Actes de la journée d'étude sur la robustesse des systemes de TAL

Caio Corro, Gaël Lejeune

Actes de la journée d'étude sur la robustesse des systèmes de TAL

Avec le soutien de l'ATALA et du laboratoire STIH

Éditeurs

Caio Corro
Université Paris-Saclay, CNRS, LISN

Gaël Lejeune
Sorbonne Université, STIH

25 novembre 2022
Maison de la recherche, 28 rue Serpente, 75006 Paris

Préface

Les méthodes d'apprentissage par transfert sont omniprésentes en traitement automatique des langues (TAL). Grâces aux développements récent à la fois en termes de matériel (existence de grandes fermes de GPU et TPU) et d'architectures neuronales (mécanisme d'attention trivialement parallélisable), la communauté a pu mettre à disposition de nombreux modèles pré-entraînés sur de grands corpus de données, en général soit des modèles de langues (par exemple GPT et autres dérivés) soit des modèles contextuels pré-entraînés via masquage (BERT et autres dérivés). Cela a conduit à des résultats impressionnants sur de nombreux *benchmarks* qui ont repoussé les performances état de l'art ce qui peut amener à questionner l'utilité des *pipelines* traditionnels du TAL (comprenant par exemple des étapes intermédiaires comme l'analyse syntaxique).

Dans ce contexte, il nous a semblé important de prendre un peu de recul et de regarder ce qu'il se passe lorsque l'on applique des systèmes de TAL dans des environnements qui ne sont pas nécessairement contrôlés. Cette journée d'études vise donc à réunir les collègues s'intéressant à la robustesse des systèmes de TAL sur des données « non-standards ». Par données non-standard, nous désignons des données présentant des variations vis-à-vis d'un certain attendu en terme d'état de langue (variation de la langue en diachronie, variations régionales, variation dans l'ordre des mots, *code-switching*, *user generated content*, orthographe inconsistante, données accidentellement bruitées suite à un pré-traitement, données incomplètes, présence d'un vocabulaire de domaine spécialisé...).

L'objectif de cette journée est double :

— documenter les cas pratiques dans lesquels les systèmes de TAL existants se sont révélés peu fiables voir inutilisables, par exemple, mais sans y être limité, dans le domaine des humanités numériques ;

— documenter les solutions existantes, par exemple, mais sans y être limité, pour les systèmes fondées sur des méthodes d'apprentissage automatique.

En bref, nous espérons que cette journée permettra d'échanger sur la recherche en TAL et de ses applications en dehors des *benchmarks* standards.

Sur une note plus générale, nous espérons que cette première journée sera l'occasion de relancer un cycle régulier de journées d'étude sous l'égide de l'ATALA. Notre souhait est d'arriver à deux journées par an (une en novembre et une en mars), soit directement organisées par nous même, soit organisées par des membres extérieurs au conseil d'administration mais toujours avec le soutien de l'ATALA. Par exemple, la prochaine journée d'étude, qui aura lieu le 13 mars 2023 et dont le thème est « Similarité en santé », sera organisé par Adrien Coulet, Christel Gérardin, Aurélie Névéol et Xavier Tannier.

Le comité d'organisation.

Comité d'organisation

Caio Corro, Gaël Lejeune.

Comité de programme

Emanuela Boros, Maximin Coavoux, Caio Corro, Karën Fort, Matthieu Labeau, Gaël Lejeune, Alice Millour, Laure Soulier.

Remerciements

Nous remercions l'ATALA et le laboratoire STIH de Sorbonne Université pour le soutien financier. Nous remercions Patrick Paroubek pour avoir répondu à nos nombreuses questions concernant l'organisation des journées d'étude et l'édition des actes. Nous remercions les membres du comité de programme pour avoir soigneusement relu les résumés soumis. Enfin, nous remercions les auteurs et autrices des résumés qui rendent cette journée d'étude possible.

Programme de la journée

9:00 - 9:30	**Clustering d'entités nommées issues de sorties OCR bruitées : une voie vers la désambiguïsation morphologique automatique ?** Caroline Koudoro-Parfait
9:30 - 10:00	**Robustesse de systèmes de traductions neuronaux pour la traduction anglais-français de syntagmes nominaux complexes en langues de spécialité (médecine et traitement automatique des langues)** Maud Bénard
10:00 - 10:30	**Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français** Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier & Aurélie Névéol
	Pause café (Hall du 2ème étage)
11:00 - 11:30	**Le rapport signal/bruit dans les corpus tirés du web** Adrien Barbaresi & Gaël Lejeune
11:30 - 12:30	Présentation invitée **Le Syndrome du Jabberwocky à l'ère des larges modèles de langues et autres BERTeries : analyse morpho-syntactique en environnement hostile** Djamé Seddah
	Buffet (Hall du 2ème étage)
14:00 - 15:00	Présentation invitée **Reconnaissance d'entités nommées : des documents modernes aux documents historiques, des documents propres aux documents bruyants** Emanuela Boros
15:00 - 15:30	**Impact of Word Splitting on the Semantic Similarity between Contextualized Word Representations** Aina Garí Soler, Matthieu Labeau & Chloé Clavel
	Pause café (Hall du 2ème étage)
15:50 - 16:20	**Améliorer la qualité de l'OCR dans la TGB pour la tâche de REN** Ljudmila Petkovic
16:20 - 16:50	**Portabilité des algorithmes de phénotypage : le cas de la polyarthrite rhumatoïde dans le dossier patient informatisé en français** Thibaut Fabacher, Erik André Sauleau, Noémie Leclerc Du Sablon, Hugo Bergier, Jacques Eric Gottenberg, Adrien Coulet & Aurélie Névéol
16:50 - 17:20	**La robustesse de la traduction neuronale : les systèmes de traduction automatique neuronale à l'épreuve de la reproductibilité de l'expérience** Guillaume Wisniewski, Lichao Zhu, Jean-Baptiste Yunès & Nicolas Ballier
17:20 - 17:30	**Le mot de la fin** Caio Corro & Gaël Lejeune

Table des matières

Évaluation de la tâche de clustering pour l'alignement de formes contaminées d'entités nommées issues d'un corpus OCR bruité

Caroline Koudoro-Parfait[1,2,3]

[1]ObTIC, Sorbonne Université, 1 Rue Victor Cousin, 75005 Paris, France
[2]STIH, Sorbonne Université, 1 Rue Victor Cousin, 75005 Paris, France
[3]SCAI, Campus Pierre et Marie Curie, 4 place Jussieu 75005 Paris, France
caroline.parfait@sorbonne-universite.fr

L'extraction d'informations de grands volumes de données issus de la numérisation et de la reconnaissance optique de caractères (OCR) suscite la réflexion quant à la possibilité de récupérer des informations exploitables scientifiquement de ces données bruitées. Le bruit désigne dans ce cas toutes les erreurs produites par le système OCR : l'insertion, la suppression, mais aussi la substitution d'un ou plusieurs caractères par d'autres. Les chercheurs sont ainsi confrontés aux difficultés d'appliquer des outils informatiques généralement entraînés sur des données textuelles correctement orthographiées (Eshel et al., 2017), à des données textuelles moins standardisées. Un des remèdes consiste à corriger les données délivrées par l'OCR, idéalement automatiquement, avant de les soumettre à un outil de traitement automatique du langage (TAL). Or, si certaines erreurs produites par les dispositifs d'OCR sont systématiques (Stanislawek et al., 2019), lorsqu'il s'agit d'erreurs singulières cet exercice devient difficile à effectuer, en outre la correction peut, elle aussi, produire des erreurs (Huynh et al., 2020).

La Reconnaissance d'entités nommées (REN) est une des tâches de TAL permettant d'extraire des connaissances de ces vastes corpus de textes ((Chiron et al., 2017) et (Linhares Pontes et al., 2019)). En effet, les entités nommées (EN) et particulièrement les EN de lieux (van Strien et al., 2020) constituent la majorité des requêtes formulées par les utilisateurs. L'identification des EN serait un moyen efficace d'améliorer l'accès aux données. Dès lors, la question principale concerne l'évaluation de l'incidence des erreurs d'OCR sur la REN spatiale (Hamdi et al., 2020), et l'influence de ce bruit sur les usages consécutifs (van Strien et al., 2020) de ces données. Dans le même temps, (Koudoro-Parfait et al., 2021) produisent une analyse automatique des sorties d'outils de REN tels que spaCy (Honnibal and Montani, 2017) et stanza(Qi et al., 2020). Leurs analyses s'appuient sur le corpus français de la collection ELTeC - European collection of literary texts [1] et comparent la REN sur cette version de référence, aux sorties obtenues par transcription OCR. Ils démontrent que certains outils de REN prêts à l'emploi sont plutôt robustes face aux variabilités auxquelles les systèmes sont confrontés : contextes et EN *contaminés* (terme proposé par (Hamdi et al., 2022)) par différentes erreurs d'OCR. Du fait des problèmes d'alignement entre les EN de la version de référence et celles de la version OCR, cette analyse reste limitée à un point de vue global, qui ne permet pas de rapprocher les formes contaminées des formes de référence. Par la suite, (Koudoro-Parfait et al., 2022) proposent de produire une analyse automatique plus précise en utilisant deux méthodes pour aligner des entités contaminées à l'EN de référence, c'est-à-dire aligner des EN qui ont la même graphie ou une graphie proche, par exemple aligner "*Saint-Nizier*" avec "*Saint-Nizier.n*" et "*Saint-Nizierl*". Dans un premier temps, ils utilisent le système Nerval [2] pour l'alignement d'EN bruitées, qui s'appuie sur une distance de Levenshtein. Après avoir déterminé que cet outil, bien que plutôt efficace, possède quelques biais, ils proposent une solution utilisant une distance Cosinus moins coûteuse en temps de calcul. La tâche d'alignement des

1. European collection of literary texts : `https://www.distant-reading.net/eltec/`
2. `https://gitlab.com/teklia/nerval`

EN, nous interroge quant aux divers algorithmes de calcul de distance (Bray-Curtis, Cosinus, Jaccard) et la méthode du clustering de données en particulier ((Lin, 1998) et (Green et al., 2012)). Le clustering de données textuelles fournit une représentation numérale et condensée de celles-ci (Loustau, 2013) et pourrait être appliqué sur des données bruitées (Brunet and Loustau, 2013). L'usage de cette méthode permettrait de regrouper les formes contaminées d'une même EN sans avoir accès à l'EN de référence. Nous menons notre expérience de clustering sur les résultats de SPACY, STANZA, SEM [3] (Dupont and Tellier, 2014) et CASEN [4] (Maurel et al., 2011). Le recours à ces deux derniers outils vient consolider la réflexion autour des usages de *systèmes clé en main* pour le français. Nous utilisons le corpus français [5] constitué par (Koudoro-Parfait et al., 2021). Ce corpus comprend une dizaine de textes de références en français et leurs versions OCRisées avec kraken [6] et le modèle français de Tesseract [7] (Smith, 2007).

Notre contribution propose une évaluation manuelle de la tâche de clustering que nous envisageons comme une étape d'aide à la prise de décision (Olteanu, 2013), dans un cas d'usage pour lequel un utilisateur chercherait à trier des EN de manière plus efficace qu'un traitement occurrence par occurrence ou même forme par forme. Idéalement, l'usager procéderait par groupe comme suit : (i) un cluster dont le *centroid* ne serait pas une EN pourrait être laissé de côté, (ii) un cluster dont le *centroid* est bien une EN pourrait être pré-sélectionné avant d'être re-filtré par l'utilisateur. Dans un premier temps, nous avons utilisé l'algorithme de *propagation d'affinité* (Frey and Dueck, 2007). Cet algorithme est intéressant car il ne demande pas de déterminer a priori le nombre de clusters attendu. Il est apparu que les premiers résultats obtenus n'étaient pas concluants, les termes à l'intérieur d'un même cluster n'étaient pas suffisamment similaires. Nous avons procédé à des paramétrages tels que la vectorisation préalable des termes en bigramme de caractères, puis la comparaison des vecteurs avec une distance Cosinus au lieu de la distance de Levenshtein proposée par le système. Une fois les clusters obtenus pour chaque version des textes, nous les avons annotés comme suit : Vrai Positif (VP), Faux Positif (FP), Ambigu (AMBIG). Les termes annotés comme ambigus sont en fait de deux sortes : soit le terme est ambigu parce qu'il pourrait être une forme très contaminée d'une EN ("tiolange" qui pourrait désigner l'église "*Sainte-Solange*"), soit parce qu'il est susceptible d'appartenir à une EN composée de plusieurs termes, mais qu'il n'en est qu'un des morceaux ("*solange*" pour "Sainte-Solange"). L'annotation a révélé un autre problème, le *centroid* choisi par l'algorithme est un FP, néanmoins son cluster comprend une ou plusieurs entités VP. Nous nous questionnons sur la manière de procéder à l'identification automatique d'une entité qui soit un VP comme référent du cluster. Les représentations graphiques des clusters de chacune des versions des textes, nous ont permis de constater que plus une version OCR est bruitée, plus le système de REN va annoter des termes comme étant des EN, même s'ils n'en sont pas (FP). Par ailleurs, il est notable qu'une majorité des entités nommées, qui le sont effectivement (VP), trouvées par le système sur le texte de référence le seront aussi sur les versions OCR. Les mêmes graphiques permettent aussi de mettre en évidence que certains FP sont tout à fait identifiables, car la distance Cosinus entre eux et le *centroid* référent du cluster est très forte. À partir de l'étude de ces résultats nous pouvons établir un critère d'élimination des FP, une distance limite au-delà de laquelle l'EN récupérée est avec certitude un terme sans rapport avec une EN de référence.

Références

[Brunet and Loustau2013] Camille Brunet and Sébastien Loustau. 2013. The algorithm of noisy k-means. working paper or preprint, August.

[Chiron et al.2017] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries Towards a better

3. https://github.com/YoannDupont/SEM et https://www.lattice.cnrs.fr/sites/itellier/SEM.html
4. https://tln.lifat.univ-tours.fr/version-francaise/ressources/casen
5. https://github.com/These-SCAI2023/NER_GEO_COMPAR
6. https://github.com/mittagessen/kraken
7. https://github.com/tesseract-ocr/tesseract

access to information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, Canada, June. IEEE.

[Dupont and Tellier2014] Yoann Dupont and Isabelle Tellier. 2014. Un reconnaisseur d'entités nommées du français. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, pages 40–41, Marseille, France, July. Association pour le Traitement Automatique des Langues.

[Eshel et al.2017] Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text.

[Frey and Dueck2007] Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814) :972–976.

[Green et al.2012] Spence Green, Nicholas Andrews, Matthew R. Gormley, Mark Dredze, and Christopher D. Manning. 2012. Entity clustering across languages. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 60–69, Montréal, Canada, June. Association for Computational Linguistics.

[Hamdi et al.2020] Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition. In *Digital Libraries for Open Knowledge 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings*, pages 87–101, August.

[Hamdi et al.2022] Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2022. In-Depth Analysis of the Impact of OCR Errors on Named Entity Recognition and Linking. *Natural Language Engineering*, March.

[Honnibal and Montani2017] Matthew Honnibal and Ines Montani. 2017. spacy 2 : Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1) :411–420.

[Huynh et al.2020] Vinh-Nam Huynh, Ahmed Hamdi, and Antoine Doucet. 2020. When to Use OCR Post-correction for Named Entity Recognition ? In *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*, pages 33–42, November.

[Koudoro-Parfait et al.2021] Caroline Koudoro-Parfait, Gaël Lejeune, and Glenn Roe. 2021. Spatial named entity recognition in literary texts : What is the influence of OCR noise ? In Ludovic Moncla, Carmen Brando, and Katherine McDonough, editors, *GeoHumanities@SIGSPATIAL 2021 : Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, Beijing, China, November 2 - 5, 2021*, pages 13–21. ACM.

[Koudoro-Parfait et al.2022] Caroline Koudoro-Parfait, Gaël Lejeune, and Richy Buth. 2022. Reconnaissance d'entités nommées sur des sorties ocr bruitées : des pistes pour la désambiguïsation morphologique automatique (resolution of entity linking issues on noisy ocr output : automatic disambiguation tracks). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, pages 45–55.

[Lin1998] Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING 1998 Volume 2 : The 17th International Conference on Computational Linguistics*.

[Linhares Pontes et al.2019] Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, and Antoine Doucet. 2019. Impact of OCR Quality on Named Entity Linking. In *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia, November.

[Loustau2013] Sébastien Loustau. 2013. Anisotropic oracle inequalities in noisy quantization.

[Maurel et al.2011] Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol, and Damien Nouvel. 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Revue TAL*, 52(1) :69–96.

[Olteanu2013] Alexandru Liviu Olteanu. 2013. *On clustering in multiple criteria decision aid : theory and applications*. Theses, Télécom Bretagne, Université de Bretagne Occidentale, June.

[Qi et al.2020] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza : A python natural language processing toolkit for many human languages.

[Smith2007] Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

[Stanislawek et al.2019] Tomasz Stanislawek, Anna Wróblewska, Alicja Wójcicka, Daniel Ziembicki, and Przemyslaw Biecek. 2019. Named entity recognition - is there a glass ceiling ? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 624–633, November.

[van Strien et al.2020] D. van Strien, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *In Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1 : ARTIDIGH*, pages 484 – 496.

Robustesse de systèmes de traductions neuronaux pour la traduction anglais-français de syntagmes nominaux complexes en langues de spécialité (médecine et traitement automatique des langues)

Maud Bénard[1]

[1]Université de Paris

Si la traduction neuronale s'est imposée depuis les années 2010 grâce à un bond qualitatif indiscutable et une réduction de certains efforts de post-édition (Bentivogli et al. 2016), elle n'est pas exempte d'erreurs (Castilho et al. 2017 ; Esperança-Rodier, Becker 2018 ; Burlot, Yvon 2018). Dans le domaine scientifique, l'anglais est la langue privilégiée pour les publications et les rencontres internationales. Or, le discours scientifique anglais se caractérise par le recours important et croissant aux syntagmes nominaux complexes (syntagmes nominaux comprenant un nom tête et un ou plusieurs modifieurs), en particulier dans les articles de recherche (Biber, Conrad 2009 ; Biber, Gray 2016 ; Gledhill, Pecman 2018). Leur complexité peut être accrue par la multiplication des éléments de complexification (longueur comme dans « the end-to-end neural MT system » ou « the so-called computer-assisted translation (CAT) framework » ; enchâssements des pré- et postmodifications comme dans « the whole syntactic and semantic information of the phrasal translation rules ») pour un même groupe considéré (Rouleau 2006 ; Berlage 2014). Ces particularités d'usage et de construction constituent un obstacle important à leur compréhension, leur production et leur traduction (Chuquet, Paillard 2002 ; 2008 ; Biber, Conrad 2009 ; Kübler et al. 2022), en particulier dans le domaine médical (Maniez 2007 ; 2008). De plus, si le discours scientifique a longtemps été considéré comme unique et monolithique, il est aujourd'hui reconnu que des variations existent selon les disciplines et les contextes de communication (Halliday, Martin 1993 ; Hyland 2009). Cela est particulièrement vrai pour les articles de recherche (Fløttum et al. 2006 ; Van Bonn, Swales 2007 ; Fløttum et al. 2013), et la construction des syntagmes nominaux complexes (SNC) n'échappe pas à cette tendance (Biber et al. 1999 ; Biber, Gray 2016). Dans ce cadre, une analyse de la capacité des systèmes de TA à traiter des SNC tirés de textes spécialisés réels prend tout son sens. L'étude que nous menons dans le cadre de cette thèse repose sur une analyse de corpus d'articles de recherche, traduits simultanément par plusieurs systèmes. Pour le domaine médical, quatre systèmes sont comparés : un premier système généraliste et le même système entraîné dans le domaine médical (issus de travaux de recherche universitaire), un second système entraîné (issu d'un autre travail de recherche universitaire) et un système généraliste grand public (Systran). Pour le domaine du TAL, une étude préliminaire porte plus spécifiquement sur une analyse diachronique (2019 et 2022) de deux systèmes génériques accessibles en ligne (Systran et DeepL). L'objectif final consiste notamment à évaluer la qualité d'un système entraîné sur des corpus spécialisés par rapport à un système généraliste. Nous proposons donc de présenter la méthodologie et les principaux résultats de cette analyse comparative, à savoir la typologie et le nombre des erreurs produites par les systèmes de TA — qu'ils soient génériques ou spécialisés — (ex. : analyses en constituants erronées, ajouts ou suppressions injustifiées, erreurs terminologiques…), et si des spécificités semblent émerger en lien soit avec le degré de spécialisation des systèmes, soit avec des traits distinctifs de construction des SNC dans chaque langue de spécialité (ex. : longueur, type de modifieurs, constructions syntaxiques…).

Notre étude préliminaire reposant sur deux systèmes de TA génériques a ainsi mis en évidence que la dispersion des 16 types d'erreurs identifiées varie entre les systèmes et qu'un même type d'erreurs peut résulter de difficultés de traductions spécifiques à un système (ex. : approche de la traduction des sigles...).

Références

BENTIVOGLI, Luisa, BISAZZA, Arianna, CETTOLO, Mauro et FEDERICO, Marcello, 2016. Neural versus Phrase-Based Machine Translation Quality : a Case Study. In : Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing [en ligne]. Austin, Texas : Association for Computational Linguistics. novembre 2016. pp. 257-267. Disponible à l'adresse : `http://aclweb.org/anthology/D16-1025`.

BERLAGE, Eva, 2014. Noun Phrase Complexity in English. United Kingdom : Cambridge University Press. Studies in English Language.

BIBER, Douglas et CONRAD, Susan, 2009. Register, Genre, and Style. S.l. : Cambridge University Press. Cambridge Texbooks in Linguistics.

BIBER, Douglas et GRAY, Bethany, 2016. Grammatical Complexity in Academic English : Linguistic Change in Writing. S.l. : Cambridge University Press. Studies in English Language.

BIBER, Douglas, JOHANSON, Stig, LEECH, Geoffrey, CONRAD, Susan et FINEGAN, Edward, 1999. Chapter 8 : Complex noun phrases. In : Longman Grammar of Spoken and Written English. 7e édition (2011). S.l. : Pearson Longman.

BURLOT, Franck et YVON, François, 2018. Évaluation morphologique pour la traduction automatique : adaptation au français. In : Actes de la conférence TALN 2018 [en ligne]. Rennes, France : s.n. 14 mai 2018. pp. 61-74. Disponible à l'adresse : https ://project.inria.fr/coriataln2018/fr/biblio/.

CASTILHO, Sheila, MOORKENS, Joss, GASPARI, Federico, SENNRICH, Rico, SOSONI, Vilemini, GEORGAKOPOULOU, Panayota, LOHAR, Pintu, WAY, Andy, VALERIO MICELI BARONE, Antonio et GIALAMA, Maria, 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. 2017.

CHUQUET, Hélène et PAILLARD, Michel, 2002. Approche linguistique des problèmes de traduction. S.l. : Éditions Ophrys.

ESPERANÇA-RODIER, Emmanuelle et BECKER, Nicolas, 2018. Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs. In : 4ème journée Traitement Automatique des Langues et Intelligence Artificielle [en ligne]. Nancy : s.n. juin 2018. Disponible à l'adresse : `https://pfia2018.loria.fr/wp-content/uploads/2018/06/Talia-Esperan%c3%a7a-Rodier_Becker.pdf`.

FLØTTUM, Kjersti, DAHL, Trine, DIDRIKSEN, Anders et GJESDAL, Anje, 2013. KIAP – reflections on a complex corpus. In : Bergen Language and Linguistics Studies [en ligne]. 10 avril 2013. Vol. 3. DOI 10.15845/bells.v3i1.367.

FLØTTUM, Kjersti, DAHL, Trine et KINN, Torodd, 2006. Academic Voices : Across Languages and Disciplines. John Benjamins. S.l. : s.n. Pragmatics & Beyond, 148.

GLEDHILL, Christopher et PECMAN, Mojca, 2018. On alternating pre-modified and post-modified nominals such as aspirin synthesis vs. synthesis of aspirin : Rhetorical and cognitive packing in English science writing. In : Fachsprache : Internationale Zeitschrift für Fachsprachenforschung- didaktik und Terminologie. 2018. Vol. 40, n° 1, pp. 24-46.

HALLIDAY, M.A.K. et MARTIN, J.R., 1993. Writing Science. Reprinted 1996. London : The Falmer Press.

HYLAND, Ken, 2009. Academic Discourse : English In A Global Context [en ligne]. S.l. : Continuum Discourse.

KÜBLER, Natalie, MESTIVIER, Alexandra et PECMAN, Mojca, 2022. Using Comparable Corpora for Translating and Post-Editing Complex Noun Phrases in Specialised Texts : Insights from English-to-French Specialised Translation. In : GRANGER, Sylviane et LEFER, Marie-Aude (éd.), Extending the Scope of Corpus-Based Translation Studies. S.l. : Bloomsbury Academic. Bloomsbury Advances in Translation.

MANIEZ, François, 2007. Prémodification et coordination : quelques problèmes de traduction des groupes nominaux complexes en anglais médical. In : ASp. 2007. Vol. 51-52, pp. 71-94. DOI `https://doi.org/10.4000/asp.500`.

MANIEZ, François, 2008. Traduction automatique et ambiguïté syntaxique : le cas de la coordination dans les groupes nominaux complexes en anglais médical. In : 9e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008) [en ligne]. Lyon : s.n. 2008. pp. 765-776. Disponible à l'adresse : `http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/maniez.pdf`.

ROULEAU, Maurice, 2006. Complexité de la phrase en langue de spécialité : mythe ou réalité ? Le cas de la langue médicale. In : Panace@. 2006. Vol. VII, n° 24, pp. 298-306.

VAN BONN, Sarah et SWALES, John, 2007. English and French journal abstracts in the language sciences : Three exploratory studies. In : Journal of English for Academic Purposes. 2007. Vol. 6, pp. 93-108. DOI 10.1016/j.jeap.2007.04.001.

Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français

N. Bannour[1], P. Wajsbürt[2], B. Rance[3,4,5], X. Tannier[2], and A. Névéol[1]

[1]Université Paris-Saclay, CNRS, LISN, 91405 Orsay cedex, France
[2]Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS, 75006 Paris, France
[3]INSERM, CRC, UMRS 1138, Université de Paris, Université Sorbonne Paris Cité, 75006 Paris, France
[4]Assistance Publique - Hôpitaux de Paris, Hôpital Européen Georges Pompidou, 75015 Paris, France
[5]HeKA, Inria Paris, 75006 Paris, France

Contexte – Les Dossiers Électroniques Patient (DEPs) présentent un fort potentiel pour améliorer la recherche clinique. Cependant, la plupart des données contenues dans les DEPs sont en format texte brut (Fu et al., 2020). De plus, jusqu'à 80 % des informations cliniques cruciales ne sont disponibles que sous forme de texte non structuré (Escudié et al., 2017; **?**). Dans ce projet, nous abordons l'extraction d'information dans des compte-rendus cliniques en français, qui consiste à identifier des entités médicales tels que Maladie, Anatomie, Médicament, etc. Les modèles d'apprentissage profond offrent de bonnes performances pour cette tâche de Reconnaissance d'entités nommées (REN). Néanmoins, la disponibilité de données d'entraînement cliniques annotées est souvent limitée, en particulier pour les langues autres que l'anglais. En outre, le caractère confidentiel des textes cliniques limite la possibilité d'échange de données entre les institutions. En effet, le partage de données est difficile dans la pratique et est strictement encadré par des réglementations telles que le RGPD [1]. Ainsi, l'adaptation de modèles de REN appris sur des corpus privés à des corpus publics est nécessaire pour permettre le partage d'outils d'extraction d'information clinique. Ce résumé présente des travaux détaillés dans (Bannour et al., 2022).

Objectifs – Dans cette étude, nous étudions l'apprentissage par mimétisme (*Mimic learning*) (Baza et al., 2020) pour la REN dans les rapports cliniques écrits en français, en utilisant à la fois les jeux de données publics et privés. L'idée de l'apprentissage par mimétisme est d'annoter des données publiques non étiquetées à l'aide d'un *modèle enseignant* (*teacher model*) privé qui a été entraîné sur les données sensibles originales. Les données publiques nouvellement étiquetées sont ensuite utilisées pour entraîner des *modèles élèves* (*student models*). Ces *modèles élèves* peuvent être partagés sans révéler les données sensibles d'origine ou exposer le modèle privé directement construit avec ces données. Notre but est de proposer une architecture de modèles préservant la confidentialité des données qui permettent aux institutions hospitalières de générer des modèles partageables, lorsqu'aucun corpus annoté n'est disponible publiquement.

Méthodologie – La Figure 1 illustre l'approche que nous proposons. Les compte-rendus cliniques sensibles sont utilisés pour entraîner un *modèle enseignant*. Plusieurs études (Chang and Li, 2018; **?**) ont indiqué qu'il est possible de reconstruire approximativement une partie des données d'entraînement

1. `https://gdpr-info.eu/`

en observant simplement les prédictions. Par conséquent, le *modèle enseignant* privé ne sera utilisé que pour produire des annotations *Silver Standard* pour les données publiques, qui seront utilisées pour entraîner les *modèles élèves* partageables. En effet, le *modèle enseignant* restera privé et, comme pour les données sensibles, il ne pourra pas être partagé. Pour générer les *modèles élèves*, on utilise le *modèle enseignant* pour annoter le corpus public non étiqueté. De cette manière, nous créons un nouveau corpus annoté. Ce dernier est utilisé pour entraîner le *modèle élève*. Ce *modèle élève* partageable a pour objectif d'améliorer le transfert de connaissances sans révéler les informations de santé personnelles des patients. A partir d'un *modèle enseignant* entraîné sur le corpus privé MERLOT (Campillos et al., 2018), nous générons trois *modèles élèves* préservant la confidentialité entraînés sur trois corpus publics : DEFT (Cardon et al., 2020), CAS (Grabar et al., 2018) et CépiDC [2]. Pour entraîner ces modèles, nous augmentons les annotations *Gold Standard* dont nous disposons avec des annotations *Silver* générées par le *modèle enseignant*. Nous comparons nos modèles avec trois modèles de référence : le *modèle enseignant* privé, un modèle public entraîné sur un corpus clinique annoté disponible publiquement et un modèle utilisant des dictionnaires construits à partir des bases de connaissance UMLS et JeuxDeMots.

Résultats – Le tableau 1 présente une comparaison de nos *modèles élèves* préservant la confidentialité avec le modèle privé. Bien que les meilleurs résultats soient obtenus avec le *modèle enseignant* privé, avec un score F1 de 0,857, l'utilisation de ce modèle privé pour créer des annotations *Silver* sur le corpus public DEFT/CAS semble être une technique efficace pour améliorer les performances de la REN clinique sur des données publiques. En effet, les *modèles élèves* obtiennent de meilleures performances que les autres modèles publics. Selon le Groupe de travail européen sur la protection des personnes à l'égard du traitement des données à caractère personnel [3], les techniques de préservation de la confidentialité doivent être évaluées selon trois critères : (i) est-il possible d'identifier directement un individu (ii) est-il possible de relier diverses informations qui pourraient conduire à l'identification d'un individu et (iii) est-il possible de déduire des informations relatives à un individu. Nous évaluons ces risques et nous affirmons qu'aucune attaque potentielle ne pourrait révéler des informations sur des données privées sensibles en utilisant les annotations *Silver* générées par le *modèle enseignant* sur des données non sensibles publiquement disponibles. Par conséquent, notre solution offre un bon compromis entre la performance et la préservation de la confidentialité. Mesurer l'impact des expériences menées pourrait être la première étape vers une prise de conscience et un contrôle de leur impact environnemental. Le tableau 1 présente l'empreinte carbone en termes d'équivalent CO2 en grammes. L'émission de CO2 résultant de l'entraînement du *modèle enseignant* et de notre meilleur *modèle élève* CAS est estimée équivalente à 2,52 km parcourus en voiture.

	Précision	**Rappel**	**F-Mesure**	**Équivalent CO_2 (g.)**
Modèle privé (*MERLOT, enseignant*)	0.852	0.862	0.857	123
Modèle public (*DEFT*)	0.592	0.383	0.465	22
Dictionary-based Model (*JDM*)	0.153	0.062	0.089	-
Dictionary-based Model (*UMLS*)	0.246	0.168	0.200	-
Modèle élève (*DEFT*)	0.604	0.743	0.666	30
Modèle élève (*CAS*)	**0.628**	**0.806**	**0.706**	169
Modèle élève (*CépiDc*)	0.580	0.710	0.638	394

TABLE 1 – La performance de nos modèles sur le corpus de test

2. http://www.cepidc.inserm.fr/

3. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

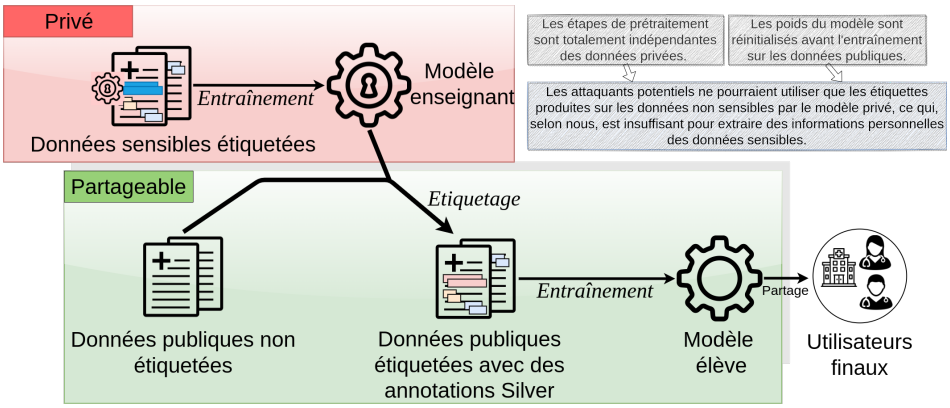


FIGURE 1 – Architecture des modèles préservant la confidentialité des données par mimétisme

Références

[Bannour et al.2022] Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, and Aurélie Névéol. 2022. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130 :104073.

[Baza et al.2020] Mohamed Baza, Andrew Salazar, Mohamed Mahmoud, Mohamed Abdallah, and Kemal Akkaya. 2020. On sharing models instead of data using mimic learning for smart health applications. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 231–236.

[Campillos et al.2018] Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A french clinical corpus with comprehensive semantic annotations : development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2) :571–601.

[Cardon et al.2020] Remi Cardon, Natalia Grabar, Cyril Grouin, and Thierry Hamon. 2020. Présentation de la campagne d'évaluation deft 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France, 6. Association pour le Traitement Automatique des Langues.

[Chang and Li2018] Shan Chang and Chao Li. 2018. Privacy in neural network learning : Threats and countermeasures. *IEEE Network*, 32 :61–67.

[Escudié et al.2017] Jean-Baptiste Escudié, Bastien Rance, Georgia Malamut, Sherine Khater, Anita Burgun, Christophe Cellier, and Anne-Sophie Jannot. 2017. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease : a case study on autoimmune comorbidities in patients with celiac disease. *BMC medical informatics and decision making*, 17(1) :1–10.

[Fu et al.2020] Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J. Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, and Hongfang Liu. 2020. Clinical concept extraction : A methodology review. *Journal of Biomedical Informatics*, 109 :103526.

[Grabar et al.2018] Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium, October. Association for Computational Linguistics.

Le rapport signal/bruit dans les corpus tirés du web

Adrien Barbaresi[1] and Gaël Lejeune[2]

[1]Académie des Sciences de Berlin-Brandenbourg, Jägerstraße 22-23, 10117 Berlin, Allemagne
[2]STIH/CERES, Sorbonne Université, 75006 Paris, France
barbaresi@bbaw.de, gael.lejeune@sorbonne-universite.fr

1 Introduction

Parmi les types de corpus populaires en TAL figurent les corpus issus du web, qui sont notamment au cœur de corpus utilisés en linguistique (Baroni and Ueyama, 2006) ou de modèles de langue à large échelle (Suárez et al., 2019).

Le problème auquel nous nous intéressons ici concerne les données textuelles qui ne sont pas accessibles sous leur forme « pure »ou native, à savoir les textes qu'il faut extraire de documents HTML qui sont conçus pour produire un rendu complexe sur un écran et non pour être directement traitées par une machine, contrairement par exemple à des API qui renvoient du JSON ou des formats de données tel XML. Ce problème est connu de longue date et pose la question de l'évaluation de la capacité des systèmes automatiques à retraduire des informations visuelles en une forme structurée (Gottron, 2007; Baroni et al., 2008).

Ceci nécessite l'utilisation d'extracteurs de contenus, une tâche plus générique souvent nommée *Web Scraping* ou nettoyage de pages Web. Nous nous plaçons dans le cas où la multiplicité des sources, voire la variation des *templates* au sein d'une même source complique de fait l'utilisation de méthodes *ad hoc* à base de règles. Les données textuelles ainsi récupérées ne seront pas exemple de bruit, des segments textuels figurant en trop comme des publicités ou des menus, ni de silence, certains paragraphes pouvant être ignorés par l'extracteur de contenu.

Cette situation impacte directement la robustesse des méthodes de TAL, tant du point de vue des données textuelles que de leur méta-données. Toutes les langues ne bénéficient pas de la même qualité d'extraction, on observe également des différences de « pratiques du Html »selon les régions qui ont des répercussions sur l'utilisabilité des outils. En parallèle, on observe des variations au fil du temps concernant la mise en forme des contenus, de sorte que le meilleur outil à un instant T n'est pas le meilleur outil à $T + 1$ ou $T - 1$.

2 Robustesse des outils en fonction de la variation dans les données traitées

Variation en pays/langues Bien que les outils disponibles librement soient pour la plupart testés sur l'anglais, dans de nombreux cas, ils ne correspondent pas au public concerné, bien plus large. Des étalonnages et comparatifs plus poussés (Barbaresi and Lejeune, 2020) permettent de mettre en lumière des différences conséquentes liées au contexte de production des textes. La langue du contenu configure à elle seule un univers de production et de mise en ligne du texte, notamment à travers des systèmes

de publication de contenu (CMS) dont la popularité et la configuration (par ex. *plugins* WordPress) varient selon les régions géographiques.

JusText (Pomikálek, 2011) représente ainsi une exception car il a été développé et évalué sur du tchèque et non de l'anglais. Il s'avère que JusText ne donne pas des résultats d'aussi bonne qualité sur des textes en anglais ou en français, mais il est en revanche nettement plus adapté au contexte d'Europe centrale.

Par ailleurs, des seuils quantitatifs fixés afin de paramétrer l'extraction perdent de leur pertinence dans d'autres contextes linguistiques (par ex. longueur d'un paragraphe en anglais ou an chinois), ce qui fait que les acquis en termes d'extraction ne sont pas transférables (toutes choses égales par ailleurs). Ceci a notamment été démontré par des évaluations extrinsèques (Lejeune and Zhu, 2018) qui restent toutefois difficiles à mettre en place.

Variation et diachronie Les CMS décrits plus hauts évoluent aussi en fil du temps, tant en termes de popularité que de fonctionnalité. Aux pages saturées d'informations (*frames*, balise HTML `<blink>`) ou de publicités a succédé l'ère du contenu importé d'une multiplicité de sources (*embedding*) et/ou produit/rendu par des scripts (souvent *JavaScript*).

Dans l'ensemble, les pages HTML sont de plus en plus lourdes et invoquent de plus en plus de types de contenus différents. L'outil le plus efficace sur les données du web d'aujourd'hui, ne sera pas nécessairement le plus efficace sur des données du web des années 2000. Un exemple très concret est par exemple que dans les articles de presse des années 2000, les commentaires sont généralement gérés par les producteurs de contenus eux-mêmes là où aujourd'hui, ils sont de plus en plus pris en charge par les réseaux sociaux. Ceci implique que les stratégies permettant de différencier le contenu produit par le journaliste du commentaire du lecteur diffère selon l'évolution des pratiques du web.

3 Impact sur les applications en aval

La ruée vers les corpus tirés du web n'est pas sans susciter des questions quant à la fiabilité et la reproductibilité du travail effectué (Tanguy, 2013). Nous proposons dans ce travail de réaliser et de commenter des expériences qui illustrent les problèmes mentionnés dans le cadre d'une extraction non-supervisée sans vérité de terrain. Ceci rejoint des problématiques évoquées, tant par des chercheurs que par des industriels, lors d'un tutoriel sur le sujet à la conférence TALN (Barbaresi et al., 2021)

Nous nous intéresserons notamment aux différences quantitatives entre une extraction brute comportant tout le texte potentiel d'une page web (html2txt) et une extraction ciblée du texte central ou des éléments centraux. Leur impact s'avère mesurable à travers les fréquences lexicales (comptage de mots, lexicalité), par exemple par un changement du profil de fréquence pour un mot ou parmi les mots les plus fréquents. Le ciblage d'entités nommées permet également de repérer des différences.

Références

[Barbaresi and Lejeune2020] Adrien Barbaresi and Gaël Lejeune. 2020. Out-of-the-box and into the ditch ? multilingual evaluation of generic text extraction tools. In *Web as Corpus workshop WAC-XII, LREC 2020*, pages 5–13.

[Barbaresi et al.2021] Adrien Barbaresi, Emmanuel Giguet, and Gaël Lejeune. 2021. X-COTE– Extraction de Contenus Textuels du Web. In *TALN-RECITAL 2021.*

[Baroni and Ueyama2006] Marco Baroni and Motoko Ueyama. 2006. Building general- and special-purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium, Language corpora : Their compilation and application*, pages 31–40.

[Baroni et al.2008] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. 2008. Cleaneval : a Competition for Cleaning Web Pages. In *Proceedings of LREC*, pages 638–643. ELRA.

[Gottron2007] Thomas Gottron. 2007. Evaluating content extraction on HTML documents. In *Proceedings of the 2nd International Conference on Internet Technologies and Applications*, pages 123–132.

[Lejeune and Zhu2018] Gaël Lejeune and Lichao Zhu. 2018. A new proposal for evaluating web page cleaning tools. *Computación y Sistemas*, 22(4).

[Pomikálek2011] Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora.* Ph.D. thesis, Masaryk University.

[Suárez et al.2019] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Challenges in the Management of Large Corpora (CMLC-7) 2019*, pages 9–16.

[Tanguy2013] Ludovic Tanguy. 2013. La ruée linguistique vers le Web. *Texto ! Textes et Cultures*, 18(4).

Impact of Word Splitting on the Semantic Similarity between Contextualized Word Representations

Aina Garí Soler[1], Matthieu Labeau[1], and Chloé Clavel[1]

[1]LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{*aina.garisoler,matthieu.labeau,chloe.clavel*}*@telecom-paris.fr*

September 2022

Word similarity estimation is a key task in lexical semantics which has become the standard way of intrinsically evaluating the semantic quality of word representations (Landauer and Dumais, 1997; Hill et al., 2015). With the appearance of modern pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), there has been an interest in extracting, analyzing, and using contextualized word representations derived from these models, for example to understand how well they represent the meaning of words (Garí Soler et al., 2019) or to predict diachronic semantic change (Giulianelli et al., 2020). These studies often rely directly on the similarity estimations obtained from these representations.

In this line of research, the units to be studied are most often words, or word instances in context; but most of these models operate at the subword level. Subword tokenization algorithms such as WordPiece (Schuster and Nakajima, 2012; Wu et al., 2016) or Byte Pair Encoding (Sennrich et al., 2016) have advantages over character- or word-based approaches: with a fixed, reasonably-sized vocabulary, models can account for out-of-vocabulary words by simply splitting them into smaller units. At the same time, it has been noted that these tokenization algorithms tend to split words in a way that disregards language morphology (Hofmann et al., 2021), and some of them favor splittings with more subword units than would be necessary (Church, 2020). This can be problematic for out-of-domain and rare words as well as morphologically complex words. In fact, it has been shown that a more morphologically-aware segmentation can improve performance on out-of-context lexical tasks (Hofmann et al., 2021). However, when it comes to word similarity, it is not clear how word splitting affects estimation quality.

A subword vocabulary implies that word representations are not all created equally. When extracting representations from these models, words that are split (*split-words*) need a special treatment, different from words that have a dedicated embedding (*full-words*). The most common approach found in the literature to derive a single vector representation for a split-word is to average the representations of all its subwords (Vulić et al., 2020; Bommasani et al., 2020). Other strategies involve converting the word to lemma form first to avoid word form bias (Laicher et al., 2021) or using the first token only (Martin et al., 2020), but it is still unclear what the best approach is for split-word representation.

Our goal is, then, to look into these two questions which current studies do not yet give an answer to:

- What is the best strategy to combine the contextualized subword representations of a word into a single contextualized, word-level representation for semantic similarity estimation?
- Given such a strategy, how reliable are word similarity estimations involving split-words in comparison to those involving full-words?

We design experiments that allow us to answer these and related questions for the BERT model. First, we carefully build a word similarity dataset relying on a WordNet-based similarity metric (Fellbaum, 1998; Wu and Palmer, 1994). We use this dataset to test different ways of representing split-words. Alongside the now-standard practice of averaging representations of all subwords (`AVG`) and simple strategies such as using the representation of the longest subword (`LNG`), we experiment with alternatives proposed in the literature: CharacterBERT (El Boukkouri et al., 2020), a modified BERT model with a character CNN module intended for building representations for complex tokens in order to improve results in specialized domains; and the FLOTA tokenizer (Hofmann et al., 2022), a simple alternative tokenization method that preserves the morphological structure of words better than the default BERT tokenizer and which can be used at inference time.

We separately investigate the quality of BERT similarity estimations on three kinds of word pairs: those where no word is split by BERT's tokenizer (0-SPLIT), those where only one word in the pair is split (1-SPLIT) and those where the two words are split (2-SPLIT). Contrary to related work which examines representations of words in isolation (Nayak et al., 2020), we extract word instance representations from sentential contexts.

Preliminary experiments have allowed us to make several interesting observations. First, with the default BERT model, `AVG` positions itself as the best strategy for representing split-words. Performance on pairs involving split-words is worse than on 0-SPLIT word pairs, but additional analyses reveal that this can mainly be explained by frequency. Split-words tend to be rarer than full-words, and when controlling for frequency, the quality of similarity estimations obtained with `AVG` is comparable across word pair types. When examining the similarity predictions made with `AVG`, we note that values obtained for 2-SPLIT pairs tend to be higher than 0- and 1-SPLIT similarities. This has important implications for the interpretation of similarity values: similarities across different types of word pairs are not directly comparable, as their values lie in different ranges.

We also investigate the effect of the number of subwords on similarity estimation. Our initial hypothesis, based on the reasoning that representations of shorter subwords are not able to encode lexical semantics as well as longer ones, was that the more subwords a word is split into, the worse the performance will be. Surprisingly, however, results point clearly in the opposite direction: the more subwords are involved in a 1- or 2-SPLIT word pair, the better the predictions. In a split-word, all subword tokens contain "##" (we refer to them as *subpieces*) except for the first one (which we call a *fullpiece*). This is to ensure that the original string can be recovered unambiguously. One difference between words split into few or many pieces is, thus, the ratio of fullpieces to subpieces. We find that, in fact, omitting the first subword is much less detrimental to similarity estimation than omitting other subwords.

In this talk we will describe our experimental setup and preliminary findings in more detail. We will also discuss other important and related questions that we plan to explore, such as the similarity between inflected forms of the same word, the importance of using enough contexts, and the effect that the presence of split-words in the context may have on word instance representations. Investigating split-word representation quality can improve our understanding of the limitations (or not) of models that rely on subword tokenization, especially when these are to be used on a new domain.

Références

[Bommasani et al.2020] Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online, July. Association for Computational Linguistics.

[Church2020] Kenneth Ward Church. 2020. Emerging trends: Subwords, seriously? *Natural Language Engineering*, 26(3):375–382.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[El Boukkouri et al.2020] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

[Fellbaum1998] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

[Garí Soler et al.2019] Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2019. Word Usage Similarity Estimation with Sentence Representations and Automatic Substitutes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 9–21, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[Giulianelli et al.2020] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July. Association for Computational Linguistics.

[Hill et al.2015] Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December.

[Hofmann et al.2021] Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online, August. Association for Computational Linguistics.

[Hofmann et al.2022] Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland, May. Association for Computational Linguistics.

[Laicher et al.2021] Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online, April. Association for Computational Linguistics.

[Landauer and Dumais1997] Thomas K Landauer and Susan T Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review*, 104(2):211.

[Liu et al.2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

[Martin et al.2020] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

[Nayak et al.2020] Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online, November. Association for Computational Linguistics.

[Schuster and Nakajima2012] Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

[Sennrich et al.2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

[Vulić et al.2020] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November. Association for Computational Linguistics.

[Wu and Palmer1994] Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA, June. Association for Computational Linguistics.

[Wu et al.2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint:1609.08144*.

Impact de la correction automatique de l'OCR/HTR sur la tâche de reconnaissance d'entités nommées dans un corpus bruité

Ljudmila Petkovic[1]

[1]ObTIC - Sorbonne Université, 4 place Jussieu, 75005 Paris, France
ljudmila.petkovic@sorbonne-universite.fr

Le volume des données numérisées en lettres et sciences humaines et sociales ne cesse de croître. Face à ce phénomène, la question de leur qualité devient centrale. Dans quelle mesure peut-on directement exploiter ces données ? Doit-on toujours procéder à leur correction, et par quel moyen ? Quel est l'impact d'une telle correction sur les tâches d'extraction d'information et de fouille textuelle ?

Notre objectif dans ce travail vise à améliorer la qualité d'archives numérisées, dans le cadre du projet patrimonial de la Très Grande Bibliothèque (TGB) [1], par le biais des algorithmes d'apprentissage automatique et de traitement automatique des langues (TAL). Les documents ont été numérisés grâce à la technologie OCR [2]. Certains systèmes d'OCR, comme ABBYY ou Calamari, présentent des performances de transcription très solides, dont le taux d'erreur de caractères ne dépasse pas 0.01% (Reul et al., 2018) [3]. Malgré ces performances et le gain du temps de transcription apporté par ces technologies, l'exploitation directe des documents historiques reste problématique à cause du bruit issu de l'OCR (insertion, substitution, suppression des caractères, etc.) lié à la mauvaise qualité des scans, ainsi qu'à la différence entre les polices présentes dans les textes à océriser et dans les données d'entraînement.

Il est alors possible de mener des actions d'amélioration du processus OCR lui-même (par le biais de l'entraînement de modèles d'OCR), voire des actions correctives post-OCR qui nécessitent soit une intervention humaine, soit un traitement informatique. Nous nous plaçons dans le second cadre en expérimentant des outils de correction automatique de la couche texte des documents numérisés, pour faciliter les tâches de TAL en aval, en l'occurrence, la reconnaissance d'entités nommées (REN). En effet, l'extraction d'information, et plus particulièrement la REN nous semble un point d'entrée privilégié pour mesurer la qualité des textes océrisés (Sagot and Gábor, 2014; Koudoro-Parfait et al., 2021; Hamdi et al., 2022).

Notre approche consiste à comparer la qualité de REN effectuée à l'aide de la librairie SpaCy [4] avant et après l'utilisation d'un correcteur automatique d'orthographe (librairie JamSpell) [5] sur des archives

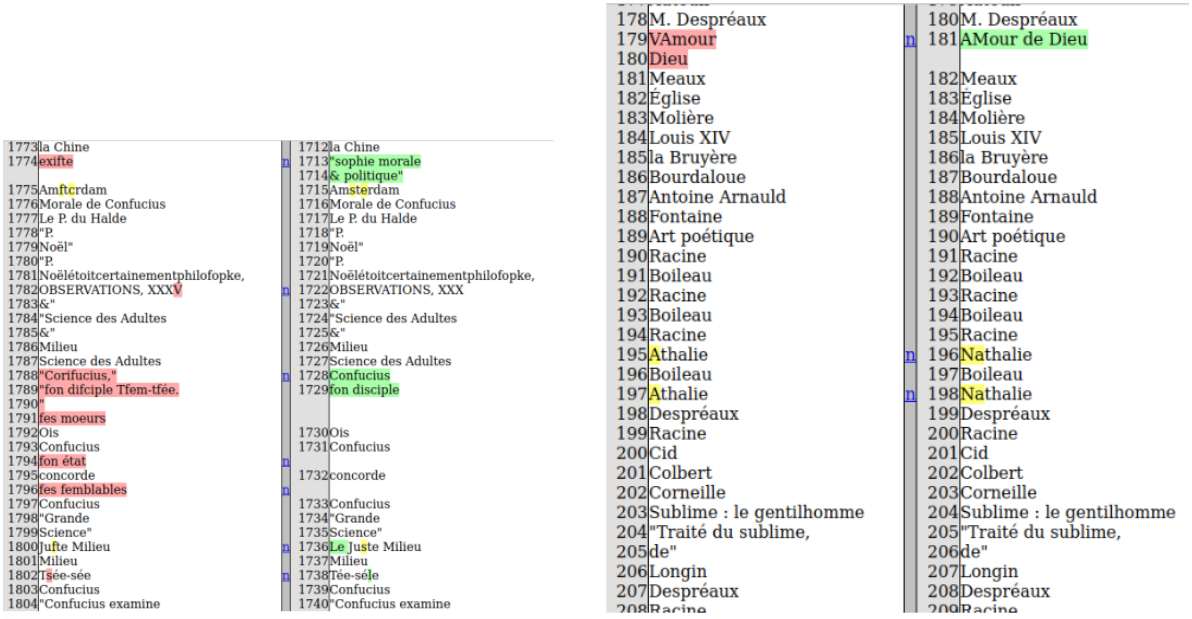
1. `http://obvil.lip6.fr/tgb/`. La TGB est une bibliothèque de 128 441 documents français (imprimés) en mode texte (reconnaissance optique des caractères non relue), issus des collections Gallica de la Bibliothèque nationale de France. Le corpus en XML-TEI provient majoritairement de l'édition du XIXe siècle et couvre différentes thématiques (littérature, histoire, droit, philosophie, etc.).

2. Angl. *Optical Character Recognition*. Nous soulignons également la méthode d'HTR (angl. *Handwritten Text Recognition*) qui se réfère soit à la méthode de reconnaissance du texte manuscrit, soit au traitement des blocs de texte (notamment des lignes) pour extraire du texte `https://datascience.unige.ch/application/files/2316/3248/7621/3.3._Simon_Gabay__Jean-Luc_Falcone.pdf`, contrairement à l'OCR qui se base sur la reconnaissance des caractères individuels. Ici, pour des raisons de brièveté, sous le terme OCR nous englobons les transcriptions automatiques des textes en général, indépendamment de la méthode sous-jacente du modèle (OCR ou HTR).

3. Score obtenu sur les données in-domain issues des documents historiques en écriture Fraktur du 19e siècle.

4. `ttps://spacy.io/models/fr`

5. `https://github.com/bakwc/JamSpell`



(a) Extrait 1.

(b) Extrait 2.

FIGURE 1 – Visualisations des EN et des changements orthographiques dans la sortie HTML de la librairie Difflib. Pour chacune des deux figures : (À gauche) mots non corrigés. À droite : mots corrigés. En rouge : caractères / mots supprimés des fichiers non corrigés ; (En jaune) changements à l'intérieur des mots. En vert : insertions / corrections dans les fichiers corrigés.

océrisées. L'évaluation menée par le biais d'extraction et de visualisation comparative des EN avec la librairie Difflib [6] (Figure 1) nous a permis de repérer celles identifiées avant et après correction, ainsi que de dresser une typologie de modifications apportées par cette correction (insertions, substitutions et suppressions). [7]

Les résultats de notre étude montrent une certaine robustesse du système de REN utilisé sur les textes bruités, ainsi que les avantages et les inconvénients de l'utilisation d'un correcteur orthographique en amont de la tâche de REN. Dans un premier temps, nous avons pu observer manuellement les corrections justes ou les vrais positifs (ex. : Térehce → Térence). D'autres EN ont été identifiées uniquement grâce à la correction automatique (ex. : Eugène). Néanmoins, le correcteur orthographique a effectué de nombreuses sur-corrections, c'est-à-dire des modifications des formes orthographiques déjà correctes dans le texte océrisé (ex. : Empédocle en L'Empésocle). Ces résultats obtenus corroborent d'autres études basées sur des approches différentes de la nôtre (Huynh et al., 2020). Ces auteurs utilisent un correcteur SymSpell [8], et ils affirment que la correction post-OCR pouvait également dégrader les scores F1 de la REN, en particulier lorsque le taux d'erreur OCR est très faible.

Le nombre d'EN reconnues avant et après la correction orthographique des textes est résumé dans le Tableau 1. Pour calculer et évaluer l'impact des corrections orthographiques sur la REN, nous avons construit le tableau de contingence (matrice de confusion, Tableau 2), à partir duquel nous avons dérivé les métriques standards de précision (16%), de rappel (47%) et de mesure F1 (24%).

Les diagrammes à barres de la Figure 2 montrent la répartition des types d'EN extraites. Pour avoir un aperçu plus clair sur la répartition des éléments en question, nous avons exclu les vrais négatifs et

6. `https://docs.python.org/3/library/difflib.html`

7. Voir la visualisation HTML complète `https://github.com/ljpetkovic/corr_OCR_NER/blob/main/diffs/diff_million.html`, ainsi que le tableur des changements `https://github.com/ljpetkovic/corr_OCR_NER/blob/main/diffs/changements_types.csv`

8. `https://github.com/wolfgarbe/SymSpell`

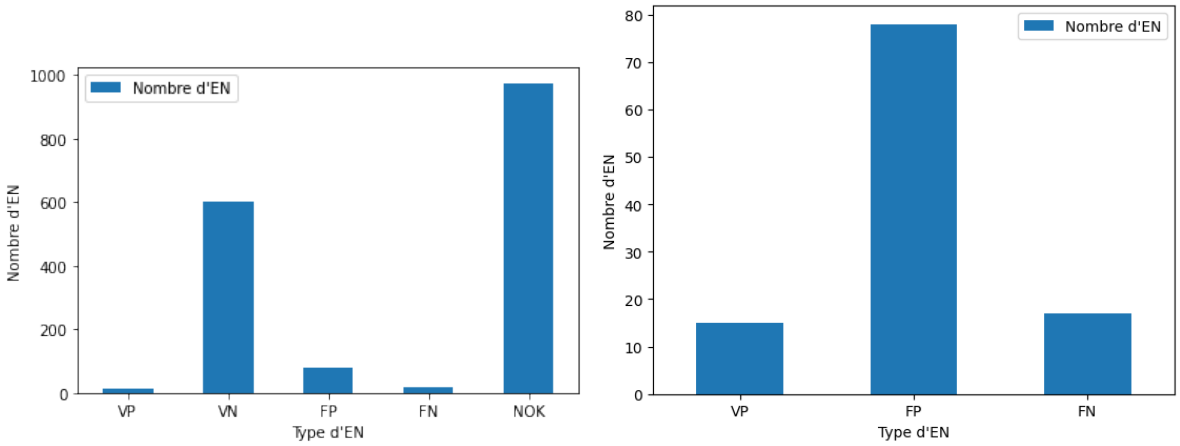
OCR corrigé	EN uniques	EN en total
Non	3918	6428
Oui	3901	6349

TABLE 1 – Le nombre d'EN reconnues avant et après la correction d'OCR.

	Cible		Non-cible		
Corrigés	Vrais positifs	15	Faux positifs	78	
Non corrigés	Vrais négatifs	600	Faux négatifs	17	
En total	Positifs	93	Négatifs	617	710

TABLE 2 – Matrice de confusion qui mesure l'impact des corrections orthographiques sur la REN.

les tokens incorrectement classifiés comme les EN de la deuxième sous-figure.



(a) Diagramme 1.

(b) Diagramme 2.

FIGURE 2 – Répartition des vrais/faux positifs/négatifs et les tokens incorrectement classifiés comme les EN.

Toutes les ressources utilisées dans ce travail sont librement accessibles en ligne. [9] Dans la suite de notre travail, nous aimerions nous pencher sur l'entraînement d'un système de correction post-OCR basé sur des réseaux de neurones (du type NeuSpell [10] ou autre), sur un corpus bien plus grand que celui déjà utilisé.

Références

[Hamdi et al.2022] Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2022. In-depth analysis of the impact of ocr errors on named entity recognition and linking. *Natural Language Engineering*, page 1–24.

[Huynh et al.2020] Vinh-Nam Huynh, Ahmed Hamdi, and Antoine Doucet. 2020. When to use ocr post-correction for named entity recognition ? In Emi Ishita, Natalie Lee San Pang, and Lihong Zhou,

9. https://github.com/ljpetkovic/corr_OCR_NER/tree/main/scripts
10. https://github.com/neuspell/neuspell

editors, *Digital Libraries at Times of Massive Societal Transition*, pages 33–42, Cham. Springer International Publishing.

[Koudoro-Parfait et al.2021] Caroline Koudoro-Parfait, Gaël Lejeune, and Glenn Roe. 2021. Spatial named entity recognition in literary texts : What is the influence of ocr noise ? In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, GeoHumanities'21, page 13–21, New York, NY, USA. Association for Computing Machinery.

[Reul et al.2018] Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. State of the art optical character recognition of 19th century fraktur scripts using open source engines.

[Sagot and Gábor2014] Benoît Sagot and Kata Gábor. 2014. Détection et correction automatique d'entités nommées dans des corpus OCRisés. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, pages 437–442, Marseille, France, July. Association pour le Traitement Automatique des Langues.

Portabilité des algorithmes de phénotypage: le cas de la polyarthrite rhumatoïde dans le dossier patient informatisé en français

T. Fabacher[1,2,3,4], E.-A. Sauleau[1,2], N. Leclerc du Sablon[1], H. Bergier[1], J.-E. Gottenberg[1], A. Coulet[3,4], and A. Névéol[5]

[1]University hospital, Strasbourg, France
[2]Icube Laboratory, Strasbourg, France
[3]Inria Paris, Paris, France
[4]Centre de Recherche des Cordeliers, Inserm, Paris, France
[5]Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France

Introduction Les dossiers patients informatisés (DPI) permettent une réutilisation secondaire des données des hôpitaux, et en particulier la conception et la réalisation d'études cliniques rétrospectives. L'une des premières étapes de ces études cliniques est la définition d'une cohorte de patients qui partagent une caractéristique ou une pathologie particulière. Cette tâche est généralement appelée *phénotypage électronique* (ou phénotypage) et se révèle souvent plus complexe qu'une simple requête par mot clef (Newton et al., 2013; Weng et al., 2020).

Une des difficultés rencontrées pour la définition des cohortes vient de la nature complexe des DPI, qui contiennent des données hétérogènes, incomplètes, structurées et non structurées, sur des périodes de temps de longueur variable et discontinues. Ainsi, la recherche d'un trait phénotypique peut nécessiter la considération à la fois des champs structurés, des textes non structurés et des marqueurs temporels. La composante temporelle est importante à prendre en compte, elle permet de définir un niveau de granularité supplémentaire des traits phénotypique, au niveau par exemple d'un patient, ou d'un séjour. Une autre difficulté vient du fait que les algorithmes de phénotypage peuvent ne pas être bien transférables d'un contexte clinique à un autre. En effet, les variations dans la collecte des données, la pratique clinique, le codage des actes médicaux, les langues font qu'un algorithme de phénotypage développé pour un lieu peut nécessiter une adaptation importante pour être transféré dans un nouveau cadre clinique.

Le phénotypage soulève plusieurs défis en matière de traitement automatique du langage (TAL). Il nécessite l'extraction d'éléments d'information rares à partir de grandes quantités de textes, ainsi que la définitions de phénotypes à partir de ces éléments d'information . L'extraction d'information à partir de texte libre demeure une tâche complexe et cette complexité est majorée parle caractère très spécifique des textes cliniques. En effet, le langage clinique se différencie du langage présent dans les articles de presse ou les romans notamment par la présence de nombreuses négations, nombreuses abréviations, le vocabulaire scientifique utilisé, des phrases ne suivant pas une construction grammaticale classique et l'importante présence de liste d'items. De plus l'aspect très sensible des données utilisées rend leur partage difficile. Il est donc nécessaire de développer des processus transférables et reproductibles (Digan et al., 2020).

Dans notre travail, nous étudions particulièrement la portabilité d'algorithmes de phénotypage de la polyarthrite rhumatoïde (PR), une pathologie auto-immune chronique qui affecte principalement les articulations. Cette pathologie est majoritairement suivie à l'hôpital lors de consultations spécialisées pour les patients complexes. Nous nous intéressons à la PR parce qu'il s'agit d'une pathologie fréquente,

qu'elle est actuellement associée à de nombreuses questions cliniques qui pourraient bénéficier d'outil d'aide à la décision clinique s'appuyant sur le TAL (par exemple prédire le pronostic du patient ou les meilleures options de traitement). De plus, plusieurs algorithmes de phénotypage pour la PR ont été décrits dans la littérature et la question se pose sur leur capacité à être transférable (Carroll et al., 2015; Ferté et al., 2021).

Plus précisément, notre objectif est d'évaluer l'adaptabilité des algorithmes de phénotypage de la PR à un nouvel hôpital, tant au niveau d'un patient que d'un séjour à l'hôpital (hospitalisation ou consultation médicale).

Méthodes Deux algorithmes sont adaptés au contexte du CHU de Strasbourg et évalués à l'aide d'un nouveau corpus de référence de la PR annoté manuellement sur une période allant de 2015 à 2020. Pour l'évaluation des performances des modèles, une cohorte de validation annotée manuellement a été réalisée. Un ensemble de 140 patients a été annoté au niveau de la séjour. Pour chacun des séjours ($\sim$ 1000) de chacun des patients, deux médecins ont annoté si la séjour était en rapport direct ou non avec une PR. Si au moins un des séjours entre 2015 et 2020 est en rapport avec une PR, le patient est considéré comme étant PR+.

Les deux algorithmes sont comparés à un algorithme naïf de phénotypage (appelé Baseline) qui s'appuye seulement sur des mots clés et sur les codes CIM-10 du PMSI. PMSI (Programme de Médicalisation des Systèmes d'Information). Dans le cadre du PMSI, des codes CIM-10 sont attribués à chaque séjour hospitalier pour décrire les pathologies, ce qui donne une information « gros grain »sur la raison principale et les raisons secondaires du séjour d'un patient. Cet algorithme naif est également enrichis par la détection simple du contexte des mots clefs recherchés.La considération du contexte permet de filtrer les variantes hypothétiques, en rapport avec un autre membre de la famille ou négativés des termes recherchés.

Le premier algorithme, appelé Carroll suit une approche supervisée (Carroll et al., 2015) dont le modèle est pré entraîné sur les données d'hôpitaux américains. L'algorithme supervisé consiste à l'utilisation d'une régression logistique qui à partir d'information extraite du DPI donne une probabilité pour chaque patient d'être atteint de PR. Les données comprennent des données de biologies et des codes CIM-9 renseignés de façon structurée et des données extraites des textes cliniques de chaque patient. Pour l'extraction d'information des données à partir de texte libre, un ensemble d'expression régulière à été développé dans les hôpitaux américains. Afin d'adapter l'algorithme au contexte local, une traduction codifie CIM-9 vers code CIM-10 et une traduction des expressions régulière a été nécessaire. Cette traduction des expressions régulières s'est faite en deux temps. Le premier consistait en une traduction littérale des expressions régulières à l'aide de terminologie bilingue (UMLS). Cette étape a été suivie d'une évaluation des expressions sur les données locales et une adaptation de ces expressions par rapport aux données cliniques. Le second algorithme, appelé PheVis, suit une approche semi-supervisée.(Ferté et al., 2021) Il prend en entrée des entités nommées extraites du dpi et des codes CIM-10. Il se base sur une approche semi-supervisée où un premier score simple (à base de règle) est calculé pour l'ensemble des patients et les patients ayant une très haute ou très faible probabilité d'être PR+ sont utilisés pour entraîner un modèle de classification utilisant l'ensemble des données d'entrée. L'extraction des entités des textes cliniques a été faite par une approche à base de dictionnaire (Cossin et al., 2018). Nous proposons également une amélioration dans la définition du silver standard.

Résultats Les algorithmes adaptés offrent des performances comparables entre eux. Pour le phénotypage au niveau du patient sur le nouveau corpus les résultats sont prometteurs (F1 0.71 à 0.79). Mais les performances sont plus faibles pour le phénotypage au niveau du séjour (F1 0.54 à 0.57).

Discussion Malgré des performances légèrement supérieures à des procédures simples de recherche de patient, les deux algorithmes testés presentes des performances similaire. Le gain de performance est à mettre en perspective du coût d'adaptation de ces algorithmes à un nouveau contexte. Le premier algorithme adapté depuis des hôpitaux américains présente une charge d'adaptation plus lourde, car il

Methods	Prec.	NPV	Spe.	Rec.	bal Acc.	Acc.	F1*	AUC*
CIM-10 seul (≥1 code)	0.53	0.90	0.52	0.90	0.66	0.71	0.67 (0.58-0.77)	N/A
Baseline algo.	0.55	0.89	0.58	0.88	0.69	0.73	0.67 (0.58-0.76)	N/A
Baseline algo., plus contexte	0.64	0.76	0.58	0.81	0.72	0.69	0.68 (0.59-0.78)	N/A
Carroll's algo.	0.56	**0.98**	0.55	**0.98**	0.77	0.71	0.71 (0.64-0.80)	0.91 (0.86-0.95)
PheVis (setting a)	0.62	0.90	0.68	0.87	0.76	0.75	0.72 (0.63-0.82)	0.88 (0.82-0.93)
PheVis modifié (setting b)	0.68	0.88	0.77	0.83	0.78	**0.79**	**0.75** (0.66-0.85)	0.85 (0.78-0.92)
Carroll's algo. (Selon Carroll *et al.*(Carroll et al., 2012))	**0.90**	N/A	0.65	N/A	N/A	N/A	N/A	**0.95**
PheVis (Selon Ferté *et al.*(Ferté et al., 2021))	0.65	0.96	**0.94**	0.74	N/A	N/A	N/A	0.94

Table 1 – Performances pour le phénotyping au niveau patient. PheVis setting a is $\omega = 10$, half−life = 365 ; PheVis modifié setting b is $\omega = 2$, half−life = 60. * intervals de confiance calculés par bootstrap.

nécessite un travail manuel pour l'adaptation des règles d'extraction d'information. L'adaptation est indispensable pour obtenir des résultats satisfaisants. Cependant, il est moins gourmand en ressources informatiques que le second algorithme, semi-supervisé. Cet algorithme est cependant plus facilement extrapolable à d'autres environnements (autre clinique, autre langue) et d'autres pathologies.

Références

[Carroll et al.2012] Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, Elizabeth W Karlson, Raul G Perez, Vivian S Gainer, Shawn N Murphy, Eric M Ruderman, Richard M Pope, Robert M Plenge, Abel Ngo Kho, Katherine P Liao, and Joshua C Denny. 2012. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1) :e162–e169, 02.

[Carroll et al.2015] Robert J. Carroll, Anne E. Eyler, and Joshua C. Denny. 2015. Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis, mar.

[Cossin et al.2018] Sebastien Cossin, Vianney Jouhet, Fleur Mougin, Gayo Diallo, and Frantz Thiessard. 2018. IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates. *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, 2125 :94.

[Digan et al.2020] William Digan, Aurélie Névéol, Antoine Neuraz, Maxime Wack, David Baudoin, Anita Burgun, and Bastien Rance. 2020. Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, 28(3) :504–515, 12.

[Ferté et al.2021] Thomas Ferté, Sébastien Cossin, Thierry Schaeverbeke, Thomas Barnetche, Vianney Jouhet, and Boris P Hejblum. 2021. Automatic phenotyping of electronical health record : Phevis algorithm. *Journal of Biomedical Informatics*, 117 :103746.

[Newton et al.2013] Katherine M Newton, Peggy L Peissig, Abel Ngo Kho, Suzette J Bielinski, Richard L Berg, Vidhu Choudhary, Melissa Basford, Christopher G Chute, Iftikhar J Kullo, Rongling Li, et al. 2013. Validation of electronic medical record-based phenotyping algorithms : results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*, 20(e1) :e147–e154.

[Weng et al.2020] Chunhua Weng, Nigam H Shah, and George Hripcsak. 2020. Deep phenotyping : embracing complexity and temporality—towards scalability, portability, and interoperability. *Journal of biomedical informatics*, 105 :103433.

La robustesse de la traduction neuronale : les systèmes de traduction automatique neuronale à l'épreuve de la reproductibilité de l'expérience

Guillaume Wisniewski[1], Lichao Zhu[1], Jean-Baptiste Yunès[1], and Nicolas Ballier[1]

{*guillaume.wisniewski,lichao.zhu,jean.baptiste.yunes,nicolas.ballier*}*@u-paris.fr*

1 Introduction et problématique

Il existe aujourd'hui de nombreuses implémentations de modèles de traduction automatique neuronale (TAN), chacune implémentant différents modèles de traduction et proposant de très nombreux paramètres allant du choix de l'architecture (nombres de couches cachées, taille des représentations, ...) au paramètre de l'algorithme d'apprentissage (choix du pas d'apprentissage, de la méthode d'optimisation, ...). Ces implémentations dépendent également de nombreuses bibliothèques (typiquement `pytorch`, `tensorflow` ou `cuda`) qui évoluent sans cesse et dont le comportement peut parfois changer significativement d'une version à l'autre.

La multiplication de ces paramètres et le nombre d'implémentations disponibles, s'ils traduisent le dynamisme de la recherche en TAN et les différents tâtonnements des chercheurs, soulèvent plusieurs problèmes et limitent la robustesse des expériences : le grand nombre de paramètres complique la description précise des expériences qui ont été réalisées, puisqu'il est difficile de savoir quels sont les paramètres ayant véritablement une influence sur le résultat d'une expérience et dont la valeur doit donc être documentée. Elle limite également l'accès à l'ingénierie de la traduction automatique, les non-spécialistes (typiquement, des traducteurs professionnels) pouvant être désemparés devant le nombre de réglages à effectuer sans véritablement savoir le rôle de chaque paramètre ni l'impact que celui-ci aura sur la qualité des traductions obtenues.

L'objectif de ce travail préliminaire est d'apporter un premier élément de réponse à ces problèmes, en comparant la qualité des traductions obtenues par différents systèmes de traduction automatique afin d'évaluer la robustesse de celles-ci au changement d'implémentation et, de ce fait, à certains choix techniques et à certains paramètres (dont la valeur n'est pas la même dans les différentes implémentations). Pour cela, nous proposons de comparer les performances obtenues par trois systèmes de traduction de l'état de l'art aussi bien à l'aide de métriques automatiques que de manière plus qualitative.

2 Expériences

Conditions expérimentales Pour comparer la robustesse de l'apprentissage d'un système de traduction neuronale, nous considérons trois implémentations d'une architecture `Transformer` aujourd'hui au cœur de tous les systèmes de traduction de l'état de l'art : `JoeyNMT` (Kreutzer et al., 2019), `OpenNMT` (Klein et al., 2017) et `Nematus` (Sennrich et al., 2017). L'expérience que nous proposons au vu de la problématique décrite dans la section précédente consiste à comparer la qualité des traductions obtenues en utilisant une version « sur étagère » de ces implémentations, c'est-à-dire sans régler aucun paramètre, exceptés ceux permettant de décrire le modèle de traduction utilisé. Nous avons utilisé la même architecture que celle utilisée dans le papier introduisant le modèle `Transformer` à savoir : un

	Bleu	chrF2
JoeyNMT		
run 1	41,6	64,5
run 2	41,1	64,4
openNMT		
run 1	37,5	60,7
Nematus		
run 1	43,9	65,8
run 2	44,3	65,9

Table 1: Évaluation sur notre ensemble de test `TedTalk` de la qualité des traductions obtenues par plusieurs implémentations d'un modèle `Transformer`

décodeur et un encodeur composés chacun de 6 couches avec 8 têtes d'attention, une représentation des unités lexicales sur 512 dimensions et une couche *feed-forward* de dimension 2 048. Les trois logiciels que nous avons considérés offrent la possibilité de modifier de nombreux paramètres en plus des 8 que nous venons de décrire. Par exemple, le fichier de configuration « de base » de `JoeyNMT` permet de spécifier 107 paramètres, une quinzaine correspondant à la définition des entrées/sorties (typiquement, nom des fichiers, des modèles, paramètres du tokenizer, fréquence du calcul de l'erreur en validation, ...), une quinzaine décrivant l'architecture `Transformer` à proprement parler et les autres correspondant aux paramètres de l'algorithme d'optimisation (il y a notamment une dizaine de paramètres permettant de configurer le pas d'apprentissage et le *dropout*) ou les paramètres du décodeur (typiquement la taille du faisceau).

Pour cette étude pilote, nous considérons un corpus de traduction de l'anglais vers le français issues du corpus `TED2020` (Reimers and Gurevych, 2020), qui est constitué par les transcriptions des conférences « TED Talks » et des traductions de celles-ci par des volontaires de ce projet (Segal et al., 2015). Ce corpus est divisé en un ensemble d'apprentissage (395 849 phrases pour 8 millions de mots), de validation (2 000 phrases) et de test (2 000 phrases également). Toutes les données ont été segmentées en unités sous-lexicales à l'aide de `SentencePiece` (Kudo and Richardson, 2018) (pour l'entraînement de `JoeyNMT` et `OpenNMT`) et de `Subword-NMT` (Sennrich et al., 2016) (pour l'entraînement de `Nematus`). Nous avons choisi un vocabulaire de 32 000 tokens comme cela se fait habituellement en traduction automatique.

Résultats Nous avons reporté à la Table 1, les scores Bleu obtenus par les différents systèmes sur notre corpus de test, ainsi que les scores chrF2 (Popović, 2015). Ce dernier score correspond à score F_1 calculé sur les 6-grams de caractères. Utiliser une métrique au niveau des caractères permet de réaliser une évaluation qui est (en grande partie) indépendante de la segmentation en mot. Ces deux scores ont été calculés en utilisant SacreBLEU (Post, 2018) sur les hypothèses (et les références !) segmentées en unités sous-lexicales. L'impact de la segmentation sur le calcul du score Bleu a été précisément documenté par (Marie, 2022) et la comparaison des scores Bleu obtenus par deux systèmes différents doit toujours être fait avec précaution.

Nous avons réalisé, pour JoeyNMT et Nematus, deux entraînements afin de mesurer l'impact du caractère aléatoire de la méthode d'optimisation et de pouvoir comparer la variabilité *intra-système* (différence de performance entre deux apprentissages avec une même implémentation) à la variabilité *inter-système* (différence de performance obtenue par deux systèmes de traduction différents). Notons que, étant donné les différents paramètres qui sont choisis aléatoirement lors de l'apprentissage (choix des neurones qui sont ignorés à cause du *dropout*, constitution des batch, ...) deux apprentissage d'un même système peuvent aboutir à des paramètres (et donc des performances) différents, même si, pour assurer la reproductbilité des expériences les systèmes fixent généralement la graine du générateur de

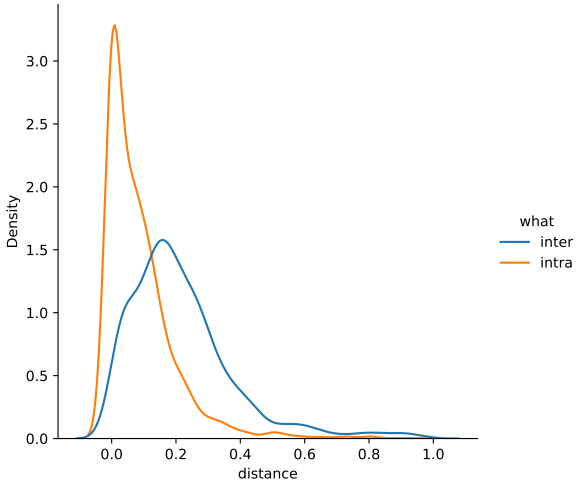


Figure 1: Distribution des distances d'édition intra-système (entre les deux entraînements de JoeyNMT) et inter-système (entre les prédictions de JoeyNMT et de openNMT

nombre aléatoire. C'est notamment le cas pour JoeyNMT : pour mesurer la variabilité intra-système, nous avons du explicitement fixer la graine à des valeurs différentes.

Les résultats de la table 1 montrent que le choix d'une implémentation n'est pas anodin : les scores obtenus par présentent une variabilité forte d'une implémentation à l'autre. Il faut toutefois noter que ces scores ont été obtenus en utilisant les valeurs par défaut de nombreux paramètres et qu'il est probable qu'un réglage plus fin permette de réduire l'écart entre les systèmes. La comparaison des performances obtenues sur les deux entraînements de JoeyNMT montre que les différences de performances observées entre les différents systèmes sont bien dues soit à des différences d'implémentation soit au choix de certaines paramètres et non au caractère aléatoire de l'apprentissage.

Pour obtenir une image plus précise des différences entre les traductions obtenues par les différents systèmes, nous avons également calculé la distance d'édition (au niveau des caractères, pour ne pas dépendre de la segmentation en mots) entre les hypothèses des différents systèmes. La distance d'édition est une mesure de similarité (de « différence » pour être plus précis) entre chaînes de caractères qui peut être interprétée comme le plus petit nombre de caractères à modifier pour transformer une chaîne en une autre. L'objectif de cette deuxième série de mesures est de déterminer si les différences entre les scores Bleu reportées ci-dessus ont un véritable impact sur sur la qualité de la traduction. Nous avons représenté à la figure 1 la distribution des distances d'édition entre les hypothèses prédites par OpenNMT et les deux apprentissages de JoeyNMT. Cette figure montre que, à quelques exceptions, les hypothèses générées ne sont pas trop différentes.

Remerciements Nous remercions le CNRS/TGIR HUMA-NUM et le Centre de calcul IN2P3 (Lyon - France) pour la fourniture des ressources informatiques et de traitement d'une partie des données.

Références

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit

for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China, November. Association for Computational Linguistics.

Taku Kudo and John Richardson. 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*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Benjamin Marie. 2022. Science left behind.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Natalia Segal, Hélène Bonneau-Maynard, and François Yvon. 2015. Traduire la parole: le cas des ted talks. *Revue TAL*, 55:13–45.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April. Association for Computational Linguistics.

Transversalité des méthodes de correction post-ASR et de correction post-OCR

Solveig Poder[1], Cyrille Suire[1], and Antoine Doucet[1]

[1]La Rochelle Université, Laboratoire L3i, 17000, La Rochelle, France

1 Introduction

La correction post-ASR (ASR pour Automatic Speech Recognition) et la correction post-OCR (Optical Character Recognition) sont deux tâches du traitement automatique des langues (TAL) visant à corriger la sortie souvent bruitée d'un système de conversion en texte d'un input donné sous forme d'image (pour l'OCR) ou de son (pour l'ASR). Ces deux tâches ont un objectif similaire et les méthodes proposées sont souvent assez proches. Pourtant, paradoxalement, les littératures scientifiques des deux domaines semblent s'ignorer. L'objectif de notre travail est de tester la robustesse des méthodes proposées dans les deux domaines en vérifiant si ces méthodes pensées pour corriger un certain type d'erreurs (erreurs de speech-to-text ou d'océrisation) sont capables de corriger des erreurs légèrement différentes de celles prévues. Pour cela, nous avons sélectionné quelques méthodes récentes de correction post-ASR et de correction post-OCR, et nous les avons appliquées aux deux tâches en français et en anglais (en les adaptant légèrement si besoin) afin de comparer les résultats obtenus.

2 État de l'art

Avant l'explosion récente du deep learning, aujourd'hui massivement utilisé pour la plupart des tâches de TAL, les méthodes de correction post-ASR étaient nombreuses et variées. On associait généralement un modèle d'erreur (canal bruité, matrice de confusion, ontologie etc.) à un modèle de langue (modèles statistiques, entropie maximale, etc.).

La correction post-ASR est souvent considérée comme une tâche de traduction automatique où la sortie bruitée du modèle d'ASR correspondrait à l'énoncé en langue source et la transcription de référence à l'énoncé en langue cible. Les premiers travaux en ce sens utilisent des systèmes statistiques de traduction automatique, comme (D'Haro and Banchs, 2016), qui utilisent le modèle open-source **Thot**, que nous avons choisi de tester. Aujourd'hui, les modèles neuronaux sont privilégiés (sequence-to-sequence, transformeur, etc.). Très récemment, (Leng et al., 2021) ont implémenté une solution sur Fairseq, appelée *FastCorrect*, qui consiste à utiliser un modèle non autorégressif dans lequel est intégré un module d'alignement permettant de prédire le nombre de tokens cibles correspondant à chaque token source.

Un état de l'art récent de la correction post-OCR a été effectué par (Nguyen et al., 2021). Ils y distinguent deux types d'approches : les approches centrées sur le mot isolé (utilisant lexiques, matrices de confusion, modèle de canal bruité, d'entropie maximale ou de Markov caché...) et celles prenant en considération le contexte (utilisant algorithmes de Machine Learning ou de Deep Learning, par exemple sequence-to-sequence) : certaines de ces méthodes sont communes à la correction post-ASR, mais on y trouve plus d'approches basées sur le caractère comme unité minimale plutôt que sur le mot.

Nous avons testé deux approches proposées par des articles récents. (Poncelas et al., 2020) utilisent un lexique et génèrent une liste de mots candidats pour chaque mot absent de ce lexique puis un

modèle de langue de 5-grammes leur permet ensuite de sélectionner le meilleur candidat. (Schaefer and Neudecker, 2020) proposent une méthode en deux étapes : un détecteur (LSTM bi-directionnel) qui indique pour chaque caractère s'il est erroné ou non, et un traducteur (sequence-to-sequence avec attention, basé sur un modèle de type LSTM) qui corrige les phrases contenant au moins un caractère erroné d'après le détecteur.

Enfin, nous avons décidé de tester BART (*Bidirectional and Auto-Regressive Transformers*), un modèle de langue sequence-to-sequence de type Transformeur pré-entraîné pour reconstituer un texte corrompu, proposé à la fois pour la correction post-OCR par (Soper et al., 2021), qui l'entraînent (*finetuning*) sur des données issues d'un sytème OCR, mais aussi plus récemment pour la correction post-ASR par (Dutta et al., 2022). Ces derniers, qui appellent leur modèle RoBART, ajoutent aux énoncés leur transcription phonétique produite à l'aide d'un outil G2P (grapheme-to-phoneme) et appliquent en dernier lieu l'outil ROVER ((Fiscus, 2000)) permettant de combiner les résultats du modèle d'ASR et du modèle de correction en alignant les deux hypothèses afin de produire un réseau de confusion menant à un rescoring.

3 Données

Données orales transcrites. Pour produire un corpus de transcriptions effectuées par un outil d'ASR, nous avons téléchargé le jeu de données en français du site Common Voice (677 020 fichiers mp3 et l'ensemble des transcriptions de référence dans un fichier tsv). Nous avons effectué la transcription automatique d'une partie des énoncés, à l'aide de la librairie SpeechRecognition de Python, en utilisant l'API de Google. Notre set d'entraînement compte 52 320 énoncés, tandis que nos données de validation et de test comptent respectivement 6 541 et 6 540 énoncés. Nous avons produit un corpus en anglais de taille équivalente en suivant la même procédure.

Données océrisées. Concernant les données océrisées, nous avons utilisé le premier des trois corpus en français fournis pour la compétition de correction post-OCR ayant eu lieu à l'occasion de la 15e édition de la conférence ICDAR (International Conference on Document Analysis and Recognition) de l'année 2019. Il s'agit du dataset HIMANIS Guérin, un ensemble de 1 172 documents datant de 1881 à 1919 numérisés par la Bibliothèque Nationale de France. Le corpus est disponible ici. Nous avons également utilisé le corpus fourni par ICDAR 2019 pour l'anglais (ce dernier est cependant de taille beaucoup plus réduite).

Typologies d'erreurs. Une majorité des erreurs OCR porte sur un caractère, souvent mal reconnu et remplacé par un autre caractère (*drocese* pour *diocèse*), parfois pas reconnu du tout et supprimé (*prmier* pour *premier*), plus rarement ajouté (*leurdite* pour *leurdit*). Il peut, plus rarement, porter sur un bigramme ou trigramme de caractères, ou encore sur des séquences de caractères plus longues (lors d'une forte détérioration du manuscrit), et ponctuellement au-delà de la frontière du mot (notamment dans le cas d'une erreur au niveau des espaces occasionnant un mauvais découpage de mots). Les erreurs ainsi engendrées sont la plupart du temps de nature purement orthographique (qui peuvent s'apparenter à des coquilles et donner lieu à des mots existant ou non).

Les erreurs d'ASR, en revanche, portent souvent sur des mots entiers, des morphèmes, parfois des séquences de plusieurs mots, toujours remplacés par des mots, morphèmes ou séquences de mots phonétiquement proches (*et* pour *est*, *méprisé* pour *méprisées*, *souffrait* pour *s'offrait*...). Il s'agit parfois d'homophones (ou quasi-homophones), parfois de mots à l'usage relativement rare, qui ont été remplacés par des mots plus fréquents (*volume* pour *voilure*). Certaines occurrences sont supprimées (locuteur qui parle trop doucement et n'est pas entendu par l'outil d'ASR, bruit ambiant, mauvaise qualité de l'audio...). Les éléments erronés sont presque toujours orthographiquement corrects (à l'exception de quelques noms propres mal orthographiés) mais engendrent des erreurs de syntaxe ou rendent la phrase incompréhensible.

4 Résultats et conclusion

Quelques résultats. La méthode FastCorrect, conçue pour la correction post-ASR, est efficace pour la tâche de correction post-ASR en français avec une amélioration relative du Word Error Rate (WER) de 5.05% et du Character Error Rate (CER) de 7.97%, mais fait très légèrement baisser la qualité des données océrisées (le WER augmente de 0.07% et le CER de 1.76%).

En revanche, les résultats de la méthode de (Schaefer and Neudecker, 2020), conçue pour la correction post-OCR et utilisant un lexique et un modèle de langue en 5-grammes, sont satisfaisants pour les deux tâches, avec une augmentation du WER de 9.94% pour la correction post-ASR en français et de 13.45% pour la correction post-OCR en français.

Les autres résultats vont globalement dans le même sens.

Conclusion. L'ensemble de nos résultats semble finalement indiquer que, malgré une typologie d'erreurs à corriger sensiblement différente pour les deux tâches, les méthodes de correction post-OCR utilisant le Deep Learning fonctionnent très bien pour la correction post-ASR bien que l'inverse ne soit pas vrai. Cependant, nos données sont imparfaites (tailles et qualités différentes selon les corpus) et ces conclusions restent à confirmer. D'autre part, les métriques classiques ne nous semblent pas pleinement satisfaisantes pour évaluer ces tâches car elles ne tiennent pas compte de la typologie d'erreurs corrigées (dans quelle mesure chaque correction améliore la compréhension du texte) et il faudrait vérifier que les améliorations apportées par la tâche de correction aient réellement un impact intéressant sur la réalisation de tâches subséquentes.

Références

[D'Haro and Banchs2016] Luis Fernando D'Haro and Rafael E. Banchs. 2016. Automatic correction of ASR outputs by using machine translation. In Nelson Morgan, editor, *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 3469–3473. ISCA.

[Dutta et al.2022] Samrat Dutta, Shreyansh Jain, Ayush Maheshwari, Ganesh Ramakrishnan, and Preethi Jyothi. 2022. Error correction in asr using sequence-to-sequence models.

[Fiscus2000] Jonathan Fiscus. 2000. A post-processing system to yield reduced word error rates : Recognizer output voting error reduction (rover). *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 08.

[Leng et al.2021] Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiang-Yang Li, Ed Lin, and Tie-Yan Liu. 2021. Fastcorrect : Fast error correction with edit alignment for automatic speech recognition.

[Nguyen et al.2021] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6), jul.

[Poncelas et al.2020] Alberto Poncelas, Mohammad Aboomar, Jan Buts, James Hadley, and Andy Way. 2020. A tool for facilitating ocr postediting in historical documents.

[Schaefer and Neudecker2020] Robin Schaefer and Clemens Neudecker. 2020. A two-step approach for automatic OCR post-correction. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 52–57, Online, December. International Committee on Computational Linguistics.

[Soper et al.2021] Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021. BART for post-correction of OCR newspaper text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online, November. Association for Computational Linguistics.