Journée commune AFIA/TLH - ATALA N°2 sur le thème "Technologies du Langage Humain et Santé"

Corinne Fredouille, Jose G. Moreno, Aurélie Névéol, Christophe Servan

AfIA

Association française
pour l'Intelligence Artificielle

Association
pour le Traitement
Automatique
des Langues

Journée N° 2

***T**echnologies du **L**angage **H**umain et **S**anté*

Collège TLH et ATALA

AFIA
Association française pour l'Intelligence Artificielle

ATALA
Association pour le Traitement Automatique des Langues

Préface

La journée «technologies du langage humain (TLH) et santé» organisée conjointement par le Collège TLH de l'AFIA et par l'Association pour le Traitement Automatique des Langues (ATALA [1]) avait pour objectif de proposer un panorama des recherches réalisées par les laboratoires francophones sur la thématique du **traitement du langage humain et de ses applications dans le domaine de la santé**. Ainsi, nous avons souhaité réunir des collègues issus d'instituts de recherche menant des travaux à l'intersection du traitement automatique des langues, de la recherche d'information, de la communication parlée, de l'informatique médicale et de la santé publique.

Cette journée fait suite à d'autres journées communes AFIA/ATALA organisées entre les années 2012 et 2018 sur les thèmes suivants :

- Intelligence artificielle et traitement automatique des langues se retrouvent (Mars 2012 - Paris INALCO) [2]
- Langue, apprentissage automatique et fouille de données (Mars 2014 - Paris INALCO)
- Représentation (Mars 2016 - Paris IHP/UPMC) [3]
- Apprentissage profond pour le Traitement Automatique des Langues (Juillet 2018 - Nancy) [4]

On peut également noter que la conférence TALN 2019 a été organisée à Toulouse dans le cadre de la plateforme AFIA (1-5 juillet 2019) [5], ce qui montre l'importance des liens entre les deux associations.

Suite à un appel à participation communiqué sur les listes de diffusion françaises des domaines de recherche en TLH et santé fin 2020, nous avons reçu 12 contributions, issues de laboratoires académiques en Belgique, Italie, France et Suisse. La diversité des recherches présentées ainsi que la qualité et la quantité des contributions reçues démontrent à la fois une dynamique importante des TLH dans la communauté francophone. Ainsi, nous avons pu articuler l'organisation de la journée autour de quatre sessions thématiques, d'une présentation invitée et d'une table ronde. Les sessions thématiques ont porté sur la traduction de la parole médicale, le traitement automatique de la langue clinique, le traitement automatique de la langue pathologique et la recherche d'information en santé.

Du fait de l'évolution de la situation sanitaire, la journée s'est déroulée en distanciel uniquement. Nous remercions l'université d'Avignon pour avoir fourni l'infrastructure BBB de la journée, ainsi que pour son soutien technique. Nous avons reçu plus de 120 demandes d'inscription à la journée. En terme de participation effective, cela s'est traduit par la présence de nombreux participants répartis sur la journée, avec un pic de présence à plus de 70 participants en fin de matinée. Par ailleurs, une dizaine de participants ont profité de l'environnement Gathertown [6] proposé pour offrir un moment de convivialité pendant la pause déjeuner. La journée a en outre bénéficié d'un public varié issu de laboratoires d'informatique, de linguistique et d'informatique médicale. Nous sommes ravis d'avoir pu réunir à cette occasion la communauté scientifique intéressée par le traitement de la langue biomédicale dans toute sa diversité. Nous remercions à ce titre l'ATALA, l'AFIA, le comité scientifique, les intervenants et le public.

Corinne FREDOUILLE, José G. MORENO, Aurélie NÉVÉOL et Christophe SERVAN

1. https://www.atala.org/
2. https://perso.limsi.fr/pz/ia-et-tal/
3. https://www.atala.org/content/tal-ia-3%C3%A8me-journ%C3%A9e-traitement-automatique-des-langues-et-intelligence-artificielle
4. https://afia.asso.fr/journee-traitement-automatique-des-langues-i-a-2018/
5. https://www.irit.fr/pfia2019/taln-recital/
6. https://gather.town/

AFIA
Association française pour l'Intelligence Artificielle

ATALA
Association pour le Traitement Automatique des Langues

Présentation de la journée

TLH et santé
4 février 2021, *en ligne*
journée commune AFIA/ATALA
https://www.irit.fr/TLH-Sante2021/

Présentation générale

Les Technologies du Langage Humain (TLH) proposent des méthodes permettant une communication homme-machine naturelle, pouvant s'étendre à une interaction homme-homme médiée. Les TLH permettent d'analyser, d'interpréter et de produire des actes du langage écrit, parlé ou signé, mais aussi d'interagir avec des données langagières. Ainsi, les TLH englobent traditionnellement le Traitement Automatique des Langues (TAL), la Communication Parlée (CP) et leurs applications les plus emblématiques comme la Recherche d'Information (RI) et la Traduction Automatique (TA).

Dans le cadre de sa mission d'animation, le Collège TLH de l'AFIA organise des journées dans le but de renforcer les interactions entre TAL, CP, RI et TA grâce à des méthodologies d'IA. La première journée du Collège, intitulée *TLH et Multimodalité*, a mis le focus sur les problématiques liées aux aspects multimodaux du langage. En effet, afin de collecter, gérer, fouiller, analyser les données hétérogènes et multi-sources, de nouvelles méthodes et outils doivent prendre en compte l'aspect multimodal à travers les langues écrites, parlées et/ou signées.

L'objectif de cette deuxième journée du Collège TLH est de réunir, en collaboration avec l'ATALA, chercheur·euse·s en Intelligence Artificielle et en Traitement Automatique des Langues travaillant sur une grande variété de thématiques liant Santé et Langage : analyse de la langue écrite, parlée et signée des patients en tant que signe pathologique mais également analyse de textes issus de la littérature, des dossiers électroniques patients ou des réseaux sociaux en tant que sources d'information sur la santé humaine. Dans ces contextes, les problématiques sont nombreuses et comportent des aspects méthodologiques de représentation des connaissances en santé exprimées en langue naturelle, des aspects méthodologiques de compréhension des informations, des aspects applicatifs en recherche clinique et santé publique. Par ailleurs, les techniques d'apprentissage profond ont récemment été mises à profit sur différentes problématiques de santé, mais peuvent également se heurter à des obstacles de disponibilité des données, de passage à l'échelle ou d'adaptation à des sous-domaines de la santé.

Cette journée a été l'occasion de confronter différentes approches, différents domaines d'application et de caractériser les éléments de contexte qui les rendent favorables. Le contexte sanitaire récent ayant fourni de nombreuses opportunités de recherche et corpus associés, la journée s'attachera à refléter les activités récentes de la communauté sur la thématique de la COVID-19.

Programme de la journée

9h15-9h30 : Accueil virtuel et ouverture de la journée (Corinne Fredouille et Aurélie Névéol)

9h30-9h40 : Mots des présidents des associations AFIA et ATALA (Domitile Lourdeaux, représentant Benoît LeBlanc, et Christophe Servan)

9h40 – 10h40 : Transcription et Traduction de la parole médicale Modérateur : Christophe Servan

9h40 – Pierrette Bouillon : Présentation du projet BabelDr
9h50 – Magali Norré, Pierrette Bouillon, Johanna Gerlach et Hervé Spechbach. BabelDr : un système de

traduction médicale avec des pictogrammes pour les patients allophones aux urgences et dans un secteur de dépistage COVID-19
10h08 – Bastien David, Pierrette Bouillon, et Hervé Spechbach. Vers une communication médicale adaptée aux personnes sourdes en période de confinement
10h26 – Lucía Ormaechea Grijalba, Pierrette Bouillon, Johanna Gerlach, Benjamin Lecouteux, Didier Schwab et Hervé Spechbach. Reconnaissance vocale du discours spontané pour le domaine médical

10h45-11h : Pause

11h – 12h20 : Traitement de la Langue Humaine dans la pratique clinique - Modératrice : Aurélie Névéol

11h00 – Ali Can Kocabiyikoglu, François Portet, Jean-Marc Babouchkine et Hervé Blanchon. Vers un système de dialogue oral pour la saisie de prescriptions médicales
11h20 – William Digan, Alice Rogier, David Baudoin, Bastien Rance et Antoine Neuraz. PyMedExt, un couteau suisse pour le traitement des textes médicaux
11h40 – Antoine Neuraz, Ivan Lerner, William Digan, Nicolas Garcelon, Rosy Tsopra, Alice Rogier, David Baudoin, Anita Burgun et Bastien Rance. Traitement de la langue naturelle pour une réponse rapide aux maladies émergentes: COVID-19
12h00 – Nesrine Bannour, Aurélie Neveol, Xavier Tannier et Bastien Rance. Traitement Automatique de la Langue et Intégration de Données pour les Réunions de Concertations Pluridisciplinaires en Oncologie

12h30-13h30 : Pause Déjeuner - Gathertown

13h30 – 14h30 : Présentation invitée – Pierre Zweigenbaum - Traitement automatique des langues pour la santé : travaux récents au LISN - Modérateur : Mathieu Roche

14h40 – 15h40 : Traitement de la Langue Humaine appliqué aux pathologies langagières - Modératrice : Corinne Fredouille

14h40 – Chuyuan Li, Maxime Amblard, Chloé Braud, Caroline Demily, Nicolas Franck et Michel Musiol. Investigation des marqueurs langagiers non-lexicaux et spécifiques des personnes souffrant de schizophrénie dans des conversations spontanées
15h00 – Rachid Riad, Lucas Gautheron, Emmanuel Dupoux, Anne-Catherine Bachoud-Lévi et Alejandrina Cristia. Measurements of turn-taking and linguistic behaviors in clinical settings
15h20 – Frédérique Brin-Henry. Exploration de la temporalité dans la désignation des pathologies du langage en orthophonie : aspects cliniques et termino-ontologiques

15h45 – 16h25 : Traitement de la Langue Humaine appliqué à l'information de santé Modérateur : José Moreno

15h45 – Elise Bigeard, Aman Sinha, Marianne Clausel et Mathieu Constant. Fouille de la littérature médicale à l'aide de graphes
16h05 – Silvia Calvi et Klara Dankova. La communication en santé publique au temps du Covid-19 dans les contextes français, québécois et tunisien

16h30 – 18h00 : Table ronde - Modérateurs : Aurélie Névéol et Christophe Servan

La table ronde a vu la participation de différents intervenants permettant d'avoir un regard croisé pluridisciplinaire sur la problématique spécifique des TLH sur les données de santé. Cette table ronde a

mis en relief les nouveaux défis scientifiques en considérant les dimensions juridiques et éthiques associées. En effet, la prise en compte de ces problématiques constitue un verrou à prendre en compte dans les travaux de recherche en TLH sur lesquels les intervenants ci-dessous ont pu échanger. Les points saillants évoqués ont été : l'accès aux données de santé textuelles dans des langues autres que l'anglais, le biais des données collectées et des modèles qui en découlent, le coût des modèles de langue modernes.

- **Sandra Bringay** a obtenu un doctorat en informatique de l'Université de Picardie Jules Verne. Maître de conférences à l'Université Paul Valery Montpellier de 2007 à 2016, elle est actuellement professeur dans cet établissement depuis septembre 2016. Elle réalise ses recherches dans le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM-CNRS-UM). Ses intérêts de recherche actuels incluent la science des données et l'intelligence artificielle appliquées au domaine de la santé.
- **Eric Brunet-Gouet** est praticien hospitalier en psychiatrie adulte au Centre Hospitalier de Versailles. Il a conduit des recherches essentiellement en neurosciences cognitives portant sur les troubles de la cognition sociale dans la schizophrénie et leurs répercussions fonctionnelles : troubles de la communication, difficultés dans la vie courante. Dans le cadre du réseau de centres experts FondaMental, il a travaillé à l'exploitation des bases de données de recherche et aussi cliniques portant sur des cohortes de patients souffrant de troubles bipolaires et schizophréniques. Ce travail, dans le cadre de la supervision d'une thèse en santé publique, a conduit à utiliser des modèles statistiques appropriés à des nombre de sujets plus élevés (équations structurales), et donc à s'intéresser aux possibilités d'accroitre massivement le recueil d'informations en employant des techniques innovantes comme le traitement du langage naturel appliqué aux données cliniques. Rattaché au Centre Epidémiologie et Santé des Populations (Inserm CESP) et à l'équipe PsyDev (Pr Speranza & Passerieux), il travaille notamment à l'application de modèles pré-entrainés type BERT à des corpus de dossiers cliniques ou de textes, dans la perspective de concevoir des modèles applicables et pertinents en recherche clinique.
- **Hélène Guimiot-Bréaud** est docteur en droit. Elle a exercé des fonctions de juriste dans le secteur de l'industrie pharmaceutique et au sein d'un Centre hospitalier universitaire. Elle a rejoint la CNIL en 2015 et est chef du service de la santé depuis février 2018.
- **Antoine Neuraz** est praticien hospitalier universitaire à l'Université de Paris et à l'hôpital Necker – Enfants Malades depuis septembre 2020 (après 4 ans comme assistant hospitalier universitaire). Médecin de santé publique depuis 2015, il a obtenu un doctorat en informatique médicale de l'Université de Paris en 2020 et est membre de l'équipe de recherche INSERM « Information Sciences to support Personalized Medicine » (Pr. A. Burgun). Ses travaux portent sur la réutilisation des données de soins pour la recherche et le machine learning, notamment via l'utilisation du traitement automatique de la langue pour faciliter la recherche d'information et l'accès à l'information dans les dossiers patients informatisés. Un autre axe de travail porte sur les méthodes de générations d'hypothèse (fouille de données) via des études d'association à large spectre dans les dossiers patients.
- **Pierre Zweigenbaum** est directeur de recherche au CNRS depuis 2006 au LIMSI, maintenant Laboratoire Interdisciplinaire des Sciences du Numérique (LISN). Il a passé vingt ans à l'Assistance publique – Hôpitaux de Paris et à l'Inserm, et a été dix ans Professeur associé à l'Inalco. Il mène des recherches sur l'extraction d'information à partir de textes avec des applications dans le domaine médical. Il est auteur ou co-auteur de méthodes et d'outils pour la détection de divers types d'entités médicales, l'expansion d'abréviations, la résolution de coréférences, la détection de relations, et la normalisation d'entités, notamment pour le codage CIM-10. Il a aussi conçu des méthodes pour l'acquisition automatique de connaissances linguistiques à partir de corpus et de thésaurus, pour aider à étendre des lexiques et des terminologies, y compris bilingues. Ses travaux lui ont valu la reconnaissance de l'American College of Medical Informatics (Fellow ACMI, 2014) et de l'International Academy of Health Sciences Informatics (Fellow IAHSI, 2019).

Co-organisation et comité scientifique

La journée est co-organisée par Corinne Fredouille, Aurélie Névéol et José Moreno du collège TLH de l'AFIA et Christophe Servan de l'ATALA et soutenue par le comité scientifique suivant :

Florian Boudin (L2SN, Université de Nantes)
Davide Buscaldi (LIPN, Université Paris XIII)
Gaël Dias (GREYC, Université de Caen Normandie)
Corinne Fredouille (LIA, Avignon Université)
Lorraine Goeuriot (LIG, Université Grenoble-Alpes)
Natalia Grabar (Université de Lille, CNRS, STL)
Cyril Grouin (Université Paris Saclay, CNRS, LISN)
José Moreno (IRIT, Université Paul Sabatier)
Aurélie Névéol (Université Paris Saclay, CNRS, LISN)
Damien Nouvel (Inalco, ERTIM)
Yannick Parmentier (LORIA, Université de Lorraine)
François Portet (LIG, Institut Polytechnique de Grenoble)
Mathieu Roche (TETIS, CIRAD)
Didier Schwab (LIG, Université Grenoble-Alpes)
Christophe Servan (QWANT)
Serena Villata (I3S, Université Côte d'Azur)

Résumé des présentations

BabelDr : un système de traduction médicale avec des pictogrammes pour les patients allophones aux urgences et dans un secteur de dépistage COVID-19

Magali Norré[1, 2] Pierrette Bouillon[2] Johanna Gerlach[2] Hervé Spechbach[3]
(1) CENTAL/ILC, Université catholique de Louvain, Belgique
(2) FTI/TIM, Université de Genève, Suisse
(3) Hôpitaux Universitaires de Genève (HUG), Suisse
magali.norre@uclouvain.be, pierrette.bouillon@unige.ch,
johanna.gerlach@unige.ch, herve.spechbach@hcuge.ch

RESUME

Contexte – BabelDr (https://babeldr.unige.ch/, Spechbach *et al.*, 2019), est un système de traduction de phrases fixes spécialisé pour le triage, le diagnostic dans un contexte d'urgences, qui permet au médecin de parler librement. Le système lie le résultat de la reconnaissance vocale à la phrase pré-traduite la plus proche avec des techniques neuronales (Mutal *et al.*, 2019). Il contient actuellement 10 000 phrases reliées à des millions de phrases, organisées par domaine. Depuis plus de deux ans, il est utilisé aux urgences des HUG pour le triage des patients avec six langues différentes, dont la langue des signes. Il répond à trois critères principaux : sécurité des données, fiabilité de la traduction et portabilité à de nouvelles langues (peu dotées), phrases ou domaines. Un des objectifs du système est de collecter des données et comparer différentes méthodes de communication.

Objectif – Nous avons étudié l'interaction de patients avec le système en réalisant des tests utilisateurs avec douze arabophones afin de comparer l'interface unidirectionnelle qui oblige le patient à répondre avec des gestes et l'interface bidirectionnelle de BabelDr qui permet aux patients de répondre en choisissant des pictogrammes Arasaac et Sclera (Norré *et al.*, 2020). Nous voulions savoir si les patients pouvaient répondre avec les deux types d'interfaces et s'ils étaient satisfaits de leurs interactions, en particulier, en utilisant les pictogrammes. Les participants n'avaient aucune connaissance de BabelDr. Ils vivaient tous depuis moins de neuf ans en Europe : onze vivaient en Belgique, un en France. Ils étaient tous de langue maternelle arabe et avaient différents niveaux de français.

Méthodologie – Ces tests se sont déroulés à distance par visioconférence (Zoom). L'expérience était divisée en trois parties au cours desquelles l'expérimentateur posait oralement des questions en français avec BabelDr. Les phrases reconnues étaient ensuite traduites et synthétisées en arabe. Les deux premières parties ont servi à comparer l'interaction avec les deux interfaces sur une série de questions de diagnostic ouvertes et fermées. La troisième consistait à reproduire une anamnèse du COVID-19 sur la base du questionnaire des HUG. Les patients devaient simuler qu'ils avaient ou non les symptômes du COVID-19 en répondant à des questions ouvertes et fermées (par exemple : « à quelle date ont commencé les symptômes ? », « avez-vous été en contact avec une personne confirmée positive au coronavirus dans les quatorze derniers jours ? », etc.). Les participants ont ensuite répondu à des questionnaires de satisfaction (dont le SUS), traduits en arabe pour chacune des deux interfaces.

Résultats – Les enregistrements ont montré que les patients n'ont pas su répondre à toutes les questions ouvertes avec l'interface unidirectionnelle même s'ils étaient tous globalement très satisfaits des deux interfaces. Au total, dix participants sur onze ont répondu avoir préféré utiliser l'interface avec les pictogrammes, certains ont rapporté que les images leur avaient permis de mieux répondre aux questions car ils ont par exemple pu montrer plus facilement où se situait la douleur. Les participants ont choisi divers pictogrammes en fonction du type de questions (par exemple : le pictogramme 'gorge', mais aussi 'poumons' pour la localisation de la douleur en cas de COVID-19). Les résultats montrent qu'utiliser des pictogrammes pour répondre aux questions ouvertes du médecin peut être une solution pour clarifier un problème médical rencontré par des patients pour les urgences ou dans de nouvelles situations comme celle du COVID-19. D'autres études sur les pictogrammes seront menées dans le cadre du projet FNS-ANR PROPICTO (PRojection du langage Oral vers des unités PICTOgraphiques).[1]

MOTS-CLES : traduction médicale, pictogramme, anamnèse, test utilisateur, COVID-19.

Références

MUTAL, J. D., BOUILLON, P., GERLACH, J., ESTRELLA, P. & SPECHBACH, H. (2019). Monolingual backtranslation in a medical speech translation system for diagnostic interviews - a NMT approach. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, p. 196–203.

NORRE, M., BOUILLON, P., GERLACH, J. & SPECHBACH, H. (2020). Évaluation de la compréhension de pictogrammes Arasaac et Sclera pour améliorer l'accessibilité du système de traduction médicale BabelDr. In *Actes de la 11e conférence de l'Institut Fédératif de Recherche sur les Aides Techniques pour personnes Handicapées (IFRATH). Technologies pour l'autonomie et l'inclusion*, p. 177–182.

SPECHBACH, H., GERLACH, J., KARKER, S. M., TSOURAKIS, N., COMBESCURE, C. & BOUILLON, P. (2019). A Speech-Enabled Fixed-Phrase Translator for Emergency Settings: Crossover Study. In *JMIR Medical Informatics*, 7(2), e13167.

[1] Ce travail a partiellement bénéficié d'un financement du Fond National Suisse (No. 197864) et de l'Agence Nationale de la Recherche, via le projet PROPICTO (ANR-20-CE93-0005).

AfIA Association française pour l'Intelligence Artificielle
ATALA Association pour le Traitement Automatique des Langues

Vers une communication médicale adaptée aux personnes sourdes. Le projet BabelDr et les personnages virtuels en langue des signes française de Suisse romande.

Bastien David[1], Pierrette Bouillon[1], Hervé Spechbach[2]
(1) TIM, boulevard du Pont-d'Arve 40, 1211 Genève 4, Suisse
(2) Hôpitaux universitaires de Genève, Suisse

bastien.david@unige.ch, pierrette.bouillon@unige.ch, herve.spechbach@hcuge.ch

RESUME

Nous présentons ici nos recherches dans le développement de technologies de communication accessible en milieu médical, plus précisément de personnages virtuels en langue des signes pour un service d'urgence ambulatoire. L'étude que nous menons se situe dans le cadre du projet d'assistant de communication multilingue BabelDr, qui est le fruit d'une collaboration entre le service ambulatoire des Hôpitaux universitaires de Genève (HUG) et la Faculté de Traduction et d'Interprétation (FTI) de l'Université de Genève (Strasly et al., 2018).

Le manque de communication efficace et fiable entre médecin et patient sourd crée une barrière encore difficilement franchissable. Celle-ci est encore accrue actuellement avec la crise sanitaire par le manque d'interprètes professionnel et de sensibilisation du personnel médical à la culture sourde, ainsi que l'emploi généralisé du masque chirurgical. Notre constat est qu'il est nécessaire de transmettre de manière efficace et sécurisée le discours médical, essentiel aux personnes sourdes, de l'accueil au diagnostic. La solution développée est la mise en place d'un *video player* en langue des signes, intégré au système de traduction de phrases fixes BabelDr. Deux corpus sont en cours de réalisation : le pré-enregistrement d'humains signeurs et la génération de personnages signeurs virtuels, ou avatars, à partir des vidéos humaines.

L'emploi d'interprètes ou de signeurs natifs est la garantie d'une grande précision et fluidité des phrases canoniques. Cependant, le contenu d'une vidéo ne peut être modifié après sa production. Les scénarios envisagés ne peuvent être dynamiques. Un certain coût humain (interprète, intermédiaire, cameraman, monteur) et logistique (salle d'enregistrement, camera et logiciels) sont à prendre en compte. En revanche, les avatars signeurs sont à la fois flexibles dans leur forme et anonyme. Leur coût de production est faible. Leur modification en post-production est facile.

En ce qui concerne la technologie utilisée, nous avons porté notre attention sur le potentiel du G SiGML pour la génération d'un personnage JASigning (Glauert & Elliott, 2011). Les expressions manuelles sont construites à partir du langage HamNoSys (HNS), tandis que les expressions non-manuelles proviennent de constructions SiGML issues du projet ViSiCast. Les expressions labiales sont décrites en phonétique SAMPA. A partir de notre glossaire HNS de 540 gloses et d'une grammaire, qui permet leur composition de manière productive et l'alignement avec les informations non-manuelles, le corpus généré se compose actuellement de plusieurs millions de constructions signées en langue des signes française de Suisse romande (LSF-SR).

Si des projets d'avatars signeurs ont déjà été testés dans le secteur ferroviaire, postal et hôtelier, peu d'initiatives se sont concentrées sur l'utilisation de ces outils automatiques dans le milieu hospitalier (Chiriac, Stoicu-Tivadar & Podoleanu, 2016).

A l'avenir, au moyen de nos corpus en LSF-SR, il sera possible, avec des utilisateurs sourds d'évaluer les propositions signées avec les humains et les avatars, d'émettre leurs critiques sur l'accessibilité des services d'urgences médicales et d'améliorer leur satisfaction en situation réelle. Les corpus pourront ensuite être disponibles dans l'outil BabelDr, intégrer d'autres services/domaines médicaux et satisfaire un besoin jusqu'à maintenant insatisfait.

MOTS-CLÉS : Langue des signes – Avatar animé – Système de communication médicale – G-SiGML

Références

CHIRIAC, I.A., STOICU-TIVADAR, L. & PODOLEANU, E. (2016). Romanian Sign Language Oral Health Corpus in Video and Animated Avatar Technology. In V. Balas, C. L. Jain, B. Kovačević, Éds. *Soft Computing Applications (SOFA 2014). Advances in Intelligent Systems and Computing*, vol. 356, p. 279-293. Springer, Cham. DOI: 10.1007/978-3-319-18296-4_24.

GLAUERT, J. & ELLIOTT, R. (2011). Extending the SiGML Notation: A Progress Report. In *Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Dundee, Ecosse.

STRASLY, I., SEBAÏ, T., RIGOT, E., MARTI, V., GONZALEZ, J.-M., GERLACH, J., SPECHBACH, H. & BOUILLON, P. (2018). Le projet BabelDr : rendre les informations médicales accessibles en Langue des Signes de Suisse Romande (LSF-SR). In P. Bouillon, S. Rodríguez Vázquez & I. Strasly, Éds. *Proceedings of the 2nd Swiss Conference on Barrier-free Communication: Accessibility in educational settings (BFC 2018), Geneva (Switzerland) - 9-10 November 2018*, p. 92-96.

AfIA Association française pour l'Intelligence Artificielle
ATALA Association pour le Traitement Automatique des Langues

Reconnaissance vocale du discours spontané pour le domaine médical

Lucía Ormaechea Grijalba[1, 2] Pierrette Bouillon[1] Johanna Gerlach[1]
Benjamin Lecouteux[2] Didier Schwab[2] Hervé Spechbach[3]

(1) FTI/TIM, Université de Genève, Genève, Suisse
(2) Laboratoire d'Informatique de Grenoble, Saint-Martin-d'Hères, France
(3) Hôpitaux Universitaires de Genève, Genève, Suisse
{lucia.ormaecheagrijalba, pierrette.bouillon, johanna.gerlach}@unige.ch
{benjamin.lecouteux, didier.schwab}@univ-grenoble-alpes.fr
herve.spechbach@hcuge.ch

RESUME

Contexte – Dans le domaine médical, et plus particulièrement dans les services d'urgence, les barrières linguistiques constituent un problème important. Une mauvaise communication entre un médecin et un patient qui ne partagent aucune langue peut mettre en danger la santé et la sécurité du patient (Hacker *et* al., 2015). Pour faire face à cette situation, nous avons développé, pour le triage aux Hôpitaux Universitaires de Genève (HUG), le système de traduction BabelDr (FR vers ES, AR, TI, FA, PRS et LSF). Celui-ci repose sur un ensemble fini de phrases pré-traduites (mémoire de traduction, MT, environ 10 000 phrases canoniques), mais permet au spécialiste de s'exprimer oralement, pour en améliorer l'ergonomie (Boujon *et* al., 2018). Le système lie le résultat de la reconnaissance vocale avec l'une des phrases pré-traduites au moyen de techniques neuronales (Mutal *et* al., 2020).

Objectifs – Actuellement, BabelDr repose sur le système de reconnaissance commercial NTE (*Nuance Transcription Engine*), spécialisé avec des données artificielles écrites. Celles-ci sont générées à partir de la grammaire BabelDr (*Synchronized Context-free Grammar, SCFG*) (Rayner *et* al., 2018), qui met en correspondance les phrases canoniques de la MT avec des variations syntaxiques et sémantiques possibles (p. ex. *avez-vous de la fièvre ?* avec *êtes-vous fiévreux ?*, *est-ce que vous avez de la température ?*, etc.). Le but principal de cette étude est de voir quelle performance peut être atteinte pour ce type de discours oral spécialisé spontané, en utilisant la boîte à outils *open source* Kaldi (Povey *et* al., 2011).

Méthodologie – Pour développer le système de reconnaissance de la parole, nous avons privilégié une approche temps réel (basée sur la version online de Kaldi). Par ailleurs, nous avons eu recours à des modèles acoustiques hybrides HMM-DNN et une modélisation linguistique appuyée sur un modèle de langage générique interpolé avec une grammaire adaptée au discours médical et se basant sur des données générées avec la grammaire SCFG BabelDr. À l'aide d'un corpus oral de questions d'anamnèses et d'instructions médicales collecté aux HUG avec des médecins via BabelDr, nous avons évalué la performance du prototype Kaldi. Cela nous a permis, en outre, de mettre en regard les résultats obtenus avec les deux technologies.

Résultats – À la lumière des résultats globaux observés, une amélioration significative du système basé sur Kaldi est observée par rapport à NTE en termes de WER (14,37% vs 22,93% pour le corpus de test, cf. Table 1). Compte tenu du contexte spécialisé visé, où une erreur de traduction n'est pas acceptable, il est nécessaire d'avoir recours à des évaluations complémentaires, le WER

étant une mesure d'évaluation globale. Les résultats en termes de SemER (*Semantic Error Rate*, pourcentage de phrases orales incorrectement liées à la phrase canonique de la MT) montrent ainsi, sur le test, une meilleure précision sémantique des transcriptions effectuées par le système Kaldi. Les deux systèmes atteignent des taux d'erreur similaires sur le corpus de dev, mais nous observons une grande différence entre les résultats de Kaldi et ceux de NTE dans le corpus de test (14,73% vs 35,52%).

		WER(%)		SemER(%)[1]	
Corpus	***Phrases***	**Nuance**	**Kaldi**	**Nuance**	**Kaldi**
Dev	**2864**	20,99	14,15	17,59	21,11
Test	**2708**	22,93	14,37	35,52	14,73

TABLE 1 : Résultats en termes de Word Error Rate (WER) et Semantic Error Rate (SemER) obtenus avec les systèmes de reconnaissance vocale Nuance et Kaldi.

Conclusion et perspectives – Ces premières expériences montrent qu'un système spécialisé de reconnaissance automatique de la parole peut être compétitif en termes de performance par rapport à des systèmes plus généralistes. L'ajout de grammaires spécialisées au sein du décodeur permet ainsi d'atteindre des performances exploitables en production. La mise en production est prévue dans le courant de l'année 2021. À plus long terme nous souhaiterions développer un système qui adapte dynamiquement sa grammaire en fonction des évolutions de son utilisation. Une autre piste serait de générer non pas une transcription, mais directement la forme canonique. Nous envisageons également d'exploiter ce système dans le cadre d'un système de traduction automatique de la parole vers des pictogrammes. En effet, cette étude s'inscrit dans le projet FNS-ANR[2] PROPICTO (*PRojection du langage Oral vers des unités PICTOgraphiques*), qui vise à développer des ressources et des outils pour la transcription automatique de la parole française et sa traduction en pictogrammes.

MOTS-CLES : reconnaissance automatique de la parole – modélisation acoustique – modélisation linguistique – Kaldi – BabelDr – discours médical

Références

BOUJON, V., BOUILLON, P., SPECHBACH, H., GERLACH, J., & STRASLY, I. (2018). Can speech-enabled phraselators improve healthcare accessibility ? A case study comparing BabelDr with MediBabble for anamnesis in emergency settings. *Proceedings of the 1st Swiss Conference on Barrier-Free Communication*, 50-65. https://doi.org/10.21256/zhaw-2018

HACKER, K., ANIES, M. E., FOLB, B., & ZALLMAN, L. (2015). Barriers to health care for undocumented immigrants : A literature review. *Risk Management and Healthcare Policy*, 175-183. https://doi.org/10.2147/RMHP.S70173

[1] Notons que l'évaluation en termes de SemER n'est effectuée que sur un sous-ensemble du corpus de développement, à savoir, 1222 phrases.

[2] Ce travail a bénéficié d'un financement du Fond National Suisse (No. 197864) et de l'Agence Nationale de la Recherche, via le projet PROPICTO (ANR-20-CE93-0005).

MUTAL, J., GERLACH, J., BOUILLON, P., & SPECHBACH, H. (2020). Ellipsis Translation for a Medical Speech to Speech Translation System. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 281-290. https://www.aclweb.org/anthology/2020.eamt-1.30/

POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., GLEMBEK, O., GOEL, N., HANNEMANN, M., MOTLICEK, P., QIAN, Y., SCHWARZ, P., SILOVSKY, J., STEMMER, G., & VESELY, K. (2011). *The Kaldi Speech Recognition Toolkit*. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. http://publications.idiap.ch/index.php/publications/show/2265

RAYNER, M., BOUILLON, P., TSOURAKIS, N., SPECHBACH, H., & GERLACH, J. (2018). Handling Ellipsis in a Spoken Medical Phraselator. In *Statistical Language and Speech Processing* (Vol. 11171, p. 140-152). Springer International Publishing. https://doi.org/10.1007/978-3-030-00810-9_13

Vers un système de dialogue oral pour la saisie de prescriptions médicales

Ali Can KOCABIYIKOGLU[1, 2] François Portet[1] Jean-Marc Babouchkine[2]
Hervé Blanchon[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG F-38000 Grenoble France
(2) Calystene SA, 38320 Eybens, France
a.kocabiyikoglu@calystene.com, francois.portet@imag.fr,
jm.babouchkine@calystene.com, herve.blanchon@imag.fr

Résumé :
Les Systèmes d'Information Hospitalier (SIH) se sont imposés dans les établissements de santé pour améliorer leur organisation ainsi que la qualité et la traçabilité des soins. L'un des composants majeurs des SIH est le Logiciel d'Aide à la Prescription (LAP) qui permet de limiter les événements indésirables médicamenteux (EIM). Cependant, la saisie informatique d'une prescription est laborieuse et réduit le temps consacré aux soins. Dans cette présentation, nous présentons un système de dialogue permettant aux cliniciens de réaliser des prescriptions médicamenteuses en langage naturel sur le lieu de soins via leur smartphone. Nous présentons la démarche, la modélisation du dialogue et un prototype initial du système de dialogue. Nous décrivons des expérimentations qui évaluent l'approche et permettent de faire ressortir des pistes d'amélioration possibles.

MOTS-CLÉS : Système de dialogue oral, Compréhension automatique du langage naturel, Informatique de santé, Traitement automatique des langues naturelles.

KEYWORDS: Spoken dialogue system, Natural Language Understanding, Health informatics, Natural Language Processing.

Références

HAUTE AUTORITÉ DE SANTÉ (2016). *Référentiel de certification par essaie de type des logiciels d'aide à la prescription en médecine ambulatoire*. Rapport interne, Haute Autorité de Santé.

KOCABIYIKOGLU A. C., PORTET F., BABOUCHKINE J.-M. & BLANCHON H. (2020). Spoken Medical Prescription Acquisition Through a Dialogue System on Smartphone : Perspective of a Healthcare Software Company. In *LREC 2020 Industry Track Language Resources and Evaluation Conference 11–16 May 2020*, Marseille, France. HAL : hal-02996728.

KOCABIYIKOGLU A. C., PORTET F., BLANCHON H. & BABOUCHKINE J.-M. (2019). Towards Spoken Medical Prescription Understanding. In *10th Conference on Speech Technology and Human-Computer Dialogue*, Timişoara, Romania. HAL : hal-02317503.

AfIA Association française pour l'Intelligence Artificielle
ATALA Association pour le Traitement Automatique des Langues

PyMedExt, un couteau suisse pour le traitement des textes médicaux

William Digan[1, 2] Alice Rogier[1, 2] David Baudoin[1, 2] Bastien Rance[1, 2]
Antoine Neuraz[1, 3]
(1) INSERM, Centre de Recherche des Cordeliers, UMRS 1138. Paris, France.
(2) Hôpital Européen Georges Pompidou, AP-HP, 75015 Paris, France
(3) Hôpital Necker - Enfants malades, AP-HP, 75015 Paris, France
william.digan@institutimagine.org, alice.rogier@inria.fr,
david.baudoin@aphp.fr, bastien.rance@aphp.fr, antoine.neuraz@aphp.fr

RÉSUMÉ

Le paysage du traitement des textes est vertigineux pour les non-spécialistes du traitement des langues. La prise en main d'outils existants, la contribution à leur évolution et le partage de proposition locale sont rendus complexes par l'absence de formats d'annotateurs à la fois faciles d'utilisation et simplement déployables. Il existe cependant de nombreuses ressources d'intérêts développées par la communauté. Nous proposons PyMedExt, un couteau suisse pour l'annotation de textes cliniques. PyMedExt a pour objectif de fluidifier la mise en place et la communication entre les différents composants nécessaires en traitement des textes cliniques : les formats de représentation des annotations, les annotateurs et des pipelines simples.

PyMedExt est construit autour de trois contributions principales :

1. Formats d'annotation et convertisseurs. PyMedExt peut consommer en entrée des fichiers textes bruts, des flux FHIR de textes, mais aussi des fichiers dans des formats classiques de TAL (BioC, CoNLL...). Il peut prendre en charge les conversions de format de et vers ces formats, ainsi que vers une représentation interne exportable en JSON. Cette représentation interne étend le format BioC en permettant l'héritage entre entités. PyMedExt propose également l'envoi vers les outils d'annotation et de visualisation d'annotations BRAT (Stenetorp *et al.*, 2011) et Doccano (Nakayama *et al.*, 2018), ainsi que la possibilité d'exporter les textes et les annotations vers une base de données au format CDM OMOP (Hripcsak *et al.*, 2015).
2. Un format de représentation d'annotateurs. PyMedExt inclut un squelette d'annotateur qui peut être facilement étendu pour créer ses propres fonctionnalités, ou afin de développer un adaptateur (wrapper) pour des outils existants. Quelques exemples d'annotateurs PyMedExt natifs ou d'adaptateurs, dont QuickUMLS (Soldaini & Goharian, 2016) et HeidelTime (Strötgen & Gertz, 2010), sont disponibles sur le dépôt du projet (`https://github.com/equipe22/pymedext_public`). Le format d'annotateur est volontairement simple et peu contraignant pour favoriser une adoption par une communauté large.
3. Gestionnaire de pipelines simples. Enfin, PyMedExt propose un système de gestion de pipelines linéaires élémentaires d'annotateurs, permettant le déploiement pour un usage reproductible. PyMedExt est distribué sous la forme d'une bibliothèque Python, utilisable en ligne de commande pour les fonctionnalités de conversion, et embarquée dans un programme pour les fonctionnalités plus avancées (conversion, annotateurs et pipelines).

PyMedExt est utilisé par les équipes d'informatique de l'hôpital Necker et de l'HEGP pour standardiser les pratiques et simplifier le partage d'outils. Une bibliothèque d'annotateurs open-source publics au format PyMedExt est proposée en ligne https://github.com/equipe22/pymedext_core.

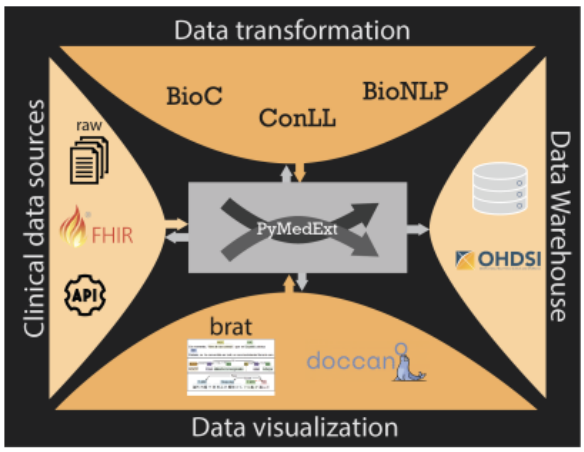

FIGURE 1 – Résumé graphique de PyMedExt

Références

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

HRIPCSAK G., DUKE J., SHAH N., REICH C., HUSER V., SCHUEMIE M., SUCHARD M., PARK R., WONG I., RIJNBEEK P., VAN DER LEI J., PRATT N., NORÉN G., LI Y., STANG P., MADIGAN D. & RYAN P. (2015). Observational Health Data Sciences and Informatics (OHDSI) : Opportunities for Observational Researchers. In *Stud Health Technol Inform*, volume 216, p. 574–8.

NAKAYAMA H., KUBO T., KAMURA J., TANIGUCHI Y. & LIANG X. (2018). Doccano : Text Annotation Tool for Human.

SOLDAINI L. & GOHARIAN N. (2016). QuickUMLS : a fast, unsupervised approach for medical concept extraction. In *SMedIR Workshop, SIGIR*.

STENETORP P., TOPIĆ G., PYYSALO S., OHTA T., KIM J.-D. & TSUJII J. (2011). BRAT. BioNLP Shared Task 2011 : Supporting Resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*.

STRÖTGEN J. & GERTZ M. (2010). HeidelTime : High quality rule-based extraction and normalization of temporal expressions. In *SemEval '10 : Proceedings of the 5th International Workshop on Semantic Evaluation*, volume 216, p. 321–324.

AfIA Association française pour l'Intelligence Artificielle
ATALA Association pour le Traitement Automatique des Langues

Traitement de la langue naturelle pour une réponse rapide aux maladies émergentes: COVID-19

Neuraz Antoine[1] Lerner Ivan[1] Digan William[2]
Garcelon Nicolas[3] Tsopra Rosy[2] Rogier Alice[2] Baudoin David[2]
Burgun Anita[1, 2] Rance Bastien[2]
(1) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université de Paris, Hôpital Necker Enfant Malade, Assistance Publique - Hôpitaux de Paris
(2) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université de Paris, Hôpital Européen Georges Pompidou, Assistance Publique - Hôpitaux de Paris
(3) INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Institut Imagine, INSERM U1163, Université Paris Descartes, Université de Paris, Paris, France
antoine.neuraz@aphp.fr

RÉSUMÉ

La fouille de données biomédicales dans les dossiers patients informatisés (DPI) a souvent été proposée comme méthode pour convertir les données non structurées vers les données structurées nécessaires pour la santé publique. Bien que cela ait souvent été suggéré (Elkin *et al.*, 2012), l'occasion ne s'était encore jamais présentée de pouvoir tester cette hypothèse en temps réel. Ainsi, la crise du coronavirus, malgré toutes ses tragédies, présente également l'opportunité d'améliorer l'informatique de santé publique. Durant la crise, l'APHP a mis en place une base de données au format OMOP CDM (Hripcsak *et al.*, 2015) contenant les données des DPI de tous les patients testés COVID-19. Voici un résumé de la méthode que nous avons utilisé pour traiter les textes clinique (Figure 1) : (1) un pré-traitement classique (*i.e.*, nettoyage du texte, détection des phrases) a été appliqué sur l'ensemble du dataset ; (2) l'extraction des noms de médicaments et des détails de prescription (dose, voie d'administration, fréquence, durée) a été effectuée à l'aire de modèles de deep-learning basés sur des embeddings contextuels de type BERT (Devlin *et al.*, 2018) fine-tuné sur 10M de textes cliniques et un modèle BiLSTM-CRF (Lample *et al.*, 2016) ($NLP_{medication}$) ; (3) l'extraction de phénotypes spécifiques associés au COVID-19 (*e.g.*, obésité, fumeur), de scores (*e.g.*, IGS2), et de mesures physiologiques (*e.g.*, Body Mass Index), a été effectuée via une liste d'expressions régulières spécialement développées ; (4) l'extraction de tous signes, symptômes, comorbidités présentes dans le Unified Medical Language System (UMLS) (Lindberg *et al.*, 1993), a été effectuée avec l'algorithme quickUMLS (Okazaki & Tsujii, 2010)). L'utilisation d'outils TAL a permis d'augmenter, de manière importante, la quantité d'informations, concernant les médicaments et les phénotypes, disponibles pour l'analyse. Le nombre de points de données pour les médicaments a été multiplié par 7.2 et le nombre de phénotypes par 15.2. Parmi les 84,966 dossiers présents dans la base EDS-COVID, 53% des patients avaient des informations sur les médicaments dans les textes cliniques contre seulement 23% dans les champs structurés. Pour les phénotypes spécifiques avec des codes CIM10 existant, l'information était disponible uniquement dans le texte libre pour une majorité de patients : 7,133/8,526 (83%) pour le diabète et 2,138/2,871 (74%) pour l'obésité. Certains items étaient absents des données structurées mais ont pu être récupérés via le TAL, comme l'agueusie ou l'anosmie, 2,449 et 2,732 patients respectivement. $NLP_{medication}$ a montré une F1-mesure brute à 93.8% (91.6% après normalisation) pour l'extraction des noms de médicament sur l'ensemble des sections ; 96.7%

(96% après normalisation) sur les sections traitement à l'admission et traitement de sortie.

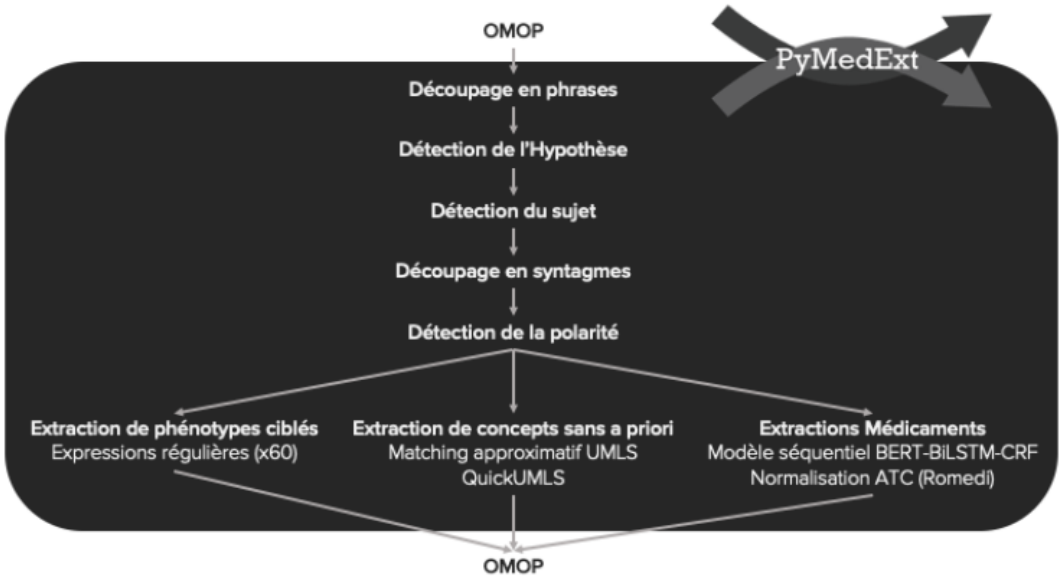

FIGURE 1 – Pipeline de traitement des textes cliniques

MOTS-CLÉS : Dossier patient informatisé, extraction d'information, COVID-19, deep-learning.

Références

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv :1810.04805 [cs]*.

ELKIN P. L., FROEHLING D. A., WAHNER-ROEDLER D. L., BROWN S. H. & BAILEY K. R. (2012). Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. *Annals of Internal Medicine*, **156**(1_Part_1), 11–18.

HRIPCSAK G., DUKE J. D., SHAH N. H., REICH C. G., HUSER V., SCHUEMIE M. J., SUCHARD M. A., PARK R. W., WONG I. C. K., RIJNBEEK P. R., VAN DER LEI J., PRATT N., NORÉN G. N., LI Y.-C., STANG P. E., MADIGAN D. & RYAN P. B. (2015). Observational Health Data Sciences and Informatics (OHDSI) : Opportunities for Observational Researchers. *Studies in Health Technology and Informatics*, **216**, 574–578.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT 2016*, volume 5805, p. 260–270, San Diego, California, June 12-17, 2016 : ACL.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The Unified Medical Language System. *Methods Archive*, **32**, 281–291.

OKAZAKI N. & TSUJII J. (2010). Simple and Efficient Algorithm for Approximate Dictionary Matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 851–859, Beijing, China : Coling 2010 Organizing Committee.

Traitement Automatique de la Langue et Intégration de Données pour les Réunions de Concertations Pluridisciplinaires en Oncologie

Nesrine Bannour[1] Aurélie Névéol[1] Xavier Tannier[2] Bastien Rance[3]
(1) Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), rue John Von Neumann, 91400 Orsay, France
(2) Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS), 15 Rue de l'École de Médecine, 75006 Paris , France
(3) Hôpital Européen Georges Pompidou (HEGP) AP-HP, 20 Rue Leblanc, 75015 Paris, France
nesrine.bannour@limsi.fr, aurelie.neveol@limsi.fr,
xavier.tannier@sorbonne-universite.fr, bastien.rance@aphp.fr

RÉSUMÉ

Les réunions de concertations pluridisciplinaires (RCP) en oncologie permettent aux experts des différentes spécialités de choisir les meilleures options thérapeutiques pour les patients. Les données nécessaires à ces réunions sont souvent collectées manuellement, avec un risque d'erreur lors de l'extraction et un coût important pour les professionnels de santé. Plusieurs travaux scientifiques portant sur des documents en anglais se sont intéressés à l'extraction automatique d'informations (telles que la localisation de la tumeur, les classifications histologiques, TNM, ...) dans les rapports cliniques des dossiers médicaux (Nguyen *et al.*, 2010; Savova *et al.*, 2017; Gupta *et al.*, 2019). Dans le cadre du projet ASIMOV (ASsIster la recherche en oncologie par le Machine Learning, l'intégration de dOnnées et la Visualisation), nous utiliserons le traitement automatique de la langue et l'intégration de données pour l'extraction d'informations liées au cancer dans les entrepôts de données et les textes cliniques en français, comme illustré en figure 1.

MOTS-CLÉS : Traitement Automatique de la Langue, Intégration de Données, Analyse Temporelle, Parcours de Soin, Oncologie.

Références

GUPTA K., THAMMASUDJARIT R. & THAKKINSTIAN A. (2019). NLP automation to read radiological reports to detect the stage of cancer among lung cancer patients. In *WNLP@ACL*.

NGUYEN A., LAWLEY M., HANSEN D., BOWMAN R., CLARKE B., DUHIG E. & COLQUIST S. (2010). Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association : JAMIA*, **17 4**, 440–5.

SAVOVA G., TSEYTLIN E., FINAN S., CASTINE M., MILLER T., MEDVEDEVA O., HARRIS D., HOCHHEISER H., LIN C., CHAVAN G. & JACOBSON R. S. (2017). DeepPhe - a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer research*, **77 21**, e115–e118.

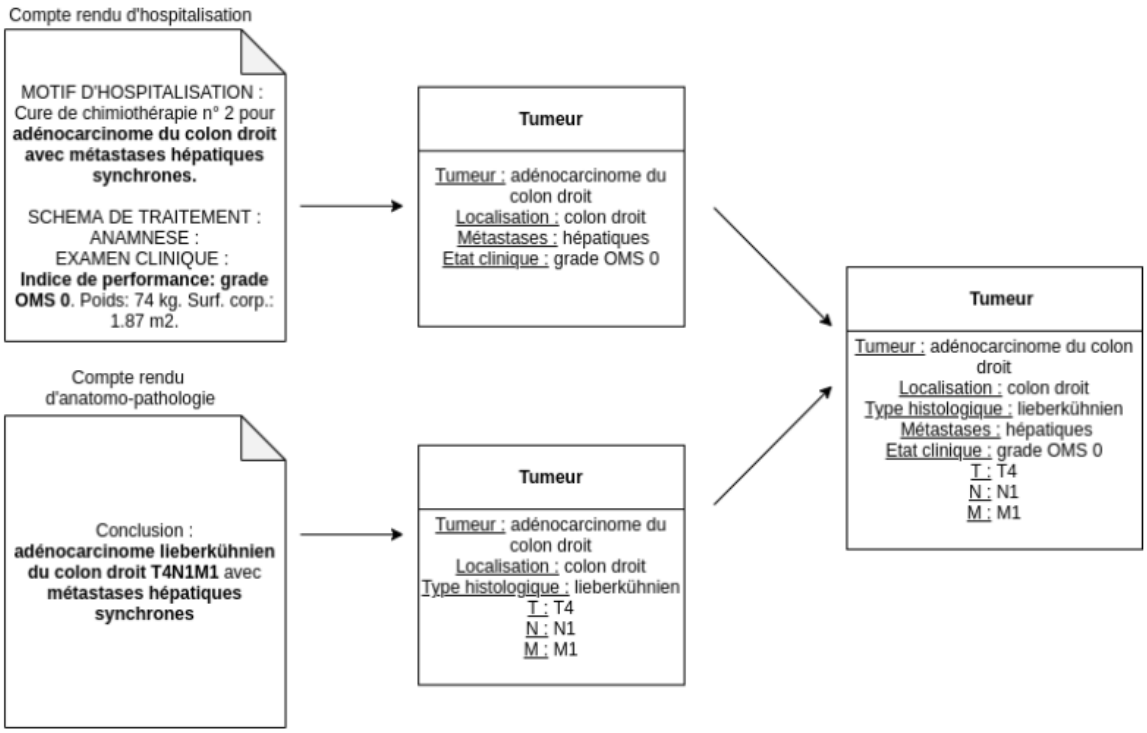

FIGURE 1 – Extraction d'informations cancer d'un dossier patient en français

AfIA
Association française
pour l'Intelligence Artificielle

ATALA
Association
pour le Traitement
Automatique
des Langues

Traitement automatique des langues pour la santé : travaux récents au LISN

Pierre Zweigenbaum
Universite Paris-Saclay, CNRS, LISN, F-91405 Orsay, France
pz@limsi.fr

RÉSUMÉ

La médecine est un champ d'expérimentation ancien pour l'intelligence artificielle, dans l'aide à la décision comme dans la compréhension de textes. Je présenterai un panorama de travaux qui mobilisent différents champs du traitement automatique des langues pour contribuer à des tâches concernant divers acteurs du monde de la santé, des professionnels aux patients en passant par les chercheurs et les étudiants en médecine. Une caractéristique récurrente est la prise en compte d'un vocabulaire spécialisé varié et la mise en relation avec des bases termino-ontologiques de grande taille, qui ont été traitées aussi bien par des méthodes à base de connaissances que par des représentations distribuées apprises à partir de données.

Références

ABDUL-RAUF S., ROSALES J. C., PHAM M. Q. & YVON F. (2020). LIMSI @ WMT 2020. In *Conference on Machine Translation*, Online, United States. HAL : hal-03013198.

CAMPILLOS LLANOS L., THOMAS C., BILINSKI E., ZWEIGENBAUM P. & ROSSET S. (2019). Designing a virtual patient dialogue system based on terminology-rich resources : Challenges and evaluation. *Natural Language Engineering*, p. 1–38. DOI : 10.1017/S1351324919000329.

GROUIN C. & ZWEIGENBAUM P. (2013). Automatic de-identification of French clinical records : Comparison of rule-based and machine-learning approaches. In *Proc MEDINFO 2013*, Studies in Health Technology and Informatics, p. 476–480 : Amsterdam IOS Press. DOI : doi :10.3233/978-1-61499-289-9-476.

KOROLEVA A. & PAROUBEK P. (2019). Extracting relations between outcomes and significance levels in randomized controlled trials (RCTs) publications. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 359–369, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5038.

LLANOS L. C., GROUIN C., LOUËT A. L.-L. & ZWEIGENBAUM P. (2019). Initial experiments for pharmacovigilance analysis in social media using summaries of product characteristics. In *MEDINFO 2019*, volume 264 de *Studies in Health Technology and Informatics*, p. 60–64, Lyon, France : Amsterdam : IOS Press.

NORMAN C., LEEFLANG M., ZWEIGENBAUM P. & NÉVÉOL A. (2018). Automating Document Discovery in the Systematic Review Process : How to Use Chaff to Extract Wheat. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), p. 3681–3687, Miyazaki, Japan : European Language Resources Association (ELRA).

RANDRIATSITOHAINA T. & HAMON T. (2019). Extracting Food-Drug Interactions from Scientific Literature : Tackling Unspecified Relation. In *Conference on Artificial Intelligence in Medecine Europe*, Poznan, Poland. HAL : hal-02122580.

ZHENG Y., MENG X., ZWEIGENBAUM P., CHEN L. & XIA J. (2020). Hybrid phenotype mining method for investigating off-target protein and underlying side effects of anti-tumor immunotherapy. *BMC Med Inform Decis Mak*, **20**(Suppl 3), 133. DOI : 10.1186/s12911-020-1105-4.

ZWEIGENBAUM P. & LAVERGNE T. (2017). Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. In *CLEF 2017 Evaluation Labs and Workshop : Online Working Notes*, CEUR Workshop Proceedings, Dublin, Ireland.

AfIA
Association française
pour l'Intelligence Artificielle

ATALA
Association
pour le Traitement
Automatique
des Langues

Investigation des marqueurs langagiers non-lexicaux et spécifiques des personnes souffrant de schizophrénie dans des conversations spontanées

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

RÉSUMÉ

Notre travail s'intéresse aux marqueurs linguistiques potentiellement spécifiques du discours des patients souffrant de schizophrénie. La production langagière associée à ce trouble psychiatrique grave a fait l'objet de nombreuses études montrant une moindre richesse lexicale et syntaxique d'une part, et d'autre part une difficulté à maintenir la cohérence du discours et du dialogue. Nous nous intéressons ici à des dialogues entre cliniciens et patients enregistrés en situation naturelle, contrairement à la majorité des études précédentes qui utilisent soit des tâches expérimentales particulières, soit des données issues des réseaux sociaux. Ces dialogues correspondent à une situation plus réaliste, mais nécessitent l'usage d'outils et de stratégies d'analyse singulière étant donné la rareté des phénomènes appréhendés. Précisément, nous extrayons de ces conversations cliniques, entre un sujet (contrôle ou patient avec schizophrénie) et un psychologue, des données à partir desquelles nous construisons des modèles langagiers de classification distinguant les deux groupes de locuteurs, de manière à identifier leurs spécificités linguistiques et/ou psycho-linguistiques.

Nous proposons différentes approches pour représenter les données dialogiques, en faisant varier le contexte : d'abord fusionner l'ensemble des tours de parole de chaque interlocuteur (*Full*) et considérer la contribution comme un unique document, puis faire varier la taille de la fenêtre en la restreignant à n tokens, possiblement sur plusieurs tours de paroles (*W-n*), jusqu'à considérer les tours de parole individuels (*Indiv.*). Concernant la modélisation, nous avons tout d'abord entraîné des modèles de classification utilisant des informations lexicales, morphologiques, syntaxiques, discursives et dialogiques.

Les résultats préliminaires (Amblard *et al.*, 2020) présentent de bonnes performances avec par exemple une précision à 93,7%. Cependant, l'analyse des modèles produits montre que les plus performants utilisent des traits lexicalisés dont les plus caractéristiques sont fortement biaisés, par exemple l'utilisation du champ lexical de la maladie et du traitement pour une population et pas l'autre. Les systèmes ainsi développés ne généralisent pas suffisamment et manquent de robustesse. Nous étendons notre étude avec la définition de modèles utilisant des traits non-lexicaux en nous intéressant à d'autres niveaux linguistiques : i) n-grammes de POS tags et de relations syntaxiques

(*treelets*) (Johannsen *et al.*, 2015) ; ii) traits dialogiques de “haut niveau" comme les OCR (*Open Class Repair* - “pardon ?",“huh ?") et les *Backchannels* (expressions phatiques comme “hum mmh", “yeah") (Howes *et al.*, 2012a,b, 2013) ; iii) traits discursifs *via* les connecteurs identifiés dans LexConn (Roze *et al.*, 2012). Notre meilleur modèle atteint une précision à 77,9% avec la combinaison des POS tag (3-grammes) et des *Backchannels* pour des fenêtres de 512 tokens (*W*-512), suivi par la combinaison des 2-grammes de POS tag et des *Backchannels* dans la configuration *Full*. Le meilleur score pour un modèle sans combinaison de traits est à 74,19% avec les 2-grammes de *treelet* également pour des fenêtres *W*-512. Nous constatons que combiner les *Backchannels* avec d'autres traits augmente généralement les résultats. Par ailleurs, les connecteurs donnent également de bonnes performances, avec un meilleur score à 73,6%. La taille de la fenêtre semble avoir également une incidence car les meilleurs résultats sont atteints avec des fenêtres de grande taille.
Dans cette étude nous comparons l'utilisation de traits linguistiques non-lexicaux dans des contextes de taille variable. C'est la première du genre sur le français et nous obtenons des résultats comparables à celles sur l'anglais (Kayi *et al.*, 2017; Allende-Cid *et al.*, 2019). Nos premières expériences tendent à montrer que la parole des personnes avec schizophrénie se distingue de celle des témoins avec des particularités langagières caractéristiques du dialogue. Les perspectives que nous souhaitons donner sont : i) de complexifier la notion de contexte en considérant l'alternance des tours de paroles et non plus seulement la production d'un seul locuteur, ii) améliorer l'identification de la structure du dialogue via la désambiguïsation des connecteurs identifiés, iii) investiguer l'utilisation de plongements lexicaux de grande taille et des méthodes neuronales. Pour toutes ces approches, notre question est d'extraire des traits non-lexicalisés.

MOTS-CLÉS : Dialogue, schizophrénie, apprentissage automatique, traits linguistiques, traits non-lexicaux.

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., BRAUD C., LI C., DEMILY C., FRANCK N. & MUSIOL M. (2020). Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie (investigating learning methods applied to language specificity of persons with schizophrenia). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e 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*, p. 12–26.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

HOWES C., PURVER M. & MCCABE R. (2013). Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomedical informatics insights*, **6**, BII–S11661.

HOWES C., PURVER M., MCCABE R., HEALEY P. & LAVELLE M. (2012a). Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 79–83.

HOWES C., PURVER M., MCCABE R., HEALEY P. G. & LAVELLE M. (2012b). Helping the medicine go down : Repair and adherence in patient-clinician dialogues. In *Proceedings of SemDial 2012 (SeineDial) : The 16th Workshop on the Semantics and Pragmatics of Dialogue*, p. 155.

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.

ROZE C., DANLOS L. & MULLER P. (2012). Lexconn : a french lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (10).

AfIA Association française pour l'Intelligence Artificielle
ATALA Association pour le Traitement Automatique des Langues

Measurements of turn-taking and linguistic behaviors in clinical settings

Rachid Riad[1, 2, 4] Lucas Gautheron[1] Emmanuel Dupoux [1, 3] Anne-Catherine Bachoud-Lévi [4] Alejandrina Cristia[1]

(1) Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France

(2) CoML Team, INRIA, Paris, France

(3) Facebook Artificial Intelligence Research, Paris, France

(4) Laboratoire de Neuropsychologie Interventionnelle, Département d'Etudes cognitives, ENS, INSERM, UPEC, PSL University, Paris, France

riadrachid3@gmail.com, alecristia@gmail.com*

RÉSUMÉ

Les conversations impliquant des patients, dans des conditions naturelles, sont de précieuses sources d'informations pour le suivi médical. Il est devenu plus facile de collecter de grands corpus de données vocales. Cependant, obtenir des indicateurs pertinents à partir d'enregistrements audio est actuellement hors de portée pour la plupart des cliniciens et des chercheurs. *Alors, comment pouvons-nous obtenir des mesures du tour de paroles et des marqueurs linguistiques pour des applications cliniques ?*. Les méthodes envisagées doivent être *fiables*, *peu coûteuses* et garantir la *vie privée*. Nous discuterons des avantages et des inconvénients de plusieurs approches qui sont actuellement développées et examinées dans notre laboratoire : 1) Développement de nouveaux outils pour l'annotation par des experts de fichiers audio (Titeux* *et al.*, 2020) , 2) Crowdsourcing avec des citoyens qui veulent aider dans l'avancée des recherches scientifiques (Semenzin *et al.*, 2020), 3) Développement d'algorithmes pour détecter et identifier les tours de parole durant les entretiens cliniques (Riad *et al.*, 2020).

ABSTRACT

Measurements of turn-taking and linguistic behaviors in clinical settings

Conversations involving patients, in natural conditions, are valuable sources of information for the medical follow-up. Thanks to recent technological advances, it became easier to collect large naturalistic corpora of speech data. However, getting meaningful insights from these long naturalistic datasets is currently out of reach for most clinicians and researchers. *Then, how do we obtain measurements of naturalistic turn-taking and linguistic behaviors for clinical applications ?* Especially, for scientific endeavor and clinical practice, the methods should be *reliable*, *cost-effective* and guarantee the *privacy* of the subjects. Indeed, these long recordings can vary greatly by their duration, by their recording conditions and by the variability of speaker traits (that we might want to capture and measure for as clinical outcomes). In this presentation, we will discuss advantages and drawbacks of several approaches that are currently being developed and examined in our laboratory to process audio recordings in clinical settings. On the one hand, researchers and clinicians can rely on expert annotators of speech and language data. We developed standardized ways of assigning annotation

. We acknowledge ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017; and the J. S. McDonnell Foundation, ANR-10-IDEX-0001-02 PSL, ANR-19-P3IA-0001 PRAIRIE 3IA Institute, Neuratris, Facebook AI Research (Research Gift), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).

to expert annotators (Titeux* *et al.*, 2020), with the evaluation of their agreements. After training the annotators, the quality of annotations is high but it becomes quickly infeasible to annotate huge corpora. On the other hand, to overcome this scaling issue, we also investigated the capabilities of the crowd-sourcing project on Zooniverse to leverage citizens' help to solve simple classification tasks on short audio chunks drawn randomly from the daylong recordings (Semenzin *et al.*, 2020) of children with and without Angelman syndrome. Finally, we witnessed an increase in performance of the speech processing algorithms thanks to neural networks trained on large open datasets. Yet, it is not known how these algorithms transfer and perform on speech productions of individuals with speech and language deficits. We evaluated the state-of-the-art pipelines to detect and identify speaker turns for conversational clinical interviews between neuropsychologists and patients with Huntington's Disease (Riad *et al.*, 2020). We found that these algorithms required a sufficient amount of well annotated data to reach significant performance. Finally, these algorithms provided halfway satisfactory results concerning speech features relevant for clinical practice.

MOTS-CLÉS : tours de parole ; daylong recordings ; annotations ; traitement de la parole pathologique.

Références

RIAD R., TITEUX H., LEMOINE L., MONTILLOT J., SLIWINSKI A., BAGNOU J. H., CAO X. N., DUPOUX E. & BACHOUD-LÉVI A.-C. (2020). Comparison of speaker role recognition and speaker enrollment protocol for conversational clinical interviews. *arXiv preprint arXiv :2010.16131*.

SEMENZIN C., HAMRICK L., SEIDL A., KELLEHER B. & CRISTIA A. (2020). Towards large-scale data annotation of audio from wearables : Validating zooniverse annotations of infant vocalization types. In *SLT*.

TITEUX* H., RIAD* R., CAO X.-N., HAMILAKIS N., MADDEN K., CRISTIA A., BACHOUD-LÉVI A.-C. & DUPOUX E. (2020). Seshat : A tool for managing and verifying annotation campaigns of audio data. In *LREC*, Marseille. * Equal contribution.

AfIA Association française pour l'Intelligence Artificielle
ATALA Association pour le Traitement Automatique des Langues

Exploration de la temporalité dans la désignation des pathologies du langage en orthophonie : aspects cliniques et termino-ontologiques

Frédérique Brin-Henry
Laboratoire ATILF-UMR 7118 CNRS-Université de Lorraine, 44 av Libération 54000 Nancy
Centre Hospitalier, 1 Boulevard d'Argonne 55012 Bar-le-Duc
frederique.henry@atilf.fr

RESUME

L'orthophonie est centrée autour de la prévention, l'identification, l'évaluation et le traitement des difficultés de langage, de la communication et des fonctions oromyofaciales des personnes de tous âges. Au cours du processus de la pose du diagnostic orthophonique, les orthophonistes procèdent au choix d'une étiquette diagnostique (Plug et al 2010) à l'issue d'un temps élaboré et structuré de bilan. Ces étiquettes diagnostiques sont variables dans le temps et d'une langue à l'autre. Le processus onomasiologique de la pose du diagnostic a été étudié pour ce qui concerne l'orthophonie française depuis une dizaine d'années, dans l'objectif de comprendre ce que révèlent les variantes des unités polylexicales d'un point de vue sémantico-syntaxique, d'éclairer sur le lien entre les termes et les concepts des pathologies du langage en orthophonie, et également afin de démontrer la spécificité des représentations des orthophonistes concernant ces pathologies.

Il est apparu que la temporalité était un critère important et particulièrement pertinent pour élaborer une conceptualisation spécifique des pathologies du langage. Cette représentation a été objectivée dans un modèle permettant de raisonner sur les concepts ainsi sélectionnés et organisés. Cette présentation aura pour but de justifier la focalisation sur ce critère de temporalité, en présentant des exemples selon deux axes complémentaires :

- un axe clinique grâce à l'étude de la temporalité dans la description de l'aphasie (Dolveck 2019). Dans cette étude sur 32 dossiers médicaux et orthophoniques de patients présentant une aphasie après un accident vasculaire cérébral, des critères temporels ont été extraits (délais d'intervention, qualité de la récupération spontanée, durée de séjour…) montrant que le délai de récupération et la progression d'évolution de l'aphasie pourraient représenter des critères distinctifs pertinents pour classer les aphasies.

- un axe terminologique grâce à la présentation de la présence de la temporalité en tant que propriété sémantique, révélée dans les formes des termes dédiés aux pathologie du langage (ex : bégaiement physiologique/bégaiement chronique, retard de langage/dysphasie, troubles développementaux du langage/ troubles phonologiques acquis). Ces exemples sont issus de l'analyse de deux corpus de spécialité (CRBO de 180 000 mots et OrthoCorpus[1] de 5 millions de mots).

[1] Analyse et traitement informatique de la langue française - UMR 7118 (ATILF) (2020). *OrthoCorpus* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v2, https://hdl.handle.net/11403/orthocorpus/v2.

AFIA Association française pour l'Intelligence Artificielle
ATALA Association pour le Traitement Automatique des Langues

Fouille de la littérature médicale à l'aide de graphes

Elise Bigeard[1] Aman Sinha[1, 2] Marianne Clausel[1] Mathieu Constant[3]
(1) Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France
(2) Indian Institute of Technology, Dhanbad, Jharkhand 826004, India
(3) Université de Lorraine, CNRS, ATILF UMR 7118, 44 Avenue de la Libération, 54000 Nancy
elise.bigeard@univ-lorraine.fr, marianne.clausel@univ-lorraine.fr

MOTS-CLÉS : Fouille de texte, littérature médicale, graphes, embeddings.

La littérature scientifique est abondante, au point que l'explorer efficacement est maintenant une tâche majeure. Il existe de nombreuses plateformes répertoriant des publications : Arxiv, DBLP... et bien entendu Pubmed pour la littérature médicale. Ces plateformes sont une ressource clé, mais contiennent un très large contenu, et peuvent être difficiles à explorer.

Nous proposons d'analyser la littérature scientifique, et en particulier médicale, à l'aide de graphes de connaissance (Ji *et al.*, 2020). L'objectif est double : d'une part, proposer un outil graphique permettant d'explorer plus facilement et intuitivement la littérature. D'autre part, d'utiliser des plongements (embeddings) de graphes pour réaliser diverses tâches d'apprentissage automatique telles que la recommandation : étant donné par example une publication, suggérer des publications similaires, ou suggérer des co-auteurs potentiels.

Nous présentons une méthode full stack, allant d'une collection de publications au format PDF jusqu'à une représentation graphique accessible en ligne sur notre démonstrateur : https://gremie-demonstrator.atilf.fr

Nous testons notre méthode sur un ensemble de publications médicales au format PDF provenant d'un même établissement de recherche. Nous utilisons également des corpus non médicaux : ACM, ACL Anthlogy et DBLP.

En ce qui concerne la partie représentation des connaissances : nous créons un graphe de notre corpus, où sont présents les types de nœuds suivants : publication, auteur et mot-clé. Les mots-clé sont détectés automatiquement dans le texte de la publication, auxquels s'ajoutent les mots-clé indiqués explicitement par les auteurs. Nous nous basons sur la terminologie MESH et sur des synonymes issus de Wikidata pour détecter les mots-clé. Gephi est utilisé pour la représentation graphique.

En ce qui concerne les plongements, nous utilisons conjointement deux sources de données : le graphe, et le texte des articles.

Le graphe nous permet de comparer des nœuds en fonction de leurs liens. Ainsi, des auteurs peuvent être considérés comme similaires s'ils sont liés aux mêmes mot-clé, ou s'ils ont tous les deux un grand nombre de publications. Nous nous basons sur Deepwalk (Perozzi *et al.*, 2014), GraphSage (Hamilton *et al.*, 2017) et GCN Kipf2017.

Le texte de l'article nous permet de rapprocher des articles sur un sujet commun, ou utilisant des méthodes similaires. Nous utilisons plusieurs représentations de texte classiques : TF-IDF, plongements de mots (Mikolov *et al.*, 2013) et plongements de documents (Le & Mikolov, 2014).

Remerciement

Nous remercions Cancéropôle Est et l'INIST pour leur contribution, notamment leur corpus de publications.

Nous remercions le Proket Olki et l'agence AMIES pour leur financement.

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.

Références

HAMILTON W. L., YING R. & LESKOVEC J. (2017). Inductive representation learning on large graphs. *CoRR*, **abs/1706.02216**.

JI S., PAN S., CAMBRIA E., MARTTINEN P. & YU P. S. (2020). A survey on knowledge graphs : Representation, acquisition and applications.

LE Q. & MIKOLOV T. (2014). Distributed representations of sentences and documents. In E. P. XING & T. JEBARA, Éds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 de *Proceedings of Machine Learning Research*, p. 1188–1196, Bejing, China : PMLR.

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, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems*, volume 26, p. 3111–3119 : Curran Associates, Inc.

PEROZZI B., AL-RFOU R. & SKIENA S. (2014). Deepwalk. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. DOI : 10.1145/2623330.2623732.

La communication en santé publique au temps du Covid-19 dans les contextes français, québécois et tunisien

Silvia Calvi[1] Klara Dankova[2]
(1) Università degli Studi di Verona, Via S. Francesco 22, 37129 Verona, Italie
(2) Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italie
silvia.calvi@univr.it, klara.dankova@unicatt.it

MOTS-CLES : terminologie, discours argumentatif, Covid-19, santé publique, communication aux citoyens.

La pandémie du Covid-19 a mis en évidence le rôle crucial de la communication en santé publique pour la protection de la population : entre autres, sa fonction est de fournir des consignes de prévention et de renseigner sur des actions à accomplir si on soupçonne d'avoir contracté la maladie. Pour ce faire, la communication en santé publique doit adopter un langage qui soit non seulement simple et compréhensible, mais aussi précis et univoque. Plus particulièrement, les sites officiels chargés de cette communication doivent avoir un format convivial et les argumentations présentées doivent être claires et convaincantes.

Notre communication se propose d'analyser les sites web de trois pays/régions francophones contenant la communication officielle sur le déroulement et la gestion locale et internationale de la pandémie. En particulier, nous nous concentrons sur la France, le Québec et la Tunisie. Nous avons décidé d'examiner la communication aux citoyens dans ces territoires culturellement éloignés parce que nous supposons qu'elle présente des stratégies communicatives différentes. De plus, en les choisissant on a vérifié que ces territoires ont été, et le sont encore, particulièrement touchés par la pandémie du Covid-19[1]. L'objectif de cette communication est donc d'étudier ces sites pour évaluer les stratégies communicatives adoptées dans chaque contexte culturel.

[1] Les données sur l'impact de la pandémie en France et en Tunisie sont disponibles sur le site du *Le Monde* : <https://www.lemonde.fr/les-decodeurs/article/2020/05/05/coronavirus-age-mortalite-departements-pays-suivez-l-evolution-de-l-epidemie-en-cartes-et-graphiques_6038751_4355770.html> [consulté le 4 décembre 2020]. En ce qui concerne le Québec, les informations sur la crise sanitaire sont repérables sur le site : https://www.quebec.ca/sante/problemes-de-sante/a-z/coronavirus-2019/situation-coronavirus-quebec/ [consulté le 4 décembre 2020].