont difficiles à circonscrire. Dès lors, le repérage automatique des catégories émotionnelles dans un texte paraît être une tâche délicate. Les quelques travaux de TAL qui s'y attellent ont recours à des lexiques émotionnels et n'intègrent pas, à notre connaissance, de marqueurs linguistiques autres que strictement lexicaux. (Mohammad, 2011) emploie par exemple un lexique émotionnel (le NRC Emotion Lexicon) afin de repérer et de quantifier automatiquement les termes de 8 catégories émotionnelles (colère, dégoût, joie, peur, surprise, tristesse, confiance et anticipation/attente) dans des contes de fées et romans anglophones. Il est à noter par ailleurs que, même si fortement connexes, les travaux sur la détection d'émotions

sont à différencier selon nous des travaux sur le repérage de sentiments (catégorisés principalement en positifs, négatifs ou neutres) ou d'opinions (catégorisés principalement en favorables, défavorables ou neutres). (Hamon et al., 2015) illustrent cette distinction en présentant différents travaux de l'édition 2015 du défi fouille de texte (DEFT), visant notamment à repérer automatiquement dans des tweets francophones des sentiments (satisfaction, insatisfaction), des opinions (accord, valorisation, désaccord, dévalorisation) et des émotions (plaisir, apaisement, amour, surprise positive, déplaisir, dérangement, mépris, surprise négative, peur, colère, ennui, tristesse). Ces travaux ont utilisé des lexiques et des méthodes d'apprentissage automatique. Si nous nous situons dans le champ de la détection d'émotions, des correspondances pourront malgré tout être faites entre les catégories émotionnelles que nous annotons et les polarités négatives et positives, établissant ainsi un pont avec le domaine de l'analyse de sentiments. Enfin, l'annotation des émotions se révèle également être une tâche complexe, comme le montrent (Bostan et Klinger, 2018) en comparant différentes ressources annotées en émotions. Ils mettent ainsi en évidence des divergences aussi bien dans la nature des textes annotés (contes, blogs, actualité ...) que dans le choix des catégories émotionnelles (en lien avec diverses approches psychologique des émotions, notamment celle de (Ekman, 1992)) ou les procédures d'annotations employées (annotation par experts, *crowdsourcing* ...) - diversité à laquelle vient s'ajouter le présent travail.

A ce jour, dans le champ du TAL, aucun travail de caractérisation linguistique fine de l'expression des émotions dans les textes pour enfants ni aucune ressource permettant d'effectuer un tel travail ne semblent avoir été produits. Notre travail vise à combler, en partie du moins, ces lacunes.

3 Constitution du corpus annoté en émotions

Notre méthodologie générale peut être résumée aux trois étapes suivantes : (1) Trouver des pistes théoriques issues de la littérature linguistique et psycholinguistique pour explorer l'expression des émotions dans les textes pour enfants ; (2) Procéder à des analyses linguistiques fines d'extraits de notre corpus afin de vérifier la pertinence des pistes théoriques (catégories émotionnelles, marqueurs linguistiques) ; (3) Systématiser l'étude du corpus *via* un schéma d'annotation. L'objectif de (3) est double : confirmer à l'échelle du corpus entier la pertinence des catégories et des critères retenus d'après (2) ; et obtenir des données quantitatives pour dégager des marqueurs linguistiques diversifiés (i.e. autres que strictement lexicaux) des émotions dans les textes pour enfants. Nous avons étudié un corpus de 97 numéros (octobre 2015 à mars 2019) du journal d'actualité en ligne *Le P'tit Libé*, adressé aux 7-12 ans, et qui vise à expliquer "l'actu des grands" aux enfants. Nous disposions donc de 97 fichiers (un numéro par fichier), soit environ 216K tokens.

3.1 Pistes théoriques explorées et choix opérés

La constitution de notre corpus annoté en émotions vise à rendre compte de la diversité des marqueurs linguistiques des émotions (ex. structures syntaxiques) mais aussi à mettre en évidence des liens entre différentes émotions (ex. apparition conjointe fréquente de la peur et de la tristesse) ou l'utilisation préférentielle de certains modes d'expression pour verbaliser certaines émotions. Comme pour toute tâche d'annotation, nous avons établi un schéma d'annotation en définissant les unités linguistiques à repérer au sein du corpus ainsi que les catégories à leur associer. Notre schéma d'annotation devait permettre le repérage de termes du lexique émotionnel (ex. le mot "triste"), mais aussi de structures syntaxiques (ex. dislocations), de marques typographiques (ex. points d'exclamation) voire de structures textuelles (ex. groupe d'énoncés averbaux). La table 1 reprend et met en regard des travaux de linguistique (Micheli, 2014) et en psycholinguistique

(Blanc, 2010, 2017) sur les modes d'expression linguistique des émotions. En nous appuyant sur ces deux types de travaux, nous avons décidé de retenir la distinction terminologique suivante pour les catégories d'émotions : "désignée", "montrée", "étayée" et "comportementale".

	Catégories			
Micheli (2014)	**Émotion « dite »** *ex. Paul est heureux.*		**Émotion « montrée »** *ex. « Ah ! Quel endroit merveilleux », dit Paul.*	**Émotion « étayée »** *ex. Après un long voyage, Paul arrive enfin dans la maison de vacances de ses rêves.*
Blanc *et al.* (2010 ; 2017)	**Émotion « désignée »** *ex. Paul est heureux.*	**« Expression comportementale de l'émotion »** *ex. Paul sourit.*		**Émotion « suggérée »** *ex. Après un long voyage, Paul arrive enfin dans la maison de vacances de ses rêves.*
Catégories retenues	**Émotion « désignée »** *ex. Paul est heureux.*	**Émotion « comportementale »** *ex. Paul sourit.*	**Émotion « montrée »** *ex. « Ah ! Quel endroit merveilleux », dit Paul.*	**Émotion « étayée »** *ex. Après un long voyage, Paul arrive enfin dans la maison de vacances de ses rêves.*

TABLE 1 : Comparatif des typologies des modes d'expression des émotions et catégories retenues

L'étude linguistique de l'expression des émotions en français souligne l'importance du schéma actanciel (i.e. qui fait quoi à qui) des segments textuels émotionnels (Mathieu, 2006 ; Bresson et al., 1995 ; Micheli, 2014). D'après ces travaux et l'analyse d'extraits de notre corpus, nous avons émis l'hypothèse que l'entité affectée par l'émotion (i.e. l'individu ressentant l'émotion) occupe une place centrale dans l'expression et la compréhension des émotions. Nous avons donc décidé d'annoter ces entités, ainsi que le lien qui les unit aux unités émotionnelles. Selon nous, des informations sur ces entités seront pertinentes pour l'analyse linguistique et doivent figurer dans notre schéma d'annotation : la nature du segment textuel désignant l'individu affecté par l'émotion ; et des caractéristiques de l'individu lui-même, avec les traits [±animé,humain] et [±collectif] (i.e. emploi générique d'un référent et non utilisation du pluriel). Ces caractéristiques pourraient fournir des pistes pour régler (en partie) le problème de la polysémie comme dans "Marie est isolée", où "isolée" prend un sens émotionnel (tristesse) grâce au trait [+animé,humain] (vs. "La maison est isolée" où "maison" est [-animé]). Un encodage plus complet du schéma actanciel des émotions nécessiterait une unité pour la cause des émotions (i.e. ce qui déclenche l'émotion chez l'individu). Cependant, d'après l'analyse d'extraits de notre corpus, nous avons supposé que la cause d'une émotion sera moins aisément délimitable que l'entité qui la ressent. Cet élément ne figure donc pas dans notre schéma. Enfin, nous avons inclus une distinction entre les émotions de base vs. complexes, pertinente pour l'étude de la compréhension de textes par les enfants. Afin d'analyser plus finement la réalisation du champ sémantique des émotions, nous avons distingué 10 catégories émotionnelles : 6 émotions "de base" (colère, dégoût, joie, peur, surprise et tristesse) et 4 émotions complexes (culpabilité, embarras, fierté et jalousie), reprises de (Davison, 2006 ; Blanc et Quenette, 2017).

3.2 Schéma d'annotation

Notre schéma aborde les émotions en se focalisant sur le mode d'expression linguistique, avec un type d'unité pour chacun des quatre modes d'expression des émotions et un cinquième pour l'entité qui ressent l'émotion. Des traits associés à chaque type d'unité permettent d'affiner les annotations. Pour les unités émotionnelles, il s'agit du "nom de l'émotion", soit la catégorie émotionnelle

exprimée par le segment textuel (ex. *peur* pour “apeuré”, “Quelle horreur !” et “mort”) et de la “nature du segment” textuel. Cette expression, volontairement peu précise, vise à englober la variété des marqueurs linguistiques à repérer. Cette information est particulièrement utile pour les émotions “montrées” car cette catégorie très hétérogène inclut aussi bien des mots isolés (ex. interjections), que des structures syntaxiques (ex. dislocations) ou des groupes de phrases (ex. enchaînement d’énoncés averbaux). L’unité “entité” a trois traits : [± animé,humain] ; [± collectif] ; et “Nature du segment” textuel désignant l’entité ressentant l’émotion. Enfin, notre schéma possède une relation (nommée “Affecte”), typée par le trait “Syntaxe”, servant à relier une unité émotionnelle à l’unité textuelle désignant l’individu affecté par l’émotion.

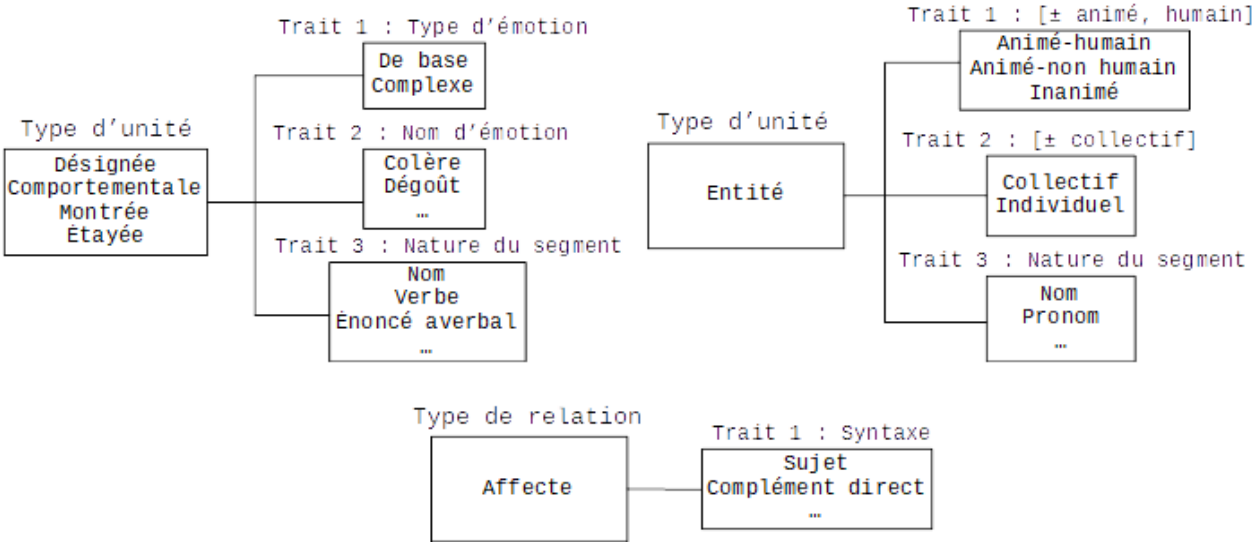


FIGURE 1 : Architecture de notre schéma d’annotation

L’objectif de l’annotation du corpus est de croiser les informations émotionnelles (de base ou complexe, nom de l’émotion) avec d’autres types d’informations (mode d’expression, nature du segment textuel) pour montrer que l’expression des émotions ne repose pas uniquement sur le lexique émotionnel. L’annotation manuelle du corpus a été effectuée *via* la plateforme Glozz (Widlöcher et Mathet, 2012) qui présente plusieurs avantages pour notre travail : (i) grande liberté de schéma d’annotation (traits pour caractériser les unités) ; (ii) diversité des annotations possibles (unités allant du caractère au paragraphe ; chevauchement d’unités ; relations entre unités) ; (iii) pour la suite de nos travaux, calcul d’accord inter-annotateurs aisé grâce à la structure des fichiers Glozz. Une pré-annotation automatique du lexique émotionnel a été opérée par l’application d’une adaptation du lexique EMOTAIX (Piolat & Bannour, 2009) sur le corpus non-lemmatisé *via* un script Python. L’adaptation du lexique a consisté principalement en un remaniement des catégories (conservation des catégories liées à nos 10 émotions ; fusion de catégories proches, ex. *terreur*, *inquiétude*, *angoisse* et *peur* fusionnées en *peur*) et une augmentation en formes fléchies grâce au Lefff (Sagot, 2010) à partir des lemmes du lexique. Cette pré-annotation a ensuite été vérifiée et affinée manuellement, sur Glozz.

4 Exploration du corpus annoté : quelques résultats

Au total, nous avons délimité 2368 unités émotionnelles dans le corpus, dont environ 41,5% d’émotions étayées, 37,1% d’émotions désignées, 11,8% d’émotions comportementales et 9,6% d’émotions montrées. La majorité des unités expriment une émotion de base (89,1%), notamment à cause du nombre plus élevé de catégories d’émotions de base que complexes (6 contre 4). Cependant, une catégorie d’émotion de base est réalisée par 351 unités en moyenne contre 65 pour

une catégorie d'émotion complexe, ces dernières sont donc ici bien moins représentées. Sur l'ensemble des données, l'émotion la plus fréquente est la peur (29,1% des unités). Viennent ensuite la colère (20,9%), la joie (15,2%) et la surprise (13,7%). Les autres catégories représentent moins de 10% des unités (tristesse 9,4%, fierté 5,3%, embarras 4,1%, culpabilité 1,4%, dégoût 0,8% et jalousie 0,1%). La répartition change lorsqu'on regarde les proportions de chaque catégorie émotionnelle selon le mode d'expression (cf. figure 2).

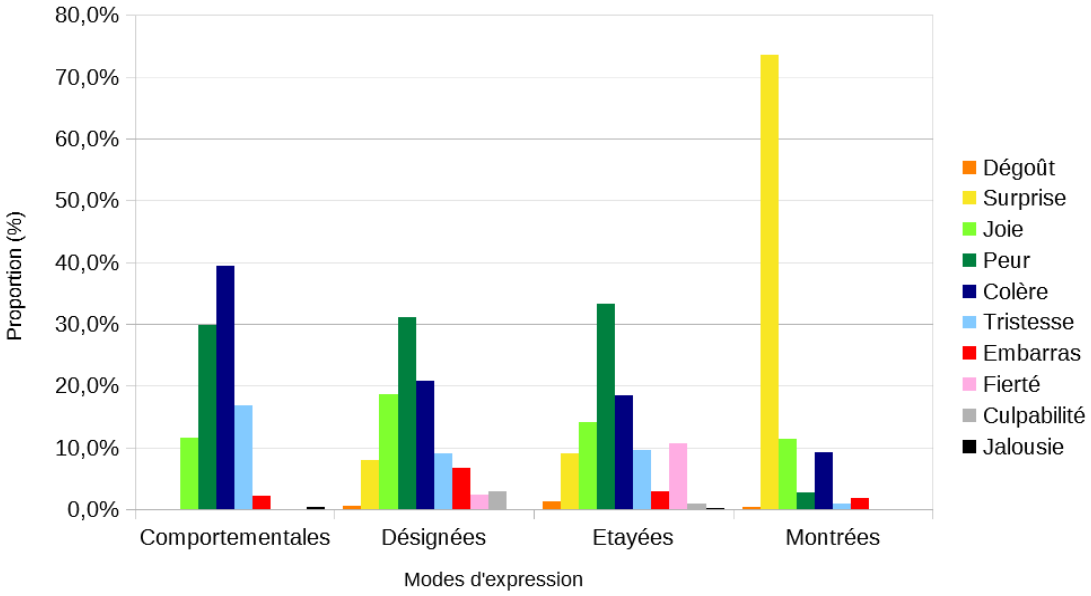


FIGURE 2 : Proportions des catégories émotionnelles selon le mode d'expression

La différence de répartition observée à l'échelle du corpus et en fonction des modes d'expression laisse supposer que des modes d'expression sont utilisés de manière préférentielle pour verbaliser certaines émotions (*ex. montrer la surprise*). Nous avons délimité 1110 unités désignant une entité affectée par une émotion. Il y a moins d'entités que d'unités émotionnelles car soit l'entité n'est pas exprimée, soit plusieurs unités émotionnelles affectent la même entité. Les entités sont réalisées principalement par des pronoms (48,5%) ou des syntagmes nominaux (noms : 37%, noms propres : 11,3%) puis quelques déterminants (3,2%) et un adjectif (dans *à la surprise générale*). Les émotions du corpus affectent surtout des humains (97,3%) et des entités collectives (59,1%).

L'annotation manuelle du corpus a souligné diverses difficultés inhérentes à notre tâche : repérer l'entité qui ressent l'émotion ; identifier précisément l'émotion en jeu parmi les 10 retenues. L'annotation ayant été réalisée par une seule personne, il sera en outre nécessaire de faire une campagne d'annotation pour évaluer la pertinence du schéma d'annotation et la faisabilité de la tâche par un calcul d'accord inter-annotateurs, selon un protocole encore à définir. Nous souhaiterions par ailleurs étendre notre corpus à d'autres genres textuels (encyclopédie, roman).

5 Conclusion

Nous avons constitué une ressource jusqu'ici absente : un corpus de 97 textes journalistiques pour enfants annotés en 10 types d'émotions et selon divers modes d'expression linguistiques mobilisés, soit au total 2368 unités émotionnelles de taille et de nature diverse annotées. Ce corpus annoté contribuera à des travaux dans trois domaines : en linguistique, pour l'exploration de la diversité de l'expression des émotions, en dehors du lexique stricto sensu ; en TAL, pour des systèmes d'apprentissage automatique ; en psycholinguistique, comme matériau de test au sein d'un

protocole expérimental visant à évaluer la compréhension de différents types d'émotions auprès de populations d'enfants, tâche que nous poursuivons actuellement au sein du projet ANR TextToKids. De plus, nous envisageons plusieurs pistes pour approfondir notre étude préliminaire, notamment une comparaison avec l'expression des émotions dans des textes non destinés aux enfants, la mise en regard de l'expression des émotions selon les différentes thématiques des articles et selon leurs auteurs ou encore l'enrichissement de notre schéma d'annotation en intégrant la cause des émotions.

Remerciements

Ce travail a bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR) dans le cadre du projet TREMoLo (ANR-16-CE23-0019) et du projet TextToKids (ANR AAPG 2019).

Références

BLANC, N. (2010). La compréhension des contes entre 5 et 7 ans: Quelle représentation des informations émotionnelles?. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *64*(4), 256–265. DOI : 10.1037/a0021283.

BLANC, N., & QUENETTE, G. (2017). La production d'inférences émotionnelles entre 8 et 10 ans : quelle méthodologie pour quels résultats ?. *Enfance*, 4(4), 503-511. DOI : 10.4074/S0013754517004141.

BOSTAN L.-A.-M. & KLINGER R. (2018, August). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2104-2119).

BRESSON, D., & DOBROVOL'SKIJ, D. (1995). Petite syntaxe de la « peur ». Application au français et à l'allemand. *Langue française*, 105, 107-119. DOI : 10.3406/lfr.1995.5297.

CREISSEN, S., & BLANC, N. (2017). Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10 ans? Apports d'une étude multimédia. *Psychologie française*, 62(3), 263-277. DOI : 10.1016/j.psfr.2015.07.006.

DIJKSTRA, K., ZWAAN, R. A., GRAESSER, A. C., & MAGLIANO, J. P. (1995). Character and reader emotions in literary texts. *Poetics*, *23*(1-2), 139-157. DOI : 10.1016/0304-422X(94)00009-U.

DAVIDSON, D., LUO, Z., & BURDEN, M. J. (2001). Children's recall of emotional behaviours, emotional labels, and nonemotional behaviours: Does emotion enhance memory?. *Cognition & Emotion*, *15*(1), 1-26. DOI : 10.1080/0269993004200105.

DAVIDSON, D. (2006). The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, *30*(3), 232-242. DOI : 10.1007/s11031-006-9037-6.

DIWERSY, S., GOOSSENS, V., GRUTSCHUS, A., KERN, B., KRAIF, O., MELNIKOVA, E., & NOVAKOVA, I. (2014). Traitement des lexies d'émotion dans les corpus et les applications d'EmoBase. *Corpus*, 13, 269-293.

EKMAN, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6:3-4, 169–200. DOI : 10.1080/02699939208411068.

GROSS, M. (1995). Une grammaire locale de l'expression des sentiments. *Langue française*, 105, 70-87. DOI : 10.3406/lfr.1995.5294.

HAMON, T., FRAISSE, A., PAROUBEK, P., ZWEIGENBAUM, P., and GROUIN, C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de

l’édition 2015 du défi fouille de texte (DEFT). Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015), Jun 2015, Caen, France. HAL : hal-01617180.

LEWIS, M. (2008). Self-conscious emotions: embarrassment, pride, shame and guilt. In M. Lewis, J. M. Haviland-Jones, & L.F. Barrett (Eds.), *Handbook of emotions* (pp. 742–756). The Guilford Press.

MATHIEU Y.Y. (2006) A Computational Semantic Lexicon of French Verbs of Emotion. In: SHANAHAN J.G., QU Y., WIEBE J., Éds., *Computing Attitude and Affect in Text: Theory and Applications*. The Information Retrieval Series, vol 20., chapitre 10, p. 109-124, Springer, Dordrecht. DOI : 10.1007/1-4020-4102-0_10.

MICHELI R. (2010). L’émotion argumentée. L’abolition de la peine de mort dans le débat parlementaire français, Paris : Cerf.

MICHELI, R. (2014). *Les émotions dans les discours : modèle d'analyse et perspectives empiriques*. De Boeck.

MOHAMMAD, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 105-114). Association for Computational Linguistics.

PIOLAT, A., & BANNOUR, R. (2009). EMOTAIX: un scénario de Tropes pour l’identification automatisée du lexique émotionnel et affectif. *L’Année psychologique*, *109*(4), 655-698. DOI : 10.4074/S0003503309004047.

SAGOT (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th international conference on Language Resources and Evaluation* (LREC 2010). Istanbul, Turkey. HAL : inria-00521242.

WIDLÖCHER, A., & MATHET, Y. (2012, September). The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering* (pp. 171-180). HAL : hal-01023774.

ZWAAN, R. A., & RADVANSKY, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162-85. DOI : 10.1037/0033-2909.123.2.162.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Traduction automatique pour la normalisation du français du XVII^{e} siècle

Simon Gabay[1] Loïc Barrault[2]
(1) Université de Neuchâtel, Espace Louis-Agassiz 1, 2000 Neuchâtel, Suisse
(2) University of Sheffield, 211 Portobello, Sheffield S1 4DP, Royaume-Uni
prenom.nom@unine.ch, initiale.nom@sheffield.ac.uk

RÉSUMÉ

L'étude des états de langue anciens se heurte à un double problème : d'une part la distance d'avec l'orthographe actuelle, qui empêche de recourir aux solutions standards de TAL, et d'autre part l'instabilité des systèmes graphiques, qui complexifie l'entraînement de solutions directement sur le texte original. Reprenant ce problème d'un point de vue des humanités numériques, nous partons du raisonnement philologique qui sous-tend la création du corpus d'entraînement, avant de recourir aux méthodes traditionnelles de TAL pour comparer deux systèmes de traduction automatique (statistique et neuronale) et offrir un outil fonctionnel pour la normalisation du français classique qui corresponde aux besoins des philologues.

ABSTRACT

Machine Translation for the Normalisation of 17th c. French

The study of old state of languages is facing a double problem : on the one hand the distance with contemporary spelling prevents scholars from using standard NLP solutions, on the other hand the instability of the *scriptae* complexifies the training of solutions directly on the original source text. Returning to this problem with a DH perspective, we start with the philological reasoning behind the creation of the training corpus, and use traditional NLP methods to compare two machine translation systems (statistical and neural) and offer a functional tool for the normalisation of 17th c. French answering the needs of philologists.

MOTS-CLÉS : Normalisation, XVII^{e} siècle, traduction automatique neuronale, traduction automatique statistique, humanités numériques.

KEYWORDS : Normalisation, 17th c. French, Neural Machine Translation (NMT), Statistical Machine Translation (SMT), digital humanities.

1 Introduction

Le français pré-orthographique est un problème pour deux raisons. D'une part, même si on peut le déplorer (Gabay, 2014), les lecteurs ont pris l'habitude de lire la langue de Molière avec une orthographe contemporaine. D'autre part, la variation graphique complique l'approche computationnelle de la langue : elle altère le rapport token/type (désormais TTR) pour les études stylométriques (Kestemont, 2012 ; Pinche *et al.*, 2019), complique la lemmatisation (Manjavacas *et al.*, 2019) et l'extraction d'information (Pettersson, 2016)… Pour toutes ces raisons, plusieurs solutions ont été testées afin de normaliser les données : les systèmes à base de règles (Bollmann *et al.*, 2011), la traduction

Volume 2 : Traitement Automatique des Langues Naturelles, pages 213–222.
HAL : hal-02784770.

automatique statistique (*Statistical Machine Translation*, désormais SMT) (Pettersson *et al.*, 2013 ; Sánchez-Martínez *et al.*, 2013 ; Scherrer & Erjavec, 2013) ou, plus récemment, la traduction automatique neuronale (*Neural Machine Translation*, désormais NMT)(Bollmann & Søgaard, 2016).

De nombreuses évaluations ont montré l'efficacité des deux dernières solutions (Pettersson *et al.*, 2014 ; Bollmann, 2019), dont le défaut est cependant bien connu : la quantité de données nécessaires pour l'entraînement d'un modèle. De telles données existent pour de nombreuses langues (anglais, allemand, hongrois, islandais, portugais, slovène, espagnol ou suédois), mais pas pour le français. Nous nous proposons donc de reprendre l'approche comparatiste de Bollmann & Søgaard (2016) mais, contrairement à ces derniers, nous travaillons sur la langue française classique, et surtout nous n'opérons pas sur des mots isolés pris dans un dictionnaire mais sur des phrases entières. Il nous est ainsi possible de traiter les mots en contexte, et donc de résoudre des ambiguïtés où des homographes auraient différentes versions modernisées, mais aussi de fournir une évaluation en condition réelle d'utilisation.

Afin de mener à bien cette tâche, nous avons décidé de créer un corpus parallèle français classique/français contemporain pour entraîner un normaliseur. Ce corpus, en constante évolution, a été pensé pour être aussi représentatif que possible, et tente de couvrir le XVII^{e} siècle dans toute sa diversité. Après une présentation détaillée des enjeux philologiques soulevés par la question de la normalisation, nous explicitons la démarche ayant mené à la création d'un modèle immédiatement fonctionnel pour les chercheurs en humanités numériques spécialisés dans la philologie.

2 Normalisation

Il existe une multitude de possibilités pour normaliser un état de langue pré-orthographique : transcription interprétative (Stutzmann, 2011), canonicalisation phonétique (Jurish, 2011), lemmatisation et étiquetage morpho-syntaxique, alignement avec le français contemporain… Chacune de ces solutions a ses avantages et ses inconvénients, mais nous nous concentrerons ici, pour les raisons évoquées au tout début de cet article, sur la dernière solution.

En pratique, normaliser le français classique implique la correction de plusieurs caractéristiques : la segmentation (*sur tout* → *surtout*), le trait d'union (*long-temps* → *longtemps*), les lettres calligraphiques (*j'ay* → *j'ai*), les archaïsmes alphabétiques (*fecours* → *secours*), les accents (*Ame* → *Âme*), les lettres étymologiques (lat. *Voster* → frm. *voſtre* →*votre*), les lettres ramistes (*vn* → *un*, *Ie* → *Je*), les changements phonétiques (*craignois* → *craignais*)…

Source	Cible
Sur tout ie redout**o**is cette Mélancolie	**Surtout j**e redout**a**is cette Mélancolie
Où j'a**y** v**eu f**i long**-**temps vo**f**tre **A**me enseuelie.	Où j'a**i** v**u s**i longtemps votre **Â**me ensevelie.
Ie craign**o**is que le Ciel, par **v**n cruel ſecours,	**J**e craign**a**is que le Ciel, par **u**n cruel secours,

TABLE 1 – Exemple de normalisation

Il est important de noter que, concernant la normalisation de sources littéraires, il ne peut y avoir un alignement complet avec le français contemporain. Certaines graphies affectant la prosodie ne peuvent être changées, comme *jusques*+voyelle (aujourd'hui *jusqu'*+voyelle) pour conserver la versification (par. exp. "Portez-vous, s'il le faut, juſques à le haïr." où "juſques à" nécessite trois voyelles : [ʒys-kə-za]). Le problème est identique pour *encor(e)*.

Ajoutons que la liste des opérations nécessaires à la normalisation présentée *supra*, comme la fréquence des types de variation dépendent fortement de l'extension diachronique et diatopique des données d'entraînement. L'hétérogénéité du corpus source grandit donc exponentiellement avec l'extension d'un corpus censé être représentatif d'un état de langue.

3 Corpus

La construction d'un outil fonctionnel nécessite un travail important sur la création du corpus. Puisque l'efficacité du normaliseur dépend des données d'entraînement, elles doivent être soigneusement choisies afin de couvrir la langue classique dans toute sa diversité. Trois points ont tout particulièrement attiré notre attention :

— L'extension du lexique est extrêmement importante. Si on peut difficilement espérer traiter les textes littéraires, médicaux, théologiques… avec la même efficacité, il est possible d'éviter de grosses déconvenues en veillant à ce que ces différents genres, et donc leur lexique, soient représentés dans les données d'entraînement.
— Le système graphique est instable tout le XVII^{e} siècle et cette variation évolue dans le temps (Biedermann-Pasques, 2017) : en une centaine d'année, le français passe rapidement d'une multitude de *scriptae* [1] à un état quasi-stable (Bonhomme, 2011). En plus de la variation diachronique, on note une importante polarisation diastratique entre les Anciens et les Modernes, qui défendent des options graphiques totalement opposées (Pellat, 1995b).
— L'utilisation des majuscules est extrêmement différente de la nôtre et ne semble pas suivre de règle claire. S'il est possible de contourner ce problème en abaissant la casse de tous les caractères, agir ainsi nous fait perdre de la lisibilité et de l'information.

Concernant le lexique, nous avons décidé de créer un corpus primaire et un corpus secondaire. Le corpus primaire se concentre sur les textes littéraires avec la poésie (Viau), le roman (La Fayette), la tragédie (Racine), la comédie (Molière), la correspondance (Guez de Balzac). Le corpus secondaire étend le lexique en ajoutant des documents traitant de physique (Pascal), de médecine (Ellain), de théologie (Sales, Bossuet), de philosophie (Descartes).

Parmi tous les textes, nous avons délibérément augmenté le nombre de pièces de théâtre pour deux raisons. La première est que les textes dramatiques utilisent plus la majuscule (acte, scène, tours de parole) que les autres genres. La seconde est que le théâtre est l'un des genres les mieux connus en ce qui concerne l'histoire des livres au XVII^{e} siècle (Riffaud, 2009) et que l'on connaît donc le nom de l'imprimeur pour chaque livre – une information importante pour équilibrer au mieux le corpus.

Concernant les systèmes graphiques, les textes de ce double corpus sont distribués chronologiquement, avec au moins deux imprimés par décennie. Si la plupart sont imprimés à Paris, certains proviennent des Flandres (Bussy-Rabutin à Bruxelles) ou de Hollande (Descartes à Leyde), où le système graphique est différent (Pellat, 1995a). Cette distribution chronologique et géographique n'est pas parfaite car elle dépend de documents OCRisés avec un modèle en cours de construction (Gabay, 2019).

1. Une *scripta* est une koinè graphique, *i.e.* une langue écrite partagée par un large groupe de scripteurs d'une même langue.

4 Entraîner et évaluer les modèles

La création de données d'entraînement a été faite avec l'aide d'un système à base de règles et d'expressions régulières pour accélérer le processus (Gabay *et al.*, 2019) : chaque phrase est pré-normalisée automatiquement avant d'être manuellement corrigée pour garantir la qualité des résultats. Ainsi, un corpus de référence de c. 140 000 tokens a pu être produit rapidement pour effectuer des premiers tests.

Deux formats de sortie sont disponibles : tsv, mais aussi XML-TMX. Ce dernier permet de conserver les données dans un format aisément manipulable, mais aussi de réutiliser les données avec de logiciels acceptant des mémoires de traduction comme MateCat (Federico *et al.*, 2014) ou des outils de textométrie comme TXM (Heiden *et al.*, 2010). Pour l'entraînement, chaque texte a été segmenté en lignes, constituées dans la majorité des cas de phrases, mais aussi de propositions dans le cas où les phrases se sont avérées trop longues (plus de soixante mots). 10% des lignes de chaque texte a été prélevé aléatoirement pour évaluer le modèle, et le reste a été utilisé pour l'entraînement.

	Lignes	Tokens	Caractères
Train	8 962	128 146	571 290
Test	814	9 068	40 253
Total	9 776	137 342	611 543

TABLE 2 – Constitution du corpus

Tous les résultats sont évalués avec l'exactitude de mots (*word accuracy* ou *Wacc*) selon les recommandations de M. Bollmann (Bollmann, 2018), mais nous fournissons aussi les scores BLEU (Papineni *et al.*, 2002) et METEOR (Denkowski & Lavie, 2014) car ces mesures sont toujours utilisées par d'autres collègues (Domingo & Casacuberta, 2018b) et qu'elles permettent donc des comparaisons avec d'autres publications. Il doit aussi être noté que, dans notre corpus d'entraînement, nous avons délibérément augmenté l'hétérogénéité des données, ce qui ne peut qu'abaisser les scores finaux.

Nous avons décidé de comparer les traductions automatiques neuronale et statistique de niveau caractère (*character level Statistical Machine Translation*, désormais cSMT).

4.1 cSMT

L'efficacité de la traduction statistique de niveau caractère ayant été plusieurs fois démontrée (Ljubešić *et al.*, 2016), nous avons décidé de ne pas tester différents degrés de granularité (mot, sous-mot, caractère…) comme avec la NMT (cf. *infra*) mais de tenter l'ajout de données avec un modèle de langue (Scherrer & Ljubešić, 2016) et la rétro-traduction (Domingo & Casacuberta, 2018a). Concernant les modèles de langue, nous avons utilisé deux corpus différents :

— Un premier constitué de 3 151 778 tokens tirés de textes classiques normalisés proches de ceux des données d'entraînement : du théâtre (Molière, Racine, Corneille), des recueils (*Nouvelles nouvelles* de Donneau de Visé), des romans (*Histoire amoureuse* de Bussy), des nouvelles (*Historiettes* de Tallemant des Réaux), des sermons (Bossuet), des essais philosophiques (les œuvres complètes de Descartes et de Pascal), de la correspondance (Sévigné, La Fayette) et des textes religieux (La Bible de Lemaître de Sacy, Saint Augustin et Saint Thomas) pour anticiper de possibles références religieuses.

— Un second corpus fait de 89 972 791 tokens en ajoutant au premier corpus la totalité des données en français du projet Gutenberg.

Nous avons aussi essayé la rétro-traduction : les données normalisées deviennent la source et les données originales la cible. Un nouveau jeu de données de 121 012 tokens (8 905 lignes) composées de pièces (Corneille), d'essais (La Rochefoucauld), de sermons (Bossuet) et textes variés (Donneau de Visé) ont ainsi été convertis en textes pseudo-anciens. Pour ce faire un troisième modèle de langue a été créé avec des transcriptions non-normalisées provenant de wikisource (La Fayette, Racine, La Fontaine, Tristan l'Hermite) et de deux grosses éditions publiées en ligne (*L'Astrée* d'Honoré D'Urfé et *Artamène* des Scudéry).

	BLEU (4-grammes)	METEOR	wAcc
Normal	77,667	87,891	86.68496
+ML	77,108	87,308	86.4022
+trad. inv.	76,680	87,024	86.06121
+trad. inv.+ML	75,766	86,257	85.57052

TABLE 3 – Evaluation des modèles avec cSMT

Les résultats sont bons, mais il semble que les différentes techniques utilisées pour améliorer les scores aient un effet neutre, voire négatif sur le score final.

4.2 NMT

En ce qui concerne la traduction automatique neuronale, nous avons décidé d'utiliser NMTPy-Torch (Caglayan *et al.*, 2017). Le modèle de base est composé d'un encodeur GRU bidirectionnel (Cho *et al.*, 2014) et d'un décodeur (GRU conditionnel à deux couches (Sennrich *et al.*, 2017)) avec mécanisme d'attention de type perceptron multicouche (Bahdanau *et al.*, 2015). L'encodeur et le décodeur ont tous les deux 400 unités cachées et leur état initial caché est initialisé à zéro. Les plongements lexicaux ont une taille fixée à 200.

Trois versions du système ont été entraînées : une première au niveau mot, une seconde avec des unités *byte pair encoding* ou *BPE* (Sennrich *et al.*, 2015) opérant au niveau du sous-mot, et une troisième au niveau du caractère. Le tableau 4 montre le résultats de tels pré-traitements. Chaque système a été entraîné quatre fois avec une initialisation différente, deux fois avec *drop out*, deux fois sans. Seuls les modèles avec les meilleurs scores BLEUS sont conservés.

Les résultats sont particulièrement encourageants, car ils améliorent de précédents tests (BLEU de 82.96) malgré un doublement du nombre de données d'entraînement (c. 60 000 vs 140 000 tokens) et une nette augmentation de leur diversité (2 imprimés vs 30) (Gabay *et al.*, 2019).

Granularité	Version	Phrase
Mot	Source	Cherchons avec empreſſement
	Target	Cherchons avec empressement
BPE	Source	Ch@@ er@@ ch@@ ons avec em@@ preſ@@ ſement
	Target	Ch@@ er@@ ch@@ ons avec em@@ pr@@ ess@@ ement
Caractères	Source	C h e r c h o n s • a v e c • e m p r e ſ ſ e m e n t
	Target	C h e r c h o n s • a v e c • e m p r e s s e m e n t

TABLE 4 – Exemple de phrase pré-traitée à différentes granularités. '@@' après une unité sous-mot spécifie que l'unité ne termine pas le mot. '•' représente un espace du texte initial pour le traitement au niveau des caractères.

	BLEU (4-grammes)	METEOR	wAcc
Word	83,647	90,783	91.41786
BPE	60,200	71,836	71.11808
characters	71,283	76,816	77.86646

TABLE 5 – Évaluation des modèles avec NMT

5 Analyse des résultats

En plus des scores précédemment présentés, revenons concrètement sur le type de résultat produit par chacun des systèmes. Malgré des scores similaires, on peut noter quelques différences assez importantes :

Version	Exemple
Original	En cet eſtat **la**, Monſeigneur, il n'y a point d'apparence de ſonger à la Valtoline, ny de ietter les yeux ſur le bien d'autruy cependant qu'on nous **diſpute** le noſtre, & qu'il faut que nous le **tenions** auecque les deux mains de peur qu'il ne nous **eſchappe**.
NMT	en cet état **la**, Monseigneur, il n' y a point d'apparence de songer à la Valteline, ni de jeter les yeux sur le bien d'autrui cependant qu'on nous **dispute** le nôtre, et qu' il faut que nous le **commun** avec les deux mains de peur qu'il ne nous **apportée**.
cSMT	en cet état, monseigneur, il n'y a point d'apparence de songer à la valteline, ni de jeter les yeux sur le bien d'autrui cependant qu'on nous **disputer** le nôtre, et qu'il faut que nous le **contentions** avec les deux mains de peur qu'il ne nous **échappe**.

TABLE 6 – Comparaison des résultats

Phénomème attendu, les deux systèmes commettent des fautes sensiblement sur les mêmes passages. Les erreurs de NMTPyTorch (*commun*) sont néanmoins beaucoup plus éloignées de l'original (*tenions*) que celles de cSMTiser (*contentions*), dont les bévues, plus nombreuses, sont souvent peu conséquentes (*dispute* vs *disputer*). Ce dernier système a en revanche tendance à "oublier" (plus rarement "rajouter") des mots (*la*).

Ajoutons que le passage du français classique au français contemporain réduit la richesse lexicale (passage d'un TTR de 0.11 à 0.092) : la question des homographes est donc peu importante. On remarque cependant dans cet exemple que les deux systèmes ont su, comme partout ailleurs, distinguer sur la base du contexte (présence du déterminant *le*) le pronom *nôtre* de l'adjectif *notre* en dépit d'une forme unique en français classique (*noſtre*).

6 Conclusion

Il est donc relativement simple d'approcher les 90% de *wAcc* avec les systèmes actuels, sans configuration spéciale. Malgré des résultats plus que satisfaisants, le cSMT est moins performant que le NMT, et cela même avec un corpus d'entraînement de petite taille. Le score de 90%, obtenu uniquement par le NMT, est proche de l'état de l'art (Bollmann, 2019) malgré des données d'entraînement

d'une grande hétérogénéité, ce qui démontre l'extrême robustesse du système.

Cette grande hétérogénéité, volontairement introduite, permet de garantir l'efficacité du modèle pour les philologues. La performance n'est en effet pas évaluée sur des listes de mots uniques, comme cela se fait souvent pour les bancs d'essai, mais sur des données proches de celle que rencontrent les philologues au quotidien.

Du point de vue informatique, les futurs recherches doivent s'orienter vers les solutions neuronales, dont l'efficacité sera clairement améliorée par les techniques que nous avons utilisées dans cet article avec le cSMT. Il conviendra donc, parallèlement à une augmentation significative des données d'entraînement, de tester la rétro-traduction et l'utilisation d'un modèle de langue pré-entraîné comme CamemBERT (Martin *et al.*, 2019) pour améliorer encore plus les résultats.

Du point de vue philologique, il serait intéressant de tenter un partitionnement des données d'entraînement afin de tester l'efficacité de modèles entraînés sur des jeux de données plus restreints, mais plus homogènes. Les cadrages chronologiques qui produiraient les modèles les plus efficaces seraient porteurs d'une information linguistique permettant de repenser la périodisation de la langue moderne sur d'autres critères que l'histoire des idées.

Du point de vue ecdotique enfin, il conviendrait de réfléchir à l'utilité d'autres types de normalisation, car il n'est pas certain que l'alignement sur l'orthographe actuelle soit le seul choix souhaitable. La création de modèles permettant une normalisation plus légère, comme les dissimilations *i* vs *j* et *u* vs *v*, permettrait de produire facilement des textes aisément lisibles qui ne perdraient pas toute leur richesse linguistique.

Remerciements

Merci à A. Baillot, organisatrice de la *Human-Machine Translation German-French Summer School*, et à Y. Scherrer pour son aide dans l'optimisation des résultats de cSMTiser.

Références

Bahdanau D., Cho K. & Bengio Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

Biedermann-Pasques L. (2017). *Les grands courants orthographiques au XVII^e siècle et la formation de l'orthographe moderne, Impacts matériels, interférences phoniques, théories et pratiques (1606–1736)*. Berlin, Boston : De Gruyter, reprint 2017 édition. doi : 10.1515/9783110938593.

Bollmann M. (2018). *Normalization of Historical Texts with Neural Network Models*. Thèse de doctorat, Ruhr-Universität Bochum, Bochum.

Bollmann M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3885–3898, Minneapolis, Minnesota : Association for Computational Linguistics. doi : 10.18653/v1/N19-1389.

Bollmann M., Petran F. & Dipper S. (2011). Rule-Based Normalization of Historical Texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage (DigHum 2011)*, p. 34–42, Hissar, Bulgaria.

Bollmann M. & Søgaard A. (2016). Improving Historical Spelling Normalization With Bi-Directional LSTMs and Multi-Task Learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 131–139, Osaka, Japan. Anthologie ACL : C16-1013.

Bonhomme M. (2011). La standardisation du français au XVII^e siècle. Le cas des observations sur la langue françoise de ménage. In *Du système linguistique aux actions langagières : Mélanges en l'honneur d'Alain Berrendonner*, Champs linguistiques. Bruxelles : De Boeck Supérieur.

Caglayan O., García-Martínez M., Bardet A., Aransa W., Bougares F. & Barrault L. (2017). NMTPY : A Flexible Toolkit for Advanced Neural Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*, **109**(1), 15–28. arXiv : 1706.00457, doi : 10.1515/pralin-2017-0035.

Cho K., van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. & Bengio Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1724–1734, Doha, Qatar. arXiv : 1406.1078.

Denkowski M. & Lavie A. (2014). Meteor universal : Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Domingo M. & Casacuberta F. (2018a). A Machine Translation Approach for Modernizing Historical Documents Using Back Translation. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, p. 39–47, Bruges, Belgium.

Domingo M. & Casacuberta F. (2018b). Spelling Normalization of Historical Documents by Using a Machine Translation Approach. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, p. 129–137, Alicante, Spain.

Federico M., Bertoldi N., Cettolo M., Negri M., Turchi M., Trombetti M., Cattelan A., Farina A., Lupinetti D., Martines A., Massidda A., Schwenk H., Barrault L., Blain F., Koehn P., Buck C. & Germann U. (2014). The MateCat Tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : System Demonstrations*, p. 129–132, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

Gabay S. (2014). Pourquoi moderniser l'orthographe ? Principes d'ecdotique et littérature du XVII^e siècle. *Vox Romanica*, **73**(1).

Gabay S. (2019). OCRising 17th French prints. *E-ditiones*.

Gabay S., Riguet M. & Barrault L. (2019). A Workflow For On The Fly Normalisation Of 17th c. French. In *DH2019*, Utrecht, Netherlands : ADHO.

Heiden S., Magué J.-P. & Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, volume 2, p. 1021–1032, Rome, Italy : Edizioni Universitarie di Lettere Economia Diritto.

Jurish B. (2011). *Finite-state canonicalization techniques for historical German*. Thèse de doctorat, Universität Potsdam, Potsdam.

Kestemont M. (2012). Stylometry for Medieval Authorship Studies : An Application to Rhyme Words. *Digital Philology : A Journal of Medieval Cultures*, **1**(1), 42–72. doi : 10.1353/dph.2012.0002.

Ljubešić N., Zupan K., Fišer D. & Erjavec T. (2016). Normalising Slovene data : historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, p. 146–155, Bochum, Germany.

Manjavacas E., Kádár Á. & Kestemont M. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 1493–1503, Minneapolis, Minnesota : Association for Computational Linguistics. doi : 10.18653/v1/N19-1153.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., Villemonte de la Clergerie É., Seddah D. & Sagot B. (2019). CamemBERT : a Tasty French Language Model. *arXiv e-prints*, p. arXiv :1911.03894. arXiv preprint : 1911.03894.

Papineni K., Roukos S., Ward T. & Zhu W.-J. (2002). BLEU : A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 311–318, Philadelphia, USA. doi : 10.3115/1073083.1073135.

Pellat J.-C. (1995a). L'évolution de l'orthographe des imprimés au XVIIe s. (libraires français et hollandais). In *« Ces mots qui sont nos mots ». Mélanges d'Histoire de la Langue Française, de Dialectologie et d'Onomastique offerts au professeur Jacques Chaurand*, p. 83–96. Institut Charles Bruneau.

Pellat J.-C. (1995b). Norme et variation orthographique au XVII^{e} siècle. In *Rencontres linguistiques en pays rhénan 5/6*, volume 3 de Sciences Cognitives, Linguistiques & Intelligence Artificielle, p. 245–260, université des sciences humaines de Strasbourg : Université des sciences humaines Strasbourg.

Pettersson E. (2016). *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*. Studia Linguistica Upsaliensia. Acta Universitatis Upsaliensis.

Pettersson E., Megyesi B. & Nivre J. (2013). Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, p. 163–179, Oslo, Norway.

Pettersson E., Megyesi B. & Nivre J. (2014). A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, p. 32–41, Gothenburg, Sweden. doi : 10.3115/v1/W14-0605.

Pinche A., Camps J.-B. & Clérice T. (2019). Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, Netherlands : ADHO and Utrecht University.

Riffaud A. (2009). *Répertoire du théâtre français imprimé entre 1630 et 1660 - Librairie Droz*. Travaux du Grand Siècle. Genève : Droz.

Sánchez-Martínez F., Martínez-Sempere I., Ivars-Ribes X. & Carrasco R. C. (2013). An open diachronic corpus of historical spanish. *Language resources and evaluation*, **47**(4), 1327–1342. doi : 10.1007/s10579-013-9239-y.

Scherrer Y. & Erjavec T. (2013). Modernizing Historical Slovene Words with Character-Based SMT. In *4th Biennial Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, p. 58–62, Sofia, Bulgaria.

Scherrer Y. & Ljubešić N. (2016). Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *KONVENS*.

Sennrich R., Firat O., Cho K., Birch A., Haddow B., Hitschler J., Junczys-Dowmunt M., Läubli S., Barone A. V. M., Mokry J. & Nădejde M. (2017). Nematus : a Toolkit for Neural Machine Translation. In *Proceedings of the EACL 2017 Software Demonstrations*, p. 65–68, Valencia, Spain. Anthologie ACL : E17-3017.

Sennrich R., Haddow B. & Birch A. (2015). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 1715–1725, Berlin, Germany.

Stutzmann D. (2011). Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, p. 247–277.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Prédire le niveau de langue d'apprenants d'anglais

Natalia Grabar[1,2] Thierry Hamon[3,4] Bert Cappelle[2] Cyril Grandin[2]
Benoît Leclercq[2] Ilse Depraetere[2]
(1) CNRS, UMR 8163, F-59000 Lille, France
(2) Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
(3) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
(4) Université Paris 13, Sorbonne Paris Cité, F-93430, Villetaneuse, France
{natalia.grabar;bert.cappelle;cyril.grandin}@univ-lille.fr
{benoit.leclercq;ilse.depraetere}@univ-lille.fr
thierry.hamon@limsi.fr

RÉSUMÉ

L'apprentissage de la deuxième langue (L2) est un processus progressif dans lequel l'apprenant améliore sa maîtrise au fur et à mesure de l'apprentissage. L'analyse de productions d'apprenants intéresse les chercheurs et les enseignants car cela permet d'avoir une meilleure idée des difficultés et les facilités d'apprentissage et de faire des programmes didactiques plus adaptés. Cela peut également donner des indications sur les difficultés cognitives à maîtriser les notions grammaticales abstraites dans une nouvelle langue. Nous proposons de travailler sur un corpus de productions langagières d'apprenants d'anglais provenant de différents pays et donc ayant différentes langues maternelles (L1). Notre objectif consiste à catégoriser ces productions langagières selon six niveaux de langue (A1, A2, B1, B2, C1, C2). Nous utilisons différents ensembles de descripteurs, y compris les verbes et expressions modaux. Nous obtenons des résultats intéressants pour cette catégorisation multiclasse, ce qui indique qu'il existe des différences linguistiques inhérentes entre les différents niveaux.

ABSTRACT

Predict the language level for English learners.

Second language (L2) learning is a progressive process during which the learner improves his/her language proficiency as the learning process progresses. The analysis of linguistic productions of learners attracts the attention of researchers and language teachers because this helps to get a better idea on learning difficulties and easiness, and to prepare more approriate didactic programs. This can also provide indications on cognitive difficulties to master grammatical and abstract notions in a new language. We propose to work with a corpus of language productions of English learners coming from different countries and having different mother tongues (L1). Our purpose is to categorize these language productions according to six language levels (A1, A2, B1, B2, C1, C2). We exploit different sets of descriptors, including modal verbs and expressions. We obtain interesting results for this multiclass categorization, which indicates that these language levels contain inherent linguistic features.

MOTS-CLÉS : Catégorisation supervisée, apprentissage L2, anglais, formules de lisibilité, n-grammes, verbes et expressions modaux.

KEYWORDS: Supervised categorization, L2 learning, English, readability scores, n-grams, modal verbs and expressions.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 223–231.
HAL : hal-02784771.

1 Introduction

Les chercheurs distinguent en général deux catégories en acquisition de langues (Robertson & Ford, 2009) : l'acquisition de la première langue (L1) et de la deuxième langue (L2). L'acquisition de la première langue est un processus universel, inconscient et indépendant de la langue. Ainsi, de jeunes enfants commencent très rapidement à imiter les productions langagières de leurs parents et de l'entourage. Cependant, l'acquisition de L2 suppose la connaissance et la maîtrise de la première langue. Ce processus suppose également que la personne apprenne consciemment les éléments d'une nouvelle langue, comme le vocabulaire, les composants phonologiques, les structures grammaticales et l'écriture. L'apprentissage de L2 est donc un processus progressif dans lequel l'apprenant améliore sa maîtrise au fur et à mesure de l'apprentissage. Les productions d'apprenants L2 intéressent les chercheurs qui veulent comprendre les difficultés d'apprentissage pour une langue donnée, faire des programmes d'apprentissage plus appropriés ou pour étudier les capacités cognitives des élèves à maîtriser les notions plus ou moins abstraites, par exemple.

Les productions langagières en L2 sont étudiées de différents points de vue : étudier un aspect langagier donné (Gibbs, 1990; Moloi, 1998; Watanabe & Iwasaki, 2009; Mortelmans & Anthonissen, 2016; Murakami *et al.*, 2016; Ayoun & Gilbert, 2017; Römer, 2019), faire le parallèle entre l'apprentissage de L1 et de L2 (Laufer & Eliasson, 1993; Chenu & Jisa, 2009; Ipek, 2009; Rabinovich *et al.*, 2016), identifier automatiquement la L1 des apprenants en L2 (Jiang *et al.*, 2014; Malmasi & Dras, 2015; Nisioi, 2015) ou définir le niveau de maîtrise d'apprenants de L2 (Granfeldt & Nugues, 2007; Pilan *et al.*, 2016; Arnold *et al.*, 2018; Balikas, 2018). Les deux premières tâches sont en général étudiées manuellement par les linguistes et didacticiens, alors que les deux autres tâches attirent l'attention des chercheurs en TAL. Les travaux effectués manuellement concernent typiquement l'étude de catégories abstraites, comme la notion de modalité et l'usage de modaux. Tout d'abord, notons qu'il a été observé que, chez les apprenants d'anglais L1, l'usage de modaux apparaît à partir de 2 ans avec des modaux déontiques (comme *can*) et se diversifie progressivement avec l'apparition de modaux épistémiques (comme *must* ou *might*) vers 3 ans (Shatz & Wilcox, 1991; Papafragou, 1998; Cournane, 2015). En ce qui concerne l'apprentissage de modaux d'anglais L2, dans une étude (Gibbs, 1990), les chercheurs ont analysé la compréhension de la valeur des modaux principaux (*can, could, may, might*) par les apprenants d'anglais parlant Panjabi, une langue indienne. Les apprenants devaient définir la valeur sémantique des modaux parmi quatre valeurs possibles : capacité, permission, possibilité et possibilité hypothétique. Dans un autre travail, les chercheurs analysaient la maîtrise des fonctions grammaticale et modale des modaux chez des enfants apprenants d'anglais (Moloi, 1998). L'auteur indique entre autre qu'il existe des similarités dans l'apprentissage de l'anglais comme L1 ou L2. De rares travaux ont porté sur l'acquisition de la modalité dans d'autres langues, comme par exemple en japonais (Watanabe & Iwasaki, 2009) ou en allemand (Mortelmans & Anthonissen, 2016).

Concernant la prédiction du niveau d'apprenants, dans un travail, les productions d'apprenants de français d'origine suédoise ont été analysées du point de vue syntaxique (Granfeldt & Nugues, 2007), en mettant l'accent sur leur étiquetage et analyse syntaxique automatique. 142 descripteurs ont ainsi pu être exploités, comme par exemple le pourcentage de séquences déterminant-nom avec accord, de mots inconnus, de GNs avec accord en genre, de prépositions, de séquences nom-adjectif avec accord, d'accord sujet-verbe avec des verbes modaux et la longueur moyenne des phrases. Les productions ont ensuite été classées automatiquement en cinq stades d'apprentissage. L'utilisation des 10 meilleurs descripteurs avec C4.5 montre une F-mesure entre 0,46 et 0,53 selon les stades. Dans un autre travail, les productions d'apprenants de suédois comme L2 sont analysées de différents points de vue (lexique, syntaxe, morphologie, sémantique) grâce à leur comparaison avec des manuels de

langue (Pilan *et al.*, 2016). Il s'avère que les descripteurs lexicaux apportent le plus de gain dans la définition de la maîtrise du suédois. Actuellement, la plupart des travaux qui cherchent à définir le niveau de maîtrise de L2 sont effectués sur le corpus EFCAMDAT avec six niveaux d'apprenants (Geertzen *et al.*, 2013; Huang *et al.*, 2018). Ce corpus est construit et maintenu à l'université de Cambridge. Il contient des productions d'apprenants adultes d'anglais comme L2 et de différentes langues maternelles L1 (portugais brésilien, chinois, russe, espagnol mexicain, allemand, français, italien, arabe saoudien, taiwanais et japonais). Plusieurs travaux qui cherchent à définir le niveau de maîtrise de L2 ont été effectués dans le cadre de la campagne d'évaluation de la conférence CAp en 2018 [1], avec une mesure d'évaluation spécifique basée sur l'erreur et une matrice de coût spécifique (Ballier *et al.*, 2018). L'erreur est calculée comme $E = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}N_{ij}$, où N est la matrice de confusion (N_{ij} compte le nombre de fois où un exemple de la classe i a été classé j), n le nombre d'exemples classés et C la matrice de coût, calculée comme une mesure d'entropie croisée pondérée entre les classes. Les probabilités prises en compte sont les probabilités d'apparition des classes telles qu'elles apparaissent dans les échantillons d'apprentissage et de test. Les poids des classes et des erreurs ont été donnés par les experts du domaine pour prendre en compte l'importance de chacune des classes. Sur les 14 équipes participantes de cette compétition, nous avons pu trouver la description de deux systèmes seulement. Le système gagnant (Balikas, 2018) met en place une large palette de descripteurs (formules de lisibilité, modèles probabilistes de langues, plongements lexicaux, topic model, étiquettes morpho-syntaxiques, n-grammes de mots). Avec Gradient Boosted trees et tous les descripteurs, le système obtient 98,2 d'*accuracy* avec l'*erreur* de 4,97. Un autre système (Arnold *et al.*, 2018) utilise une palette plus restreinte de descripteuts (diversité lexicale, complexité syntaxique, formules de lisibilité). Gradient Boosted trees est utilisé pour effectuer la classification binaire entre deux niveaux de langues. Selon les paires de niveaux, l'*AUC*, qui est la seule métrique présentée dans ce travail, varie entre 0,587 et 0,916.

Nous travaillons avec les productions langagières d'apprenants d'anglais comme L2 provenant de différents pays et de différentes langues L1. Notre objectif consiste également à prédire le niveau de ces apprenants sur la base de leurs productions écrites. Nous exploitons différents types de descripteurs, y compris les verbes et expressions modaux. Comme les modaux sont des catégories abstraites de la langue, ils peuvent être plus difficiles à acquérir et maîtriser par les apprenants. Selon les observations des chercheurs, lors des étapes initiales d'apprentissage d'anglais L2, l'usage de modaux est pauvre qu'il s'agisse de leur fréquence et diversité, les modaux déontiques sont dominants, certains modaux peuvent être sous-utilisés ou sur-utilisés (Biewer, 2011). Par ailleurs, il existe peu de travaux qui s'intéressent à cette catégorie d'unités linguistiques (Jabbari & Sedghi, 2015) chez les apprenants d'anglais L2. Nous proposons donc d'étudier le rôle que les modaux pourraient jouer dans la détection automatique du niveau de langue, à côté d'autres descripteurs.

2 Approche et données

Nous travaillons avec le corpus EFCAMDAT [2] (Geertzen *et al.*, 2013; Huang *et al.*, 2018). La partie exploitée du corpus contient 27 287 productions contenant presque 2M occurrences de mots. Une production correspond au texte écrit par un apprenant d'anglais pour répondre à une question donnée. Une production peut contenir une ou plusieurs phrases et véhiculer des idées plus ou moins complexes, en fonction de la maîtrise de la langue. La figure 1 présente des exemples de productions.

1. http://cap2018.litislab.fr/competition.html
2. https://corpus.mml.cam.ac.uk/efcamdat2/

C2	*In France, robots are commonly used in industrial fields. They replace humans for heavy duties, dangerous or repetitive tasks. Also, we find more and more robots in houses, like vacums that move by themselves, same idea for lawn mawers. But in my opinion, these kind of robots are making people lazy. I admit that we can save time by using them, but on the other hand, do we really use this spare time doing good stuff ? I am not sure.*
C1	*In general I admire succesfull persons who keep their humility and stay simple. This kind of person are not self oriented, they have all they could wish and they are not selfish. Being focused on the others and attentive to his/her friends, relatives and colleagues give to succesful something more, making them more human and noble.*
B2	*I don't practise any extreme sport. I am definitely not a thrill-seeker. I'm scared of deep water, I have the vertigo... However, I am awestruck by people who pratise this kind of activities. They still search to overtook their limits. I can understand them. When I was younger, I was passionate about pacey carousels. It was fantastic to fly in the air even though I was just in a seat with a harness. I always got in them alone because my friends were so afraid. I could feel such a rush. One time, I tried one carousel who was going very fast. I felt like I was going to pass out. Now I always wimp out.*
B1	*I saw this girl at the swimming pool. I found her very attractive. She had a necklace with her first name. I introduced me and I asked her few questions to know her better. After we met, I invited her for a date the next day. Unfortunately, I was late. She left when I arrived. The next day, I tried my luck again and we had a new date for a tennis party. Of course I was careful to be on time. We played tennis but the next time she left on holiday for a week. We phoned us during this week and I welcomed her for her back at the station. It was 27 years ago. We got married since.*
A2	*Dear friends, excuse me but I couldn't come to a mariage. I feel awful and I think that I'm sick. I have a cold, a headache and a fever. I went to the doctor and I should stay in bed a few days. The pharmacist gave me some medecine. I'm sorry, let' go and have fun.*
A1	*Hi Sue, Sorry, I'm busy. Right now I'm working in my office. Then, I have to clean my house. And after, cook for my parents. See you another day. xoxo.*

FIGURE 1 – Exemples de productions d'apprenants d'anglais pour chaque niveau

Ces productions sont catégorisées selon six niveaux allant de A1 (le plus bas niveau de maîtrise) à C2 (le plus haut niveau de maîtrise) selon CECRL, un cadre européen de référence pour les langues. Le tableau 1 indique le nombre de productions par niveau, le nombre d'occurrences de mots que cela représente et la longueur moyenne des productions pour un niveau donné. Nous pouvons voir que les niveaux C, et surtout le niveau C2, contiennent très peu de productions. La longueur moyenne des productions tous niveaux confondus est de 68 mots. Cependant, la longueur moyenne des productions par niveau montre une tendance à augmenter avec l'amélioration de la maîtrise de l'anglais.

Pour la prédiction du niveau d'anglais, nous exploitons les algorithmes d'apprentissage de la bibliothèque Scikit-learn [3] (Pedregosa *et al.*, 2011). Plusieurs descripteurs sont exploités :

— un ensemble de 59 formules et indicateurs de lisibilité fournis avec le corpus EFCAMDAT. Il s'agit de formules classiques de lisibilité, comme celles proposées dans les travaux existants (Flesch, 1948; McLaughlin, 1969; Kincaid *et al.*, 1975). Leur calcul est basé le plus souvent sur des indicateurs de surface des textes (nombre de mots, nombre de phrases, nombre de

3. https://scikit-learn.org/stable/

Niveau	*# productions*	*# occurrences*	*Moyenne d'occurrences*
A1	11 346	432 442	38
A2	7 680	503 246	66
B1	5 383	511 356	95
B2	2 337	308 433	132
C1	491	80 786	165
C2	50	8 317	166
Total	27 287	1 844 580	68

TABLE 1 – Nombre de productions selon les six niveaux d'anglais et leurs taille

syllabes, longueur moyenne de phrases, taille moyenne de mots, etc.). Ces formules associent les productions avec les niveaux scolaires et universitaires et reflètent la complexité des textes. Les formules de lisibilité fournissent des scores associés aux productions ;

— les n-grammes de 2, 3 et 4 mots (formes) provenant de deux corpus de référence : COCA (Corpus of Contemporary American English) (Davies, 2010) et BNC (British National Corpus) (Burnard, 2000). COCA contient plus de 560M mots et couvre la période entre 1990 et 2017. BNC contient 100M mots provenant de productions écrites et orales produites entre les années 1980 et 1990. La motivation d'utiliser des n-grammes vient du fait que les cooccurrences de mots peuvent également indiquer le niveau de connaissance d'une langue. Par exemple, on dit *commit atrocities* plutôt que *do atrocities* ou *perform atrocities*. Ainsi, plus on maîtrise une langue et plus on a tendance à utiliser des expressions standards, comme par exemple l'usage d'expressions plus ou moins figées, de prépositions ou de temps verbaux. L'utilisation de ces descripteus va dans le même sens que l'exploitation de manuels de langue effectué dans un travail existant (Pilan *et al.*, 2016). Les n-grammes sont extraits de la même manière à partir de corpus et à partir de chaque production, ce qui permet de calculer ensuite les n-grammes communs (en nombre et pourcentage) ;

— un ensemble de 17 verbes modaux (*may, might, can, could, shall, should, will, would, must, have to, got to, need to, be supposed to, had better, be allowed to, be able to*), qui véhiculent les caractéristiques modales principales en anglais. Il s'agit de verbes modaux utilisés le plus souvent par les locuteurs natifs de la langue. Les travaux existants se focalisent le plus souvent sur l'emploi de ces verbes modaux (Gibbs, 1990; Moloi, 1998; Saeed, 2009; Elturki & Salsbury, 2016). Comme déjà indiqué, nous pensons que, comme les valeurs modales sont des notions abstraites, leur maîtrise et utilisation peuvent être indicatives du niveau d'avancement dans l'apprentissage de la langue ;

— un ensemble d'autres expressions modales (verbes, adjectifs, noms, adverbes) qui véhiculent une sémantique similaire, comme par exemple *possible, probably* ou *seem*.

Pour les modaux, nous calculons la fréquence de leur utilisation dans les productions.

Nous exploitons ces ensembles de descripteurs séparément pour voir leur pertinence pour la tâche mais aussi en combinaison car nous pensons que les niveaux de langue correspondent aux catégories complexes et reposent sur différents aspects liés à la maîtrise de la langue.

Notre tâche consiste donc à effectuer une catégorisation multiclasse et à assigner les productions écrites d'apprenants à l'un des six niveaux de CECRL. Nous utilisons plusieurs algorithmes d'apprentissage supervisé avec leurs paramètres par défaut en validation croisée à 10 plis. Les résultats sont évalués avec trois mesures standards : précision P, rappel R et F-mesure F dans leur version macro au niveau des catégories.

3 Détection du niveau d'apprenants

	DT			*RF*			*SVM*		
Descripteurs	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Lisibilité (Lis)	0,63	0,63	0,63	0,67	0,59	0,61	0,67	0,65	0,66
BNC	0,53	0,53	0,53	0,61	0,57	0,59	0,47	0,43	0,44
COCA	0,53	0,54	0,54	0,78	0,58	0,59	0,49	0,44	0,45
17 modaux	0,35	0,28	0,28	0,35	0,29	0,29	0,32	0,26	0,25
Autres modaux	0,25	0,19	0,15	0,24	0,19	0,15	0,20	0,19	0,14
Lis+17 modaux	0,63	0,64	0,64	0,63	0,59	0,60	0,69	0,67	0,68
Lis+autres modaux	0,64	0,63	0,63	0,62	0,58	0,59	0,68	0,66	0,67
Lis+tous les modaux	0,63	0,63	0,63	0,62	0,57	0,59	0,70	0,67	0,69
Tous+BNC	0,71	0,71	0,71	0,81	0,66	0,69	**0,74**	**0,70**	**0,72**
Tous+COCA	0,71	0,71	0,71	0,86	0,66	0,69	0,73	0,69	0,71

TABLE 2 – Résultats de catégorisation selon les descripteurs exploités (version macro des mesures)

Les résultats globaux de quelques expériences obtenus avec trois algorithmes (arbres de décision DT, RandomForest RF et SVM linéaire SVM) sont présentés dans le tableau 2. Nous avons plusieurs ensembles de descripteurs : différents descripteurs utilisés séparément (*lisibilité, BNC, COCA, 17 modaux, autres modaux*) et leurs combinaisons (formules de lisibilité avec 3 ensembles de modaux (*17 modaux, autres modaux* et *tous les modaux*) et la combinaison de tous les descripteurs (*tous+BNC* et *tous+COCA*)). Nous voyons que tous les algorithmes montrent de meilleurs résultats avec la combinaison de tous les descripteurs. SVM se détache des autres algorithmes lorsqu'il est exploité avec les scores de lisibilité et les combinaisons de descripteurs, alors que DT et RF montrent aussi de bons résultats avec les n-grammes de mots. Les 17 modaux principaux permettent de catégoriser correctement un peu moins d'un tiers des productions (F-mesure entre 0,25 et 0,29). Les autres expressions modales, sans doute parce qu'elles sont utilisées moins fréquemment par les apprenants, montrent les performances les plus faibles (F-mesure entre 0,14 et 0,15). L'exploitation de modaux avec les formules de lisibilité améliore la F-mesure obtenue avec les formules de lisibilité seules de 0,20 points avec DT et de 0,25 points avec SVM. Cela indique donc que les modaux apportent des informations importantes sur le niveau de maîtrise de la langue.

Le tableau 3 indique les résultats par niveau de langue obtenues avec SVM et tous les descripteurs combinés avec les n-grammes BNC : 0,72 de F-mesure macro (pour information, cela correspond à 0,82 de F-mesure micro). Le niveau A1 montre les résultats les plus élevés (F-mesure de 0,89), ce qui peut être dû aux facteurs quantitatifs (grand nombre de productions) et qualitatifs (les apprenants ont

Niveau	*P*	*R*	*F*
A1	0,89	0,90	0,89
A2	0,76	0,76	0,76
B1	0,78	0,79	0,79
B2	0,79	0,75	0,77
C1	0,72	0,66	0,69
C2	0,47	0,36	0,40

TABLE 3 – Catégorisation par niveau (SVM et tous les descripteurs avec les n-grammes de BNC)

Ref/Predit	*A1*	*A2*	*B1*	*B2*	*C1*	*C2*	
A1	**10 154**	1 045	131	13	3	0	11 346
A2	1 238	**5 827**	576	36	1	2	7 680
B1	80	713	**4 266**	307	17	0	5 383
B2	12	31	438	**1 763**	85	8	2 337
C1	4	17	29	106	**325**	10	491
C2	1	2	2	8	19	**18**	50
tous	11 489	7 635	5 442	2 233	450	38	27 287

TABLE 4 – Matrice de confusion (SVM et tous les descripteurs avec les n-grammes de BNC)

des productions très éloignées de la langue standard et des productions d'autres niveaux plus avancés). Le niveau C2 montre les performances les plus pauvres, sans doute à cause du très faible nombre de productions. Le niveau C1 a une F-mesure de 0,69. Les trois autres niveaux (A2, B1 et B2) montrent une F-mesure entre 0,76 et 0,79. Le tableau 4 présente la matrice de confusion de la même expérience. Nous voyons en diagonal le nombre de productions associées correctement aux niveaux de langue.

Il est difficile de comparer nos résultats avec le système de Arnold *et al.* (2018) car nous n'effectuons pas le même type de catégorisation (bi-classe *vs* multiclasse). Nos résultats (valeurs macro) sont inférieurs à ceux de (Balikas, 2018) (98,2 d'*accuracy* mais il n'est pas indiqué s'il s'agit de valeurs micro ou macro). Ce travail utilise les descripteurs différents des nôtres. Nous préférons cependant exploiter les descripteurs observés directement dans les productions d'apprenants (n-grammes, modaux...) car nous pensons que les modèles probabilistes ou inductifs, comme les plongements lexicaux, ne reflettent pas forcément la compétence d'un apprenant donné. À notre avis, ces modèles correspondent à la performance collective des apprenants. Par exemple, les plongements lexicaux peuvent grouper ensemble les verbes comme *commit, perform* et *do*. Cependant, si un apprenant n'utilise pas le bon verbe dans une expression comme *commit atrocities*, les n-grammes de mots sont plus susceptibles de refléter correctement son niveau de langue que les plongements lexicaux.

4 Conclusion

Nous avons présenté quelques expériences de prédiction du niveau de langue des apprenants d'anglais sur la base de leurs productions écrites. Nous exploitons pour ceci plusieurs ensembles de descripteurs : formules de lisibilité classiques, n-grammes de mots et expressions et verbes modaux. La catégorisation montre des résultats intéressants et souligne l'importance des modaux pour cette tâche. Ces résultats peuvent être améliorés par d'autres expériences (ajout d'autres descripteurs et de leurs combinaisons, exploitation d'autres algorithmes). Par ailleurs, le rôle de notions plus abstraites, comme les valeurs modales, pourra être étudié encore plus en détail dans les travaux futurs.

Remerciements

Cette publication s'inscrit dans le projet *REM (Re-thinking English Modal Constructions)* financé par l'ANR franco-suisse sous la référence ANR-16-CE93-0009. Nous remercions les relecteurs pour leurs remarques constructives.

Références

ARNOLD T., BALLIER N., GAILLAT T. & LISSÓN P. (2018). Predicting CEFRL levels in learner english on the basis of metrics and full texts. In *Conférence sur l'Apprentissage Automatique (CAp)*, p. 31–38.

AYOUN D. & GILBERT C. (2017). *The acquisition of modal auxiliaries in English by advanced Francophone learners*, In M. HOWARD & P. LECLERCQ, Éds., *Tense-Aspect-Modality in a Second Language : Contemporary perspectives*, p. 183–212.

BALIKAS G. (2018). Lexical bias in essay level prediction. In *CAp*, p. 1–5.

BALLIER N., CANU S., GAILLAT T., GASSO G., PETITJEAN C. & RAKOTOMAMONJY A. (2018). *Appel à participation à la compétition « my taylor is rich » de CAp 2018. Prédiction du niveau en anglais à partir de production écrite d'apprenants*. Rapport interne, CAP 2018.

BIEWER C. (2011). *Modal auxiliaries in second language varieties of English : A learner's perspective*, In J. MUKHERJEE & M. HUNDT, Éds., *Exploring second-language varieties of English and learner Englishes : Bridging a paradigm gap*, p. 7–33. John Benjamins : Amsterdam.

BURNARD L. (2000). *The British National Corpus Users Reference Guide*. Rapport interne, Oxfort university. http://www.natcorp.ox.ac.uk/docs/userManual/.

CHENU F. & JISA H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development. *Acquisition et interaction en langue étrangère*, (1), 1–22.

COURNANE A. (2015). *Modal development : Input-divergent L1 acquisition in the direction of diachronic reanalysis*. Thèse de doctorat, University of Toronto, Toronto, Canada.

DAVIES M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, **25**(4), 447–65.

ELTURKI E. & SALSBURY T. (2016). A cross-sectional investigation of the development of modality in English language learners' written narratives : A corpus-driven study. *Issues in Applied Linguistics*, **20**(1), 51–72.

FLESCH R. (1948). A new readability yardstick. *Journ Appl Psychol*, **23**, 221–233.

GEERTZEN J., ALEXOPOULOU T. & KORHONEN A. (2013). Automatic linguistic annotation of large scale L2 databases : The EF-Cambridge open language database (EFCAMDAT). In *31st Second Language Research Forum (SLRF)*.

GIBBS D. A. (1990). Second language acquisition of the English modal auxiliaries can, could, may, and might. *Applied Linguistics*, **11**(3), 297–314.

GRANFELDT J. & NUGUES P. (2007). Évaluation des stades de développement en français langue étrangère. In *TALN*, p. 1–10.

HUANG Y., MURAKAMI A., ALEXOPOULOU T. & KORHONEN A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, **23**(1), 28–54.

IPEK H. (2009). Comparing and contrasting first and second language acquisition : Implications for language teachers. *English Language Teaching*, **2**(2), 155–163.

JABBARI A. & SEDGHI M. (2015). Acquisition of English modality by Persian EFL learners. *International Journal of Educational Investigations*, **2**(5), 23–45.

JIANG X., GUO Y., GEERTZEN J., DORA ALEXOPOULOU AND L. S. & KORHONEN A. (2014). Native language identification using large, longitudinal data. In *LREC*, p. 1–4.

KINCAID J., FISHBURNE R., ROGERS R. & CHISSOM B. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Rapport interne, Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

LAUFER B. & ELIASSON S. (1993). What causes avoidance in L2 learning. L1-L2 difference, L1-L2 similarity or L2 complexity ? *Studies in Second Language Acquisition*, (15), 35–48.

MALMASI S. & DRAS M. (2015). Large-scale native language identification with cross-corpus evaluation. In *Annual Conference of the North American Chapter of the ACL*, p. 1403–1409.

MCLAUGHLIN G. H. (1969). SMOG grading – a new readability formula. *Journal of reading*, **12**(8), 639–646.

MOLOI F. (1998). Acquisition of modal auxiliaries in English L2. *Southern African Journal of Applied Language Studies*, **6**(2), 1–22.

MORTELMANS T. & ANTHONISSEN L. (2016). *German modals in second language acquisition : A constructionist approach*, In A. STEFANOWITSCH & T. HERBST, Éds., *Yearbook of the German Cognitive Linguistics Association*, p. 9–30.

MURAKAMI A., MICHEL M., ALEXOPOULOU T. & MEURERS D. (2016). Analyzing learner language in task contexts : A study case of linguistic complexity and accuracy in EFCAMDAT. In *European Second Language Association Conference*.

NISIOI S. (2015). Feature analysis for native language identification. In *CICLING*, p. 1–15.

PAPAFRAGOU A. (1998). The acquisition of modality : Implications for theories of semantic representation. *Mind and Language*, **13**(3), 370–399. DOI : 10.1111/1468-0017.00082.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PILAN I., VOLODINA E. & ZESCH T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Int Conf on Computational Linguistics*, p. 2101–2111.

RABINOVICH E., NISIOI S., ORDAN N. & WINTNER S. (2016). On the similarities between native, non-native and translated texts. In *Annual Meeting of the Association for Computational Linguistics*, p. 1870–1881.

ROBERTSON K. & FORD K. (2009). *Language Acquisition : An Overview*. Rapport interne, Colorin Colorado.

RÖMER U. (2019). A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, **24**(3), 270–292.

SAEED A. T. (2009). Arab EFL learners' acquisition of modals. *Research in Language*, **7**, 75–98.

SHATZ M. & WILCOX S. (1991). *Constraints on the acquisition of English modals*, In S. GELMAN & J. BYRNES, Éds., *Perspectives on language and thought*, p. 319–353. Cambridge University Press : Cambridge.

WATANABE S. & IWASAKI N. (2009). *The Acquisition of Japanese Modality during Study Abroad*, In B. PIZZICONI & M. KIZU, Éds., *Japanese Modality*, p. 231–258.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

TArC
Un corpus d'*arabish* tunisien

Elisa Gugliotta[1,2] Marco Dinarelli[1]

(1) LIG - Bâtiment IMAG, 700 Avenue Centrale 38401 Saint-Martin-d'Hères, Grenoble, France

(2) Université Sapienza di Roma, 82 Viale dello Scalo S. Lorenzo 00159, Roma, Italia

elisa.gugliotta@uniroma1.it, marco.dinarelli@univ-grenoble-alpes.fr

Résumé

Cet article décrit la procédure de constitution du premier corpus d'*arabish* tunisien (TArC) annoté avec des informations morpho-syntaxiques. L'*arabish* est la transcription spontanée des dialectes arabes en caractères latins et *arythmographies*, c'est à dire avec des chiffres utilisées comme lettres. Ce système d'encodage a été développé par les utilisateurs arabes des réseaux sociaux afin de faciliter l'écriture dans les communications informelles. L'*arabish* diffère pour chaque dialecte arabe et il est sous-doté en termes de ressources, de la même façon que la plupart des dialectes arabes. Dans les dernières années, l'attention des travaux de recherche en TAL sur les dialectes arabes est augmentée de façon remarquable. En prenant ceci en compte, *TArC* serait un support utile pour plusieurs types d'analyses, computationnelles ainsi que linguistiques, et pour l'apprentissage d'outils informatiques. Nous décrivons le travail fait pour mettre en place une procédure d'acquisition semi-automatique du corpus TArC, ainsi que certaines analyses faites sur les données collectées. Afin de montrer les difficultés rencontrées pendant la procédure de constitution du corpus, nous présentons également les caractéristiques principales du dialecte tunisien, ainsi que sa transcription en *arabish*.

Abstract

***TArC* : Incrementally and Semi-Automatically Collecting a Tunisian arabish Corpus**

This article describes the collection process of the first morpho-syntactically annotated Tunisian arabish Corpus (TArC). Arabish is a spontaneous coding of Arabic Dialects (AD) in Latin characters and *arithmographs* (numbers used as letters). This code-system was developed by Arabic-speaking users of social media in order to facilitate the communication on digital devices. Arabish differs for each Arabic dialect and each arabish code-system is under-resourced. In the last few years, the attention of NLP on AD has considerably increased. TArC will be thus a useful support for different types of analyses, as well as for NLP tools training. In this article we will describe preliminary work on the TArC semi-automatic construction process and some of the first analyses on the corpus. In order to provide a complete overview of the challenges faced during the building process, we will present the main Tunisian dialect characteristics and its encoding in Tunisian arabish.

Mots-clés : Corpus d'arabish tunisien, Dialecte arabe, Arabizi.

Keywords : Tunisian arabish Corpus, Arabic Dialect, Arabizi.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 232–240.
HAL : hal-02784772.

1 Introduction

L'*arabish*[1] est la romanisation des dialectes arabes (DAs) utilisée pour des messages écrits informels, particulièrement dans les réseaux sociaux. Ce système d'écriture fournit un terrain très intéressant pour des recherches linguistiques, computationnelles, mais aussi socio-linguistiques, principalement grâce au fait qu'il s'agit d'une représentation écrite spontanée des DAs, et il est en constante expansion dans le web.[2] Malgré ce potentiel, peu de recherches de ce type on été dédiées à l'arabish tunisien (TA). Dans cet article nous décrivons notre travail pour le développement d'une ressource en TA riche, avec plusieurs niveaux d'annotation. Celle-ci inclut un corpus en TA annoté, mais aussi des outils informatiques pour à la fois analyser les données et étendre éventuellement le corpus avec des nouvelles données. En premier lieu, la ressource permet de donner une vue générale du TA. Au même temps elle constitue une représentation significative de l'évolution du dialecte tunisien (TUN) dans le cadre de la communication numérique (CMC, de l'anglais *Computer-Mediated Communication*) à travers les dix dernières années. En effet, les textes collectés datent de 2010 jusqu'à l'année courante.[3] Pour cette raison, notre corpus se prête à des études phonologiques, morphologiques, syntaxiques et aussi sémantiques, et à la fois dans un contexte de linguistique et de TAL. Nous avons donc décidé de construire un corpus qui peut mettre en évidence les caractéristiques structurelles du TA grâce à plusieurs niveaux d'annotation, comprenant parties du discours (POS) et lemmes. De plus, afin de faciliter la correspondance avec d'autres études, ainsi que des outils déjà existants pour le traitement du langage arabe, nous avons transcrit les tokens en arabish en caractères arabes du dialecte tunisien, suivant les lignes directrices du *Conventional Orthography for Dialectal Arabic* (CODA*) (Habash *et al.*, 2018).[4] Enfin, même si la traduction de notre corpus n'est pas l'objectif principal de nos recherches, nous avons décidé de traduire les textes de notre corpus en italien.[5]

Le reste de l'article est organisé comme suit : dans la section 2 nous décrivons des études sur l'arabish tunisien (TA) et les corpus accessibles du dialecte tunisien (TUN) et TA ; la section 3 décrit le TA ; dans la section 4 nous présentons la procédure de construction du corpus *TArC* ; la section 5 montre des expériences préliminaires pour la transcription et l'annotation semi-automatiques des données, adoptés pour une construction plus facile et rapide du corpus ; nous tirons nos conclusions dans la section 6.

2 État de l'art

Le travail décrit dans cet article concerne surtout la transcription semi-automatique du corpus en caractères arabes. En revanche pour la réalisation d'un des niveaux d'annotation du corpus, nous décrivons ici quelques travaux représentatifs de l'état de l'art sur le codage de l'arabish tunisien. Nous décrivons également les corpus tunisiens accessibles librement. Pendant ces dernières années,

[1] Aussi appelé *Arabizi* (de la combinaison des mots arabes 'Arab', ['ʕarab] et 'English' [i ːn'ʒli ːzi ː], Franco-Arabe, *Arabic Chat Alphabet*, ACII-ized Arabic, entre autres. 'arabish' est probablement le résultat de l'union entre ['ʕarab] et 'English'. Ce mot a été choisi pour être lisible par tout le monde. Le terme 'arabizi' ne serait en fait compréhensible que par ceux qui connaissent la langue arabe.

[2] Pour des informations sur le pourcentage d'utilisation, veuillez vous référer à (Tobaili, 2016 ; Younes *et al.*, 2015).

[3] Cette sélection vise à observer le degré d'évolution diachronique d'une convention d'écriture arabish.

[4] Le CODA* est un ensemble unifié de lignes directrices pour la transcription de 28 dialectes arabes.

[5] Nous avons pris en considération une traduction à moindre coût, donc vers notre langue maternelle, laissant une traduction vers l'anglais comme éventuelle option future.

le nombre de travaux sur l'arabish a considérablement augmenté.[6] La plupart des recherches sur l'arabish se focalisent sur des variétés très différentes de l'arabe tunisien, par exemple l'égyptien (Darwish, 2014 ; Al-Badrashiny *et al.*, 2014). À cause d'un manque de place, nous nous focaliserons uniquement sur les travaux sur l'arabish tunisien. (Younes *et al.*, 2018b) est le seul travail qui propose une transcription automatique de l'arabish tunisien vers les caractères arabes et vice versa, basée sur des modèles d'apprentissage profond dits *sequence-to-sequence*. En transcrivant une collection de textes de Facebook, après l'avoir nettoyé des éléments non linguistiques (émojis, émoticônes, etc.) ou qui n'étaient pas dialectales, ils ont obtenu une précision de plus de 95.59%.[7] La même ressource a été utilisée dans (Younes *et al.*, 2020) pour une étude approfondie de la transcription automatique. En utilisant uniquement les mots tunisiens, une transcription en caractères arabes a été faites par des internautes tunisiens. La transcription d'un mot arabish donné est effectuée en utilisant une approche à base de modèles CRF, BLSTM et BLSTM-CRF, ce dernier donnant les meilleurs résultats quantitatifs. Le taux d'erreur global est de 2,7%, alors que le taux d'erreur de contexte est de 0,68%. Le seul travail qui prévoit une transcription de l'arabish tunisien dans la convention CODA est (Masmoudi *et al.*, 2015). Ce travail utilise un système à base de règles qui, pour chaque mot en arabish, génère toutes les transcriptions possibles en arabe, parmi lesquelles le meilleur candidat est ensuite sélectionné manuellement. Le système a un taux d'erreur de 10%. Cependant, à l'exception du recueil de textes de (Younes *et al.*, 2015), aucune de ces ressources n'est accessible au public. Les corpus en dialecte tunisien disponible au public sont cinq (Younes *et al.*, 2018a). Le *PADIC* (Meftouh *et al.*, 2015), composé de 6 400 phrases en six dialectes arabes, traduites en arabe moderne standard (MSA).[8] Deux autres corpus sont le *TuDiCoI* (*Tunisian Dialect Corpus Interlocutor*) (Graja *et al.*, 2010) et le *STAC* (*Spoken Tunisian Arabic Corpus*) (Zribi *et al.*, 2015), qui sont annotés avec des informations morpho-syntaxiques. Le premier est un corpus de dialogues entre les clients et le personnel collectés à la station des trains. Il est composé de 21 682 mots (Graja *et al.*, 2013).[9] Le corpus STAC est constitué de 42 388 mots transcrits de fichiers audio téléchargé du web (émissions télé et radio) (Zribi *et al.*, 2015). Le *TARIC* (Masmoudi *et al.*, 2014a) contient 20 heures de TUN oral, transcrites en caractères arabes et correspondant à 71 684 mots (Masmoudi *et al.*, 2014b). Un dernier corpus est le *TSAC*, composé de 17 000 commentaires de Facebook, annotés manuellement en polarité positive et négative pour la fouille d'opinion (Medhaffar *et al.*, 2017). Ce dernier corpus est le seul qui contient à la fois des textes transcrits en TA et en caractères arabes. À notre connaissance il n'y pas de corpus en TA transcrit aussi en caractères arabes et annoté avec des informations morpho-syntaxiques. Notre travail se propose de combler ce manque en offrant à la communauté scientifique une ressource d'arabish tunisien ouverte au public, c'est-à-dire le *TArC*, le premier corpus en TA annoté avec des informations morpho-syntaxiques (POS et lemmes) et transcrit aussi en caractères arabes en suivant la convention CODA*. Compte tenu également de l'indisponibilité des systèmes de transcription de l'arabish tunisien,[10] il est intéressant de construire un système *ad hoc* en fonction de nos besoins. Nous avons fait le choix, en effet, de ne pas exclure les termes étrangers ou les éléments para-linguistiques (émoticônes et emojis), et d'utiliser les conventions CODA.

[6] Pour une description détaillée veuillez consulter (Guellil *et al.*, 2019).

[7] Le corpus utilisé est disponible sous requête.

[8] Les dialectes arabes dans le corpus PADIC sont : le TUN (Sfax), deux dialectes de l'Algérie, le syrien, le palestinien, et le marocain (Meftouh *et al.*, 2018).

[9] L'annotation en revanche a été faite uniquement pour 7 814 mots.

[10] Pour la distinction entre transcription et translittération veuillez vous référer à (Coulmas, 2003).

3 Caractéristiques de l'*arabish* tunisien

Le dialecte tunisien (TUN) est la langue parlé dans la vie tunisienne de tous les jours, appelé généralement الدَّارِجَة, *ad-dārija*, العَامِّيَّة, *'āmmiyya*, or التُّونْسِي, *ət-tūnsī*. Conformément à la classification diatopique traditionnelle, le TUN appartient à la zone de l'arabe maghrébin, duquel il constitue une des variantes principales avec le libyen, l'algérien, le maroquin, et le hassanya de la Mauritanie (Durand, 2009). L'arabish est une transposition des DAs, qui sont des systèmes essentiellement oraux, dans une forme écrite qui n'est pas réalisé avec caractères arabes et par conséquence il n'est pas sujet aux règles orthographiques de l'arabe standard. Comme résultat, on peut considérer le TA comme un écrit spontané, fidèle à la réalisation orale de TUN, ou en d'autres termes un système quasi-oral.

La notion de *quasi-oralité* décrit des formes d'écriture typiques de la communication numérique (CMC), caractérisée par un ton informel, par la dépendance du contexte, par le manque d'attention sur la façon d'écrire, et ayant spécialement la capacité de créer un sens de collectivité (Hert, 1999). TA et TUN n'ont pas une orthographie standard, avec l'exception du CODA. Cependant, le TA est un système d'écriture utilisé depuis plus que dix ans, et il subit donc une conventionnalisation spontanée à travers son utilisation. Dans le tableau 1 nous montrons un schéma du système d'écriture TA. Il est possible d'observer qu'il n'y a pas une correspondance un à un entre les caractères en TA et ceux en arabes, et souvent le TA présente une ambiguïté dans la possibilité d'écriture.[11] Le problème principal est constitué par le manque d'une représentation propre en TA pour les phonèmes emphatiques : [ðˤ], [tˤ] et [sˤ]. D'un autre côté, puisque le TA n'est pas codifié à travers l'alphabet arabe, il peut

1	[ðˤ]	[a ː]	[ʔ]	[b]	[θ]	[ʒ]	[ħ]	[x]	[d]	[ð]	[s]	[ʃ]	[sˤ]
2	ض	ة	ء	ب	ث	ج	ح	خ	د	ذ	س	ش	ص
3	dh th d	a e h	2	b p	th	j	7 h	5 kh	d	dh	s	ch, (sh)	s
1	[a][a ː]	[tˤ]	[ðˤ]	[ʕ]	[ɣ]	[q]	[k]	[l]	[m]	[n]	[h]	[w][u ː]	[j][i ː]
2	ا ى	ط	ظ	ع	غ	ق	ك	ل	م	ن	ه	و	ي
3	a e é è	6 t	th dh	3 a	4 gh	9 q	k	l	m	n	8, h	ou, w	i, y

Tab. 1: Exemples de correspondance entre TA et TUN. **1** indique la représentation phonétique des graphèmes. **2** représente les caractères arabe, **3** les caractères arabish correspondant.

bien représenter la réalisation phonétique du TUN, comme il est montré dans l'exemple suivant : **1.** L'alphabet arabe est généralement utilisé pour des conversations formelles en arabe moderne standard (MSA), l'arabe des contextes formels, ou pour l'arabe classique (CA), l'arabe du *Saint Qur'ān*. De la même façon que le MSA et le CA, les dialectes arabes également peuvent être écrits avec l'alphabet arabe, mais dans ce cas il est possible d'observer une auto-correction spontanée de l'orateur pour respecter les règles d'écriture du MSA. Par exemple, dans les textes en TUN écrits avec l'alphabet arabe, il est possible de trouver une voyelle muette (ʾalif <ا> épenthétique, ou additionnel) au début des mots qui commencent par la séquence '#CCv', ce qui n'est pas possible en MSA. **2.** En écrivant le TUN avec l'alphabet arabe, l'écriture de mots étrangers dans leur alphabet est très forcée, comme, par exemple, dans l'usage de mots empruntés d'autres langues. **3.** Comme nous le montrons dans le tableau 1, l'alphabet arabe fournit trois voyelles courtes, qui correspondent à leur version longue [a ː], [u ː], [i ː], mais le TUN présente un ensemble plus étendu de voyelles. En effet, l'ensemble des voyelles du TA offre une meilleure possibilité de représenter la phonétique du TUN.

[11] Pour ces raisons, la conversion de TA à TUN ne peut être traitée comme une simple translittération.

3.1 Collecte des données - trois étapes

1. Détection des catégories thématiques. Afin de construire un corpus qui soit représentatif du TUN, il nous semblait utile d'identifier des catégories thématiques larges, qui pourraient représenter les sujets de discussion plus courant dans les CMC. Dans cette perspective, nous avons employé deux instruments avec une organisation thématique similaire : Un dictionnaire arabe avec fréquence des mots (Buckwalter & Parkinson, 2014)[12] et la *Loanword Typology Meaning List* (LTML) (Haspelmath & Tadmor, 2009), une liste de 1 460 sens. 15 catégories ont été identifiées grâce à ces documents. Pour une description détaillée nous renvoyons à (Gugliotta & Dinarelli, 2020).

2. Construction des correspondances entre les catégories et les mots clefs du TA en relation sémantique. Nous avons associé à chaque catégorie un ensemble de mots-clefs en TA, appartenant au vocabulaire de base tunisien. Nous avons trouvé que trois sens pour chaque catégorie sémantique étaient suffisants pour obtenir un nombre significatif de mots-clefs pour chaque catégorie. Pour une analyse plus détaillée de cette procédure nous renvoyons à (Gugliotta & Dinarelli, 2020).

3. Extraction des textes et des méta-données. À travers ces mots-clés, nous avons effectué une recherche des textes sur les réseaux sociaux. Nous avons collecté des textes pour l'équivalent d'environ 40 000 mots, et leur méta-données associées, comme première partie de notre corpus.[13] Concernant les méta-données, nous avons extrait les informations publiées par les utilisateurs, en nous focalisant sur trois types d'information généralement utilisées dans les études ethnographiques : genre, tranche d'âge et ville d'origine.

4 Constitution du corpus *TArC*

Pour créer notre corpus, nous avons appliqué une annotation au niveau des mots. Cette phase a été précédée de quelques étapes de pré-traitement des données, en particulier la tokenisation. Chaque token a été associé à ses annotations et métadonnées (tableau 2). Afin d'obtenir la correspondance entre les transcriptions de morphèmes arabes et arabish, les tokens ont été segmentés en morphèmes. Cette segmentation a été effectuée manuellement pour un premier groupe de tokens.[14] Dans sa version finale, chaque token est associé à 11 annotations différentes, correspondant au nombre de niveaux d'annotation que nous avons choisi. Un extrait du corpus avec annotation est présenté dans le tableau 2.

Comme le TA est une écriture spontanée du TUN, nous avons jugé important d'adopter les directives CODA* comme modèle pour produire des lemmes et une transcription unifiées pour chaque token (colonnes *Lem* et *Tra* dans la tableau 2). Afin de garantir une transcription et une lemmatisation précises, nous avons annoté manuellement les 6 000 premiers tokens avec tous les niveaux d'annotation. Concernant les mots étrangers, nous avons transcrit les mots arabish en caractères arabes, à l'exception des termes étrangers. En fait, ces mots seront analysés dans un second temps, en faisant la distinction entre les mots étrangers intégrés et le mélange de langues. Les premiers seront transcrits en caractères arabes, les autres seront lemmatisés dans leur langue étrangère. Cette identification sera également utile pour la construction d'un analyseur morphologique qui, après une phase d'identifica-

[12] En particulier a été utilisé son vocabulaire thématique (TVL, de l'anglais *Thematic Vocabulary List*).

[13] Nous avons planifié d'augmenter la taille du corpus dans un second temps.

[14] L'arabe, en général, est une langue à haut niveau de synthèse, ce qui signifie qu'elle peut concentrer dans un token plusieurs informations grammaticales grâce à l'ajout de différents morphèmes.

A	B	C	D	E	F	G	H	I	J	K	L
Cor	Textco	Par	W	Arabiʃ	Tra	Ita	Lem	POS	Var	Age	Gen
3fE	150902	2	1	kifech	كيفاش	come	كيفاش	adv	Bnz	25-35	M
3fE	150902	2	2	tchou-fou	تشوفوا	vi pare	شاف	verb	Bnz	25-35	M
3fE	150902	2	3-4	l3icha	العيشة	la vita	عيشة	noun	Bnz	25-35	M
3fE	150902	2	3	l	ال	-	ال	det	Bnz	25-35	M
3fE	150902	2	4	3icha	عيشة	-	عيشة	noun	Bnz	25-35	M
3fE	150902	2	5-6	fil	فال	all'	في	prep	Bnz	25-35	M
3fE	150902	2	5	f	ف	-	في	prep	Bnz	25-35	M
3fE	150902	2	6	il	ال	-	ال	det	Bnz	25-35	M
3fE	150902	2	7	4orba	غربة	estero	غربة	noun	Bnz	25-35	M
3fE	150902	2	8	?	؟	?	؟	pct	Bnz	25-35	M

Tab. 2: Un extrait de la structure du corpus TArC. Parmi les colonnes plus significatives, la colonne E (*Arabiʃ*) correspond au token en arabish. La colonne F (*Tra*) est la transcription en caractères arabes. La colonne G (*Ita*) est la traduction en italien. La colonne H (*Lem*) est le lemme. La colonne I le *POS*, etc. Pour plus de détails voir (Gugliotta & Dinarelli, 2020).

tion des mots à transcrire et du *code-switching*, contribuera à la tâche de transcription en elle-même. Ce travail est toujours en cours.

5 Procédure incrémentale et semi-automatique de constitution du corpus

Afin de rendre la collecte du corpus plus facile et plus rapide, nous avons adopté une procédure semi-automatique basée sur des modèles neuronaux séquentiels (Dinarelli & Grobol, 2019b,a). Puisqu'il est plus facile d'obtenir automatiquement certaines annotations une fois que les tokens arabish sont transcrits en caractères arabes, puisque la transcription de l'arabish en arabe est une information très importante pour étudier le système arabish, et puisque elle est aussi la plus coûteuse, la procédure semi-automatique ne concerne que la transcription de l'arabish en écriture arabe.[15] Pour cela, nous avons utilisé le premier groupe de 6 000 tokens transcrits manuellement comme ensemble de données d'entraînement et de test dans un cadre de validation croisée. Comme nous l'avons expliqué dans la section précédente, les tokens français ont été retirés des données puisque ils créent du bruit pour un modèle automatique et probabiliste basé sur l'orthographie arabish. Après l'élimination des tokens français, les données ont été réduites à environ 5 000 tokens. Nous constatons qu'en combinant l'index de la phrase, du paragraphe et l'index du token dans le corpus, des phrases entières, voire des paragraphes, peuvent être reconstruites. Cependant, à partir des 5 000 tokens seulement 300 séquences ont pu être reconstruites, ce qui n'est pas suffisant pour l'apprentissage d'un modèle neuronal.[16] Au lieu de cela, puisque les tokens sont transcrits au niveau des morphèmes, nous avons divisé les tokens arabish en caractères, et les tokens arabes en morphèmes, et nous avons traité chaque token comme

[15] En réalité nous sommes en train de développer des systèmes équivalant pour l'annotation en *POS* et pour la lemmatisation. Ce travail est en cours.

[16] Les expériences préliminaires ont donné des mauvais résultats : 50% de précision.

une séquence. Notre modèle apprend donc à transcrire les caractères arabish en morphèmes arabes. Dans ces conditions la validation croisée a donné une précision moyenne d'environ 65%. Ce résultat n'est pas satisfaisant dans l'absolu, mais il est plus qu'encourageant compte tenu de la petite taille de nos données. Ce résultat signifie que moins de 4 tokens, en moyenne, sur 10 doivent être corrigés manuellement. Avec ce modèle, nous avons automatiquement transcrit en morphèmes arabes environ 700 tokens supplémentaires. Ces tokens ont été corrigés manuellement et ont été ajoutés aux données d'entraînement de notre modèle neuronal, et une nouvelle validation croisée a été effectuée. Le résultat a été maintenant d'environ 70% en moyenne. Cette procédure a été ré-itéré 3 fois au total, pour transcrire 4 bloques de tokens sur 5. La précision moyennes sur le quatrième bloque a été d'environ 76%.[17]

6 Conclusions

Dans ce document, nous avons présenté TArC, le premier corpus d'arabish tunisien annoté avec des informations morpho-syntaxiques. Concernant la procédure de construction, nous avons décrit les étapes de constitution et notre effort visant à rendre le corpus le plus représentatif possible du TA et du TUN. Nous avons décrit l'étape de collecte des textes, ainsi que la construction du corpus et la procédure semi-automatique adoptée pour transcrire le TA en écriture arabe, en tenant compte des directives CODA*. Au stade actuel de la recherche, le TArC est constitué de 40 000 tokens, une partie desquels a été transcrite et annoté manuellement, le reste est en cours de transcription semi-automatique, qui a déjà montré des résultats encourageants avec une précision de transcription de 76% en moyenne.

Références

Al-Badrashiny M., Eskander R., Habash N. & Rambow O. (2014). Automatic transliteration of romanized dialectal arabic. In *Proceedings of the eighteenth conference on computational natural language learning*, p. 30–38.

Buckwalter T. & Parkinson D. (2014). *A frequency dictionary of Arabic : Core vocabulary for learners*. Routledge.

Coulmas F. (2003). *Writing systems : An introduction to their linguistic analysis*. Cambridge University Press.

Darwish K. (2014). Arabizi detection and conversion to Arabic. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, p. 217–224. DOI : 10.3115/v1/W14-3629.

Dinarelli M. & Grobol L. (2019a). Hybrid neural models for sequence modelling : The best of three worlds. *CoRR*. arXiv preprint 1909.07102.

Dinarelli M. & Grobol L. (2019b). Seq2biseq : Bidirectional output-wise recurrent neural networks for sequence modelling. *CoRR*. arXiv preprint 1904.04733.

Durand O. (2009). *Dialettologia araba*. 'Sapienza' University of Rome, 'Studi Orientali' Faculty.

[17] Au moment de la soumission finale, 2 bloques de tokens additionnels ont été téléchargés et ajoutés au corpus. Ces bloques sont donc à transcrire et à corriger avec le bloque 5. Nous sommes en train de développer un système d'apprentissage multi-tâche pour effectuer tous les niveaux d'annotations avec un seul modèle.

Graja M., Jaoua M. & Hadrich-Belguith L. (2010). Tunisian dialect corpus interlocutor (tudicoi). In *Arabic Natural Language Processing Research Group (ANLP)* : MIRACL Laboratory, Sfax (Tunis).

Graja M., Jaoua M. & Hadrich-Belguith L. (2013). Discriminative framework for spoken tunisian dialect understanding. In *International Conference on Statistical Language and Speech Processing*, p. 102–110: Springer.

Guellil I., Saâdane H., Azouaou F., Gueni B. & Nouvel D. (2019). Arabic natural language processing : an overview. *Journal of King Saud University-Computer and Information Sciences*. doi : 10.1016/j.jksuci.2019.02.006.

Gugliotta E. & Dinarelli M. (2020). Tarc : Incrementally and semi-automatically collecting a tunisian arabish corpus. *cs.CL*. arXiv preprint 2003.09520.

Habash N., Eryani F., Khalifa S., Rambow O., Abdulrahim D., Erdmann A., Faraj R., Zaghouani W., Bouamor H., Zalmout N., Hassan S., Al-Shargi F., Alkhereyf S., Abdulkareem B., Eskander R., Salameh M. & Saddiki H. (2018). Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

Haspelmath M. & Tadmor U. (2009). *Loanwords in the world's languages : a comparative handbook*. Walter de Gruyter.

Hert P. (1999). Quasi-oralité de l'écriture électronique et sentiment de communauté dans les débats scientifiques en ligne. *Réseaux*, **17**(97).

Masmoudi A., Ellouze Khemakhem M., Estève Y. & Hadrich-Belguith L. (2014a). Tunisian arabic railway interaction corpus. In *Arabic Natural Language Processing Research Group (ANLP)* : MIRACL Laboratory, Sfax (Tunis).

Masmoudi A., Habash N., Ellouze M., Estève Y. & Hadrich-Belguith L. (2015). Arabic transliteration of romanized tunisian dialect text : A preliminary investigation. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), p. 608–619: Springer-Verlag. doi : 10.1007/978-3-319-18111-0_46.

Masmoudi A., Khmekhem M. E., Estève Y., Belguith L. H. & Habash N. (2014b). A corpus and phonetic dictionary for Tunisian Arabic speech recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 306–310, Reykjavik, Iceland : European Language Resources Association (ELRA).

Medhaffar S., Bougares F., Estève Y. & Hadrich-Belguith L. (2017). Sentiment analysis of Tunisian Dialects : Linguistic ressources and experiments. In *Proceedings of the third Arabic Natural Language Processing Workshop*, p. 55–61: Association for Computational Linguistics. doi : 10.18653/v1/W17-1307.

Meftouh K., Harrat S., Jamoussi S., Abbas M. & Smaili K. (2015). Machine Translation Experiments on PADIC : A Parallel Arabic DIalect Corpus. In *The 29th Pacific Asia Conference on Language, Information and Computation*, shanghai, China. hal : hal-01261587.

Meftouh K., Harrat S. & Smaïli K. (2018). PADIC : extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey. hal : hal-01718858.

Tobaili T. (2016). Arabizi identification in twitter data. In *Proceedings of the ACL 2016 Student Research Workshop*, p. 51–57.

Younes J., Achour H. & Souissi E. (2015). Constructing linguistic resources for the tunisian dia-

lect using textual user-generated contents on the social web. In *International Conference on Web Engineering*, p. 3–14: Springer.

Younes J., Achour H., Souissi E. & Ferchichi A. (2018a). Survey on corpora availability for the tunisian dialect automatic processing. In *2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO : TICET-ICCA-GECO)*, p. 1–7: IEEE.

Younes J., Achour H., Souissi E. & Ferchichi A. (2020). Romanized tunisian dialect transliteration using sequence labelling techniques. *Journal of King Saud University-Computer and Information Sciences*.

Younes J., Souissi E., Achour H. & Ferchichi A. (2018b). A sequence-to-sequence based approach for the double transliteration of tunisian dialect. *Procedia computer science*, **142**, 238–245.

Zribi I., Ellouze M., Hadrich-Belguith L. & Blache P. (2015). Spoken tunisian arabic corpus "stac" : Transcription and annotation. *Research in Computing Science*, **90**, 123–135.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Segmentation automatique en périodes pour le français parlé

Natalia Kalashnikova[1], Iris Eshkol-Taravella[2], Loïc Grobol[3,4], François Delafontaine[1]

(1) LLL UMR 7270, 10 Rue de Tours, 45065 Orléans, France
(2) MoDyCo UMP 7114, 200 Avenue de la République 401B, 92001 Nanterre, France
(3) Lattice, 1 Rue Maurice Arnoux, 92120 Montrouge, France
(4) LLLF (Université de Paris, CNRS)

RÉSUMÉ

Nous proposons la comparaison de deux méthodes de segmentation automatique du français parlé en périodes macro-syntaxiques, qui permettent d'analyser la syntaxe et la prosodie du discours. Nous comparons l'outil Analor (Avanzi *et al.*, 2008) qui a été développé pour la segmentation des périodes prosodiques et les modèles de segmentations utilisant des CRF et des traits prosodiques et / ou morpho-syntaxiques. Les résultats montrent qu'Analor divise le discours en plus petits segments prosodiques tandis que les modèles CRF détectent des segments plus larges que les périodes macro-syntaxiques. Cependant, les modèles CRF ont de meilleurs résultats qu'Analor en termes de F-mesure.

ABSTRACT

Automatic Period Segmentation of Oral French

Natural Language Processing in oral speech segmentation is still looking for a minimal unit for analyze. In this work, we propose a comparison of two methods of automatic segmentation in macro-syntactic periods which allows to take into account syntactic and prosodic components of speech. We compare the performances of an existing tool Analor (Avanzi *et al.*, 2008) developed for automatic segmentation of prosodic periods and of CRF models relying on syntactic and / or prosodic features. We find that Analor tends to divide speech into smaller segments and that CRF models detect larger segments than macro-syntactic periods. However, in general CRF models perform with better results than Analor in terms of F-measure.

MOTS-CLÉS : français oral, segmentation automatique, périodes, CRF, unités macro-syntaxiques.

KEYWORDS: spoken language, automatic segmentation, period, oral french, CRF, macro-syntactic units.

1 Introduction

Dans le domaine du traitement automatique des langues, la segmentation des données linguistiques est une étape préalable à la plupart des tâches. Pour le traitement du langage écrit, l'unité de base est la phrase. Or ce type de segmentation est d'une pertinence limitée pour le langage parlé. C'est la raison pour laquelle Lacheret & Victorri (2002) proposent une unité de segmentation appelée la période prosodique. Cette notion est fondée sur les observations et les analyses de l'oral. Ainsi, cette approche ne tient pas compte de la syntaxe ni de la sémantique.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 241–248.
HAL : hal-02784773.

Notre travail s'inscrit dans le cadre du projet SegCor dont le but est le développement de plusieurs outils de segmentation automatique des unités linguistiques. Dans ce travail, nous nous intéressons aux périodes dans le cadre du modèle fribourgeois de la macro-syntaxe. Ce modèle définit les périodes du point de vue prosodique et syntaxique. L'objectif est de créer un outil automatique pour la segmentation de périodes macro-syntaxiques.

Notre approche aborde la tâche de segmentation comme un problème d'étiquetage des séquences. Pour cela, nous proposons d'utiliser un algorithme d'apprentissage automatique utilisant les *Conditional Random Fields* (Lafferty *et al.*, 2001) en nous appuyant sur des traits lexicaux, syntaxiques et prosodiques.

La suite de cet article est divisée en 7 parties. Les sections 2, 3 et 4 présentent un état des lieux de la recherche sur les notions de périodes, corpus et d'annotation manuelle qui sert de référence pour les méthodes automatiques. Les sections 5, 6 et 7 décrivent les expériences d'annotation automatique, leurs résultats, la conclusion et les perspectives pour de futures recherches.

2 Notion de période

Lacheret & Victorri (2002) définissent la période comme la structure prosodique qui lie plusieurs constructions syntaxiques dans un seul bloc discursif. Plusieurs périodes prosodiques peuvent aussi être incluses dans une seule structure syntaxique. Les périodes sont définis selon les paramètres prosodiques suivants : 1) Pause d'au moins 300 millisecondes ; 2) Différence de hauteur entre la valeur moyenne de la F0 de tout le signal acoustique et la dernière valeur de la F0 avant la pause ; 3)Différence de hauteur entre la dernière valeur de la F0 avant la pause et la première valeur de la F0 après la pause ; 4) Absence des signes d'hésitation (« euh ») avant et après la pause.

(1) et vous logez euh () la le la façade du théâtre (0.72)

Dans l'exemple (1) la durée de la pause marque la fin de la période après le mot "théâtre". Analor (Avanzi *et al.*, 2008) est un outil semi-automatique développé dans le cadre de cette théorie.

Une autre approche est celui de Berrendonner (2012) qui considère les périodes comme une unité prosodique autonome définie par son contour mélodique conclusif (Berrendonner, 2017). Les approches macro-syntaxiques s'appuient sur la prosodie pour analyser la structure syntaxique de l'oral (Blanche-Benveniste *et al.*, 1990; Cresti *et al.*, 2011). Pour cette raison, la période constitue potentiellement à la fois la structure complète et l'unité maximale de monologue (Blanche-Benveniste, 2012; Berrendonner, 2012, 34-35). A notre connaissance il n'existe aucun outil pour la segmentation automatique des périodes macro-syntaxiques.

Notre étude vise à trouver la méthode la plus performante pour la segmentation automatique des périodes macro-syntaxiques. Dans ce but, nous comparons deux méthodes. La première utilise Analor, qui ne nécessite pas d'entraînement et qui n'analyse pas la syntaxe des périodes. La deuxième reformule cette segmentation comme une tâche d'étiquetage de séquences — une modélisation déjà utilisée avec succès pour d'autres tâches de segmentation en français (Tellier *et al.*, 2012, 2014; Eshkol-Taravella *et al.*, 2019; Tellier *et al.*, 2013). Pour cette deuxième méthode, nous nous appuyons sur un algorithme d'apprentissage automatique bien connu : les CRF (*Conditional Random Fields*) (Lafferty *et al.*, 2001) en utilisant des traits syntaxiques et prosodiques.

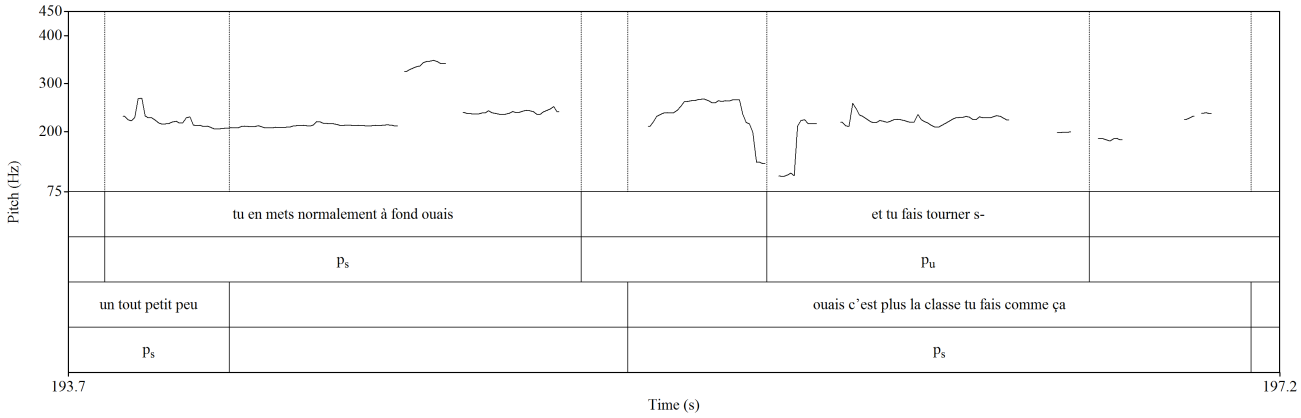


FIGURE 1 – Exemple de courbe intonative.

3 Corpus

Dans cette étude, nous travaillons sur un corpus pilote composé de 10 transcriptions de dialogues de 10 minutes et de 1 transcription de monologue de 20 minutes. Les extraits ont été sélectionnés pour représenter les différents types de discours du point de vue de l'environnement conversationnel, des relations entre les locuteurs, etc. Le corpus pilote est manuellement annoté en périodes macro-syntaxiques (voir Section 4).

4 Annotation manuelle

L'annotation manuelle réalisée par un annotateur en utilisant le logiciel Praat (Boersma & Weenik, 2002) s'est appuyée sur la définition de la période proposée par le modèle de la macro-syntaxe fribourgeoise (Berrendonner, 2012). L'un de ces deux critères devait être rempli pour reconnaître une frontière de période : la détection d'un contour intonatif conclusif ou un changement effectif de locuteur.

La détection de contours intonatifs a reposé sur une approche perceptive, par l'écoute du locuteur. L'annotation manuelle suit les mêmes propriétés dégagées par la théorie de Lacheret & Victorri (2002) qui servent à l'annotation automatique. Quant au changement de locuteur, il s'agit soit de cas où le locuteur interrompt son discours du fait de l'intervention d'un interlocuteur, soit de cas où le locuteur abandonne une structure incomplète après une longue pause (plus de 0.8 seconde).

(2) ELI tu en mets normalement à fond ouais (0.5)

BEA ouais [c'est plus classe tu fais] comme ça

ELI [et tu fais tourner]

La figure 1 illustre l'exemple (2) et le complète avec la courbe de la fréquence fondamentale.

La 1ère période d'ELI est suivie d'un changement de locuteur accompagné par la perception d'un contour conclusif montant que la mesure acoustique échoue à reproduire. La seconde période d'ELI est interrompue par le chevauchement, sans contour conclusif mais suivie d'une longue pause. La

période de BEA telle que reportée dans l'exemple 2 présente par ailleurs une structure composée de multiples unités micro-syntaxiques.

Au-delà du changement de locuteur, la syntaxe a également joué un rôle prédominant dans la gestion des pauses qui peuvent être soit une simple suspension du discours, soit l'abandon même momentané du tour par le locuteur, par exemple pour une requête lexicale (Lerner, 1991). Cette distinction devient essentielle en contexte monologique, y compris pour des pauses inférieures à 0.8 seconde. Dans de tels cas les indices prosodiques seuls s'avèrent insuffisants.

5 Expériences

5.1 Analor

Le corpus-pilote est d'abord annoté semi-automatiquement avec l'outil Analor (Avanzi *et al.*, 2008). Cet outil prend en compte les paramètres prosodiques développés dans le cadre de la théorie de Lacheret & Victorri (2002). Il semble important d'explorer la performance de l'outil pour l'annotation des périodes macro-syntaxiques due à la différence des définitions des périodes.

5.1.1 Pré-traitement

Le lancement du processus d'annotation sur Analor nécessite les fichiers PitchTier et TextGrid (format Praat). Les fichiers TextGrid contiennent 3 tires : token (un intervalle par token de l'enregistrement), locuteur (un intervalle par tour de parole) et l'annotation manuelle des périodes (un intervalle par période pour chaque locuteur).

5.1.2 Expériences

Analor est un outil pré-entraîné, on peut donc utiliser toutes les données du corpus-pilote pour la segmentation. Après la procédure d'annotation, Analor crée un autre fichier TextGrid qui contient une nouvelle tire de l'annotation automatique pour chaque fichier son. Analor réalise l'étiquetage des périodes sur une seule tire, tandis que le fichier TextGrid de l'annotation manuelle contient une tire par locuteur pour chaque fichier son. Nous avons résolu ce problème par la séparation manuelle de cette tire en une tire par locuteur. Un autre problème rencontré est le nombre différent de périodes entre l'annotation manuelle et l'annotation automatique. La solution pour cela est la tokenisation des périodes de l'annotation manuelle et de l'annotation automatique.

5.2 Modèles CRF

Les Conditional Ramdom Fields (CRF, parfois « Champs Markoviens Conditionnels » en français), (Lafferty *et al.*, 2001) sont des modèles d'étiquetage de séquences conçus pour permettre un apprentissage automatique robuste vis-à-vis d'effets à longue distance. En particulier, leur usage pour des tâches d'étiquetage modélisant une segmentation est bien connu pour des tâches de chunking

(Eshkol-Taravella *et al.*, 2019; Tellier *et al.*, 2012) et la reconnaissance d'entités nommées (Dupont & Tellier, 2014).

5.2.1 Traits

Pour le développement des modèles CRF nous utilisons deux types de traits : prosodiques et morpho-syntaxiques. Les traits prosodiques sont la fréquence fondamentale (les valeurs maximale, minimale et moyenne), l'intensité (les valeurs maximale, minimale et moyenne) et la durée pour chaque token.

Les traits morpho-syntaxiques sont les POS étiquetés par TreeTagger (Schmid, 1994) et les chunks issus d'un outil développé par Eshkol-Taravella *et al.* (2019) pour le projet SegCor. Ainsi, on construit 3 types de modèles CRF : le premier est créé uniquement sur les traits prosodiques, le deuxième est basé sur les traits prosodiques et morpho-syntaxiques et le troisième est entraîné uniquement sur les traits morpho-syntaxiques.

Nous faisons cette répartition dans le but de répondre à la question suivante : est-ce que les traits prosodiques contiennent assez d'information pour réaliser la segmentation des périodes macro-syntaxiques ou avons-nous aussi besoin des traits morpho-syntaxiques ?

5.2.2 Pré-traitement

TABLE 1 – Exemple de traits prosodiques et d'étiquettes pour le modèle CRF

mot	$f_0^{\max}$	f_0^{mean}	$f_0^{\min}$	durée	$\text{int}_{\max}$	int_{mean}	$\text{int}_{\min}$	BILU
ça	9	10	9	82	40	41	39	pA_B
va	8	9	7	83	41	43	39	pA_L
dis	9	14	6	82	41	41	40	pA_B
je	13	14	10	81	42	43	41	pA_I
voulais	8	10	7	81	42	43	42	pA_I
te	7	9	6	80	43	43	42	pA_I
demander	9	10	6	86	42	42	42	pA_L
demain	9	14	6	84	39	40	38	pA_B

Pendant la phase de pré-traitement, les traits acoustiques sont extraits pour chaque token en utilisant le logiciel Praat. Les valeurs prosodiques sont divisées en groupes de valeurs pour faciliter l'entraînement des modèles CRF. Les valeurs d'intensité sont discrétisées avec un pas de 10, de la F0 avec un pas de 20 et de la durée avec un pas de 0.1. Les fichiers TextGrid contenant les tokens et les traits morpho-syntaxiques sont convertis au format tabulaire.

Les données sont organisées selon les séquences de tours de parole. Un tour de parole peut donc contenir plusieurs périodes. En tenant compte du fait que le corpus pilote contient peu de données, notamment des tours de parole avec plusieurs périodes, nous avons décidé d'élargir le corpus en multipliant les données existantes. On a gardé les mêmes valeurs de traits mais en remplaçant les tokens par des faux mots. Les faux mots ont été créés par tirage aléatoire des caractères. Cela a permis au système d'avoir plus de données prosodiques pour l'entraînement tout en l'empêchant de procéder par simple mémorisation des mots.

Nous analysons également l'influence de l'intensité et de la fréquence fondamentale en entraînant les modèles uniquement sur un de ces traits prosodiques. Cette méthode permet d'observer quel paramètre prosodique est le plus important pour la segmentation des périodes.

Le corpus est divisé en 3 parties : 60 % pour l'entraînement, 30 % pour l'évaluation et 10 % pour le développement. Au total, nous avons 6 configurations différentes pour entraîner nos modèles CRF.

6 Expériences et résultats

Nous utilisons le logiciel Wapiti (Lavergne *et al.*, 2010) pour construire les modèles CRF.

TABLE 2 – Comparaison des résultats avec les différents traits.

Modèle	P	R	F
Analor	0.52	0.22	0.31
Prosodie	0.56	0.78	0.66
Morphosyntaxe	0.54	0.68	0.60
Prosodie + morphosyntaxe	0.56	0.78	0.66
Prosodie + morphosyntaxe + augmentation	0.56	0.70	0.62
f_0	0.55	0.76	0.64
Intensité	0.72	0.55	0.62

Le tableau 2, compare les performances de nos deux méthodes en termes de précision, de rappel et de F-mesure, calculés à partir des étiquettes BILU des périodes annotées automatiquement et manuellement.

Les valeurs rapportées sont des valeurs de détection des périodes (et non simplement des étiquettes) en considérant qu'une période est détectée si le modèle a correctement identifié ses frontières gauches et droites.

Pour Analor, le score de précision est plus haut que le score de rappel. Ceci est dû au fait qu'Analor détecte des segments plus petits que les périodes macro-syntaxiques. De plus, dans la plupart des cas, les meilleurs résultats correspondent aux locuteurs ayant le moins de temps de conversation et inversement.

Pour les modèles CRF, il semble que bien que l'utilisation uniquement des traits morpho-syntaxiques donne déjà de meilleurs résultats que l'utilisation directe d'Analor, les ajouter à un modèle ayant accès aux traits prosodiques n'améliore pas les performances, les traits prosodiques seuls obtenant déjà les meilleurs résultats.

Pour définir l'importance de chaque trait prosodique, nous comparons les résultats des modèles construits uniquement sur les valeurs de la F0 et de l'intensité. Les résultats obtenus montrent une grande complémentarité entre ces traits, chacun contribuant à un des aspects de la détection, et leur combinaison donnant de meilleurs résultats que leurs usages en isolation.

7 Conclusions et perspectives

Dans ce travail, nous avons présenté une nouvelle méthode pour la segmentation automatique de l'oral. Nous avons analysé les périodes macro-syntaxiques qui permettent de tenir compte du contenu prosodique et morpho-syntaxique du discours. La performance d'Analor n'est pas assez satisfaisante pour l'annotation des périodes macro-syntaxiques.

Tous les modèles CRF ont montré de meilleurs résultats qu'Analor. La F-mesure varie entre 0,54 et 0,66 parmi les modèles CRF différents. Si l'on compare la performance de chaque modèle CRF et que l'on tient compte du temps de pré-traitement, le modèle le plus performant est développé sur le corpus initial de la 1ère échelle des valeurs.

Nous avons également trouvé que le modèle fondé sur les traits de la F0 montre de meilleurs résultats que le modèle développé sur les traits de l'intensité.

Étant donné la quantité limitée des données, il serait intéressant de procéder à une validation croisée lors de l'entraînement des modèles. De plus, on pourrait appliquer les tests de significativité sur les résultats obtenus par les modèles.

Une piste de recherche pourrait être une évaluation de l'importance des différentes caractéristiques prosodiques sur la performance des modèles CRF. Il serait également envisageable de calculer la différence entre les valeurs d'un mot et du mot précédent lors du pré-traitement.

Références

AVANZI M. (2005). Quelques hypothèses à propos de la structuration interne des périodes. In *Actes Du Symposium Interface Discours-Prosodie*, Aix-en-Provence, France.

AVANZI M. (2012). *L'interface prosodie/syntaxe en français : dislocations, incises et asyndètes*. GRAMM-R. Bruxelles, Belgique : Peter Lang.

AVANZI M., LACHERET A. & VICTORRI B. (2008). Analor, un outil d'aide pour la modélisation de l'interface prosodie-grammaire. *Travaux linguistiques du CerLiCo*, **21**, 27–46.

BALTHASAR L. & BERT M. (2005). La plate-forme "Corpus de langues parlées en interaction" (CLAPI) : historique, états des lieux, perspectives. *LIDIL - Revue de linguistique et de didactique des langues*, **31**, 13–33.

BAUDE O. & DUGUA C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*, **10**, 99–118. HAL : hal-01162479.

BERRENDONNER A., Éd. (2012). *Grammaire de La Période*. Sciences pour la communication. Peter Lang. DOI : 10.3726/b11424.

BERRENDONNER A. (2017). La notion de période (note terminologique). *Encyclopédie grammaticale du français*.

BLANCHE-BENVENISTE C. (2012). Postface. In (Berrendonner, 2012), p. 341–355. DOI : 10.3726/b11424.

BLANCHE-BENVENISTE C., BILGER M., ROUGET C., VAN DEN EYNDE K. & WILLEMS D. (1990). *Le Français parlé : études grammaticales*. Paris, France : Centre national de la recherche scientifique.

BOERSMA P. & WEENIK D. (2002). Praat, a system for doing phonetics by computer. *Glot International*, **5**(9/10), 341–345.

CRESTI E., MONEGLIA M. & TUCCI I. (2011). Annotation de l'entretien d'Anita Musso selon la Théorie de la langue en acte. *Langue francaise*, **170**(2), 95–110.

DUPONT Y. & TELLIER I. (2014). Un reconnaisseur d'entités nommées du Français. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, volume 3, p. 40–41, Marseille, France : Association pour le Traitement Automatique des Langues.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral « disponible » : Le corpus d'Orléans 1 1968-2012. *Traitement Automatique des Langues*, **53**(2), 17–46.

ESHKOL-TARAVELLA I., MAAROUF M., BADIN F. & SKROVEC M. (2019). Chunker différents types de discours oraux : défis pour l'apprentissage automatique. In *Actes de La 26eConférence sur le Traitement Automatique des Langues Naturelles*, Toulouse, France : ATALA.

LACHERET A. & VICTORRI B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum*, **24**, 55–72.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *18th International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical Very Large Scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 504–513, Uppsala, Sverige : Association for Computational Linguistics.

LERNER G. H. (1991). On the syntax of sentences-in-progress*. *Language in Society*, **20**(3), 441–458. DOI : 10.1017/S0047404500016572.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

TELLIER I., DUCHIER D., ESHKOL I., COURMET A. & MARTINET M. (2012). Apprentissage automatique d'un chunker pour le français. In G. S. GEORGES ANTONIADIS, HERVÉ BLANCHON, Éd., *Conférence Conjointe JEP-TALN-RECITAL 2012*, volume 2, p. 431–438, Grenoble, France.

TELLIER I., DUPONT Y., ESHKOL I. & WANG I. (2013). Adapt a Text-Oriented Chunker for Oral Data : How Much Manual Effort Is Necessary ? In H. YIN, K. TANG, Y. GAO, F. KLAWONN, M. LEE, T. WEISE, B. LI & X. YAO, Éds., *Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning*, Lecture Notes in Computer Science, p. 226–233, Berlin, Heidelberg : Springer. DOI : 10.1007/978-3-642-41278-3_28.

TELLIER I., ESHKOL-TARAVELLA I., DUPONT Y. & WANG I. (2014). Peut-on bien chunker avec de mauvaises étiquettes POS ? In B. BIGI, Éd., *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, p. 125–136, Marseille, France.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Les avis sur les restaurants à l'épreuve de l'apprentissage automatique

Hyun Jung Kang Iris Eshkol-Taravella
MoDyCo UMR7114, 200 Avenue de la République, 92001 Nanterre, France
hyunjung.kang@parisnanterre.fr, ieshkolt@parisnanterre.fr

RÉSUMÉ

Dans la fouille d'opinions, de nombreuses études portent sur l'extraction automatique des opinions positives ou négatives. Cependant les recherches ayant pour objet la fouille de suggestions et d'intentions sont moins importantes, malgré leur lien profond avec l'opinion. Cet article vise à détecter six catégories (opinion positive/mixte/négative, suggestion, intention, description) dans les avis en ligne sur les restaurants en exploitant deux méthodes : l'apprentissage de surface et l'apprentissage profond supervisés. Les performances obtenues pour chaque catégorie sont interprétées ensuite en tenant compte des spécificités du corpus traité.

ABSTRACT

An Empirical Examination of Online Restaurant Reviews

In opinion mining, many works focus on the automatic extraction of positive or negative opinions. However, researches on suggestion and intention mining haven't been addressed much, despite their strong connection to opinion. This article aims to detect six categories (positive/negative/mixed opinion, suggestion, intention, description) in online restaurant reviews using two methods : traditional supervised learning and deep learning. We then interpret the performances obtained for each category by taking into account the specificities of the corpus treated.

MOTS-CLÉS : fouille d'opinions, avis en ligne, apprentissage supervisé, apprentissage profond, suggestion, intention.

KEYWORDS: opinion mining, online reviews, machine learning, deep learning, suggestion, intention.

1 Introduction

Les avis en ligne sont des sources d'analyses dans différents domaines : le marketing, l'informatique, la linguistique, le TAL où ils sont souvent considérés comme relevant de la notion d'opinion. De nombreux travaux portent sur la fouille d'opinions dont l'objectif est de classifier un document (Pang *et al.*, 2002; Turney, 2002) ou une phrase (Wiebe *et al.*, 1999) selon sa polarité. L'opinion peut être aussi extraite comme un tuple (ABSA ; *Aspect Based Sentiment Analysis*) qui se compose des éléments suivants : entité, aspect, sentiment, porteur et temps (Hu & Liu, 2004; Liu, 2012; Hamon *et al.*, 2015; Lark, 2017).

D'autres notions apparaissent en lien avec les avis comme les suggestions ou les intentions. (Benamara *et al.*, 2017), par exemple, affirment que la détection des intentions permet de compléter l'analyse de

Volume 2 : Traitement Automatique des Langues Naturelles, pages 249–257.
HAL : hal-02784774.

sentiments et d'opinions. Cependant les recherches ayant pour objet la détection de suggestions et d'intentions sont moins nombreuses. Le travail de (Ramanand *et al.*, 2010) est, selon nous, le premier à s'intéresser à la détection de suggestions. Il distingue deux types de souhaits (*wish*) : le souhait d'améliorer un produit et le souhait d'acheter celui-ci. (Brun & Hagège, 2013) analysent les avis sur les produits et établissent un ensemble de règles de leur détection en se fondant sur des éléments linguistiques. (Negi & Buitelaar, 2015) étudient les avis sur les hôtels et les produits électroniques dans lesquels les auteurs formulent des conseils ou offrent des suggestions aux futurs utilisateurs. (Negi *et al.*, 2016) évaluent différentes méthodes de détection de suggestion telles que les règles linguistiques élaborées manuellement, les machines vectorielles de support (SVM) et l'apprentissage profond. L'une des tâches proposées à l'atelier SemEval-2019 (Negi *et al.*, 2019) avait pour objectif d'extraire les suggestions dans les avis et les forums sur Internet. La détection d'intentions est abordée dans le travail de (Carlos & Yalamanchi, 2012) qui catégorise l'intention dans le domaine du marketing et du service clientèle. (Chen *et al.*, 2013) effectuent une classification des intentions explicites. (Ding *et al.*, 2015) proposent un modèle, basé sur les réseaux de neurones convolutifs (CNN, *Convolutional Neural Network*), pour identifier si l'utilisateur manifeste une intention de consommation.

Les avis traités dans ce travail concernent les restaurants. Leur analyse linguistique a permis de proposer un modèle conceptuel qui dépasse la notion d'opinion positive/négative/mixte et qui intègre une dimension linguistique (Eshkol-Taravella & Kang, 2019). Le modèle propose trois nouvelles classes (i.e. suggestion, intention, description) qui sont moins étudiées dans la fouille d'opinion.

La suggestion est un conseil émis par un visiteur. Ses marqueurs linguistiques sont les verbes de parole comme « recommander », « conseiller », le mode conditionnel et impératif, les pronoms personnels (« vous ») et les adjectifs possessifs de la deuxième personne (« votre », « vos »). Les suggestions peuvent être adressées à la fois aux restaurants (afin qu'ils prennent conscience des problèmes) et aux autres clients potentiels (futurs visiteurs) comme dans les exemples suivants : « Je conseille le tiramisu. », « Une lumière un peu plus tamisée aurait été parfaite ». Pourtant, les travaux précédents consacrés à leur détection ne prennent en compte que l'un des deux destinataires.

L'intention est un souhait exprimé explicitement de revenir ou de ne pas pas revenir dans un restaurant, elle montre un engagement volontaire du visiteur (« On reviendra ! », « Nous ne reviendrons pas ! »). L'intention est marquée de manière explicite à travers les pronoms « je », « nous » et « on », les verbes au futur et le préfixe verbal d'itération « re- ».

La description concerne les informations factuelles associées à l'expérience vécue comme « Soirée pour notre anniversaire de mariage » ou « Nous y étions un midi », qui sont peu reconnues dans la fouille d'opinion. Malgré sa nature objective et sa faible fréquence, la description est néanmoins une information intéressante car elle permet aux lecteurs de découvrir l'arrière-plan de l'expérience comme la raison pour laquelle les visiteurs se rendent dans le restaurant, les personnes qui les accompagnent, *etc*.

(Eshkol-Taravella & Kang, 2019) ont présenté la détection automatique de ces catégories fondée sur l'apprentissage supervisé en comparant différents modèles (*Naïve Bayes, Support Vector Machine, Logistic Regression*) tout en tenant compte du déséquilibre entre les classes d'évaluation créées. Le score F-mesure 0,88 a été obtenu en utilisant le sur-échantillonnage de ADASYN associé à l'algorithme SVM. Le travail présenté ici vise deux objectifs : (1) comparer les résultats de deux techniques d'apprentissage supervisé utilisées (l'apprentissage de surface et l'apprentissage profond) ; (2) examiner et comparer les performances obtenues pour chaque catégorie détectée. La méthodologie appliquée est décrite dans la section 2. Les expériences de détection automatique des six catégories

proposées exploitant les techniques de l'apprentissage de surface et de l'apprentissage profond ainsi que l'analyse de leurs performances sous diverses facettes sont présentées dans la section 3.

2 Méthodologie

2.1 Annotations et prétraitements

Données traitées. Nous avons collecté 21 158 avis sur 87 restaurants situés à Paris depuis un site internet [1] dont 6 287 avis (17 268 phrases, avec une moyenne de dix mots par phrase) ont été annotées. Les prétraitements réalisés sont décrits dans (Eshkol-Taravella & Kang, 2019). Nous nous contentons de présenter ici leur synthèse.

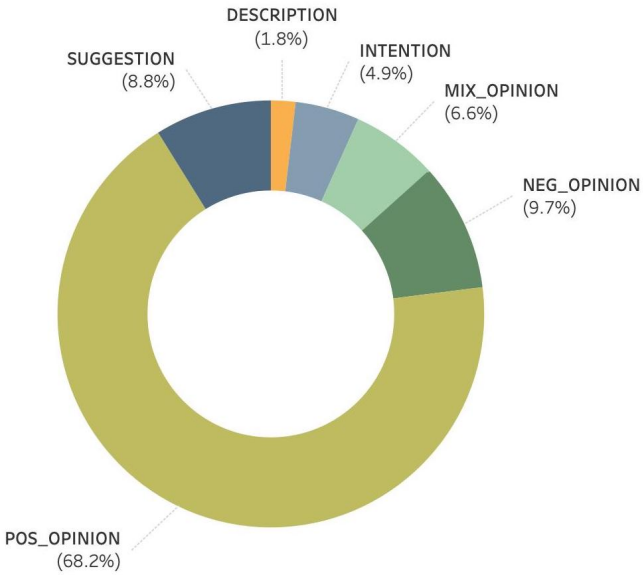


FIGURE 1 – Répartition des classes

Annotation. Chaque phrase a été annotée selon l'une des six catégories : POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION. Dans les cas d'ambiguïté où plusieurs catégories sont possibles, SUGGESTION ou INTENTION ont été privilégiées car celles-ci sont peu représentées dans les données. La phrase « Toujours aussi excellent, nous y retournerons c'est certain » par exemple, peut être classée dans les deux catégories POS_OPINION et INTENTION. Nous l'avons cependant annotée comme INTENTION. La tâche d'annotation a été évaluée par trois annotateurs (i.e. les doctorants en linguistique). Selon la mesure Kappa de Fleiss, nous avons obtenu 0,90, un accord considéré comme « presque parfait » (Landis & Koch, 1977). La distribution des catégories dans le corpus annoté s'est avérée non homogène (voir figure 1) : l'étiquette POS_OPINION constitue la majorité des catégories, représentant 68,2%, alors que DESCRIPTION et INTENTION n'en font que 1,8% et 4,9%.

Normalisation. Pour normaliser et nettoyer le corpus, certains traitements ont été effectués : le remplacement des émoticônes par « emoPOS » ou « emoNEG » selon la polarité [2], le passage des mots en

1. La Fourchette, https://www.lafourchette.com/
2. Les émoticônes se distinguent généralement par la polarité : ceux qui ont une polarité positive comme « :) », « :-) »

minuscules ; l'élimination de la ponctuation [3] ; la normalisation des mots consistant en la transformation des abréviations par leurs variations complètes (e.g. « resto » par « restaurant ») ; le remplacement des chiffres par une étiquette « NUM » [4] ; la lemmatisation en utilisant StanfordCoreNLP [5].

2.2 Classification

La classification automatique des phrases dans les six catégories prédéfinies a été effectuée en utilisant trois méthodes : SVM (*Support Vector Machine*) linéaire, CNN (*Convolutional Neural Network*) et LSTM (*Long Short-Term Memory network*). Pour les expériences basées sur SVM, nous avons utilisé la librairie scikit-learn [6] et pour celles de CNN et LSTM, nous avons employé la librairie Keras [7] avec TensorFlow [8]. Pour évaluer toutes les expériences, une validation croisée stratifiée à 5 plis a été effectuée. Les techniques pour gérer l'équilibre de la répartition des catégories n'ont pas été appliquées dans cette étude.

SVM linéaire. Deux approches de représentation de texte ont été exploitées : l'approche du sac de mots et celle du plongement de mots (*embedding*). Pour la première approche, nous avons supprimé les mots-outils [9] et employé *CountVectorizer* et *TfidfVectorizer* de scikit-learn. Nous avons testé deux paramètres dont ces méthodes disposent, *n_gram* et *max_feature*. La meilleure combinaison des paramètres a été obtenue avec une procédure de grille de recherche (*GridSearch* [10]).

La deuxième approche concerne le plongement de mots, méthode améliorée par rapport à celle du sac de mots car capable de prendre en compte les similarités contextuelles entre les mots. Pour ce faire, nous avons entraîné Word2vec (Mikolov *et al.*, 2013) du type de CBOW (sac de mots continus) avec Gensim [11]. La taille de la fenêtre a été fixée à 6. Les traits (*features*) pris en compte lors de l'apprentissage sont : les catégories morphosyntaxiques jugées pertinentes (e.g. les noms, les verbes, les adjectifs) proposées par StanfordCoreNLP, les différentes variations des verbes (e.g. les verbes au futur, les verbes au conditionnel, les verbes possédant le préfixe 're-'), la négation, les mots positifs et négatifs [12], les scores de polarité et de subjectivité [13], le connecteur « mais », le symbole € (*euro*), les chiffres, les émoticônes, les multiples ponctuations en cascade, les mots en majuscule [14], la longueur de la phrase, la diversité et la densité lexicales. En nous basant sur ces traits définis, l'algorithme SVM linéaire a été appliqué.

CNN. Pour exploiter la technique des CNN, nous nous sommes appuyés sur la proposition de configurations proposée dans (Zhang & Wallace, 2017) pour une tâche de classification automatique

et ceux qui sont plutôt de polarité négative « :(». Une liste comprenant les émoticônes positifs et négatifs a été créée préalablement.

3. Pour éviter des interférences entre les traitements réalisés, leur ordre a été prédéfini.
4. Les informations factuelles (le nombre de personnes, l'heure, le prix, etc.) sont souvent présentées en chiffres.
5. https://stanfordnlp.github.io/CoreNLP/download.html
6. http://scikit-learn.org/stable/
7. https://keras.io/
8. https://www.tensorflow.org/
9. La liste des mots-outils proposée par NLTK (https://www.nltk.org/) a été modifié.
10. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
11. https://radimrehurek.com/gensim/ ; Les modèles Word2vec entraînés sur le corpus frWiki (Wikipédia) par (Fauconnier, 2015) ont été testés, or, le résultat des modèles pré-entraînés a été moins intéressant que nos propres modèles.
12. Proposés par Textblob (https://textblob.readthedocs.io/en/dev/)
13. Obtenus par Textblob
14. Puisque les données ont été normalisées, la fréquence des multiples ponctuations en cascade et des mots en majuscule sont extraites des données originales.

des textes. Nous avons pris en entrée une matrice d'embedding à l'aide de Word2vec entraîné précédemment sur notre corpus. Par la suite, nous avons appliqué un filtre de convolution de 32 neurones ainsi qu'un kernel de taille 4 suivie par la fonction d'activation ReLU (Unité de Rectification Linéaire). Une couche de *Max pooling* a ensuite été appliquée sur la sortie de la couche de convolution, divisant par 2 la sortie de la couche précédente. Enfin, nous avons exploité l'aplatissement et la réduction de dimension, en appliquant une couche dense à 10 unités, suivie par la fonction d'activation ReLU, avant d'exploiter une activiation softmax sur la couche finale composée de 6 neurones.

LSTM. Les LSTMs sont une variante de réseaux de neurones récurrents (RNN) considérés comme très performantes sur de longues séquences (Osinga, 2018). Nous avons utilisé comme entrée la même matrice d'embeddings que celui de CNN, suivie d'une couche avec 100 unités, envoyée ensuite à une couche dense et finissante par une activation softmax. L'avantage des LSTMs est de mieux prendre en compte les dépendances entres mots distants.

Pour ces types de réseaux de neurons, CNN et LSTM, les hyperparamètres choisis sont l'optimiseur Adam (*Adaptive Moments*) et une perte d'entropie. La taille du batch est de 5, avec 7 époques.

3 Résultats

Les performances des différents modèles ont été évaluées en calculant pour chacun d'eux la moyenne pondérée de la précision, du rappel, de la F-mesure et la matrice de confusion. La macro F-mesure, donnant un poids identique à chaque catégorie, ne tient pas compte de la répartition déséquilibrée des classes. Étant donné que cette répartition est asymétrique dans les données, la moyenne pondérée de la F-mesure est jugée pertinente [15], c'est la raison pour laquelle cette mesure est utilisée pour l'évaluation. Le tableau 1 illustre la comparaison des performances entre les méthodes employées. Le SVM linéaire utilisé avec le sac de mots (i.e. l'apprentissage de surface) a donné le meilleur résultat avec une moyenne pondérée de F-mesure égale à 0,88. L'apprentissage de surface a donc produit dans le cadre de cette étude un meilleur résultat que l'apprentissage profond. Nous considérons pourtant que ce dernier peut encore être amélioré au moyen de réseaux de neurones plus complexes. Bien que Word2vec soit présenté comme une approche optimale, nos résultats montrent que l'approche du sac de mots peut s'avérer plus performante que Word2vec dans les cas où les données ne sont pas importantes.

sac de mots+linearSVM	Word2vec+linearSVM	CNN	LSTM
0.88	0.80	0.85	0.84

TABLE 1 – La comparaison des moyennes pondérées de la F-mesure entre les méthodes employées

La matrice de confusion offre une vision globale de la meilleure performance, comme le montre la figure 2. Chaque ligne correspond à la classe réelle et chaque colonne à la classe prédite. Les cellules de la diagonale principale indiquent celles qui sont classifiées correctement. La cellule DESCRIPTION est légèrement plus claire car l'information réunie sous cette étiquette était plus difficile à être classifiée. Ce constat peut s'expliquer en partie par le manque d'échantillons de DESCRIPTION et donc par le faible nombre de traits fournis durant l'apprentissage. Par ailleurs, cette catégorie est très hétérogène et varie en fonction du profil de l'internaute, ce qui donne lieu à un

15. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score

large éventail de vocabulaires et de contextes. Par conséquent, DESCRIPTION a une tendance à être rangée dans la classe majoritaire, c'est-à-dire POS_OPINION (0,41) et occasionnellement comme NEG_OPINION (0,14). Nous observons également une tendance similaire pour MIX_OPINION, dont le mauvais score (0,66) vient du fait que la catégorie implique à la fois POS_OPINION (0,19) et NEG_OPINION (0,12) [16].

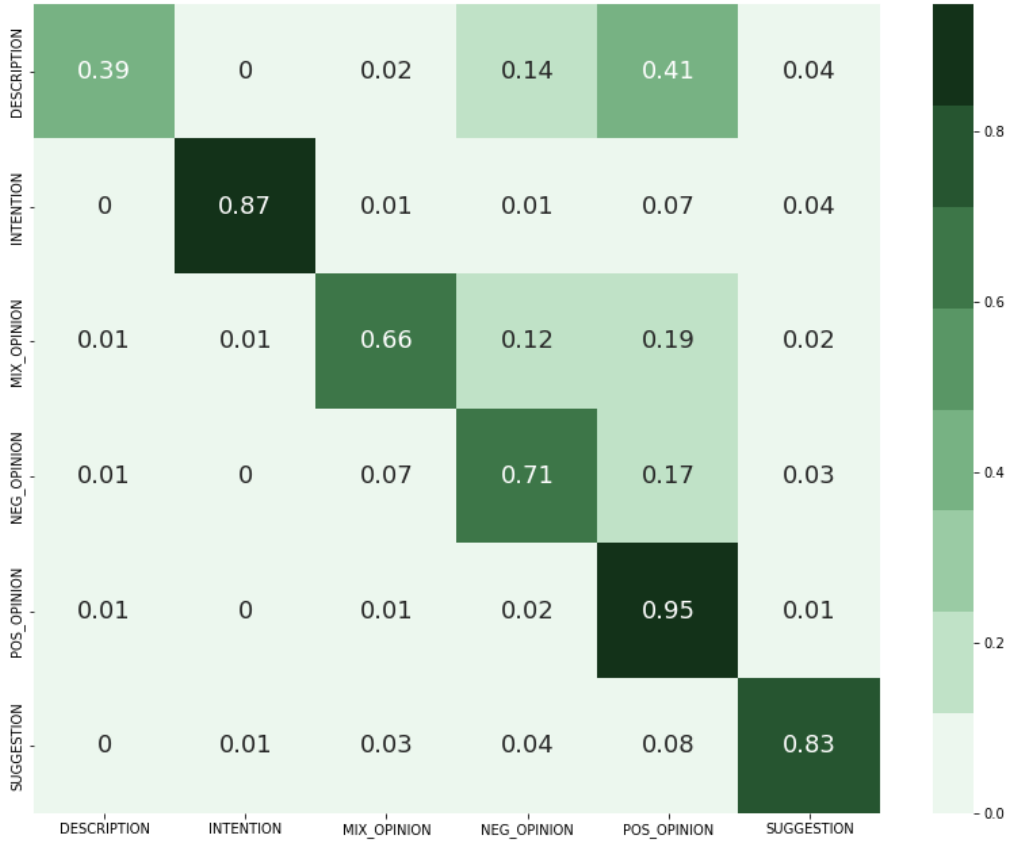


FIGURE 2 – La matrice de confusion normalisée de SVM linéaire utilisé avec le sac de mots

La figure 3 présente une comparaison de la moyenne pondérée de la précision, du rappel et de la F-mesure entre l'apprentissage de surface (sac de mots+SVM linéaire) et l'apprentissage profond (CNN) pour chaque catégorie d'évaluation. Une étiquette POS_OPINION est détectée avec la meilleure performance (la F-mesure étant d'environ 0,94) et l'INTENTION, obtient le deuxième meilleur score (0,86-0,88). Une étiquette DESCRIPTION est détectée avec la plus mauvaise performance (0,34-0,46) ayant le plus grand écart entre la précision et le rappel ce qui correspond à 0,16 pour chaque méthode d'apprentissage. Ce résultat est dû à sa faible fréquence dans le corpus. Par ailleurs, elle semble s'appuyer sur peu de marqueurs lexicaux car sa nature est très hétérogène.

D'une manière générale, les performances de l'apprentissage de surface sont supérieures à celle de l'apprentissage profond dans la majorité des cas. Les étiquettes DESCRIPTION et SUGGESTION ont une tendance à être mieux détectées avec l'apprentissage de surface ce qui montre que les traits linguistiques retenus sont utiles dans l'apprentissage. L'INTENTION a une précision supérieure mais un rappel inférieur à la précision lorsque l'apprentissage profond est exploité, donnant au final une F-mesure similaire à l'apprentissage de surface.

16. Par exemple, « Vraiment très bien, juste un peu trop bruyant. ».

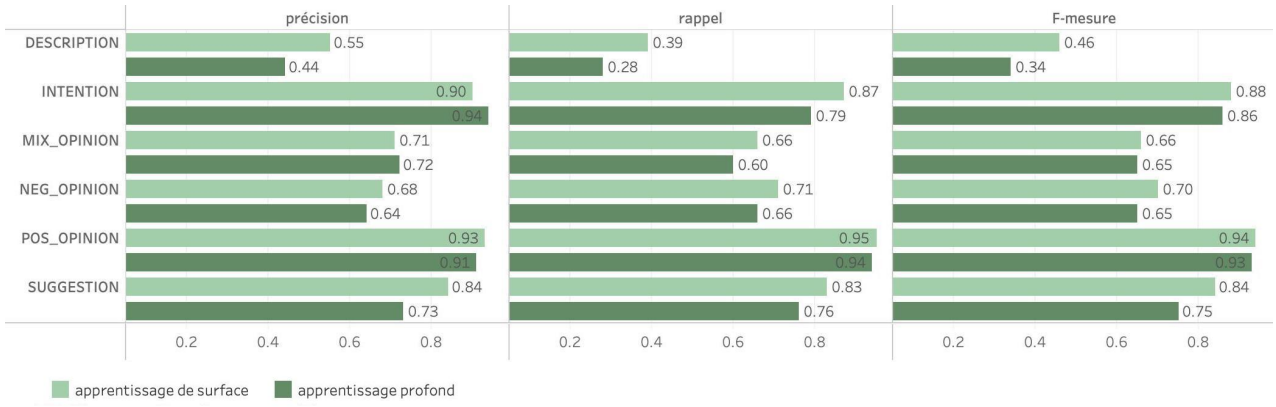


FIGURE 3 – Précision, Rappel et F-mesure d'apprentissage supervisé (sac de mots+linearSVC) et d'apprentissage profond (CNN)

Une catégorie MIX_OPINION semble poser également des difficultés aux classifieurs. L'observation manuelle du corpus montre que le changement de polarité est marqué souvent par les conjonctions comme « mais » (e.g. « Le restaurant était complet et très bruyant mais la cuisine excellente. », « Beau restaurant, très bons plats mais service à revoir. »). La segmentation du corpus selon les conjonctions pourrait améliorer la performance de la détection de cette catégorie mais aussi pourrait permettre sa suppression. La segmentation de la phrase « Le restaurant était complet et très bruyant mais la cuisine excellente. » en deux propositions « Le restaurant était complet et très bruyant » et « la cuisine excellente » permet de classer la première proposition dans une catégorie NEG_OPINION et la seconde dans POS_OPINION. L'approche de (Lark, 2017), qui emploie les règles lexicosyntaxiques appliquées sur les connecteurs, semble pouvoir résoudre cette difficulté.

4 Conclusion

Cet article décrit l'expérience portant sur la classification automatique des avis selon six catégories prédéfinies en exploitant deux techniques de l'apprentissage supervisée : l'apprentissage de surface et l'apprentissage profond. L'approche de surface obtient la meilleure moyenne pondérée de la F-mesure (0,88), qui chute légèrement dans le cas de l'apprentissage profond (0,86). Notons que l'apprentissage de surface est plus coûteux en termes de temps et d'efforts à cause des traits à fournir aux algorithmes. Lorsque l'approche de surface est appliquée, les moyennes pondérées de la F-mesure pour chaque catégorie sont : POS_OPINION (0,94), INTENTION (0,88), SUGGESTION (0,84), NEG_OPINION (0,70), MIX_OPINION (0,66) et DESCRIPTION (0,46). Parmi les trois nouvelles catégories proposées : INTENTION, SUGGESTION, DESCRIPTION, c'est la détection de cette dernière qui obtient les résultats moins satisfaisants. Pour améliorer son score, il faudrait augmenter la taille du corpus de référence. Par ailleurs, il serait intéressant de mesurer la généricité des catégories INTENTION et SUGGESTION dans d'autres corpus.

Références

BENAMARA F., TABOADA M. & MATHIEU Y. (2017). Evaluative language beyond bags of words :

Linguistic insights and computational applications. *Computational Linguistics*, **43**(1), 201–264. DOI : 10.1162/COLI_a_00278.

BRUN C. & HAGÈGE C. (2013). Suggestion mining : Detecting suggestions for improvement in users' comments. *Research in Computing Science*, **70**, 199–209.

CARLOS C. S. & YALAMANCHI M. (2012). Intention analysis for sales, marketing and customer service. In *Proceedings of COLING 2012 : Demonstration Papers*, p. 33–40, Mumbai, India : The COLING 2012 Organizing Committee.

CHEN Z., LIU B., HSU M., CASTELLANOS M. & GHOSH R. (2013). Identifying intention posts in discussion forums. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1041–1050, Atlanta, Georgia : Association for Computational Linguistics.

DING X., LIU T., DUAN J. & NIE J.-Y. (2015). Mining user consumption intention from social media using domain adaptive convolutional neural network. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

ESHKOL-TARAVELLA I. & KANG H. J. (2019). Observation de l'expérience client dans les restaurants. In *TALN 2019*.

FAUCONNIER J.-P. (2015). French word embeddings. http://fauconnier.github.io.

HAMON T., FRAISSE A., PAROUBEK P., ZWEIGENBAUM P. & GROUIN C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (deft).

HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 168–177.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

LARK J. (2017). *Construction semi-automatique de ressources pour la fouille d'opinion*. Thèse de doctorat, Nantes.

LIU B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv : 1301.3781.

NEGI S., ASOOJA K., MEHROTRA S. & BUITELAAR P. (2016). A study of suggestions in opinionated texts and their automatic detection. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, p. 170–178, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/S16-2022.

NEGI S. & BUITELAAR P. (2015). Towards the extraction of customer-to-customer suggestions from reviews. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. DOI : 10.18653/v1/D15-1258.

NEGI S., DAUDERT T. & BUITELAAR P. (2019). Semeval-2019 task 9 : Suggestion mining from online reviews and forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, p. 783–883.

OSINGA D. (2018). *Deep Learning Cookbook*. Practical Recipes to Get Started Quickly. O'Reilly Media.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? sentiment classification using machine learning techniques. *EMNLP*, **10**. DOI : 10.3115/1118693.1118704.

RAMANAND J., BHAVSAR K. & PEDANEKAR N. (2010). Wishful thinking : Finding suggestions and 'buy' wishes from product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, p. 54–61, Stroudsburg, PA, USA : Association for Computational Linguistics.

TURNEY P. D. (2002). Thumbs up or thumbs down ? : Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 417–424, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073153.

WIEBE J. M., BRUCE R. F. & O'HARA T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, p. 246–253, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.3115/1034678.1034721.

ZHANG Y. & WALLACE B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 253–263, Taipei, Taiwan : Asian Federation of Natural Language Processing.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Recherche de similarité thématique en temps réel au sein d'un débat en ligne

Mathieu Lafourcade[1, 2] Noémie-Fleur Sandillon-Rezer[1, 2]
(1) LIRMM, 161 rue Ada, 34095 Montpellier Cedex 5, France
(2) Université de Montpellier, 163 rue Auguste Broussonnet, 34090 Montpellier, France
mathieu.lafourcade@lirmm.fr, noemie-fleur.sandillon-rezer@lirmm.fr

RÉSUMÉ

Cet article se focalise sur l'utilisation d'un large réseau lexico-sémantique français pour le calcul de similarité thématique d'interventions au cours d'un débat en ligne dans les lycées, proche du temps réel. Pour cela, notre système extrait des informations sémantiques du réseau et crée à la volée des vecteurs enrichis pour chaque fragment de texte. Les données récupérées sont contextualisées via un algorithme de propagation. Les vecteurs résultat permettent aux fragments de texte d'être comparés. Notre méthode aide à trouver les thématiques émergentes des débats et à identifier des clusters d'opinion. La contrainte temps réel nous force à sélectionner précisément les informations que nous incluons, aussi bien pour les temps de calcul des vecteurs créés que la qualité de ceux-ci.

ABSTRACT

Thematic similarity real-time computation during an online debate

This paper describes the use of a large French lexical and semantic network for text embedding computation for thematic similarity, as close as possible to real time, in the context of in-school online debates. To this purpose, our system creates on the fly enriched vectors that embed thematic aspects of text fragments. Semantic information associated to textual contents are retrieved from a knowledge base, then contextualised by a graph propagation algorithm. Those lexicalized vectors allow texts to be thematically compared. The system helps teachers by finding emergent topics of discussions or identifying clusters of opinions. The real-time constraint forces us to choose precisely which semantic processing we include in vector building, as they can have a crucial impact.

MOTS-CLÉS : proximité thématique, réseau lexico-sémantique, vecteurs lexicalisés.

KEYWORDS: thematic similarity, lexical and semantic network, lexicalized vectors.

1 Introduction

Cet article présente une utilisation d'un réseau lexico-sémantique français pour le calcul de vecteurs textuels enrichis, cela dans le but de trouver des similarités thématiques entre des fragments de texte. Nous nous plaçons dans le cadre du projet AREN (ARgumentation Et Numérique), où notre méthode est appliquée en temps réel à des débats en ligne entre lycéens. Ce projet, soutenu par le ministère de l'Éducation Nationale, est un des lauréats de l'appel e-FRAN [1]. Son but principal est d'apprendre les mécanismes du débat aux lycéens via une plateforme mise en ligne et utilisée depuis

1. Espaces de Formation, de Recherche et d'Animation Numérique

Volume 2 : Traitement Automatique des Langues Naturelles, pages 258–267.
HAL : hal-02784775.

2017 (voir figure 1). Un des objectifs secondaires est d'exploiter les techniques du TAL pour assister élèves et enseignants durant un débat ou pendant l'étape de restitution. En effet, un débat en classe (dans le cadre du projet) se divise en trois parties : la préparation en classe, où les élèves acquièrent des données et connaissances sur le sujet à débattre, le débat en ligne, d'une durée de 50 minutes en général et la restitution du débat. Lors de celle-ci, un travail consiste à résumer les différents arguments du débat. Pour cela, pouvoir rassembler les arguments par thème est d'une grande aide pour les enseignants, leur permettant de vérifier ce que propose le système plutôt que lire parfois jusqu'à 300 arguments pour commencer le tri. Durant le débat, cela peut également permettre à l'enseignant de mettre l'accent sur un thème qu'il souhaiterait voir abordé et qui n'a pas encore assez de contributions. Les textes analysés sont donc courts (généralement une phrase), et écrits sur le vif.

L'idée soutenant notre méthode est de créer des vecteurs pour chaque contribution textuelle, d'enrichir ceux-ci via le réseau lexico-sémantique et de les comparer par un produit scalaire. Il est ainsi possible de déterminer automatiquement quelles contributions sont proches thématiquement les unes des autres, et ce même si le vocabulaire utilisé diverge.

FIGURE 1 – Interface d'AREN : texte débattu à gauche, commentaires à droite découpés en 3 parties : sélection (extrait à commenter), reformulation (de la sélection), argumentation (où on s'exprime).

Il est important de garder à l'esprit deux aspects du projet. La contrainte temps-réel nous impose d'utiliser des approches rapides et nous ne vérifions pas que les interventions sont sémantiquement convergentes (les négations, par exemple, ne sont pas prises en compte), mais simplement qu'elles portent sur le même thème. En outre, les vecteurs sont lexicalisés (ensemble de paires mot-poids) pour une interprétation humaine et machine facilitée, dans l'esprit de Panigrahi *et al.* (2019). Les vecteurs sont en dimension ouverte. Ils peuvent être composés d'autant de paires mot-poids que souhaité. Pour être comparés deux à deux, le fait de ne pas avoir la même dimension n'est pas problématique.

Les données utilisées sont les débats réalisés au cours des 4 ans de projet. Celles-ci seront mises à disposition au terme du projet, ainsi que le code produisant les vecteurs et la plateforme utilisée, le tout sous licence libre.

Dans cet article, nous commencerons par décrire rapidement certains aspects de la base de connaissance JeuxDeMots dont nous nous servons pour l'augmentation sémantique. Après un rapide aperçu des méthodes récentes sur les plongements de mots et de textes, nous décrirons notre méthodologie, puis l'évaluerons avant de conclure.

2 Utiliser une grande base de connaissances lexicalisées

Le projet JeuxDeMots [2] (Lafourcade, 2007) (JDM) a pour cœur des GWAP (un jeu en ligne, voir Ahn (2006)) où des joueurs s'affrontent pour capturer des mots, combiné à des mécanismes d'inférences. Le ressort principal (Lafourcade *et al.*, 2018) est de leur faire produire des associations entre termes selon une consigne (par exemple : donner des synonymes de "chat"). Ainsi, le réseau lexical JDM, dont la structure est composée de nœuds connectés par des relations (voir Collins & Quillian (1969), Sowa & Zachman (1992), Gaume *et al.* (2007) et Polguère (2014)), se développe en fonction de l'activité des participants (environ 4 millions de termes et 310 millions de relations en janvier 2020). Il existe environ 120 types de relations pouvant être organisés selon les catégories suivantes :

Relations lexicales - Focalisées sur le vocabulaire et la lexicalisation, cela correspond à la synonymie, antonymie, champ lexical, etc.
Relations ontologiques - Centrées sur les connaissances de la langue, il s'agit des génériques (hyperonymie), spécifiques (hyponymie), parties de (meronymie), lieux spécifiques, etc.
Relations associatives - Plus subjectives, elle font appel à la culture générale : associations libres, sentiments associés, gloses, objets similaires, objets souvent présents ensembles.
Relations prédicatives - Associées à un vembe ou un nom d'action, aussi bien qu'aux valeurs des arguments : agent, patient, lieux où une action se déroule, etc.
Sens des termes permettant de représenter des raffinements spécifiques, par exemple : frégate>navire de frégate>oiseau.

En plus d'être typée, une relation est pondérée, le poids pouvant être négatif, indiquant une relation fausse ou impossible. Enfin, les relations peuvent être annotées de façon ouverte avec diverses informations : fréquences, pertinence, prépositions (pour les relations de lieux), subjectivité, etc. La base est très lexicalisée, en ce sens que les verbes arrivent avec leur formes conjuguées ; les groupes nominaux avec leur pluriel, etc. Un grand nombre de formes verbales infinitives arrivent avec leur version négative (manger / ne pas manger / ne plus manger, etc.) et leurs associations sémantiques respectives. Cela résulte d'un ajout de contributeurs hors jeu, et permet lors d'une analyse sémantique de rejeter certains sens.

3 Calcul de vecteurs avec une large base de connaissances

Contrairement aux approches récentes, où les plongements textuels sont calculés à partir d'un grand corpus – souvent Wikipedia Ein Dor *et al.* (2018)–, on s'appuie sur une large base de connaissances.

Les techniques de fouilles textuelles cherchant à extraire des données pertinentes de texte sont souvent utilisées pour mesurer la similarité (Vijaymeena1 & Kavitha (2016), Sumathy & Chidambaram (2016), Peinelt *et al.* (2019) et Gong *et al.* (2018)). L'enjeu des modèles à base de vecteurs est de construire lesdits vecteurs. En contrepartie, on adopte généralement ces modèles pour la facilité de comparaison des vecteurs. On peut citer à la volée : les méthodes LSA - Latent Semantic Analysis - (Magerman *et al.*, 2011), LDA - Latent Dirichlet Allocation - (Liu *et al.*, 2015), LSTM - Siamese Long Short Term Memory - (Melamud *et al.*, 2016) ou encore Doc2Vec, foncé sur Word2Vec (Le & Mikolov, 2014). Les réseaux neuronaux convolutionnels commencent également à être utilisés (Zheng *et al.*, 2019). Généralement, les approches sont entraînées sur larges corpus, vu qu'il est délicat de créer des

2. http://www.jeuxdemots.org

vecteurs pertinents sans aucune source de connaissance (Park *et al.*, 2018). Cela rend la récupération des informations sous-entendues plus complexe (Smalheiser *et al.*, 2019) , alors qu'un lecteur humain le fait aisément tant que lesdites informations viennent de son environnement culturel. Nous avons délaissé les méthodes telles que ELMo (Peters *et al.*, 2018) BERT (Devlin *et al.*, 2019) car nous souhaitions une méthodologie en contrôle, où nous choisissons les relations sémantiques où sont puisés les termes à ajouter.

Actuellement, nous exploitons la proximité thématique (par produit scalaire entre vecteurs) selon deux circonstances : (a) vérifier automatiquement que la reformulation est (thématiquement) similaire à la sélection (leurs vecteurs ont un produit scalaire élevé) et (b) identifier les commentaires (thématiquement) proches d'un groupe de mot pouvant jouer le rôle de thème spécifique. Cela permet de recentrer le débat facilement, de vérifier si certaines idées sont exprimées et combien de fois.

3.1 Schéma général de construction de vecteurs lexicalisés

L'idée maîtresse est de déclarer différentes étapes et de construire un pipeline avec. On différencie les étapes sémantiques – qui font appel aux connaissances issues de JeuxDeMots et ajoutent des informations importantes, comme les synonymes (Abdalgader & Skabar, 2011), et améliorent la qualité des vecteurs (Espinosa Anke *et al.*, 2019), avec les co-locations – et celles qui ne le sont pas (étapes de régularisation), mais qui sont indispensables pour construire des vecteurs (où les mots sont associés à leur poids). Le pipeline complet est illustré figure 2. Nos choix se fondent sur un compromis entre de temps de calcul et qualité. Il nous a bien sûr fallu faire des choix, principalement en terme d'augmentation, pour rester dans l'optique du temps réel (un débat en classe dure 50 minutes, il est nécessaire que les calculs soient terminés à la fin de celui-ci pour faire le point avec les élèves).

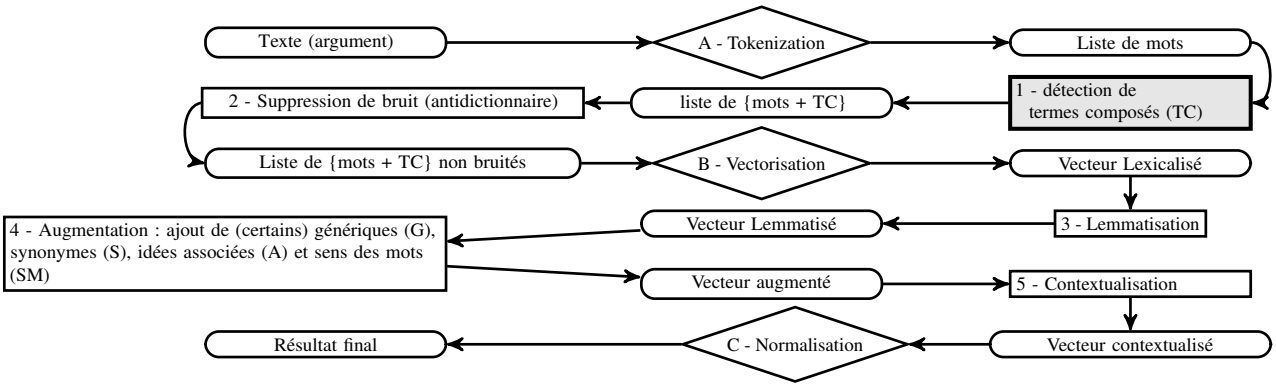


FIGURE 2 – Processus global ; les étapes sémantiques (1 - 5) exploitent la base de connaissances.

3.2 Étapes de régularisation

Elles peuvent être utilisées seules pour obtenir des vecteurs élémentaire très rapidement. Ces vecteurs peuvent être utilisés comme références (baseline). On considère un fragment de texte (ou fragment) comme une suite de mots $w_1 s_1 w_2 s_2 ... s_k w_n$ ($n, k \in \mathbb{N}$), où w_i représente un mot et s_j un séparateur.

A - Tokenisation : transforme un fragment en tokens, en détectant les séparateurs : $(w_1, w_2, ..., w_n)$.
B - Vectorisation : crée un vecteur en comptant et fusionnant les tokens dupliqués. L'ensemble pondéré obtenu est appelé vecteur lexicalisé : $\{w_1 : 2, w_2 : 1, ..., w_{n'} : 3\}$ où $n' \in \mathbb{N}$ et $n' \leq n$.
C - Normalisation : modifie les poids avec une norme euclidienne. Soit un vecteur (x, y), une fois normé il devient $(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}})$). Cela permet la comparaison de deux vecteurs.

3.3 Étapes sémantiques

Ces étapes sémantiques ajoutent des informations sémantiques au vecteur, en atténuant les fluctuations lexicales de langage utilisé, et surtout en identifiant les concepts sous-tendus tels que identifiés a priori par les locuteurs (dans le cadre du projet JeuxdeMots) aussi bien qu'en réduisant l'ambiguïté en identifiant les termes composés.

1 - Détection des termes composés (TC) : est nécessaire pour éviter les erreurs d'interprétation aussi bien que pour le raffinement. Par exemple *île flottante* est plus certainement un dessert qu'une île qui flotte et *véhicule autonome* est un concept précis qui a un sens et un champ lexical bien à lui.

Les termes composés sont extraits sous forme de liste de JDM ; celle-ci permet de créer un automate à états finis, qui lira les fragments (en prenant en compte les séparateurs) et concatènera les séquences correspondant à un TC avec "_" : *véhicule autonome* deviendra *véhicule_autonome*. Lorsque deux TCs se superposent, nous sélectionnons le terme le plus à droite (heuristique correspondant à ce qu'on observe plus souvent en français). La détection de TCs, pour être complète, doit être effectuée en trois passes, accompagnée entre chaque d'une *lemmatisation progressive*. Cela demande de construire le graphe des lemmatisations possibles et devient coûteux en temps. Les trois passes sont donc :
Sans lemmatisation : couvre les cas où les traits grammaticaux comptent (*monteur de câbles d'avions*).
Lemmatisation verbale : pour reconnaître des expressions telles que *mettre les pieds dans le plat*.
Lemmatisation totale : détecte les TCs enregistrés sous forme lemmatique, tel que *véhicule autonome*.

2 - Retrait du bruit : à partir d'une liste de mots vides – du bruit – on crée une expression rationnelle qui retire ces mots des fragments.

3 - Lemmatisation : récupère les lemmes des mots, depuis JDM. Leur poids est le même que celui du mot initial. Lorsque l'on effectue la détection des termes composés complète, cette étape y est incluse. Cependant, pour une exécution plus rapide, on peut réduire la détection des termes composés à sa première étape et il est alors nécessaire d'effectuer la lemmatisation à part.

4 - Augmentation : demande à JDM les mots associés (génériques, synonymes, idées associées et sens des mots) et récupère une liste de termes pondérés. Cette liste est triée par poids de relation, les négatifs sont exclus, et nous récupérons les k premiers, qui sont ajoutés au vecteur avec un facteur d'atténuation f qui s'applique sur le poids du terme initial. Le facteur d'atténuation ainsi que le nombre de termes sélectionnés peuvent être modifiés à l'appréciation de l'utilisateur. Empiriquement et dans le cadre de l'utilisation que nous faisons de la similarité thématique dans le projet AREN, nous avons fixé $f = 0.8$, pour les synonymes $k = 10$ et pour les autres augmentations $n = 3$. Si les résultats existent déjà dans le vecteur, les poids se cumulent.

5 - Contextualisation : a pour but de garder les termes ajoutés les plus pertinents, en fonction du contexte textuel – Chapuis & Lafourcade (2017) pour une approche similaire –. L'algorithme prend en entrée le vecteur construit jusqu'à l'étape 4, et le considère comme un ensemble de termes pondérés $S = \{t_1/w_1 ; ...t_n/w_n\}$. On normalise les poids de manière à ce que le plus élevé soit égal à 1 ; on construit S' de la même manière. On crée un graphe en ajoutant toutes les relations possibles entre les éléments de $S \cup S'$ trouvées dans JDM (celles-ci peuvent être positives, donc vraies, ou négatives, fausses). On passe à la phase d'initialisation. À chaque terme de départ (de S), on assigne une valeur d'activation a égale à son poids dans S. Les termes de S' ont une valeur d'activation de 0. Vient ensuite l'étape de propagation : on propage la valeur d'activation de chaque nœud N à chaque voisin de S', via la relation R d'un poids $w(R)$ tel que $a(S'_i) \leftarrow (a(S'_i) \times a(N) \times w(R))^{1/3}$. Un nœud N est donc considéré comme un neurone qui transmet son activation $a(N)$ si celle-ci est au dessus

d'un certain seuil (empiriquement, ce seuil est de 0.5). L'étape suivante, d'itération, ajoute à la valeur d'activation de chaque terme initial (de S) son poids dans S, avant de repasser à l'étape de propagation. On répète jusqu'à convergence des poids ou jusqu'au maximum d'étapes autorisées. L'algorithme est prouvé non convergent en général, mais converge dans les cas où il n'y a pas d'interprétations multiples au texte d'entrée. En contextualisant notre exemple, on obtient :

Rafinements gardés	**Rafinements écartés**	voiture>automobile>jouet: -284
voiture>automobile : 228	voiture>train: -182	piéton>facteur: -333
piéton>personne se déplaçant à pied: 98	voiture>véhicule de transport à roues: -204	piéton>soldat: -333
danger>péril: 74	voiture>mode de transport: -284	danger>marine: -343
		danger>inconvénient: -343

Par exemple, on trouve dans le graphe le chemin *voiture>automobile* → *voiture autonome* → *danger>péril*, ce qui permet de renforcer ces sélections. Plus des mots ont des liens, plus ils sont renforcés et sont susceptibles d'être identifiés comme les sens les plus probables. Cependant, cette méthode est dépendante des informations présentes dans la base de connaissances.

3.4 Exemple

En partant de la phrase "les voitures autonomes sont un danger pour les piétons" et en y appliquant toutes les étapes décrites ci-dessus nous obtenons :

```
{voitures_autonomes:0.25, automobile:0.25, voiture_autonome:0.25, danger:0.25
;voiture>automobile:0.25, voiture:0.25, être un danger:0.25, piéton>personne se
déplaçant à pied:0.25, risque:0.17, inquiétude:0.15, menace:0.12, difficulté:0.12,
piéton:0.25, passant:0.23, marcheur:0.23, individu:0.12 personne:0.12;
véhicule:0.17, véhicule terrestre:0.23, accident de la circulation:0.95 ;écraser un
piéton:0.23, renverser un piéton:0.23, accident:0.19, rue:0.16
```

On notera que les termes composés sont reconnus comme tels (*voitures autonomes* et *être un danger*) et que, à l'aide de la contextualisation, le verbe *piéter* a été écarté comme lemme pour *piétons*.

4 Évaluation et discussion

Les données du projet, correspondant à 6481 arguments de 77 débats, nous ont permis d'évaluer notre méthode : en premier lieu, au niveau des temps de calcul des vecteurs, sur un ordinateur personnel (16 GO de RAM, processeur 2,2 GHz Intel Core i7 quatre cœurs). Avec la baseline (étapes de régularisation), il faut **0.339s** pour calculer les 6481 vecteurs. En ajoutant les étapes sémantiques, on a au total **5159.42s**, moyen **0.79s**, minimal **4 10^{-5}s** et maximal **6.44s** (correspond à une intervention très longue (355 mots), particulièrement rare lors des débats). Les points les plus coûteux sont les appels à la base de connaissance et la contextualisation (**0.58s**). La base de connaissance étant codée sous forme de base de données indexée, sa taille n'a que peu d'influence sur la complexité, c'est le nombre de requêtes effectuées qui joue.

L'évaluation qualitative sans corpus de référence est délicate. Nous avons donc créé notre propre Gold Standard : pour chaque débat et chaque argument du débat, nous avons ordonné les 5 plus proches. Il est donc ensuite possible de comparer avec les résultats de notre méthode mais également avec les vecteurs calculés de façon alternative, ici en particulier selon la méthode de Bojanowski *et al.* (2017) qui utilise FastText (dimension de 300 et modèle skip-gram). Après adaptation (somme vectorielle normée des vecteurs de chaque terme des segments textuels, termes pour lesquels un vecteur existe), nous avons pu l'appliquer à nos fragments textuels et comparer les résultats selon les approches.

Méthode de construction	Précision/GSM	Gain / MSs	Gain / Sans CT
Mots Simples (MSs)	42.08%		
MSs + Termes Composés (TCs)	67.83%	+25.7	
MSs + TCs + Lemmatisation (L)	78.41%	+36.2	
MSs + TCs + L + Génériques (G)	84.23%	+42.3	
MSs + TCs + L + Synonymes (S)	85.71%	+43.6	
MSs + TCs + L + G + S + Termes associés	91.31%	+49.2	
Mots Simples (MSs) + Contextualisation (CT)	92.44%		+50.2
MSs + TCs + CT	97.08%	+4.7	+29.3
MSs + TCs + Lemmatisation (L) + CT	98.31%	+6	+21.1
MSs + TCs + L + Génériques (G) + CT	98.67%	+6.3	+14.3
MSs + TCs + L + G + CT	98.64%	+6.2	+13.1
MSs + TCs + L + G + S + Sens des mots + termes associés + CT	99.92%	+7.5	+8.6
MSs + Bojanowski *et al.*	45.21%		
MSs + TCs + L + G + S + B. *et al.*	56.71%		
PWs + TCs + L + G + S + Sens des mots + A + CT + B. *et al.*	61.34%		

TABLE 1 – Pourcentage d'arguments ordonnés correctement par rapport à notre Gold Standard manuel (GSM), en fonction des différentes méthodes utilisées.

La table 1 montre la précision des calculs pour chaque pipeline et il apparaît clairement que l'augmentation et la contextualisation ont un effet très positif sur la qualité des vecteurs (et donc du rapprochement thématique). Les cas d'échec ont systématiquement été identifiés comme des relations manquantes dans la base de connaissance, qui peut être complétée de façon rapide. Enfin, la forte lexicalisation de la base induit que le rapprochement est souvent sémantique, en particulier selon l'usage ou non de la négation (notamment, car les formes verbales négatives sont présentes dans la base). Par exemple, les segments "une voiture autonome peut ne pas polluer", "la voiture autonome est verte" et "la voiture intelligente est écologique" sont identifiés comme très proches.

Nous avons également comparé notre approche avec de la méthode de Bojanowski *et al.*, dont les résultats sont assez proches de notre baseline, ce qu'on peut expliquer par le fait que le plongement de termes isolés n'est pas équivalent à l'augmentation + contextualisation. Par ailleurs, il est possible que l'entraînement en utilisant Wikipédia ne soit pas adéquat pour un débat en ligne, où les arguments sont assez spontanés.

5 Conclusion et travail futur

En nous écartant des méthodes actuelles de plongements de termes, nous avons mis au point un calcul de vecteurs à la volée qui s'appuie fortement sur la base de connaissances JeuxDeMots ainsi qu'une procédure de contextualisation thématique. Nos résultats sont adaptables et explicables, et exhibe une ambiguïté sémantique réduite. Les ressources machines nécessaires pour les calculs sont raisonnables (l'ensemble tourne sur un ordinateur de bureau classique) et nous sommes proches du temps réel (sous la seconde), ce qui était requis. Nous avons effectivement testé notre méthode dans les classes lors du projet AREN, et obtenus des retours très positifs sur la qualité des rapprochements thématiques obtenus. De plus, si le corpus d'arguments utilisé pour tester notre méthode n'est actuellement pas disponible au public, il sera anonymisé et rendu disponible à la fin du projet.

Bien que n'étant pas un objectif premier, le graphe de relations produit par la contextualisation est lisible par un être humain et peut être utilisé pour expliquer le résultat exprimé ; fonctionnalité d'explication qui sera ajoutée dans la plate-forme. Nous comptons également améliorer plus avant notre précision, en testant l'exploitation d'autres informations de JeuxDeMots et élargir notre cercle d'utilisateurs à la société civile pour construire un Gold Standard plus large et librement accessible.

Références

ABDALGADER K. & SKABAR A. (2011). Short-text similarity measurement using word sense disambiguation and synonym expansion. In J. LI, Éd., *AI 2010 : Advances in Artificial Intelligence*, p. 435–444, Berlin, Heidelberg : Springer Berlin Heidelberg.

AHN L. V. (2006). Games with a purpose. *Computer*, **39**(6), 92–94. DOI : 10.1109/MC.2006.196.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

CHAPUIS M. & LAFOURCADE M. (2017). Identifying Polysemous Words and Inferring Sense Glosses in a Semantic Network. In *IWCS : International Conference on Computational Semantics*, Montpellier, France. HAL : lirmm-01763423.

COLLINS A. M. & QUILLIAN M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **8**(2), 240 – 247. DOI : 10.1016/S0022-5371(69)80069-1.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EIN DOR L., MASS Y., HALFON A., VENEZIAN E., SHNAYDERMAN I., AHARONOV R. & SLONIM N. (2018). Learning thematic similarity metric from article sections using triplet networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 49–54, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-2009.

ESPINOSA ANKE L., SCHOCKAERT S. & WANNER L. (2019). Collocation classification with unsupervised relation vectors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5765–5772, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1576.

GAUME B., DUVIGNAU K. & VANHOVE M. (2007). Semantic associations and confluences in paradigmatic networks. In M. VANHOVE, Éd., *From polysemy to semantic change - towards a typology of lexical semantic associations* : John Benjamins Publishing Company. DOI : 10.1075/slcs.106.11gau, HAL : hal-01321894.

GONG H., SAKAKINI T., BHAT S. & XIONG J. (2018). Document similarity for texts of varying lengths via hidden topics. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2341–2351, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1218.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing*, Pattaya, Chonburi, Thailand. HAL : lirmm-00200883.

LAFOURCADE M., MERY B., MIRZAPOUR M., MOOT R. & RETORÉ C. (2018). Collecting Weighted Coercions from Crowd-Sourced Lexical Data for Compositional Semantic Analysis. In *isAI : International Symposium on Artificial Intelligence*, volume LNCS de *New Frontiers in Artificial Intelligence*, p. 214–230, Tokyo, Japan. DOI : 10.1007/978-3-319-93794-6_15, HAL : lirmm-01916209.

LE Q. V. & MIKOLOV T. (2014). Distributed Representations of Sentences and Documents. *arXiv e-prints*, p. arXiv :1405.4053. arXiv : 1405.4053.

LIU Y., LIU Z., CHUA T.-S. & SUN M. (2015). Topical word embeddings. In *Proceedings AAAI Conference on Artificial Intelligence*.

MAGERMAN T., VAN LOOY B., BAESENS B. & DEBACKERE K. (2011). Assessment of latent semantic analysis (lsa) text mining algorithms for large scale mapping of patent and scientific publication documents. *Katholieke Universiteit Leuven Department of Mangerial Economics Strategy and Innovation, Working Paper 1114*, p. 1–7.

MELAMUD O., GOLDBERGER J. & DAGAN I. (2016). context2vec : Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, p. 51–61, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-1006.

PANIGRAHI A., SIMHADRI H. V. & BHATTACHARYYA C. (2019). Word2Sense : Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5692–5705, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1570.

PARK S., BYUN J., BAEK S., CHO Y. & OH A. (2018). Subword-level word vector representations for Korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2429–2438, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1226.

PEINELT N., LIAKATA M. & NGUYEN D. (2019). Aiming beyond the obvious : Identifying non-obvious cases in semantic similarity datasets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2792–2798, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1268.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1202.

POLGUÈRE A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, **27**(4), 396—418.

SMALHEISER N. R., COHEN A. M. & BONIFIELD G. (2019). Unsupervised low-dimensional vector representations for words, phrases and text that are transparent, scalable, and produce similarity metrics that are not redundant with neural embeddings. *Journal of Biomedical Informatics*, **90**, 103096. DOI : https ://doi.org/10.1016/j.jbi.2019.103096.

SOWA J. F. & ZACHMAN J. A. (1992). Extending and formalizing the framework for information systems architecture. *IBM Systems Journal*, **31**(3), 590–616. DOI : 10.1147/sj.313.0590.

SUMATHY M. K. L. & CHIDAMBARAM D. (2016). A hybrid approach for measuring semantic similarity between documents and its application in mining the knowledge repositories. *International Journal of Advanced Computer Science and Applications*, **7**(8).

THONGTAN T. & PHIENTHRAKUL T. (2019). Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 407–414, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-2057.

VIJAYMEENA1 M. & KAVITHA K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications, Machine Learning and Applications : An International Journal (MLAIJ)*, **3(1)**, 19–28.

ZHENG T., GAO Y., WANG F., FAN C., FU X., LI M., ZHANG Y., ZHANG S. & MA H. (2019). Detection of medical text semantic similarity based on convolutional neural network. *BMC Medical Informatics and Decision Making*, **19**(1), 156. DOI : 10.1186/s12911-019-0880-2.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français

Hang Le[1] Loïc Vial[1] Jibril Frej[1] Vincent Segonne[2] Maximin Coavoux[1]
Benjamin Lecouteux[1] Alexandre Allauzen[3] Benoît Crabbé[2] Laurent Besacier[1]
Didier Schwab[1]
(1) Univ. Grenoble Alpes, CNRS, LIG
(2) Université Paris Diderot
(3) E.S.P.C.I, CNRS LAMSADE, PSL Research University
{thi-phuong-hang.le, loic.vial, jibril.frej}@univ-grenoble-alpes.fr
{maximin.coavoux, benjamin.lecouteux, laurent.besacier, didier.schwab}@univ-grenoble-alpes.fr
{vincent.segonne@etu, bcrabbe@linguist}.univ-paris-diderot.fr, alexandre.allauzen@espci.fr

RÉSUMÉ

Les modèles de langue pré-entraînés sont désormais indispensables pour obtenir des résultats à l'état-de-l'art dans de nombreuses tâches du TALN. Tirant avantage de l'énorme quantité de textes bruts disponibles, ils permettent d'extraire des représentations continues des mots, contextualisées au niveau de la phrase. L'efficacité de ces représentations pour résoudre plusieurs tâches de TALN a été démontrée récemment pour l'anglais. Dans cet article, nous présentons et partageons FlauBERT, un ensemble de modèles appris sur un corpus français hétérogène et de taille importante. Des modèles de complexité différente sont entraînés à l'aide du nouveau supercalculateur *Jean Zay* du CNRS. Nous évaluons nos modèles de langue sur diverses tâches en français (classification de textes, paraphrase, inférence en langage naturel, analyse syntaxique, désambiguïsation automatique) et montrons qu'ils surpassent souvent les autres approches sur le référentiel d'évaluation FLUE également présenté ici.

ABSTRACT

FlauBERT : Unsupervised Language Model Pre-training for French.

Language models have become a key step to achieve state-of-the art results in many NLP tasks. Leveraging the huge amount of unlabeled texts available, they provide an efficient way to pre-train continuous word representations that can be fine-tuned for downstream tasks, along with their contextualization at the sentence level. This has been widely demonstrated for English. In this paper, we introduce and share FlauBERT, a model learned on a very large and heterogeneous French corpus. We train models of different sizes using the new CNRS *Jean Zay* supercomputer. We apply our French language models to several NLP tasks (text classification, paraphrasing, natural language inference, parsing, word sense disambiguation) and show that they often outperform other pre-training approaches on the FLUE benchmark also presented in this article.

MOTS-CLÉS : FlauBERT, FLUE, BERT, français, modèles de langue, évaluation, classification de textes, analyse syntaxique, désambiguïsation lexicale, inférence en langue naturelle, paraphrase.

KEYWORDS: FlauBERT, FLUE, BERT, French, language model, NLP benchmark, text classification, parsing, word sense disambiguation, natural language inference, paraphrase.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 268–278.
HAL : hal-02784776.

1 Introduction

En 2018, l'introduction de représentations linguistiques profondes contextuelles, obtenues à partir de textes bruts, a conduit à un changement de paradigme pour plusieurs tâches du TALN. Alors que les approches fondées sur des représentations continues telles que word2vec (Mikolov *et al.*, 2013) ou GloVe (Pennington *et al.*, 2014) apprennent un vecteur unique pour chaque mot, les modèles introduits récemment produisent des *représentations contextuelles* qui dépendent de la séquence de mots d'entrée complète. Initialement fondées sur des réseaux neuronaux récurrents (Dai & Le, 2015; Ramachandran *et al.*, 2017; Howard & Ruder, 2018; Peters *et al.*, 2018), ces approches ont peu à peu intégré des modèles *Transformer* (Vaswani *et al.*, 2017) comme c'est le cas pour GPT (Radford *et al.*, 2018), BERT (Devlin *et al.*, 2019), XLNet (Yang *et al.*, 2019b), RoBERTa (Liu *et al.*, 2019), ALBERT (Lan *et al.*, 2019) et T5 (Raffel *et al.*, 2019). L'utilisation de ces modèles pre-entraînés a permis des avancées de l'état-de-l'art pour de nombreuses tâches du TALN. Cependant, ceci a surtout été montré pour l'anglais, même si des variantes multilingues sont également disponibles, prenant en compte plus d'une centaine de langues dans un seul modèle : mBERT (Devlin *et al.*, 2019), XLM (Lample & Conneau, 2019), XLM-R (Conneau *et al.*, 2019). Dans cet article [1], nous décrivons notre méthodologie pour construire et partager FlauBERT (**F**rench **La**nguage **U**nderstanding via **B**idirectional **E**ncoder **R**epresentations from **T**ransformers), un modèle BERT pour la français. FlauBERT surpasse le modèle multilingue mBERT dans plusieurs tâches. Nous proposons aussi un référentiel d'évaluation nommé FLUE (**F**rench **L**anguage **U**nderstanding **E**valuation) similaire au benchmark GLUE (Wang *et al.*, 2018) pour l'anglais. FlauBERT et FLUE sont disponibles en ligne pour la communauté TALN. [2]

Étant donné l'impact des modèles contextuels pre-entrainés, plusieurs auteurs ont récemment publié des modèles disponibles pour d'autres langues que l'anglais. Par exemple, ELMo (Peters *et al.*, 2018, ELMo) existe pour le portugais, le japonais, l'allemand et le basque, [3] tandis que BERT a été récemment entraîné pour plusieurs langues (allemand, chinois, espagnol, finnois, italien, néerlandais, suédois). [4] Pour le français, en parallèle du modèle que nous proposons, une équipe jointe INRIA et Facebook a développé CamemBERT (Martin *et al.*, 2019). Une autre tendance considère des modèles estimés sur plusieurs langues avec un vocabulaire commun, comme par exemple une version de BERT multilingue pour 104 langues [5]. Mentionnons également les travaux récents utilisant des données parallèles comme LASER (Artetxe & Schwenk, 2019) pour 93 langues, XLM (Lample & Conneau, 2019) et XLM-R (Conneau *et al.*, 2019) pour 100 langues.

Par ailleurs, l'existence d'un référentiel d'évaluation tel que GLUE (Wang *et al.*, 2018) pour l'anglais est très utile pour stimuler des recherches reproductibles. Les bonnes performances obtenues avec des modèles contextuels pre-entrainés sur la plupart des tâches de TALN couvertes par GLUE ont conduit à son extension. SuperGLUE (Wang *et al.*, 2019) est un nouveau référentiel construit sur les mêmes principes, incluant un ensemble de tâches plus difficiles et variées. Une version chinoise de GLUE [6] est aussi développée pour évaluer la performance du modèle sur cette langue. À ce jour, nous n'avons pas connaissance d'un tel référentiel pour le français, d'où la proposition détaillée en section 3.

1. Cet article est une version traduite et raccourcie de l'article de Le *et al.* (2019), accepté à LREC 2020.
2. https://github.com/getalp/Flaubert
3. https://allennlp.org/elmo
4. Une liste de modèles, en constante évolution, est disponible sur https://huggingface.co/models
5. https://github.com/google-research/bert
6. https://github.com/chineseGLUE/chineseGLUE

2 Apprentissage du modèle FlauBERT

Données d'apprentissage et pré-traitements Nous agrégeons 24 sous-corpus de types divers (wikipedia, livres, *Common Crawl*, ...). Nos trois sources principales sont (1) les textes monolingues des campagnes d'évaluation WMT19 (Li *et al.*, 2019, 4 sous-corpus), (2) les textes en français de la collection OPUS (Tiedemann, 2012, 8 sous-corpus), (3) le projet Wikimedia [7] (8 sous-corpus). La taille totale (non compressée) des textes ainsi agrégés est de 270GB. Après un prétraitement consistant en différents filtrages (enlever les phrases très courtes, les séquences de numéros ou d'adresses électroniques, etc.), une normalisation de l'encodage des caractères, et une tokenisation à l'aide de Moses (Koehn *et al.*, 2007), nous obtenons un corpus de 71GB. Notre code pour télécharger et pré-traiter les données est publiquement disponible. [8]

	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	CamemBERT	$\text{FlauBERT}_{\text{BASE}}$/$\text{FlauBERT}_{\text{LARGE}}$
Langue	Anglais	Anglais	Français	Français
Données d'apprentissage	13 GB	160 GB	138 GB†	71 GB‡
Objectifs de pré-entraînement	NSP et MLM	MLM	MLM	MLM
Nombre total de paramètres	110 M	125 M	110 M	138 M/ 373 M
Tokenisation	WordPiece 30K	BPE 50K	SentencePiece 32K	BPE 50K
Masque	Statique + sous-mots	Dynamique + sous-mots	Dynamique + mot entier	Dynamique + sous-mot

†, ‡ : 282 GB, 270 GB before filtering/cleaning.

TABLE 1 – Comparaison entre FlauBERT et d'autres modèles de langue pré-entraînés.

Objectif de l'entraînement et optimisation Le pré-entraînement du Bert original consiste en deux tâches supervisées : (1) un *modèle de langue masqué* (MLM) qui apprend à prédire des jetons masqués de façon aléatoire ; et (2) une *prédiction de la prochaine phrase* (NSP - *Next Sentence Prédiction*) dans laquelle le modèle apprend à prédire si `B` est une phrase qui suit effectivement `A`, étant donné une paire de phrases d'entrée `A,B`.

Devlin *et al.* (2019) a observé que la suppression de NSP nuit considérablement aux performances sur certaines tâches. Cependant, le contraire a été démontré dans des études ultérieures, notamment Yang *et al.* (2019b, XLNet), Lample & Conneau (2019, XLM), et Liu *et al.* (2019, RoBERTa). [9] Par conséquent, nous avons utilisé seulement l'objectif MLM dans FlauBERT.

Pour optimiser cette fonction objectif, nous avons suivi Liu *et al.* (2019) et utilisé l'optimiseur Adam (Kingma & Ba, 2014) avec les paramètres suivants :

— $\text{FlauBERT}_{\text{BASE}}$: étapes de mise en route (ou *warm up*) de 24k, taux d'apprentissage maximal de $6\text{e}{-4}$, $\beta_1 = 0,9$, $\beta_2 = 0,98$, $\epsilon = 1\text{e}{-6}$ et perte de poids de 0,01.

— $\text{FlauBERT}_{\text{LARGE}}$: étapes de mise en route de 30k, taux d'apprentissage maximal de $3\text{e}{-4}$, $\beta_1 = 0,9$, $\beta_2 = 0,98$, $\epsilon = 1\text{e}{-6}$ et perte de poids de 0,01.

Modèles et configuration d'apprentissage Nous utilisons la même architecture que BERT (Devlin *et al.*, 2019). Un vocabulaire de 50K unités sous-lexicales est construit en utilisant l'algorithme *Byte Pair Encoding* (Sennrich *et al.*, 2016, BPE). Nous entraînons deux principaux modèles (transformers

7. https://meta.wikimedia.org/w/index.php?title=Data_dumps&oldid=19312805
8. https://github.com/getalp/Flaubert
9. Liu *et al.* (2019) ont émis l'hypothèse que l'implantation originale de BERT pourrait avoir supprimé la fonction de coût associée au NSP tout en conservant le format d'entrée consistant en des paires de phrases.

bi-directionnels) : FlauBERT$_{\text{BASE}}$ (12 blocs de dimension cachée 768, 12 têtes pour l'attention) et FlauBERT$_{\text{LARGE}}$ (24 blocs de dimension cachée 1024, 12 têtes). Le critère d'apprentissage est de type *masked language model* : il consiste à prédire des tokens d'une phrase ayant été préalablement et aléatoirement masqués. FlauBERT$_{\text{BASE}}$ est appris sur 32 GPU Nvidia V100 SXM2 32 GB en 410h et FlauBERT$_{\text{LARGE}}$ est appris sur 128 de ces mêmes GPU en 390h.

3 FLUE

Le référentiel d'évaluation FLUE est composé de 7 tâches correspondant à différents niveaux d'analyse (syntaxique, sémantique) du traitement automatique du français.

Classification de texte Le corpus d'analyse de sentiments translingue CLS (Prettenhofer & Stein, 2010) est constitué de critiques issues du site Amazon pour trois catégories de produits (livres, DVD et musique) en quatre langues : anglais, français, allemand et japonais. Chaque échantillon contient une critique associée à une note allant de 1 à 5. Suivant Blitzer *et al.* (2006) et Prettenhofer & Stein (2010), les évaluations avec 3 étoiles sont écartées et la note est binarisée avec un seuil de 3. Pour chaque catégorie de produit, nous construisons des ensembles d'apprentissage et de test qui sont équilibrés. Les données de test contiennent ainsi 2000 avis en français.

Identification de paraphrases Cette tâche consiste à identifier si des paires de phrases sont sémantiquement équivalentes ou non. PAWS-X est un ensemble de données multilingues pour l'identification des paraphrases (Yang *et al.*, 2019a). Il s'agit de l'extension de la tâche PAWS (Zhang *et al.*, 2019) pour l'anglais à six autres langues : français, espagnol, allemand, chinois, japonais et coréen. Yang *et al.* (2019a) ont utilisé la traduction automatique pour créer les corpus de ces autres langues mais les ensembles de développement et de test pour chaque langue sont traduits manuellement. Nous prenons à nouveau la partie française pour FLUE.

Natural Language Inference (NLI) Cette tâche, également connue sous le nom de reconnaissance d'implications textuelles (RTE), considère une prémisse (p) et une hypothèse (h) et consiste à déterminer si p implique, contredit ou n'implique ni ne contredit h. Le corpus *Cross-lingual NLI Corpus* (Conneau *et al.*, 2018, XNLI) étend l'ensemble de développement et de test du corpus *Multi-Genre Natural Language Inference* corpus (Williams *et al.*, 2018, MultiNLI) à 15 langues. Les ensembles de développement et de test pour chaque langue consistent en 7 500 exemples annotés manuellement, soit un total de 112 500 paires de phrases annotées avec les étiquettes *entailment*, *contradiction* ou *neutre*. FLUE intègre la partie française de ce corpus.

Analyse syntaxique et étiquetage morphosyntaxique Nous considérons deux tâches d'analyse syntaxique : analyse en constituants et en dépendances, ainsi que l'étiquetage morphosyntaxique. Pour cela, nous utilisons le *French Treebank* (Abeillé *et al.*, 2003), une collection de phrases du *Monde* annotées manuellement en constituants et dépendances syntaxiques. Nous utilisons la version de ce corpus de la campagne d'évaluation SPMRL 2014 décrite par Seddah *et al.* (2013), qui contient 14759, 1235 et 2541 phrases pour respectivement l'entraînement, le développement et l'évaluation.

Désambiguïsation lexicale des verbes et des noms La désambiguïsation lexicale consiste à assigner un sens, parmi un inventaire donné, à des mots d'une phrase. Pour la désambiguïsation lexicale de verbes, nous utilisons les données de FrenchSemEval (Segonne *et al.*, 2019). Il s'agit d'un corpus d'évaluation dont les occurrences de verbes ont été annotées manuellement avec les sens de Wiktionary. [10] Pour la désambiguïsation lexicale des noms, nous utilisons la partie française de la tâche de désambiguïsation multilingue de SemEval 2013 (Navigli *et al.*, 2013). Nous adaptons l'inventaire de sens de BabelNet utilisé par Navigli & Ponzetto (2010) pour WordNet 3.0 (Miller, 1995), en convertissant les étiquettes de sens lorsqu'une projection est présente dans BabelNet, et en les supprimant dans le cas contraire. Ce processus de conversion donne un corpus d'évaluation composé de 306 phrases et 1 445 mots français annotés en sens WordNet, et vérifiés manuellement. Les données d'apprentissage sont obtenues par transfert selon la méthode décrite par Hadj Salah (2018), qui consiste à traduire des corpus annotés en sens puis transférer leurs annotations. Nous rendrons disponibles à la fois nos données d'entraînement et d'évaluation.

4 Expériences et résultats

Dans cette section, nous présentons les résultats de FlauBERT sur le référentiel d'évaluation FLUE. Nous comparons les performances de FlauBERT avec BERT multilingue (Devlin *et al.*, 2019, mBERT) et CamemBERT (Martin *et al.*, 2019) sur toutes les tâches. Nous comparons également avec le meilleur modèle non contextuel pour chaque tâche. Nous utilisons les bibliothèques open-source XLM (Lample & Conneau, 2019) et Transformers (Wolf *et al.*, 2019). Nous renvoyons à Le *et al.* (2019) pour une description détaillée des expériences.

Classification de texte Nous avons suivi le processus de réglage fin (*fine tuning*) standard de BERT (Devlin *et al.*, 2019). Le bloc de classification ajouté au dessus du model BERT est composé des couches suivantes : dropout, linéaire, activation $\tanh$, dropout et linéaire. Les dimensions de sortie des couches linéaires sont respectivement égales à la taille cachée du Transformer et au nombre de classes (2). La valeur de dropout a été fixée à 0.1. Nous entraînons le modèle pendant 30 époques, par lots de 8 exemples. Nous testons 4 valeurs de *learning rate* ($1e{-}5$, $5e{-}5$, $1e{-}6$ et $5e{-}6$). Nous utilisons comme ensemble de validation un échantillon aléatoire de 20% des données, pour sélectionner le meilleur modèle. Le tableau 2 présente l'exactitude finale sur l'ensemble de test pour chaque modèle. Les résultats mettent en évidence l'importance d'un modèle monolingue en français pour la classification des textes : CamemBERT et $\text{FlauBERT}_{\text{BASE}}$ surpassent largement mBERT.

Identification de paraphrases La configuration de cette tâche est presque identique à la précédente, la seule différence étant que la séquence d'entrée est maintenant une paire de phrases `A, B`. La performance finale de chaque modèle est indiquée dans le tableau 2. On peut observer que notre modèle monolingue français ne fonctionne que légèrement mieux qu'un modèle multilingue (mBERT), ce qui pourrait être attribué aux caractéristiques de l'ensemble de données PAWS-X. En effet, cet ensemble de données contient des paires de phrases avec une forte proportion de chevauchement lexical, ce qu'un modèle multilingue peut détecter aussi bien qu'un modèle monolingue.

10. Version du 20-04-2018 incluse dans le jeu de donnée.

Tâche	Classification			Paraphrase	NLI	Constituants		Dépendances		Désambiguïsation	
Section	Livres	DVD	Musique							Noms	Verbes
Mesure	Acc.	Acc.	Acc.	Acc.	Acc.	F_1	POS	UAS	LAS	F_1	F_1
État de l'art ant.	91.25[c]	89.55[c]	93.40[c]	66.2[d]	80.1/**85.2**[e]	87.4[a]		89.19[b]	85.86[b]	-	43.0[h]
Sans pré-entr.	-	-	-			83.9	97.5	88.92	85.11	50.0	-
FastText	-	-	-			83.6	97.7	86.32	82.04	49.4	34.9
mBERT	86.15[c]	86.9[c]	86.65[c]	89.3[d]	76.9[f]	87.5	98.1	89.5	85.86	56.5	44.9
CamemBERT	**93.40**	**92.70**	94.15	89.8	81.2	88.4	**98.2**	91.37	88.13	56.1	**51.1**
FlauBERT$_{\text{BASE}}$	**93.40**	92.50	**94.30**	**89.9**	81.3	**89.1**	98.1	**91.56**	**88.35**	54.9/**57.9**[g]	47.4

TABLE 2 – Résultats finaux sur les tâches de FLUE. [a]Kitaev *et al.* (2019). [b]Constant *et al.* (2013). [c]Eisenschlos *et al.* (2019, MultiFiT). [d]Chen *et al.* (2017, ESIM). [e]Conneau *et al.* (2019, XLM-F BASE/LARGE). [f]Martin *et al.* (2019). [g]Utilise FlauBERT$_{\text{LARGE}}$. [h]Segonne *et al.* (2019).

Natural Language Inference (NLI) Comme cette tâche a également été considérée par Martin *et al.* (2019, CamemBERT), nous utilisons la même configuration expérimentale pour que nos résultats soient comparables. L'entrée du modèle pour cette tâche est aussi une paire de phrases. Nous présentons la performance pour chaque modèle dans le tableau 2. Les résultats confirment la supériorité des modèles français par rapport aux modèles multilingues (mBERT) pour cette tâche. FlauBERT$_{\text{BASE}}$ fonctionne légèrement mieux que CamemBERT. Les deux dépassent clairement XLM-R$_{\text{BASE}}$, bien qu'ils ne puissent pas dépasser XLM-R$_{\text{LARGE}}$. Il convient de noter que XLM-R$_{\text{LARGE}}$ employait une architecture beaucoup plus profonde.

Analyse syntaxique en constituants et étiquetage morphosyntaxique Nous réalisons de manière conjointe l'analyse en constituants et l'étiquetage morphosyntaxique, à l'aide de l'analyseur [11] décrit par Kitaev & Klein (2018) et Kitaev *et al.* (2019). La table 2 présente les résultats. Sans pré-entraînement, nous reproduisons le résultat de Kitaev & Klein (2018). FastText n'améliore pas les résultats. CamemBERT obtient un meilleur résultat que mBERT, grâce à son entraînement monolingue. FlauBERT obtient un encore meilleur résultat (+0.7). Les trois analyseurs utilisant un modèle de langue contextuel obtiennent des résultats similaires en étiquetage morphosyntaxique (98.1-98.2).

Analyse syntaxique en dépendances Pour l'analyse en dépendances, nous utilisons une réimplémentation de l'algorithme de Dozat & Manning (2017) avec décodage par arbre couvrant de poids maximal. L'analyseur prend en entrée des phrases étiquetées en partie du discours. Nous utilisons les tags prédits fournis par la campagne SPMRL. Notre représentation lexicale est une concaténation de plongements lexicaux et de plongements de tags appris avec le reste du modèle d'analyse sur le French Treebank ainsi que d'un vecteur préentrainé. Les résultats sont donnés en table 2. Tous les modèles utilisant les vecteurs BERT font au moins aussi bien que l'état de l'art sur cette tâche et les deux modèles monolingues sont état de l'art avec les vecteurs FlauBERT$_{\text{BASE}}$ qui donnent un résultat marginalement meilleur que les vecteurs CamemBERT. On remarque également que les deux modèles monolingues apportent des résultats substantiellement meilleurs que le modèle mBERT multilingue.

Désambiguïsation lexicale des noms Nous utilisons le réseau de neurones décrit par Vial *et al.* (2019a,b) dont le code est fourni. [12] Il prend, en entrée, les vecteurs issus d'un modèle de langue

11. https://github.com/nikitakit/self-attentive-parser
12. https://github.com/getalp/disambiguate

pré-entraîné, qui restent fixes, puis il est composé de plusieurs couches d'encodeur *Transformer* et d'une couche linéaire qui sont entraînées. La couche linéaire réalise une projection sur l'ensemble des *synsets* vus pendant l'entraînement. Enfin, le *synset* qui obtient le plus haut score est choisi. Nous donnons le résultat issu d'un ensemble de 8 modèles entraînés indépendamment, qui moyenne la sortie du *softmax*. Dans la Table 2, on observe d'abord des performances largement meilleures avec les modèles BERT qu'avec des vecteurs statiques. mBERT obtient de meilleures performances que CamemBERT ainsi que FlauBERT$_{\text{BASE}}$, ce que nous pensons être dû à la nature translingue des corpus d'entraînement, mais FlauBERT$_{\text{LARGE}}$ obtient les meilleurs résultats sur la tâche.

Désambiguïsation lexicale des verbes Nous suivons la méthode décrite par Segonne *et al.* (2019). Nous utilisons les plongements contextuels fournis par les modèles FlauBERT/mBERT/CamemBERT pour les representations vectorielles des occurences (l'inventaire de sens et données d'évaluation). Nous comparons également nos résultats à une représentation plus simple qui consiste à moyenner les plongements lexicaux des mots entourant le mot cible. Pour cette expérience nous avons utilisé les plongements lexicaux issus de FastText avec une fenêtre de mots de taille 5. Les résultats de nos expériences sont présentés dans la table 2. On observe que l'utilisation des modèles BERT pour cette tâche apporte un gain conséquent par rapport à l'état de l'art, les meilleurs résultats étant obtenus par CamemBERT. De plus, nos expériences confirment l'intérêt des modèles spécifiquement entraînés sur le français puisque les deux modèles CamemBERT et FlauBERT$_{\text{BASE}}$ surpassent le modèle multilingue mBERT.

5 Conclusion

Nous avons présenté et partagé FlauBERT, un ensemble de modèles de langues pre-entraînés pour le français, accompagné de FLUE, un dispositif d'évaluation. FlauBERT obtient des résultats à l'état de l'art sur un certain nombre de tâches de TALN. Il est aussi compétitif avec CamemBERT (Martin *et al.*, 2019) – un autre modèle pour le français développé en parallèle – bien qu'il ait été entraîné sur presque deux fois moins de données textuelles. Nous espérons que cette contribution stimulera les recherches sur le TALN en français. [13]

6 Remerciements

Ce travail a bénéficié du programme « Grand Challenge Jean Zay » (projet 100967) et a également été partiellement soutenu par MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). Nous remercions Guillaume Lample et Alexis Conneau pour leur soutien technique pour l'utilisation du code XLM.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Springer Netherlands : Dordrecht. DOI : 10.1007/978-94-010-0201-1_10.

13. FlauBERT est notamment disponible sur https://huggingface.co/models.

ARTETXE M. & SCHWENK H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, **7**, 597–610.

BLITZER J., MCDONALD R. & PEREIRA F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, p. 120–128 : Association for Computational Linguistics.

CHEN Q., ZHU X., LING Z.-H., WEI S., JIANG H. & INKPEN D. (2017). Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1657–1668.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

CONNEAU A., RINOTT R., LAMPLE G., WILLIAMS A., BOWMAN S., SCHWENK H. & STOYANOV V. (2018). Xnli : Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2475–2485.

CONSTANT M., CANDITO M. & SEDDAH D. (2013). The ligm-alpage architecture for the spmrl 2013 shared task : Multiword expression analysis and dependency parsing. In *Proceedings of the EMNLP Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*.

DAI A. M. & LE Q. V. (2015). Semi-supervised sequence learning. In *Advances in neural information processing systems*, p. 3079–3087.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186.

DOZAT T. & MANNING C. D. (2017). Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* : OpenReview.net.

EISENSCHLOS J., RUDER S., CZAPLA P., KARDAS M., GUGGER S. & HOWARD J. (2019). Multifit : Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543.

HADJ SALAH M. (2018). *Arabic word sense disambiguation for and by machine translation*. Theses, Université Grenoble Alpes ; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion. HAL : tel-02139438.

HOWARD J. & RUDER S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 328–339.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

KITAEV N., CAO S. & KLEIN D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3499–3505, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1340.

KITAEV N. & KLEIN D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long*

Papers), p. 2676–2686, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1249.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, p. 177–180.

LAMPLE G. & CONNEAU A. (2019). Cross-lingual language model pretraining. In *Advances in neural information processing systems*.

LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2019). Albert : A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

LI X., MICHEL P., ANASTASOPOULOS A., BELINKOV Y., DURRANI N., FIRAT O., KOEHN P., NEUBIG G., PINO J. & SAJJAD H. (2019). Findings of the first shared task on machine translation robustness. *Fourth Conference on Machine Translation (WMT19)*, p. 91–102.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2019). CamemBERT : a Tasty French Language Model. *arXiv preprint arXiv:1911.03894*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 3111–3119, USA : Curran Associates Inc.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

NAVIGLI R., JURGENS D. & VANNELLA D. (2013). SemEval-2013 Task 12 : Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 222–231.

NAVIGLI R. & PONZETTO S. P. (2010). Babelnet : Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 216–225 : Association for Computational Linguistics.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *In EMNLP*.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, p. 2227–2237.

PRETTENHOFER P. & STEIN B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 1118–1127.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). Improving language understanding by generative pre-training. *Technical report, OpenAI*.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

RAMACHANDRAN P., LIU P. & LE Q. (2017). Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 383–391.

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & VILLEMONTE DE LA CLERGERIE E. (2013). Overview of the SPMRL 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

SEGONNE V., CANDITO M. & CRABBÉ B. (2019). Using wiktionary as a resource for wsd : the case of french verbs. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, p. 259–270.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in opus. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.

VIAL L., LECOUTEUX B. & SCHWAB D. (2019a). Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL)*, Toulouse, France. HAL : hal-02092559.

VIAL L., LECOUTEUX B. & SCHWAB D. (2019b). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, Wroclaw, Poland. HAL : hal-02131872.

WANG A., PRUKSACHATKUN Y., NANGIA N., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. R. (2019). Superglue : A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WILLIAMS A., NANGIA N. & BOWMAN S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1112–1122.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M. & BREW J. (2019). Huggingface's transformers : State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

YANG Y., ZHANG Y., TAR C. & BALDRIDGE J. (2019a). Paws-x : A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

YANG Z., DAI Z., YANG Y., CARBONELL J., SALAKHUTDINOV R. & LE Q. V. (2019b). Xlnet : Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*.

ZHANG Y., BALDRIDGE J. & HE L. (2019). Paws : Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 1298–1308.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Relation, es-tu là ? Détection de relations par LSTM pour améliorer l'extraction de relations

Cyrielle Mallart[1, 2, 3] Michel Le Nouy[1] Guillaume Gravier[3] Pascale Sébillot[2]
(1) SIPA Ouest-France, 10 rue du Breil, 35000 Rennes, France
(2) INSA Rennes, IRISA, Campus de Beaulieu, 35042 Rennes, France
(3) CNRS, IRISA, Campus de Beaulieu, 35042 Rennes, France
cyrielle.mallart@ouest-france.fr, michel.lenouy@ouest-france.fr,
guig@irisa.fr, pascale.sebillot@irisa.fr

RÉSUMÉ

De nombreuses méthodes d'extraction et de classification de relations ont été proposées et testées sur des données de référence. Cependant, dans des données réelles, le nombre de relations potentielles est énorme et les heuristiques souvent utilisées pour distinguer de vraies relations de co-occurrences fortuites ne détectent pas les signaux faibles pourtant importants. Dans cet article, nous étudions l'apport d'un modèle de détection de relations, identifiant si un couple d'entités dans une phrase exprime ou non une relation, en tant qu'étape préliminaire à la classification des relations. Notre modèle s'appuie sur le plus court chemin de dépendances entre deux entités, modélisé par un LSTM et combiné avec les types des entités. Sur la tâche de détection de relations, nous obtenons de meilleurs résultats qu'un modèle état de l'art pour la classification de relations, avec une robustesse accrue aux relations inédites. Nous montrons aussi qu'une détection binaire en amont d'un modèle de classification améliore significativement ce dernier.

ABSTRACT

Relation, are you there ? LSTM-based relation detection to improve knowledge extraction

Various methods for relation extraction and classification have been proposed and benchmarked on standard academic datasets. In real-life data however, the number of potential relations is enormous and the heuristics and count-based methods often used to separate actual relations from meaningless co-occurrences fail to detect weak signals of importance. In this paper, we investigate the use of a computationally-light binary detection model to identify whether a couple of entities in a sentence bears a relation, as a preliminary step prior to relation classification. Our model is based on the shortest dependency path between two entities analyzed with a LSTM recurrent network and combined with information on the entities types. On the binary relation detection task, we achieve results better than a state-of-the-art relation classification model adapted to detection with increased robustness to relations unseen in training. We finally show that binary detection as a pre-processing step to classification of relations is effective in significantly improving the latter.

MOTS-CLÉS : extraction d'informations, détection de relations, classification de relations, LSTM, plus court chemin de dépendances.

KEYWORDS: information extraction, relation detection, relation classification, LSTM, shortest dependency path.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 279–287.
HAL : hal-02784777.

1 Motivation

Les articles de journaux sont une source particulière de données, de par leur quantité – des milliers d'articles écrits chaque jour – et le grand nombre d'entités qui interagissent dans leurs textes au sein de multiples relations. Cette masse de textes, souvent confinée aux archives, regorge d'informations, de connaissances, de détails qui peuvent aider à la mise en contexte de nouveaux événements. Afin de pouvoir exploiter une telle richesse, l'utilisation de bases de connaissances est un atout majeur. Il s'agit de remplir une base contenant des entités et les relations qu'elles entretiennent pour mettre en lien des concepts à plus grande échelle que celle d'un seul texte. Pour ce faire, de nombreux aspects de l'extraction d'informations sont mis en œuvre, un aspect crucial étant l'extraction de relations entre entités.

Une entité est un objet, soit tangible soit un concept, auquel on peut référer par un nom propre, comme « Barack Obama » ou « Paris ». Une relation est une paire d'entités issues d'une même phrase et entre lesquelles la phrase exprime un lien. Elle est spécifiée avec un nom, comme « capitale de » ou « époux de ». Extraire cette relation permet de structurer l'information, en expliquant en quoi les entités d'une phrase sont liées, si elles le sont. Ceci est un défi dans le cas de données journalistiques, où une grande masse d'entités cohabite dans les articles et peu sont réellement en relation. L'extraction doit donc être opérée de façon sûre, rapide, sans omettre les signaux faibles qui sont potentiellement porteurs d'informations inédites pour les journalistes, tout en étant capable de s'adapter au flux constant d'articles aux thématiques variées.

De nombreux extracteurs et classifieurs de relations existent déjà, issus principalement de quatre types d'approches : les schémas universels (Riedel *et al.*, 2013), l'*open information extraction* (Banko *et al.*, 2007; Mesquita *et al.*, 2013; Del Corro & Gemulla, 2013), la classification non supervisée (Hasegawa *et al.*, 2004; Wang *et al.*, 2011; Takase *et al.*, 2015), et la classification supervisée (Kambhatla, 2004; Xu *et al.*, 2015; Wang *et al.*, 2016; Cai *et al.*, 2016; Zhang *et al.*, 2018). Certaines approches combinent d'ailleurs plusieurs de ces paradigmes (Banko & Etzioni, 2008). Cependant, ces diverses méthodes ne sont souvent testées et évaluées que sur des données de référence, en anglais, propres et peu bruitées. Quand elles sont testées sur des données réelles, le nombre de relations candidates est habituellement fortement réduit par l'utilisation d'heuristiques, pour ne garder que les relations les plus fréquentes. Dans le cas du journalisme, où les relations potentielles dans les textes sont extrêmement nombreuses tout en cachant de précieux signaux faibles, il devient d'une part coûteux d'appliquer des modèles de classification sur toutes les relations candidates ; d'autre part, les signaux faibles se retrouvent noyés dans de telles quantités de données que les systèmes de classification ne les distinguent pas : des relations potentiellement cruciales sont classifiées dans de très vastes classes « poubelle », ou tout simplement ignorées.

Les systèmes existants ne peuvent donc pas être appliqués directement à notre cas d'application journalistique. Nous proposons par conséquent de procéder en deux étapes, en ajoutant un modèle préliminaire pour détecter si une phrase contient ou non une relation. Les phrases pour lesquelles un couple d'entités a été identifié comme étant bien en relation peuvent ensuite être traitées par un modèle de classification de relations. Cette division permet tout d'abord d'augmenter la qualité des données fournies au modèle de classification, en réduisant le bruit apporté par des entités sans relation, et ensuite de diminuer le temps d'exécution d'une classification de relations, puisque seule une partie des données est traitée par un modèle de classification. Nous inspirant de l'état de l'art, nous proposons un modèle simple mais efficace de détection binaire de relations, fondé sur un réseau de neurones LSTM exploitant la syntaxe des phrases, que nous testons sur des données réelles que

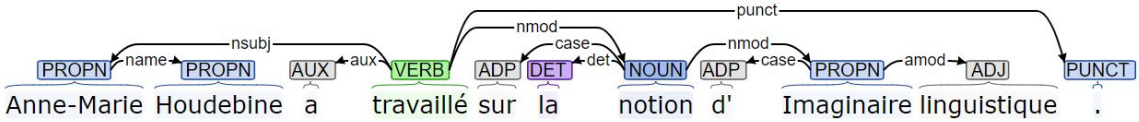


FIGURE 1 – Exemple de graphe de dépendances pour la phrase « Anne-Marie Houdebine a travaillé sur la notion d'Imaginaire linguistique. ».

nous avons collectées. Nous montrons la performance de notre modèle sur la détection de relations, ses capacités de généralisation et la possibilité qu'il offre d'améliorer les performances d'un modèle de classification de relations avec de la détection de relations en amont, démontrant ainsi la pertinence de notre approche en deux temps.

2 Méthodologie

Notre modèle doit être capable de détecter des relations au sein d'articles de style journalistique, c'est-à-dire identifier si deux entités sont ou non en relation, en analysant la phrase dans laquelle elles apparaissent. Dans chaque phrase, tous les couples d'entités potentiellement en relation (les relations candidates) sont donc extraits. Notre modèle doit apprendre si les mots se trouvant entre deux des entités signifient qu'une relation incluant ces deux entités existe ou non.

La caractéristique principale de notre modèle est le recours au plus court chemin de dépendances entre deux entités, c'est-à-dire le chemin syntaxique liant deux entités dans le graphe de dépendances qui traverse le moins de mots. Un graphe de dépendances consiste en un graphe où les nœuds correspondent aux mots, et les arêtes aux dépendances syntaxiques entre les mots (voir un exemple en figure 1). Suivant (Bunescu & Mooney, 2005), nous faisons l'hypothèse que le plus court chemin entre deux entités du graphe de dépendances contient l'essentiel de la relation, car il traverse les mots nécessaires à la compréhension de la phrase, omettant l'information auxiliaire et réduisant ainsi le bruit potentiel. Ce plus court chemin de dépendances est produit par l'algorithme de Dijkstra appliqué sur le graphe de dépendances obtenu avec le *Stanford dependency parser* (Manning *et al.*, 2014), dont nous avons modifié quelques aspects (regroupé les entités « multi-mots » (« Barack Obama » ou « université d'Harvard ») en un seul nœud, réparé les phrases auxquelles il manque un lien, et retiré les doublons créés à cause d'une erreur d'implémentation de l'analyseur). Sur ce plus court chemin de dépendances, nous collectons des variables sur chaque mot : les plongements du mot obtenu grâce à un modèle *skip-gram* pré-entraîné (Mikolov *et al.*, 2013), la catégorie morpho-syntaxique et l'étiquette de dépendance. La moyenne des plongements de mots est prise pour les entités formées de plusieurs mots. À ceci, nous ajoutons le type des entités donné par un système de détection d'entités, tel « parti politique », « ong », « personne » ou encore « autre ». Le plus court chemin exclut donc les entités elles-mêmes, car elles sont déjà prises en compte dans notre modèle *via* leur type.

Le modèle est constitué de deux branches, l'une modélisant les entités à travers leur type avec une couche cachée entièrement connectée, et l'autre prenant en compte l'information syntaxique sur le plus court chemin entre les deux entités avec une couche LSTM. Une variante du LSTM proposée dans (Graves *et al.*, 2013) est utilisée. Les deux branches sont fusionnées par un produit scalaire, puis une couche entièrement connectée dotée d'une activation *sigmoïde* prédit la probabilité qu'il y ait une relation entre les deux entités. Une telle couche de sortie a été préférée à une activation *softmax* avec deux neurones de sortie, car cette dernière approche nécessite de comparer les deux probabilités de sortie et garder la classe correspondant à la plus grande, tandis que, dans notre cas,

une unique probabilité de sortie permet d'appréhender le degré de certitude de la relation. La figure 2 présente l'architecture de notre modèle appliqué à une phrase exemple. Le choix d'intégrer les types des entités est motivé par l'observation que certaines relations n'existent pas entre certains types d'entités. Ainsi une relation potentielle « fils de » ne peut exister entre une personne et un lieu. Nous cherchons donc à confronter les mots du chemin et les types des entités pour s'assurer de leur compatibilité. Nous avons de plus séparé l'information sur les entités de l'information séquentielle par souci de simplicité et de rapidité. Nos résultats préliminaires ont en effet montré qu'entraîner une couche entièrement connectée uniquement pour les types des entités réduit le temps d'entraînement par rapport à un LSTM où les entités sont incorporées dans le plus court chemin de dépendances, pour des performances identiques.

3 Expériences

3.1 Création d'un jeu de données

Les données de référence standards pour l'extraction de relations ne sont pas compatibles avec notre volonté de détecter des relations pour des données réelles : premièrement, elles ne sont généralement pas en français ; ensuite, ces données sont déjà propres, annotées à la main, et les effectifs dans les classes de relations sont équilibrés, ce qui est éloigné des données réelles de notre problème initial.

Nous avons donc créé un jeu de données dérivé de Wikipédia, car le « style Wikipédia » est assez proche de celui des journaux d'information, les phrases sont généralement syntaxiquement correctes et de nombreux sujets sont abordés. De plus, ces données sont *open-source* et sont disponibles pour une future utilisation [1]. 200 000 articles ont ainsi été pris aléatoirement dans le *dump* français de Wikipédia en date d'avril 2019. Après nettoyage, découpage en phrases, liage des entités et appariement des paires d'entités comme relations candidates, un étiquetage distant des relations présentes dans les phrases est rendu possible grâce à la base de connaissances Wikidata, en partant du principe qu'une relation existe lorsqu'un triplet connectant les deux entités dans Wikidata existe. Le jeu de données est ensuite sous-échantillonné en termes des types d'entités, pour équilibrer les couples de types d'entités, ce qui réduit drastiquement le nombre total de relations candidates. Ce sous-échantillonnage assure qu'aucun couple de types d'entités (personne-personne, ong-lieu, etc.) ne soit majoritaire. Nous évitons ainsi que nos données soient dominées par des relations entre types « autres », ou entre lieux (communes, départements, lieux, pays), bien plus courantes dans Wikidata que les relations entre autres types (par ex. ong-parti politique). Nos modèles n'apprennent donc pas que les relations existent uniquement entre lieux et ne rejettent pas le reste des données sur ce seul critère. Nous n'équilibrons en revanche pas les types de relations, car certaines relations (par ex. « capitale de » ou « pays ») sont plus communes que d'autres (par ex. « employeur »), et certaines si peu fréquentes qu'elles sont considérées signaux faibles (par ex. « éditeur scientifique » ou « élève de »). Au final, nous obtenons 230 exemples positifs et 876 exemples négatifs, pour un total de 1106 paires. Ces données sont réparties en 55% de données d'entraînement, 20% de validation et 25% de test. Malgré cette réduction, notre jeu de données a été créé pour respecter les propriétés de signaux faibles et de déséquilibre des relations positives et négatives qui sont les défis réels que notre modèle doit relever.

1. https://github.com/CMallart/RelationDetectionFrench

3.2 Modèle de référence

Nous comparons tous nos résultats à ceux du modèle proposé par Xu *et al.* (2015). Ce modèle a été choisi comme base pour une comparaison équitable car ce système est l'état de l'art pour la classification de relations parmi les modèles ayant recours à un réseau LSTM exploitant les informations sur le plus court chemin de dépendances. Notre modèle s'inscrit dans le même esprit syntaxique, ce qui ne le rend pas fondamentalement différent de ce modèle de référence, à l'exception toutefois notable des modifications que nous avons dû faire pour prendre en compte les types des entités. L'option que nous avons retenue d'effectuer une comparaison entre des modèles similaires a été guidée par la possibilité de juger non seulement la pertinence des choix de modélisation spécifiques – tels l'ajout des types d'entités – mais surtout celle de la tâche de détection de relations par rapport à la tâche de classification.

Le modèle de Xu *et al.* est fondé sur la séparation du plus court chemin de dépendances en deux, d'un côté et de l'autre de l'ancêtre commun des deux entités sur l'arbre de dépendances, afin de prendre en compte la direction des dépendances. Le long du chemin de dépendances sont collectées la catégorie morpho-syntaxique, l'étiquette de dépendance, les plongements et les hyperonymes WordNet de chaque mot. Chaque variable est traitée par un LSTM distinct. Pour chaque variable, deux LSTM collectent l'information sur chaque moitié de chemin, puis les couches cachées de ces LSTM sont fusionnées par *max-pooling*. Les quatre canaux, un pour chaque variable, sont ensuite aussi fusionnés par *max-pooling*. Une couche entièrement connectée avec une fonction d'activation *softmax* renvoie la probabilité d'appartenir à chaque catégorie de relation.

Bien que relevant donc de la même philosophie syntaxique, notre modèle diffère du modèle précédent selon cinq aspects. Premièrement, nous voyons les dépendances syntaxiques comme un graphe plutôt qu'un arbre, ce qui permet de mieux réparer les erreurs de l'analyseur en ajoutant des arêtes. De plus, nous ne nous soucions pas de la direction des dépendances, et ne séparons donc pas le plus court chemin en deux morceaux de part et d'autre d'un ancêtre commun, afin de garder entières les expressions figées et figures de style. En troisième lieu, le plongement du mot et les catégories morpho-syntaxiques et de dépendance sont corrélées et donnent ensemble son sens global au mot ; ces variables ne sont donc pas séparées par canaux dans le LSTM. Quatrièmement, les hyperonymes WordNet, qui auraient pu être obtenus en français grâce à une traduction de la base lexicale telle que WoNeF (Pradet *et al.*, 2013), sont omis pour éviter le recours à une ressource externe qui prend du temps supplémentaire et ajoute du bruit, dû à la granularité très fine des sens de WordNet, à une incomplétude dans certains domaines et à la difficulté de choisir le sens approprié. Enfin, seul le type des entités est pris en compte, au lieu de traiter ces entités comme des mots du chemin de dépendances, ce qui permet de réduire le temps de traitement pour les entités et de généraliser.

3.3 Résultats

Nous rapportons nos résultats pour quatre expériences conduites sur les données que nous avons construites pour la tâche. Ces expériences ont pour but de prouver que notre modèle répond aux défis de l'extraction de relations dans des articles, à savoir qu'il sépare correctement les couples d'entités en relation de co-occurrences fallacieuses des autres, en étant capable de généraliser à des relations rares ou inédites, et qu'il apporte de la valeur ajoutée par rapport à une classification de relations.

Le but de notre première expérience est de vérifier qu'une architecture simple comme celle de notre modèle permet d'atteindre des performances au moins aussi élevées que le modèle plus complexe

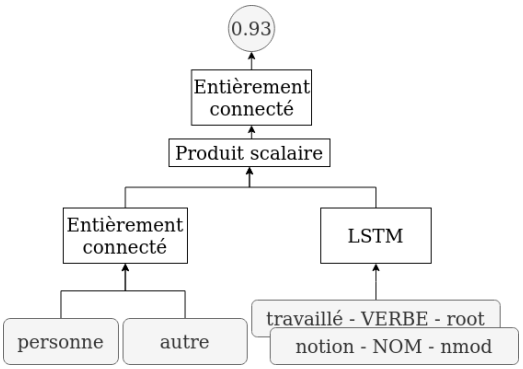


FIGURE 2 – Architecture générale du modèle appliqué à la phrase « Anne-Marie Houdebine a travaillé sur la notion d'Imaginaire linguistique. ».

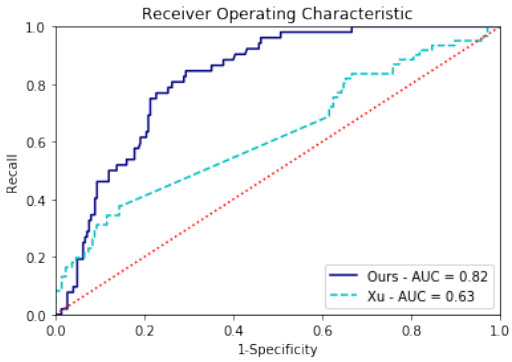


FIGURE 3 – Courbes ROC pour la détection binaire de relations (Xu-binaire et Nous).

de Xu *et al.* sur une tâche de détection. Pour ce faire, ce dernier est transformé d'un modèle de classification multi-classes à un modèle de détection binaire (Xu-binaire) en changeant l'activation de la dernière couche d'un *softmax* en un *sigmoïde*. Xu-binaire et notre modèle rendant tous deux des probabilités, le seuil pour décider qu'une relation existe est la valeur qui maximise le score *rappel+spécificité*. La courbe ROC est donnée en figure 3. Notre rappel de 0,73 est un peu plus bas que le rappel de Xu-binaire de 0,75, mais l'apport en précision est important, notre modèle atteignant 0,84 tandis que Xu-binaire obtient une précision de 0,67. En outre, l'aire sous la courbe ROC est plus grande pour notre modèle, indiquant qu'il a de meilleures performances que Xu-binaire indépendamment du seuil choisi. Un saut dans la courbe ROC de Xu-binaire montre que ce modèle est très sensible au seuil, à la différence du nôtre, plus robuste à une variation dans le choix de celui-ci. Notre modèle est donc particulièrement pertinent pour détecter l'existence ou non de relations. De plus, ce modèle spécialisé, créé pour la détection de relations, détecte mieux les cas fallacieux que Xu-binaire, qui classifie de nombreuses relations négatives dans d'autres catégories que « none ».

Une deuxième expérience vérifie la capacité à détecter des relations inédites, qui n'ont pas été vues lors de l'entraînement. Notre modèle de détection binaire est entraîné avec des labels binaires. L'architecture originale de Xu *et al.* (2015) (Xu-multiclass) n'est pas modifiée, et le modèle est entraîné avec les noms des relations, les différentes étiquettes en sortie étant binarisées *a posteriori*. Nous avons conscience que ce protocole défavorise Xu-multiclass pour ce qui est des performances en classification binaire, car il s'agit d'un modèle entraîné à spécifier et non à détecter. Cependant, le but ici n'est pas de comparer les performances de deux modèles binaires – ce qui a été fait dans l'expérience précédente – mais de montrer qu'un modèle spécialisé dans la détection de relations permet de mieux découvrir des relations inédites qu'un modèle plus « généraliste » de classification, et donc de mieux les traiter ultérieurement. Le jeu d'entraînement consiste en notre jeu de données auquel on a ôté les relations dont le nom est « capitale » et « capitale de ». Notre modèle détecte ces relations inédites (« capitale » et « capitale de ») avec un rappel de 0,65 et une précision de 0,8, alors que le modèle de Xu *et al.* (2015) a un rappel de 0,57 et une précision de 0,71. Notre modèle a donc rempli sa promesse initiale de pouvoir mieux détecter les relations inédites, et donc de pouvoir mieux s'adapter à un flux de textes variés qu'un modèle de classification.

Notre troisième expérience teste la détection de signaux faibles. Nous cherchons à nouveau à prouver l'intérêt d'un modèle de détection binaire pour ce type de tâche par rapport à un modèle de classification dont le rôle principal n'est pas la reconnaissance des signaux faibles. Dans les mêmes conditions que l'expérience précédente, nous avons ôté les signaux faibles des données d'entraînement de l'expérience 1, et redistribué les exemples de signaux faibles dans les données de test. Notre modèle, pour

l'ensemble des données de test, a un rappel de 0,71, égal à celui de Xu-multiclass, mais une meilleure précision, à savoir 0,67 contre 0,72. Mais, lorsque nous considérons uniquement les exemples de signaux faibles, nous obtenons un rappel de 0,34, presque le triple du rappel de Xu-multiclass, 0,12. La précision pour ces résultats est égale à 1, ce qui est normal car elle mesure la proportion d'exemples classifiés comme relation qui sont réellement des relations ; or nous nous concentrons ici sur les relations rares qui existent toutes, donc toutes les relations candidates classifiées comme « en relation » sont bien pertinentes. La détection des relations rares par notre modèle est donc significativement plus élevée, avec un score F1 de 0,51 contre 0,21 pour Xu-multiclass ; ceci est certainement dû au fait qu'il repose sur la syntaxe en utilisant le plus court chemin de dépendances sans le scinder, permettant de reconnaître des constructions syntaxiques en passant outre le vocabulaire inédit.

Notre modèle étant pensé comme une étape préliminaire à la classification de relations, une dernière expérience teste le bénéfice d'ajouter ce modèle de détection comme étape de pré-traitement à un modèle de classification. Les deux systèmes testés et comparés sont un *pipeline* combinant notre modèle et le modèle de classification (Nous+Xu-multiclass) d'une part et, d'autre part, le modèle de classification de relations Xu-multiclass seul. Pour Nous+Xu-multiclass, lors de l'inférence, les relations prédites comme existantes par notre modèle sont passées en entrée de Xu-multiclass, tandis que les relations fallacieuses sont assignées à la classe « none » ; 39% des relations en entrée de notre modèle de détection sont considérées comme pertinentes pour être classifiées. Xu-multiclass prédit, lui, l'entièreté du jeu de test. Nous obtenons pour Nous+Xu-multiclass une précision de 0,74 et un rappel de 0,62, contre une précision de 0,68 et un rappel de 0,62 pour Xu-multiclass seul. Les relations étant en nombre très inégal, le modèle de classification de relations n'apprend à classifier que dans les classes les plus peuplées, les exemples de test n'étant jamais classifiés dans les autres classes. Nous+Xu-multiclass reconnaît six classes, tandis que Xu-multiclass n'en reconnaît que 4. Les rappels présentés pour les deux modèles testés, bien qu'identiques, sont donc issus de phénomènes différents. Plus précisément, le rappel de Nous+Xu-multiclass sur la classe « none » est inférieur à celui de Xu-multiclass seul (0,66 contre 0,75), mais pour une précision accrue (0,93 contre 0,25). Nous+Xu-multiclass obtient également des valeurs de précision et rappel supérieures à Xu-multiclass pour les 4 classes reconnues par les deux systèmes, à l'exception de la classe « contient les subdivisions territoriales administratives » pour laquelle la précision est plus faible (0,18 contre 0,22) mais le rappel plus que doublé. Ces résultats montrent qu'ôter les relations qui ne sont pas pertinentes aide le classifieur, et apporte plus de subtilité dans la classification des relations, tout en limitant l'attribution erronée de la classe « none ».

4 Conclusion

Nous avons créé un système de détection de relations efficace dans le cas de textes journalistiques, et avons expérimentalement prouvé sa pertinence pour le problème de détection et de classification de relations sur des données réelles. Notre modèle peut être utilisé en amont d'un système de classification dont il améliore les résultats, tout en étant capable de repérer des signaux faibles et des relations inédites. Ce modèle seul permet, lorsqu'un journaliste recherche des informations sur une entité, de pouvoir éviter les contresens d'une recherche par co-occurrences : nous ne présentons pas les cas où les relations ne sont pas avérées, évitant d'associer des concepts qui ne sont liés que de façon lointaine. Notre modèle peut de plus être intégré dans la chaîne de traitement des articles d'un journal, juste avant le typage des relations, dans le but d'améliorer à la fois la qualité et la rapidité de l'alimentation d'une base de connaissances journalistique.

Références

BANKO M., CAFARELLA M. J., SODERLAND S., BROADHEAD M. & ETZIONI O. (2007). Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, p. 2670–2676 : Morgan Kaufmann Publishers Inc. DOI : 10.1145/1409360.1409378.

BANKO M. & ETZIONI O. (2008). The tradeoffs between open and traditional relation extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 28–36 : Association for Computational Linguistics. Anthologie ACL : P08-1004.

BUNESCU R. & MOONEY R. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 724–731 : Association for Computational Linguistics. DOI : 10.3115/1220575.1220666.

CAI R., ZHANG X. & WANG H. (2016). Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1-Long Papers)*, p. 756–765 : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1072.

DEL CORRO L. & GEMULLA R. (2013). Clausie : Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, p. 355–366 : Association for Computing Machinery. DOI : 10.1145/2488388.2488420.

GRAVES A., MOHAMED A. & HINTON G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 6645–6649 : IEEE Signal Processing Society. DOI : 10.1109/ICASSP.2013.6638947.

HASEGAWA T., SEKINE S. & GRISHMAN R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* : Association for Computational Linguistics. DOI : 10.3115/1218955.1219008.

KAMBHATLA N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions* : Association for Computational Linguistics. DOI : 10.3115/1219044.1219066.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 55–60 : Association for Computational Linguistics. DOI : 10.3115/v1/P14-5010.

MESQUITA F., SCHMIDEK J. & BARBOSA D. (2013). Effectiveness and efficiency of open relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 447–457 : Association for Computational Linguistics. Anthologie ACL : D13-1043.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations, Workshop Track Proceedings*. arXiv : 1301.3781.

PRADET Q., BAGUENIER-DESORMEAUX J., DE CHALENDAR G. & DANLOS L. (2013). WoNeF, an improved, extended and evaluated automatic French translation of WordNet (WoNeF : amélioration, extension et évaluation d'une traduction française automatique de WordNet) [in French]. In *Proceedings of TALN 2013 (Volume 1 : Long Papers)*, p. 76–89 : ATALA. HAL : cea-00932340.

RIEDEL S., YAO L., MCCALLUM A. & MARLIN B. M. (2013). Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 74–84 : Association for Computational Linguistics. Anthologie ACL : N13-1008.

TAKASE S., OKAZAKI N. & INUI K. (2015). Fast and large-scale unsupervised relation extraction. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, p. 96–105. Anthologie ACL : Y15-1012.

WANG L., CAO Z., DE MELO G. & LIU Z. (2016). Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1-Long Papers)*, p. 1298–1307 : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1123.

WANG W., BESANÇON R., FERRET O. & GRAU B. (2011). Filtering and clustering relations for unsupervised information extraction in open domain. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, p. 1405–1414, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2063576.2063780.

XU Y., MOU L., LI G., CHEN Y., PENG H. & JIN Z. (2015). Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1785–1794 : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1206.

ZHANG Y., QI P. & MANNING C. D. (2018). Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2205–2215 : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1244.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Analyse automatique en cadres sémantiques pour l'apprentissage de modèles de compréhension de texte

Gabriel Marzinotto[2] Delphine Charlet[2] Géraldine Damnati[2] Frédéric Béchet[1]
(1) Aix-Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
(2) Orange Labs, Lannion
(1) {prenom.nom}@lis-lab.fr
(2) {prenom.nom}@orange.com

RÉSUMÉ

Dans le cadre de la compréhension automatique de documents, cet article propose une évaluation intrinsèque et extrinsèque d'un modèle d'analyse automatique en cadres sémantiques (*Frames*). Le modèle proposé est un modèle état de l'art à base de GRU bi-directionnel, enrichi par l'utilisation d'embeddings contextuels. Nous montrons qu'un modèle de compréhension de documents appris sur un corpus de triplets générés à partir d'un corpus analysé automatiquement avec l'analyseur en cadre sémantique présente des performances inférieures de seulement 2.5% en relatif par rapport à un modèle appris sur un corpus de triplets générés à partir d'un corpus analysé manuellement.

ABSTRACT

Semantic Frame Parsing for training Machine Reading Comprehension models

In the framework of Machine Reading Comprehension this paper presents an intrinsic and extrinsic evaluation of a Semantic Frame parser. The proposed model is based on a state of the art bi-directionnal GRU enhanced by the use of transformer-based contextual embeddings. We show that a Machine Reading Comprehension model trained on a corpus of triplets generated from an automatically parsed corpus with our semantic frame parser only yields a 2.5% relative decrease in performance with respect to a model trained on triplets generated from a manually annotated corpus.

MOTS-CLÉS : Analyse en cadres sémantiques, Génération automatique de questions, Compréhension automatique de texte.

KEYWORDS: Semantic Frame Parsing, Question Generation, Machine Reading Comprehension.

1 Introduction

Les systèmes de *Question/Réponse* à partir de documents ont pour but de sélectionner un ou plusieurs passages d'un texte constituant la ou les réponses possibles à une question de compréhension sur le document. Cette tâche de *compréhension de texte* a été traitée avec deux types de modèles : d'une part des modèles basées sur des méthodes de *Recherche d'Information* utilisant un appariement entre requête et document reposant sur des représentations explicites du *sens* des requêtes et des *connaissances* contenues dans les textes avec éventuellement l'accès à des bases externes de connaissances (Kolomiyets & Moens, 2011; Shen & Lapata, 2007) ; d'autres part des méthodes d'appariement direct entre questions et passages de documents utilisant des apprentissages de type *end-to-end* rendu possibles grâce à la disponibilité de très grands corpus contenant à la fois documents, questions et réponses tells que SQuAD (Rajpurkar *et al.*, 2016).

Volume 2 : Traitement Automatique des Langues Naturelles, pages 288–295.
HAL : hal-02784778.

Récemment il a été proposé dans (Béchet *et al.*, 2019b) de combiner représentation explicite du sens et systèmes de question/réponse appris par appariement direct en utilisant un corpus annoté en cadres sémantiques pour générer le corpus d'apprentissage nécessaire à l'adaptation d'un système générique tel que BERT à un nouveau cas d'utilisation. Cette approche, évaluée dans (Béchet *et al.*, 2019c), a cependant une limitation importante : elle nécessite la disponibilité de corpus annotés manuellement en cadres sémantiques afin de générer le corpus d'adaptation. L'étude présentée dans cet article vise à relâcher cette contrainte en développant un analyseur sémantique de type *FrameNet* spécifiquement pour cette tâche de génération de corpus de type Question/Réponse sur documents. Nous étudions les performances de cet analyseur dans ce contexte particulier et proposons une analyse détaillée des résultats en fonction du type de questions posées. Nos contributions visent d'abord l'amélioration des modèles d'analyse en Frames pour le Français, puis l'étude sur la viabilité de la génération des corpus de Question/Réponse à partir des annotations automatiques en cadres sémantiques.

2 Génération de corpus de questions

La tâche de compréhension automatique de texte rencontre un succès grandissant, préfigurant de nouvelles façons d'accéder à l'information contenue dans des documents. Les corpus disponibles sont composés de triplets *(document, question, passage de document constituant la réponse)*, ils sont très majoritairement en anglais (SQuAD (Rajpurkar *et al.*, 2016), MS MARCO (Nguyen *et al.*, 2016)) et leur construction est coûteuse. Dans (Béchet *et al.*, 2019c), les auteurs proposent une approche alternative pour la constitution de corpus d'apprentissage, utilisant un corpus annoté en cadres sémantiques pour générer automatiquement ces triplets à l'aide de patrons. Dans ce protocole, lorsqu'une phrase est annotée en cadre sémantique pour la *Frame* F, pour chaque *Frame Element* E, on peut générer une question dont la réponse est E, à partir de patrons appliqués sur F et les autres *Frames Elements* présents dans la phrase.

Un exemple de phrase annotée en cadre sémantique et les questions générées à partir des annotations est donné ci-dessous :

M. Wildon a laissé plus clairement entendre que si l'Allemagne exécutait sa menace contre le commerce neutre, [*l'Amérique*]$_{\texttt{Speaker}}$ [*lui*]$_{\texttt{Addressee}}$ [*déclarerait*]$_{\texttt{Statement}}$ [*la guerre*]$_{\texttt{Message}}$ *et* [*il*]$_{\texttt{Speaker}}$ *a* [*demandé*]$_{\texttt{Request}}$ [*aux neutres*]$_{\texttt{Addressee}}$ [*de se joindre à lui dans son action*]$_{\texttt{Message}}$.

Questions générées

— *Qui est-ce qui a demandé de se joindre à lui dans son action ?*
— *À qui est-ce que l'Amérique a déclaré la guerre ?*

Dans cet exemple la première question porte sur l'élément `Speaker` de la Frame `Request` ; la deuxième question porte sur l'élément `Addressee` de la Frame `Statement`.

Dans ce protocole, le seul travail manuel requis pour produire les corpus de questions est de définir les patrons générateurs de questions pour chaque *Frame*. Le corpus `CALOR-QUEST` (Béchet *et al.*, 2019a) ainsi généré à partir d'un corpus annoté manuellement en cadres sémantiques a été utilisé pour entraîner un modèle de compréhension de lecture et a donné des résultats très encourageants. Nous proposons dans cet article de développer un analyseur automatique en cadres sémantiques spécifiquement mis au point pour permettre de générer automatiquement, à partir de textes sans annotation manuelle, un corpus d'apprentissage pour les modèles de compréhension de lecture.

3 Analyseur sémantique pour la génération de questions

Dans cette étude nous avons développé un analyseur en cadre sémantique se basant sur un étiqueteur de séquence tel que proposé dans (Marzinotto *et al.*, 2018b). C'est un modèle `biGRU` avec 2 couches de `GRU` bidirectionnelles dans lequel les cadres sémantiques sont codés à l'aide de structures plates reprenant le codage *Begin, Inside, Outside* (`BIO`). Les couches de `biGRU` ont 150 neurones dans chaque direction, et un dropout de 30% entre chaque couche. Nous utilisons Adam comme optimiseur, avec un taux d'apprentissage de $lr = 0.00005$ et des mini-batchs de taille 32. Les séquences d'apprentissage ont une longueur maximale de 120 tokens/word-pieces. Cette taille est suffisante pour tenir compte de plus de 99% des exemples annotées dans CALOR.

Notre système utilise en entrée soit des plongements de mots de type `word2vec` soit des plongements contextuels issus de `BERT` (Devlin *et al.*, 2019). Ces plongements contextuels apparus récemment ont apporté des gains considérables sur des tâches similaires à l'analyse en cadres sémantiques, comme l'analyse en rôles sémantiques de type PropBank (Peters *et al.*, 2018). Cependant, l'impact de ces représentations des mots dans la tâche d'analyse FrameNet a été étudiée uniquement au niveau de la sélection du cadre sémantique (Tan & Na, 2019), et l'étude se limite au corpus FrameNet en anglais. Notre étude étend l'utilisation de BERT à toute la chaîne d'analyse sémantique et l'applique à des corpus en français. Dans tous nos modèles, nous incorporons également des traits linguistiques comme les dépendances syntaxiques, POS, morphologie, capitalisation, préfixes et suffixes des mots de la phrase. Les analyses syntaxiques et morphologiques ont été faits avec un modèle UDPipe (Straka & Straková, 2017) appris sur la FTB d'Universal Dependencies 2.0. Lors de l'apprentissage, nous ajustons les plongements des mots `BERT` ou `word2vec`.

L'originalité de notre approche est d'appliquer au moment du décodage n fois notre analyseur sur chaque phrase sur les n occurrences de déclencheurs potentiels de Frame au sein de la phrase. Les paires { phrase, déclencheur } sont traitées séparément par le réseau, qui prend en entrée une feature indicateur du mot déclencheur. C'est ainsi que chaque paire génère une probabilité de distribution sur les Frames et Frame Elements pour chaque mot de la phase. A partir de l'ensemble des hypothèses produites une dernière phase de décodage implémentant une stratégie de décodage A* similaire à celle proposée par (He *et al.*, 2017) est appliquée. Cette dernière étape permet de garantir la cohérence à la fois des étiquettes `BIO`, mais aussi des relations sémantiques entre Frame et Frame Elements.

L'avantage de cette approche est de permettre facilement l'optimisation de notre modèle par rapport à un point de fonctionnement particulier en terme de précision et rappel pour la détection des Frame et Frame Elements, il suffit pour cela de filtrer parmi toutes les hypothèses produites par l'analyseur de séquence. Nous présentons dans le paragraphe suivant une évaluation intrinsèque de cet analyseur se focalisant sur le type de *questions* qui peuvent être générées par les analyses produites et proposant plusieurs points de fonctionnement qui seront évalués dans l'évaluation extrinsèque sur la tâche de compréhension de documents.

3.1 Évaluation intrinsèque

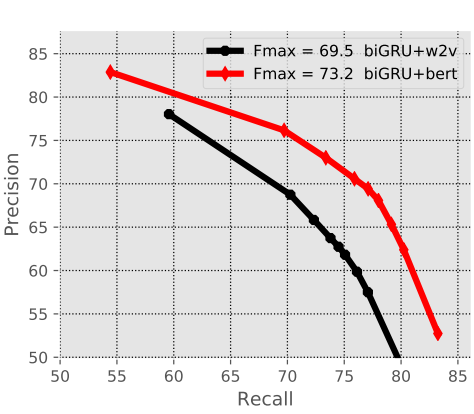
Le modèle est appris et évalué sur le corpus CALOR (Marzinotto *et al.*, 2018a) [1], un corpus de textes encyclopédiques en français annotés en cadres sémantiques (Frames) selon le formalisme FrameNet (Fillmore *et al.*, 2004). Ce corpus est annoté sur un ensemble de 53 Frames différentes pour un total de 31440 occurrences de déclencheurs. Pour cette première série d'expériences le corpus est séparé

1. Corpus disponible : `https://gitlab.lis-lab.fr/alexis.nasr/calor-public/`

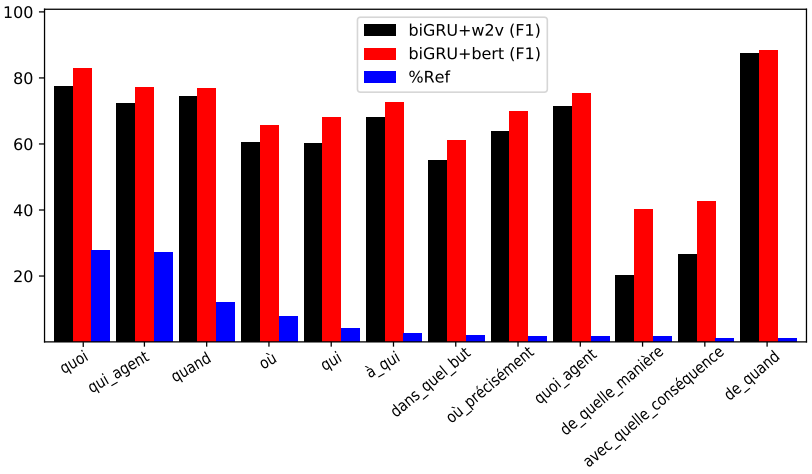
selon une partition de 70% pour l'apprentissage, 10% pour la validation et 20% pour le test. Pour évaluer notre modèle nous utilisons la *F-mesure* sur la tâche de détection et classification des Frame Elements. Nous considérons une détection comme correcte si le recouvrement entre la référence et l'hypothèse contient au moins un mot en commun. Dans l'évaluation, nous propageons les erreurs faits dans l'étape de sélection de la Frame. C'est-à-dire, si un Frame est mal sélectionne, ces Frame Elements seront également faux.

Compromis entre précision et rappel Nous présentons les courbes précision/rappel (P/R) en utilisant différents seuils d'acceptation sur les hypothèses de cadres et de rôles sémantiques produites par nos modèles. Pour dessiner ces courbes, nous utilisons un paramètre $\delta \in (-1; 1)$ qui est soustrait à la probabilité de sortie de l'étiquette *nulle* (ou *Outside*) $P(y_t = O)$ de chaque mot. Par défaut, avec $\delta = 0$, l'hypothèse non nulle la plus probable est sélectionnée si sa probabilité est supérieure à $P(y_t = O)$. Faire varier $\delta < 0$ (ou $\delta > 0$) équivaut à être plus strict (ou moins strict) sur l'hypothèse non nulle la plus élevée. Nous pouvons ainsi étudier le compromis P/R de nos modèles.

Deux variantes du modèle `biGRU` sur le corpus CALOR ont été apprises et évaluées : l'une avec `word2vec` et l'autre avec `BERT`. Le modèle `word2vec` est un modèle cbow de dimension 300 appris sur Wikipedia en français. Le modèle `BERT` utilisé est le modèle multilingue multi_cased_L-12_H-768_A-12 (Devlin *et al.*, 2019). Dans la figure 1a nous observons les courbes précision et rappel pour ces deux variantes. Le modèle `BERT` est supérieur au modèle `word2vec` classique, il atteint une performance de $Fmax = 73.2\%$, avec un $\delta = 0$, soit trois points d'amélioration absolue sur la $F1$ par rapport au précédent modèle à base de `word2vec`. Ce point de fonctionnement est proche du point d'égale erreur ($P \approx R \approx Fmax$). La figure 1a montre aussi plusieurs points de fonctionnement possibles. Les deux points extrêmes pour le modèle basé sur `BERT` présentent respectivement une précision maximale de $P = 83\%$ pour rappel minimal de $R = 55\%$ ($\delta = -0.9$) et un rappel maximal de $R = 84\%$ pour une précision minimale de $P = 53\%$ ($\delta = +0.9$).



(a) Courbe P/R du modèle `biGRU` avec W2V ou `BERT`



(b) Performace du `biGRU` pour chaque type de question

FIGURE 1

Analyse des résultats Afin de mieux percevoir les performances du modèles, nous avons établi un regroupement des *Frame Elements* (FE) selon une question prototypique à laquelle ils pourraient potentiellement répondre. Par exemple, pour la Frame `Hiding`, l'élément Hiding_place est associé à la question prototypique `où`, `Hiding_object` est associé à `quoi` et `Agent` à `qui_agent`.

Pour "qui" et "quoi", nous traitons séparément le cas où le FE est agent de la Frame, auquel cas la question prototypique est `qui_agent` ou `quoi_agent`. Cette distinction permet d'évaluer les cas comme "qui attaque qui". Cette association a été établie pour les 53 Frames du corpus, conduisant à un regroupement des 495 Frame Elements en 64 catégories selon la question prototypique à laquelle ils peuvent répondre.

Les graphiques ne reprennent que les catégories les plus représentées. Ainsi, la figure 1b montre à la fois la performance de l'analyse automatique en Frame par type de question, ainsi que la distribution des types de questions associées aux FEs dans le corpus CALOR. Nous observons que les types les plus fréquents sont *quoi*, *quoi_agent*, *quand*, *où* et *qui*. Les FEs répondant à des questions plus abstraites comme *dans_quel_but*, *de_quelle_manière* et *avec_quelle_conséquence*, sont beaucoup plus rares mais aussi plus difficiles à repérer. Nous observons aussi que `BERT` donne systématiquement de meilleurs résultats que ceux obtenus avec `word2vec` dans tous les types de FEs et que les apports les plus significatifs sont pour les catégories les plus difficiles.

3.2 Evaluation extrinsèque

Pour répondre à la tâche de détection de la réponse à une question donnée dans un texte, nous utilisons une version adaptée sur cette tâche d'un modèle de langue contextuel, ici le modèle BERT multilingue (multi_cased_L-12_H-768_A-12 (Devlin *et al.*, 2019)), avec les hyperparamètres utilisés par ce auteurs pour l'entraînement sur le corpus SQUAD. Afin de se placer dans des conditions équivalentes à SQUAD, les documents de CALOR sont découpés en paragraphes d'une longueur proche de la longueur moyenne des paragraphes de SQUAD (environ 120 tokens).

L'évaluation standard proposée dans SQUAD consiste à comparer, en supprimant les articles, l'ensemble des mots présents dans la réponse détectée à l'ensemble des mots présents dans la réponse de référence. Cette comparaison est faite de façon stricte ("exact-match") pour donner une valeur binaire, ou par le calcul d'une F-mesure issue de la précision et du rappel sur la comparaison des ensemble de mots. Nous reprenons ici cette évaluation standard, en adaptant la liste d'articles au français, ainsi que le protocole expérimental proposé sur CALOR-QUEST par (Béchet *et al.*, 2019a).

Pour entraîner le modèle de compréhension de lecture, nous considérons les corpus suivants : le corpus des questions générées par patrons à partir de l'annotation manuelle en cadres sémantiques (Gold), et les corpus des questions générées par patrons à partir de l'annotation automatique en cadres sémantiques par le meilleur modèle obtenu précédemment (BiGRU-BERT), à différents points de fonctionnement de l'analyseur automatique (de façon à pouvoir traiter l'ensemble du corpus CALOR, les analyses automatiques sont produites par un mécanisme de k-Fold avec k=9).

Le corpus d'évaluation est constitué de 2069 triplets (paragraphe,question,réponse) produits par des annotateurs humains. Ces annotateurs observaient une `Frame`, un `FE Reponse` et des `FEs Contexte`), ensuite ils produisent une question qui a comme réponse le FE indiqué. La phrase originale n'était pas affichée pour laisser plus de liberté aux annotateurs dans les choix lexicaux effectués pour rédiger les questions. Les annotateurs étaient libres de choisir les FE du contexte qu'ils allaient inclure dans leurs questions. Même si les questions sont limités aux Frames et Frame Elements de CALOR, il faut clarifier que ces Frames ont été sélectionnées pour être les plus représentatifs des documents (Marzinotto *et al.*, 2018a). Par ailleurs, le grand nombre de Frame Elements et le dégrée de détail des annotations FrameNet induit une variabilité importante dans les questions produites. Afin de réduire au maximum le biais dû au fait que les questions sont restreints aux Frames du corpus CALOR. Nous générons les exemples d'apprentissage à partir des documents distincts à ceux du qui

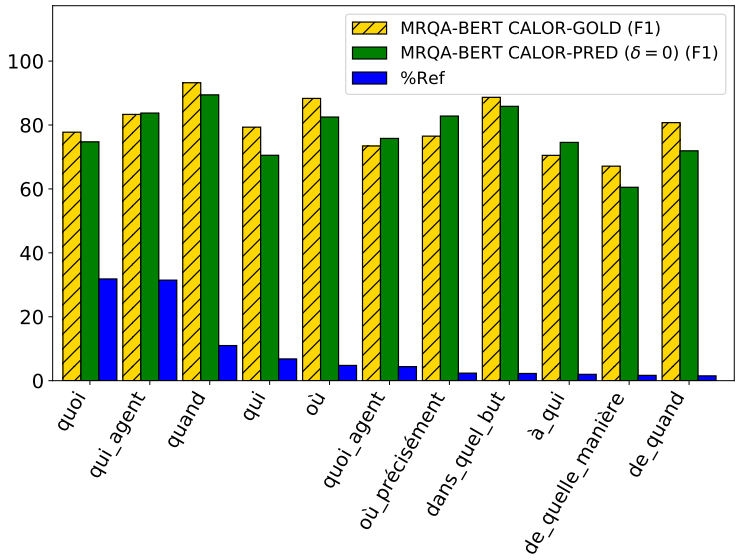


FIGURE 2: Performance du MRQA pour chaque type de question

ont servi pour générer les question manuelles.

Le tableau 1 présente les performances obtenues selon les corpus d'apprentissage utilisés. Pour le corpus d'apprentissage issu de l'analyse automatique, nous considérons les 3 points de fonctionnement suivants : maxP, maxF et maxR. maxP est le point de fonctionnement favorisant la précision de l'annotation en cadre sémantique (80%), maxF celui favorisant la Fmesure (73%) et maxR celui favorisant le rappel (80%). Pour maxP, le corpus est relativement restreint mais avec peu d'erreurs, tandis que pour maxR, le corpus est plus gros mais avec plus d'erreurs. On constate qu'il vaut mieux ne pas favoriser la précision de l'analyse en cadres sémantiques pour la génération du corpus, et qu'il est préférable d'avoir plus de questions, même entachées d'erreurs. En effet, les points de fonctionnement maxF et maxR donnent des performances sensiblement équivalentes. Finalement, les performances obtenues avec un corpus d'apprentissage issu de l'annotation automatique ne sont que faiblement dégradées par rapport aux performances obtenues par le corpus d'apprentissage Gold. Les résultats sur la configuration Gold montrent que même si le corpus de Frames induit un biais sur l'annotation du corpus de test, le corpus obtenu est loin d'être trivial car les modèles de MRQA BERT ont encore une marge d'amélioration considérable.

point de fonctionnement analyseur	#nbtrain	Exact-Match	F-mesure
maxP	8779	62.6	75.6
maxF	12254	67.4	78.6
maxR	13692	67.0	78.7
Gold	17423	69.9	80.6

TABLE 1: Performance de réponse aux questions selon la qualité du corpus d'apprentissage

La figure 2 montre le détail de ces performances de compréhension de lecture selon le type de questions. Le type de question est obtenu par le même protocole d'association que celui présenté à la section 3.1. Il s'agit en effet de la catégorie associée au FE répondant à la question. Les types de questions sont triés par fréquence décroissante et seuls ceux ayant plus de 30 occurrences dans le corpus de test sont présentés ici. La figure permet de comparer, pour chaque type de question, les performances obtenues par apprentissage sur annotations Gold et par apprentissage sur annotations

automatiques. On peut constater que les performances pour chaque type de questions sont assez proches, que le corpus d'apprentissage soit manuel ou automatique.

Une analyse plus détaillée des erreurs produites par le modèle d'analyse en Frames peut permettre d'expliquer pourquoi les performances ne sont que peu dégradées avec le corpus produit automatiquement. Pour le point de fonctionnement `maxF` par exemple, les erreurs commises sur l'identification des FE sont pour 54.2% d'entre elles des insertions et pour 32.0% d'entre elles des omissions. Seules 13.8% des erreurs sont des substitutions dues principalement à des confusions entre deux cadres sémantiques, or les substitutions sont les erreurs les plus enclines à générer des questions erronées. Les erreurs d'omission sont présentes pour tous les types de FE, mais sont plus fréquentes pour les arguments sémantiques abstraits et difficiles. Les omissions n'ayant d'autre impact sur le processus d'apprentissage que de réduire le nombre d'exemples d'apprentissage, les conséquences sur ces types de questions sont moindres.

4 Conclusion

Nous avons présenté un nouveau modèle d'analyse en cadres sémantiques basé sur un modèle bi-GRU associé à des embeddings contextuels de type BERT. Une évaluation intrinsèque détaillée originale sous l'angle de questions prototypiques a permis de révéler une typologie de rôles sémantiques plus ou moins difficiles à détecter et identifier. De façon complémentaire, une évaluation extrinsèque est proposée ou le corpus analysé automatiquement est utilisé pour générer un corpus d'apprentissage pour une tâche de compréhension de lecture. Les expériences montrent qu'une analyse automatique en Frames peut permettre efficacement de générer un corpus d'apprentissage conduisant à des modèles fauiblement dégradés par rapport à l'utilisation d'une annotation manuelle. Ces résultats encourageants pourront être confortés par la suite par des expériences complémentaires sur des données issues de domaines applicatifs différents, au delà des textes encyclopédiques.

Références

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019a). CALOR-QUEST : generating a training corpus for Machine Reading Comprehension models from shallow semantic annotations. In *MRQA : Machine Reading for Question Answering - Workshop at EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. HAL : hal-02317018.

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019b). CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes. In *TALN 2019*, Toulouse, France. HAL : hal-02377119.

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019c). Calor-quest : generating a training corpus for machine reading comprehension models from shallow semantic annotations. In *MRQA2019, second workshop on machine reading comprehension, satellite workshop EMNLP2019*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. arXiv preprint : 1810.04805.

FILLMORE C. J., BAKER C. F. & SATO H. (2004). FrameNet as a "net". In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA). Anthologie ACL L04-1221.

HE L., LEE K., LEWIS M. & ZETTLEMOYER L. (2017). Deep semantic role labeling : What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

KOLOMIYETS O. & MOENS M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, **181**(24), 5412–5434.

MARZINOTTO G., AUGUSTE J., BECHET F., DAMNATI G. & NASR A. (2018a). Semantic Frame Parsing for Information Extraction : the CALOR corpus. In *LREC2018*, Miyazaki, Japan. HAL : hal-01959187.

MARZINOTTO G., BÉCHET F., DAMNATI G. & NASR A. (2018b). Sources of Complexity in Semantic Frame Parsing for Information Extraction. In *International FrameNet Workshop 2018*, Miyazaki, Japan. HAL : hal-01731385.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). Ms marco : A human generated machine reading comprehension dataset. arXiv preprint : 1611.09268.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

SHEN D. & LAPATA M. (2007). Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, p. 12–21.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

TAN S.-S. & NA J.-C. (2019). Positional attention-based frame identification with bert : A deep learning approach to target disambiguation and semantic frame selection.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Analyse de sentiments des vidéos en dialecte algérien

Mohamed Amine Menacer[1] Karima Abidi[1] Nouha Othman[1,2] Kamel Smaïli[1]
(1) LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France
(2) LARODEC, Institut Supérieur de Gestion de Tunis, 2000 Bardo, Tunisia
{mohamed-amine.menacer,karima.abidi,nouha.othman,kamel.smaili}@loria.fr

RÉSUMÉ

La plupart des travaux existant sur l'analyse de sentiments traitent l'arabe standard moderne et ne prennent pas en considération les spécificités de l'arabe dialectal. Cet article présente un système d'analyse de sentiments de textes extraits de vidéos exprimées en dialecte algérien. Dans ce travail, nous avons deux défis à surmonter, la reconnaissance automatique de la parole pour le dialecte algérien et l'analyse de sentiments du texte reconnu. Le développement du système de reconnaissance automatique de la parole est basé sur un corpus oral restreint. Pour pallier le manque de données, nous proposons d'exploiter des données ayant un impact sur le dialecte algérien, à savoir l'arabe standard et le français. L'analyse de sentiments est fondée sur la détection automatique de la polarité des mots en fonction de leur proximité sémantique avec d'autres mots ayant une polarité prédéterminée.

ABSTRACT

Sentiment analysis of videos in Algerian dialect

Most of the existing works on sentiment analysis deal only with Modern Standard Arabic (MSA), and do not take into account the dialects. This article presents a system for analyzing the sentiments of the utterances extracted from videos, in which the language used is Algerian dialects. We have two challenges to overcome, the automatic speech recognition for the Algerian dialect and the sentiment analysis of the recognized text. A spoken corpus has been recorded in order to develop a baseline system for recognizing the videos. This system is then improved by taking advantage of the acoustic data having an impact on the Algerian dialect, namely standard Arabic and French. The sentiment analysis is based on the automatic detection of the polarity of words according to their semantic proximity to other words with a predetermined polarity.

MOTS-CLÉS : Analyse de sentiments, Dialecte algérien, Vidéos, Reconnaissance automatique de la parole.

KEYWORDS: Sentiment analysis, Algerian dialect, Videos, Automatic speech recognition.

1 Introduction

Plusieurs recherches ont été conduites sur la langue arabe. En revanche, la majorité des travaux destinés au traitement automatique de la langue arabe écrite s'est focalisée, de façon presque exclusive, sur l'arabe moderne standard, en laissant de côté les formes vernaculaires. En effet, l'arabe moderne standard est la langue officielle dans le monde arabe. Elle se trouve principalement dans les livres, les journaux, les magazines, et les médias officiels. Elle représente la forme de l'arabe universel enseignée dans les écoles et utilisée dans les discussions formelles. Cependant, la communication

Volume 2 : Traitement Automatique des Langues Naturelles, pages 296–304.
HAL : hal-02784779.

dans la vie quotidienne se fait à travers le dialecte qui est propre à chaque région du monde arabe. Cette forme parlée est essentiellement basée sur l'arabe moderne standard en relâchant plusieurs contraintes morpho-syntaxiques de la langue d'origine pour laisser place à une langue informelle plus simple d'usage. Le dialecte est parfois combiné avec d'autres langues étrangères comme le français ou l'anglais et il ne s'agit pas de simples emprunts, mais d'utilisation de phrases entières en langues étrangères.

Depuis l'apparition des réseaux sociaux, la communauté TAL s'est lancée dans une activité de recherche accrue sur les dialectes arabes. En effet, les internautes expriment leurs sentiments et opinions à propos de différents sujets dans les réseaux sociaux essentiellement en dialecte. L'analyse de sentiments qu'elle soit parlée ou textuelle est un domaine riche en publications (Kiritchenko *et al.*, 2016; Barhoumi *et al.*, 2018; Brahimi *et al.*, 2019). Néanmoins, très peu de travaux dans ce domaine ont été réalisés sur les dialectes.

Dans ce travail, nous nous intéressons à l'étude de sentiments dans les réseaux sociaux où le dialecte algérien est utilisé comme support de communication. Pour ce faire, nous proposons un système de détection de la polarité (sentiment positif ou négatif) pour une collection de vidéos en dialecte algérien. Ces vidéos sont collectées à partir des chaînes algériennes disponibles sur YouTube. Les vidéos sont transcrites à l'aide d'un système de reconnaissance automatique de la parole (SRAP) pour le dialecte algérien, et ensuite l'étude de sentiments est effectué sur les transcriptions.

Le dialecte algérien est l'un des dialectes les plus difficiles à reconnaître par un SRAP. Cela est dû au fait que cette variante de la langue arabe utilise de nombreuses séquences de mots empruntées (principalement de la langue française). En outre, dans ce dialecte les mots de l'arabe standard sont altérés phonologiquement afin d'en faciliter la prononciation (Harrat *et al.*, 2017, 2018). Par ailleurs, les mots empruntés peuvent être utilisés tels quels, ou ils peuvent être modifiés afin de respecter la structure morphologique de la langue arabe.

Pour construire un SRAP robuste, il faut disposer d'une grande quantité de données orales et écrites de la langue à reconnaître. Malheureusement, ce type de données n'existe pas pour le dialecte algérien puisqu'il est principalement parlé de plus, il n'existe pas de normes ni de règles pour l'écrire ce qui rend le traitement des textes existant plus complexe. Notre approche pour reconnaître le dialecte algérien est d'exploiter des données d'autres langues ayant un impact sur le dialecte, à savoir le MSA et le français. Une autre ressource primordiale dans les SRAP est le dictionnaire de prononciation. L'approche la plus simple pour le générer se base sur la décomposition en caractères de chaque mot pour avoir sa prononciation (Le & Besacier, 2009; Killer *et al.*, 2003; Gizaw, 2008). Une autre approche consiste à utiliser des méthodes statistiques pour convertir les graphèmes en phonèmes (Cucu *et al.*, 2011; Karanasou & Lamel, 2010; Harrat *et al.*, 2014; Masmoudi *et al.*, 2018). C'est cette approche que nous avons adoptée pour notre système.

Une fois la transcription des vidéos est générée par le SRAP, nous procédons ensuite à l'analyse de sentiments qui est basée sur la détection de polarité des mots dialectaux composant cette transcription. Cette polarité est déterminée en fonction des mots proches ayant une orientation prédéterminée.

2 Les corpus

Afin de développer et évaluer un système permettant l'analyse de sentiments de vidéos en dialecte algérien, nous avons utilisé plusieurs sources de données qui sont décrites ci-dessous :

YouTubAlg : nous utilisons ce corpus pour calculer l'orientation sémantique des mots du dialecte algérien et pour apprendre le modèle de langage du SRAP. Il comporte des commentaires collectés à partir de YouTube en utilisant l'API [1] de Google. Pour récupérer un maximum de données correspondant au dialecte algérien, nous avons utilisé une liste de mots-clés spécifiques dressée au préalable. Ces mots-clés correspondent principalement à des évènements ou à des personnalités connues relatives à l'actualité algérienne et ne présentant aucun intérêt au niveau international. En effet, ce principe a été utilisé pour éliminer l'éventualité de collecter des commentaires d'arabophone autres qu'algériens. Le corpus obtenu est composé de 18,3M de mots (Abidi *et al.*, 2017)).

ADIC : l'apprentissage du modèle acoustique dans les SRAP est basé sur une collection de données orales avec leur transcription. ADIC (Algerian DIalect Corpus) a été construit en enregistrant, à l'aide d'un microphone unidirectionnel professionnel, 4,6k phrases par 7 locuteurs natifs algériens. Les phrases ont été sélectionnées à partir de deux corpus : YouTubAlg et PADIC (Meftouh *et al.*, 2015, 2018). Ce dernier est une collection de 6,4K phrases en arabe standard avec leurs traductions dans plusieurs dialectes arabes dont deux dialectes algériens. Le corpus obtenu contient 6 heures de parole réparties comme suit : 240 minutes sont utilisées pour l'apprentissage, 40 minutes pour la validation et 70 minutes pour le test.

SentAlgVid : nous utilisons ce corpus pour l'évaluation finale de notre modèle d'analyse de sentiments de vidéos en dialecte. *SentAlgVid* est une collection de vidéos en dialecte diffusées par des chaînes de télévision algériennes comme *Ennahar TV, Echorouk TV, et El Bilad TV*. Le nombre total de vidéos est égale à 30 vidéos d'une durée moyenne de 2 minutes. Les vidéos de ce corpus ont été annotées manuellement en termes de polarité (positive et négative) par des locuteurs natifs.

3 Modèles proposés

Dans ce travail, nous avons deux défis à surmonter, la RAP du dialecte algérien et l'analyse de sentiments de ce dernier. Le modèle final est basé sur une architecture *pipeline* où la sortie du SRAP est utilisée comme entrée de système de l'analyse de sentiments. Dans ce qui suit, nous présentons chaque composant du modèle final proposé.

3.1 Reconnaissance automatique de la parole pour le dialecte algérien

Le développement d'un SRAP est basé sur trois composants : le modèle acoustique modélisant le système phonologique de la langue, le modèle de langage assurant le respect des règles grammaticales et le modèle de prononciation définissant le vocabulaire et les différentes variantes de prononciation.

3.1.1 La modélisation acoustique

Le modèle acoustique est basé sur les réseaux de neurones de type perceptrons multicouches. Ces modèles sont entraînés pour estimer la probabilité d'associer chaque observation acoustique à un triphone. Les observations acoustiques sont des vecteurs fMLLR (feature-space Maximum Likelihood

1. Disponible sur : https ://developers.google.com/YouTube

Linear Regerssion) (Gales, 1998) de dimension 40. Ces observations sont souvent utilisés pour l'apprentissage adaptatif (*speaker Adaptive Training (SAT)* (Anastasakos *et al.*, 1996)) qui vise à rapprocher les observations acoustiques initiaux et cibles par une transformation linéaire. L'architecture est basée sur 6 couches cachées de 2048 neurones chacune. La couche en entrée est composée de 440 neurones représentant la concaténation de 11 observations acoustiques. L'estimation des paramètres du réseau de neurones nécessite une grande quantité de données acoustiques, en revanche, on ne dispose que de 4 heures du dialecte pour l'apprentissage. Pour cette raison, nous avons décidé de tirer profit des langues influençant le dialecte algérien, à savoir : le MSA et le français. C'est pourquoi, ADIC a été étendu en ajoutant progressivement 4 heures de chaque langues jusqu'à arriver à 44 heures. Les données de l'arabe standard sont extraites de deux corpus NEMLAR (Yaseen *et al.*, 2006) et NetDC (Choukri *et al.*, 2004), tandis que les données de la langue française ont été collectées à partir du corpus ESTER (Galliano *et al.*, 2005). La quantité optimale de données acoustiques de chaque langue a été déterminée en minimisant le WER (Word Error Rate) sur la partie de validation de ADIC. Nous sommes arrivés à la conclusion qu'en ajoutant plus de 12 heures du MSA et plus de 12 heures du français aux données dialectales, les performances du SRAP se dégradent.

3.1.2 La modélisation du langage

L'apprentissage du modèle de langage pour le dialecte algérien n'est pas limitée aux données dialectales (les deux corpus PADIC et YouTubAlg), nous utilisons également des données de l'arabe standard. Comme la quantité des différentes données textuelles est déséquilibrée, le modèle de langage, que nous proposons, est une combinaison linéaire de quatre modèles bi-grammes. Deux d'entre eux ont été entraînés sur des données textuelles de l'arabe standard : la version arabe de Gigaword (1 milliard de mots) et la transcription des données acoustiques utilisées pour enrichir ADIC (315 000 mots), les deux autres ont été entraînés sur des données dialectales : PADIC et YouTubAlg. Les poids de l'interpolation linéaire ont été estimés pour maximiser la probabilité d'un corpus de développement composé d'un mélange de données du MSA et du dialecte. Les poids de pondération, calculés sur le corpus de développement, pour chaque corpus sont les suivants : 0,48 pour YouTubAlg, 0,22 pour Gigaword, 0,19 pour la transcription des données orales du MSA et 0,11 pour PADIC.

3.1.3 La modélisation de la prononciation

Le lexique de prononciation est composé de l'union des mots les plus fréquents extraits à partir de chaque ensemble de données utilisé pour l'apprentissage du modèle de langage. Pour chaque mot du lexique, il faut disposer de toutes ses variantes de prononciation. La question est de savoir comment produire toutes les variantes de prononciation possibles pour les mots arabes, et plus particulièrement les mots dialectaux, sachant que les textes arabes sont écrits sans aucune diacritique. Nous avons utilisé un lexique externe (Ali *et al.*, 2014) comme une table de recherche à partir de laquelle les prononciations des mots arabes sont extraites. Malheureusement, nous ne disposons pas de l'équivalent de cette ressource pour le dialecte algérien. Pour remédier à ce problème, nous avons adopté une approche de type G2P (*grapheme-to-phoneme*) afin de produire les variantes de prononciation pour les mots dialectaux. Pour ce faire, nous avons adapté l'approche proposée dans (Harrat *et al.*, 2014). Le processus de conversion G2P commence par la restitution des diacritiques avec un processus automatique basé sur une approche statistique. Ce problème est considéré comme un problème de traduction automatique où la langue source est un ensemble de phrases non voyellées et la langue cible est un ensemble de phrases avec voyelles. Une fois que les diacritiques sont

restituées, un ensemble de règles est utilisé pour produire la prononciation de mots dialectaux (Harrat *et al.*, 2014). Le lexique final contient 125k mots et 538k variantes de prononciation.

3.2 L'analyse de sentiments

Pour déterminer la polarité des mots dialectaux, nous nous sommes basés sur la proximité de leur orientation avec celle de mots de base que nous appellerons des mots-germes. Pour ce faire, nous proposons une méthode qui s'inspire des travaux de (Turney & Littman, 2003) et (Htait *et al.*, 2017) tous deux appliqués à l'anglais. La méthode que nous proposons est composée de deux sous-tâches.

3.2.1 L'identification des mots germes

L'idée consiste à estimer la polarité d'un mot en se basant sur celles des mots d'une liste établie en amont (mots-germes) (Turney & Littman, 2003). Dans ce travail, les mots germes sont ceux dont la polarité est évidente. Dans (Htait *et al.*, 2017), en plus des mots identifiés par Turney, les auteurs ont ajouté une liste d'une quarantaine de mots-germes identifiés à partir des mots les plus fréquents. Ces mots ont été étiquetés, en termes de polarité, manuellement par les auteurs.

Pour ce qui nous concerne, nous avons choisis 80 mots germes à partir d'une liste des mots les plus fréquents de *YouTubAlg*. Dans le tableau 3.2.1 nous donnons quelques exemples de ces mots-germes retenus.

Mots-germes positifs	chaba (*jolie*), bravo, هَايل (*super*), الصحة (*la santé*), belle, شكرًا (*merci*)
Mots-germes négatifs	شيّات (*lèche botte*), harki (traître), جَاهل (*ignorant*), زعفَان (*énervé*), mafia

TABLE 1 – Quelques exemples de mots germes positifs et négatifs.

3.2.2 L'estimation de la polarité des mots du lexique

Dans (Turney & Littman, 2003) les auteurs estiment qu'un mot positif est plus proche des mots germes positifs que des mots germes négatifs s'il est proche des mots-germes positifs et inversement. L'orientation d'un mot est calculée sur la base de la différence entre entre sa similitude avec les mots-germes positifs et les mots-germes négatifs, comme le montre l'équation 1 :

$$SO(w) = \sum_{w_p \in MGP} sim(w, w_p) - \sum_{w_n \in MGN} sim(w, w_n) \quad (1)$$

Où MGP et MGN correspondent respectivement à la liste des mots-germes positifs et négatifs. Le calcul de la similarité est effectué par les auteurs de (Turney & Littman, 2003) en utilisant l'information mutuelle. Pour ce qui nous concerne, nous avons utilisé une représentation distribuée (*word embedding*) apprise sur le corpus YouTubAlg. Ensuite, nous avons calculé la similarité cosinus entre les vecteurs représentatifs de ces mots. La méthode proposée nous a permis de construire un lexique de polarité pour le dialecte algérien comportant 11,2k entrées.

4 Expérimentations

La démarche expérimentale que nous mettons en place consiste à évaluer chaque composant séparément, à savoir : le SRAP et le système d'analyse de sentiments, pour évaluer enfin la sortie finale.

4.1 La reconnaissance automatique de la parole

L'évaluation de notre système de reconnaissance de la parole est basée sur la partie test de ADIC (70 minutes de parole). Nous n'avons pas utilisé le corpus *SentAlgVid* qui est destiné pour l'évaluation de la sortie finale car on ne dispose pas de la transcription de vidéos de ce corpus. Les résultats en terme du WER sont présentés dans le tableau 2.

Système	**Données acoustiques**	**WER(%)**	**OOV (%)**
S_{base}	ADIC	40.0	6.8
S_1	ADIC+44hMSA+40hFr	38.8	
S_2	ADIC+12hMSA+12hFr	**37.7**	

TABLE 2 – Les résultats de reconnaissance de la parole sur la partie test de ADIC.

Le WER du système de base S_{base} entraîné avec seulement les 4 heures de la partie d'apprentissage de ADIC est de 40%. En intégrant toutes les données acoustiques (système S_1) du MSA (44 heures) et du français (40 heures), les performances de S1 sont meilleures de 1,2% par rapport au système de base. Mieux encore, en optimisant la taille des données acoustiques provenant du MSA et du français, nous avons obtenu une amélioration absolue de 2,3% (S_2). Il est à noter que l'intervalle de confiance pour le système de base est de $\pm$1,2%, ce qui signifie que S_2 atteint une amélioration significative par rapport au système de base. Cela montre également que la taille des données utilisées pour apprendre le modèle acoustique pour le dialecte algérien affecte les performances du système de reconnaissance.

Il est à noter que les travaux de recherches sur la RAP pour le dialecte algérien sont relativement moins nombreux pour pouvoir comparer nos résultats. Cependant, dans la dernière édition de la compétition MGB, MGB5 (Ali *et al.*, 2019), il y avait une tâche de RAP pour le dialecte marocain. Ce dernier est relativement proche du dialecte algérien, ils partagent plusieurs aspects linguistiques et acoustiques. Le meilleur système a obtenu un WER de 37,6%, sachant que 13 heures de la parole dialectale ont été utilisées avec 1200 heures de l'arabe standard pour apprendre le modèle acoustique. Cela montre la difficulté de reconnaître les dialectes maghrébins en particulier le dialecte algérien et que les résultats de notre système sont acceptables.

4.2 L'analyse de sentiments

Afin d'évaluer le lexique que nous avons construit d'une manière automatique, il faut disposer d'un corpus de commentaires en dialecte algérien où chacun d'entre eux est associé à une polarité. Ensuite, il faut utiliser le lexique de polarité que nous avons développé pour calculer la polarité sur ce corpus. Malheureusement, ce type de corpus n'existe pas pour le dialecte algérien. Par conséquent, nous avons dû en construire un. Pour ce faire, nous avons annoté manuellement 750 commentaires extraits de YouTube. Cela a donné lieu à 390 commentaires positifs et à 360 commentaires négatifs, avec une

moyenne de 9 mots par commentaire. Nous avons ensuite estimé la qualité du lexique construit en utilisant ce corpus que nous avons nommé *SentAlg*. Pour calculer la polarité d'un commentaire nous sommons la polarité de chacun de ses termes. Dans le tableau 4 nous donnons les résultats obtenus en terme de rappel et de précision. De ces résultats on peut constater que la méthode proposée est

Corpus	Rappel	Précision
SentAlg	88.11%	88.64%

TABLE 3 – Résultats expérimentaux sur le corpus *SentAlg*.

pertinente et a conduit à la construction d'un lexique de polarité pertinent.

Nous avons testé ce lexique également sur les transcriptions automatiques des vidéos de *SentAlgVid*. Les résultats du tableau 4 sont intéressants, même s'ils ne sont pas de la même qualité que ceux obtenus sur des corpus de textes simples. En effet, rappelons que ces résultats sont calculés sur des transcriptions automatiques obtenus à l'aide d'un SRAP dont le WER du système de reconnaissance de la parole est de 37,7%. Nous considérons ce résultat comme très encourageant étant donné le taux d'erreur élevé.

Corpus	Rappel	Précision
SentAlgVid	60%	64.28%

TABLE 4 – Résultats expérimentaux sur le corpus de vidéos de dialecte algérien *SentAlgVid*.

5 Conclusion

Dans cet article, nous avons proposé un système d'analyse de sentiments de vidéos en dialecte algérien. Nous y avons abordé deux problèmes critiques, à savoir la reconnaissance automatique de la parole pour le dialecte algérien et l'analyse de sentiments du texte reconnu. Pour surmonter le problème de manque de données dialectales nécessaires aux différents modèles du SRAP, nous avons exploité des données de langues ayant un impact sur le dialecte, à savoir l'arabe standard et le français. Nous avons montré qu'il est important de doser la quantité de données à utiliser de chaque langue étrangère afin d'améliorer le SRAP. En ce qui concerne l'analyse de sentiments, une méthode a été proposée pour construire automatiquement un lexique de polarité qui a permis d'analyser le contenu de vidéos en dialecte algérien.

Références

ABIDI K., MENACER M. A. & SMAILI K. (2017). CALYOU : A Comparable Spoken Algerian Corpus Harvested from YouTube. In *18th Annual Conference of the International Communication Association (Interspeech)*, Conference of the International Communication Association (Interspeech), Stockholm, Sweden.

ALI A., SHON S., SAMIH Y., MUBARAK H., ABDELALI A., GLASS J., RENALS S. & CHOUKRI K. (2019). The mgb-5 challenge : Recognition and dialect identification of dialectal arabic speech.

ALI A., ZHANG Y., CARDINAL P., DAHAK N., VOGEL S. & GLASS J. (2014). A complete KALDI recipe for building Arabic speech recognition systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, p. 525–529. DOI : 10.1109/SLT.2014.7078629.

ANASTASAKOS T., MCDONOUGH J., SCHWARTZ R. & MAKHOUL J. (1996). A compact model for speaker-adaptive training. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, p. 1137–1140 vol.2. DOI : 10.1109/ICSLP.1996.607807.

BARHOUMI A., ALOULOU C., CAMELIN N., ESTÈVE Y. & BELGUITH L. (2018). Arabic Sentiment analysis : an empirical study of machine translation's impact. In *Language Processing and Knowledge Management international conference (LPKM2018)*, Sfax, Tunisia.

BRAHIMI B., TOUAHRIA M. & TARI A. (2019). Improving sentiment analysis in arabic : A combined approach. *Journal of King Saud University - Computer and Information Sciences*.

CHOUKRI K., NIKKHOU M. & PAULSSON N. (2004). Network of data centres (NetDC) : BNSC - an Arabic broadcast news speech corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

CUCU H., BESACIER L., BURILEANU C. & BUZO A. (2011). Investigating the role of machine translated text in ASR domain adaptation : Unsupervised and semi-supervised methods. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, p. 260–265. DOI : 10.1109/ASRU.2011.6163941.

GALES M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, **12**(2), 75–98.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F. & GRAVIER G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Ninth European Conference on Speech Communication and Technology*.

GIZAW S. (2008). Multiple pronunciation model for Amharic speech recognition system. In *Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*.

HARRAT S., MEFTOUH K., ABBAS M. & SMAÏLI K. (2014). Grapheme to phoneme conversion - an Arabic dialect case. In *Spoken Language Technologies for Under-resourced Languages*.

HARRAT S., MEFTOUH K. & SMAÏLI K. (2017). Creating Parallel Arabic Dialect Corpus : Pitfalls to Avoid. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Budapest, Hungary.

HARRAT S., MEFTOUH K. & SMAÏLI K. (2018). Maghrebi Arabic dialect processing : an overview. *Journal of International Science and General Applications*, **1**.

HTAIT A., FOURNIER S. & BELLOT P. (2017). Identification semi-automatique de mots-germes pour l'analyse de sentiments et son intensité. In *COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference, Marseille, France, March 29-31, 2017. Proceedings.*, p. 415–424.

KARANASOU P. & LAMEL L. (2010). Comparing SMT methods for automatic generation of pronunciation variants. In *International Conference on Natural Language Processing*, p. 167–178 : Springer.

KILLER M., STUKER S. & SCHULTZ T. (2003). Grapheme based speech recognition. In *Eighth European Conference on Speech Communication and Technology*.

KIRITCHENKO S., MOHAMMAD S. & SALAMEH M. (2016). Semeval-2016 task 7 : Determining sentiment intensity of english and arabic phrases. In S. BETHARD, D. M. CER, M. CARPUAT, D. JURGENS, P. NAKOV & T. ZESCH, Éds., *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 42–51 : The Association for Computer Linguistics.

LE V.-B. & BESACIER L. (2009). Automatic speech recognition for under-resourced languages : application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, **17**(8), 1471–1482.

MASMOUDI A., BOUGARES F., ELLOUZE M., ESTÈVE Y. & BELGUITH L. (2018). Automatic speech recognition system for Tunisian dialect. *Language Resources and Evaluation*, **52**(1), 249–267. DOI : 10.1007/s10579-017-9402-y.

MEFTOUH K., HARRAT S., JAMOUSSI S., ABBAS M. & SMAILI K. (2015). Machine translation experiments on PADIC : A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, p. 26–34.

MEFTOUH K., HARRAT S. & SMAÏLI K. (2018). PADIC : extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey. HAL : hal-01718858.

TURNEY P. D. & LITTMAN M. L. (2003). Measuring praise and criticism : Inference of semantic orientation from association. *ACM Transactions on Information Systems*, **21**(4). DOI : 10.1145/944012.944013.

YASEEN M., ATTIA M., MAEGAARD B., CHOUKRI K., PAULSSON N., HAAMID S., KRAUWER S., BENDAHMAN C., FERSØE H., RASHWAN M., HADDAD B., MUKBEL C., MOURADI A., AL-KUFAISHI A., SHAHIN M., CHENFOUR N. & RAGHEB A. (2006). Building annotated written and spoken Arabic LRs in NEMLAR project. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy : European Language Resources Association (ELRA).

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

VerNom : une base de paires morphologiques acquise sur très gros corpus

Alice Missud[1, 2] Pascal Amsili[2] Florence Villoing[1]
(1) Modyco (UMR 7114 CNRS/Paris Nanterre)
(2) Lattice (UMR 8094 CNRS/Paris 3/ENS)
missud.a@parisnanterre.fr, pascal.amsili@gmx.fr,
villoing@parisnanterre.fr

RÉSUMÉ

Alors qu'une part active de la recherche en morphologie dérivationnelle s'intéresse à la compétition qui oppose les suffixations construisant des noms d'événement à partir de verbes (*-age, -ment, -ion, -ure, -ance, -ade, -aison*), l'accès à des données en large quantité devient nécessaire pour l'application de méthodes quantitatives. Dans l'optique de réunir des paires de verbes et de noms morphologiquement reliés dans le cadre de ces suffixations rivales, nous présentons VerNom, une base morphologique comprenant 25 857 paires verbe-nom, construite automatiquement à partir d'un corpus massif issu du web.

ABSTRACT

VerNom : a French derivational database acquired on a massive corpus

While a significant part of the litterature in word formation revolves around the competition between suffixations that construct event nouns from verbs in French (*-age, -ment, -ion, -ure, -ance, -ade, -aison*), accessing massive data is necessary for applying quantitative methods. With the purpose of gathering pairs of verbs and nouns that are morphologically related in the case of this competition, we present VerNom, a lexical database consisting of 25 875 verb-noun pairs acquired automatically from a massive web corpus.

MOTS-CLÉS : morphologie dérivationnelle, compétition morphologique, nominalisation, base lexicale.

KEYWORDS: word formation, morphological rivalry, nominalization, lexical database.

1 Introduction

Nous présentons VerNom, une ressource morphologique comprenant un ensemble de paires verbe-nom (et leurs fréquences d'apparition) collectées dans un large corpus issu du web. Les paires ont été appariées morphologiquement de façon automatique, et elles recouvrent les principales dérivations construisant des noms d'événement à partir de verbes : les suffixations en *-ion, -age, -ment, -ure, -ance, -ade* et *-aison*.

Les données utilisées proviennent de frCOW (Schäfer & Bildhauer 2012, Schäfer 2015), un corpus de 9 milliards de mots issu du web francophone. La constitution de cette ressource a pour objectif de donner accès à une grande quantité de données "authentiques" du français écrit, pour permettre

Volume 2 : Traitement Automatique des Langues Naturelles, pages 305–313.
HAL : hal-02784780.

une analyse quantitative et qualitative de la compétition entre nominalisations rivales du français (construisant des noms déverbaux événementiels), qui constitue une part active de la recherche actuelle en morphologie dérivationnelle (Lüdtke 1978 ; Martin 2010 ; Ferret *et al.* 2010 ; Uth 2010 ; Ferret & Villoing 2012 ; Uth 2016 ; Fradin 2014, 2019 ; Dal *et al.* 2018 : Wauquier *et al.* 2018, 2019).

VerNom s'inscrit dans un contexte marqué par une relative carence, pour le français, de ressources purement dérivationnelles permettant des recherches avec des méthodes de modélisation quantitative. En effet, alors que les travaux sur la compétition morphologique tendent de plus en plus à s'appuyer sur ce type de méthode (par exemple, Wauquier *et al.* 2018, Bonami & Thuilier 2019), on recense principalement deux ressources exploitables pour le français, le corpus Nomage (Balvet *et al.* 2011) et la base lexicale Démonette (Hathout & Namer 2014, Namer *et al.* 2019), qui s'appuie elle-même sur Verbaction (Hathout & Tanguy 2002). Bien que ces ressources soient extrêmement précieuses, en particulier parce qu'elles fournissent des annotations fines tant syntaxiquement que sémantiquement, voire morphologiquement et phonologiquement pour la seconde (qui en outre construit tout un réseau dérivationnel), elles ne répondent pas entièrement aux besoins d'une analyse quantitative à gros grain. En effet, la minutie des informations renseignées par ces ressources les conduit à traiter un jeu de données assez réduit (736 noms événementiels et 679 verbes pour Nomage et 8848 paires V/N d'action pour Démonette). Par ailleurs, leurs lexiques n'enregistrent que peu de néologismes, indispensables au calcul de la productivité des schémas morphologiques. Cette lacune tient au fait qu'ils sont issus soit du journal *Le Monde* (cf. Nomage, basé sur le French Treebanck d'Abeillé *et al.* 2003), soit de dictionnaires (cf. Démonette qui compte 6765 paires V/N d'action tirées de TLFnome). L'ensemble de ces contraintes nous a conduit à élaborer une nouvelle ressource qui réponde à ces objectifs spécifiques.

À la suite de Hathout *et al.* (2009), qui ont exploité le web pour collecter une liste extensive de paires verbe-nom d'action (notamment les suffixations en *-age, -ment* et *-ion*) en procédant à l'aide de règles, nous proposons de réactualiser les méthodes en nous basant sur un corpus plus massif et plus récent (2016) afin de récolter une plus grande quantité de données, plus diverses, regroupant toutes les suffixations déverbales construisant des noms d'événement, et qui permette en outre d'explorer de nombreux néologismes.

Dans cet article, nous décrivons la ressource en détail, les méthodes d'extraction des noms et des verbes et leur appariement, ainsi que l'évaluation des paires collectées. Enfin, nous présentons une brève étude de la productivité des schémas morphologiques basée sur la ressource.

2 Méthodologie

2.1 Extraction des données

La base a été constituée à partir de frCOW16 (Schäfer & Bildhauer 2012, Schäfer 2015), un corpus massif du français issu du web datant de 2016 et comprenant 9 milliards de mots. Les corpus COW sont catégorisés et lemmatisés, généralement avec TreeTagger, de telle sorte que tous les tokens reçoivent une catégorie (il n'y a pas de POS-tag 'unknown') mais pas forcément toujours un lemme. Nous avons fait le choix d'exploiter directement ces informations de catégorie et lemmatisation, sans en contrôler au préalable la qualité. Puisque nous cherchions à apparier des verbes et des noms reliés morphologiquement, nous avons extrait toutes les formes catégorisées 'verbe' ou 'nom', que ces formes aient été lemmatisées ou non (les formes non lemmatisées nous intéressent particulièrement

puisqu'elles peuvent correspondre à des néologismes qui échappent aux lexiques construits à partir de dictionnaires). La table 1 présente des exemples d'entrées du corpus frCOW comprenant des formes lemmatisées et non lemmatisées. Ces deux groupes de données ont été traités séparément dans les phases suivantes.

forme	catégorie	lemme
mangera	VER	manger
voiture	NOM	voiture
pourcenter	VER	(unknown)
pipolisation	NOM	(unknown)

TABLE 1 – Exemples d'entrées de frCOW

2.2 Nettoyage des formes

Nous avons dû procéder à toutes les étapes à l'élimination de lemmes/formes contenant des caractères spéciaux, de la ponctuation ou des séquences impossibles.

Formes lemmatisées Pour les verbes lemmatisés, les formes associées à des lemmes se terminant en *-er*, *-ir* ou *-re* ont été récupérées. Pour les noms lemmatisés, les formes extraites ont un lemme se terminant par l'une des 7 suffixations constructrices de noms d'événement à savoir *-age, -ment, -ion, -ure, -ance, -ade* et *-aison*. La suffixation en *-erie*, susceptible de construire des noms d'événement également, a cependant été mise de côté en raison de la difficulté de différencier automatiquement son homonyme plus productif dérivant des noms de lieux à partir de noms. Au total, 25 209 verbes et 23 200 noms lemmatisés ont été extraits.

Formes non lemmatisées La non-lemmatisation de certaines formes dans frCOW peut dépendre de plusieurs problèmes : les formes peuvent être mal catégorisées, et donc ne pas correspondre avec un lemme (par exemple : *expressément*, catégorisé 'nom' plutôt que 'adverbe'), elles peuvent également correspondre à des variantes orthographiques qui les distancient d'un lemme (par exemple : *develloppé* plutôt que *développé*), ou encore constituer des néologismes non répertoriés (comme *pipolisation*). Nous avons tenté d'une part de rassembler les variantes orthographiques avec les bons lemmes, et d'autre part de lemmatiser les formes susceptibles de correspondre à des néologismes.

Seules les formes étiquetées 'verbe' avec une finale en *-er* ont été extraites parmi les verbes non lemmatisés, car la construction de verbes irréguliers du deuxième et troisième groupe est peu probable parmi les néologismes. Pour les noms, seules les formes en *-ion, -age, -ment, -ure, -ance, -ade* et *-(a)ison* ont été récoltées. Afin d'obtenir les fréquences des formes fléchies de ces entrées, nous avons utilisé des expressions régulières pour retrouver dans frCOW des formes fléchies pour chaque verbe et chaque nom. Les fréquences des formes fléchies trouvées ont été ajoutées aux lemmes. Ceci a permis, par exemple pour *wikifier*, de passer d'une fréquence de 11 à 18. Pour réunir les variantes orthographiques d'un même lemme, nous avons d'abord tenté d'ignorer les diacritiques, sujets à de nombreuses erreurs d'orthographe en français, afin de regrouper ensemble les formes non lemmatisées avec les bons lemmes. Nous avons supprimé tous les diacritiques des deux lexiques, et avons regardé si parmi les formes non préalablement lemmatisées se trouvait un équivalent identique chez les formes

lemmatisées. Ceci a permis d'ajouter des fréquences aux noms et verbes lemmatisés comme *étudier* par exemple, dont la fréquence a été augmentée de 5 occurrences. Au total, 96 770 verbes et 130 341 noms non préalablement lemmatisés ont été récoltés. L'ensemble des verbes et des noms décrits a ensuite servi à l'appariement morphologique. L'influence des différentes étapes de nettoyage sur les proportions des lexiques des formes non lemmatisées est détaillée dans la table 2.

Mots-forme non lemmatisés	VERBES	NOMS
Se terminent en *-er*	140 036	-
Se terminent en *-ion, -age, -ment, -ure, -ance, -ade, -aison*	-	225 177
Retrait des caractères spéciaux	97 275	132 498
Regroupement des lemmes doublons	96 937	132 300
Regroupement des lemmes mal orthographiés (diacritiques)	96 770	130 341

TABLE 2 – Nettoyage des formes non lemmatisées

2.3 Appariement morphologique

De nombreux travaux se sont intéressés à la question de l'appariement morphologique automatique, en particulier dans le domaine de la recherche d'information. Les méthodes consistent principalement à constituer des ensembles de formes partageant les mêmes racines, en réunissant à la fois les formes fléchies, les lexèmes simples et les lexèmes construits (dérivation ou composition). Par exemple, il s'agira de regrouper *chanter* avec *chant*, *chanteur* et leurs formes fléchies. Etant donné la prise en compte d'allomorphes et de multiples schémas de construction morphologique (préfixation, suffixation, conversion, composition), les travaux proposent l'utilisation de règles symboliques (Gaussier 1999, pour la dérivation en français), de mesures de similarité sémantique entre lexèmes (Schone & Jurafsky 2000), ou encore d'algorithmes de clusterisation non-supervisés (Singh & Gupta 2019) qui ne nécessitent pas de connaissances linguistiques préalables. En ce qui concerne la constitution de ressources dérivationnelles, les approches à base de règles induites à partir de connaissances ont montré leur intérêt pour l'allemand (Zeller *et al.* 2013).

Dans notre cas, nos objectifs sont moins ambitieux ; nous ciblons des suffixes de nominalisation spécifiques et cherchons à extraire des paires plutôt que des ensembles. En ce sens, et parce que nous cherchons à réunir un très large ensemble de paires avec le moins de bruit possible, pour l'appariement morphologique, nous optons pour une approche à base de règles.

Troncation des formes Afin d'apparier les noms et les verbes, nous avons tronqué les formes de leurs suffixes (*-age, -ment, -ion, -ure, -ance, -ade* et *-(a)ison*) ou de leur finale verbale (*-er*, *-ir* ou *-re*) dans le but de faire correspondre leurs radicaux. Nous avons également généré des radicaux susceptibles d'être sujets à allomorphie. Toutes les troncations possibles ont été gardées pour un même lemme, à condition que la longueur de la forme tronquée ne soit pas inférieure à 2 caractères.

Appariement Les formes tronquées verbales et nominales ont été appariées sur la base d'une identité immédiate. De nombreux noms en *-ion* correspondant à des verbes à la deuxième personne du pluriel de l'imparfait (*entendion* plutôt que *entendions*) ont été évincés en cherchant pour chaque forme en *-ion* si un équivalent avec un *-s* existait parmi les formes fléchies du verbe base dans le

Glàff (Sajous *et al.* 2013). Pour les formes en *-ment* qui se trouvaient être des adverbes faussement étiquetés 'nom', nous nous sommes servis de l'ensemble des formes étiquetées 'adverbe' dans frCOW. Si la fréquence de la forme étiquetée 'adverbe' était plus élevée que la fréquence de la forme étiquetée 'nom', nous avons supprimé la paire comprenant le nom en *-ment* de la base. Afin de récupérer des paires incluant des radicaux allomorphiques et supplétifs que nos méthodes ne parvenaient pas à collecter, nous avons exploité la base lexicale Démonette (Hathout & Namer, 2014). Seules les paires qui ne figuraient pas déjà dans notre base ont été conservées. L'ajout de ces données a permis d'enrichir la base de 1 380 nouvelles paires (avec leurs fréquences dans frCOW quand les verbes et noms s'y trouvent). Enfin, en raison de nombreuses paires appariées à la suite d'erreurs orthographiques, nous avons retiré toutes les paires comprenant des noms non préalablement lemmatisés lorsque des équivalents lemmatisés avec une distance de Levenshtein de 1 étaient présents dans la base. Le détail des proportions de paires obtenues selon les étapes est décrit dans la table 3.

	Nombre de paires
Identité immédiate	40 940
Nettoyage des lemmes mal orthographiés	36 519
Retrait des noms en *-ment* mal étiquetés	33 478
Ajout des paires de Démonette	34 858
Distances de Levenshtein (1)	27 857

TABLE 3 – Etapes pour l'appariement verbe-nom

3 Description de la ressource

Au total, la base est constituée de 25 857 paires verbe-nom dont les dérivés nominaux sont issus des suffixations en *-ion, -age, -ment, -ance, -ure, -ade* et *-aison*. Pour chaque paire sont données les fréquences du verbe et du nom dans frCOW, leur provenance (préalablement lemmatisé 'lemmatized' ou non 'nolemma' dans frCOW ou bien issu de Démonette) ainsi que le suffixe à l'origine de la dérivation. La table 4 présente 2 entrées : la paire comprenant le dérivé le plus fréquent, et une paire comprenant un verbe et un nom n'apparaissant chacun qu'une fois dans le corpus.

verbe	freq_verbe	origine_verbe	nom	freq_nom	suffixe	origine_nom
former	662730	lemmatized	formation	3966933	ion	lemmatized
kikouloler	1	nolemma	kikoulolage	1	age	nolemma

TABLE 4 – Exemples d'entrées de la base

Distribution Le détail des proportions de paires par suffixe est donné dans la table 5. La base réunit une majorité de dérivés en *-ion, -age* et *-ment*, ceux-ci étant à eux-seuls à l'origine de 85% des paires. Les suffixations en *-ance* et *-ure*, bien moins nombreuses, représentent un peu plus de 10% des données, tandis que seuls 4,3% des paires concernent les suffixations en *-ade* et *-aison*.

Evaluation La base a été évaluée en procédant à 4 tirages aléatoires de 100 paires (table 6, gauche). Pour chaque tirage, les paires ont été annotées "correcte" ou "incorrecte" par un seul annotateur. Deux

	Nombre de paires	Proportion
-ion	10 558	40,8%
-age	6 588	25,4%
-ment	4 865	18,8%
-ance	1 439	5,5%
-ure	1 252	4,8%
-ade	771	2,9%
-aison	384	1,4%
Total	25 857	100%

TABLE 5 – Distribution des paires selon le suffixe

évaluations différentes ont été réalisées sur chaque tirage. Pour la première (*évaluation stricte*), les paires incorrectes pouvaient répondre à des erreurs de catégorisation de la base (nom plutôt que verbe, le plus souvent), à des erreurs d'orthographe sur l'une des deux formes ou sur les deux, des erreurs de langue (italien, créoles à base française, anglais, latin, ancien ou moyen français), ou encore à une correspondance sémantique trop opaque. Pour la deuxième évaluation (*évaluation relâchée*), nous avons considéré que les fautes d'orthographe ne constituaient pas des erreurs d'appariement si les deux formes comportaient les mêmes erreurs (par exemple : *anhiler* → *anhilation*). Les mêmes évaluations ont également été réalisées pour chaque suffixe (table 6, droite, 100 paires par suffixe).

	évaluation stricte	évaluation relâchée
1	75%	88%
2	79%	90%
3	84%	89%
4	78%	81%
moyenne	79%	87%

	évaluation stricte	évaluation relâchée
-ion	75%	89%
-age	86%	96%
-ment	77%	91%
-ance	54%	72%
-ure	62%	68%
-ade	48%	62%
-aison	53%	61%

TABLE 6 – Scores d'exactitude selon le type d'évaluation

Productivité des schémas morphologiques Grâce à la quantité de paires qu'elle regroupe, la base permet en outre de calculer la productivité globale des suffixes, notamment en appliquant la mesure de Baayen (1994). Avec l'idée que les suffixes les plus productifs vont former de nombreux dérivés peu fréquents, cette mesure calcule la productivité d'un suffixe E en calculant le rapport entre le nombre d'hapax suffixés par E et le nombre total d'hapax dérivés dans le même corpus. Plutôt que de donner cette mesure de productivité, nous donnons à la figure 1 les comptages des différents hapax dans notre base, ce qui permet de dessiner les premiers contours de la répartition du lexique par ces schémas en compétition : les suffixations en *-ion, -age* et *-ment* apparaissent comme les plus productives dans nos données, tandis que les suffixations en *-ance, -ure, -ade* et *-aison* se démarquent par leur faible productivité en comparaison.

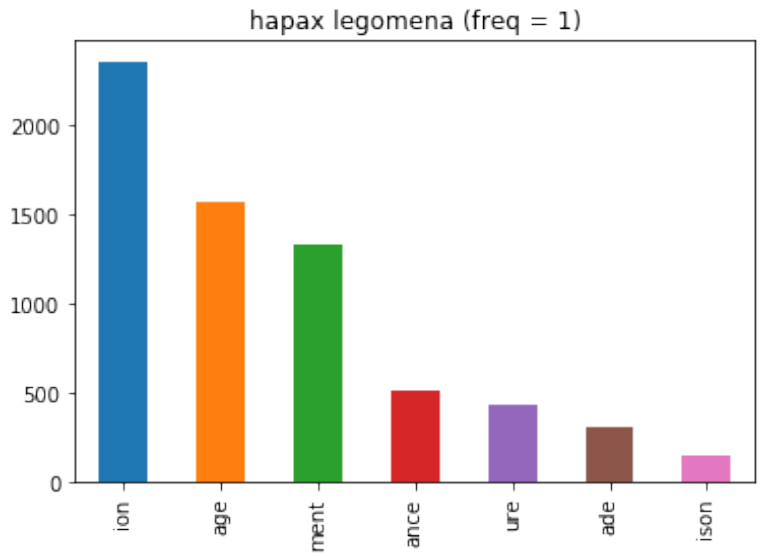


FIGURE 1 – Productivité des suffixes (Baayen (1994), hapax)

4 Conclusion

Cette ressource a été constituée dans le but d'explorer quantitativement les raisons de la coexistence de schémas morphologiques rivaux en français dans le cadre de la compétition entre les suffixes constructeurs de noms d'événement à partir de verbes. A notre connaissance, il s'agit de la base lexicale la plus large construite intégralement à partir de données issues du web pour les nominalisations issues du français écrit. Sa qualité pourrait toutefois faire l'objet d'améliorations. Les évaluations de la base ont montré que les erreurs orthographiques constituaient la majorité des erreurs uniquement pour les suffixations les plus fréquentes et les plus productives, à savoir *-ion*, *-age* et *-ment*. En ce sens, le nettoyage des lexiques pourrait être amélioré afin de rassembler au mieux les diverses variantes orthographiques d'une forme sous un même lemme. L'appariement morphologique pourrait en outre bénéficier de méthodes issues de la sémantique distributionnelle (Schone & Jurafsky 2000, Wauquier *et al.* 2018 pour les suffixations en *-age, -ment* et *-ion*). Etant donné les analyses proposées dans la littérature (Tribout 2010, Tribout & Villoing 2014), la base mériterait également d'être enrichie par l'ajout de paires verbe-nom issues de la conversion (*marcher* → *marche*, *découvrir* → *découverte*, *arriver* → *arrivée*), elle aussi en compétition dans la construction de noms d'événement en français. La conversion, non régulière sur le plan formel, pose d'autres enjeux pour l'extraction et l'appariement automatiques, qui constitueront la prochaine étape de ce travail. La ressource est disponible sur le site d'Ortolang sous le nom VerNom.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Éd., *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Dordrecht : Springer Netherlands. DOI : 10.1007/978-94-010-0201-1_10.

BAAYEN R. H. (1994). Productivity in language production. *Language and Cognitive Processes*, **9**(3), 447–469. DOI : 10.1080/01690969408402127.

BALVET A., BARQUE L., CONDETTE M.-H., HAAS P., HUYGHE R., MARÍN R. & MERLO A. (2011). Nomage : an electronic lexicon of French deverbal nouns based on a semantically annotated corpus. In *WoLeR 2011 at ESSLLI, International Workshop on Lexical Resources*, p. 8–15, Ljubljana, Slovenia. HAL : halshs-01078047.

BONAMI O. & THUILIER J. (2019). A statistical approach to rivalry in lexeme formation : French-iser and-ifier. *Word Structure*, **12**(1), 4–41.

DAL G., HATHOUT N., LIGNON S., NAMER F. & TANGUY L. (2018). Toile versus dictionnaires : Les nominalisations du français en-age et en-ment. In *SHS Web of Conferences*, volume 46, p. 08003 : EDP Sciences.

FERRET K., SOARE E. & VILLOING F. (2010). Rivalry between french–age and–ée : the role of grammatical aspect in nominalization. In M. ALONI, H. BASTIAANSE, T. DE JAGER & K. SCHULTZ, Éds., *Logic, language and meaning, 17th Amsterdam Colloquium, The Netherlands, December 2009, Revised Selected Papers, Lecture Notes in Computer Science (Vol. 6042)*, p. 284–295. Berlin : Springer.

FERRET K. & VILLOING F. (2012). L'aspect grammatical dans les nominalisations en français : les déverbaux en -age et -ée. *Lexique (20)*, p. 73–127.

FRADIN B. (2014). La variante et le double. In F. VILLOING, S. DAVID & L. SARAH, Éds., *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, p. 109–147. Presses Universitaires de Paris Ouest.

FRADIN B. (2019). Competition in derivation : What can we learn from french doublets in-age and-ment ? In F. RAINER, F. GARDANI, W. U. DRESSLER & H. C. LUSCHÜTZKY, Éds., *Competition in Inflection and Word-Formation. Studies in Morphology, vol 5.*, p. 67–93. Springer, Cham.

GAUSSIER É. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, p. 24–30.

HATHOUT N. & NAMER F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, **11**(5), 125–168.

HATHOUT N., SAJOUS F. & TANGUY L. (2009). Looking for french deverbal nouns in an evolving web (a short history of wac). In *Proceedings of the fifth workshop on Web As Corpus (WAC), San Sebastian, September 7th*, p. 37–44.

HATHOUT N. & TANGUY L. (2002). Webaffix : Discovering Morphological Links on the WWW. In *LREC 2002*, Proceedings of LREC, Las Palmas, Spain. HAL : halshs-01322326.

LÜDTKE J. (1978). *Prädikative Nominalisierungen mit Suffixen im Katalanischen, Spanischen und Französischen*. Tübingen : Niemeyer.

MARTIN F. (2010). The semantics of eventive suffixes in French. In M. RATHERT & A. ALEXIADOU, Éds., *The Semantics of Nominalizations across Languages and Frameworks*, p. 109–141. Berlin : Mouton de Gruyter.

NAMER F., BARQUE L., BONAMI O., HAAS P., HATHOUT N. & TRIBOUT D. (2019). Demonette2 – A large scale derivational database for French : first results. In *TALN*, Toulouse, France. HAL : halshs-02275652.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.

SCHONE P. & JURAFSKY D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, p. 67–72 : Association for Computational Linguistics.

SCHÄFER R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. BANSKI, H. BIBER, E. BREITENEDER, M. KUPIETZ, H. LÄNGEN & A. WITT, Éds., *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster : UCREL IDS.

SCHÄFER R. & BILDHAUER F. (2012). Building large corpora from the web using a new efficient tool chain. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOÄŸAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, p. 486–493, Istanbul, Turkey : European Language Resources Association (ELRA).

SINGH J. & GUPTA V. (2019). A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics. *Knowledge-Based Systems*, **180**, 147–162.

TRIBOUT D. (2010). *Noun to verb and verb to noun conversions in French*. Thèse de doctorat, Université Paris Diderot (Paris 7). HAL : tel-01577528.

TRIBOUT D. & VILLOING F. (2014). La composition VN et la conversion V>N en français : un nouveau cas de concurrence morphologique ? In F. VILLOING, S. DAVID & S. LEROY, Éds., *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Presses Universitaires de Paris Ouest. HAL : halshs-01690227.

UTH M. (2010). The rivalry of the French nominalization suffixes -age and -ment from a diachronic perspective. In M. RATHERT & A. ALEXIADOU, Éds., *The Semantics of Nominalizations across Languages and Frameworks*, p. 215–244. Berlin : Mouton de Gruyter.

UTH M. (2016). The competition of event nominalization procedures of French, in comparison with German. *Zeitschrift für Romanische Philologie*, **132**(1), 58–89.

WAUQUIER M., FABRE C. & HATHOUT N. (2018). Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. In *Congrès Mondial de Linguistique Française (CMLF)*, volume 46 de *6e Congrès Mondial de Linguistique Française*, Mons, Belgium : EDP Sciences. DOI : 10.1051/shsconf/20184608006, HAL : hal-01876027.

WAUQUIER M., HATHOUT N. & FABRE C. (2019). Contributions of distributional semantics to the semantic study of french morphologically derived agent nouns. In *Mediterranean Morphology Meetings*, volume 12, p. 111–121.

ZELLER B., ŠNAJDER J. & PADÓ S. (2013). Derivbase : Inducing and evaluating a derivational morphology resource for german. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1201–1211.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude des variations sémantiques à travers plusieurs dimensions

Syrielle Montariol[1, 2] Alexandre Allauzen[3]
(1) LIMSI, CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, F-91405 Orsay, France
(2) Société Générale, 17 Cours Valmy 92043 Puteaux, France
(3) ESPCI, Univ. Paris Dauphine - PSL, 75016 Paris
syrielle.montariol@limsi.fr, alexandre.allauzen@espci.psl.eu

RÉSUMÉ

Au sein d'une langue, l'usage des mots varie selon deux axes : diachronique (dimension temporelle) et synchronique (variation selon l'auteur, la communauté, la zone géographique...). Dans ces travaux, nous proposons une méthode de détection et d'interprétation des variations d'usages des mots à travers ces différentes dimensions. Pour cela, nous exploitons les capacités d'une nouvelle ligne de plongements lexicaux contextualisés, en particulier le modèle BERT. Nous expérimentons sur un corpus de rapports financiers d'entreprises françaises, pour appréhender les enjeux et préoccupations propres à certaines périodes, acteurs et secteurs d'activités.

ABSTRACT

Studying semantic variations through several dimensions

In a language, word usage can vary across two axis : diachronic (time evolution), and synchronic (variation across sources, authors, communities...). In this work, we propose to leverage the capacity of contextualised embeddings models to analyse financial texts along these two axes of variation. Starting from a corpus of annual company reports spanning 20 years, we explore the ability of the language model BERT to extract interpretable variations in word usage in order to understand the stakes and concerns of specific periods, financial actors and sectors.

MOTS-CLÉS : Diachronie, Variation sémantique, Plongements lexicaux contextualisés, Clustering.

KEYWORDS: Diachrony, Semantic variation, Contextualised Embeddings, Clustering.

1 Introduction

L'usage des mots varie selon deux axes : diachronique (dimension temporelle) (Aitchison, 2001) et synchronique (en fonction de l'auteur, la communauté, la zone géographique...). Ces variations peuvent être d'envergure et d'ordre générationnel et culturel ; ou d'ampleur plus réduite, pour des variations à court terme et entre individus ou entités. Dans ce cas, elles peuvent être le résultat d'événements qui, sans altérer la signification du mot, changent ponctuellement son usage et sa connotation. Elles révèlent alors des divergences d'intérêts et de préoccupations entre les individus.

Dans cet article, nous nous concentrons sur le domaine financier et un corpus de rapports financiers d'entreprises. Dans ce cadre, modéliser les évolutions temporelles d'usage des mots peut permettre une meilleure appréhension des enjeux et préoccupations de chaque période (Matthew Purver & Pollak, 2018) tandis que les opinions, comportements et préoccupations des différents acteurs financiers peuvent transparaître à travers la façon dont ils utilisent les mots. En d'autres termes, nous cherchons

Volume 2 : Traitement Automatique des Langues Naturelles, pages 314–322.
HAL : hal-02784781.

à détecter des “signaux faibles" en analysant les variations d'usage des mots. Un signal faible est une information observée à partir de données, qui peut avoir une interprétation et des conséquences ambiguës, mais peut revêtir de l'importance pour la compréhension d'événements présents ou futurs. Les variations diachroniques et synchroniques au sein d'un corpus de documents sont difficilement repérables par les analystes financiers ; mais elles peuvent révéler des informations précieuses en tant que potentiels signaux faibles de changement dans l'opinion ou la situation d'un acteur financier. Par exemple, l'évolution de la connotation du vocabulaire employé dans les communications des banques centrales (BCE et Fed) est liée à la situation économique de la période (Buechel *et al.*, 2019).

Avec la numérisation de textes historiques, les méthodes de traitement automatique des langues (TAL) pour l'analyse diachronique se sont développées rapidement ces dernières années (Tahmasebi *et al.*, 2018). De nombreux modèles reposent sur des plongements de mots statiques comme Word2Vec (Mikolov *et al.*, 2013). Ils rassemblent les différents usages d'un mot dans un unique vecteur, ce qui rend difficile une analyse des variations de contexte et d'usage. Plus récemment, les plongements de mot contextualisés sont apparus comme BERT (Devlin *et al.*, 2019) ou ELMO (Peters *et al.*, 2018). Ce type de modèles renouvellent les perspectives pour l'analyse des variations sémantiques. Dans cet article, nous utilisons le modèle BERT qui montre une nette supériorité pour la tâche de désambiguïsation sémantique par rapport à ELMO et Flair (Wiedemann *et al.*, 2019). Nous nous appuyons sur FlauBERT, une version de BERT pour le français, pour proposer une méthode de détection et d'interprétation des variations d'usages de mots dans un corpus selon plusieurs dimensions. La méthode est décrite dans la section 3 puis appliquée (section 4) sur un corpus français de rapports financiers annuels d'entreprises (section 4.1).

2 État de l'art

Avant que l'emploi de représentations vectorielles de mots ne se généralise, la mesure de changement sémantique reposait sur la détection de variation dans les co-occurences de mots (Sagi *et al.*, 2009). Puis des modèles de plongements de mots diachroniques se sont développés, reposant sur l'hypothèse qu'un changement dans le contexte d'un mot reflète une évolution de sa signification et son usage. Le plus souvent, ils impliquent de diviser un corpus en strates temporelles puis d'apprendre des plongements lexicaux (Mikolov *et al.*, 2013) pour chaque mot, dans chaque strate. Pour cela, deux méthodes populaires sont l'apprentissage incrémental (Kim *et al.*, 2014) et l'alignement d'espaces vectoriels (Hamilton *et al.*, 2016). Néanmoins, ces méthodes rassemblent l'ensemble des significations et usages possibles d'un mot dans un unique vecteur, à chaque strate temporelle.

En parallèle, l'analyse de variations sémantiques dans le cas synchronique est faite en majorité à partir de méthodes de désambiguïsation sémantique. Certains auteurs utilisent des mesures de similarité entre plongements de mots pour analyser les variations de vocabulaire entre plusieurs communautés (Tredici & Fernández, 2017). Plus récemment, Schlechtweg *et al.* (2019) analysent à la fois les variations synchroniques et diachroniques dans des corpus en utilisant des plongements lexicaux et une méthode d'alignement d'espaces de représentations.

Les plongements de mots contextualisés comme BERT (Devlin *et al.*, 2019) et ELMO (Peters *et al.*, 2018) permettent à chaque occurrence d'un mot d'être représentée par un vecteur dépendant de son contexte. Pré-entraînés sur des corpus volumineux, ils améliorent l'état de l'art dans de nombreuses tâches de TAL et commencent à être utilisés pour la détection de changement sémantique. Dans un premier temps, BERT fut utilisé de façon supervisés (Hu *et al.*, 2019) pour déterminer l'évolution de

la distribution des sens d'un mot au cours du temps. Néanmoins, cette méthode implique de travailler sur un nombre réduit de mots dont l'ensemble des sens est connu l'avance. Nous explorons l'usage de méthodes de partitionnement sur l'ensemble des représentations des occurrences d'un mot pour extraire automatiquement ses sens possibles (Giulianelli *et al.*, 2019; Martinc *et al.*, 2020a).

3 Modélisation des variations avec FlauBERT

BERT (Bidirectional Encoder Representations from Transformers) est un modèle de représentation des mots dans leurs contextes d'utilisation (typiquement une phrase). Son utilisation repose sur l'apprentissage par transfert : pré-entraîner un modèle sur une tâche non-supervisée à partir d'un très grand volume de données, avant de le raffiner sur une nouvelle tâche. Ici, nous utilisons ce type de modèle directement pour représenter l'occurrence d'un mot dans une séquence. Notre étude portant sur des données français, nous utilisons le modèle FlauBERT (Le *et al.*, 2020) qui est une variante des modèles BERT et RoBERTa (Liu *et al.*, 2019) entraînée sur des textes français.

3.1 Détection des variations

FlauBERT permet de prédire la représentation spécifique au contexte de chaque occurrence d'un mot. Afin d'en détecter les variations d'usage, nous proposons une méthode en quatre étapes.

Extraction des plongements contextualisés
Le corpus est d'abord segmenté en phrases, chacune d'elle étant assortie de méta-données décrivant son document d'origine, l'auteur et la date. Pour chaque mot du vocabulaire sélectionné, nous extrayons l'ensemble des plongements lexicaux contextualisés de ses occurrences dans le corpus à l'aide de FlauBERT. Le vecteur extrait correspond à la dernière couche du modèle pré-entraîné.

Détection des mots subissant des variations
Les travaux précédents sélectionnent manuellement une faible quantité de mots à analyser (Giulianelli *et al.*, 2019; Hu *et al.*, 2019). À l'inverse, nous souhaitons détecter automatiquement les mots qui subissent une variation significative. Pour cela, nous utilisons une approche similaire à celle de Martinc *et al.* (2020b). Pour chaque dimension, nous calculons une métrique reposant sur la moyenne des représentations vectorielles des occurrences d'un mot calculée sur l'ensemble du corpus. Prenons le cas des variations diachroniques, la moyenne des représentations observées sur tous le corpus u_{corpus} est comparée à la moyenne estimée sur chaque strate temporelle u_{t} en utilisant la distance cosinus. La sélection se fait par seuillage sur la moyenne des distances.

Partitionnement des usages
Pour chaque mot sélectionné, nous appliquons un algorithme de partitionnement sur l'ensemble des représentations vectorielles de ses occurrences. Nous retenons pour ce faire deux algorithmes en particulier : K-Means et propagation par affinité. La méthode de propagation par affinité (Frey & Dueck, 2007), moins classique que la méthode K-Means, a fait ses preuves dans la littérature de la désambiguïsation du sens des mots (Alagić *et al.*, 2018), une tâche proche de notre objectif. Elle a l'avantage de ne pas nécessiter de sélectionner manuellement le nombre de clusters.

On notera que les représentations inférées par FlauBERT contiennent des informations sémantiques mais aussi syntaxiques (Coenen *et al.*, 2019). Par construction, les clusters obtenus ne regroupent donc pas les différents sens d'un mot, mais plus largement les différentes manières dont il est utilisé.

Matrice de distributions d'usages
Pour chaque mot, une matrice est déduite afin de représenter les distribution normalisée des clusters (donc des usages) au travers d'une dimension de variation. Par exemple, la figure 1 représente une telle matrice pour les variations temporelles et sectorielles.

3.2 Interprétation des variations

Les matrices de distributions d'usages de mots permettent d'extraire différentes informations :
1) À quel point l'usage d'un mot varie pour une dimension donnée ?
2) À quelle période, pour quel auteur, quelle classe, se produit la variation ?
3) Quel usage apparaît / disparaît ? Comment interpréter ce changement ?

Pour le premier élément, nous utilisons la divergence de Jensen-Shannon (JSD), une mesure de comparaison de deux distributions de probabilité, ainsi que sa généralisation à n distributions de probabilités $d_1, d_2, \ldots, d_n$ proposée par Ré & Azad (2014) (H est la fonction entropie) :

$$\mathrm{JSD}(d_1, d_2, \ldots, d_n) = H\left(\frac{\sum_{i=1}^{n} d_i}{n}\right) - \frac{\sum_{i=1}^{n} H(d_i)}{n} \quad (1)$$

Pour le second élément, chaque distribution des usages est comparée avec la distribution moyenne de toute la dimension. Par exemple dans le cadre diachronique, la moyenne élément-à-élément des distributions de l'ensemble des périodes est calculée. Cette distribution moyenne est comparée à celle de chaque période à l'aide de la JSD. Puis il convient de déterminer les clusters / usages concernés : ceux qui varient le plus, apparaissent ou disparaissent au cours du temps, ou sont propres à certains acteurs uniquement. Pour cela, on recherche les clusters qui se répartissent de manière inégale selon la dimension de variation.

Pour finir, une fois les clusters cibles identifiés, interpréter les usages qui y sont associés se fait de deux façons. Tout d'abord, nous faisons l'hypothèse que l'exemple (la phrase, dans notre cas) le plus proche du centroïde est représentatif du contexte des occurrences du mot dans le cluster analysé. Nous comparons donc ces phrases entre les différents clusters pour avoir une idée préliminaire de l'usage du mot dans son contexte.

Ensuite, nous mettons en place une méthode de détection de mots-clés pour caractériser les différents clusters les uns par rapport aux autres. La méthode repose sur le principe du tf-idf (Term Frequency - Inverse Document Frequency). Un cluster étant constitué d'un ensemble de phrase, nous considérons chaque cluster comme un document et l'ensemble des clusters comme un corpus. Le but est de déterminer les mots ayant de l'importance pour un document mais pas dans le reste du corpus, c'est-à-dire les mots les plus discriminants pour chaque cluster. Les mots-outils sont exclus ; nous éliminons aussi tous les mots qui apparaissent déjà dans plus de 50% des clusters. Nous calculons ensuite les scores de tf-idf des mots dans chaque cluster. Les mots ayant les scores les plus importants sont utilisés comme mots-clés pour caractériser et faciliter l'interprétation des clusters.

4 Expérimentation

Nous appliquons la méthode décrite à un corpus du domaine financier, pour détecter les variations d'usages de mots à travers plusieurs dimensions en plus de celle du temps. Pour nos expériences, nous

utilisons le modèle pré-entraîné et librement disponible FlauBERT-base-uncased (Le *et al.*, 2020).

4.1 Corpus

Les données utilisées sont issues du corpus financier CoFiF [1] (Daudert & Ahmadi, 2019). Il est composé des rapports financiers des 60 plus grandes entreprises françaises appartenant aux indices boursiers CAC40 et CAC Next 20. Il comporte plus de 5 millions de phrases dans 2655 rapports de différents types (rapports trimestriel, semestriel, annuel, et document de référence), de 1995 à 2018. Une particularité du corpus est la présence de tableaux de données au format brut dans le texte. Afin d'écarter ces éléments de l'analyse, lors de la division du corpus en phrases, nous excluons les phrases composées de moins de 70% de lettres (plus de 30% de chiffres, symboles et espaces). Pour finir, nous nous concentrons sur les documents de référence (DR), qui constituent presque 85% du volume de données. Ils sont publiés chaque année par les entreprises et résument leur situation financière et perspectives. Ainsi, nous aboutissons à un corpus constitué d'environ 2.7 millions de phrases. Chacune de ces phrases est associée aux méta-données du document dont elle est extraite : le nom de l'entreprise, et l'année de publication du rapport. Ce sont deux dimensions pour l'analyse de variations d'usage des mots : diachronique (par année) et synchronique (par entreprise). L'axe synchronique est étoffé en collectant des informations sur les entreprises : leur domaine d'activité (luxe, transport, chimie, ...) et leur secteur (secondaire ou tertiaire). Les deux secteurs sont équilibrés (1.4 millions de phrases pour le secteur tertiaire, et 1.3 millions pour le secondaire).

Parmi l'ensemble du vocabulaire, nous conservons les 10 000 mots les plus fréquents. Nous en excluons les mots-outils et sélectionnons les mots qui sont uniquement des noms [2]. En effet, le modèle BERT est fortement influencé par la catégorie grammaticale d'un mot (Coenen *et al.*, 2019) ; réduire ainsi l'analyse permet de limiter l'impact de la catégorie grammaticale lors du partitionnement.

4.2 Résultats

Certains mots ont plus de 500 000 occurrences dans le corpus ; c'est autant de phrases dont il faut extraire le plongement lexical avec FlauBERT. Pour alléger ce processus, nous échantillonnons 5000 phrases pour chaque mot, à partir desquelles nous mesurons la variation pour chaque dimension (par année, entreprise, domaine d'activité et secteur) en se ramenant à des plongements lexicaux statiques. Pour chaque dimension, nous conservons comme mots-cibles les 10% de mots ayant la mesure de variation la plus élevée.

Puis pour chaque mot-cible retenu, nous appliquons les algorithmes de partitionnements K-means et propagation par affinité sur l'ensemble des représentations vectorielles de ses occurrences. Afin d'évaluer la qualité du partitionnement, nous calculons le coefficient de silhouette pour chaque mot et algorithme. Puis, nous nous extrayons les distributions de probabilité pour chaque strate temporelle (analyse diachronique) et pour chaque entreprise / secteur d'activité (analyse synchronique). Enfin, pour chaque dimension, nous calculons la JSD généralisée sur l'ensemble des distributions de probabilité afin de mesure le niveau de variation sémantique du mot dans la dimension étudiée. Les valeurs moyenne pour tous les mots-cibles du coefficient de silhouette et de la JSD par secteur et par année figurent en Table 1. Rappelons que le coefficient de silhouette se situe entre 0 et 1 (proche de

1. https ://github.com/CoFiF/Corpus
2. En utilisant l'outil Wolf (Sagot & Fišer, 2008), une ressource lexicale et sémantique pour le français.

Method	S-score	JSD-secteur	JSD-année
aff-prop	0.118	1.722	1.265
kmeans3	0.094	0.137	0.075
kmeans5	0.088	0.234	0.124
kmeans7	0.071	0.230	0.167

TABLE 1: Valeurs moyenne pour tous les mots-cibles du coefficient de silhouette et de la JSD par secteur et par année

	Année	Secteur
1	écologie	magasin
2	climat	écologie
3	biodiversité	luxe
4	syndicats	syndicats
5	gouvernement	publicité

TABLE 2: Top 5 des mots ayant les plus fortes JSD selon les dimensions temporelle et sectorielle, avec la propagation par affinité.

zéro indique une faible qualité) ; et que si la JSD se situe entre 0 et 1 pour deux distribution, la version généralisée à n distributions est bornée par $\log_2(n)$. Pour la dimension temporelle par exemple, notre période de 18 ans mène à une borne supérieure valant $\log_2(18) \approx 4.17$.

Selon la table 1, l'algorithme de propagation par affinité mène au coefficient de silhouette moyen le plus élevé. Nous utilisons donc cet algorithme de partitionnement pour comparer les variations des mots-cibles entre eux. Les 5 mots-cibles ayant les plus fortes JSD selon les dimensions temporelle et sectorielle dans le corpus sont listés dans la Table 2. On note que les 3 mots les plus variables par année font tous partie du champs lexical du climat. De plus, le mot *écologie* varie fortement dans les deux dimensions ; nous allons analyser et interpréter ses variations.

Étude de cas : le mot *écologie*.

Pour ce mot, le coefficient de silhouette le plus élevé est obtenu à partir de l'algorithme K-Means avec $k = 7$, pour un score de 0.231. La distribution normalisée des clusters issus de ce partitionnement pour les dimensions sectorielle et temporelle est représentée sur la Figure 1. En comparant les distributions de chaque période ou secteur avec la distribution moyenne sur le corpus pour la dimension associée, nous repérons les périodes et secteurs qui se démarquent. Puis nous quantifions la variation de chaque cluster au sein d'une dimension. Cela nous mène à la dernière étape : l'interprétation des clusters. Nous extrayons les phrases les plus proches des centroïdes ; puis à partir de la méthode d'extraction de mots-clés, nous associons un thème à chaque cluster (Table 3).

Par exemple, le cluster 6 connaît une forte variation temporelle, avec une proportion croissante à partir de 2007 ; il est associé à des problématiques de financement et de coût (Figure 1). Le cluster 2 n'est propre qu'à quelques secteurs et se concentre sur les idées de métier propre à l'écologie ; il apparaît assez tardivement dans les documents. À l'inverse, les clusters 1 et 5 propres aux transports et à l'aménagement des territoires pour l'un, et à l'énergie pour l'autre, sont communs à la plupart des secteurs. Le cluster 1 est bien résumé par sa phrase-centroïde, "*ces obligations concernent pour l'essentiel l'écologie, l'aménagement du paysage et l'archéologie pour les sites de développement associés*". Le cluster 3, plutôt associé au territoire du point de vue des ressources, est présent sur toute la période étudiée mais n'est propre qu'à quelques secteurs tels ceux du pétrole et de la chimie. Pour finir, le cluster 4 qui contient le champs lexical du risque et du danger, apparaît et prend progressivement de l'ampleur mais reste minoritaire, probablement dû au fait que les analystes financiers évitent d'employer des termes négatifs lors de la rédactions des rapports d'activité afin de ne pas inquiéter les investisseurs.

N^o	Titre	Exemple de mots-clés
0	pratique	éco, concept , logement, économique, raisonné, préservant, préfabrication
1	transport	directeur, énergie, impacts, transports, aviation, initiatives, territoire, aménagement
2	métier	apprendre, structure, métiers, collaborateurs, réseau, professionnels, management
3	territoire	industrielle, sites, flux, déchets, échanges, territoriale, eaux, circulaire, ressources
4	danger	groupe, fondation, prix, intégrer, péril, polluante, excessive, concernés
5	énergie	émissions, énergie, fessenheim, industrielle, biodiversité, slovénie, co2, nucléaire
6	coût	énergie, arrêté, coût, mer, prix, stockage, économiques, milliards, aménagement

TABLE 3: Liste des clusters et interprétations pour le mot *écologie*.

Dans l'ensemble, les disparités d'usage et de connotation des mots que nous détectons sont encourageantes. La détection de variations propres à une période temporelle permettrait à un analyste de relier le résultat avec des événements de la vie réelle, tandis que les variations de connotation entre les clusters ouvrent la voie à une analyse de sentiments plus poussée.

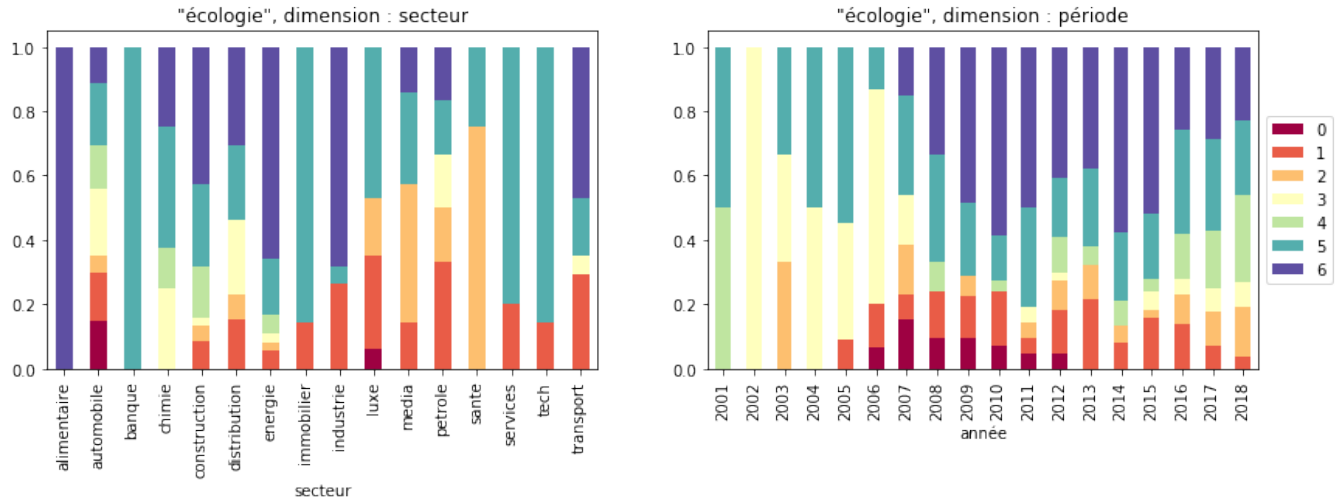


FIGURE 1: Distribution des clusters pour le mot *écologie*, par secteur d'activité (à gauche) et par année (à droite).

5 Conclusion

Cet article est une investigation préliminaire de la capacité des plongements lexicaux contextualisés de BERT à détecter des variations diachroniques et synchroniques d'usages de mots. Nous montrons sur une étude de cas que notre méthode permet de détecter et d'interpréter de façon fine les variations d'usage dans plusieurs dimensions.

L'étape suivante est de proposer une méthode d'évaluation de notre processus. Peu de corpus annotés de variations sémantiques étant disponibles, nous nous tournons vers la génération de variations synthétiques (Shoemark *et al.*, 2019). Nous définissons différents scénarios de variations d'usage d'un mot selon différentes dimensions, puis les simulons en générant des corpus comportant ces variations. Le but est d'évaluer les capacités de détection de notre méthode sur ces données synthétiques, pour chaque scénario.

Références

AITCHISON J. (2001). Language change : Progress or decay ? In *Cambridge Approaches to Linguistics*. Cambridge University Press.

ALAGIĆ D., ŠNAJDER J. & PADÓ S. (2018). Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence* : link.

BUECHEL S., JUNKER S., SCHLAAK T., MICHELSEN C. & HAHN U. (2019). A time series analysis of emotional loading in central bank statements. In *Proceedings of the Second EcoNLP Workshop*, Hong Kong : D19-5103.

COENEN A., REIF E., YUAN A., KIM B., PEARCE A., VIÉGAS F. B. & WATTENBERG M. (2019). Visualizing and measuring the geometry of bert. In *NeurIPS* : NIPS2019_9065.

DAUDERT T. & AHMADI S. (2019). CoFiF : A corpus of financial reports in French language. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, p. 21–26, Macao, China : W19-5504.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT* : N19-1423.

FREY B. J. & DUECK D. (2007). Clustering by passing messages between data points. *Science*, **315**(5814), 972–976. DOI : 10.1126/science.1136800.

GIULIANELLI M., FERNANDEZ R. & DEL TREDICI M. (2019). Contextualised word representations for lexical semantic change analysis. In *EurNLP* : link.

HAMILTON W., LESKOVEC J. & JURAFSKY D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. p. 1489–1501. DOI : 10.18653/v1/P16-1141.

HU R., LI S. & LIANG S. (2019). Diachronic sense modeling with deep contextualized word embeddings : An ecological view. p. 3899–3908. DOI : 10.18653/v1/P19-1379.

KIM Y., CHIU Y.-I., HANAKI K., HEGDE D. & PETROV S. (2014). Temporal analysis of language through neural language models. DOI : 10.3115/v1/W14-2517.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of LREC 2020* : arXiv : 1912.05372.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *arXiv : 1907.11692*.

MARTINC M., MONTARIOL S., ZOSA E. & PIVOVAROVA L. (2020a). Capturing evolution in word usage : Just add more clusters ? In *Companion Proceedings of the Web Conference 2020*, WWW '20, p. 343–349, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3366424.3382186.

MARTINC M., NOVAK P. K. & POLLAK S. (2020b). Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of LREC 2020* : arXiv : 1912.01072.

MATTHEW PURVER, ALJOŠA VALENTINČIČ M. P. & POLLAK S. (2018). Diachronic lexical changes in company reports : An initial investigation. In M. EL-HAJ, P. RAYSON & A. MOORE, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France : lrec2018-9_W27.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, p. 3111–3119. NIPS2013_5021.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. *arXiv : 1802.05365*.

RÉ M. A. & AZAD R. K. (2014). Generalization of entropy based divergence measures for symbolic sequence analysis. *PLoS ONE*, **9**. DOI : 10.1371/journal.pone.0093532.

SAGI E., KAUFMANN S. & CLARK B. (2009). Semantic density analysis : Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, Athens, Greece : W09-0214.

SAGOT B. & FIŠER D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco. HAL : inria-00614708.

SCHLECHTWEG D., HÄTTY A., DEL TREDICI M. & SCHULTE IM WALDE S. (2019). A wind of change : Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 732–746, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1072.

SHOEMARK P., LIZA F. F., NGUYEN D., HALE S. & MCGILLIVRAY B. (2019). Room to Glo : A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 EMNLP-IJCNLP Conference*, p. 66–76, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1007.

TAHMASEBI N., BORIN L. & JATOWT A. (2018). Survey of computational approaches to diachronic conceptual change. *arXiv : 1811.06278*.

TREDICI M. D. & FERNÁNDEZ R. (2017). Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers* : W17-6804.

WIEDEMANN G., REMUS S., CHAWLA A. & BIEMANN C. (2019). Does bert make any sense ? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of KONVENS 2019*, Erlangen, Germany : link.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Identification des problèmes d'annotation pour l'extraction de relations

Tsanta Randriatsitohaina[1] Thierry Hamon[1,2]
(1) LIMSI, CNRS, Université Paris-Saclay, Campus universitaire d'Orsay, 91405 Orsay cedex, France
(2) Université Paris 13, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
tsanta@limsi.fr, hamon@limsi.fr

RÉSUMÉ

L'annotation d'un corpus est une tâche difficile et laborieuse, notamment sur des textes de spécialité comme les textes biomédicaux. Ainsi, dans un contexte comme l'extraction des interactions aliment-médicament (FDI), l'annotation du corpus POMELO a été réalisée par un seul annotateur et présente des risques d'erreur. Dans cet article, nous proposons d'identifier ces problèmes d'annotation en utilisant un corpus Silver Standard (CSS) que nous établissons à partir d'un vote majoritaire parmi les annotations proposées par des modèles entraînés sur un domaine similaire (interaction médicament-médicament – DDI) et l'annotation manuelle à évaluer. Les résultats obtenus montrent que l'annotation dans POMELO est considérablement éloignée du CSS. L'analyse des erreurs permet d'en identifier les principales causes et de proposer des solutions pour corriger l'annotation existante.

ABSTRACT

Identification of annotation problem for the relation extraction.

Annotating a corpus is a difficult and time-consuming task, especially on texts of specific field such as biomedical texts. Thus, in the context of the extraction of food-drug interactions (FDI), the POMELO corpus annotation has been performed by a single annotator and presents risks of errors. In this article, we propose to identify these annotation problems by referring to a Silver Standard (CSS) corpus that we establish from a majority vote among the annotations proposed by models trained on a similar domain (drug-drug interaction) and the manual annotation to be evaluated. The obtained results show that the annotation in POMELO is far from the CSS. Error analysis helps to identify the main causes and to suggest solutions to correct the existing annotation.

MOTS-CLÉS : Annotation, extraction de relation, corpus biomédical, interaction aliment-médicament.

KEYWORDS: Annotation, relation extraction, biomedical corpus, food-drug interaction.

1 Introduction

Il peut être difficile dans de nombreuses circonstances d'établir un corpus selon les pratiques habituelles, c'est-à-dire une double annotation avec un consensus. C'est notamment le cas dans les domaines spécialisés où mobiliser plusieurs experts pour réaliser des annotations peut être compliqué. En pratique, les corpus annotés sont développés lors des travaux préliminaires, comme preuve de concept ou pour définir le guide d'annotation. Étant donné qu'aucun consensus n'a été effectué, on peut s'attendre à ce que de tels corpus comportent des problèmes d'incohérence ou d'ambiguïté.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 323–331.
HAL : hal-02784782.

Dans ce contexte, il est possible d'envisager d'améliorer l'annotation effectuée par un seul annotateur pour réduire le coût de l'annotation. Pour atteindre cet objectif, nous considérons la tâche d'extraction de relations spécifiques au domaine, en particulier l'extraction de l'interaction entre aliments et médicaments (*Food-drug interaction* – FDI) pour laquelle il existe très peu de corpus annotés ou de modèles d'extraction automatique.

À l'instar des interactions entre médicaments (*Drug-drug interaction* – DDI), les FDI correspondent à l'absorption, distribution ou l'effet inattendu d'un médicament, provoqués par l'ingestion d'un aliment (Doogue & Polasek, 2013). Par exemple, le pamplemousse est connu pour avoir un effet inhibiteur sur une enzyme impliquée dans le métabolisme de plusieurs médicaments (Hanley *et al.*, 2011).

Toutefois, à la différence des DDI (Aagaard & Hansen, 2013) ou les effets indésirables des médicaments (Aronson & Ferner, 2005) qui sont répertoriés dans des bases de données telles que DrugBank [1] ou Thériaque [2], les FDI figurent rarement dans les bases de connaissances et sont généralement dispersées dans des sources hétérogènes (Wishart *et al.*, 2017). En outre, dans DrugBank par exemple, les informations sont principalement stockées sous forme de phrases.

Des travaux précédents ont été réalisés pour l'extraction automatique de ces informations (Hamon *et al.*, 2017) et ont conduit à la définition d'un ensemble de 639 résumés Medline (corpus POMELO) annotés par un externe en pharmacie selon 21 types de relation avec un grand déséquilibre en termes de nombre d'exemples par type. Les tentatives d'entraînement d'un système de détection automatique des interactions entre aliments et médicaments sur ce corpus ont conduit à de faibles performances. ((Randriatsitohaina & Hamon, 2019)) Nous émettons l'hypothèse que l'annotation du corpus, étant faite par une seule personne, peut présenter des incohérences. Afin de les identifier, nous proposons de tirer parti du corpus existant sur les DDI qui est un domaine similaire afin d'évaluer l'annotation POMELO selon un schéma d'annotation de plus haut niveau (4 types de relations DDI). Nos contributions se concentrent sur la définition d'un corpus Silver Standard (CSS) à partir d'un consensus avec les annotations des modèles d'extraction des DDI, l'analyse de l'annotation POMELO par rapport au CSS, l'identification des incohérences et la proposition de solutions pour corriger l'annotation existante.

2 État de l'art

Le développement d'un corpus Silver Standard (CSS) a été introduit pour réduire les coûts d'annotation de corpus. L'idée est d'appliquer des systèmes d'annotation sur un corpus et d'harmoniser l'annotation finale en utilisant un système de vote. Rebholz-Schuhmann *et al.* (2010) a produit un corpus pour l'identification des entités nommées à partir de textes biomédicaux. Lu *et al.* (2011) utilisent un algorithme espérance-maximisation (Expectation Maximisation) pour déduire la vérité terrain, basé uniquement sur les soumissions des équipes participantes à un défi sur le normalisation de gènes. Suivant le même principe, nous établissons un CSS obtenu par un système de vote sur les annotations des modèles d'extraction des DDI pour identifier ensuite les problèmes d'annotation des relations.

Certaines approches utilisent l'apprentissage actif afin d'améliorer l'annotation d'un corpus bruyant

1. https://www.drugbank.ca/
2. http://www.theriaque.org

((Rehbein & Ruppenhofer, 2017)). Le processus est basé sur le choix d'une instance mal étiquetée par détection d'erreur, en donnant les instances sélectionnées à l'oracle (l'annotateur humain), pour être manuellement désambiguïsées et ajoutées aux données d'apprentissage.

Une approche de détection des erreurs d'annotation est proposée par Ménard & Mougeot (2019), basée sur le regroupement des instances correspondant à chaque valeur d'étiquette "silver" unique, puis sur la considération des valeurs aberrantes comme des étiquettes éventuellement bruyantes qui doivent être ré-annotées par l'expert. Dans notre cas, nous utilisons le résultat de modèles entraînés sur les DDI pour identifier les éventuelles erreurs d'annotation.

Différents types d'approches ont été explorés pour extraire les DDI. Parmi elles, Kolchinsky *et al.* (2015) se concentrent sur l'identification des phrases pertinentes et des résumés pour l'extraction des preuves pharmacocinétiques. Étant donné que nous nous concentrons sur la correction d'annotations, nous proposons de travailler sur les phrases du corpus POMELO qui ont été annotées comme pertinentes. Ben Abacha *et al.* (2015) proposent une approche basée sur un SVM combinant : (i) les caractéristiques décrivant les mots dans le contexte des relations à extraire, (ii) les noyaux composites utilisant des arbres de dépendance. Leur approche permet d'obtenir des F1-mesures de 0,53 et 0,40 sur des résumés Medline et 0,83 et 0,68 sur des documents de DrugBank. Dans notre tâche, nous utilisons un SVM linéaire combiné avec des descripteurs sémantiques et syntaxiques qui nous permettent d'obtenir des résultats comparables. Kim *et al.* (2015) ont construit deux classifieurs pour l'extraction DDI : un classifieur binaire pour extraire les paires de médicaments en interaction et un classifieur de types DDI pour identifier les catégories de l'interaction. Cejuela *et al.* (2018) considèrent l'extraction de la relation de localisation des protéines comme une classification binaire. Liu *et al.* (2016) proposent une méthode basée sur un CNN pour l'extraction des DDI. D'autres travaux utilisent un modèle de réseau neuronal récurrent avec plusieurs couches d'attention pour la classification DDI (Yi *et al.*, 2017; Zheng *et al.*, 2017), ou utilisant des récurrences au niveau des mots et des caractères (Kavuluru *et al.*, 2017) produisant une performance de 0,72. Sun *et al.* (2019) proposent une méthode hybride combinant un réseau de neurones récurrent et convolutif induisant une amélioration de 3%. Le réseau convolutif profond de (Dewi *et al.*, 2017) permet de couvrir de longues phrases qui ont des jeux de données de DDI typiques et d'obtenir une performance de 0,86. BERT (Devlin *et al.*, 2019), un modèle de représentation de mots contextualisés permet obtenir les meilleurs résultats sur les tâches de TAL y compris l'extraction de relations. Lee *et al.* (2019) ont entraîné et adapté BERT sur des corpus biomédicaux pour être efficaces dans les tâches d'exploration de textes biomédicaux. Nous avons adapté BioBert pour l'extraction des DDI.

3 Corpus

Notre approche est mise en œuvre sur le corpus POMELO qui contient des annotations de FDI mais dont la qualité doit être améliorée, et le corpus DDI fournissant des annotations d'interactions entre médicaments et dont la qualité est avérée suite à son utilisation dans des campagnes d'évaluation.

Corpus POMELO Des études ont déjà été menées sur les FDI durant lesquelles l'ensemble de données POMELO a été développé (Hamon *et al.*, 2017). Ce corpus est constitué de 639 résumés d'articles scientifiques du domaine médical (269 824 mots, 5 752 phrases) en anglais, collectés à partir

du portail PubMed [3]. Les 639 résumés sont annotés selon 9 types d'entités et 21 types de relations avec Brat (Stenetorp *et al.*, 2012) par un étudiant en pharmacie. Les annotations se concentrent sur des informations sur la relation entre aliments, médicaments et pathologies. Étant donné que nous examinons les interactions aliment-médicament, nous avons construit notre ensemble de données en tenant compte de tous les couples de *drug* et *food* ou *food-supplement* à partir des données POMELO. L'ensemble de données qui en résulte est composé de 900 phrases étiquetées avec 13 types de relations (avec le nombre d'instances) : *decrease absorption* (53), *slow absorption* (15), *slow elimination* (15), *increase absorption* (39), *speed up absorption* (1), *new side effect* (4), *negative effect on drug* (88), *worsen drug effect* (8), *positive effect on drug* (21), *improve drug effect* (6), *no effect on drug* (109), *without food* (13), *unspecified relation* (528).

Corpus DDI Le corpus DDI [4] est composé de 4 037 instances de relations issues de la base de données DrugBank et des résumés Medline. Il a été utilisé lors de SemEval 2013 (Segura-Bedmar *et al.*, 2013) et propose quatre types d'annotation : *Conseil* (819 instances) pour une recommandation concernant l'utilisation concomitante de deux médicaments, *Effet* (1 700 instances) pour l'effet de la DDI, *Mécanisme* (1322 instances) pour la pharmacodynamique et la pharmacocinétique du médicament, *Interaction* (188 instances) si aucune information sur l'interaction n'est fournie.

4 Méthode

Afin d'évaluer la qualité de l'annotation du corpus POMELO, nous proposons une approche en trois étapes décrite à la figure 1 .

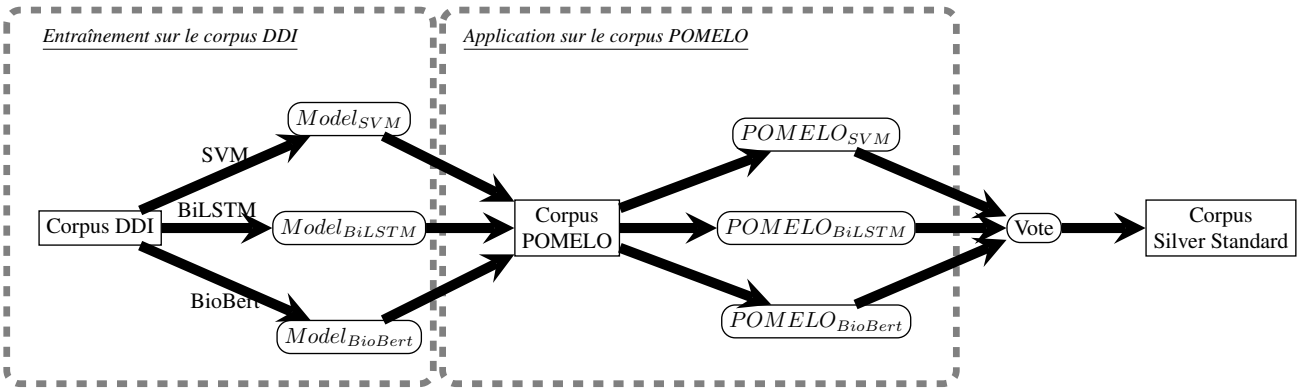


FIGURE 1 – Architecture de l'approche.

Étape 1 : Apprentissage sur le corpus DDI Nous proposons de tirer parti des méthodes d'extraction de DDI pour obtenir des annotations sur le corpus POMELO. Pour ce faire, nous avons entraîné 3 types de modèles sur le corpus DDI :

— **SVM.** SVM (Vapnik, 1995) ayant montré de bonnes performances dans l'état de l'art, nous proposons d'entraîner un modèle SVM linéaire avec les TF/IDF des formes fléchies, lemmes,

3. https://www.ncbi.nlm.nih.gov/pubmed/
4. https://github.com/dbmi-pitt/public-PDDI-analysis/blob/master/PDDI-Datasets/DDI-Corpus2013

catégories morpho-syntaxiques des mots, catégories sémantiques des termes (aliment, médicament, supplément alimentaire) des contextes avant, entre et après les deux arguments de la relation.

— **BiLSTM.** Nous entraînons un modèle récurrent BiLSTM (Gers *et al.*, 2000) avec les plongements de mots Skipgram fastText (Bojanowski *et al.*, 2017) de taille 300 et les plongements de position.

— **BioBert.** Nous utilisons BioBert version 1.1 Lee *et al.* (2019) entraîné avec un vocabulaire de 4,5 milliards de mots issus de PubMed.

Étape 2 : Application des modèles sur le corpus POMELO Une fois les modèles entraînés sur les données DDI, nous les appliquons sur le corpus POMELO. Cette étape permet d'obtenir des annotations dites artificielles des phrases pertinentes de POMELO selon les types de relations DDI. À ce niveau, il nous faut une correspondance des types de relations DDI et FDI pour pouvoir évaluer les annotations manuelles faites sur POMELO. Pour ce faire, nous avons établi une correspondance entre les types de DDI et les types de FDI, validée par les experts, comme suit : les interactions impliquant la pharmacocinétique des médicaments (absorption, élimination, métabolisme) sont étiquetées *Mécanisme* (123 instances), les interactions impliquant les effets des médicaments (positifs, négatifs, effet secondaire) sont étiquetées *Effet* (236 instances), le type de relation *Without food* qui est une contre-indication à la prise d'aliment avec un médicament est étiqueté *Conseil* (13 instances) et les interactions *Unprecised relation* sont étiquetées *Interaction* (528 instances). Nous appellerons cette correspondance POMELO-DDI.

Étape 3 : Vote majoritaire parmi les annotations obtenues Dans cette section, nous proposons d'établir une annotation de corpus Silver Standard (CSS) en appliquant un système de vote majoritaire sur les annotations obtenues par les modèles SVM, BiLSTM et BioBert ainsi que l'annotation POMELO-DDI décrite à la section 4. Nous prenons en compte l'annotation POMELO-DDI étant donnée qu'elle a été faite par un expert donc peut apporter des informations pertinentes sur certains types de relation. Ainsi, pour chaque instance de relation, nous gardons le type de relation le plus fréquent parmi les quatre types d'annotations. En cas d'absence d'accord, nous privilégions l'annotation POMELO-DDI qui est basée sur l'annotation humaine.

Étape 4 : Analyse des erreurs d'annotation dans POMELO-DDI Nous utilisons l'annotation Silver Standard obtenue à la section 4 comme pré-annotation et nous identifions les problèmes d'annotation des relations où l'annotation manuelle et le Silver Standard diffèrent.

5 Expériences et résultats

Pour l'étape d'apprentissage sur les données DDI, nous évaluons les modèles selon la macro-moyenne des F1-mesures obtenues par validation croisée en 3 échantillons. De manière générale, les résultats obtenus par les modèles sont prometteurs, avec des F1-mesures allant de 0,78 avec SVM à 0,91 avec BioBert. C'est d'autant plus important que ces performances décriront la fiabilité des annotations que les modèles fourniront. Les performances du modèle BioBert surpassent de loin celle des autres modèles. Cela confirme l'efficacité des plongements de mot contextualisé appris sur des documents du domaine spécifique.

Une fois les modèles entraînés, nous les appliquons sur le corpus POMELO en gardant les mêmes configurations (paramètres et descripteurs). Cela nous permet d'obtenir 4 annotations de type DDI sur les données POMELO selon les modèles SVM, BiLSTM et BioBert ainsi que l'annotation POMELO-DDI décrite à la section 4. Le vote majoritaire appliqué sur ces 4 annotations nous permet d'avoir une sorte de consensus (corpus Silver Standard) que nous utilisons comme annotation de référence pour évaluer la performance des modèles obtenu à l'étape 1 sur l'annotation du corpus POMELO selon les types DDI. Les résultats de cette évaluation indiquent que BioBert fournit l'annotation la plus en accord avec le consensus (kappa de Cohen 0,86) tandis que les annotations manuelles POMELO-DDI en semblent être les plus éloignées (kappa de Cohen 0,15).

Nous proposons alors d'effectuer une analyse des erreurs afin d'identifier la source de ces différences d'annotation. Une des erreurs les plus fréquentes est l'annotation des titres ou des descriptifs des expériences avec le type *interaction* alors que la phrase contient des vocabulaires spécifiques au mécanisme ou aux effets de médicaments, d'où la faible précision (cf. tableau 1. Par exemple l'instance *"Nous décrivons ici deux études pharmacocinétiques pour quantifier l'impact des **aliments** sur l'absorption de la **ziprasidone** chez des volontaires sains."* a été annotée manuellement *interaction* et *mécanisme* par le vote. Étant donné que la conclusion de l'étude qui devrait fournir plus d'information sur le médicament et l'aliment considérés est donnée un peu plus tard dans le résumé, il est possible de ne pas annoter les titres ou les descriptifs réduisant ainsi l'ambiguïté qui peut avoir lieu. Nous avons également observé que le type *interaction* est utilisé comme type par défaut et n'est pas précisé dans certains cas alors qu'il est préférable de privilégier des types plus précis, ce que font les modèles. Par exemple, l'instance *"Bloc cardiaque auriculo-ventriculaire complet développé en raison de l'utilisation du **vérapamil** et de la consommation de **miel**"* a été annoté manuellement *interaction* et *effet* par le vote. Dans certains cas, l'annotateur utilise le type *interaction* faute de type plus précis plus adapté. Par exemple l'instance *"En ce qui concerne les autres facteurs alimentaires pouvant interagir avec les **anticoagulants oraux**, le patient doit être mis en garde concernant les suppléments de **vitamines A, E et C et d'alcool** utilisés de façon chronique ou ingérés en grande quantité."* a été annoté manuellement *interaction* et *conseil* par le vote. Ce type de relation pourrait donc être intégré dans le guide d'annotation, ce qui permettrait d'améliorer le rappel (cf. tableau 1).

Nous avons également observé que certaines instances impliquent plusieurs types de relation simultanément, ce qui provoque une divergence entre les différentes annotations. Par exemple l'instance *"Ces résultats démontrent que l'administration de **ziprasidone** avec des **aliments** est cruciale pour assurer une biodisponibilité optimale, fiable et dépendante de la dose et donc un contrôle et une tolérabilité prévisibles des symptômes."* mentionne à la fois le mécanisme du médicament (biodisponibilité) ainsi que son effet (contrôle, tolérabilité). Les 2 types de relation ont généralement un lien de cause à effet, donc plusieurs solutions sont possibles : (1) privilégier une relation par rapport aux autres, (2) annoter les différents types en même temps (multi-étiquette), (3) scinder l'instance en plusieurs instances de relation différentes.

6 Conclusion

Dans cet article, nous proposons d'identifier les problèmes durant l'annotation des relations dans des textes biomédicaux, en particulier les interactions médicament-aliment (FDI) rassemblées dans le corpus POMELO. Pour ce faire, nous proposons de tirer parti des données d'un domaine similaire (interactions médicament-médicament – DDI) afin d'établir un corpus Silver Standard CSS) obtenu par

Relation	P	R	F1	Nb instance
Conseil	0,54	0,11	0,19	61
Effet	0,64	0,57	0,60	264
Interaction	0,12	0,98	0,22	65
Mécanisme	0,95	0,23	0,37	510
moy / total	0,56	0,48	0,52	900

TABLE 1 – Macro précision, rappel, F1-mesure entre le corpus POMELO-DDI et le corpus Silver Standard.

un vote majoritaire parmi les annotations des modèles d'extraction automatique ainsi que l'annotation à évaluer. Entraînés sur le corpus DDI, les performances des modèles allant de 0,78 jusqu'à 0,91 de F1-mesure obtenues avec des modèles SVM, BiLSTM et BioBert sont prometteuses et indicatrices de la fiabilité des annotations que ces modèles fourniront. Une fois les modèles entraînés, nous les appliquons sur les instances POMELO puis nous appliquons un système de vote majoritaire pour garder le type de relation affecté le plus souvent à chaque instance constituant ainsi le CSS. Le CSS servira de référence afin d'identifier les sources de divergences avec les annotations manuelles faites sur le corpus POMELO. Les résultats obtenus ont montré que l'annotation dans le corpus POMELO est considérablement éloignée du corpus Silver Standard. L'analyse des erreurs a permis d'en identifier les principales causes et de proposer des solutions à savoir limiter les annotations aux phrases informatives et fiables de FDI, privilégier les types de relation précis au type générique, ajouter d'autres types de relation comme *conseil*, dans le cas de l'implication de plusieurs types de relation dans une instance, privilégier une relation par rapport à l'autre ou annoter les types en même temps (multi-étiquette) ou scinder l'instance en plusieurs instances de relation différentes. Dans la suite du travail, nous proposons d'appliquer ces solutions dans une phase de correction d'annotation en utilisant le Silver Standard comme pré-annotation et entraîner des modèles d'extraction automatique sur l'annotation obtenue.

Remerciements

Ce travail est financé par l'ANR dans le cadre du projet MIAM (ANR-16-CE23-0012).

Références

AAGAARD L. & HANSEN E. (2013). Adverse drug reactions reported by consumers for nervous system medications in europe 2007 to 2011. *BMC Pharmacology & Toxicology*, **14**, 30.

ARONSON J. & FERNER R. (2005). Clarification of terminology in drug safety. *Drug Safety*, **28**(10), 851–70.

BEN ABACHA A., CHOWDHURY M. F. M., KARANASIOU A., MRABET Y., LAVELLI A. & ZWEIGENBAUM P. (2015). Text mining for pharmacovigilance : Using machine learning for drug name recognition and drug-drug interaction extraction and classification. *Journal of Biomedical Informatics*, **58**, 122–132.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

CEJUELA J. M., VINCHURKAR S., GOLDBERG T., PRABHU SHANKAR M. S., BAGHUDANA A., BOJCHEVSKI A., UHLIG C., OFNER A., RAHARJA-LIU P., JENSEN L. J. & ROST B. (2018). LocText : relation extraction of protein localizations to assist database curation. *BMC Bioinformatics*, **19**(1), 15. DOI : 10.1186/s12859-018-2021-9.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DEWI I. N., DONG S. & HU J. (2017). Drug-drug interaction relation extraction with deep convolutional neural networks. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 1795–1802.

DOOGUE M. & POLASEK T. (2013). The abcd of clinical pharmacokinetics. *Therapeutic Advances in Drug Safety*, **4**(1), 5–7. PMID 25083246, DOI : 10.1177/2042098612469335.

GERS F. A., SCHMIDHUBER J. & CUMMINS F. A. (2000). Learning to forget : Continual prediction with lstm. *Neural Computation*, **12**(10), 2451–2471.

HAMON T., TABANOU V., MOUGIN F., GRABAR N. & THIESSARD F. (2017). Pomelo : Medline corpus with manually annotated food-drug interactions. In *Proceedings of Biomedical NLP Workshop associated with RANLP 2017*, p. 73–80, Varna, Bulgaria.

HANLEY M. J., CANCALON P., WIDMER W. W. & GREENBLATT D. J. (2011). The effect of grapefruit juice on drug disposition. *Expert Opinion on Drug Metabolism & Toxicology*, **7**(3), 267–286. PMID 21254874, DOI : 10.1517/17425255.2011.553189.

KAVULURU R., RIOS A. & TRAN T. (2017). Extracting drug-drug interactions with word and character-level recurrent neural networks. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, p. 5–12 : IEEE.

KIM S., LIU H., YEGANOVA L. & WILBUR W. J. (2015). Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics*, **55**, 23–30.

KOLCHINSKY A., LOURENÇO A., WU H.-Y., LI L. & ROCHA L. M. (2015). Extraction of pharmacokinetic evidence of drug–drug interactions from the literature. *PLOS ONE*, **10**(5), e0122199. DOI : 10.1371/journal.pone.0122199.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. DOI : 10.1093/bioinformatics/btz682.

LIU S., TANG B., CHEN Q. & WANG X. (2016). Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, **2016**. DOI : 10.1155/2016/6918381.

LU Z., KAO H.-Y., WEI C.-H., HUANG M., LIU J., KUO C.-J., HSU C.-N., TSAI R. T.-H., DAI H.-J., OKAZAKI N., CHO H.-C., GERNER M., SOLT I., AGARWAL S., LIU F., VISHNYAKOVA D., RUCH P., ROMACKER M., RINALDI F., BHATTACHARYA S., SRINIVASAN P., LIU H., TORII M., MATOS S., CAMPOS D., VERSPOOR K. M., LIVINGSTON K. M. & WILBUR W. J. (2011). The gene normalization task in BioCreative III. *BMC Bioinformatics*, **12**(S2). DOI : 10.1186/1471-2105-12-S8-S2.

MÉNARD P. A. & MOUGEOT A. (2019). Turning silver into gold : error-focused corpus reannotation with active learning. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, p. 758–767, Varna, Bulgaria. DOI : 10.26615/978-954-452-056-4_088.

RANDRIATSITOHAINA T. & HAMON T. (2019). Extracting food-drug interactions from scientific literature : relation clustering to address lack of data. In *International Conference on Intelligent Text Processing and Computational Linguistics*, La Rochelle, France. HAL : hal-02122766.

REBHOLZ-SCHUHMANN D., JIMENO YEPES A. J., VAN MULLIGEN E. M., KANG N., KORS J., MILWARD D., CORBETT P., BUYKO E., TOMANEK K., BEISSWANGER E. & HAHN U. (2010). The CALBC silver standard corpus for biomedical named entities — a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

REHBEIN I. & RUPPENHOFER J. (2017). Detecting annotation noise in automatically labelled data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1160–1170, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1107.

SEGURA-BEDMAR I., MARTÍNEZ P. & HERRERO ZAZO M. (2013). Semeval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 341–350 : Association for Computational Linguistics.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, p. 102–107, Stroudsburg, PA, USA : Association for Computational Linguistics.

SUN X., DONG K., MA L., SUTCLIFFE R. F. E., HE F., CHEN S.-S. & FENG J. (2019). Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, **21**, 37.

VAPNIK V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.

WISHART D., DJOUMBOU Y., GUO A. C., LO E., MARCU A., GRANT J., SAJED T., JOHNSON D., LI C., SAYEEDA Z., ASSEMPOUR N., IYNKKARAN I., LIU Y., MACIEJEWSKI A., GALE N., WILSON A., CHIN L., CUMMINGS R., LE D. & WILSON M. (2017). Drugbank 5.0 : A major update to the drugbank database for 2018. *Nucleic Acids Research*, **46**(D1). DOI : 10.1093/nar/gkx1037.

YI Z., LI S., YU J., TAN Y., WU Q., YUAN H. & WANG T. (2017). Drug-drug interaction extraction via recurrent neural network with multiple attention layers. In *International Conference on Advanced Data Mining and Applications*, p. 554–566 : Springer.

ZHENG W., LIN H., LUO L., ZHAO Z., LI Z., ZHANG Y., YANG Z. & WANG J. (2017). An attention-based effective neural model for drug-drug interactions extraction. *BMC bioinformatics*, **18**(1), 445. DOI : 10.1186/s12859-017-1855-x.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Simplification automatique de texte dans un contexte de faibles ressources

Sadaf Abdul Rauf[1], Anne-Laure Ligozat[1, 2], Francois Yvon[1],
Gabriel Illouz[1] and Thierry Hamon[1, 3]
(1) Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
(2) Université Paris-Saclay, CNRS, ENSIIE, LIMSI 91400, Orsay, France
(3) Université Sorbonne Paris Nord, 93430, Villetaneuse, France
{firstName.lastName}@limsi.fr

RÉSUMÉ

La simplification de textes a émergé comme un sous-domaine actif du traitement automatique des langues, du fait des problèmes pratiques et théoriques qu'elle permet d'aborder, ainsi que de ses nombreuses applications pratiques. Des corpus de simplification sont nécessaires pour entrainer des systèmes de simplification automatique ; ces ressources sont toutefois rares et n'existent que pour un petit nombre de langues. Nous montrons ici que dans un contexte où les ressources pour la simplification sont rares, il reste néanmoins possible de construire des systèmes de simplification, en ayant recours à des corpus synthétiques, par exemple obtenus par traduction automatique, et nous évaluons diverses manières de les constituer.

ABSTRACT

Automatic Text Simplification : Approaching the Problem in Low Resource Settings for French

Sentence simplification has emerged as an active area of research in recent years owing to its utility in natural language processing as well as human learning studies. Simplification corpora are required to build such automatic simplification systems. We show that where resources for simplification are scarce, it is still possible to build simplification systems. We show the effectiveness of translations as a synthetic corpus for the simplification task and present an analysis of the two metrics often used to evaluate simplification, i.e. BLEU and SARI, in terms of their ability to predict complexity level of a text.

MOTS-CLÉS : Simplification de textes, compression de texte, corpus synthétique, apprentissage par transfert cross-lingue.

KEYWORDS: Text Simplification, Sentence Compression, Synthetic Corpus, Cross-Lingual Transfert Learning.

1 Introduction

La lecture et la compréhension constituent une contribution majeure à la courbe d'apprentissage d'une langue pour un apprenant. La complexité d'une phrase empêche souvent la bonne compréhension et constitue un obstacle majeur dans la chaîne d'apprentissage. Ceci s'applique spécifiquement à certains groupes de personnes, par exemple les enfants (De Belder & Moens, 2010), les personnes ayant des difficultés d'apprentissage (Rello *et al.*, 2013a; Huenerfauth *et al.*, 2009; Fajardo *et al.*, 2013) ou des difficultés spécifiques comme des formes de dyslexie (Rello *et al.*, 2013b; McCarthy & Swierenga,

Volume 2 : Traitement Automatique des Langues Naturelles, pages 332–341.
HAL : hal-02784783.

2010) ou d'aphasie (Canning & Tait, 1999), ou encore des troubles du spectre autistique (Evans *et al.*, 2014; Barbu *et al.*, 2015). Disposer de versions simplifiées d'un texte peut grandement en faciliter la lisibilité et la compréhension pour ces populations, et pourrait aussi bénéficier à des apprenants d'une langue étrangère.

La simplification automatique de texte (SAT) construit une version plus simple de textes complexes afin qu'ils soient plus facilement compréhensibles et intelligibles. La simplification implique des transformations élaborées du texte d'origine telles que le fractionnement, la suppression et la paraphrase. Nous nous limitons ici à des transformations qui s'appliquent à des phrases isolées.

La plupart des approches de simplification automatique de texte considèrent le processus de simplification comme une tâche de traduction monolingue, l'algorithme de traduction apprenant la simplification en réécrivant à partir du corpus parallèle simple-complexe (Daelemans *et al.*, 2004; Zhu *et al.*, 2010; Zhang & Lapata, 2017; Nisioi *et al.*, 2017).

Pour qu'un algorithme d'apprentissage automatique puisse apprendre ces transformations, il faut lui fournir suffisamment d'exemples de simplifications appariant des couples de phrases simples et complexes. La qualité d'un système de simplification automatique de texte dépendra fortement de la qualité et de la quantité des corpus de simplification ainsi que la qualité de l'algorithme d'apprentissage.

Il existe de nombreuses ressources linguistiques pour le français, mais en ce qui concerne la tâche de simplification, il s'agit d'une langue peu dotée. Nous ne connaissons aucun corpus de grande taille disponible gratuitement. Nous avons collecté des corpus de diverses sources (Grabar & Cardon, 2018; Brouwers *et al.*, 2014), mais ils restent trop petits pour développer des systèmes de simplification basés sur des méthodes d'apprentissage ; nous les utilisons uniquement comme corpus de test.

Nous proposons d'utiliser des traductions automatiques pour composer un corpus parallèle synthétique. Les corpus synthétiques sont souvent utilisés en traduction automatique (Lambert *et al.*, 2011; Abdul Rauf *et al.*, 2016; Sennrich *et al.*, 2016; Burlot & Yvon, 2018) mais leur utilisation reste peu explorée pour la simplification automatique. (Aprosio *et al.*, 2019) exploitent des données synthétiques pour créer des systèmes de simplification, mais en s'appuyant sur des données de simplification de "gold standard" en italien qui leur permettent d'entraîner le système "complexificateur", qu'ils utilisent pour traduire les phrases simples en complexes. Dans notre étude, nous faisons l'hypothèse qu'aucune donnée n'est initialement disponible et nous étudions s'il est possible de construire un système de simplification raisonnable à partir uniquement de traductions automatiques de phrases complexes et simples. À cet égard, nous présentons la première tentative d'utilisation d'un corpus synthétique pour aborder le problème de la simplification automatique des phrases et montrons qu'il s'agit d'une approche viable pour réaliser des systèmes de simplification dans des contextes à faibles ressources.

2 État de l'art

La plupart des approches de simplification automatique de textes considèrent le problème comme une tâche de traduction monolingue, l'algorithme de traduction apprenant la réécriture de la simplification à partir du corpus parallèle complexe-simple (Daelemans *et al.*, 2004; Zhu *et al.*, 2010; Zhang & Lapata, 2017; Nisioi *et al.*, 2017). Considérant la simplification comme un problème d'apprentissage automatique, la traduction monolingue a été opérée en utilisant toutes les techniques de traduction automatique, y compris la traduction à base de syntaxe (Zhu *et al.*, 2010), la traduction à base

de segments (Wubben *et al.*, 2012), l'hybridation du modèle de simplification avec la traduction automatique (Narayan & Gardent, 2014) et plus récemment diverses architectures neuronales (Zhang & Lapata, 2017; Nisioi *et al.*, 2017; Zhao *et al.*, 2018; Korhonen *et al.*, 2019; Surya *et al.*, 2019). Cependant, aucune de ces études n'a abordé la simplification dans un scénario de ressources limitées. Une exception est (Aprosio *et al.*, 2019) qui sélectionne des phrases simples à partir de corpus monolingues et produit les phrases complexes correspondantes en utilisant un "complexificateur" entrainé sur le corpus de simplification italien gold standard.

La qualité du corpus d'entraînement est un facteur important et a été largement discutée, en particulier pour le corpus *Wikipedia simple* (Xu *et al.*, 2015; Scarton *et al.*, 2018). Newsela (Xu *et al.*, 2015) est devenu une des principales ressources en la matière, car il propose des simplifications manuelles à plusieurs niveaux, du niveau 0 comprenant le texte original au niveau 4 comprenant le texte le plus simplifié ; cette propriété reste cependant sous-exploitée et aucun des travaux de l'état de l'art ne donne d'analyse approfondie des niveaux de simplification les plus utiles. Des travaux antérieurs comme (Alva-Manchego *et al.*, 2017) et (Scarton *et al.*, 2018) ont pris en compte ces niveaux et les ont utilisés en exploitant les niveaux de simplification adjacents (par exemple 0-1, 1-2), mais n'ont pas étudié les différences entre les niveaux. (Zhang & Lapata, 2017) ont utilisé des niveaux non-adjacents (par exemple 0-2, 1-4), tandis que (Scarton & Specia, 2018) ont utilisé les différents niveaux de lisibilité pour construire des systèmes de simplification pour différents publics. Dans cet article, nous présentons des systèmes de simplification et des analyses avec toutes les paires de niveaux et étudions également l'effet combiné de différents niveaux sur la simplification.

Un autre objectif de notre travail est d'étudier si les résultats obtenus pour l'anglais peuvent être reproduits pour d'autres langues. Pour le français, des approches basées sur des règles existent (Brouwers *et al.*, 2014) mais aucun système basé sur l'apprentissage automatique ne peut être construit en raison du manque de corpus de simplification. Un troisième objectif est d'utiliser des données synthétiques, comme cela a été fait pour la traduction automatique dans (Lambert *et al.*, 2011; Abdul Rauf *et al.*, 2016; Sennrich *et al.*, 2016; Burlot & Yvon, 2018) et d'étudier les niveaux de simplification et l'impact des corpus synthétiques.

3 Corpus de simplification synthétique

Nous avons construit des corpus de simplification français synthétiques en traduisant deux corpus de simplification anglais avec l'outil Google translate (version de mai 2019) [1]. Pour obtenir une mesure de la qualité des traductions, nous avons traduit le corpus de test anglais-français WMT14 [2] dans les mêmes conditions et avons obtenu un score BLEU (Papineni *et al.*, 2002) de $35,6$ sur l'ensemble de test tokenisé, comparable au meilleur système de cette évaluation, mais significativement moins bon que l'état de l'art actuel de la traduction automatique anglais-français.

Le corpus anglais qui nous sert de point de départ est Newsela (Xu *et al.*, 2015), qui a été créé pour fournir du matériel de lecture destiné à l'enseignement pré-universitaire. Chaque article a été réécrit quatre fois pour des enfants de différents niveaux. Le niveau 0 correspond à l'article original, et le même article apparaît pour 4 niveaux, du niveau 1 au niveau 4, le niveau 4 étant le plus simple. Nous avons couplé le corpus dans toutes les configurations possibles de $complex : simple$, où $complex$ comprend tous les textes d'un niveau de complexité donné l et simple comprend les textes ayant un niveau de complexité inférieur à l. Les corpus ainsi construits associent le niveau 0 aux niveaux

1. https://translate.google.com
2. https://www.statmt.org/wmt14/translation-task.html

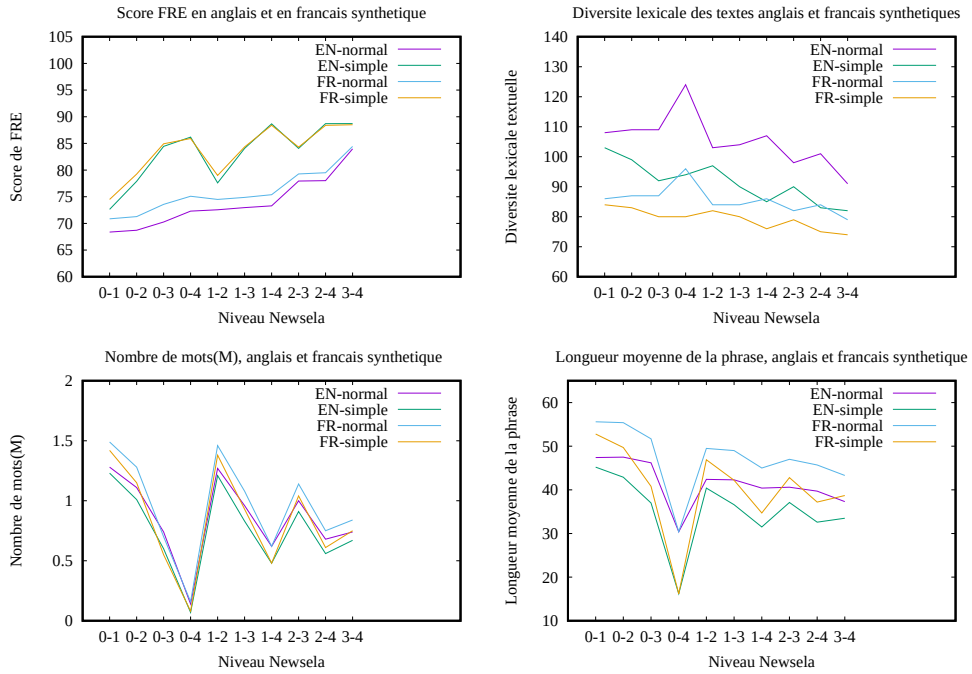


FIGURE 1 – Propriétés lexicales et de complexité des corpus originaux et traduits

$\{1, 2, 3, 4\}$, 1 à $\{2, 3, 4\}$, 2 à $\{3, 4\}$ et 3 à 4 comme indiqué dans le tableau 2.

La figure 1 présente l'analyse lexicale et la complexité des textes originaux et traduits. Nous constatons que les textes conservent leur niveau de lisibilité après traduction, comme le montrent les paramètres. Le comportement suivant est manifesté par les corpus de simplification anglais et français synthétique, où 0 est la phrase complexe et 4 le niveau le plus simple :

Niveau de complexité : Nous utilisons le score Flesch Reading Ease (Flesch, 1948) pour l'anglais et son adaptation en français proposée par (Kandel & Moles, 1958), implantés respectivement par les équations (1) et (2) ci-dessous. Ces scores varient entre 0 et 100. Une valeur plus élevée signifie que le texte est facile à lire et une valeur plus faible signifie que la difficulté est plus élevée.

$$FRE(Anglais) = 206.835 - 1.015(\frac{\frac{total}{words}}{\frac{total}{sentences}} - 84.6(\frac{\frac{total}{syllables}}{\frac{total}{words}}) \quad (1)$$

$$FRE(Francais) = 207 - 1.015(\frac{\frac{total}{words}}{\frac{total}{sentences}} - 73.6(\frac{\frac{total}{syllables}}{\frac{total}{words}}) \quad (2)$$

Comme on le voit sur la figure 1 en haut à gauche, après traduction, le corpus simple français synthétique se situe au même niveau de lisibilité que l'anglais simple original. Il en est de même pour les phrases complexes. On peut donc conclure qu'au moins en surface, les textes traduits conservent la classe de complexité des textes originaux.

Diversité lexicale textuelle : La diversité lexicale (DL) [3] est une mesure fréquemment utilisée pour évaluer la niveau de complexité d'un texte, un indice de DL est élevé indiquant qu'il est sera plus difficile à lire. Le score DL identifie clairement la classe de complexité du corpus comme le montre le sous-graphique en haut à droite (figure 1). On observe que du côté complexe, le score de diversité

3. Les mesures sont calculées avec https://pypi.org/project/lexicalrichness/

lexicale des textes augmente, alors que du côté simple, il diminue d'un niveau à l'autre, ainsi pour le côté complexe (1-2 = 103, 1-3 = 104, 1-4 = 107) et le côté simple (1-2 = 97, 1-3= 90, 1-4 = 85).

Longueur moyenne de la phrase et nombre de mots : Le nombre de mots diminue à mesure que le niveau de simplicité augmente. Ce comportement s'applique aussi bien aux mots complexes qu'à leur équivalent simple parmi toutes les paires de niveaux. La longueur moyenne des phrases est également beaucoup plus faible pour les textes simples. Ceci est illustré graphiquement dans les deux derniers sous-graphes.

On note que les phrases parallèles peuvent être très différentes dans chaque paire de niveaux, car elles sont sélectionnées par l'aligneur de phrases (Moore, 2002) en fonction du score de l'aligneur.

4 Évaluation

4.1 Protocole d'évaluation

Nous évaluons le système de simplification en utilisant les scores BLEU et SARI. BLEU (K. Papineni & Ward, 1998) est une métrique de traduction automatique qui s'appuie sur une comparaison de surface entre la sortie du système et la phrase de référence, en utilisant des correspondances de n-grammes et une pénalité de brièveté. Il donne une mesure de *l'adéquation*. SARI (Xu *et al.*, 2015) compare la phrase de sortie avec la phrase d'entrée ainsi qu'avec les phrases de référence. Il récompense les fragments qui sont validés par une des références, alors qu'ils n'apparaissaient pas dans l'entrée. BLEU est bien corrélé aux scores humains pour l'évaluation de la grammaticalité et, dans une moindre mesure, du transfert sémantique, tandis que SARI est fortement corrélé aux scores humains pour l'évaluation de la simplicité.

Pour illustrer le fonctionnement des deux métriques, le tableau 1 présente un exemple et des scores obtenus à partir de (Xu *et al.*, 2015). Pour cet exemple, BLEU attribue le même score à OUTPUT-2 et OUTPUT-3 puisque les deux mots "now" et "currently" apparaissent dans les phrases de référence, SARI en revanche donne un meilleur score à OUTPUT-2 sur la base de l'utilisation du mot "now", qui n'était pas présent dans la phrase originale.

INPUT	About 95 species are currently accepted .		
REF-1	About 95 species are currently known .		
REF-2	About 95 species are now accepted .		
REF-3	95 species are now accepted .		
	System Output	**B**LEU	**S**ARI
OUTPUT-1	About 95 you now get in .	0.1562	0.2683
OUTPUT-2	About 95 species are now agreed .	0.6435	0.7594
OUTPUT-3	About 95 species are currently agreed.	0.6435	0.5890

TABLE 1 – BLEU par rapport à SARI : exemple de score de (Xu *et al.*, 2015).

Pour l'évaluation du système, deux corpus de tests ont été utilisés pour l'anglais et le français : le corpus *self-test*, composé de 10% de phrases du système en cours de construction (non inclus dans l'entraînement) ; et des corpus de test standard construits par des humains. Pour l'anglais, ce corpus est le corpus Turk (Xu *et al.*, 2015) ayant 8 phrases de référence et, pour le français, le corpus de simplification (Grabar & Cardon, 2018) a été utilisé (Fr-test). Turk est basé sur la Wikipédia simple en anglais et Fr-test est basé sur la Wikipédia française et sur Vikidia.

	English						Synthetic French					
	nombre de	*Self-test*			*Turk-test*		*nombre de*	*Self-test*			*Fr-test*	
Niveau	*phrases*	**BLEU**	**SARI**	**FRE**	**BLEU**	**SARI**	*phrases*	**BLEU**	**SARI**	**FRE**	**BLEU**	**SARI**
0-1	27047	63.66	33.51	71.70	60.51	32.74	26809	49.28	31.50	74.62	6.64	34.96
0-2	23556	43.81	33.40	78.77	48.20	29.02	23175	8.20	18.75	82.19	3.41	34.03
0-3	16114	22.52	31.77	87.92	21.84	21.82	13483	8.33	24.96	87.35	2.96	33.73
0-4	4644	4.67	34.25	86.23	2.13	13.26	4587	4.01	33.91	83.90	1.13	31.50
1-2	29973	55.30	33.96	77.47	51.67	30.45	29613	8.86	18.82	87.73	3.14	33.73
1-3	22728	30.55	34.28	85.03	33.88	26.06	22103	8.49	24.05	88.17	3.82	34.01
1-4	15292	17.26	33.71	93.78	13.56	19.28	13899	11.50	32.49	94.52	2.36	33.98
2-3	24668	40.58	33.74	84.36	43.40	28.24	24238	35.50	33.14	86.29	4.75	34.68
All	226351	50.10	37.69	85.72	47.10	26.22	218079	43.46	36.88	86.37	4.27	34.18

TABLE 2 – Scores BLEU et SARI pour l'anglais et le français traduit, obtenus en faisant varier les niveaux du corpus Newsela utilisés à l'apprentissage.

4.2 Évaluation expérimentale

Notre cadre expérimental a été conçu pour répondre aux questions suivantes :

— En l'absence de ressources suffisantes pour la simplification, est-il possible de construire un système de simplification automatique « raisonnable » pour une langue ?
— Quelle est l'efficacité de l'utilisation de corpus synthétiques pour la tâche de simplification ?

Nous avons utilisé OpenNMT-py (Klein *et al.*, 2017) pour développer des modèles de simplification en anglais et en français. Un réseau neuronal récurrent (RNN) seq2seq à deux couches a été utilisé avec 500 unités de mémoire à long et court terme (LSTM) dans chaque couche. L'apprentissage a été optimisé via l'optimiseur Adam avec un taux d'apprentissage de 0,001 pour toutes les expériences. Nous avons utilisé une taille de lot de 64 et une taille de lot de 128 pour les corpus plus grands. Une validation est effectuée tous les 5000 pas d'apprentissage.

5 Discussion et conclusion

Le tableau 2 présente les résultats obtenus pour tous les systèmes de simplification pour l'anglais et le français et pour tous les niveaux de simplification. Nous utilisons le score FRE comme mesure supplémentaire pour évaluer la simplicité de la production. Les scores FRE ne sont calculés que pour l'"auto-test", car il s'agit d'un ensemble de tests communs à l'anglais et au français qui nous permet d'établir une comparaison entre les systèmes.

On observe que les résultats sont très variés. Pour les systèmes de simplification construits à partir de l'anglais Newsela, une première observation est que les scores BLEU baissent de manière drastique lorsque l'écart entre niveaux de simplicité augmente. Par exemple, pour le self test pour 0 $\{0-1 \Rightarrow 63.66, 0-2 \Rightarrow 43.81, 0-3 \Rightarrow 22.52, 0-4 \Rightarrow 4.67\}$ et ensuite un bond pour $1-2 \Rightarrow 55.30$ qui baisse encore pour $1-3 \Rightarrow 30.55$ et encore pour $1-4 \Rightarrow 17.26$ et enfin un bon score pour $2-3 \Rightarrow 40.58$ mais inférieur à $1-2$. Les scores BLEU supérieurs à 40 sont principalement obtenus pour des niveaux de simplicité consécutifs, c'est-à-dire $0-1$, $0-2$, $1-2$, $1-3$. Cette tendance se retrouve à

la fois dans les corpus Self et Turk.

Cependant, ce phénomène n'est pas aussi unanimement démontré par les systèmes construits à partir de la Newsela française synthétique. Cette tendance est illustrée par les résultats de l'évaluation des systèmes utilisant le test français, à l'exception du score pour $1-3$, et par self-test, à l'exception du score pour $1-2$. Ici aussi, les bons systèmes sur self-test sont les systèmes basés sur des niveaux consécutifs de simplicité, par exemple $0-1$, $1-2$, $1-3$.

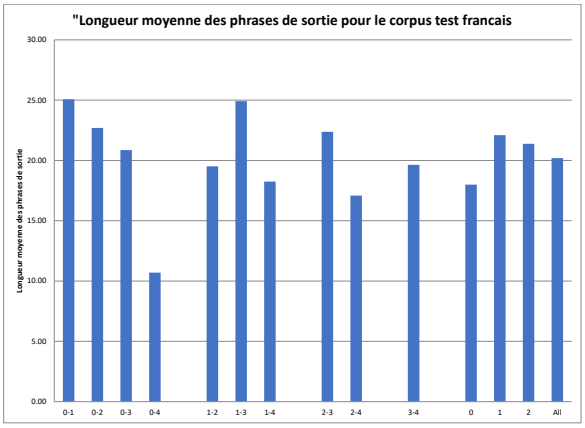


FIGURE 2 – Les longueurs moyennes des phrases de sortie pour le test de français sur différents niveaux de complexité sur Newsela.

SARI, en revanche, donne des résultats relativement stables à tous les niveaux de simplification pour le corpus self-test en anglais et en français. La même tendance à la baisse du score par niveau de simplicité est observée pour le corpus Turk-test en anglais, mais à une échelle moindre. Les scores du SARI sur le corpus self-test pour le français présentent une tendance opposée, c'est-à-dire que le score augmente avec le niveau de simplicité, à l'exception du score pour $0-1$. Contrairement à BLEU, SARI présente une variabilité moindre entre les niveaux de simplicité, mais ce score seul n'est pas assez concluant pour définir la simplicité du texte. Les scores FRE, à l'inverse, donnent une idée complète du niveau de simplicité.

Pour approfondir ces résultats, nous donnons sur la figure 2 la longueur moyenne des phrases pour chaque paire de niveaux. Nous observons des tendances claires : ainsi la longueur moyenne des phrases diminue à mesure que le niveau de simplification augmente. Cela s'observe pour presque tous les niveaux.

La tendance à la baisse est à nouveau évidente pour 2-$\{3, 4\}$. La longueur moyenne de phrase la plus faible concerne les simplifications produites par le système formé en utilisant un corpus de paires de niveaux 0-4, qui est la forme la plus simple.

Nous avons enfin réalisé des évaluations humaines des simplifications produites et avons observé que seule la sortie relative à des niveaux consécutifs de simplification est acceptable. On observe le même phénomène dans les scores BLEU. Les phrases dans les parties les plus simples, par exemple 1-3 et 1-4, sont vraiment très concises du côté simple et les modèles sont incapables d'apprendre les schémas de simplification complexes étant donné les intrications de suppression et de résumé impliquées dans le processus. Cependant, pour les niveaux consécutifs, par exemple 1-2, 2-3 et 3-4, la simplification était acceptable, ce qui indique qu'une simplification automatique basée sur un corpus synthétique est une démarche viable. Plus important encore, la qualité des systèmes appris avec de l'anglais est comparable à celle des systèmes appris avec du français synthétique et ces systèmes montrent un comportement similaire, ce qui tend à montrer que l'utilisation de corpus de simplification synthétique est une démarche viable pour les situations dans lesquelles les ressources sont peu nombreuses.

Références

ABDUL RAUF S., SCHWENK H., LAMBERT P. & NAWAZ M. (2016). Empirical use of information retrieval to build synthetic data for SMT domain adaptation. *ACM/IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, **24**(4), 745–754.

ALVA-MANCHEGO F., BINGEL J., PAETZOLD G., SCARTON C. & SPECIA L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 295–305.

APROSIO A. P., TONELLI S., TURCHI M., NEGRI M. & DI GANGI M. A. (2019). Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, p. 37–44.

BARBU E., MARTÍN-VALDIVIA M. T., MARTÍNEZ-CÁMARA E. & UREÑA-LÓPEZ L. A. (2015). Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, **42**(12), 5076–5086.

BROUWERS L., BERNHARD D., LIGOZAT A.-L. & FRANÇOIS T. (2014). Syntactic sentence simplification for french. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, p. 47–56, Gothenburg, Sweden.

BURLOT F. & YVON F. (2018). Using monolingual data in neural machine translation : a systematic study. In *Proceedings of the Third Conference on Machine Translation*, p. 144–155, Belgium, Brussels : Association for Computational Linguistics. DOI : 10.18653/v1/W18-64015.

CANNING Y. & TAIT J. (1999). Syntactic simplification of newspaper text for aphasic readers. In *ACM SIGIR'99 Workshop on Customised Information Delivery*, p. 6–11.

DAELEMANS W., HÖTHKER A. & SANG E. F. T. K. (2004). Automatic sentence simplification for subtitling in dutch and english. In *LREC*.

DE BELDER J. & MOENS M.-F. (2010). Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, p. 19–26 : ACM ; New York.

EVANS R., ORĂSAN C. & DORNESCU I. (2014). An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, p. 131–140, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.3115/v1/W14-1215.

FAJARDO I., TAVARES G., ÁVILA V. & FERRER A. (2013). Towards text simplification for poor readers with intellectual disability : When do connectives enhance text cohesion ? *Research in developmental disabilities*, **34**(4), 1267–1279.

FLESCH R. (1948). A new readability yardstick. *Journal of applied psychology*, **32**(3), 221.

GRABAR N. & CARDON R. (2018). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, p. 3–9, Tilburg, the Netherlands : Association for Computational Linguistics. DOI : 10.18653/v1/W18-7002.

HUENERFAUTH M., FENG L. & ELHADAD N. (2009). Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, p. 3–10 : ACM.

K. PAPINENI S. R. & WARD R. (1998). Maximum likelihood and discriminative training of direct translation models. In *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, p. 189–192.

KANDEL L. & MOLES A. (1958). Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, **19**(1958), 253–274.

KLEIN G., KIM Y., DENG Y., SENELLART J. & RUSH A. (2017). OpenNMT : Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, p. 67–72, Vancouver, Canada : Association for Computational Linguistics.

KORHONEN A., TRAUM D. & MÀRQUEZ L. (2019). Proceedings of the 57th annual meeting of the association for computational linguistics. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

LAMBERT P., SCHWENK H., SERVAN C. & ABDUL-RAUF S. (2011). Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 284–293, Edinburgh, Scotland : Association for Computational Linguistics.

MCCARTHY J. E. & SWIERENGA S. J. (2010). What we know about dyslexia and web accessibility : a research review. *Universal Access in the Information Society*, **9**(2), 147–152.

MOORE R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In S. D. RICHARDSON, Éd., *Proc. Association for Machine Translation in America (AMTA O2)*, Lecture Notes in Computer Science 2499, p. 135–144, Tiburon, CA, USA : Springer Verlag.

NARAYAN S. & GARDENT C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, p. 435–445.

NISIOI S., ŠTAJNER S., PONZETTO S. P. & DINU L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, p. 85–91.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

RELLO L., BAEZA-YATES R., BOTT S. & SAGGION H. (2013a). Simplify or Help ? Text Simplification Strategies for People with Dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2461121.2461126.

RELLO L., BAEZA-YATES R., DEMPERE-MARCO L. & SAGGION H. (2013b). Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, p. 203–219 : Springer.

SCARTON C., PAETZOLD G. & SPECIA L. (2018). Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

SCARTON C. & SPECIA L. (2018). Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 712–718.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*, p. 371–376, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W16-2323.

SURYA S., MISHRA A., LAHA A., JAIN P. & SANKARANARAYANAN K. (2019). Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2058–2068.

WUBBEN S., VAN DEN BOSCH A. & KRAHMER E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, p. 1015–1024 : Association for Computational Linguistics.

XU W., CALLISON-BURCH C. & NAPOLES C. (2015). Problems in current text simplification research : New data can help. *Transactions of the Association of Computational Linguistics*, **3**(1), 283–297.

ZHANG X. & LAPATA M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 584–594.

ZHAO S., MENG R., HE D., SAPTONO A. & PARMANTO B. (2018). Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3164–3173.

ZHU Z., BERNHARD D. & GUREVYCH I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, COLING 10, p. 1353–1361, Beijing, China.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Représentation sémantique des familles dérivationnelles au moyen de frames morphosémantiques

Daniele Sanacore[1] Nabil Hathout[1] Fiammetta Namer[2]
(1) CLLE, CNRS & Université de Toulouse
(2) ATILF, CNRS & Université de Lorraine

RÉSUMÉ

L'article présente un formalisme de représentation des relations morphologiques dérivationnelles inspiré de la Sémantique des Frames. La description morphosémantique y est réalisée au niveau des familles dérivationnelles au moyen de frames morphosémantiques dans lesquels les lexèmes sont définis les uns relativement aux autres. Les frames morphosémantiques permettent de rendre compte de la structure paradigmatique du lexique morphologique par l'alignement des familles qui présentent les mêmes oppositions de sens. La seconde partie de l'article est consacrée aux données qui seront utilisées pour produire (semi-) automatiquement ces représentations.

ABSTRACT

Semantic representation of derivational families by means of morphosemantic frames

In this paper, we propose a formalism for the morphosemantic description of the morphological families inspired by Frame Semantics.In this formalism, the lexemes are typed semantically and ontologically and are defined with respect to the other members of their family. The morphosemantic frames describe the semantic structure of sets of morphological families. Having identical semantic structures, these families can be aligned into derivational paradigms. In the last section of the paper, we review the data that will be used to (semi-) automatically generate these frames.

MOTS-CLÉS : morphologie dérivationnelle, morphologie paradigmatique, sémantique des frames, types morphosémantiques, types ontologiques.

KEYWORDS: derivational morphology, paradigmatic morphology, frame semantics, morphosemantic types, ontological types.

1 Introduction

Nous proposons dans cet article un nouveau formalisme pour la représentation sémantique des relations dérivationnelles. [1] Ce travail est une contribution au développement de Démonette (Hathout & Namer, 2014, 2016; Namer *et al.*, 2019) une ressource morphologique dérivationnelle du français. Les entrées de Démonette sont des relations dérivationnelles entre des lexèmes de la même famille dérivationnelle. Ces relations peuvent être directes (*rédiger* - *rédacteur*) ou indirectes (*rédaction* - *rédacteur*). Nous considérons dans Démonette que le sens d'un lexème construit est déterminé par l'ensemble des relations sémantiques dans lesquelles il est impliqué. Le formalisme proposé, que nous appellerons **frames morphosémantiques**, permet donc de décrire l'ensemble des relations

1. Ce travail bénéficie du soutien de l'ANR 17-CE23-0005.

HAL : hal-02784784.

morphosémantiques qui existent entre les membres d'une famille dérivationnelle, mais aussi leurs relations argumentales, la et la catégorie ontologique des lexèmes. Les relations morphosémantiques sont décrites au moyen de **gloses morphosémantiques** qui définissent le sens des membres de chaque famille les uns relativement aux autres.

2 Fondements théoriques

La morphologie dérivationnelle abandonne progressivement le morphème comme unité d'analyse en faveur du lexème (Aronoff, 1976; Anderson, 1992; Fradin, 2003). Le morphème est en effet trop rigide et ne permet pas de décrire les très nombreux décalages entre forme et sens, comme dans le cas des constructions dites parasynthétiques (*continent* → *intercontinental*) (Hathout & Namer, 2018; Namer & Hathout, 2019). Un autre développement plus récent est l'adoption d'approches paradigmatiques qui rendent compte d'une manière plus complète des très nombreuses régularités qui existent dans le lexique dérivationnel (Van Marle, 1984; Stump, 1991; Bauer, 1997; Booij, 2008, parmi d'autres). C'est dans ce cadre que s'inscrit notre proposition. Les paradigmes dérivationnels sont basées sur trois structures : les familles, les séries et les paradigmes. Elles sont illustrées en Figure 1.

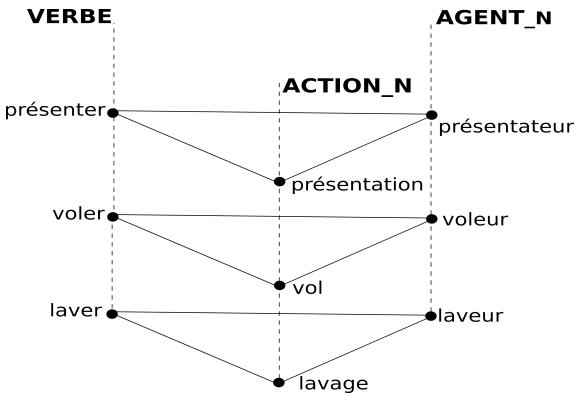


FIGURE 1 – Extrait d'un paradigme dérivationnel

Une **famille dérivationnelle** est un ensemble de lexèmes connectés par des relations dérivationnelles directes ou indirectes (Hathout, 2011). Une relation dérivationnelle directe connecte un dérivé à sa base (*laver* - *lavage*) ou vice-versa. les autres relations dérivationnelles sont indirectes comme *lavage* - *laveur* qui dérivent tous les deux de *laver*. Une **série dérivationnelle** se compose de lexèmes qui se trouvent dans les mêmes relations de contraste avec les autres membres de leurs familles. Par exemple, *coiffage*, *tranchage* et *calage* appartient à la série composé des noms d'action en *-age*. Un **paradigme dérivationnel** est un empilement de **sous-familles** (c'est-à-dire de sous-ensembles de familles dérivationnelles) dont les membres sont liés par les mêmes relations, et qui de ce fait sont superposables. Dans cet article, nous nous intéressons aux relations sémantiques qui s'établissent dans les familles et les paradigmes dérivationnels et notamment à la manière dont la caractérisation sémantique des familles morphologiques met en évidence l'organisation paradigmatique du lexique dérivationnel.

3 Démonette

Démonette est une base de données morphologique dérivationnelle alimentée par des ressources existantes de nature variée (Namer *et al.*, 2019). Dans cette ressource, chaque relation dérivationnelle détermine un ensemble de propriétés sémantiques des lexèmes qu'elle connecte. Le sens d'un dérivé est conçu comme la combinaison des propriétés sémantiques induites par toutes les relations dérivationnelles dans lesquelles il est impliqué. Une entrée dans Démonette décrit un couple de lexèmes d'une famille dérivationnelle, Lex_1 et Lex_2. Elle fournit notamment les propriétés formelles, morphologiques et sémantiques des deux lexèmes et les propriétés formelles, morphologiques et sémantiques de leur relation. La description sémantique des lexèmes et des relations est illustrée dans la Table 1. Elle se compose d'une description ontologique de chacun des deux lexèmes (Typ_1 et Typ_2) et d'une caractérisation sémantique de la relation entre Lex_1 et Lex_2 (trois dernières colonnes).

			Relation sémantique		
Lex_1 - Lex_2	**Typ_1**	**Typ_2**	**Type général**	**Sém. Lex_1**	**Schéma Lex_1**
$danseur_N$ - $danser_V$	person	sit.dyn	entité-situation	agent	ce(lui) qui Lex_2
$admirer_V$ - $admirateur_N$	sit.stat	person	situation-entité	jugement	ressentir ce que ressent Lex_2

TABLE 1: Description sémantique des couples de lexèmes dans Démonette

4 La sémantique des frames

Le sémantique des frames est une théorie de la représentation des connaissances basée sur les frames. Un frame est une structure conceptuelle qui décrit une situation, un objet ou un évènement ainsi que ses participants et leurs propriétés. Les participants d'un frame, appelés **frame elements (FE)**, sont décrits par des rôles sémantiques. Il existe plusieurs implémentations de la sémantique des frames comme FrameNet (Fillmore, 2006; Johnson *et al.*, 2003; Ruppenhofer *et al.*, 2006) pour l'anglais ou Asfalda (Candito *et al.*, 2014) pour le français.

Un frame se compose (*i*) d'une glose globale qui décrit la situation représentée et la manière dont les différents FE interagissent à l'intérieur de celle-ci ; (*ii*) des gloses partielles qui décrivent la situation relativement à chacun des FE ; (*iii*) d'un ensemble d'unités lexicales qui évoquent le frame ; (*iv*) d'un ensemble de phrases qui réalisent le frame. Par exemple, dans Asfalda, le frame CONTACTING est défini par la glose globale en anglais (1) ; le FE **Addressee** est décrit par la glose partielle (2) ; la liste des unités lexicales du français qui évoquent la situation exprimée par le frame CONTACTING est donnée en (3) ; la phrase (4) est un exemple de réalisation de ce frame.

(1) CONTACTING : A **Communicator** directs a communication to an **Addressee** and/or at a particular **Address**, for a particular **Purpose** and/or about a particular **Topic**. A **Location_of_communicator** can be expressed.

(2) **Addressee** : The person that receives the message from the **Communicator**.

(3) *appel téléphonique.n*, *appel.n*, *appeler.v*, *contact.n*, *contacter.v*, *coup de fil.n*, *coup de téléphone.n*, *écrire.v*, *joindre.v*, *téléphoner.v*, *toucher.v*

(4) FR : *Faisant contre mauvaise fortune bon coeur , l' amiral Higgins a alors* ***appelé*** *au téléphone les journalistes britanniques pour les délier de leurs engagements*

5 Adaptation des frames sémantiques à la morphosémantique

Les correspondances qui existent entre les frames sémantiques et les paradigmes dérivationnels permettent une adaptation immédiate des premiers pour décrire les seconds. En effet, les unités lexicales qui évoquent les frames correspondent aux (sous-) familles qui composent les paradigmes. Les FE correspondent aux séries dérivationnelles, c'est-à-dire à des lexèmes jouent le même rôle dans leurs familles. La description du frame correspond au schéma du paradigme dérivationnel, c'est-à-dire à la description des relations qui existent entre les séries. Cette description peut se faire au moyen d'un ensemble de gloses morphosémantiques. Schématiquement, la correspondance entre les frames et les paradigmes peut être résumée comme suit :

frame sémantique	**paradigme**
description du frame	schéma du paradigme
frame element	série de lexèmes
unités lexicales du frame	familles dérivationnelles

La description des FE comporte trois niveaux : le niveau relationnel réunit les **gloses morphosémantiques** ; le niveau argumental décrit les rôles des membres des séries au sein de leur familles ; le niveau ontologique donne la catégorie des membres de ces séries. La dimension ontologique est indépendante de la dimension argumentale. En revanche, les dimensions relationnelles et argumentales sont interdépendantes tout en étant distinctes. Pour présenter plus en détail ces trois niveaux de description, considérons la sous-famille dérivationnelle (5).

(5) *laver* $_{\text{V}}$; *lavage* $_{\text{N}}$; *lavoir* $_{\text{N}}$; *laverie* $_{\text{N}}$; *laveur* $_{\text{N}}$; *laveuse* $_{\text{N}}$; *lavette* $_{\text{N}}$; *lavable* $_{\text{A}}$

Au **niveau relationnel**, on trouve les gloses morphosémantiques qui décrivent les relations qui connectent les lexèmes : les éléments d'une famille y sont définis mutuellement au moyen d'énoncés non orientés comme en (6), de sorte que la définition globale de la famille est composée de l'union de toutes ces gloses simples. Notons que plusieurs membres d'une famille sont susceptibles d'occuper la même position dans une glose de la même manière qu'ils sont susceptibles d'occuper la même case dans le paradigme. C'est le cas de *laveur* et *laveuse* en (6a), puisque les deux noms sont les agents de *laver*. De même, *lavoir* et *laverie* en (6c) décrivent des lieux scéniques où se déroule l'activité *laver*.

(6) a. Un **laveur** (une **laveuse**) **lave** quelque chose

b. Quelque chose est **lavable** si on peut la **laver**

c. On **lave** quelque chose dans un **lavoir** (une **laverie**)

Pour obtenir des gloses morphosémantiques intelligibles et suffisamment naturelles, le nombre d'éléments dans chaque glose est limité à deux ou trois. De ce fait, il n'est pas envisageable de réunir l'ensemble des lexèmes de la famille dans une glose globale unique. Une glose morphosémantique associe en réalité deux ou plusieurs séries dérivationnelles et décrit une superposition de couples au sein des familles du paradigme. Les couples qui instancient la glose se trouvent tous dans la même relation comme l'illustre la Table 2.

Au **niveau argumental**, un rôle sémantique est attribué à chaque membre de la (sous-) famille (ou plus exactement à chaque FE instancié par ce membre), en fonction de la position que le lexème occupe dans la structure argumentale des prédicats de ses gloses du niveau relationnel. Par exemple, pour le frame qui contient la famille de *laver* (5), les rôles sémantiques des FE instanciés par les membres de cette famille sont donnés en Table 3.

	PRÉD.VSUP		PRÉDICAT	
On réalise un(e)	**lavage**	quand on	**lave**	qqc
On réalise un(e)	**vol**	quand on	**vole**	qqc
On réalise un(e)	**gonflement**	quand on	**gonfle**	qqc
On réalise un(e)	**présentation**	quand on	**présente**	qqc
On réalise un(e)	**défense**	quand on	**défend**	qqc
On réalise un(e)	**découverte**	quand on	**découvre**	qqc

TABLE 2: Instances d'une glose morphosémantique qui relient deux séries dérivationnelles. La première est composée de noms d'action et la seconde des verbes correspondants

FE	**rôle**
laver	prédicat
lavage	prédicat à verbe support (pratiquer/réaliser)
laveur / laveuse	agent

FE	**rôle**
lavoir / laverie	lieu
lavette	instrument
lavable	potentialité

TABLE 3: Niveau argumental des FE du frame morphosémantique qui contient la famille de *laver*

Au **niveau ontologique**, nous utilisons pour les noms la version du projet *Fr-SemCor* (Barque *et al.*, 2020) des *Unique Beginners for Nouns* (UB) de *WordNet* (Miller *et al.*, 1990). Cette ontologie est donnée en Table 4. Son niveau de granularité relativement fin permet la caractérisation des FE qui participent à certaines relations dérivationnelles spécifiques comme la suffixation en *-aie* qui construit des noms de plantations à partir de noms de plantes (*oranger* → *orangeraie* ; *palmier* → *palmeraie*). Cette ontologie est complétée par une catégorie SITUATION pour les verbes et une catégorie MODIFIER pour les adjectifs. Nous n'avons pas utilisé les 15 premières catégories de verbes de WordNet (*bodily function and care*, *change*, *communication*, *competition*, etc.) comme nous l'avons fait pour les noms parce que les contraintes imposées par les procédés dérivationnels qui mettent en jeu des verbes sont très générales.

Entity			**Situation**	
Animate Entity	**Non Animate Entity**	**Abstract Entity**	**Stative Situation**	**Dynamic situation**
animal	artifact	cognition	attribute	act
person	groupxartifact	groupxcognition	state	event
groupxanimal	food	communication	feeling	
groupxperson	substance	group		
	object	part		
	plant	quantity		
	body	possession		
		relation		
		phenomenon		

TABLE 4: Catégories ontologiques utilisées dans le projet *Fr-SemCor*

Un frame morphosémantique se compose d'un ensemble de gloses qui définissent ses FE les uns relativement aux autres, d'une caractérisation ontologique et argumentale des FE et d'un ensemble de familles qui instancient ces FE. La Figure 2 présente une partie du frame morphosémantique qui contient la famille de *laver* (5). Ce frame réunit trois FE (prédicat verbal, prédicat nominal,

nom d'agent) définit mutuellement par 3 gloses morphosémantiques pour l'une des familles, en l'occurrence celle de *laver*. Par exemple, la première glose décrit la relation qui existe entre le prédicat nominal exemplifié par *lavage* et le prédicat verbal exemplifié par *laver*. Les catégories grammaticales et ontologiques et les rôles de ces FE sont donnés dans la seconde partie du frame. La dernière partie du frame liste les familles qui composent le frame, en l'occurrence, celles de *laver*, de *voler* et de *présenter*. Chaque membre de chaque famille est assigné à un FE. Par exemple, *voleur* et *voleuse* sont des instances de FE3.

Gloses MorphoSémantiques
On réalise un(e) **lavage**=FE2 quand on **lave**=FE1 qqc
Un(e) **laveur**=FE3 réalise un(e) **lavage**=FE2
Un(e) **laveur**=FE3 **lave**=FE1 qqc

	PoS.Onto	**rôle**
FE1	V.SITUATION	PRÉDICAT
FE2	N.ACTIVITY	PRED.VSUP
FE3	N.PERSON	AGENT

FE1	**FE2**	**FE3**
laver	lavage	laveur/laveuse
voler	vol	voleur/voleuse
présenter	présentation	présentateur/présentatrice

FIGURE 2 – Partie du frame morphosémantique qui contient la famille de *laver*

6 Génération (semi-) automatique des frames morphosémantiques

Nous venons de voir dans ce qui précède la motivation et les principes qui gouvernent la conception de frames morphosémantiques. Différentes ressources existantes peuvent être utilisées pour les produire (semi-)automatiquement pour un ensemble significatif de familles et de paradigmes de Démonette. Une première ressource est Glawinette, un lexique de familles et de séries dérivationnelles du français (Hathout *et al.*, 2020) qui décrit 160113 couples de lexèmes regroupés en 15895 familles. Les couples de lexèmes sont caractérisés par des couples de schémas dérivationnels qui permettent de déduire les FE auxquels ils peuvent se rattacher dans un frame morphosémantique. Nous envisageons d'utiliser les définitions lexicographiques de dictionnaires comme GLAWI (Sajous & Hathout, 2015) pour la génération de gloses morphosémantiques standardisées et la détermination des rôles argumentaux et des catégories ontologiques.

Malgré les variations qu'elles peuvent présenter, les définitions sont une source de connaissance riche et adaptée à la production de frames morphosémantiques comme l'illustrent les exemples sous (7). Les définitions des lexèmes dérivés contiennent souvent un autre membre de leurs famille dérivationnelle ; ces mots sont soulignés dans les exemples de (7) ; ces éléments sont fournis par Glawinette ou peuvent à défaut être identifiés sur la base des analogies qu'ils permettent de former. Le genre prochain dans les définitions permet par ailleurs d'identifier la catégorie ontologique de l'entrée ; ces genres sont encadrés dans les exemples de (7) ; leur extraction peut se faire directement à partir de GLAWI qui

fournit des analyses syntaxiques en dépendance de toutes les définitions. Une fois le genre prochain identifié, la catégorie ontologique peut être déduite en utilisant des correspondances comme en (8).

(7) a. OBSERVATEUR : *personne* *qui observe*

b. SOIGNEUR : *personne* *qui donne des soins et s'occupe de l'état physique d'un sportif*

c. OBSERVATOIRE : *lieu* *d'où l'on peut observer, surveiller l'ennemi*

d. DÉMAIGRISSEMENT : *action* *de démaigrir*

e. ÉTIQUETEUSE : *machine* *qui fait l'étiquetage*

(8) *personne* → person, human being
lieu → artifact OR location
action → act, activity
machine → artifact

La structure syntaxique de la définition peut aussi servir de matrice pour la génération des gloses morphosémantiques. Par exemple, la définition donnée dans la première ligne des des exemples en (9) permet de produire la glose de la 2e ligne et de déduire les rôles et les catégories ontologiques des FE listées dans la 3e ligne. Le nombre des paradigmes du français permet de réaliser cette tâche au moyen de patrons. Elle peut également être réalisées au moyen de réseaux de neurones en adaptant la méthode proposée par (Mickus *et al.*, 2020).

(9) a. DONATEUR : *personne qui a fait une donation*
*Un **donateur** fait une **donation***
AGENT = **donateur** ; PRÉDICAT NOMINAL = **donation** ; VERBE SUPPORT = **faire**

b. EXPOSANT : *personne qui expose un fait*
*Un **exposant expose** quelque chose*
AGENT = **exposant** ; PRÉDICAT = **exposer**

c. OBSERVABLE : *qui peut être observé*
*On peut **observer** quelque chose ou quelqu'un qui est **observable***
MODIFIEUR = **observable** ; PRÉDICAT = **observer**

7 Conclusion

Nous avons montré dans cet article comment des structures inspirées de la sémantique des frames peuvent être adaptées pour permettre la description des relations morphosémantiques qui s'établissent dans les familles dérivationnelles. Les frames morphosémantiques rendent compte également de la superposition de familles et de ces relations à l'intérieur des paradigmes dérivationnelles. Les frames morphosématiques peuvent être produits de manière (semi-)automatique à partir de données structurées provenant des dictionnaires, comme GLAWI et Glawinette.

Références

ANDERSON S. R. (1992). *A-morphous morphology*, volume 62. Cambridge University Press.

ARONOFF M. (1976). *Word Formation in Generative Grammar*. Linguistic Inquiry Monographs. Cambridge, MA : MIT Press.

BARQUE L., HAAS P., HUYGHE R., TRIBOUT D., CANDITO M., CRABBÉ B. & SEGONNE V. (2020). FrSemCor : Annotating a French corpus with supersenses. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

BAUER L. (1997). Derivational paradigms. In *Yearbook of Morphology 1996*, p. 243–256. Springer.

BOOIJ G. (2008). Paradigmatic morphology. *La raison morphologique. Hommage à la mémoire de Danielle Corbin*, p. 29–38.

CANDITO M., AMSILI P., BARQUE L., BENAMARA ZITOUNE F., DE CHALENDAR G., DJEMAA M., HAAS P., HUYGHE R., YANNICK MATHIEU Y., MULLER P., SAGOT B. & VIEU L. (2014). Developing a French Framenet : Methodology and first results. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

FILLMORE C. J. E. A. (2006). Frame semantics. *Cognitive linguistics : Basic readings*, **34**, 373–400.

FRADIN B. (2003). *Nouvelles approches en morphologie*. Presses Universitaires de France.

HATHOUT N. (2011). Morphonette : a paradigm-based morphological network. *Lingue e linguaggio*, **10**(2), 245–264.

HATHOUT N. & NAMER F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, **11**(5), 125–168.

HATHOUT N. & NAMER F. (2016). Giving lexical resources a second life : Démonette, a multi-sourced morpho-semantic network for french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

HATHOUT N. & NAMER F. (2018). La parasynthèse à travers les modèles : des RCL au ParaDis. In O. BONAMI, G. BOYÉ, G. DAL, H. GIRAUDO & F. NAMER, Éds., *The lexeme in descriptive and theorical morphology*, p. 365–399. Langage Sciences Press.

HATHOUT N., SAJOUS F., CALDERONE B. & NAMER F. (2020). Glawinette : a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

JOHNSON C. R., SCHWARZER-PETRUCK M., BAKER C. F., ELLSWORTH M., RUPPENHOFER J. & FILLMORE C. J. (2003). *Framenet : Theory and practice*. Rapport interne, International Computer Science Institute, Berkeley, CA.

MICKUS T., CONSTANT M. & PAPERNO D. (2020). Génération automatique de définition pour le français. In *Actes de la 27e conférence annuelle sur le traitement automatique des langues naturelles (TALN-2020)*.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. J. (1990). Introduction to WordNet : An on-line lexical database. *International Journal of Lexicography*, **3**(4), 335–391.

NAMER F., BARQUE L., BONAMI O., HAAS P., HATHOUT N. & TRIBOUT D. (2019). Démonette2 – une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de la 26e conférence annuelle sur le traitement automatique des langues naturelles (TALN-2019)*, p. 233–243, Toulouse.

NAMER F. & HATHOUT N. (2019). ParaDis and Démonette : From theory to resources for derivational paradigms. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, p. 5–14.

RUPPENHOFER J., ELLSWORTH M., PETRUCK M. R., JOHNSON C. R. & SCHEFFCZYK J. (2006). *FrameNet II : Extended Theory and Practice*. Berkeley, California : International Computer Science Institute. Distributed with the FrameNet data.

SAJOUS F. & HATHOUT N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, p. 405–426, Herstmonceux, England.

STUMP G. T. (1991). A paradigm-based theory of morphosemantic mismatches. *Language*, p. 675–725.

VAN MARLE J. (1984). *On the paradigmatic dimension of morphological creativity*. Dordrecht : foris.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Modèle neuronal pour la résolution de la coréférence dans les dossiers médicaux électroniques

Julien Tourille[1] Olivier Ferret[1] Aurélie Névéol[2] Xavier Tannier[3]
(1) CEA, LIST, Laboratoire Analyse Sémantique Texte et Image, Gif-sur-Yvette, F-91191, France
(2) LIMSI, CNRS, Université Paris-Saclay
(3) Sorbonne Université, Inserm, LIMICS
{julien.tourille,olivier.ferret}@cea.fr, aurelie.neveol@limsi.fr,
xavier.tannier@sorbonne-universite.fr

RÉSUMÉ

La résolution de la coréférence est un élément essentiel pour la constitution automatique de chronologies médicales à partir des dossiers médicaux électroniques. Dans ce travail, nous présentons une approche neuronale pour la résolution de la coréférence dans des textes médicaux écrits en anglais pour les entités générales et cliniques en nous évaluant dans le cadre de référence pour cette tâche que constitue la tâche 1C de la campagne i2b2 2011.

ABSTRACT

Neural approach for coreference resolution in electronic health records

Coreference resolution is an essential step for clinical timeline extraction from electronic health records. Herein, we present a neural approach for coreference resolution in clinical documents written in English for both general and clinical entities and we evaluate it in the reference evaluation framework of the task 1C of the i2b2 2011 campaign.

MOTS-CLÉS : TAL clinique, réseaux de neurones, résolution de la coréférence.

KEYWORDS: clinical NLP, deep learning, coreference resolution.

1 Introduction

La résolution de la coréférence consiste à identifier toutes les mentions d'entités ou d'événements et à les regrouper en classes d'équivalence (Pradhan *et al.*, 2011). Cette tâche n'implique pas de déterminer à quels entités ou événements ces mentions font référence. Il s'agit de déterminer si plusieurs mentions font référence à la même entité ou événement. La résolution de la coréférence a principalement été explorée pour des textes de nature journalistique, avec de premières campagnes d'évaluation dans les années 1990 (Sundheim, 1995; Hirschman & Chinchor, 1998). Ce n'est qu'au cours des dix dernières années que des travaux se sont intéressés à la résolution de la coréférence dans le domaine clinique. La première campagne d'évaluation sur le sujet a été proposée par la fondation i2b2 (Uzuner *et al.*, 2012) et a permis le développement de plusieurs modèles (Jindal & Roth, 2013; Hinote *et al.*, 2011; Grouin *et al.*, 2011; Chowdhury & Zweigenbaum, 2013). L'intérêt pour cette tâche répond au besoin grandissant d'explorer et d'utiliser les données contenues dans les rapports et autres documents textuels qui composent les dossiers médicaux électroniques. Parmi ces données, la

Volume 2 : Traitement Automatique des Langues Naturelles, pages 351–360.
HAL : hal-02784785.

chronologie médicale, c'est-à-dire la suite d'événements médicaux qui ont lieu au cours de la vie d'un patient, est une information importante. Être en mesure d'extraire automatiquement ces chronologies permettrait de mieux comprendre certains phénomènes médicaux tels que l'évolution des maladies et les effets longitudinaux des médicaments (Lin *et al.*, 2016; Sun *et al.*, 2013). Or, les événements médicaux sont mentionnés plusieurs fois dans les dossiers, rendant difficile la construction de ces chronologies sans une étape de résolution de la coréférence.

Dans ce travail, nous nous intéressons à la résolution de la coréférence dans des dossiers médicaux électroniques écrits en anglais. Nous proposons une approche neuronale fondée sur les travaux récents appliqués aux textes journalistique (Wiseman *et al.*, 2016; Clark & Manning, 2016).

2 Données

Dans ce travail, nous utilisons le corpus i2b2 tâche 1C (Uzuner *et al.*, 2012) et plus précisément, la partie i2b2/VA ne contenant pas de documents de University of Pittsburgh Medical Center (UPMC). Nous reproduisons ainsi un des contextes expérimentaux proposés lors de la campagne d'évaluation. Le corpus contient 194 documents cliniques du Beth Israel Deaconess Medical Center (BETH) et 230 documents cliniques de Partners Healthcare (PARTNERS) (cf. tableau 1). Le tableau 2 présente le nombre de chaînes de coréférence ainsi que leur longueur moyenne et maximale.

Institution	Train	Test	Total
BETH	115	79	194
PARTNERS	136	94	230
Combinés	251	173	424

TABLE 1: Statistiques concernant le corpus i2b2/VA tâche 1C

Institution	Nombre chaînes	Long. moy.	Long. max.
BETH	1 816	4,2	122
PARTNERS	1 395	4,4	105

TABLE 2: Statistiques concernant le nombre et la longueur des chaînes de coréférence du corpus i2b2/VA tâche 1C

Cinq types d'éléments sont annotés dans le corpus en y incluant les singletons, *i.e.* les éléments non coréférents : personnes, pronoms, tests, traitements et problèmes. Les chaînes de coréférence peuvent être divisées en deux groupes : les chaînes relatives aux *événements* (tests, traitements et problèmes) et les chaînes relatives aux *personnes*. Les deux groupes présentent des caractéristiques différentes : les chaînes *personnes* sont plus longues (moyenne de 12,44 contre environ 2,5 pour les événements) tandis que les mentions composant les chaînes *personnes* prennent généralement la forme de pronoms personnels, même si des pronoms peuvent également faire partie des chaînes *événements*. Enfin, les événements médicaux ont une structure argumentale implicite que les mentions de personnes n'ont pas (Styler IV *et al.*, 2014). Ces trois différences nous ont amené à considérer la résolution de la coréférence pour ces deux ensembles d'entités comme des sous-tâches distinctes. En conséquence, nous apprenons deux modèles distincts, un pour chaque sous-ensemble. Nous faisons l'hypothèse que le modèle sera capable de distinguer quels pronoms appartiennent aux deux sous-ensembles et nous les incluons dans les modèles caractérisant chaque sous-ensemble.

3 Description du modèle

Notre modèle s'inspire des approches neuronales récemment développées (Lee *et al.*, 2017; Wiseman *et al.*, 2016). Le composant principal de notre approche est un modèle de type *mention-ranking* (Denis & Baldridge, 2008) dont l'objectif est d'ordonnancer l'ensemble des antécédents possibles pour une mention donnée et de choisir le premier. Le cas non-anaphorique, c'est-à-dire l'absence d'antécédent pour une mention donnée, est géré par l'utilisation d'un *dummy antecedent* (Durrett & Klein, 2013; Wiseman *et al.*, 2016).

Notre approche diffère des modèles locaux pour la résolution de la coréférence dans lesquels les paires de mentions sont considérées séparément. Nous proposons d'utiliser des traits globaux extraits des chaînes de coréférence en cours de construction. Ainsi, notre approche s'inspire et s'inscrit dans une lignée de travaux récents examinant l'utilisation de ce type de traits dans des approches neuronales (Clark & Manning, 2015, 2016; Wiseman *et al.*, 2016, 2015).

Plus spécifiquement, nous incorporons de l'information concernant les chaînes de coréférence en cours de construction dans notre modèle *mention-ranking*. Cette information est construite en utilisant un LSTM (Hochreiter & Schmidhuber, 1997) qui examine les différentes mentions des chaînes par ordre d'apparition dans le texte. Le principal avantage de cette approche est qu'elle facilite l'inférence en ne requérant qu'une seule passe sur les mentions (de gauche à droite).

Plongements en entrée Les plongements utilisés en entrée de notre modèle sont construits en concaténant une représentation dense des caractères et un vecteur de mot pré-calculé sur un grand corpus. La représentation dense des caractères est construite suivant la méthode proposée par Lample *et al.* (2016). Un plongement aléatoire est d'abord généré pour chaque caractère présent dans le corpus. Ensuite, les caractères des différents tokens passent à travers un Bi-LSTM. Les deux représentations denses résultantes sont enfin concaténées pour former la représentation finale.

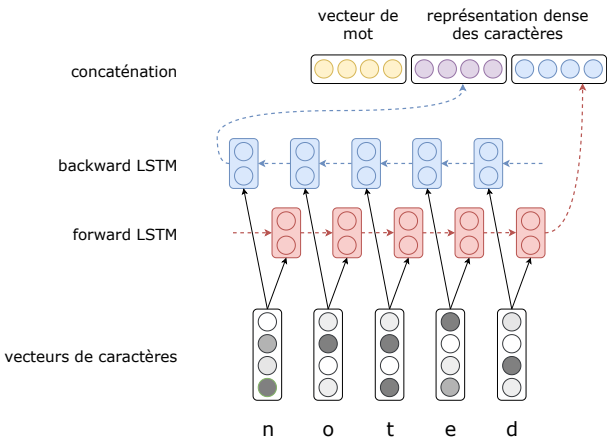


FIGURE 1: Construction des plongements en entrée de notre modèle

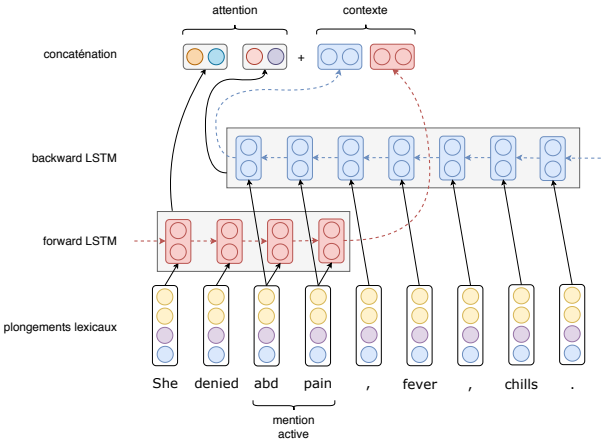


FIGURE 2: Construction des représentations des mentions

Représentation des mentions Le processus de construction des représentations des mentions est présenté à la figure 2. Dans notre exemple, la mention considérée est *abd pain* dans le contexte de la

phrase *She denied abd pain, fever, chills*. Tout d'abord, notre modèle calcule deux représentations contextuelles de la mention considérée grâce à un Bi-LSTM. Le *forward* LSTM prend en entrée le segment allant du premier token de la phrase jusqu'au dernier token de la mention. Le *backward* LSTM prend en entrée le segment allant du dernier token de la phrase jusqu'au premier token de la mention. Les deux représentations denses forment la première partie de la représentation finale.

Nous ajoutons ensuite une représentation dense issue d'un mécanisme d'attention. Ce mécanisme, qui permet d'accorder une importante différenciée aux deux éléments constituant la représentation contextuelle d'une mention, consiste en une somme pondérée des états cachés intermédiaires des deux LSTMs. Les poids utilisés dans la somme pondérée sont calculés en utilisant un réseau de neurones *feed-forward*.

Représentation des chaînes de coréférence Pour le calcul des représentations denses des chaînes de coréférence, nous utilisons, à l'instar de Wiseman *et al.* (2016), un LSTM prenant en entrée les différentes mentions composant les chaînes de coréférence, dans l'ordre de leur apparition dans le document. Un aperçu du processus complet est présenté à la figure 3. Bien entendu, nous maintenons au cours de l'analyse d'un document autant de ces représentations que de chaînes de coréférence, représentations construites en utilisant le même LSTM (*i.e.* les mêmes poids).

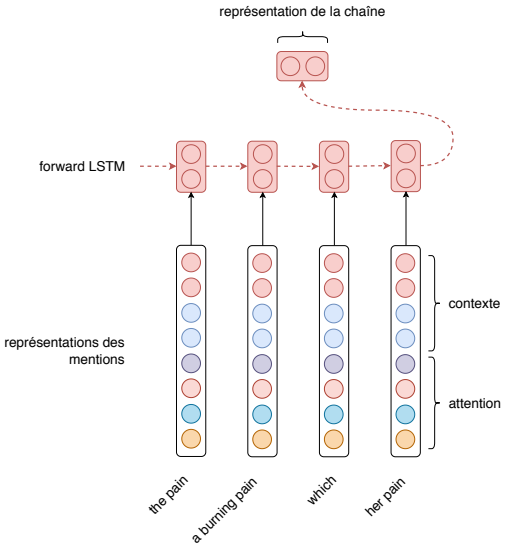


FIGURE 3: Calcul des représentations des chaînes de coréférence

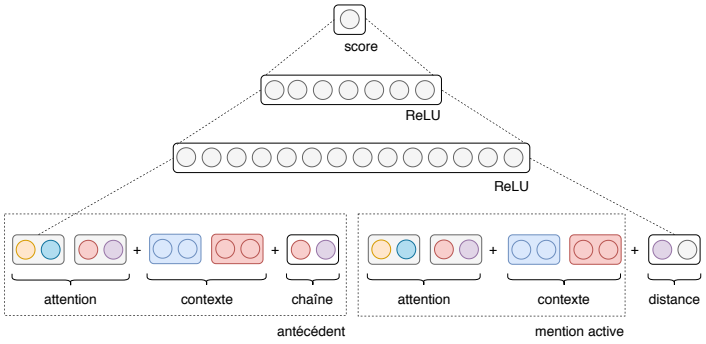


FIGURE 4: Classifieur de paires de mentions

Classifieur de paires de mentions Le rôle du classifieur est d'assigner un score à chaque paire (mention, antécédent). Pour cela, nous utilisons un réseau *feed-forward* qui prend en entrée la concaténation des représentations des mentions, de la chaîne de coréférence dont l'antécédent fait partie et d'une représentation dense de la distance qui sépare la mention active de l'antécédent considéré. La distance est discrétisée en un nombre fixe de catégories : $[1, 2, 3, 4, 5-7, 8-15, 16-31, 32-63, 64+]$ (Clark & Manning, 2016). Le réseau feed-forward est composé de deux couches cachées, chaque couche ayant pour taille la moitié de la taille de la couche précédente. Les scores obtenus pour chaque paire (mention, antécédent) sont concaténés dans un vecteur et nous appliquons une fonction *softmax*. Nous optimisons la vraisemblance de tous les antécédents inclus dans la chaîne de coréférence (Lee *et al.*, 2017).

4 Contexte expérimental

Dans ce travail, nous nous intéressons à la situation dans laquelle nous disposons d'une détection parfaite des mentions dans les documents en utilisant les annotations *gold* du corpus pour ces mentions. Nous testons notre modèle dans plusieurs configurations. Tout d'abord, nous implémentons un modèle de base en enlevant certaines parties de notre architecture. Les plongements présentés en entrée du modèle sont composés uniquement d'un vecteur de mot pré-entraîné. Le mécanisme d'attention ainsi que les représentations denses des chaînes de coréférence et des caractères ne sont pas utilisés.

En partant de cette configuration, nous activons chaque composant pour mesurer son effet sur la performance du système. Par ailleurs, nous implémentons une stratégie de pré-entraînement dans laquelle nous entraînons notre modèle uniquement sur les mentions coréférentes. La tâche se résume alors à trouver l'antécédent correct parmi les mentions précédentes. Cette stratégie permet généralement d'améliorer les performances des systèmes (Clark & Manning, 2015, 2016; Wiseman *et al.*, 2015, 2016). Ensuite, nous implémentons une spécialisation du classifieur de paires de mentions. Le réseau feed-forward qui le compose est unique pour chaque type d'entité. Notre hypothèse est que les éléments permettant d'identifier des mentions coréférentes peuvent différer selon le type d'entité considéré. Enfin, nous implémentons un filtrage des antécédents selon le type de la mention considérée par le modèle. Pour le cas des pronoms, nous considérons tous les antécédents.

Nous implémentons notre modèle avec PyTorch (Paszke *et al.*, 2017). La taille des LSTMs utilisés pour calculer les représentations des mentions et des chaînes de coréférence est fixée à 100. Celle du Bi-LSTM utilisé pour la représentation des caractères est fixée à 25. Les plongements de caractères ont une taille de 25 et sont initialisés aléatoirement. Notre modèle est entraîné sur des mini-lots de 1 document. Nous utilisons l'algorithme d'optimisation Adam et fixons le taux d'apprentissage à 0,001. Nous implémentons une décroissance du taux d'apprentissage de 1 % à chaque itération. Les plongements lexicaux pré-entraînés sont appris sur le corpus MIMIC-III (Johnson *et al.*, 2016) avec une taille de 100. Nous appliquons un *dropout* sur les couches cachées de notre classifieur de mentions avec un taux de 0,2. Un *dropout* est aussi appliqué sur les plongements en entrée avec un taux de 0,5. Nous implémentons l'apprentissage des états initiaux des LSTMs (Gers *et al.*, 2002). Finalement, nous répétons chaque expérience 10 fois pour prendre en considération l'aspect non-déterministe de notre modèle (Reimers & Gurevych, 2017).

5 Résultats

Le tableau 3 présente les résultats de nos expériences, obtenus grâce à la mesure CoNLL calculée avec l'outil de référence de Pradhan *et al.* (2014) [1]. Nous considérons uniquement les clusters regroupant au moins deux mentions et excluons ainsi, à l'instar d'une grande partie des travaux en domaine général, les singletons lors du calcul de la performance. L'utilisation d'un mécanisme d'attention

1. Nous avons préféré cet outil à celui de la campagne i2b2 dans la mesure où suite à divers travaux méthodologiques réalisés sur l'évaluation de la coréférence en domaine général (Pradhan *et al.*, 2011, 2014), il s'y est imposé comme un standard. En outre, Pradhan *et al.* (2014) soulignent que l'outil i2b2 s'appuie sur l'approche de Cai & Strube (2010), dont l'outil associé présente une erreur de mise en œuvre, peut-être également présente dans l'outil i2b2.

permet d'améliorer la performance des modèles *personnes* et *événements*, mais avec une contribution minime comme le suggère la faible différence par rapport au réseau initial. Aucun des deux modèles ne semble bénéficier de l'information issue des représentations denses des caractères, malgré leur contribution dans d'autres tâches (*e.g.* reconnaissance d'entités nommées). La prise en compte des chaînes de coréférence en cours de construction n'améliore pas la performance de notre modèle, avec une baisse supérieure à un point pour les événements. Ces résultats sont en contradiction avec ceux présentés par Wiseman *et al.* (2016). Cependant, les contextes expérimentaux diffèrent. Wiseman *et al.* (2016) travaillent sur des documents journalistiques et rapportent des résultats issus d'un seul run. Or Reimers & Gurevych (2017) suggèrent que les systèmes non déterministes tels que les modèles neuronaux doivent être évalués sur plusieurs runs afin de prendre en compte cette variabilité inhérente.

condition	P	R	F1		P	R	F1		P	R	F1	
	Personnes				Événements				Combinés			
baseline	87,77 (± 0,71)	82,82 (± 0,99)	85,17 (± 0,46)		65,15 (± 1,81)	54,62 (± 1,35)	59,30 (± 0,45)		76,89 (± 1,31)	67,94 (± 0,89)	72,02 (± 0,38)	
attention	88,08 (± 0,82)	83,42 (± 0,55)	85,64 (± 0,38)	↑	66,39 (± 1,45)	54,18 (± 2,08)	59,56 (± 0,86)	↑	77,77 (± 1,01)	67,93 (± 1,27)	72,41 (± 0,40)	↑
caractères	87,87 (± 0,98)	82,55 (± 0,69)	85,08 (± 0,21)	↓	66,39 (± 1,57)	53,26 (± 1,76)	59,02 (± 0,92)	↓	77,78 (± 1,14)	67,13 (± 1,01)	71,96 (± 0,46)	↓
chaîne	87,48 (± 0,77)	82,36 (± 0,72)	84,71 (± 0,49)	↓	65,86 (± 1,03)	52,12 (± 1,10)	58,11 (± 0,58)	↓	77,29 (± 0,69)	66,55 (± 0,64)	71,37 (± 0,37)	↓
filtrage	87,60 (± 1,48)	82,82 (± 0,89)	85,08 (± 0,39)	↓	65,42 (± 1,31)	54,37 (± 1,37)	59,30 (± 0,56)	=	76,98 (± 1,00)	67,85 (± 0,90)	72,03 (± 0,29)	↑
spécialisation	88,52 (± 0,76)	82,34 (± 0,86)	85,25 (± 0,34)	↑	63,66 (± 1,12)	49,62 (± 1,60)	55,68 (± 0,72)	↓	76,64 (± 0,84)	64,92 (± 1,00)	70,17 (± 0,38)	↓
pré-entraînement	88,93 (± 0,30)	82,60 (± 0,79)	85,60 (± 0,41)	↑	68,63 (± 2,09)	56,70 (± 2,77)	61,99 (± 1,11)	↑	79,09 (± 1,43)	69,27 (± 1,53)	73,79 (± 0,53)	↑
optimal	88,65 (± 1,22)	82,91 (± 0,69)	85,62 (± 0,61)		67,72 (± 1,34)	57,51 (± 1,61)	62,16 (± 0,78)		78,36 (± 1,04)	69,86 (± 1,10)	73,82 (± 0,47)	

TABLE 3: Résultat des expériences sur les mentions *gold* du corpus de test

La stratégie de filtrage n'apporte pas d'amélioration nette : nous observons une légère baisse de performance pour le modèle *personnes* tandis que la performance du modèle *événements* reste stable. La spécialisation du classifieur de paires de mentions permet d'améliorer la performance du modèle *personnes* mais diminue fortement la performance du modèle *événements*. Ce résultat négatif pourrait s'expliquer par le volume de données d'entraînement disponible. Le corpus de référence dans le domaine général comprend plus de 2 000 documents (Pradhan *et al.*, 2011) alors que dans notre cas, nous en avons seulement 200. Enfin, le pré-entraînement permet d'améliorer fortement la performance globale de notre modèle, corroborant ainsi les résultats obtenus dans d'autres travaux (Clark & Manning, 2016; Wiseman *et al.*, 2016). L'amélioration est modeste pour le modèle *personnes*. La prévalence des singletons dans des mentions relatives aux personnes (10 %) pourrait limiter l'effet bénéfique du pré-entraînement.

La dernière ligne du tableau 3 donne la performance optimale obtenue en sélectionnant la meilleure configuration pour chaque type d'entités : attention, pré-entraînement dans les deux cas ; spécialisation pour les personnes et filtrage pour les événements. La performance globale de 73,82 en f1-mesure CoNLL est à comparer aux performances des systèmes ayant participé à la tâche 1C de la campagne d'évaluation i2b2 de 2011 pour lesquels nous rapportons les performances dans le tableau 4 (recalculées dans les conditions où nous nous plaçons). Notre système se placerait ainsi entre celui de Cai *et al.* (2011) (classé $5^{ème}$ lors de la campagne d'évaluation) et celui de (Jindal & Roth, 2013) (classé $9^{ème}$ lors de la campagne d'évaluation). Il faut noter que le calcul des performances *via* le script CoNLL a un effet sur le classement initial des systèmes comme on peut le voir dans le tableau 4.

	# i2b2	P	R	F1
Xu *et al.* (2011)	1	82,38	78,13	80,20
Cai *et al.* (2011)	5	75,27	73,96	74,60
notre modèle		78,36	69,86	73,82
Jindal & Roth (2013)	9	65,53	83,48	73,41
Dai *et al.* (2011)	8	76,08	65,65	70,48
Anick *et al.* (2011)	7	79,61	61,67	69,41

TABLE 4: Comparaison de notre système à ceux de la campagne i2b2. Les scores sont obtenus *via* le script officiel CoNLL et calculés en excluant les singletons. Nous rapportons les scores des systèmes pour lesquels une conversion entièrement automatique du format i2b2 vers CoNLL a été possible.

Les modèles développés par les participants lors de la campagne d'évaluation sont fondés sur des traits linguistiques choisis manuellement et intégrés dans des approches à base de règles, d'apprentissage automatique ou une combinaison des deux. Contrairement à ces approches, notre modèle est entièrement neuronal et ne repose pas sur des traits choisis manuellement. Parmi les pistes d'amélioration envisagées figure l'utilisation de ressources externes. Cette possibilité est explorée dans le travail de Zhang *et al.* (2019). Leurs résultats sont néanmoins difficilement comparables aux nôtres dans la mesure où Zhang *et al.* (2019) se focalisent sur les pronoms de façon exclusive et ne traitent donc qu'une partie de la tâche. Par ailleurs, l'utilisation de modèles de langue fondés sur la notion de transformer (Vaswani *et al.*, 2017), tels que le modèle BERT (Devlin *et al.*, 2019), pourrait améliorer les performances de notre modèle, à l'instar du domaine général (Joshi *et al.*, 2019).

6 Conclusion

Nous présentons un modèle neuronal pour la résolution de la coréférence dans le domaine médical, appliqué sur un corpus clinique en anglais. Nous montrons que ce type d'approche permet d'obtenir des performances intéressantes, mais qui restent toutefois inférieures à l'état de l'art à l'aide de modèles fondés sur des traits catégoriels. Dans d'autres expériences, nous avons exploré une situation réelle incluant la détection de mentions en amont de la coréférence (Tourille, 2018). La suite de ce travail pourra explorer plusieurs pistes. Nos expériences font l'hypothèse de la détection parfaite des mentions. Dans une situation d'application réelle, il est nécessaire de procéder à l'extraction des mentions. De plus, les aspects temporels des événements ne sont actuellement pas pris en compte dans notre approche. L'utilisation d'informations temporelles pertinentes permettrait d'apporter des éléments utiles à notre modèle en filtrant les mentions temporellement incompatibles. Le code développé pour convertir le corpus i2b2 du format original vers le format CoNLL est disponible à cette adresse : `https://github.com/jtourille/i2b2-coref-task1c-converter`.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche sous la référence CABeRneT ANR-13-JS02-0009-01 et d'une aide du labex DigiCosme sous la référence CÔT.

Références

ANICK P., HONG P., XUE N. & AL. (2011). Coreference resolution for electronic medical records. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

CAI J., MUJDRICZA E., HOU Y. & AL. (2011). Weakly supervised graph-based coreferencere solution for clinical texts. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

CAI J. & STRUBE M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, p. 28–36, Tokyo, Japan : Association for Computational Linguistics.

CHOWDHURY M. F. M. & ZWEIGENBAUM P. (2013). A controlled greedy supervised approach for co-reference resolution on clinical text. *Journal of Biomedical Informatics*, **46**, 506–515.

CLARK K. & MANNING C. D. (2015). Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the Joint Conference of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, p. 1405–1415 : Association for Computational Linguistics.

CLARK K. & MANNING C. D. (2016). Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 643–653 : Association for Computational Linguistics.

DAI H. J., WU C. Y., CHEN C. Y. & AL. (2011). Co-reference resolution of the medical concepts inthe patient discharge summaries. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

DENIS P. & BALDRIDGE J. (2008). Specialized Models and Ranking for Coreference Resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 660–669 : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4171–4186 : Association for Computational Linguistics.

DURRETT G. & KLEIN D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1971–1982 : Association for Computational Linguistics.

GERS F. A., SCHRAUDOLPH N. N. & SCHMIDHUBER J. (2002). Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research*, **3**, 115–143.

GROUIN C., DINARELLI M., ROSSET S., WISNIEWSKI G. & ZWEIGENBAUM P. (2011). Coreference Resolution in Clinical Reports. The LIMSI Participation in the i2b2/VA 2011 Challenge. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*.

HINOTE D., RAMIREZ C. & CHEN P. (2011). A Comparative Study of Coreference Resolution in Clinical Text. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*.

HIRSCHMAN L. & CHINCHOR N. A. (1998). MUC-7 Coreference Task Definition. In *Proceedings of the 7th Message Understanding Conference* : Morgan Kaufmann.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780.

JINDAL P. & ROTH D. (2013). Using domain knowledge and domain-inspireddiscourse model for coreference resolution forclinical narratives. *Journal of the American Medical Informatics Association*, **20**, 356–362.

JOHNSON A. E., POLLARD T. J., SHEN L., LI-WEI H. L., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., CELI L. A. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, **3**, 160035.

JOSHI M., LEVY O., WELD D. S. & ZETTLEMOYER L. (2019). Bert for coreference resolution : Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, p. 5803–5808 : Association for Computational Linguistics.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270 : Association for Computational Linguistics.

LEE K., HE L., LEWIS M. & ZETTLEMOYER L. (2017). End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 188–197 : Association for Computational Linguistics.

LIN C., DLIGACH D., MILLER T. A., BETHARD S. & SAVOVA G. K. (2016). Multilayered Temporal Modeling for the Clinical Domain. *Journal of the American Medical Informatics Association*, **23**(2), 387–395.

PASZKE A., GROSS S., CHINTALA S., CHANAN G., YANG E., DEVITO Z., LIN Z., DESMAISON A., ANTIGA L. & LERER A. (2017). Automatic Differentiation in PyTorch. In *Proceedings of the NIPS 2017 Autodiff Workshop*.

PRADHAN S., LUO X., RECASENS M., HOVY E., NG V. & STRUBE M. (2014). Scoring Coreference Partitions of Predicted Mentions : A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 30–35, Baltimore, Maryland : Association for Computational Linguistics.

PRADHAN S., RAMSHAW L., MARCUS M., PALMER M., WEISCHEDEL R. & XUE N. (2011). CoNLL-2011 Shared Task : Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning*, p. 1–27 : Association for Computational Linguistics.

REIMERS N. & GUREVYCH I. (2017). Reporting Score Distributions Makes a Difference : Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 338–348 : Association for Computational Linguistics.

STYLER IV W. F., BETHARD S., FINAN S., PALMER M., PRADHAN S., DE GROEN P. C., ERICKSON B., MILLER T., LIN C., SAVOVA G. & PUSTEJOVSKY J. (2014). Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, **2**, 143–154.

SUN W., RUMSHISKY A. & UZUNER O. (2013). Temporal Reasoning over Clinical Text : The State of the Art. *Journal of the American Medical Informatics Association*, **20**, 814–819.

SUNDHEIM B. M. (1995). Overview of Results of the MUC-6 Evaluation. In *Proceedings of the 6th Message Understanding Conference*, volume 423–442 : Morgan Kaufmann.

TOURILLE J. (2018). *Extracting Clinical Event Timelines : Temporal Information Extraction and Coreference Resolution in Electronic Health Records*. Thèse de doctorat, Université Paris-Saclay.

UZUNER Ö., BODNARI A., SHEN S., FORBUSH T., PESTIAN J. & SOUTH B. R. (2012). Evaluating the State of the Art in Coreference Resolution for Electronic Medical Records. *Journal of the American Medical Informatics Association*, **19**(5), 786–791.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, p. 5998–6008. Curran Associates, Inc.

WISEMAN S., RUSH A. M., SHIEBER S. & WESTON J. (2015). Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the Joint Conference of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, p. 1416–1426 : Association for Computational Linguistics.

WISEMAN S., RUSH A. M. & SHIEBER S. M. (2016). Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 994–1004 : Association for Computational Linguistics.

XU Y., LIU J., WU J. & AL. (2011). EHUATUO : a mention-pair coreference system by exploiting document intrinsic latent structures and world knowledge in discharge summaries : 2011 i2b2 challenge. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

ZHANG H., SONG Y., SONG Y. & YU D. (2019). Knowledge-aware Pronoun Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 867–876 : Association for Computational Linguistics.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Un corpus d'évaluation pour un système de simplification discursive

Rodrigo Wilkens [1] Amalia Todirascu[1]
(1)LiLPa – University of Strasbourg, 22, rue René Descartes, 67084 Strasbourg cedex, France
rswilkens@gmail.com, todiras@unistra.fr

RÉSUMÉ

Nous présentons un nouveau corpus simplifié, disponible en français pour l'évaluation d'un système de simplification discursive. Ce système utilise des chaînes de référence pour simplifier et pour préserver la cohésion textuelle après simplification. Nous présentons la méthodologie de collecte de corpus (via un formulaire, qui recueille les simplifications manuelles faites par des participants experts), les règles présentées dans le guide, une analyse des types de simplifications et une évaluation de notre corpus, par comparaison avec la sortie du système de simplification automatique.

ABSTRACT

An Evaluation Corpus for Automatic Discourse Simplification

We present a new public available corpus of simplified French targeting the evaluation of a discourse simplification system aware of the coreference chains able to maintain the textual cohesion. We describe the corpus collecting method based on Web questionnaire (collecting several simplified sentence variants), the rules described in the guidelines, an analysis of variants provided, and the discourse simplification evaluation by comparing system output with the evaluation corpus.

MOTS-CLÉS : simplification automatique discursive, chaînes de référence, corpus d'évaluation.

KEYWORDS: discourse simplification, reference chain, evaluation corpus.

1 Introduction

La simplification discursive est un domaine peu exploité dans le domaine de la simplification automatique. La plupart des corpus pour l'évaluation automatique sont constitués de phrases alignées extraites de corpus comparables ou des simplifications automatiques jugées par des humains en termes de simplicité, d'information préservée, de fluidité. Mais il n'y a pas de ressource comparable pour le français, à part le lexique ReSyf (Billami *et al.*, 2018) ou Cardon (2018) pour le domaine médical. Le corpus anglais le plus connu pour la simplification automatique lexicale est Specia *et al.* (2012) dans lequel listes de synonymes sans ambiguïté ont été classés par complexité par 5 anglophones non natifs. Pour la simplification syntaxique, il y a Newsela (Xu *et al.*, 2015) qui propose des adaptations graduées des articles de journal faites par leurs auteurs. Il y a peu de ressources annotées avec leur transformations, car le recours à des annotateurs non spécialisés (par exemple avec AMTurk) introduit des erreurs. Outre les erreurs d'alignement, les corpus extraits par des méthodes d'alignement (Narayan & Gardent, 2014), ne disposent pas d'une annotation des opérations de simplification.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 361–369.
HAL : hal-02784786.

L'évaluation des systèmes de simplification est une tâche difficile à cause de la rareté des corpus et des coûts importants nécessaires pour créer des corpus simplifiés manuellement. Les métriques d'évaluation doivent être corrélées avec le jugement de simplicité des humains. La plus adaptée est SAMSA (Sulem *et al.*, 2018), qui nécessite une pré-annotation sémantique des états, des événements et des participants , indisponible actuellement en français. Alternativement, BLEU compte le nombre de n-grammes communs entre le texte original et simplifié : par conséquent, plus le texte original est proche, meilleur est le score. SARI, une autre mesure utilisée dans des travaux de simplification, plus proche des jugements humains, évalue les simplifications par leur simplicité (Xu *et al.*, 2016) et prend en compte plusieurs références. En effet, plusieurs simplifications sont possibles et SARI prend en compte cet aspect. Nous adoptons cette mesure pour évaluer les références multiples.

Cependant, ces mesures ne prennent pas en compte des contraintes de discours. Les marques de cohésion textuelle, telles que les chaînes lexicales (Hirst & St-Onge, 1995) ou les chaînes de coréférence (Schnedecker, 1997), aident la compréhension (Hobbs, 1979; Charolles, 2006). En particulier, les enfants faibles lecteurs rencontrent des difficultés dans la résolution des inférences et des relations anaphoriques (Fayol, 2000; Ehrlich & Remond, 1997). Même si les marqueurs de cohésion sont fortement liés à la lisibilité et à la complexité du texte (Pitler & Nenkova, 2008), il y a peu de travaux qui les traitent en simplification. Certains systèmes les appliquent après la simplification syntaxique (Siddharthan, 2006; Canning, 2002), ou remplacent des pronoms anaphoriques par leurs antécédents (Quiniou & Daille, 2018). Si des corpus annotés pour évaluer la détection automatique des anaphores (Quiniou & Daille, 2018) ou de la coréference (Lattice *et al.*, 2019) sont disponibles, il n'y a pas encore de corpus simplifié en discours marquant ces transformations.

À cet égard, nous avons étudié les restrictions de cohésion liées aux chaînes de référence et leur applicabilité à la simplification automatique en français. Par conséquent, dans cet article, nous proposons un corpus original d'évaluation, disponible en français, pour aborder le niveau de discours parmi les différents niveaux de simplification. Le corpus présente plusieurs simplifications alternatives (annotées) pour chaque phrase, pour le niveau discursif mais aussi au niveau lexical et syntaxique.

Dans cet article, nous présentons notre projet et les choix que nous avons faits pour la simplification automatique qui prend en compte les liens de coréférence (section 2). La méthodologie de construction de corpus, la description du corpus et du guide d'annotation appliqué pour créer le corpus de simplification sont présentées dans la section 3. Ensuite, nous présentons l'évaluation du corpus et nous le comparons avec la sortie du système de simplification (section 4).

2 Le projet ALECTOR

Dans le cadre du projet [1] nous avons développé un système de simplification automatique de texte appliqué au niveau lexical, syntaxique et discursif pour proposer des contenus adaptés aux enfants dyslexiques ou faibles lecteurs. Vu le manque de ressources adaptées pour l'évaluation de la simplification automatique, nous avons construit un corpus d'évaluation. Nous adoptons une approche à base de règles pour notre système. D'une part, les corpus parallèles originaux et simplifiés sont de petite taille, ce qui rend difficile l'application des méthodes par apprentissage automatique. D'autre part, nous étudions l'influence de chaque type de transformation sur les capacités de compréhension et de lecture du public visé (faibles lecteurs et enfants dyslexiques). Dans le cadre de cet article, nous évaluons le module de simplification discursive (Wilkens *et al.*, 2020), qui réduit les inférences

1. Le projet Alector : https://alectorsite.wordpress.com.

nécessaires pour identifier liens anaphoriques et coréférentielles, en modifiant la structure des chaînes de référence (Schnedecker, 1997) [2], car ces liens sont des éléments de difficulté pour les enfants faibles lecteurs. L'architecture, ainsi que les étapes de traitement, sont représentées sur la figure 1. Le schéma regroupe les 4 modules pour la simplification : prétraitement, simplification discursive, simplification syntaxique et simplification lexicale et morphologique.

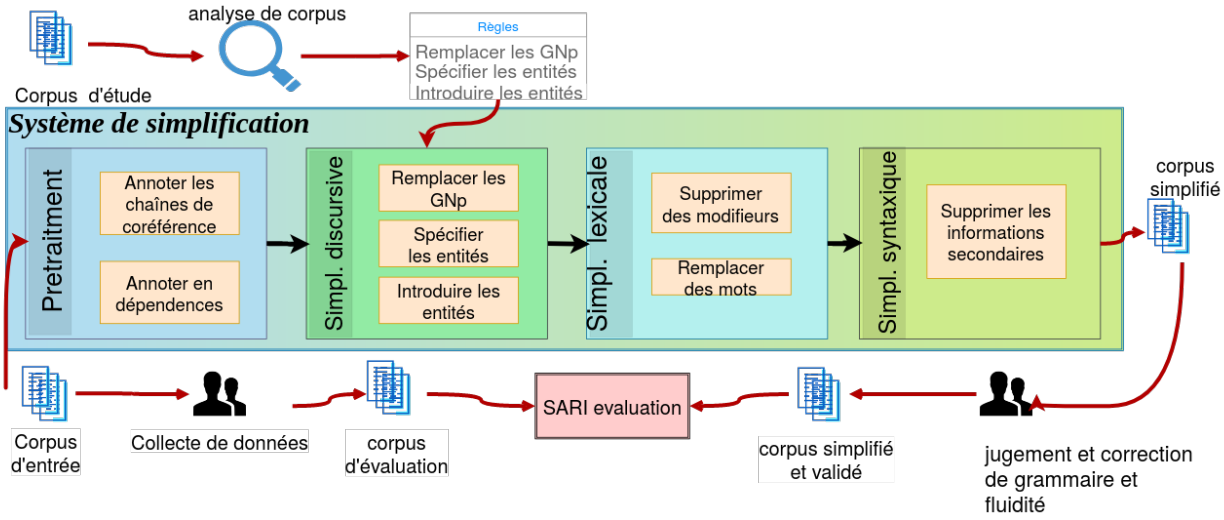


FIGURE 1 – La méthodologie et l'architecture du système

La première étape nécessite le **prétraitement** du texte brut afin d'appliquer les règles de simplification proposées. D'abord, le texte est pré-annoté en chaînes de coréférence en utilisant le système proposé par Wilkens *et al.* (2020) qui obtient le score CONLL de 85 %. Ce module délimite les expressions potentiellement référentielles (noms propres, groupes nominaux définis, déterminants démonstratifs, etc.) et identifie les liens de coréférence entre ces expressions. L'annotation en dépendances est réalisée avec l'analyseur Stanford NLP (Qi *et al.*, 2019), car ces informations sont nécessaires pour les transformations syntaxiques [3].

Ensuite, les données sont traitées par le module de **simplification discursive** qui modifie la structure des chaînes de référence annotées par le module de détection automatique de la coréférence. Ce module implémente plusieurs règles de simplification discursives : précision de l'entité (remplacement de pronoms par leur antécédent), spécification de l'entité (remplacement d'un déterminant par un autre de plus basse accessibilité pour réduire les inférences nécessaires pour trouver son antécédent), remplacement d'un groupe nominal possessif (GNp) par un référent plus explicite ou suppression de pronoms. Ces règles ont été proposées après l'analyse d'un corpus d'étude (voir la figure 1), composé de textes originaux et leurs versions simplifiées manuellement pour dyslexiques (Wilkens *et al.*, 2020), brièvement présentée dans la section 3.1.

Le troisième module, de **simplification syntaxique**, inspiré par les travaux de Siddharthan (2003); Brouwers *et al.* (2014), applique des règles permettant la suppression des informations secondaires (non-essentielles) : subordonnées relatives, participes passés et présents, compléments circonstanciels de temps, de lieu ou de manière. Les règles de simplifications syntaxiques sont décrites sur la base d'une étude de corpus de textes originaux et simplifiés pour dyslexiques (Gala *et al.*, 2020).

2. On parle de chaînes de référence à partir de 3 expressions qui indiquent le même référent (identité référentielle).
3. Ces outils sont parmi les plus performants disponibles, mais des erreurs sont inévitables.

Le dernier module de **simplification lexicale et morphologique**, vise la suppression des mots porteurs d'informations secondaires (adjectifs, adverbes) ou le remplacement d'un mot par un synonyme plus fréquent et plus simple, à l'aide du lexique gradué (Billami *et al.*, 2018). D'autres simplifications visent la forme morphologique : remplacer un mot par un autre de la même famille morphologique (remplacer un verbe par sa nominalisation) ou remplacer l'imparfait par la forme présent du verbe.

Pour évaluer ce système, en particulier le module de simplification discursive, nous avons créé un corpus d'évaluation avec les simplifications annotées, par plusieurs lecteurs experts. La sortie du système est d'abord jugée par les experts humains, en termes de fluidité et de grammaticalité. Nous comparons la sortie du système avec le corpus d'évaluation (comprenant de multiples simplifications), à l'aide de la mesure SARI. Le corpus d'étude, d'évaluation et le système de simplification sont disponibles à l'adresse `github.com/rswilkens/text-rewrite`.

3 Méthodologie de construction de corpus

Pour construire les règles de simplifications présentées dans la section précédente, nous avons constitué un corpus parallèle d'étude, constitué de contes pour enfants, simplifiés manuellement pour des enfants dyslexiques. D'abord, nous avons analysé les transformations, en particulier les transformations discursives (les propriétés des chaînes de référence). Afin de construire un corpus d'évaluation pour notre système de simplification (voir figure 1), nous avons mis en place un système de collecte de textes simplifiés. Sur la base de l'analyse du corpus d'étude, nous avons défini un guide de simplification discursive (qui résume les transformations identifiées entre les textes originaux et simplifiés). Les participants à la collecte de corpus sont des étudiants linguistes (niveau Master) qui ont reçu les mêmes textes originaux et ont suivi le guide de simplification. Nous avons mis en place la collecte de transformations à l'aide des rédacteurs ayants des bonnes connaissances linguistiques, afin d'obtenir des données plus fiables que celles qu'on pourra obtenir à l'aide d'Amazone Mechanical Turk. Les étudiants ont pris connaissance du guide et sont passés par une phase d'entraînement.

3.1 Corpus et annotation

Le corpus d'analyse est constitué de 5 textes originaux (1 969 mots) et leur version simplifiée (1 143 mots) provenant de methodolodys.ch, une association qui propose des textes aux enfants dyslexiques. Nous avons étudié ce corpus et nous avons identifié plusieurs règles de transformation de discours [4]. Pour ce faire, le corpus a été annoté manuellement en chaînes de référence, suivant le guide d'annotation du projet DEMOCRAT, à l'aide de la plateforme SACR (Oberle, 2018). Nous avons comparé plusieurs propriétés de chaînes de référence proposées par (Todirascu *et al.*, 2017). Plusieurs présentent une différence statistiquement significative ($p < 0,05$) : la longueur de la chaîne en nombre de maillons (de 10,37 dans les textes originaux à 10,86 dans les textes simplifiés), le nombre de chaînes (de 6,20 à 7,80), le coefficient de stabilité (Perret, 2000), indiquant la variation des expressions dans la chaîne (de 0,60 à 0,47). Nous avons aussi annoté toutes les catégories grammaticales des expressions référentielles. Les changements les plus frappants lors de l'analyse de la distribution statistique sont la proportion des pronoms personnels (de 36 % dans les textes originaux à 19 % dans les textes simplifiés), l'utilisation plus fréquente des GN définis (de 18 % à 36 %), et la réduction de la fréquence des déterminants possessifs (de 12 % à 10 %). L'inversion du pourcentage d'utilisation

4. Dans ce travail, nous présentons les résultats liés à la création des règles, pour plus de détails, voir Wilkens *et al.* (2020).

entre les pronoms personnels et les GN définis n'est pas un hasard, notre observation du corpus aligné montre une tendance à rendre l'entité explicite dans le corpus simplifié et à la répéter.

Sur la base de l'étude de corpus, nous avons identifié trois catégories de simplification de discours qui résument celles présentées par Wilkens *et al.* (2020) :

1. Marquer l'introduction ou la répétition des entités, permettant ainsi de limiter les inférences (par exemple, remplacer un déterminant plus accessible par un moins accessible (ce → le))
2. Expliciter les entités en remplaçant l'anaphore pronominale par le référent en cas de concurrence référentielle. En cas de répétition d'un pronom, on le remplace par un même antécédent en cascade. Par exemple, c'est le pronom *elle* qui peut avoir plusieurs antécédents possibles dans l'extrait : *«La deuxième amie dit que la soupe a une odeur agréable. Madame Dupont est en colère contre elle. Elle*$_{original}$*/Madame Dupont*$_{simplifié}$ *la trouve hypocrite.»*
3. Rendre les syntagmes nominales plus accessibles. On remplace les groupes nominaux possessifs par un référent explicite (nom propre ou une structure spécifique N1 de N2, où N2 est un référent identifié au préalable). Ainsi, *Son mari* est remplacé par *Le mari de Mme Dupont* : *«Mme Dupont fais une soupe dont l'odeur est insupportable. Son mari*$_{original}$*/Le mari de Mme Dupont*$_{simplifié}$ *n'a jamais avoué qu'il déteste sa soupe. »*

Ces règles ont été décrites dans un guide d'annotation appliqué pour la simplification et consulté par les participants à la constitution de corpus.

3.2 Collecte de données

A l'aide de la plateforme PsyToolkit [5], nous avons mis en place des questionnaires permettant de vérifier le temps de lecture et la compréhension des participants et de proposer des simplifications discursives suivant le guide d'annotation. Les participants renseignent l'âge, la langue maternelle et le niveau d'études. 25 étudiants ont répondu aux questionnaires, après avoir lu le guide. Nous avons écarté de nos données les réponses de 6 étudiants qui n'ont pas compris la tâche.

Pour créer un corpus d'évaluation conséquent, les étudiants ont modifié 5 extraits du texte original (55 phrases ou 894 mots, représentant environ 178 mots/texte) et ont proposé des alternatives, suivant le guide. Il est difficile de faire ces modifications discursives sans appliquer certaines transformations lexicales ou syntaxiques (en particulier pour la suppression de pronoms). Toutefois, les étudiants ont aussi proposé des transformations lexicales ou syntaxiques supplémentaires (découper les phrases, supprimer les modifieurs adjectivaux ou adverbiaux, ajouter des explications. . .). Nous avons gardé ces réponses dans notre corpus, pour comparer le résultat des transformations automatiques syntaxiques et lexicales. Nous avons aligné les textes originaux et les réponses des étudiants, à l'aide de Collatex [6] (Haentjens Dekker & Middell, 2011). Nous avons identifié des transformations agrammaticales et nous avons éliminé ces réponses, ainsi que celles qui étaient trop éloignées du texte original ou du guide. Nous avons corrigé certains erreurs (par exemple des fautes d'orthographe), en remplaçant la transformation par le segment de texte original. De plus, nous avons annoté les transformations effectuées par les étudiants selon la typologie présente dans le guide, ainsi que d'autres transformations syntaxiques et lexicales proposées par les étudiants, nécessaires pour respecter la fluidité et la grammaticalité des textes résultants. Ainsi, nous avons crée un corpus avec des multiples références (minimum 8 et maximum 16 variantes) étiqueté selon le guide de transformation, qui peut être utilisé

5. psytoolkit.org
6. collatex.net

pour l'évaluation du système de simplification. Dans le contexte de la simplification automatique, plusieurs règles sont applicables, d'où l'intérêt de construire un corpus avec multiple références.

3.3 Les propriétés du corpus d'évaluation

Sur un total de 1 386 transformations proposées par les étudiants, les plus nombreuses sont les simplifications discursives (686, représentant 45,89 %). Les transformations syntaxiques représentent 22,58 % (313) suivies par les simplifications lexicales (20,13 % – 279). Les transformations morphologiques représentent 5,92 % (82) et les transformations typographiques (ajout ou suppression de ponctuation) représentent 5,48 % (76). Les opérations morphologiques les plus appliquées sont le changement de mots les plus fréquents d'une famille morphologique (ex. retirer → retrait) et le changement du temps du verbe (ex. passé simple → passé composé ou présent). Les simplifications syntaxiques obtenues sont aussi celles qui suppriment des informations secondaires (phrases subordonnées relatives ou circonstancielles, etc.), suivis par les transformations qui privilégient les phrases courtes (découpage en plusieurs phrases lorsqu'il y a des conjonctions ou des signes de ponctuation). Les transformations qui privilégient l'ordre SVO sont également proposées : suppression des clivées, transformation de la diathèse passive en diathèse active, des phrases négatives en phrases positives. Nous avons calculé l'accord entre les annotateurs proposant des simplifications, qui reste faible (l'accord Krippendorff : 0,189). L'accord interannotateur est bas, mais il s'agit ici de plusieurs simplifications possibles qui s'appliquent sur la même phrase. Pour avoir un corpus d'évaluation varié, nécessaire pour SARI, on doit avoir plusieurs variantes de simplification. Pour cette raison l'accord reste bas.

	Texte 1	**Texte 2**	**Texte 3**	**Texte 4**	**Texte 5**	**TOTAL**	**TOTAL (%)**
Règle 1	11	30	18	14	20	93	14,62
Règle 2	231	36	13	68	69	417	65,57
Règle 3	24	4	1	32	3	64	10,06
Autre	2	1	15	34	10	62	9,75

TABLE 1 – La répartition des règles de transformation discursives dans le corpus simplifié

Parmi les règles de transformation discursive (présentées dans le tableau 1), la plus appliquée est la règle 2, qui remplace une expression référentielle par un référent bien identifié (65,57 % des transformations). La règle 1 qui remplace un déterminant par un déterminant moins accessible (*ces* → *les*) ou inversement (*un* → *le*) représente 14,62 %. Le remplacement des pronoms en série par des référents déjà identifiés dans le texte est la troisième règle appliquée (10,06 %). Les cas de suppressions de pronoms sont accompagnées de suppressions de subordonnées relatives ou de phrases. 9,75 % représentent des simplifications que nous n'avons pas prévu dans le guide, par exemple l'ajout d'un pronom pour expliciter un sujet zéro (qui pourrait être considéré comme une extension de la règle 2). Le remplacement des GNp a posé des problèmes aux étudiants : ils ont remplacé seulement le déterminant possessif (avec perte d'information sur le référent) ou le nom par son hyperonyme.

4 Comparaison avec le système automatique et discussions

Nous avons évalué manuellement les résultats de la simplification automatique appliquées sur les 55 phrases des textes originaux. 38,49 % des phrases contiennent au moins une erreur. Afin de

permettre une évaluation de la simplification automatique avec le minimum d'impact des outils de pré-traitement et des questions de grammaticalité, nous avons analysé les opérations qui ont généré les phrases agrammaticales. Cette analyse a montré que 18,42 % des erreurs du système sont dues au prétraitement (analyse syntaxique ou coréférence). 22,34 % des erreurs sont dues aux règles ne prenant pas en compte certaines structures polylexicales (ex. superlatif et expressions - *d'un côté vs. du côté vs. son côté*). 52,63 % sont dues à une mauvaise identification des référents complexes (ex. *le sol*) qui peuvent accepter l'indétermination selon le contexte.

Une fois le corpus d'évaluation disponible et les problèmes de grammaire résolus, nous évaluons automatiquement le système de simplification proposé. Pour ce faire, nous avons appliqué les métriques BLEU (Papineni *et al.*, 2002) et SARI. [7] SARI privilégie les modifications lexicales (insertions ou suppressions de mots). Les transformations discursives impliquent des modifications lexicales, le taux de suppression est plus important (suppression de subordonnées etc.). Le résultat élevé du BLEU est une conséquence du faible nombre de modifications apportées au texte (puisque les simplifications syntaxique et lexicale entraînent des modifications plus importantes du texte). De plus, nous sélectionnons aléatoirement les annotations d'une personne et nous les évaluons comme si elles étaient la sortie d'un système. Les résultats du système sont proches à la référence (table 2). En observant la différence de performance SARI entre la référence et le système, on peut observer une différence moyenne de 5, mais il y un texte dans lequel la différence entre la référence et le système est très importante (13 points de différence), pour les autres textes la différence moyenne entre les métriques est de 3,88. Cependant, il convient de noter que les performances du système ne sont pas loin de la valeur de référence.

Texte	BLEU	SARI
Texte généré par notre système	89,82	39,04
Texte de référence	91,98	44,72

TABLE 2 – Les scores BLEU et SARI obtenus par rapport à l'original

5 Conclusion et perspectives

Notre travail a permis le développement d'un corpus d'évaluation de simplifications discursives, crée à l'aide des experts humains. Cette ressource contient des références multiples par chaque phrase, nécessaires pour évaluer des systèmes de simplification et sera disponible en ligne, ainsi que le guide et le système de simplification. Notre corpus est composé de plus de 8 alternatives par phrase, représentant un corpus varié pour la simplification. Les opérations de simplification effectuées sont annotées et permettent de reconstituer les diverses étapes de simplification et d'identifier les erreurs possibles. Certaines erreurs (repérées manuellement dans la sortie du système) peuvent être évitées en faisant appel à une base de données d'expressions polylexicales. Etant donné que le guide sera disponible sur le site du projet ALECTOR, une nouvelle collecte de données permettra d'agrandir le corpus d'évaluation. SARI, corrélée avec les jugements humains, montre que le système automatique obtient un résultat proche des simplifications manuelles. Cette expérience peut être adaptée pour simplifier d'autres types et genres textuels (des textes juridiques ou des textes pour les apprenants) et construire des corpus plus grands.

7. Nous n'avons pas appliqué SAMSA, car il nécessite une pré-annotation sémantique qui n'est pas disponible en français.

Références

BILLAMI M. B., FRANÇOIS T. & GALA N. (2018). Resyf : a french lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2570–2581.

BROUWERS L., BERNHARD D., LIGOZAT A. & FRANÇOIS T. (2014). Syntactic sentence simplification for french. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR@EACL 2014, Gothenburg, Sweden, April 27, 2014*, p. 47–56. DOI : 10.3115/v1/W14-1206.

CANNING Y. M. (2002). *Syntactic simplification of Text*. Thèse de doctorat, University of Sunderland.

CARDON R. (2018). Approche lexicale de la simplification automatique de textes médicaux. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, p. 159.

CHAROLLES M. (2006). De la cohérence à la cohésion du discours. In F. CALAS, Éd., *Cohérence et discours*, p. 25–38. Presses de l'Université Paris Sorbonne.

EHRLICH M.-F. & REMOND M. (1997). Skilled and less skilled comprehenders : French children's processing of anaphoric devices in written texts. *British journal of developmental psychology*, **15**(3), 291–309.

FAYOL M. (2000). *Maîtriser la lecture : poursuivre l'apprentissage de la lecture de 8 à 11 ans*. Centre national de documentation pédagogique, Editions O. Jacob, Observatoire national de la lecture (France).

GALA N., TODIRASCU A., BERNHARD D., WILKENS R. & MEYER J.-P. (2020). Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés. In *Actes du Congrès Mondial de Linguistique Française*.

HAENTJENS DEKKER R. & MIDDELL G. (2011). Computer-supported collation with collatex. *Supporting Digital Humanities 2011*.

HIRST G. & ST-ONGE D. (1995). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet : An Electronic Lexical Database*, **305**.

HOBBS J. R. (1979). Coherence and coreference. *Cognitive Science*, **3**(1), 67–90. DOI : 10.1207/s15516709cog0301_4.

LATTICE, LILPA, ICAR & IHRIM (2019). Democrat. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.

NARAYAN S. & GARDENT C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1 : Long Papers*, p. 435–445. DOI : 10.3115/v1/p14-1041.

OBERLE B. (2018). SACR : A drag-and-drop based tool for coreference annotation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

PERRET M. (2000). Quelques remarques sur l'anaphore nominale aux XIV^{e} et XV^{e} siècles. *L'Information Grammaticale*, **87**. DOI : 10.3406/igram.2000.2740.

PITLER E. & NENKOVA A. (2008). Revisiting readability : A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 186–195.

QI P., DOZAT T., ZHANG Y. & MANNING C. D. (2019). Universal dependency parsing from scratch. *arXiv preprint arXiv :1901.10457*.

QUINIOU S. & DAILLE B. (2018). Towards a diagnosis of textual difficulties for children with dyslexia. In *11th International Conference on Language Resources and Evaluation (LREC)*.

SCHNEDECKER C. (1997). *Nom propre et chaînes de référence*. Recherches linguistiques. Klincksieck. HAL : hal-00808797.

SIDDHARTHAN A. (2003). Preserving discourse structure when simplifying text. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*.

SIDDHARTHAN A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, **4**(1), 77–109.

SPECIA L., JAUHAR S. K. & MIHALCEA R. (2012). SemEval-2012 task 1 : English lexical simplification. In **SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, p. 347–355, Montréal, Canada : Association for Computational Linguistics.

SULEM E., ABEND O. & RAPPOPORT A. (2018). Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 685–696, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1063.

TODIRASCU A., FRANÇOIS T., BERNHARD D., GALA N., LIGOZAT A.-L. & KHOBZI R. (2017). Chaînes de référence et lisibilité des textes : Le projet ALLuSIF. *Langue française*, **195**(3), 35–52. HAL : halshs-01665316.

WILKENS R., OBERLE B. & TODIRASCU A. (2020). Coreference-based text simplification. In *Workshop Tools and Resources to Empower People with REAding DIfficulties (READI), Conference on Language Resources and Evaluation (LREC)* : ELRA.

XU W., CALLISON-BURCH C. & NAPOLES C. (2015). Problems in current text simplification research : New data can help. *Transactions of the Association for Computational Linguistics*, **3**, 283–297. DOI : 10.1162/tacl_a_00139.

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415. DOI : 10.1162/tacl_a_00107.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La réécriture monolingue ou bilingue facilite-t-elle la compréhension ?

Yuming Zhai Gabriel Illouz Anne Vilnat
Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
{prénom.nom}@limsi.fr

RÉSUMÉ

La capacité en compréhension écrite est importante à développer pour les apprenants de langues étrangères. Cet article présente une expérience pour vérifier si les paraphrases fournies en contexte facilitent la compréhension des apprenants. Les paraphrases ont été extraites automatiquement d'un corpus parallèle bilingue. Suite à l'analyse des résultats, nous proposons des pistes d'enrichissement d'un outil conçu préalablement, pour automatiser la sélection de réécritures dans un futur travail, tout en caractérisant mieux différents types de réécritures.

ABSTRACT

Does monolingual or bilingual rewriting facilitate comprehension ?

Reading comprehension skills are important to develop for foreign language learners. This article presents an experiment to verify whether paraphrases provided in context make comprehension easier for learners. The paraphrases were automatically extracted from a bilingual parallel corpus. After analysing the results, we propose directions to enrich a previously designed tool, in order to automate the selection of rewritings in the future, while better characterizing different types of rewritings.

MOTS-CLÉS : extraction de paraphrase, compréhension écrite, TAL pour l'enseignement.

KEYWORDS: paraphrase extraction, reading comprehension, NLP for education.

1 Introduction

L'apprentissage des langues étrangères est important pour les étudiants, surtout quand ils veulent poursuivre des études dans un pays étranger. En effet, les études et l'intégration dans une société étrangère nécessitent un niveau de langue intermédiaire voire avancé. Au cours de l'apprentissage d'une langue étrangère, il est important de développer la capacité en compréhension écrite. Selon le niveau des apprenants, les difficultés en compréhension écrite évoluent. Les débutants peuvent confondre des mots homographes, sans pouvoir distinguer leur catégorie grammaticale et leur usage en contexte. Pour les apprenants dans un niveau intermédiaire ou avancé, Yilmaz Güngör (2015) a mené une enquête auprès de 23 étudiants de français langue étrangère (FLE) dans le cadre d'un cours de compréhension écrite. Cela montre leurs difficultés dans les catégories suivantes : lexical, grammatical, contextuel et socio-culturel. Nous listons ici quelques phénomènes concrets : des mots inconnus des domaines généraux ; des mots déjà vus mais qui apparaissent dans un contexte donné avec un sens inconnu ou figuré ; des termes spécifiques d'un certain domaine ; des nuances et des subtilités de langue ; différents registres de langue ; des expressions figées, des idiomes ; des structures syntaxiques complexes, des phrases longues avec une logique complexe, etc.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 370–378.
HAL : hal-02784787.

L'enquête menée par Dagilienė (2012) suggère que la traduction intégrée dans les activités quotidiennes des cours d'anglais s'avère utile pour le progrès des étudiants dans diverses compétences de langue. Selon Martinot (2012), un apprenant de langues étrangères a intérêt à prendre conscience des différentes procédures reformulatoires qu'il peut appliquer aux énoncés de la langue seconde. Compte tenu des rôles importants des traductions et des reformulations en acquisition de langues étrangères, notre problématique de recherche est d'appliquer des travaux en traitement automatique des langues (TAL) pour faciliter la compréhension écrite aux apprenants de FLE.

Dans notre publication précédente (Zhai *et al.*, 2019b), nous avons proposé la conception d'un outil dans ce but, via la proposition de réécritures des mots ou des suites de mots en contexte. Les recherches sur l'extraction de paraphrase et la reconnaissance des procédés de traduction constituent la base pour le développement de cet outil.

Dans cet article, nous décrivons une expérience utilisant des réécritures extraites automatiquement d'un corpus parallèle bilingue. Elle est menée avec la participation des apprenants chinois de FLE. Nous proposons des pistes d'enrichissement de l'outil conçu suite à l'analyse des résultats. Nous ciblons les apprenants chinois pour observer l'influence de la maîtrise d'une première langue étrangère (anglais) sur l'apprentissage d'une deuxième langue étrangère similaire (français), puisqu'une majorité des étudiants chinois apprend l'anglais depuis l'école primaire.

2 Travaux précédents

En TAL, la paraphrase est définie comme une forme alternative exprimant, dans une même langue, le même contenu sémantique, la même information ou la même idée que la forme originale (Barzilay & McKeown, 2001; Fujita, 2005; Callison-Burch, 2007; Bhagat, 2009; Madnani & Dorr, 2010; Bouamor, 2012). Dans le cadre de notre travail, nous nous focalisons sur la paraphrase linguistique au niveau sous-phrastique et nous utilisons le terme « réécriture » pour désigner les reformulations textuelles qui ne gardent pas strictement la même sémantique que le segment original.

Concernant l'extraction de paraphrase, Bannard & Callison-Burch (2005) ont proposé d'exploiter les techniques de traduction automatique pour extraire des paraphrases à partir de corpus parallèles bilingues. Leur hypothèse est que deux segments monolingues sont des paraphrases potentielles s'ils partagent des traductions communes (appelées aussi traductions "pivots") dans une autre langue. Actuellement, la plus grande ressource de paraphrases, PPDB (*ParaPhrase DataBase*) est construite en se basant sur cette méthode (Ganitkevitch *et al.*, 2013; Pavlick *et al.*, 2015b). En revanche, Pavlick *et al.* (2015a) ont révélé qu'il existe d'autres relations que l'équivalence stricte (paraphrase) dans PPDB.

Des traductions "pivots" non littérales dans des corpus parallèles bilingues peuvent influencer l'équivalence stricte entre les candidats de paraphrase extraits. Néanmoins, elles n'ont pas reçu assez d'attention pendant l'extraction des paraphrases dans des corpus parallèles bilingues. Afin d'étudier systématiquement les traductions non littérales, nous avons proposé une typologie de procédés de traduction adaptée à un corpus anglais-français de *TED Talks*, en nous fondant sur les théories développées en traduction (Vinay & Darbelnet, 1958; Chuquet & Paillard, 1989; Molina & Hurtado Albir, 2002). Nous avons aussi présenté une classification automatique des procédés de traduction en utilisant ce corpus annoté manuellement (Zhai *et al.*, 2019a).

3 Extraction automatique de paraphrases

Étant donné notre but : aider les apprenants en compréhension écrite, les paraphrases proposées en contexte doivent être contrôlées et de bonne qualité. Afin d'étudier en même temps l'utilité des traductions (réécritures bilingues), nous avons suivi l'approche d'extraction de paraphrases via la méthode par pivot dans des corpus parallèles bilingues (Bannard & Callison-Burch, 2005).

Nous avons adapté un système de traduction automatique statistique, développé par Gong *et al.* (2013). Le corpus utilisé contient celui de *TED Talks* [1] et celui de *Tatoeba* [2]. Ce corpus anglais-français contient 397k paires de phrases parallèles (sans doublon). L'alignement automatique de mots a été effectué par l'outil *FastAlign* avec les paramètres par défaut (Dyer *et al.*, 2013). Pendant cet alignement, la langue source est le français, et la langue cible est l'anglais, afin d'utiliser l'anglais comme une langue "pivot" pour générer des paraphrases françaises.

Étant donné une entrée (un mot ou une expression, *ex.* « de par le monde »), ce système peut réaliser ces trois tâches :

1) Afficher toutes les phrases françaises où l'entrée apparaît, ainsi que les phrases correspondantes en anglais. Ce concordancier bilingue peut faciliter l'examen manuel du corpus parallèle.
2) Extraire toutes les traductions anglaises possibles pour cette entrée. (*ex.* « *over the world, around the globe* », etc.)
3) À partir de chaque traduction en anglais, récupérer leur rétro-traduction en français, à savoir des candidats de paraphrase. (*ex.* « à travers le monde, aux quatre coins du monde » etc.)

Suite à cette extraction automatique, nous avons effectué une sélection manuelle des paraphrases et des traductions adéquates. Quand ce système ne fournissait pas ces informations, à cause de la taille limitée du corpus ou de la difficulté d'alignement automatique de mots, nous avons eu recours à la ressource en ligne *Linguee* [3]. Par exemple dans cette phrase : *Quand il n'écrivait pas, il racontait des récits* ***édifiants*** *sur les prouesses accomplies par des personnages dans sa tête.* Le mot « édifiant » n'existe pas dans notre corpus, mais grâce à *Linguee*, nous avons pu trouver ses traductions anglaises (dans ce contexte) « *amazing* » ou « *stunning* », ce qui peut être retraduit en français comme « étonnant », « incroyable », « époustouflant », etc. Dans certains cas, nous n'avons pu extraire que des traductions anglaises, parce qu'il n'existe pas de paraphrase (même avec l'aide de *Linguee*), par exemple les mots « pétrole », « charbon ».

Enfin, les paraphrases proposées ont été classées selon leur niveau de difficulté, avec l'aide de la ressource FLELex [4] (Francois *et al.*, 2014), qui est le premier lexique classé pour le FLE qui indique les fréquences normalisées de mots dans chaque niveau de A1 à C2 [5].

1. Nous avons utilisé deux corpus anglais-français de *TED Talks*, qui ont été publiés pour la campagne d'évaluation IWSLT en 2013 et 2016 (https://wit3.fbk.eu/).
2. *Tatoeba* est un site de collection de phrases multilingues et de leurs traductions (https://tatoeba.org/eng).
3. https://www.linguee.fr/
4. http://cental.uclouvain.be/flelex/, la version FLELex_CRF contient 2038 expressions multi-mots.
5. https://www.france-langue.fr/niveaux-de-francais/

4 Expériences avec les apprenants chinois

Pour étudier notre hypothèse que des réécritures monolingues ou bilingues facilitent la compréhension, nous avons mené deux expériences sur un test de compréhension écrite avec des étudiants chinois adultes. Les tests se sont déroulés sous forme de QCM (sur la plateforme *Moodle*). Nous avons contacté trois enseignantes de FLE de trois établissements chinois.[6] Nous avons échangé des idées sur la préparation des textes et des questions avec ces enseignantes. Après les tests, les étudiants ont dû répondre à un questionnaire sur leur profil (niveau d'étude en français et en anglais) et leur satisfaction. 15 étudiants de niveau intermédiaire (A2) et 11 étudiants de niveau avancé (B2) ont participé.[7]

Nous avons préparé un texte pour chaque niveau selon la grille pour l'auto-évaluation du CECR[8]. Le texte du niveau A2 (548 tokens) est issu d'un site de ressource pour l'éducation, et nous l'avons légèrement modifié. Celui du niveau B2 (300 tokens) a été écrit par nous à partir d'une liste de mots et de segments. Pour chaque groupe d'étudiants, le test est divisé en trois phases, avec pour chacune une version différente du texte qui a été présentée aux étudiants : 1) version originale (sans aucune information complémentaire) ; 2) les étudiants peuvent regarder les paraphrases des mots ou des expressions mises en gras (pas uniquement ceux qui sont pertinents pour répondre aux questions), qui apparaissent dans une fenêtre quand le curseur est passé dessus ; 3) en plus de ces paraphrases en français, les traductions anglaises sont ajoutées.

À chaque phase, les étudiants répondent au même ensemble de questions conçues par les auteurs. Notre hypothèse est la suivante : grâce aux informations complémentaires, les apprenants peuvent répondre correctement à de plus en plus de questions. Nous avons respecté les principes suivants pour la conception des questions et des options de réponses : 1) les questions concernent les mots ou les expressions potentiellement difficiles, mais pas la logique du texte ; 2) les paraphrases et les traductions anglaises fournies ne doivent pas contenir exactement les mêmes chaînes de caractères que celles présentes dans les réponses proposées ; 3) les options de réponse doivent être assez proches les unes des autres pour que les distracteurs soient bons. Les distracteurs ne doivent pas permettre de trouver la bonne réponse par déduction de façon évidente.

Nous avons limité le temps pour chaque test : 30 minutes pour le premier, et 15 minutes pour chacun des deux suivants. L'utilisation des dictionnaires a été interdite pendant les tests (peu importe le format). Nous mettons à disposition le matériel utilisé dans ces expériences[9] : les textes en trois versions en format HTML, les questions de compréhension, les listes des paraphrases et traductions proposées, le questionnaire et les réponses anonymisées.

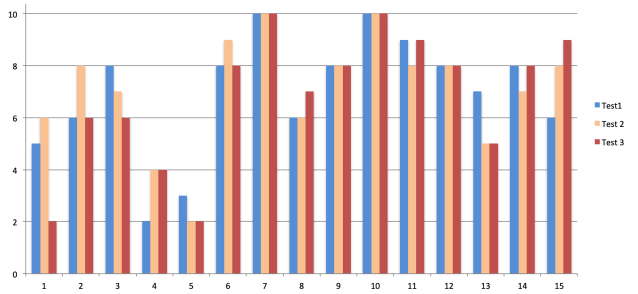
5 Analyse des résultats

Résultats du groupe A2 Les résultats par participant sont montrés dans la figure 1(a). Pour deux étudiants (N° 4, 15), des informations complémentaires semblent apporter une aide supplémentaire. La figure 1(b) montre le nombre d'étudiants ayant choisi la bonne réponse pour chaque question.

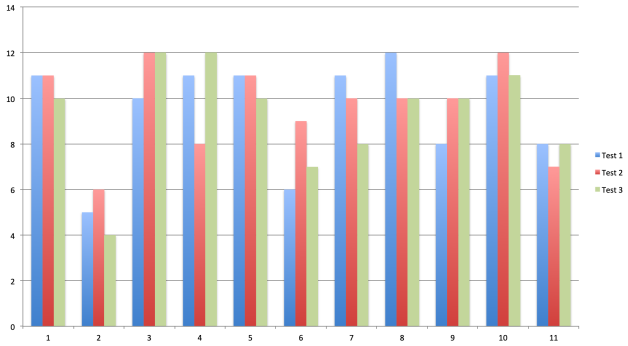
6. Université des langues étrangères de Dalian, Alliance Française de Dalian et Collège de technologie professionnelle de la ville de Ningbo.
7. 44 étudiants ont été contactés, en revanche, seuls 26 étudiants ont fait le test malgré des relances.
8. https://www.coe.int/fr/web/portfolio/self-assessment-grid
9. https://github.com/YumingZHAI/reecriture_comprehension

Nous voyons que la deuxième question est la plus difficile. Elle porte sur la phrase suivante et sur le mot « survivre » : *[...] il fera bientôt beaucoup trop chaud sur la Terre pour que certaines espèces puissent* ***survivre***.
La question est : *Quand il fera bientôt beaucoup plus chaud, certains animaux vont :*
a) *disparaître* ***b)*** *souffrir* ***c)*** *changer de lieu de vie*
Les paraphrases données sont « continuer à vivre » et « toujours en vie », et la traduction anglaise est « *survive* ». La plupart des étudiants ont choisi « changer de lieu de vie » or la bonne réponse est « disparaître ». Nous supposons que la difficulté ne concerne non seulement le mot « survivre », mais aussi l'expression « trop + adj. + pour que ... puisse ... ».

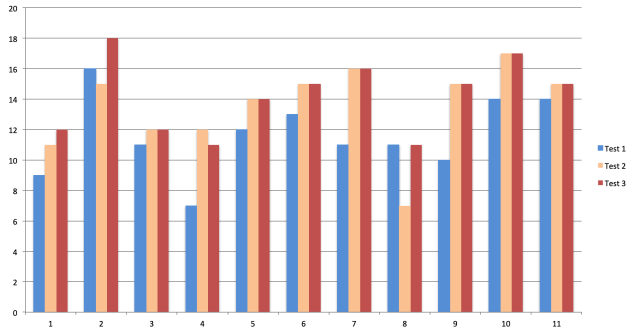
Nous avons analysé leurs réponses au questionnaire. En général, le niveau d'anglais n'est pas très élevé pour plusieurs étudiants, surtout ceux qui ont eu des moins bons résultats. Ceux qui ont le mieux répondu, ont dit qu'ils apprennent l'anglais et le français en les comparant, tandis que les autres étudiants ne veulent pas mélanger l'apprentissage des deux langues. La majorité des étudiants ont confirmé que la maîtrise de l'anglais favorise l'apprentissage du français, surtout pour le vocabulaire et la grammaire. Dix personnes (sur quinze au total) ont indiqué qu'un outil proposant des paraphrases en contexte pourrait les aider dans la lecture. Neuf personnes trouvent que la proposition des traductions anglaises est utile dans un tel outil, mais qu'il serait encore plus utile si une phrase avec un contexte était donnée en même temps.



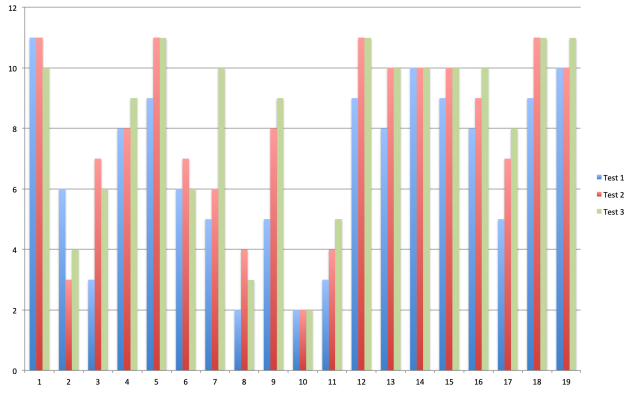
(a) Nombre de bonnes réponses par participant : 15 participants (axe X), 11 questions (axe Y)



(b) Nombre de personnes qui répondent correctement par question : 11 questions (axe X), 15 participants (axe Y)



(c) Nombre de bonnes réponses par participant : 11 participants (axe X), 19 questions (axe Y)



(d) Nombre de personnes qui répondent correctement par question : 19 questions (axe X), 11 participants (axe Y)

FIGURE 1: Résultats pour les apprenants de niveau A2 (en haut) et B2 (en bas)

Résultats du groupe B2 Les résultats par participant et par question sont montrés dans les figures 1(c) et 1(d). Pour les questions N° 1, 14, 19, la performance est déjà très élevée au test 1.

À part l'étudiant N° 8, tous les autres étudiants ont eu des meilleures notes avec des informations complémentaires par rapport au premier test. Pour trois étudiants (N° 1, 2, 8), l'ajout des traductions anglaises apporte de l'aide supplémentaire en plus des paraphrases. En général, l'évolution des performances du groupe B2 est plus homogène que celle du groupe A2.

La question la plus difficile est la question 10, peu importe la version du texte, seuls deux étudiants ont donné la bonne réponse. Elle concerne cette phrase : *On y voyait un ivrogne errer avec son calepin à la main, avec* ***dans son sillage*** *certains de ses compagnons.*
La question est : *Où voyait-on les compagnons de cet homme ?*
a) *autour de lui* ***b)*** *derrière lui* ***c)*** *devant lui* ***d)*** *ce n'est pas dit*
Cela montre que la paraphrase (« sur ses traces ») et les traductions (« *in the tracks/wake of* ») ne sont pas suffisamment claires pour aider la compréhension. Cette difficulté est aussi due aux principes de la conception de cette expérience, où les informations complémentaires ne doivent pas rendre la réponse évidente.

Dans la figure 1(d), nous constatons que pour la deuxième question, l'ajout d'informations nuit à la compréhension. Cette question porte sur la phrase suivante : *Ainsi, j'ai pu voir à la* ***lisière*** *d'une ville, dans un pâté de maisons isolé [...]*
La question est : *Où se déroule l'histoire ?*
a) *au centre-ville* ***b)*** *à la campagne* ***c)*** *aucune des deux réponses*
Nous avons fourni « frontière » et « bordure » comme paraphrase, et « *edge* » comme traduction. La bonne réponse est « aucune des deux », mais « à la campagne » a été choisie par la majorité des étudiants à la troisième phase du test. Seuls quatre étudiants ont saisi la nuance entre la « lisière d'une ville » et « la campagne », sachant que cette notion existe aussi en chinois.

Parmi ces 11 étudiants en niveau B2, trois sont en quatrième année de licence du département langue française en Chine. Les autres poursuivent leurs études en France, dont certains ont eu des bonnes notes en TOEIC.

Quatre étudiants (N° 2, 6, 7, 8) préfèrent ne pas mélanger l'anglais et le français pendant l'apprentissage, alors que les autres les apprennent par comparaison. Les étudiants confirment que la maîtrise de l'anglais favorise l'apprentissage du français, à part les similarités en vocabulaire et en grammaire, ils ont aussi mentionné l'étymologie et les expressions figées. Excepté deux étudiants qui préfèrent toujours utiliser un dictionnaire, les autres expriment leur envie d'avoir un outil qui propose des paraphrases en contexte pour élargir leur vocabulaire. L'ajout des traductions anglaises dans un tel outil est apprécié par huit étudiants, tandis que les trois autres préfèrent n'avoir affaire qu'à une seule langue étrangère à la fois.

6 Bilan et perspectives

Les résultats montrés dans la figure 1 valident notre hypothèse (et de façon plus évidente pour le groupe B2) : les paraphrases facilitent la compréhension des apprenants. Pourtant, l'hypothèse que les traductions anglaises fournissent encore de l'aide supplémentaire n'est pas encore complètement validée. Nous pensons qu'une expérience de plus large envergure, avec plus de textes et de participants, est nécessaire pour mieux étudier nos hypothèses. Nous testerons des procédures plus méthodologiquement robustes que celles utilisées dans notre première tentative. Par exemple, demander à différents groupes d'étudiants de faire le test avec différents niveaux d'informations et de

comparer les résultats doit être considéré.

Pour notre travail futur, des corpus parallèles plus volumineux seront exploités pour extraire des paraphrases. Dans cette expérience, l'extraction des paraphrases a été automatique mais le filtrage a été manuel. L'utilité des paraphrases pour faciliter la compréhension étant vérifiée, nous tenterons d'automatiser ce filtrage dans le futur. Pour cela, nous intégrerons la reconnaissance automatique des procédés de traduction (*ex.* classification entre traduction littérale et non littérale) (Zhai *et al.*, 2019a). Notre hypothèse est que pendant l'extraction de réécritures via la méthode par pivot, la reconnaissance des procédés de traduction nous permettra de caractériser les réécritures dans les trois cas suivants : 1) réécritures en équivalence sémantique stricte, à savoir des paraphrases, et la phrase après la substitution reste grammaticale ; 2) réécritures en relation d'implication (plus général ou plus spécifique), si elles existent ; 3) réécritures reliées avec le mot ou la suite de mots d'une certaine manière (par exemple, ils appartiennent au même champ sémantique ; il faut adapter la structure de la phrase pour la substitution). Cela nous permettra d'éviter les contresens, de classer les réécritures et d'étudier quel type de réécriture conviendrait mieux à chaque type d'apprenants.

Pour le développement de l'outil, nous suivrons la chaîne de traitement présentée dans la figure 2. Nous sommes aussi conscients des aspects suivants à prendre en compte pendant l'implémentation et l'évaluation d'un tel outil : l'adaptation du système face aux demandes différentes des apprenants et l'évaluation longitudinale de l'outil selon les différents niveaux des apprenants.

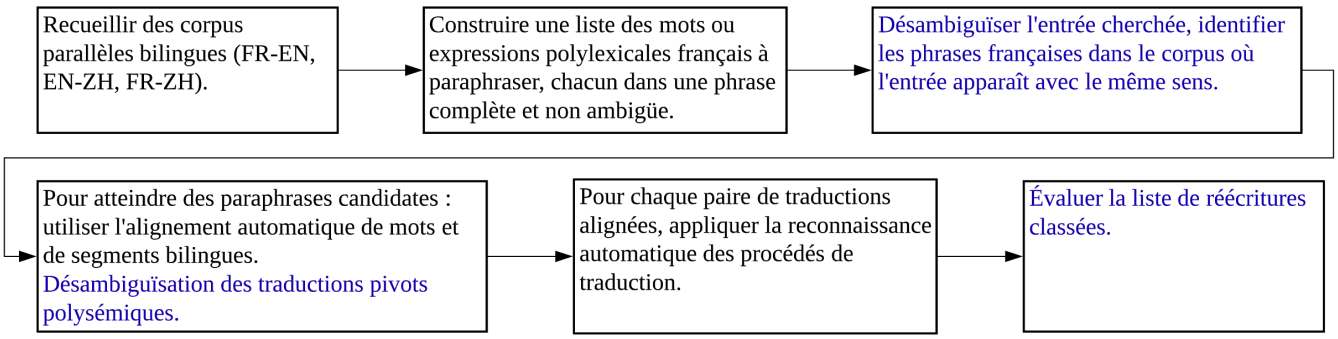


FIGURE 2: Chaîne de traitement pour le développement de l'outil (en noir : les blocs prêts ou en développement ; en bleu : les blocs à développer)

Remerciements

Nous remercions les relecteurs anonymes pour leurs remarques détaillées et constructives. Nous exprimons notre gratitude aux participants, et aux enseignantes qui ont aidé à solliciter les participants en Chine : Chunhong Yu, Jingjing Liang, Xinyan Li, Miao Wang et Xiaoya Xu. Nous remercions l'agence ANR pour son financement à travers le projet ALECTOR (ANR-16-CE28-0005).

Références

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 597–604 : Association for Computational Linguistics.

BARZILAY R. & MCKEOWN K. (2001). Extracting Paraphrases from a Parallel Corpus. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France.*, p. 50–57.

BHAGAT R. (2009). *Learning Paraphrases from Text*. Thèse de doctorat, Los Angeles, CA, USA. AAI3368694.

BOUAMOR H. (2012). *Etude de la paraphrase sous-phrastique en traitement automatique des langues. (A study of sub-sentential paraphrases in Natural Language Processing)*. Thèse de doctorat, University of Paris-Sud, Orsay, France.

CALLISON-BURCH C. (2007). *Paraphrasing and Translation*. Thèse de doctorat, University of Edinburgh, Edinburgh, Scotland.

CHUQUET H. & PAILLARD M. (1989). *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

DAGILIENĖ I. (2012). Translation as a learning method in English language teaching. *Kalbų studijos*, **21**, 124–129. DOI : 10.5755/j01.sal.0.21.1469.

DYER C., CHAHUNEAU V. & SMITH N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In L. VANDERWENDE, H. D. III & K. KIRCHHOFF, Éds., *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, p. 644–648 : The Association for Computational Linguistics.

FRANCOIS T., GALA N., WATRIN P. & FAIRON C. (2014). FLELex : a graded Lexical Resource for French Foreign Learners. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).

FUJITA A. (2005). *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Thèse de doctorat, Ph. D. thesis, Nara Institute of Science and Technology.

GANITKEVITCH J., VAN DURME B. & CALLISON-BURCH C. (2013). PPDB : The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 758–764.

GONG L., MAX A. & YVON F. (2013). Improving bilingual sub-sentential alignment by sampling-based transpotting. In *Proceedings of IWSLT*, Heidelberg, Germany.

MADNANI N. & DORR B. J. (2010). Generating phrasal and sentential paraphrases : A survey of data-driven methods. *Computational Linguistics*, **36**(3), 341–387. DOI : 10.1162/coli_a_00002.

MARTINOT C. (2012). De la reformulation en langue naturelle, vers son exploitation pédagogique en langue étrangère : pour une optimisation des stratégies d'apprentissage. *Synergies Pologne*, **9**, 63–76. http://gerflint.fr/Base/Pologne9/martinot.pdf.

MOLINA L. & HURTADO ALBIR A. (2002). Translation Techniques Revisited : A Dynamic and Functionalist Approach. *Meta*, **47**(4), 498–512. DOI : 10.7202/008033ar.

PAVLICK E., BOS J., NISSIM M., BELLER C., VAN DURME B. & CALLISON-BURCH C. (2015a). Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, volume 1, p. 1512–1522.

PAVLICK E., RASTOGI P., GANITKEVITCH J., DURME B. V. & CALLISON-BURCH C. (2015b). PPDB 2.0 : Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2 : Short Papers*, p. 425–430.

VINAY J.-P. & DARBELNET J. (1958). *Stylistique comparée du français et de l'anglais : méthode de traduction*. Bibliothèque de stylistique comparée. Didier.

YILMAZ GÜNGÖR Z. (2015). La compréhension des textes en français langue étrangère : quelles difficultés ? *Journal of International Social Research*, **8**(40).

ZHAI Y., ILLOUZ G. & VILNAT A. (2019a). Classification automatique des procédés de traduction. In *26th Conférence sur le Traitement Automatique des Langues Naturelles*, Toulouse, France. HAL : hal-02265644.

ZHAI Y., ILLOUZ G. & VILNAT A. (2019b). Conception d'un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère. In *9ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain*, p. 379–382, Paris, France. HAL : hal-02265646.