Traitement automatique des langues

Traitement automatique des langues et santé

sous la direction de
Aurélie Névéol
Berry de Bruijn
Corinne Fredouille

Vol. 61 - n°2 / 2020

ATALA

Traitement automatique des langues et santé

ATALA

Revue de l'Association pour le Traitement Automatique des Langues

TAL
Vol.
61

n°2
2020

Traitement automatique des langues et santé

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Secrétaire
Peggy Cellier - IRISA, INSA Rennes

Traitement automatique des langues

Volume 61 – n°2 / 2020

TRAITEMENT AUTOMATIQUE DES LANGUES ET SANTÉ

Table des matières

Introduction au numéro spécial « traitement automatique des langues et santé »

Aurélie Névéol* — **Berry de Bruijn**** — **Corinne Fredouille*****

* *Université Paris Saclay, CNRS, LISN,* aurelie.neveol@lisn.upsaclay.fr
** *Conseil national de recherches Canada - Centre de recherche en technologies numériques,* berry.debruijn@nrc-cnrc.gc.ca
*** *LIA, Avignon Université,* corinne.fredouille@univ-avignon.fr

RÉSUMÉ. *À l'heure où l'informatique connaît des changements rapides et où le domaine médical voit émerger de nouvelles opportunités (médecine personnalisée, recherche pharmaceutique) et de nouveaux défis (pandémies, maladies chroniques, vieillissement de la population), les interactions entre ces domaines sont plus pertinentes que jamais. Cet article introduit le numéro spécial «TAL et santé». Après une présentation rapide des problématiques du domaine que nous avons souhaité voir abordées dans ce numéro, nous résumons trois contributions qui présentent différentes facettes du TAL biomédical : la construction de ressources dans des langues autres que l'anglais, la normalisation d'entités et l'analyse de la parole pathologique.*

ABSTRACT. *As computer science goes through rapid changes and the medical field is seeing its own opportunities (precision medicine, drug discovery) and pressures (pandemics, chronic diseases, an aging population), interactions between those fields are more relevant than ever. This article introduces the special issue "NLP and health". After a brief presentation of the challenges in the field that we wanted to see addressed in this issue, we summarize three contributions that present different facets of biomedical NLP: the construction of resources in languages other than English, the standardization of entities, and the analysis of pathological speech.*

MOTS-CLÉS : *traitement automatique de la langue biomédicale, construction de ressources, liaison référentielle, traitement de la parole pathologique.*

KEYWORDS: *biomedical natural language processing, building and evaluating resources, entity normalization, pathological speech processing.*

1. Introduction

Le domaine biomédical et le traitement automatique des langues (TAL) interagissent depuis plus d'un demi-siècle, pour un bénéfice mutuel. Les méthodes de TAL ont en effet contribué à la découverte de connaissances médicales et à l'amélioration de la pratique clinique (Demner-Fushman *et al.*, 2009 ; Velupillai *et al.*, 2018). À l'inverse, le domaine médical a été une source importante de cas d'usages intéressants tant pour le langage écrit qu'oral. Dans le traitement du langage écrit, il a également apporté de vastes collections de documents, comme, par exemple, MIMIC, base de données centralisant plus de 50 000 dossiers de patients (Johnson *et al.*, 2016) ou MEDLINE[1] dédiée aux résumés scientifiques, ainsi que des ressources lexicales détaillées, telles que le Unified Medical Language System, UMLS (Lindberg *et al.*, 1993). Ces vastes collections ont contribué aux progrès dans la discipline du TAL en général (Filannino et Uzuner, 2018).

Par ailleurs, le langage est présent à tous les niveaux du parcours de soins d'un patient, fournissant ainsi autant de champs d'application pour le traitement automatique de la langue. Parmi ces applications, nous pouvons citer la recherche d'information à partir du dossier patient facilitée par les *infobuttons*, des liens contextuels cliquables (Cook *et al.*, 2017), ou l'extraction d'information épidémiologique à partir de multiples sources en ligne, utile par exemple pour la mise à jour des recommandations de santé publique ou de surveillance sanitaire (Carter *et al.*, 2020). L'analyse des productions langagières des patients est également un champ d'application dans le cadre d'une détection précoce de pathologies, pour un accompagnement des cliniciens dans leur diagnostic, mais également dans la prise en charge thérapeutique ou le suivi du patient. Ainsi, des pathologies comme les troubles psychiatriques (Low *et al.*, 2020), et plus spécifiquement, la schizophrénie (Ratana *et al.*, 2019 ; Amblard *et al.*, 2020), les troubles cognitifs affectant des patients atteints de démence (Calzà *et al.*, 2021) ou de la maladie d'Alzheimer (Petti *et al.*, 2020), ou encore les troubles dépressifs (Cummins *et al.*, 2015) sont particulièrement étudiées dans la littérature au travers du TAL (Cummins *et al.*, 2018 ; Voleti *et al.*, 2020).

De nombreux travaux se concentrent sur les agents conversationnels appliqués à la santé, avec des objectifs variés. Certains visent l'aide aux cliniciens dans leur prise en charge du patient ou dans la détection précoce de pathologies (Pacheco-Lorenzo *et al.*, 2020). Une autre part de ces travaux porte sur la formation des cliniciens en termes de prise en charge thérapeutique, de gestion du relationnel avec les patients ou encore de gestion du stress en situation critique. Du point de vue du patient, ces agents conversationnels peuvent également contribuer à leur accompagnement dans la vie quotidienne et intervenir dans leur prise en charge thérapeutique et leur suivi à domicile (Montenegro *et al.*, 2019). Outre les applications citées ci-dessus, ces études comme celles fondées sur l'analyse des réseaux sociaux et des dossiers patients peuvent également contribuer à une meilleure connaissance et compréhension des pathologies concernées (Demner-Fushman et Elhadad, 2016 ; Gonzalez-Hernandez

1. `https://www.nlm.nih.gov/bsd/medline.html`

et al., 2017).

Les productions langagières qui relèvent du domaine de la santé se caractérisent par une grande diversité, au niveau du support – langue écrite ou parlée, au niveau du registre – écrits édités en revues, prises de notes professionnelles dans les documents cliniques, production spontanée sur les réseaux sociaux – ou au niveau de la langue – anglais pour la littérature scientifique, toute langue pour les autres types de documents.

Nous vivons à une époque où l'informatique connaît des changements rapides, que ce soit en termes de production et de diffusion de données numériques massives, *via* les réseaux sociaux ou l'Internet des objets, ou en termes de traitement automatique de ces données incluant l'apprentissage profond (Wu et al., 2020). De même, le domaine médical voit l'émergence de nouvelles opportunités (Grouin et Grabar, 2020) comme l'utilisation secondaire des données de santé, la médecine personnalisée ou la recherche pharmaceutique, mais également de nouveaux défis, tels que les pandémies, les maladies chroniques ou le vieillissement de la population. Ainsi, les interactions entre la médecine et l'informatique sont plus pertinentes que jamais.

L'objectif de ce numéro de traitement automatique des langues (TAL) est de proposer un aperçu des recherches actuelles sur l'ensemble des thématiques liées à la santé, aussi bien dans leurs aspects méthodologiques qu'applicatifs. De nombreux travaux en traitement de la langue médicale ont porté sur des textes en anglais, mais d'autres langues, dont le français, sont également abordées (Névéol *et al.*, 2018a). En effet, lorsqu'il s'agit de littérature scientifique, l'anglais est un choix naturel, car c'est la langue de communication utilisée en science. Pour d'autres tâches, notamment autour des documents cliniques ou générés par les patients, il existe un enjeu fort pour toutes les langues. Aussi, nous présentons des travaux concernant le français dans la section 2 mais soulignons que des travaux similaires existent dans d'autres langues.

2. Ressources et travaux sur TAL et santé en français

En accord avec la réglementation européenne, quelques corpus de langue française, principalement issus de la littérature biomédicale, sont disponibles librement tels que le corpus QUAERO médical du français (Névéol *et al.*, 2014b) qui comporte des annotations en entités et concepts UMLS ou le corpus CAS qui comporte des annotations en entités (Grabar *et al.*, 2018). Le corpus CépiDC, qui rassemble des certificats de décès associés à un codage CIM10 (Lavergne *et al.*, 2016), utilisé lors des campagnes CLEF eHealth de 2016 à 2018 reste disponible avec l'accord de l'Inserm. Cependant, les corpus cliniques tels que le corpus MERLoT (Campillos *et al.*, 2018) et d'autres (par exemple celui décrit par Lerner *et al.* (2020)) restent inaccessibles en dehors du cadre de l'hôpital ayant fourni les données à annoter.

Les ressources termino-ontologiques comme l'UMLS ou encore Wikipédia deviennent de plus en plus multilingues par traduction, transposition ou adaptation à d'autres langues. Ainsi, de nombreuses sources peuvent être agrégées pour rassem-

bler l'ensemble des informations terminologiques médicales disponibles pour le français (Névéol *et al.*, 2014a). Les progrès de la traduction automatique de textes médicaux et la plus grande disponibilité de textes parallèles ont accéléré ce processus (Névéol *et al.*, 2018b).

Les algorithmes de TAL modernes reposent sur des méthodes statistiques et sont donc applicables à toute langue à condition de disposer de données d'entraînement adéquates. Les conditions sont rendues favorables grâce aux corpus mentionnés ci-dessus, ainsi qu'à la disponibilité de modèles préentraînés tels que FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2020) pour le français. Ces ressources ont par exemple été utilisées par les participants à la campagne DEFT avec de bons résultats pour l'extraction d'entités (Copara *et al.*, 2020 ; Wajsbürt *et al.*, 2020).

Ces progrès dans le domaine de la langue française et du traitement multilingue des textes médicaux ont donné lieu à de nombreuses études, avec un intérêt particulier pour divers aspects de l'extraction d'information. Ainsi, une série de travaux a porté sur l'extraction d'information du dossier électronique patient. On note, par exemple, la reconnaissance d'entités avec une méthode intégrant apprentissage profond et ressources terminologiques (Lerner *et al.*, 2020), l'étude de l'impact de divers types de plongements lexicaux obtenus sur des corpus cliniques ou encyclopédiques en français (Neuraz *et al.*, 2020), l'analyse temporelle avec transfert d'architecture de l'anglais vers le français (Tourille *et al.*, 2017) ou encore l'extraction d'information de négation et d'antécédents familiaux de documents cliniques en français (Garcelon *et al.*, 2017). D'autres travaux ont permis une investigation des spécificités langagières propres aux personnes avec schizophrénie à l'aide de méthodes d'apprentissage (Amblard *et al.*, 2020). L'aspect discursif de la langue médicale a également été étudié dans le cadre de systèmes de dialogues pour la formation des médecins (Campillos-Llanos *et al.*, 2020). Une autre série de travaux a porté sur l'extraction d'information des réseaux sociaux. Une étude sur la qualité de vie des patientes atteintes d'un cancer du sein a conduit à l'intégration de nouveaux éléments dans les grilles de qualité de vie utilisées dans les essais cliniques (Nzali *et al.*, 2017). Le domaine de la pharmacovigilance bénéficie également de l'analyse de *posts* de forum patients avec l'extraction d'information concernant le mésusage des médicaments (Bigeard *et al.*, 2018).

3. Contenu du numéro spécial

Ce numéro spécial de la revue TAL contient trois articles qui illustrent la variété des travaux conduits dans le domaine. Ainsi, ils abordent la construction de ressources pour le français médical, le développement de méthodes d'analyse de texte pour la simplification ou la liaison référentielle, ainsi que l'analyse de la parole en français dans un contexte de comparaison entre sujets sains et pathologiques.

Dans *Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français*, Cardon et Grabar s'intéressent à la simplification de textes médicaux en français, un problème qui répond à un enjeu sociétal

fort pour les patients francophones. Les auteurs proposent une méthode permettant de construire un corpus de simplification médicale à partir de textes issus d'articles encyclopédiques, de notices d'informations sur les médicaments et d'articles de la littérature scientifique. La méthode comporte deux étapes : le préfiltrage de paires de phrases candidates à l'alignement selon une heuristique syntaxique suivi d'une classification binaire permettant de distinguer les phrases en relation de simplification. Outre l'évaluation de divers classifieurs non neuronaux, ce travail met à disposition de la communauté un corpus de référence pour la simplification en français. Cette ressource a également vocation à être utilisée pour d'autres applications, comme l'étude de la similarité textuelle.

Robust Multi-pass Sieve for Clinical Concept Normalization, l'article de Wang *et al.*, porte sur la normalisation d'entités, ou liaison référentielle, qui consiste à mettre en correspondance les concepts rencontrés dans le texte libre avec une ressource terminoontologique en support de diverses applications (recherche et stockage d'information, facturation médicale, recherche clinique, études épidémiologiques, etc.). Cette étude particulière s'appuie sur un corpus annoté de documents cliniques en anglais (issu de la campagne n2c2), et sur des vocabulaires standard (SNOMED et RxNorm, par le biais de l'UMLS), et combine diverses approches existantes pour parvenir à des solutions de mises en correspondance, y compris la traduction à travers le français, l'allemand et le chinois. L'un des principaux points forts de ce travail est l'évaluation approfondie, notamment l'analyse des expériences d'ablation qui permet une meilleure compréhension des modèles.

Dans *Détection de la somnolence dans la voix : nouveaux marqueurs et nouvelles stratégies,* Martin, Rouas et Philip présentent deux approches bien distinctes en vue de détecter automatiquement la somnolence chez des patients souffrant de maladies neuro-psychiatriques chroniques à court et à long terme. L'originalité de la première approche repose sur la sélection de marqueurs vocaux, classiquement connus pour caractériser la qualité vocale et ayant la particularité d'être facilement explicables à des non spécialistes de la voix comme les médecins. La deuxième approche tient compte de l'analyse des erreurs de lecture que les patients pourraient commettre en phase de somnolence, notamment lors d'un suivi à long terme. Si la première approche permet d'observer un impact potentiel de la somnolence sur les processus neuromusculaires, la seconde se focalise, pour sa part, sur les processus cognitifs nécessaires à la lecture, ce qui les rend, par conséquent, complémentaires en vue d'une pratique clinique.

Remerciements

Nous remercions le comité éditorial et scientifique de la revue TAL, en particulier Emmanuel Morin, ainsi que le comité scientifique invité, en particulier les relecteurs, qui ont contribué par leur temps et leur expertise à la qualité de ce numéro : Asma Ben Abacha (National Library of Medicine, États-Unis), Gabriel Bernier-Colborne (Conseil national de recherches Canada), Sandra Bringay (LIRMM, Université de Montpellier, France), Leonardo Campillos Llanos (Universidad de Madrid,

Espagne), Jérôme Farinas (IRIT, Université de Toulouse, France), Graciela Gonzalez-Hernandez (University of Pennsylvania, États-Unis), Natalia Grabar (STL-CNRS, Université de Lille, France), Julia Ive (King's College, London, Royaume-Uni), Svetlana Kiritchenko (Conseil national de recherches, Canada), Hongfang Liu (Mayo Clinic, États-Unis), Stan Matwin (Dalhousie University, Halifax NS, Canada), Timothy Miller (Harvard University, États-Unis), Maite Oronoz (Universidad del País Vasco, Espagne), François Portet (LIG, Université de Grenoble, France), Laurianne Sitbon (Queensland University of Technology, Australie), Sumithra Vellupilai (King's College, London, Royaume-Uni), Meliha Yetisgen (University of Washington, États-Unis).

4. Bibliographie

Amblard M., Braud C., Li C., Demily C., Franck N., Musiol M., « Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie », *Actes TALN*, vol. 2, p. 12-26, 2020.

Bigeard E., Grabar N., Thiessard F., « Detection and analysis of drug misuses. A study based on social media messages », *Frontiers in pharmacology*, vol. 9, p. 791, 2018.

Calzà L., Gagliardi G., Rossini Favretti R., Tamburini F., « Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia », *Computer Speech & Language*, vol. 65, p. 101113, 2021.

Campillos L., Deléger L., Grouin C., Hamon T., Ligozat A.-L., Névéol A., « A French clinical corpus with comprehensive semantic annotations : development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT) », *Language Resources and Evaluation*, vol. 52, n° 2, p. 571-601, 2018.

Campillos-Llanos L., Thomas C., Bilinski É., Zweigenbaum P., Rosset S., « Designing a virtual patient dialogue system based on terminology-rich resources : Challenges and evaluation », *Natural Language Engineering*, vol. 26, n° 2, p. 183-220, 2020.

Carter D., Stojanovic M., Hachey P., Fournier K., Rodier S., Wang Y., De Bruijn B., « Global Public Health Surveillance Using Media Reports : Redesigning GPHIN », *Stud Health Technol Inform.*, p. 843-847, Jun 16, 2020.

Cook D. A., Teixeira M. T., Heale B. S., Cimino J. J., Del Fiol G., « Context-sensitive decision support (infobuttons) in electronic health records : a systematic review », *Journal of the American Medical Informatics Association*, vol. 24, n° 2, p. 460-468, 2017.

Copara J., Knafou J., Naderi N., Moro C., Ruch P., Teodoro D., « Contextualized French Language Models for Biomedical Named Entity Recognition », *Actes TALN-DEFT*, p. 36-48, 6, 2020.

Cummins N., Baird A., Schuller B. W., « Speech analysis for health : Current state-of-the-art and the increasing impact of deep learning », *Methods*, vol. 151, p. 41 - 54, 2018. Health Informatics and Translational Data Analytics.

Cummins N., Scherer S., Krajewski J., Schnieder S., Epps J., Quatieri T. F., « A review of depression and suicide risk assessment using speech analysis », *Speech Communication*, vol. 71, p. 10 - 49, 2015.

Demner-Fushman D., Chapman W. W., McDonald C. J., « What can natural language processing do for clinical decision support ? », *J Biomed Inform*, vol. 42, n° 5, p. 760-772, 2009.

Demner-Fushman D., Elhadad N., « Aspiring to unintended consequences of natural language processing : a review of recent developments in clinical and consumer-generated text processing », *Yearb med inform*, vol. 1, p. 224-233, 2016.

Filannino M., Uzuner Ö., « Advancing the state of the art in clinical natural language processing through shared tasks », *Yearb med inform*, vol. 27, n° 1, p. 184, 2018.

Garcelon N., Neuraz A., Benoit V., Salomon R., Burgun A., « Improving a full-text search engine : the importance of negation detection and family history context to identify cases in a biomedical data warehouse », *J Am Med Inform Assoc.*, vol. 24, n° 3, p. 607-613, 2017.

Gonzalez-Hernandez G., Sarker A., O'Connor K., Savova G., « Capturing the patient's perspective : a review of advances in natural language processing of health-related text », *Yearb med inform*, vol. 26, n° 1, p. 214, 2017.

Grabar N., Claveau V., Dalloux C., « Cas : French corpus with clinical cases », *Proc. LOUHI*, p. 122-128, 2018.

Johnson A. E., Pollard T. J., Shen L., Li-Wei H. L., Feng M., Ghassemi M., Moody B., Szolovits P., Celi L. A., Mark R. G., « MIMIC-III, a freely accessible critical care database », *Scientific data*, vol. 3, n° 1, p. 1-9, 2016.

Lavergne T., Névéol A., Robert A., Grouin C., Rey G., Zweigenbaum P., « A dataset for ICD-10 coding of death certificates : Creation and usage », *Proc. BioTxtM*, p. 60-69, 2016.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., « FlauBERT : Unsupervised Language Model Pre-training for French », *Proc LREC*, Marseille, France, p. 2479-2490, 2020.

Lerner I., Paris N., Tannier X., « Terminologies augmented recurrent neural network model for clinical named entity recognition », *J Biomed Inform*, vol. 102, p. 103356, 2020.

Lindberg D. A., Humphreys B. L., McCray A. T., « The Unified Medical Language System », *Methods of information in medicine*, vol. 32, n° 4, p. 281, 1993.

Low D. M., Bentley K. H., Ghosh S. S., « Automated assessment of psychiatric disorders using speech : A systematic review », *Laryngoscope Investigative Otolaryngology*, vol. 5, n° 1, p. 96-116, 2020.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D., Sagot B., « CamemBERT : a Tasty French Language Model », *Proc ACL*, Association for Computational Linguistics, Online, p. 7203-7219, July, 2020.

Montenegro J. L. Z., da Costa C. A., da Rosa Righi R., « Survey of conversational agents in health », *Expert Systems with Applications*, vol. 129, p. 56 - 67, 2019.

Neuraz A., Rance B., Garcelon N., Llanos L. C., Burgun A., Rosset S., « The Impact of Specialized Corpora for Word Embeddings in Natural Language Understanding », *Stud Health Med Inform*, vol. 270, p. 432, 2020.

Névéol A., Dalianis H., Velupillai S., Savova G., Zweigenbaum P., « Clinical natural language processing in languages other than english : opportunities and challenges », *Journal of biomedical semantics*, vol. 9, n° 1, p. 12, 2018a.

Névéol A., Grosjean J., Darmoni S. J., Zweigenbaum P., « Language Resources for French in the Biomedical Domain. », *Proc. LREC*, p. 2146-2151, 2014a.

Névéol A., Grouin C., Leixa J., Rosset S., Zweigenbaum P., « The QUAERO French medical corpus : A ressource for medical entity recognition and normalization », *Proc. BioTextM*, 2014b.

Névéol A., Yepes A. J., Neves L., Verspoor K., « Parallel corpora for the biomedical domain », *Proc. LREC*, p. 286-291, 2018b.

Nzali M. D. T., Bringay S., Lavergne C., Mollevi C., Opitz T., « What patients can tell us : topic analysis for social media on breast cancer », *JMIR Med Inform*, vol. 5, nº 3, p. e23, 2017.

Pacheco-Lorenzo M. R., Valladares-Rodríguez S. M., Anido-Rifón L. E., Fernández-Iglesias M. J., « Smart conversational agents for the detection of neuropsychiatric disorders : A systematic review », *J Biomed Inform*, 2020.

Petti U., Baker S., Korhonen A., « A systematic literature review of automatic Alzheimer's disease detection from speech and language », *J Am Med Inform Assoc.*, vol. 27, nº 11, p. 1784-1797, 2020.

Ratana R., Sharifzadeh H., Krishnan J., Pang S., « A Comprehensive Review of Computational Methods for Automatic Prediction of Schizophrenia With Insight Into Indigenous Populations », *Frontiers in Psychiatry*, vol. 10, p. 659, 2019.

Tourille J., Ferret O., Tannier X., Névéol A., « Temporal information extraction from clinical text », *Proc EACL*, Valencia, Spain, p. 739-745, 2017.

Velupillai S., Suominen H., Liakata M., Roberts A., Shah A. D., Morley K., Osborn D., Hayes J., Stewart R., Downs J. *et al.*, « Using clinical Natural Language Processing for health outcomes research : Overview and actionable suggestions for future advances », *J Biomed Inform*, vol. 88, p. 11-19, 2018.

Voleti R., Liss J. M., Berisha V., « A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders », *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, nº 2, p. 282-298, 2020.

Wajsbürt P., Taillé Y., Lainé G., Tannier X., « Participation de l'équipe du LIMICS à DEFT 2020 », *Actes TALN-DEFT*, p. 108-117, 6, 2020.

Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français

Rémi Cardon* — **Natalia Grabar***

* *UMR 8163 STL – CNRS / Université de Lille, F-59000 LILLE*
{remi.cardon, natalia.grabar}@univ-lille.fr

RÉSUMÉ. *La simplification automatique a pour objectif de produire une version de textes plus facile à comprendre à destination d'un public identifié. Nous nous intéressons à la simplification de textes médicaux. Le plus souvent, le lexique et les règles de simplification sont acquis à partir de corpus parallèles. Comme de tels corpus n'existent pas en français, nous proposons des méthodes pour les construire à partir de corpus comparables. Notre méthode repose sur une étape de filtrage, destinée à ne garder que les meilleures phrases candidates à l'alignement, et une étape d'alignement considérée comme un problème de catégorisation. Il s'agit de décider si une paire de phrases est alignable ou non. Nous exploitons différents types de descripteurs (essentiellement basés sur le lexique et les corpus) et obtenons jusqu'à 0,97 de F-mesure avec les données équilibrées.*

ABSTRACT. *The purpose of automatic simplification is to create version of texts which is easier to understand for a given targeted population. We aim at simplifying medical texts. Usually, lexicon and rules required for the simplification are acquired from parallel corpora. Since such corpora are not available for French, we propose methods for their creation from comparable corpora. Our method relies on filtering step, which purpose is to keep the best sentence candidates for alignment, and alignment step considered as categorization problem. The aim is to decide whether a pair of sentences is alignable or not. We exploit different types of features (mainly issued from lexicon and corpora) and get up to 0.97 F-measure with balanced data.*

MOTS-CLÉS : *simplification automatique, textes médicaux, corpus de phrases parallèles, constitution de ressources.*

KEYWORDS: *automatic simplification, medical texts, corpus with parallel sentences, resource building.*

1. Introduction

La simplification automatique vise à fournir une version simplifiée des textes à destination d'une population donnée. La simplification peut concerner le lexique, la syntaxe, la sémantique mais aussi la pragmatique et l'organisation des textes. La simplification peut être vue comme une aide fournie aux lecteurs ou comme un prétraitement dans les applications de TAL. Dans le cas d'aide aux lecteurs, la simplification vise différents types d'utilisateurs : les enfants (De Belder et Moens, 2010), les personnes non ou mal alphabétisées, les lecteurs étrangers (Paetzold et Specia, 2016), les personnes handicapées ou ayant des pathologies neurodégénératives (Chen *et al.*, 2016), ou les personnes non spécialistes face à des documents spécialisés (Leroy *et al.*, 2013). Dans le cas d'applications de TAL, la simplification produit une version de textes plus facile à traiter par d'autres modules de TAL, comme l'analyse syntaxique (Chandrasekar et Srinivas, 1997 ; Jonnalagadda *et al.*, 2009), l'annotation sémantique (Vickrey et Koller, 2008), le résumé automatique (Blake *et al.*, 2007), la traduction automatique (Stymne *et al.*, 2013 ; Štajner et Popović, 2016), l'indexation (Wei *et al.*, 2014) et la recherche et extraction d'information (Beigman Klebanov *et al.*, 2004).

Au moins deux types de connaissances sont nécessaires pour effectuer la simplification : un lexique pour la simplification au niveau lexical, et un ensemble de règles de transformation pour la simplification syntaxique. De telles ressources peuvent être construites manuellement, provenant de l'expertise d'un spécialiste, ce qui était propre aux premiers travaux de simplification, ou bien être acquises à partir de données réelles, ce qui correspond aux approches actuelles et nécessite de gros corpus parallèles (Nisioi *et al.*, 2017). Deux corpus avec des données en anglais sont fréquemment utilisés : *WikiLarge* (Zhang et Lapata, 2017), un corpus libre d'utilisation qui contient environ 300 000 paires de phrases, et *Newsela* (Xu *et al.*, 2015 ; Hwang *et al.*, 2015). Ce type de corpus n'existe pas en français, alors qu'il pourrait fournir des informations précieuses pour la simplification automatique. Ainsi, les paires de phrases, qui se différencient par leur degré de technicité, peuvent fournir des informations utiles, comme dans les exemples (1), avec le texte technique, et (2), avec le texte simplifié.

(1) *L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine sous le périchondre du pavillon. Il est souvent provoqué par un traumatisme contondant. En l'absence de traitement, il finit par entraîner une difformité couramment appelée oreille en chou-fleur ou oreille du boxeur.*

(2) *L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine dans le pavillon (oreille externe), souvent à la suite d'un traumatisme contondant. S'il n'est pas traité, il entraîne une difformité appelée oreille en chou-fleur ou oreille du boxeur.*

L'objectif principal de notre travail consiste à détecter automatiquement des phrases parallèles, avec le contenu similaire ou identique, au sein de documents monolingues

comparables en français et distingués par leur technicité. À notre connaissance, le seul travail de ce type en français a été effectué avec un alignement de phrases manuel (Brouwers *et al.*, 2014) mais les phrases alignées et les règles de transformation syntaxiques ne sont pas disponibles. Par rapport à nos travaux précédents (Cardon et Grabar, 2019 ; Cardon et Grabar, 2020), nous proposons une méthode plus complète (avec une étape de filtrage et une étape d'alignement), exploitons un ensemble plus riche de descripteurs (basés sur le lexique, les indices formels, la similarité et les corpus), effectuons plus d'expériences en alignement de phrases parallèles, et présentons globalement de meilleurs résultats.

Dans la suite de ce travail, nous présentons d'abord les travaux existants (section 2). Nous présentons ensuite notre approche (sections 3 et 4) et les résultats obtenus (section 5). Nous terminons avec une conclusion et des perspectives (section 6).

2. Travaux existants

Dans les corpus parallèles, l'alignement de phrases parallèles peut se baser sur des indices de surface comme la longueur relative des phrases (Gale et Church, 1993) ou les informations lexicales (Chen, 1993), tandis que dans les corpus comparables, les phrases ont une sémantique relativement proche et, de plus, elles ne sont pas forcément ordonnées de la même manière. D'autres difficultés proviennent du fait que le degré du parallélisme peut varier en allant des corpus presque parallèles, avec beaucoup de phrases parallèles, aux corpus assez éloignés (*very-non-parallel corpora*) (Fung et Cheung, 2004) et que de tels corpus peuvent contenir des données parallèles à différents niveaux de granularité : documents, phrases, segments sous-phrastiques (Hewavitharana et Vogel, 2011). Plusieurs travaux sont positionnés en traduction automatique : les corpus comparables bilingues sont exploités pour créer des corpus parallèles et alignés. Ces travaux requièrent l'utilisation de lexiques bilingues ou de systèmes de traduction automatique et reposent en général sur trois étapes :

1) détection de documents comparables au sein d'un corpus grâce aux métriques de similarité par exemple (Utiyama et Isahara, 2003 ; Fung et Cheung, 2004), ce qui permet de réduire l'espace de recherche de phrases parallèles ;

2) détection de phrases ou de segments candidats à l'alignement en exploitant des systèmes de recherche d'information cross-langue (Utiyama et Isahara, 2003), des arbres d'alignement de séquences (Munteanu et Marcu, 2002) ou des traductions automatiques mutuelles (Yang et Li, 2003 ; Abdul-Rauf et Schwenk, 2009) ;

3) sélection de bonnes propositions en exploitant des classifieurs binaires (Ștefănescu *et al.*, 2012), des mesures de similarité (Fung et Cheung, 2004), le taux d'erreurs (Abdul-Rauf et Schwenk, 2009), des modèles génératifs (Zhao et Vogel, 2002) ou des règles spécifiques (Yang et Li, 2003).

Plus récemment, cette tâche est également explorée dans le contexte monolingue : la similarité sémantique textuelle (*semantic text similarity - STS*) est calculée au niveau de phrases ou de segments sous-phrastiques. Cette tâche a attiré l'attention des cher-

cheurs car ce type d'information fournit des indications précieuses pour la détection du plagiat, les questions-réponses ou la pondération des réponses, par exemple. Ainsi, la compétition *SemEval* propose une tâche dédiée à la similarité sémantique textuelle (Agirre *et al.*, 2013) et poursuit l'objectif suivant : étant donné une paire de phrases, les systèmes doivent prédire si ces phrases sont similaires sémantiquement et leur attribuer un score de similarité allant de 0 (sémantique indépendante) à 5 (sémantique identique). Plusieurs types de méthodes sont exploités par les participants :

– *les méthodes basées sur le lexique*, qui exploitent les chaînes de caractères et de mots ou la traduction automatique (Clough *et al.*, 2002 ; Zhang et Patrick, 2005 ; Nelken et Shieber, 2006 ; Qiu *et al.*, 2006 ; Zhu *et al.*, 2010 ; Zhao *et al.*, 2014) ;

– *les méthodes basées sur les connaissances*, qui exploitent des sources lexicales externes, comme WordNet ou la ressource PPDB avec les paraphrases (Mihalcea *et al.*, 2006 ; Fernando et Stevenson, 2008 ; Lai et Hockenmaier, 2014) ;

– *les méthodes basées sur la syntaxe*, qui exploitent la modélisation syntaxique des phrases (Wan *et al.*, 2006 ; Severyn *et al.*, 2013 ; Tai *et al.*, 2015 ; Tsubaki *et al.*, 2016) ;

– *les méthodes basées sur les corpus*, qui exploitent les modèles distributionnels, LSA, etc. (Barzilay et Elhadad, 2003 ; Guo et Diab, 2012 ; Zhao *et al.*, 2014 ; He *et al.*, 2015 ; Mueller et Thyagarajan, 2016).

Nous nous intéressons à cette tâche car elle permet de construire des ressources (un lexique, des règles de transformation, etc.) utilisables en simplification automatique.

3. Données linguistiques

Nous exploitons un corpus monolingue comparable disponible (section 3.1), les données de référence issues de l'alignement manuel au niveau des phrases (section 3.2) et une liste de mots vides (décrite dans la section 4.1).

3.1. *Corpus monolingue comparable*

Le corpus monolingue comparable[1] contient des textes provenant de trois sources de données : (1) les articles de deux encyclopédies collaboratives disponibles en ligne Wikipédia[2] et Vikidia[3], (2) les informations sur les médicaments de la base publique de médicaments [4] gérée par le ministère de la Santé, (3) les résumés de revues systématiques de la fondation Cochrane[5]. Trois genres sont donc couverts dans ce corpus : les articles d'encyclopédie, les informations sur les médicaments et la littérature scientifique. Dans chaque source et pour un sujet donné, les textes techniques et simplifiés

1. http://natalia.grabar.free.fr/resources.php
2. https://fr.wikipedia.org
3. https://fr.vikidia.org
4. http://base-donnees-publique.medicaments.gouv.fr/
5. http://www.cochranelibrary.com/

sont disponibles. Ce corpus contient plus de 55 M de mots dans la partie technique et plus de 35 M de mots dans la partie simplifiée.

3.2. *Données de référence*

Pour créer les données de référence, nous sélectionnons aléatoirement 14 articles encyclopédiques, 12 médicaments et 13 revues Cochrane. Les documents sont segmentés en phrases. L'alignement est effectué manuellement et indépendamment par deux annotateurs (les auteurs) au niveau de la paire de documents. L'alignement des 39 paires de documents a pris environ 20 heures par annotateur, à quoi s'ajoutent environ 5 heures de consensus où toutes les annotations ont été passées en revue. L'annotation a été menée sans guide. L'accord inter-annotateur (Cohen, 1960) est de 0,76.

Un accord est compté lorsqu'un alignement est proposé par les deux annotateurs, et un désaccord lorsqu'un alignement est proposé par un seul annotateur. Les phrases non alignées ne sont pas considérées. Dans le tableau 1, nous indiquons la taille des données de référence avant (colonne « *brut* ») et après l'alignement (colonne « *aligné* »), ainsi que le taux d'alignement, c'est-à-dire le nombre de phrases d'un corpus qui trouvent un alignement par rapport au nombre total de phrases. Il s'agit des résultats d'alignement consensuel. Les séances de consensus ont été l'occasion d'identifier les types d'alignement conservés et de trouver un accord sur les données utiles pour la simplification. Nous avons observé que les désaccords venaient de la prise en compte ou non de certains types d'alignement, comme les phrases identiques ou l'intersection sémantique. Dans les exemples, la phrase technique est suivie par la phrase simplifiée :

1) les deux phrases, technique et simplifiée, doivent contenir un verbe.

2) les phrases ne sont pas identiques et diffèrent par le lexique ou la morphologie des mots, mais pas uniquement par la ponctuation ou les mots vides. Ces phrases ont une sémantique équivalente : *{Les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler}{Les sondes gastriques sont couramment utilisées pour administrer des médicaments et de la nourriture directement dans le tractus gastro-intestinal (un tube permettant de digérer les aliments) pour les personnes ne pouvant pas avaler}*

3) le sens d'une phrase est intégralement inclus dans le sens de l'autre phrase. Il s'agit de l'inclusion sémantique. Cela permet de repérer les cas de simplification syntaxique (fusion ou découpage de phrases) ainsi que les ajouts et suppressions. Dans cet exemple, la phrase technique indique le nombre de participants et la mesure d'évaluation en plus : *{Peu de données (43 participants) étaient disponibles concernant la détection d'un mauvais placement (la spécificité) en raison de la faible incidence des mauvais placements}{Cependant, peu de données étaient disponibles concernant les sondes placées incorrectement et les complications possibles d'une sonde mal placée}*

4) les cas d'intersection sémantique, où chaque phrase apporte des informations spécifiques propres, sont rejetés. L'intersection sémantique est en effet plus difficile à généraliser pour en dégager des règles de transformation : *{Des études à plus grande échelle sont nécessaires pour déterminer la possibilité d'événements indésirables lorsque les ultrasons sont utilisés pour confirmer le positionnement des sondes}{Des études à plus grande échelle sont nécessaires pour déterminer si les ultrasons pourraient remplacer les rayons x pour confirmer la mise en place d'une sonde gastrique, et pour évaluer si les ultrasons pourraient permettre de réduire les complications graves, telles que la pneumonie résultant d'un tube mal placé}*

		Technique				*Simplifié*				*Alignement*	
		Brut		*Aligné*		*Brut*		*Aligné*		*(%)*	
Corpus	*Doc.*	*Ph.*	*Occ.*	*Ph.*	*Occ.*	*Ph.*	*Occ.*	*Ph.*	*Occ.*	*Tech.*	*Simp.*
Médicaments	12 × 2	4 391	44 684	143	4 227	2 710	27 804	143	8 481	3,25	5,27
Cochrane	13 × 2	426	8 852	84	2 278	227	4 688	84	2 466	19,71	36,56
Encyclopédie	14 × 2	2 416	36 703	39	873	235	2 659	39	710	1,61	16,6

Tableau 1. *Taille des données de référence avec l'alignement consensuel*

Selon le tableau 1, nous pouvons voir que les phrases alignées sont relativement plus rares dans les corpus *Médicaments* et *Encylopédie*, alors que le corpus *Cochrane* en offre plus par rapport à sa taille : le taux d'alignement est entre 1,61 et 36,56. Ceci peut être expliqué par les spécificités des corpus : (1) la ligne directrice de rédaction des versions simplifiées des résumés *Cochrane* affiche explicitement une volonté de simplifier le contenu de ses résumés d'origine pour le grand public. Les rédacteurs prennent donc comme point de départ les résumés techniques et les simplifient ; (2) l'objectif de Vikidia est de traiter des sujets présents dans Wikipédia mais pour un public d'enfants. La création d'articles de Vikidia est rarement basée sur les articles de Wikipédia et, le plus souvent, il s'agit d'une écriture indépendante ; (3) quant au corpus *Médicaments*, en respect avec la législation, les informations sur les médicaments sont créées à destination des professionnels de santé et des patients. Certaines de ces informations sont propres à la version technique (composition plus détaillée, action sur l'organisme, molécules, détail sur les effets indésirables...), alors que d'autres sont propres à la version simplifiée (précautions d'emploi, mises en garde...).

Suite à l'alignement manuel, nous gardons deux types d'alignement :

1) *équivalence sémantique.* Les deux phrases, technique et simplifiée, ont le même sens ou des sens proches, comme dans ces phrases du corpus *Cochrane* : *{Les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler}{Les sondes gastriques sont couramment utilisées pour administrer des médicaments et de la nourriture directement dans le tractus gastro-intestinal (un tube permettant de digérer les aliments) pour les personnes ne pouvant pas avaler}*. Dans le cas d'équivalence sémantique, la simplification est essentiellement effectuée au niveau lexical. Elle repose alors sur la substitution de termes, comme c'est observable à travers les paires *{technique}{simplifié}* :

{alimentation}{nourriture}, *{entérale}{directement dans le tractus gastro-intestinal}*. La simplification peut également être effectuée grâce à l'ajout d'informations et, dans ce cas, les notions complexes sont suivies par leurs équivalents, souvent entre parenthèses, comme *le tractus gastro-intestinal (un tube permettant de digérer les aliments)*. Dans plusieurs cas, ces deux procédés (substitution et ajout d'informations) sont employés conjointement parce qu'ils apportent des informations différentes et complémentaires ;

2) *inclusion sémantique.* Le sens d'une phrase se trouve inclus dans le sens de l'autre phrase de la paire. L'inclusion est orientée : la phrase technique ou la phrase simplifiée peuvent être incluantes. Nous traitons les deux sens de l'inclusion comme un seul type d'alignement. Leur distinction n'a pas montré de différences significatives lors des expériences. Dans l'exemple qui suit, la phrase technique est incluante et indique en plus le nombre de participants et la mesure d'évaluation : *{Peu de données (43 participants) étaient disponibles concernant la détection d'un mauvais placement (la spécificité) en raison de la faible incidence des mauvais placements}{cependant, peu de données étaient disponibles concernant les sondes placées incorrectement et les complications possibles d'une sonde mal placée}*. Dans le cas d'inclusion, la simplification est effectuée également au niveau syntaxique. Typiquement, les subordonnées, les incises, les informations entre parenthèses, certains adjectifs ou adverbes peuvent être supprimés. Dans l'exemple cité, les informations entre parenthèses (*43 participants* et *la spécificité*) sont omises. Dans d'autres cas, les phrases complexes syntaxiquement sont segmentées. L'inclusion sémantique concerne également les énumérations. Ainsi, une phrase technique coordonnée peut être segmentée en une liste d'items séparés dans la version simplifiée. Notons aussi que la simplification syntaxique est souvent accompagnée par des transformations lexicales, comme *{incidence}{complications possibles}*, *{mauvais placement}{placé incorrectement}* ou *{mauvais placement}{mal placé}* dans le dernier exemple.

4. Méthodologies pour l'alignement de phrases parallèles

Dans notre corpus, les documents comparables sont déjà associés entre eux. En revanche, comme les textes techniques et simplifiés sont souvent rédigés de manière indépendante, l'ordre des phrases dans les documents n'est pas significatif. L'accent principal de la méthode est donc mis sur la recherche de phrases parallèles. Notre méthode se compose de plusieurs étapes : le prétraitement dont le filtrage de phrases, l'alignement de phrases et l'évaluation des alignements. Nous décrivons également les différentes expériences effectuées.

4.1. *Prétraitement*

Tous les documents sont étiquetés avec TreeTagger (Schmid, 1994), ce qui permet d'en obtenir leurs versions lemmatisées. Les documents sont ensuite segmentés en phrases en exploitant la ponctuation forte (. *? ! ; :*). D'autres prétraitements sont dédiés

au filtrage des phrases pour ne retenir que les meilleurs candidats à l'alignement. Nous exploitons trois méthodes pour le filtrage basées sur la forme et la syntaxe :

1) méthode basée sur le nombre de mots dans les phrases : chaque phrase candidate doit contenir au moins cinq mots, ce qui correspond à la longueur de la phrase la plus courte dans les données de référence ;

2) suppression de paires avec les phrases identiques ;

3) exploitation d'informations syntaxiques : nous nous inspirons d'un travail existant qui mesure la similarité entre les phrases dans un corpus monolingue grâce aux constituants syntaxiques (Duran *et al.*, 2014). Le score de similarité est alors calculé sur la base des nœuds syntaxiques similaires qui contiennent des mots similaires. Il est difficile d'adapter cette méthode, essentiellement parce qu'elle se base sur une table de similarité entre les constituants, alors que cette table est créée pour l'anglais et que de plus les auteurs ne donnent pas d'indications sur les principes de sa création. Nous supposons cependant que l'adoption d'une approche similaire permettra d'éliminer les paires de phrases indésirables pour l'alignement. Ainsi, au lieu de calculer le score de similarité, nous effectuons un filtrage binaire : garder ou non une paire de phrases candidates. Pour une paire donnée, nous calculons l'arbre syntaxique de chacune des phrases. Ensuite, nous comparons les feuilles (c'est-à-dire les mots) des arbres, sauf celles qui contiennent les mots vides. La liste de mots vides contient 83 entrées (mots grammaticaux comme les déterminants ou prépositions). Lorsque nous trouvons deux mots identiques, nous vérifions leurs nœuds pères : s'ils sont identiques, nous gardons la phrase comme candidate à l'alignement. Le processus est illustré par l'algorithme 1. Nous exploitons également une variante de la méthode : au lieu de nous arrêter lorsque les nœuds pères ne sont pas identiques, nous continuons de remonter l'arbre jusqu'au troisième nœud tant que les nœuds précédents n'ont pas donné de résultats positifs. La comparaison s'arrête lorsque les nœuds sont identiques et la phrase est retenue pour l'alignement ou lorsque la profondeur est supérieure à 3. Cette approche est illustrée par l'algorithme 2. La considération de nœuds parents de profondeur 3 permet également d'observer comment la profondeur de l'arbre influence le filtrage. L'analyse syntaxique des phrases est obtenue avec le Berkeley Neural Parser et le modèle de langue pour le français de la librairie python `benepar` (Kitaev et Klein, 2018). Nous utilisons la librairie NLTK `Tree` pour la manipulation d'arbres syntaxiques (Bird *et al.*, 2009).

4.2. *Alignement de phrases*

Nous abordons la recherche de phrases parallèles comme une problématique de catégorisation : pour une paire de phrases présélectionnées lors de l'étape précédente, il faut décider s'il faut les mettre dans la catégorie *aligné* ou non.

Nous utilisons plusieurs classifieurs linéaires de `scikit-learn` (Pedregosa *et al.*, 2011) avec leurs paramètres par défaut, s'il n'est pas indiqué autrement : `Perceptron` (Rosenblatt, 1958), `Perceptron multicouche (MLP)` (Rosenblatt, 1961), `Random`

Data: Deux arbres syntaxiques (T_1 et T_2), une liste de *mots vides* (*SW*)
Result: Booléen
Booléen ← False;
if *au moins un verbe est dans chaque arbre* **then**
  foreach *feuille de* T_1 *(*L_1*) absente de SW* **do**
    foreach *feuille de* T_2 *(*L_2*) absente de SW* **do**
      if L_1 *est identique à* L_2 **then**
        if *l'étiquette du père de* L_1 *est identique à l'étiquette du père de* L_2 **then**
          Booléen ← True ;
        else
          rien ;
        end
      else
        rien ;
      end
    end
  end
else
  rien ;
end
return Booléen ;
Algorithm 1: Filtrage par la comparaison des pères immédiats des feuilles

`Forest` (`RF`) (Ho, 1995) `Linear discriminant analysis` (`LDA`) (Fisher, 1936) avec le solveur `LSQR`, `Quadratic discriminant analysis` (`QDA`) (Cover, 1965), `Logistic regression` (Berkson, 1944), modèle log-linéaire appris avec `Stochastic gradient descent` (`SGD`) (Ferguson, 1982), et `SVM linéaire` (Vapnik et Lerner, 1963).

Pour avoir une méthode assez générique et pouvoir l'évaluer sur d'autres jeux de données, nous utilisons des descripteurs qui seraient facilement calculables. Nous exploitons cinq types de descripteurs. Par rapport à la typologie présentée dans la section 2, ces descripteurs sont essentiellement liés au lexique et au corpus. Les descripteurs sont calculés sur les formes et les lemmes :

1) *BL : descripteurs de base (baseline)* :

– *nombre de mots communs, hors mots grammaticaux*, ce qui permet de calculer l'intersection lexicale de base entre les phrases (Barzilay et Elhadad, 2003) ;

– *ratio longueur de la phrase la plus courte sur la longueur de la phrase la plus longue*. Ce descripteur suppose que la simplification peut impliquer une association stable avec la longueur des phrases ;

– *différence de la longueur moyenne des mots entre les deux phrases* pour estimer l'utilisation de mots longs, jugés spécifiques au langage technique ;

2) *L : descripteurs issus de la distance de chaînes d'édition* (Levenshtein, 1966) :

```
Data: Deux arbres syntaxiques (T1 et T2), une liste de mots vides (SW)
Result: Booléen
Booléen ← False;
if au moins un verbe est dans chaque arbre then
    foreach feuille de T1 (L1) absente de SW do
        foreach feuille de T2 (L2) absente de SW do
            if L1 est identique à L2 then
                if l'étiquette du père de L1 (P1) est identique à l'étiquette du
                  père de L2 (P2) then
                    Booléen ← True ;
                else
                    if l'étiquette du père de P1 (PP1) est identique à l'étiquette
                      du père de P2 (PP2) then
                        Booléen ← True;
                    else
                        if l'étiquette du père de PP1 est identique à l'étiquette
                          du père de PP2 then
                            Booléen ← True;
                        else
                            rien;
                        end
                    end
                end
            else
                rien ;
            end
        end
    end
else
    rien ;
end
return Booléen ;
```

Algorithm 2: Filtrage par la comparaison d'ancêtres des feuilles (profondeur 3)

– *distance d'édition calculée au niveau des caractères*. Il s'agit de l'acception classique de la mesure. Elle prend en compte les opérations d'édition de base (insertion, suppression et substitution). Le coût de chaque opération est de 1 ;

– *distance d'édition calculée au niveau des mots*. Ce descripteur est calculé avec des mots comme unité. Il prend en compte les mêmes opérations d'édition avec le coût de 1. Le descripteur permet de calculer le coût de la transformation lexicale ;

3) *S : descripteurs basés sur les similarités lexicales* avec la *similarité au niveau des mots calculée selon trois scores (cosinus, Dice et Jaccard)*. Ce descripteur fournit une indication plus sophistiquée sur l'intersection lexicale entre les deux phrases. Le

poids de chaque mot est de 1 ;

4) *N : descripteurs basés sur les n-grammes (bigrammes et trigrammes) de caractères en commun*, ce qui permet de prendre en compte la présence de séquences de caractères communs ;

5) *PL : descripteurs basés sur les plongements lexicaux.* Deux descripteurs sont utilisés : *WAVG* (Stajner *et al.*, 2018), où la moyenne des vecteurs de mots de chacune des deux phrases est calculée et ces vecteurs sont comparés pour attribuer un score de similarité ; et *CWASA* (Franco-Salvador *et al.*, 2016) pour (*continuous word alignment-based similarity analysis*). Ces descripteurs sont exploités avec des plongements entraînés sur le corpus CLEAR à l'aide de Word2Vec[6] (Mikolov *et al.*, 2013), alors que les scores sont calculés avec l'outil CATS (Stajner *et al.*, 2018). Nous avons mené les mêmes expériences avec des vecteurs préentraînés avec Fast Text[7] (Grave *et al.*, 2018) et n'avons pas noté de différences significatives.

4.3. *Évaluation*

L'évaluation est effectuée par rapport aux données de référence. L'entraînement du système est effectué sur 70 % de paires de phrases et le test est effectué sur le reste des données. Plusieurs classifieurs et plusieurs combinaisons de descripteurs sont testés. Les mesures d'évaluation classiques sont calculées : précision, rappel, F-mesure, EQM (erreur quadratique moyenne) et vrais positifs. Avec les données déséquilibrées, l'évaluation est effectuée sur 50 tirages différents afin de mieux évaluer les performances de l'alignement des phrases. Nous rapportons uniquement les scores pour la catégorie de phrases alignées car, d'une part, c'est le principal résultat visé et, d'autre part, avec les données déséquilibrées et une très grande quantité de phrases non alignables, les résultats globaux sont toujours très élevés.

4.4. *Expériences*

Les données de référence fournissent 266 paires de phrases parallèles comme exemples positifs et nous choisissons aléatoirement des exemples négatifs à partir des mêmes documents : 266 paires de phrases non parallèles pour les expériences avec des données équilibrées et d'autres paires de phrases pour des expériences avec des données non équilibrées. Les exemples négatifs sont obtenus en appariant aléatoirement des phrases d'un document technique et de son pendant simple, en vérifiant que ces paires ne font pas partie de la classe positive (équivalence ou inclusion). Il n'y a pas d'intersection entre les phrases alignées et non alignées. Plusieurs expériences sont effectuées, où nous étudions les effets des descripteurs et du déséquilibre.

6. Hyperparamètres de Word2Vec : `-size 300 -window 7 -sample 1e-5 -hs 1 -negative 50 -mincount 20 -alpha 0.025 -cbow 0`
7. `https://fasttext.cc/docs/en/crawl-vectors.html`

4.4.1. *Baseline*

Notre *baseline* correspond à la combinaison de descripteurs lexicaux traditionnellement utilisés pour l'alignement de phrases : la longueur des phrases et l'intersection lexicale entre les phrases. Cette expérience est effectuée avec les données équilibrées.

4.4.2. *Détection de phrases parallèles avec une distribution équilibrée*

Le nombre d'exemples positifs et négatifs est comparable, ce qui correspond à une distribution équilibrée des paires de phrases entre les deux catégories. Cette expérience permet de tester différents descripteurs et leurs combinaisons.

4.4.3. *Détection de phrases parallèles selon la sémantique des paires*

Les paires de phrases de référence sont divisées en deux sous-ensembles, selon le lien sémantique qui existe au sein de la paire :

– *E* : 130 paires avec équivalence sémantique ;

– *I* : 136 paires où le contenu de la phrase technique est compris dans la phrase simplifiée ou l'inverse. Ceci représente les cas de découpage ou de fusion de phrases, ainsi que la suppression ou l'ajout d'informations, lors de la simplification.

4.4.4. *Détection de phrases parallèles avec une distribution déséquilibrée*

Comme le montre le tableau 1, les phrases parallèles sont plutôt rares et il existe beaucoup plus de phrases non alignables. La distribution de phrases parallèles n'est donc pas élevée ni constante : le taux d'alignement varie selon les corpus, les paires de documents et le sens d'alignement. Ainsi, l'objectif de cette expérience est de voir quelles sont les performances du système lorsque les données traitées s'approchent de la distribution naturelle de phrases alignables. Pour chaque sous-ensemble (*E, I*), nous prenons d'abord autant de paires équilibrées que d'exemples négatifs sélectionnés aléatoirement. Ensuite, nous augmentons progressivement le nombre de paires non alignables jusqu'à un ratio de $200:1$, proche de celui des données réelles après le filtrage. Ceci correspond à l'ensemble déséquilibré D avec les 136 (*E*) ou 130 *I* paires alignées et le ratio croissant de paires non alignées. Le ratio et les données changent donc à chaque itération. Nous utilisons aussi l'ensemble réel R, qui comporte toute la combinatoire possible de paires de phrases après filtrage (21 428), alignées et non alignées. L'ensemble R est toujours le même. Nous procédons ainsi en raison du faible nombre d'exemples positifs. Il est donc à noter que le score de rappel en sera artificiellement augmenté. Cela dit, le score de précision évalue la robustesse du modèle à ne pas produire de faux positifs, ce qui nous semble important en raison du grand déséquilibre en faveur d'exemples négatifs. À chaque point de déséquilibre de l'ensemble D, nous faisons deux séries d'expériences :

1) *DD* : entraînement et test au sein de l'ensemble déséquilibré D ;

2) *DR* : entraînement sur l'ensemble D et test sur les données réelles R (environ 21 428 paires de phrases après filtrage).

5. Résultats

Nous présentons les résultats de différentes expériences : (1) l'effet du filtrage sur le corpus, (2) la méthode de *baseline* pour l'alignement de phrases avec l'utilisation de descripteurs basiques, (3) la détection de phrases parallèles avec une distribution équilibrée de paires de phrases parallèles et non parallèles, (4) la détection de phrases parallèles selon la sémantique des paires en distinguant l'équivalence sémantique et l'inclusion, (5) la détection de phrases parallèles avec une distribution déséquilibrée s'approchant de la distribution réelle moyenne de phrases alignables. Nous faisons également une analyse des erreurs et présentons quelques limitations actuelles.

5.1. *Filtrage*

Paires restantes	*Originales*	*IF*	*Syntaxe 1*	*Syntaxe 3*
Total	1 164 407	409 530	16 879	21 428
Équivalence	136	136	94	94
Inclusion	130	130	94	100

Tableau 2. *Effet du filtrage sur le corpus*

La première colonne du tableau 2 indique le nombre de paires de phrases originales, la seconde le nombre de paires qui restent après l'utilisation des indices formels liés à la présence du verbe et l'élimination des paires avec des phrases identiques (IF), et les deux dernières indiquent le nombre de paires qui restent après l'utilisation du filtre syntaxique, en remontant respectivement au premier et au troisième père. Les indices formels sont appliqués avant les filtres syntaxiques. Les filtres syntaxiques sont appliqués indépendamment l'un de l'autre.

Nous observons que les indices formels réduisent le nombre total de paires de 65 % : on passe de 1 164 407 à 409 530. Nous voyons qu'avec ces indices nous ne perdons aucun exemple positif. À partir des 409 530 paires obtenues après le premier filtre, nous observons une autre grande réduction du volume de paires avec chacun des filtres syntaxiques. Le filtre de profondeur 1 laisse 16 879 paires ($\sim$ 96 % de réduction) et celui de profondeur 3 laisse 21 428 paires ($\sim$ 95 % de réduction). Le défaut de ce type de filtre est qu'un nombre non négligeable d'exemples positifs est perdu : 42 sur 136 ($\sim$ 30 %) pour les couples équivalents avec les deux filtres syntaxiques, 36 sur 130 ($\sim$ 27 %) pour la profondeur 1, et 32 sur 130 ($\sim$ 24 %) pour la profondeur 3 pour les couples avec l'inclusion. Nous présentons deux exemples, où le premier est conservé alors que le deuxième est rejeté à tort après filtrage IF et syntaxe 3 :

– *{L'apparition de signes cliniques tels qu'un mal de gorge, une fièvre, une pâleur, un purpura ou un ictère pendant le traitement par la sulfasalazine peut faire suspecter une myélosuppression, une hémolyse ou une hépatotoxicité.}{L'apparition de signes cliniques tels qu'un mal de gorge, une fièvre, une pâleur, de petites taches rouges sur*

la peau ou une jaunisse pendant le traitement par la sulfasalazine peut faire suspecter une diminution du nombre de cellules du sang, une destruction des globules rouges ou une toxicité du foie.}

– {L'allaitement doit être interrompu en cas de traitement par capécitabine.}{Vous ne devez pas allaiter si vous êtes traitée par capecitabine eg.}

Nous voyons que les phrases avec des substitutions lexicales, comme *{hémolyse}{destruction des globules rouges}*, sont conservées. Ceci est important car il a été observé que ce procédé représente environ 70 % des transformations dans les textes médicaux (Koptient *et al.*, 2019). En revanche, les transformations syntaxiques, comme la modification de parties du discours *{allaitement}{allaiter}* ou de la voix du verbe *{passive}{active}*, sont plus difficiles à conserver. Pour ces cas, un travail plus poussé sur les structures syntaxiques comparables et la morphologie sera nécessaire.

Tous les traitements décrits ci-après sont effectués sur les données filtrées avec les indices formels et le filtre syntaxique de profondeur 3.

5.2. *Alignement de phrases parallèles*

Classifieur	*P*	*R*	*F1*	*EQM*	*VP*
`Perceptron`	0,90	0,93	0,92	0,08	28
`MLP`	0,93	0,93	0,93	0,06	28
`RF`	**1,00**	**0,97**	**0,98**	**0,02**	**29**
`LDA`	0,93	0,87	0,90	0,09	26
`QDA`	0,96	0,90	0,93	0,06	27
`LogReg`	0,97	**0,97**	0,97	0,03	**29**
`SGD`	0,90	0,93	0,92	0,08	28
`LinSVM`	0,97	0,93	0,95	0,04	28

Tableau 3. *Résultats d'alignement : différents classifieurs, ensemble des descripteurs, ensemble de test, texte non lemmatisé, ratio des classes* 1 : 1. *en-têtes de colonnes : précision (P), rappel (R), erreur quadratique moyenne (EQM), vrais positifs (VP)*

Les résultats globaux se trouvent dans le tableau 3. Il s'agit de l'exploitation de l'ensemble des descripteurs sur l'ensemble de test avec le texte non lemmatisé. Les résultats sont présentés en termes de rappel R, précision P, F-mesure F, erreur quadratique moyenne EQM et vrais positifs VP (sur un total de 30 paires de phrases alignées dans l'ensemble de test). Nous pouvons voir que tous les classifieurs testés sont compétitifs avec une F-mesure entre 0,92 et 0,98. Pour tous les classifieurs, nous indiquons les scores moyens de 20 itérations. La précision et le rappel sont équilibrés. `Random Forest` semble être le meilleur classifieur : F-mesure de 0,98 (précision 1 et rappel 0,97), le plus grand nombre de vrais positifs (56) et l'erreur quadratique moyenne la plus faible (0,02). `Régression Logistique` est presque aussi performant

avec 0,97 de précision, rappel et F-mesure. Les expériences qui suivent sont effectuées avec `Random Forest`.

5.2.1. *Baseline*

Pour la *baseline*, nous exploitons les descripteurs le plus souvent utilisés : longueur des phrases et intersection lexicale entre les phrases. Les résultats sont présentés dans la première ligne du tableau 4 : nous obtenons une F-mesure de 0,95, ce qui indique que les descripteurs traditionnels sont en effet assez efficaces pour cette tâche.

5.2.2. *Détection de phrases parallèles avec une distribution équilibrée*

Descripteurs	*R*	*P*	*F1*	*EQM*	*VP*	*Descripteurs*	*R*	*P*	*F1*	*EQM*	*VP*
BL	0,97	0,93	0,95	0,05	28	*BL + L + S*	1,00	0,97	0,98	0,02	29
S	0,97	0,97	0,97	0,03	29	*BL + L + N*	1,00	0,97	0,98	0,02	29
L	0,90	0,93	0,92	0,09	28	*BL + L + PL*	1,00	0,97	0,98	0,02	29
N	0,97	0,93	0,95	0,05	28	*BL + S + N*	1,00	0,97	0,98	0,02	29
PL	0,97	0,97	0,97	0,03	29	*BL + S + PL*	1,00	0,97	0,98	0,02	29
L + S	1,00	0,93	0,97	0,03	28	*BL + N + PL*	1,00	0,97	0,98	0,02	29
L + N	1,00	0,97	0,98	0,02	29	*L + S + N*	1,00	0,97	0,98	0,02	29
L + PL	0,97	0,97	0,97	0,03	29	*L + S + PL*	1,00	0,97	0,98	0,02	29
S + N	1,00	0,97	0,98	0,02	29	*L + N + PL*	1,00	0,97	0,98	0,02	29
S + PL	1,00	0,97	0,98	0,02	29	*BL + L + S + N*	1,00	0,97	0,98	0,02	29
BL + L	1,00	0,97	0,98	0,02	29	*BL + L + S + PL*	1,00	0,97	0,98	0,02	29
BL + S	1,00	0,97	0,98	0,02	29	*BL + L + N + PL*	1,00	0,97	0,98	0,02	29
BL + N	1,00	0,97	0,98	0,02	29	*BL + S + N + PL*	1,00	0,97	0,98	0,02	29
BL + PL	1,00	0,97	0,98	0,02	29	*L + S + N + PL*	1,00	0,97	0,98	0,02	29
N + PL	1,00	0,97	0,98	0,02	29	*BL + L + S + N + PL*	1,00	0,97	0,98	0,02	29

Tableau 4. *Résultats d'alignement : différents ensembles de descripteurs,* `Random Forest`*, texte non lemmatisé, ratio des classes* 1 : 1*. en-têtes de colonnes : précision (P), rappel (R), erreur quadratique moyenne (EQM), vrais positifs (VP). Rangées : baseline (BL), similarité (S), Levenshtein (L), plongements lexicaux (PL).*

Le tableau 4 présente les résultats de la détection de phrases parallèles avec une distribution équilibrée des données. Nous testons différents types de descripteurs et leurs combinaisons. Les meilleurs résultats sont obtenus par l'ensemble *S* (mesures de similarité) avec une F-mesure de 0,97.

Les moins bons résultats sont obtenus avec l'ensemble *L* (distance de Levenshtein) avec une F-mesure de 0,92. Les différentes combinaisons de descripteurs permettent d'améliorer ces résultats, ce qui indique que chaque type de descripteurs apporte des informations complémentaires. La plupart des combinaisons atteignent les résultats les plus élevés. Les expériences qui suivent sont effectuées avec tous les descripteurs.

5.2.3. *Détection de phrases parallèles selon la sémantique des paires avec des données équilibrées*

Ensemble	*P*	*R*	*F1*	*EQM*	*VP*
Équivalence E	1,00	0,97	0,98	0,02	29
Inclusion I	1,00	0,94	0,97	0,03	29

Tableau 5. *Résultats d'alignement : les deux ensembles de données équilibrées (l'équivalence sémantique et les inclusions), ensemble de test, tous les descripteurs,* `Random Forest`*, texte non lemmatisé, ratio des classes* $1:1$*. en-têtes de colonnes : précision (P), rappel (R), erreur quadratique moyenne (EQM), vrais positifs (VP).*

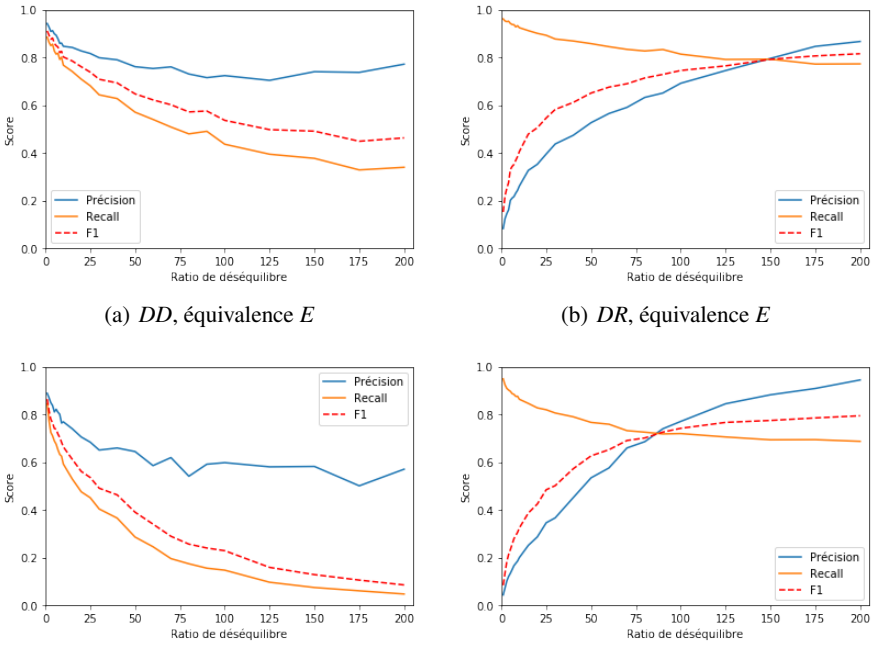
Dans cette série d'expériences sur les données équilibrées, nous voulons voir s'il existe une différence selon le type de relation sémantique au sein des paires de phrases. Dans l'ensemble de test, nous comptons 30 couples en équivalence et 31 en inclusion. Selon le tableau 5, il existe une légère différence : il est un peu plus facile de détecter les phrases en relation d'équivalence que les phrases en relation d'inclusion. Nous supposons que les paires d'inclusion couvrent une plus grande variété de situations, ce qui est plus difficile à modéliser avec le volume de données dont nous disposons.

5.2.4. *Détection de phrases parallèles avec une distribution déséquilibrée*

Comme les documents comparables peuvent contenir un taux variable de phrases parallèles, nous faisons des tests avec des données déséquilibrées. Nous testons différents taux de déséquilibre. Les résultats sont présentés à la figure 1 : l'axe x représente l'augmentation du déséquilibre (seule la première position 1 correspond aux données équilibrées), alors que l'axe y représente les scores de précision, rappel et F-mesure. Les résultats pour les deux ensembles sont présentés : équivalence (figures 1(a) et 1(b)) et inclusion (figures 1(c) et 1(d)). La colonne de gauche présente les résultats DD, lorsque l'entraînement et le test sont effectués sur des données avec le même rapport de déséquilibre D. La colonne de droite présente les résultats DR obtenus par les mêmes modèles D mais testés sur l'ensemble des données R (toutes les paires de phrases possibles). Les résultats présentés sont les moyennes de 50 itérations.

Nous pouvons faire plusieurs observations. Comme indiqué dans la section précédente, les paires équivalentes (figures 1(a) et 1(b)) sont plus faciles à catégoriser que les inclusions. Les scores de précision et de rappel sont alors plus élevés à différents points de déséquilibre.

Ce résultat est positif car les phrases équivalentes fournissent les informations les plus utiles et complètes sur les transformations requises lors de la simplification. Sans surprise, l'augmentation du déséquilibre mène vers des performances réduites durant l'entraînement. Cela signifie que le déséquilibre crée de la confusion entre les paires alignables et non. Cependant, pour atteindre notre objectif, qui consiste à identifier le peu d'exemples positifs présents dans une masse de paires non alignables, il vaut mieux utiliser un modèle plus robuste face au déséquilibre des données, même s'il



(a) *DD*, équivalence *E*

(b) *DR*, équivalence *E*

(c) *DD*, inclusion *I*

(d) *DR*, inclusion *I*

Figure 1. *Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (DD et DR) sur les données déséquilibrées*

fonctionne moins bien à l'entraînement. Il s'agit ici de modèles entraînés avec un déséquilibre important. Finalement, notons que la même tâche a été effectuée avec un réseau neuronal à propagation avant basique[8], avec différentes expériences faisant varier le nombre de couches cachées et leur taille ainsi que le nombre d'*epochs*. Les résultats étaient plus élevés lors de l'entraînement sur les données équilibrées : jusqu'à 0,98 de précision, rappel et F-mesure. Cependant, à partir du déséquilibre $5:1$, toutes les phrases étaient systématiquement classées dans la catégorie « *non aligné* », ce qui montre les limites de l'architecture neuronale utilisée. Nous avons également fait des essais avec un système d'alignement existant qui utilise un réseau de neurones récurrent bidirectionnel (Grégoire et Langlais, 2018), mais les résultats obtenus n'étaient pas exploitables. Finalement, nous avons ajouté les descripteurs à base de plongements lexicaux comme proposé dans un travail existant (Kajiwara et Komachi, 2016) mais cette expérience n'a pas montré d'évolution dans les résultats. Nous

8. Un réseau à propagation avant avec fonction d'activation `ReLU`, optimiseur `ADAM`, fonction d'erreur `BCEWithLogitsLoss` (sigmoïde + `BCELoss`).

pensons que le faible volume d'exemples à notre disposition (136 exemples d'équivalence et 130 exemples d'inclusion) représente un frein à l'utilisation des méthodes neuronales. Un travail plus avancé sur ce type de méthodes et l'accroissement du volume des données de référence font partie de nos perspectives.

5.3. *Analyse des erreurs*

	Équivalence	Inclusion	Intersection	Faux positifs
Nb de paires	75	15	2	8

Tableau 6. *Analyse des 100 premiers alignements par le modèle entraîné sur les paires équivalentes à un ratio de* $125:1$*, appliqué sur un ensemble aléatoire de paires non vues pendant l'entraînement*

Le tableau 6 présente les différents types de paires de phrases considérées comme alignables par le modèle entraîné sur un ratio de $125:1$ de paires équivalentes, qui montre le meilleur équilibre entre rappel et précision sur les données réelles. Les 100 premiers alignements, absents des données d'entraînement, sont analysés. Nous pouvons observer que 75 % de paires correspondent en effet à l'équivalence, ce qui est en adéquation avec la figure 1(b). Nous remarquons également que 15 % des paires alignées relèvent de l'inclusion, qui est une catégorie plus difficile à identifier de manière ciblée. Nous trouvons enfin que 2 % des alignements correspondent à d'intersection. Nous ne recherchons pas spécifiquement ce type de couples de phrases pour la simplification car nous le considérons plus difficile à traiter. Cependant, de tels alignements peuvent être utiles à l'identification de paraphrases, par exemple. Nous avons ainsi 90 à 92 % d'alignements exploitables et 8 à 10 % de bruit. Certaines paires des corpus *Cochrane* et *Médicament* sont plus difficiles à aligner car ces corpus combinent les transformations lexicales spécifiques du domaine, l'utilisation de l'opposition et du contraire, et des transformations syntaxiques :

– *{Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.}{Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal.}*

– *{Aucune preuve n'indique que les agents gonflants sont efficaces dans le traitement du SCI}{Nous avons observé que les agents gonflants n'étaient pas efficaces dans le traitement du SCI}*

Pour aider ce type d'alignement, il serait nécessaire de capter la similarité lexicale et sémantique avec des ressources et connaissances complémentaires. Elles peuvent venir de ressources externes ou bien être acquises sur le corpus.

Finalement, dans les faux positifs, nous trouvons de bons candidats à l'alignement qui ne sont pas dans les données de référence, comme :

– {Trois études ont rapporté des résultats mitigés concernant l'association entre le début des cours plus tardif et la vigilance des étudiants.}{Différentes études rapportaient des résultats mitigés concernant l'association entre le début plus tardif des cours et une augmentation de la fréquentation et de la vigilance des étudiants.}

5.4. *Limitations de l'étude actuelle*

La limitation principale des expériences présentées est liée aux descripteurs exploités. Parmi les quatre types de descripteurs distingués dans les travaux existants, nous exploitons les descripteurs principalement basés sur le lexique et le corpus. L'exploitation de descripteurs faciles à calculer et ne nécessitant pas de ressources externes était un des objectifs. Cependant, cet aspect doit évoluer car actuellement les similarités lexicales sont assez faibles dans les phrases différenciées par leur degré de technicité. Ce point a été relevé lors de l'analyse des erreurs d'alignement : les faux négatifs contiennent souvent des phrases sémantiquement similaires mais contenant un lexique et des structures syntaxiques différents. Une meilleure intégration de plongements lexicaux et d'une architecture neuronale fait partie des perspectives.

Une autre limitation est liée à la catégorisation binaire des paires de phrases selon qu'elles sont alignables ou non. Cette catégorisation est motivée par la tâche poursuivie, où nous avons besoin de paires de phrases parallèles pour induire des règles de transformation nécessaires pour la simplification. Cependant, comme dans les données STS, nous pouvons aussi viser de caractériser les paires de phrases sur une échelle de similarité et disposer ainsi de données de référence plus fines. Notons que nous avons effectué un travail de ce type sur des données issues du corpus CLEAR et des articles de Wikipédia et Vikidia en langue générale. Ces données ont été exploitées lors de la compétition DEFT 2020.

6. Conclusion

Nous avons proposé une série d'expériences en alignement de phrases parallèles à partir de corpus monolingues comparables en français. La dimension comparable est due à la technicité des documents et contraste les versions techniques et simplifiées des documents et des phrases. Nous exploitons un corpus comparable existant lié au domaine biomédical et contenant des documents de trois genres (encyclopédique, scientifique et notices de médicaments). Les données de référence sont construites manuellement. La recherche de phrases parallèles est abordée comme une problématique de catégorisation : nous devons décider si une paire de phrases peut être alignée ou non. Plusieurs classifieurs sont exploités. Nos résultats atteignent une F-mesure de 0,97 sur les données équilibrées en français, avec un bon équilibre entre la précision et le rappel. Les meilleurs résultats sont obtenus avec `Random Forest`. Deux autres expériences s'intéressent aux types de relations au sein des paires de phrases (les paires de phrases avec la relation d'équivalence sont plus faciles à aligner que les phrases

en relation d'inclusion) et sur l'équilibre entre les paires alignables et non alignables dans les ensembles d'entraînement et de test.

Comme nous l'avons vu, dans les données de référence, la distance lexicale entre les phrases techniques et simplifiées est assez élevée. En conséquence, d'autres descripteurs doivent être utilisés pour mieux cerner les phrases alignables. Par exemple, nous comptons utiliser des connaissances externes, comme les terminologies médicales (Côté *et al.*, 1993 ; Lindberg *et al.*, 1993) et le lexique ReSyf (Billami *et al.*, 2018), ou des ressources constituées à partir de corpus. Nous comptons également tester les représentations de phrases avec FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2020). Une étude plus poussée des descripteurs pourrait également être intéressante : les performances selon les types d'alignement (équivalence et inclusion), l'impact des descripteurs individuels et non pas par sous-ensembles. Comme le déséquilibre est une caractéristique naturelle des données que nous traitons, le travail à venir pourra également enrichir les filtres pour éliminer un maximum de phrases non alignables, *a priori* et/ou *a posteriori* de l'alignement.

Les meilleurs modèles générés sont actuellement exploités pour enrichir l'ensemble de phrases parallèles. En plus des corpus comparables liés au domaine médical, nous exploitons également un corpus similaire de la langue générale qui regroupe l'ensemble d'articles comparables de Wikipédia et de Vikidia. Nous avons également l'intention d'essayer la méthode sur des corpus spécialisés d'autres domaines que celui de la santé. La ressource constituée dans ce travail, un corpus avec plusieurs milliers de phrases parallèles, sera mise à disposition des chercheurs. En dehors de la simplification automatique, les phrases parallèles peuvent aussi être intéressantes pour d'autres applications de TAL, comme l'étude de la similarité textuelle, les systèmes de questions-réponses, la recherche d'information ou l'implication textuelle.

Remerciements

Ce travail s'inscrit dans le cadre du projet ANR-17-CE19-0016-01 CLEAR (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*) financé par l'Agence nationale de la recherche. Nous remercions les relecteurs pour leurs commentaires et remarques détaillés qui ont permis d'améliorer la qualité du présent article.

7. Bibliographie

Abdul-Rauf S., Schwenk H., « On the Use of Comparable Corpora to Improve SMT performance », *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Association for Computational Linguistics, Athens, Greece, p. 16-23, March, 2009.

Agirre E., Cer D., Diab M., Gonzalez-Agirre A., Guo W., « *SEM 2013 shared task : Semantic Textual Similarity », *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1 : Proceedings of the Main Conference and the Shared Task : Seman-*

tic Textual Similarity, Association for Computational Linguistics, Atlanta, Georgia, USA, p. 32-43, June, 2013.

Barzilay R., Elhadad N., « Sentence Alignment for Monolingual Comparable Corpora », *in* ACL (ed.), *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, p. 25-32, 2003.

Beigman Klebanov B., Knight K., Marcu D., « Text Simplification for Information-Seeking Applications », *in* R. Meersman, Z. Tari (eds), *On the Move to Meaningful Internet Systems 2004 : CoopIS, DOA, and ODBASE*, Springer, LNCS vol 3290, Berlin, Heidelberg, 2004.

Berkson J., « Application of the Logistic Function to Bio-Assay », *Journal of the American Statistical Association*, vol. 39, n° 227, p. 357-365, 1944.

Billami M. B., François T., Gala N., « ReSyf : a French lexicon with ranked synonyms », *in* ACL (ed.), *27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, p. 2570-2581, 2018.

Bird S., Klein E., Loper E., *Natural Language Processing with Python : Analyzing Text with the Natural Language Toolkit*, O'Reilly, Beijing, China, 2009.

Blake C., Kampov J., Orphanides A. K., West D., Lown C., « Unc-ch at duc 2007 : Query expansion, lexical simplification and sentence selection strategies for multi-document summarization », *Proceedings of Document Understanding Conference (DUC) Workshop*, Rochester, New York, USA, 2007.

Brouwers L., Bernhard D., Ligozat A.-L., François T., « Syntactic Sentence Simplification for French », *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Association for Computational Linguistics, Gothenburg, Sweden, p. 47-56, April, 2014.

Cardon R., Grabar N., « Parallel Sentence Retrieval From Comparable Corpora for Biomedical Text Simplification », *Proceedings of Recent Advances in Natural Language Processing*, Varna, Bulgaria, p. 168-177, september, 2019.

Cardon R., Grabar N., « Reducing the Search Space for Parallel Sentences in Comparable Corpora », *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, European Language Resources Association, Marseille, France, p. 44-48, May, 2020.

Chandrasekar R., Srinivas B., « Automatic induction of rules for text simplification », *Knowledge Based Systems*, vol. 10, n° 3, p. 183-190, 1997.

Chen P., Rochford J., Kennedy D. N., Djamasbi S., Fay P., Scott W., « Automatic text simplification for people with intellectual disabilities », *Artificial Intelligence Science and Technology*, 2016.

Chen S. F., « Aligning Sentences in Bilingual Corpora Using Lexical Information », *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, Association for Computational Linguistics, USA, p. 9–16, 1993.

Clough P., Gaizauskas R., Piao S. S., Wilks Y., « Measuring Text Reuse », *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 152-159, July, 2002.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, n° 1, p. 37, 1960.

Côté R. A., Rothwell D. J., Palotay J. L., Beckett R. S., Brochu L., *The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International*, College of American Pathologists, Northfield, 1993.

Cover T., « Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition », *IEEE Transactions on Electronic Computers*, vol. 14, nº 3, p. 326-334, 1965.

De Belder J., Moens M.-F., « Text Simplification for Children », *Workshop on Accessible Search Systems of SIGIR*, Geneva, Switzerland, p. 1-8, 2010.

Duran K., Rodriguez J., Bravo M., « Similarity of sentences through comparison of syntactic trees with pairs of similar words », *11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Campeche, p. 1-6, 09, 2014.

Ferguson T., « An inconsistent maximum likelihood estimate », *Journal of the American Statistical Association*, vol. 77, nº 380, p. 831-834, 1982.

Fernando S., Stevenson M., « A semantic similarity approach to paraphrase detection », *Comp Ling UK*, p. 1-7, 2008.

Fisher R., « The Use of Multiple Measurements in Taxonomic Problems », *Annals of Eugenics*, vol. 7, nº 2, p. 179-188, 1936.

Franco-Salvador M., Gupta P., Rosso P., Banchs R. E., « Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language », *Knowledge-Based Systems*, vol. 111, p. 87 - 99, 2016.

Fung P., Cheung P., « Mining Very-Non-Parallel Corpora : Parallel Sentence and Lexicon Extraction via Bootstrapping and EM », *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Barcelona, Spain, p. 57-63, July, 2004.

Gale W. A., Church K. W., « A Program for Aligning Sentences in Bilingual Corpora », *Comp Linguistics*, vol. 19, nº 1, p. 75-102, 1993.

Grave E., Bojanowski P., Gupta P., Joulin A., Mikolov T., « Learning Word Vectors for 157 Languages », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, May, 2018.

Grégoire F., Langlais P., « Extracting Parallel Sentences with Bidirectional Recurrent Neural Networks to Improve Machine Translation », *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, p. 1442-1453, August, 2018.

Guo W., Diab M., « Modeling Sentences in the Latent Space », *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Jeju Island, Korea, p. 864-872, July, 2012.

He H., Gimpel K., Lin J., « Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks », *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, p. 1576-1586, September, 2015.

Hewavitharana S., Vogel S., « Extracting Parallel Phrases from Comparable Data », *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, Association for Computational Linguistics, Portland, Oregon, p. 61-68, June, 2011.

Ho T. K., « Random Decision Forests », *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, IEEE Computer Society, USA, p. 278-282, 1995.

Hwang W., Hajishirzi H., Ostendorf M., Wu W., « Aligning Sentences from Standard Wikipedia to Simple Wikipedia », *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Association for Computational Linguistics, Denver, Colorado, p. 211-217, May–June, 2015.

Jonnalagadda S., Tari L., Hakenberg J., Baral C., Gonzalez G., « Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text », *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, Association for Computational Linguistics, Boulder, Colorado, p. 177-180, June, 2009.

Kajiwara T., Komachi M., « Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings », *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, The COLING 2016 Organizing Committee, Osaka, Japan, p. 1147-1158, December, 2016.

Kitaev N., Klein D., « Constituency Parsing with a Self-Attentive Encoder », *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, July, 2018.

Koptient A., Cardon R., Grabar N., « Simplification-induced transformations : typology and some characteristics », *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, Florence, Italy, p. 309-318, August, 2019.

Lai A., Hockenmaier J., « Illinois-LH : A Denotational and Distributional Approach to Semantics », *Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, p. 239-334, 2014.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., « FlauBERT : Unsupervised Language Model Pre-training for French », *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 2479-2490, May, 2020.

Leroy G., Kauchak D., Mouradi O., « A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty », *Int J Med Inform*, vol. 82, n° 8, p. 717-730, 2013.

Levenshtein V. I., « Binary Codes Capable of Correcting Deletions, Insertions and Reversals », *Soviet Physics Doklady*, vol. 10, p. 707, February, 1966.

Lindberg D., Humphreys B., McCray A., « The Unified Medical Language System », *Methods Inf Med*, vol. 32, n° 4, p. 281-291, 1993.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D., Sagot B., « CamemBERT : a Tasty French Language Model », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 7203-7219, July, 2020.

Mihalcea R., Corley C., Strapparava C., « Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity », *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, AAAI Press, Boston, Massachusetts, p. 775–780, 2006.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., « Distributed Representations of Words and Phrases and their Compositionality », *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, p. 3111-3119, 2013.

Mueller J., Thyagarajan A., « Siamese Recurrent Architectures for Learning Sentence Similarity », *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, AAAI Press, Phoenix, Arizona, p. 2786–2792, 2016.

Munteanu D. S., Marcu D., « Processing Comparable Corpora With Bilingual Suffix Trees », *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, Philadelphia, PA, USA, p. 289-295, July, 2002.

Nelken R., Shieber S. M., « Towards Robust Context-Sensitive Sentence Alignment for Monolingual Corpora », *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Trento, Italy, April, 2006.

Nisioi S., Štajner S., Ponzetto S. P., Dinu L. P., « Exploring Neural Text Simplification Models », *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Association for Computational Linguistics, Vancouver, Canada, p. 85-91, July, 2017.

Paetzold G., Specia L., « Benchmarking Lexical Simplification Systems », *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, p. 3074-3080, May, 2016.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, 2011.

Qiu L., Kan M.-Y., Chua T.-S., « Paraphrase recognition via dissimilarity significance classification », *Empirical Methods in Natural Language Processing*, Sydney, Australia, p. 18-26, 2006.

Rosenblatt F., « The Perceptron : a probabilistic model for information storage and organization in the brain », *Psychological Review*, vol. 65, n° 6, p. 386-408, 1958.

Rosenblatt F., *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington DC, 1961.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Severyn A., Nicosia M., Moschitti A., « Learning Semantic Textual Similarity with Structural Representations », *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Association for Computational Linguistics, Sofia, Bulgaria, p. 714-718, August, 2013.

Stajner S., Franco-Salvador M., Ponzetto S. P., Rosso P., « CATS : A Tool for Customized Alignment of Text Simplification Corpora », *Proceedings of the 11th Language Resources and Evaluation Conference, LREC*, Miyazaki, Japan, 2018.

Ștefănescu D., Ion R., Hunsicker S., « Hybrid Parallel Sentence Mining from Comparable Corpora », *16th Conference of the European Association for Machine Translation EAMT*, Trento, Italy, p. 137-144, 2012.

Stymne S., Tiedemann J., Hardmeier C., Nivre J., « Statistical Machine Translation with Readability Constraints », *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, Linköping University Electronic Press, Sweden, Oslo, Norway, p. 375-386, May, 2013.

Tai K. S., Socher R., Manning C. D., « Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks », *Annual Meeting of the Association for Computational Linguistics*, Beijing, China, p. 1556-1566, 2015.

Tsubaki M., Duh K., Shimbo M., Matsumoto Y., « Non-Linear Similarity Learning for Compositionality », *AAAI Conference on Artificial Intelligence*, p. 2828-2834, 2016.

Utiyama M., Isahara H., « Reliable Measures for Aligning Japanese-English News Articles and Sentences », *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Sapporo, Japan, p. 72-79, July, 2003.

Vapnik V., Lerner A., « Pattern Recognition using Generalized Portrait Method », *Automation and Remote Control*, vol. 24, p. 709-715, 1963.

Vickrey D., Koller D., « Sentence Simplification for Semantic Role Labeling », *Proceedings of ACL-08 : HLT*, Association for Computational Linguistics, Columbus, Ohio, p. 344-352, June, 2008.

Štajner S., Popović M., « Can Text Simplification Help Machine Translation ? », *Baltic J. Modern Computing*, vol. 4, n° 2, p. 230-242, 2016.

Wan S., Dras M., Dale R., Paris C., « Using Dependency-based Features to Take the "Para-farce" out of Paraphrase », *Australasian Language Technology Workshop*, p. 131-138, 2006.

Wei C.-H., Leaman R., Lu Z., « SimConcept : A Hybrid Approach for Simplifying Composite Named Entities in Biomedicine », *BCB '14*, p. 138-146, 2014.

Xu W., Callison-Burch C., Napoles C., « Problems in Current Text Simplification Research : New Data Can Help », *Transactions of the Association for Computational Linguistics*, vol. 3, p. 283-297, 2015.

Yang C. C., Li K. W., « Automatic construction of English/Chinese parallel corpora », *J. Am. Soc. Inf. Sci. Technol.*, vol. 54, n° 8, p. 730-742, 2003.

Zhang X., Lapata M., « Sentence Simplification with Deep Reinforcement Learning », *in* ACL (ed.), *Proc of the Conf on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, p. 584-594, 2017.

Zhang Y., Patrick J., « Paraphrase identification by text canonicalization », *Australasian Language Technology Workshop*, p. 160-166, 2005.

Zhao B., Vogel S., « Adaptive parallel sentences mining from web bilingual news collection », *IEEE International Conference on Data Mining*, p. 745-748, 2002.

Zhao J., Zhu T., Lan M., « ECNU : One Stone Two Birds : Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment », *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Association for Computational Linguistics, Dublin, Ireland, p. 271-277, August, 2014.

Zhu Z., Bernhard D., Gurevych I., « A Monolingual Tree-based Translation Model for Sentence Simplification », *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Coling 2010 Organizing Committee, Beijing, China, p. 1353-1361, August, 2010.

A Multi-pass Sieve for Clinical Concept Normalization

Yuxia Wang — Brian Hur — Karin Verspoor — Timothy Baldwin

School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia

ABSTRACT. *Clinical concept normalization involves linking entity mentions in clinical narratives to their corresponding concepts in standardized medical terminologies. It can be used to determine the specific meaning of a mention, facilitating effective use and exchange of clinical information, and to support semantic cross-compatibility of texts. We present a rule-based multi-pass sieve approach incorporating both exact and approximate matching based on dictionaries, and experiment with back-translation as a means of data augmentation. The dictionaries are built from the UMLS Metathesaurus as well as MCN corpus training data. Additionally, we train a multi-class baseline based on BERT. Our multi-pass sieve approach achieves an accuracy of 82.0% on the* MCN corpus*, the highest for any rule-based method. A hybrid method combining these two achieves a slightly higher accuracy of 82.3%.*

RÉSUMÉ. *La normalisation des concepts cliniques consiste à relier les mentions d'entités dans les récits cliniques à leurs concepts correspondants dans des terminologies médicales normalisées. Il peut être utilisé pour déterminer la signification spécifique d'une mention, faciliter l'utilisation et l'échange efficaces d'informations cliniques et soutenir la compatibilité sémantique des textes. Nous présentons une approche de tamisage multi-passes intégrant deux types de correspondance – exacte et approximative – basée sur des dictionnaires construits avec UMLS Metathesaurus et le corpus MCN, et expérimentons la rétro-traduction comme moyen d'augmenter les données. De plus, nous preparons une méthode de référence multi-classes basée sur BERT. Notre méthode de tamisage multi-passes atteint une précision de 82,0% sur le corpus MCN, la plus élevée de toutes les méthodes fondée sur des règles. Notre méthode hybride réalise une précision légèrement supérieure de 82,3%.*

KEYWORDS: *Clinical concept normalization, Rule-based sieve, Back-translation, Neural classifier.*

MOTS-CLÉS : *Normalisation deo concepts cliniques, Tamis basé sur des regles, Traduction arrière, Classificateur neuronal.*

1. Introduction

Free-text clinical notes and discharge summaries are a rich resource for clinical information, and have been utilized in a variety of clinical applications, such as clinical decision making, adverse drug effect analysis, and mortality prediction (Topaz *et al.*, 2016; LePendu *et al.*, 2012; Weissman *et al.*, 2018). Extraction of key clinical concepts mentioned in free-form clinical notes is an important step towards capturing patient-specific signs, symptoms, and disorders that are recorded in the course of care documentation. This requires: (1) concept recognition to identify where a relevant clinical concept is mentioned in the text; and (2) normalization of the recognized concept to a standard identifier from a controlled vocabulary, such as that provided by the Unified Medical Language System (UMLS) (Bodenreider, 2004), which enables standardization in the concept representation. Our work focuses on the second step of clinical concept normalization. This step requires handling of linguistic variation to unify different ways of referring to the same concept, as well as strategies to deal with ambiguity — a term that may refer to different concepts, depending on context — and coverage gaps — mentions that do not link to any concepts in a given knowledge base (D'Souza and Ng, 2015; Li *et al.*, 2017).

In this paper, we focus on normalizing mentions in the MCN (Medical Concept Normalization) corpus, as adopted in N2C2 2019 shared task 3 (Luo *et al.*, 2019). This task was aimed at mapping each mention in a discharge summary to a clinical concept in the form of a Concept Unique Identifier ("CUI") in UMLS 2017AB, concentrating on concepts from either SNOMED-CT (Spackman *et al.*, 1997) or RxNorm (Liu *et al.*, 2005).

In comparison to previously released clinical concept normalization corpora — such as the datasets of ShARe/CLEF eHealth 2013 Task 1 (Pradhan *et al.*, 2013), SemEval-2014 Task 7 (Pradhan *et al.*, 2014), and SemEval-2015 Task 14 (Elhadad *et al.*, 2015) — this dataset reduces the volume of "CUI-less" mentions (mentions that cannot be mapped to a CUI) by expanding the scope of the knowledge base, as well as splitting and adjusting compositional concepts. Specifically, the search space was broadened from a restricted set of 11 disorder-related semantic types in SNOMED-CT to any concept in SNOMED-CT and RxNorm, covering a large set of clinical concepts, including medical problems, treatments, and tests. Each compositional mention span was split into multiple smaller spans that can be normalized to an existing CUI. For example, given that no direct CUI exists for *left breast biopsy*, it was split into *left* and *breast biopsy*, where *breast biopsy* maps to `C0405352` in SNOMED-CT. Ultimately, only 2.7% of mentions were labelled as CUI-less in the final dataset. Furthermore, though ambiguity is abundant in the clinical domain, the restrictions applied in the MCN corpus reduce it greatly (only SNOMED-CT and RxNorm concepts). To be concrete, just 233 mentions among 6,684 instances in the training data of the MCN corpus fall into this category. Therefore, in the context of this specific dataset, the key challenge is not coverage gaps or ambiguity, but variation: mentions which vary lexically and grammatically and are linked to the same CUI.

The goal of the work described in this paper is to improve the accuracy of concept normalization in clinical discharge summaries, and empirically investigate the impact of back-translation (Sennrich *et al.*, 2016) on the clinical normalization task.

Unlike normalizing medical mentions in social media text (Limsopatham and Collier, 2016) or shorter clinical texts such as emergency department triage notes (Aamer *et al.*, 2016), discharge summaries written by clinicians or nurses are more formal. As a result, the main focus in this work is on matching mentions and their variations obtained through morphological alternation with concept names in standardized terminologies, with a particular emphasis on rule-based approaches over machine-learning models. Rule-based methods have the advantage of being redeployable to new vocabularies, as they do not rely on training data (Groza and Verspoor, 2014). We compare our rule-based method with a neural classifier based on BERT (Devlin *et al.*, 2019).

Furthermore, inspired by cross-lingual normalization, we perform back-translation over three different languages (Chinese, French, and German), on original mentions, and then perform exact matching over three dictionaries. We assume that we can take advantage of the following two features of commercial translation tools in our task: (1) high tolerance to spelling errors and abbreviations; and (2) (controlled) lexical variance in the output of back-translation.

Our contributions are three-fold: (1) we propose a multi-pass sieve approach using morphological rules based on UMLS, which we combine with neural models; (2) we are the first to apply back-translation to the clinical concept normalization task; and (3) we achieve a new benchmark accuracy of 82.0% on the MCN corpus for a rule-based method, and 82.3% for a hybrid method combining our rule-based and neural methods together.

2. Related Work

Clinical and biomedical concept normalization is an active field of research, with a broad spectrum of proposed approaches, encompassing rule-based and machine learning-based methods.

Dictionary-based methods focus on strategies for matching terms in a text to the terms of the controlled vocabulary, represented in a dictionary, generally employing rules to control the matching of terms. MetaMap (Aronson, 2001), NCBO Annotator (Shah *et al.*, 2009), and cTAKES (Savova *et al.*, 2010) are three dictionary-based concept normalization systems that have been widely adopted and shown to have good effectiveness across a number of biomedical concept recognition tasks (Funk *et al.*, 2014). Rule-based approaches tend to share a core set of rules relating to abbreviation expansion, word reordering, and punctuation removal, but equally incorporate specialist rules customized to specific datasets. For example, POS and chunking related rules were employed for the AZDC dataset (Kang *et al.*, 2013). Morphological sieves — where unmatched mentions pass through a series of "sieves", generally with increasing recall and decreasing precision, until a match occurs — were developed in previous

work for the ShARe and NCBI datasets (D'Souza and Ng, 2015). However, manual work is required to adapt such methods to a new dataset. In addition, the choice of target terminology (e.g. SNOMED-CT, RxNorm, or MEDIC) often varies across datasets due to their coverage of domain-specific terms, further limiting the direct employment of most rule-based systems. Luo *et al.* (2019) proposed to apply this sieve-based approach to MCN, achieving an accuracy of 76.35%. We build on this research in our work.

Most machine learning-based methods, such as DNorm (Leaman *et al.*, 2013) and its extensions (Leaman and Lu, 2014; Leaman and Lu, 2016), incorporate semantic information by projecting words into vector spaces, where semantic similarity between the input mention and concept names is measured by a similarity score. The score can be calculated directly via similarity metrics such as cosine similarity and Euclidean distance, or learned from the training data. Ranking is generally used as the next step, to rank the candidate concepts associated with a given mention. Before the application of word2vec (Mikolov *et al.*, 2013), TF-IDF and its variants were the dominant word representation. However, as demonstrated by Gong *et al.* (2018), both context-dependent and context-independent word embedding methods are heavily biased by the frequency of occurrence of words, resulting in clusters of rare words with little semantic similarity. Given that most words in clinical mentions and concept names are rare in general domains, they cannot be represented accurately through standard pre-training methods. That is, they tend to be clustered with other rare words rather than semantically. Hence the performance of machine learning-based methods is limited by their heavy dependence on the quality of the underlying word representations.

To overcome this bottleneck, instead of calculating cosine similarity to identify candidates, Xu *et al.* (2020) applied two approaches, one based on Lucene and the other based on fine-tuning a BERT multi-class neural classifier. As Reimers and Gurevych (2019) have shown, fine-tuning can perform much better than directly calculating the cosine similarity of BERT text representations for semantic textual similarity. The most critical component here is the neural ranker, which incorporates semantic type into the loss function as a regularizer, improving performance on multiple datasets. Specifically, on the MCN dataset, an increase in accuracy of 0.81% is obtained using semantic type regularization. While one may argue that neural models require large amounts of in-domain labelled data to perform well, making them impractical for applications in the clinical domain, recent zero-shot entity linking methods can use disposition which don't require in-domain labelled data, suggesting a promising direction for neural concept normalization (Logeswaran *et al.*, 2019).

3. The MCN Corpus

The MCN corpus (Luo *et al.*, 2019) is a publicly-available medical concept normalization dataset, which was first released as part of 2019 N2C2 Shared-Task and Workshop Track 3: N2C2/UMass Track on Clinical Concept Normalization. [1] Table 1

1. https://n2c2.dbmi.hms.harvard.edu/track3.

	Mentions	Unique concepts	CUI-less mentions	Ambiguous mentions
training	6,684	2,331	151	233
test	6,925	2,579	217	192
TOTAL	13,609	3,792	368	425

Table 1. *Numbers of mentions, Unique concepts, mentions labeled as CUI-less, and Ambiguous mentions (more than one CUI) in the training and test partitions of the MCN corpus.*

provides a statistical breakdown of the dataset. It consists of 13,609 mentions representing 10,919 distinct expressions with a total coverage of 3,792 unique concepts, split into 6,648 mentions in the training data set, and 6,925 in the test set.

Two clinical source vocabularies from the 2017AB version of UMLS (Bodenreider, 2004) were used to annotate mentions extracted from 100 discharge summaries: (1) SNOMED-CT (Spackman *et al.*, 1997), a comprehensive clinical reference term base, covering concepts from areas such as anatomy, normal and abnormal functions, symptoms and signs of diseases, diseases/diagnoses, and procedures; and (2) RxNorm (Liu *et al.*, 2005), a collection of medications (drug names). The number of unique concepts in SNOMED-CT, RxNorm, and the combination of the two, is 333,183, 114,150, and 434,056, respectively (Luo *et al.*, 2019). Note that each concept is assigned a Concept Unique Identifier (CUI), and that ambiguous concepts are assigned multiple CUIs (Bodenreider, 2004).

We highlight three features of the corpus below, which inform the development of our method.

Broad coverage of medical concepts

In contrast to disease/disorder entities in corpora such as ShARe/CLEF eHealth 2013 Task 1 (Pradhan *et al.*, 2013), SemEval-2014 Task 7 (Pradhan *et al.*, 2014), and SemEval-2015 Task 14 (Elhadad *et al.*, 2015), the MCN corpus extends the search space to all concepts in SNOMED-CT and RxNorm. This reduces the effects of coverage gaps, where a large proportion of mentions cannot be assigned CUIs due to the limited coverage of the knowledge base: just 368 (2.7%) mentions could not be assigned CUIs (i.e. were "CUI-less"). As such, there is little need to distinguish CUI-less from other mentions before normalizing.

Resolution of compositional mentions

If one span text involves more than one concept, we refer to it as a compositional mention. For example, *breast or ovarian cancer* encompasses two concepts: *breast*

cancer and *ovarian cancer*. *Left breast biopsy* is split into the largest mention span *breast biopsy* which can be normalized to C0405352, and the smaller mention span *left*. As part of the corpus construction, Luo *et al.* (2019) split and adjusted the mention spans so that the smaller spans were annotated using a single CUI.

Formal language

Clinical mentions extracted from discharge summaries are more formal and rigorous than clinically-related social media texts (Limsopatham and Collier, 2016). For example, *head spinning a little* in social media text expresses the concept of dizziness (C0012833), which typically occurs in the more canonical form of *dizzy* or *dizziness* in clinical notes. This makes concept mapping easier.

4. Methods

Based on the three dataset characteristics presented in Section 3, and the fact that rule-based methods tend to be superior to machine-learning methods under such settings (Li *et al.*, 2017; D'Souza and Ng, 2015), we focus primarily on a rule-based method. The procedure from inputting a mention to outputting the CUI is shown in Figure 1. We further hybridize our method with a neural multi-class classifer based on BERT (Devlin *et al.*, 2019).

Our approach is made up of three types of pre-processing, followed by exact match, approximate match, mention permutation, and the neural multi-class classifier, as detailed below.

4.1. *Pre-processing*

We pre-process each mention and dictionary term as follows. Steps 6–8 are applied only to mentions.

1. Lowercase
2. Remove noisy strings and common words such as *'d*, *'s*, *''*, *<*, *>*, *his*, *her*, *patient*, *an*, *a*, and *the*.
3. Remove possessives (e.g. *'s*) and punctuation (e.g. ,, ., -, and /).
4. Remove prepositions, including *of*, *in*, *to*, *for*, *with*, *on*, *at*, *from*, *by*, *about*, *as*, *into*, *like*, *through*, and *throughout*.

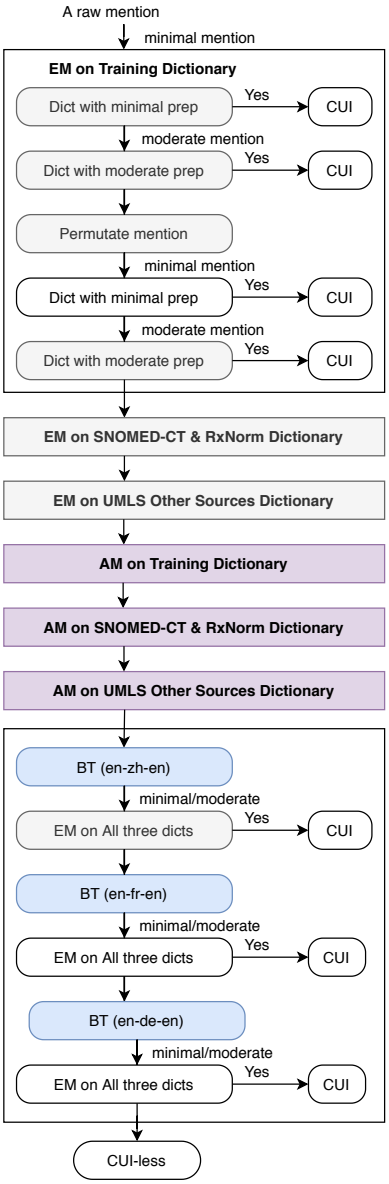


Figure 1. *Flow diagram of the method. EM, AM, BT refer to "exact match", "approximate match", and "back-translation" respectively. Every rectangular box has the same five steps as the first one.*

5. Stem with the PorterStemmer in NLTK. [2]
6. Expand abbreviations with: (1) diseases and disorders abbreviations of Wikipedia; [3] and (2) a clinical abbreviation list from NSW Health. [4]
7. Convert adverbs into adjectives, based on WordNet. [5]
8. Remove common clinical words, such as *studies*, *surgery*, *operation*, *procedure*, *preparation*, *test*, *behavior*, *well*, *phase*, *examination*, and *series*.

We refer to steps 1–2 as "**minimal** pre-processing", steps 1–5 as "**moderate** pre-processing", and steps 1–8 as "**advanced** pre-processing".

4.2. *Exact Match*

The three dictionaries mentioned in Section 5 are all made up of concept names with unique CUIs. Under pre-processing, if mention m is exactly the same as concept name n, then the corresponding CUI of n is the output of exact match.

4.3. *Approximate Match*

For mentions that do not match under exact match, two approximate matching approaches are applied: (1) contains match ("CM"); and (2) edit distance match ("ED").

4.3.1. *Contains match*

It sometimes occurs that a mention is not string-identical with its corresponding concept name, but all component tokens are contained in the concept name. For example, *nystatin ointment* cannot be exact-matched to `C1247197: Nystatin Topical Ointment`, but each token in the mention is in its corresponding concept name. Hence, we first generate a candidate list with the restriction that all tokens in the mention m must match in the concept name, and there can be at most one unmatched token in the concept name. It should be highlighted that the order of tokens is not considered during the retrieval of candidates. In the case of multiple candidate matches, we select the concept name which is shortest.

2. https://www.nltk.org/howto/stem.html.
3. https://en.wikipedia.org/wiki/List_of_abbreviations_for_diseases_and_disorders.
4. http://www.seslhd.health.nsw.gov.au/Policies_Procedures_Guidelines/Corporate/Health_Records/documents/SESLHDPR282-ClinicalAbbreviationsList.pdf.
5. https://wordnet.princeton.edu/.

4.3.2. *Edit distance*

To handle spelling errors and pluralization for single-word mentions, we calculate the character-level edit distance between each mention m and concept name n that also consists of one word, and increase the edit distance threshold up to 3 until a match is found. An empirical study to determine the optimal threshold is presented in Section 5.3.3. For time efficiency, we only compare mentions m with concept names whose length is within three characters' length of m.

4.4. *Mention Permutation*

Observing that the original token order of some mentions does not match the order of the canonical concept name, we generate a list of all possible permutations of a mention. If any variant matches the concept name, the corresponding CUI will be assigned. This permutation is applied only after failing to matching the original token order in the matching process (see Figure 1).

4.5. *Back-translation*

Inspired by recent work on translation for cross-lingual biomedical concept normalization (Perez *et al.*, 2018; Roller *et al.*, 2018), we propose a heuristic approach based on back-translation [6] ("BT") (Sennrich *et al.*, 2016). In this, we use a range of pivot languages with different linguistic properties, to help deal with derivational morphological changes, synonym replacements, and spelling errors. Our approach is straightforward: for all unmatched mentions after approximate match, we perform back translation from English to Chinese and then back to English, where Google Translate is used for en–zh, and Baidu Translate for zh–en, following Wang *et al.* (2020). In addition, to retain plural forms, we apply BT to French and German as well. [7]

4.6. *Neural Classifier*

Finally, we fine-tune a multi-class classifier, utilizing a single linear layer connected by a softmax activation function on top of a BERT encoder (Devlin *et al.*, 2019). Specifically, with the aim of providing a neural classifier baseline, the model is fine-tuned to classify an unseen mention into one of the 2,331 unique concepts available in the training data, instead of predicting over all possible concepts in SNOMED-CT and RxNorm, which contain a combined total of 434,056 CUIs. Moreover, following the BERT-based neural classification approach of Xu *et al.* (2020), we do not consider the sentence or paragraph context in which the mention occurs, but only the mention

6. Back-translation also refers to long-trip translation in prior work.
7. Using Google Translate in both directions.

text itself. These two factors lead to a big gap in performance compared with other contextualized transformer-based concept normalization systems developed for the MCN corpus (see Section 5.5), such as the TTI system (Ji *et al.*, 2020).

In training, 6,684 mention–CUI pairs are used to update the parameters of the classifier, where each mention is represented by the vector associated with the CLS-token, and each CUI is indexed by a unique ID ranging from 0 to 2,330.

This multi-class classifier is trained on the basis of the original implementation of BERT-base with 12 transformer encoders, using the pre-trained weights of Clinical-BERT (Alsentzer *et al.*, 2019), and the Adam optimizer with cross-entropy as the loss function. We apply weight decay to the optimizer with a linear scheduler and warm-up proportion of 0.1, and set the learning rate to 1e-5 and batch size to 32, thus updating 209 (6,684 / 32) steps for each epoch. Given that the accuracy improves on the dev set when we incrementally increase the training step in steps of 20k instances to 100k, we stop training at 100k steps for the final test (i.e. 479 epochs = 100k / 209). We apply the fine-tuned model to the mentions which are not matched by the rule-based approach, resulting in a hybrid system (see Section 5.4.2).

5. Experiments and Results

5.1. *Evaluation Metrics*

The standard evaluation metric used for concept normalization systems is *accuracy* (Xu *et al.*, 2020), given that the system must assign an identifier for each provided concept. Accuracy is the percentage of concept mentions that are correctly assigned CUI labels over all evaluated mentions. To assess performance of each component of the sieve, *precision* is adopted. Specifically, for a specific stage of matching, the percentage of mentions correctly assigned a CUI label, relative to the total number of matched mentions in that stage, is calculated.

5.2. *Dictionary Construction*

We constructed three dictionaries, which we employ in priority order as described below. To obtain a single CUI for a mention during matching, we apply simple disambiguation strategies. For the dictionary based on the MCN corpus training data, we retain the highest frequency CUI for an ambiguous mention. For the two dictionaries based on the UMLS Metathesaurus, we maintain the concept with the most number of concept names (synonyms). One of these dictionaries is derived from the two key dictionaries SNOMED-CT and RxNorm, while the other draws on other source vocabularies of UMLS Metathesaurus, such as MeSH, MSH, and NCI, where CUIs that are not in SNOMED-CT or RxNorm are ignored, consistent with the annotation guidelines, solely remaining concepts of SNOMED-CT or RxNorm.

Source of dictionary	Pre-processing	Unique concept name	Unique concept	UACN
① Training Data	NA	3,739	2,303	51
① Training Data	minimal	3,092	2,293	64
① Training Data	moderate	2,932	2,269	92
② SNOMED-CT & RxNorm	minimal	1,048,536	433,843	—
② SNOMED-CT & RxNorm	moderate	1,041,971	433,304	—
③ UMLS Other Sources	minimal	622,774	220,415	—
③ UMLS Other Sources	moderate	561,816	219,622	—

Table 2. *The number of unique concept names, unique concepts, and UACN ("unique ambiguous concept name": names connected to multiple concepts) in the different dictionaries. Pre-processing has three types: no pre-processing ("NA"), minimal, and moderate; see Section 4 for details.*

For convenience, we refer to the three dictionaries below according to their sources: ① Training Data, ② SNOMED-CT & RxNorm, and ③ UMLS Other Sources. Table 2 provides the statistics of these dictionaries under different pre-processing strategies.

5.3. *Optimizing the matching strategy*

In this section, we perform several ablation experiments to optimize the approach to matching. As has been demonstrated empirically, performing exact match prior to approximate (partial) match in the matching workflow results in higher precision (D'Souza and Ng, 2015).However, a number of questions remain in terms of the optimal matching approach: which dictionary should be adopted as the priority resource, which pre-processing steps should be employed in the first step, and what range of threshold value should be set for edit distance (ED) in the approximate match?In addition we should confirm the effectiveness of back-translation for the clinical concept normalization task. To answer these questions, we perform ablation studies utilizing five sample data sets derived from the training data.

We randomly split the training data into five partitions of 20% each (6,684 instances), resulting in five different groups of development and training data sets, with 1,337 and 5,347 mentions, respectively, in each group. The number of matched mentions and percentage of accurately matched mentions (precision) are used as metrics to evaluate which design choice is optimal.

5.3.1. *Dictionary priority*

Three dictionaries are leveraged during matching, derived from different resources: ① Training Data, ② SNOMED-CT & RxNorm and ③ UMLS Other Sources. Thus, there are six possible permutations to arrange the three dictionaries in order. Based on the assumption that the concept name coverage of a dictionary is independent of the matching method, these permutations are assessed in the setting of exact match, and the resulting ordering is applied consistently in all matching processes.

Dict order	Dev1	Dev2	Dev3	Dev4	Dev5	AVG
① Training	938 (97.23%)	910 (96.37%)	931 (96.89%)	950 (97.05%)	947 (96.73%)	**935 (96.85%)**
② SNOMED-CT & RxNorm	962 (81.08%)	930 (80.97%)	920 (78.59%)	963 (80.58%)	939 (83.28%)	942 (80.90%)
③ Other UMLS	1,058 (76.56%)	1,024 (75.39%)	1,014 (74.36%)	1,067 (74.98%)	1,043 (78.04%)	1,041 (75.87%)
①, ②	1,140 (95.99%)	1,118 (94.01%)	1,114 (93.99%)	1,134 (94.18%)	1,141 (95.00%)	**1,129 (94.44%)**
①, ③	1,159 (94.31%)	1,146 (91.97%)	1,146 (92.93%)	1,167 (92.72%)	1,161 (93.45%)	1,155 (93.08%)

Table 3. *Experimental results of optimizing three dictionaries priority, deciding the order during match. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

Step comb		Dev1	Dev2	Dev3	Dev4	Dev5	AVG
(a)	1–2	25 (76.00%)	33 (60.61%)	41 (75.61%)	41 (63.41%)	27 (70.37%)	33 (69.20%)
(b)	1–5	28 (75.00%)	41 (60.98%)	45 (75.56%)	47 (55.32%)	31 (70.97%)	38 (67.57%)
(c)	1–3,6,7	28 (78.57%)	36 (66.67%)	43 (79.07%)	42 (64.29%)	27 (70.37%)	**35 (71.79%)**
(d)	1–8	38 (60.53%)	43 (60.47%)	55 (65.45%)	55 (56.36%)	37 (59.46%)	45 (60.45%)

Table 4. *Ablation experiments of mentions pre-processing steps in exact match using* UMLS *Other Sources dictionary. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

In the matching phase, minimal and moderate pre-processing is applied to the five development sets. As shown in Table 3, the dictionary based on the training data ① has the highest precision although matching the smallest number of mentions. In contrast, precision using the dictionary based on other UMLS terms (beyond SNOMED-CT and RxNorm) achieves an accuracy of 75.87% on average, despite the larger coverage. As a result, the dictionary learned from the training data is set to the highest priority. Following this, the order of ①, ② and ①, ③ are evaluated, demonstrating the advantage of SNOMED-CT & RxNorm with higher average precision. Therefore, the order of dictionary is set as ① $\gg$ ② $\gg$ ③.

5.3.2. *Combinations of pre-processing steps*

After precise match using dictionaries ① and ② (row 4 in Table 3), we attempt to improve cumulative accuracy by increasing the number of matched mentions, by applying various pre-processing steps to the mention in exact match with the UMLS Other Sources dictionary. However, the approach to combining pre-processing steps and the order influences the precision. Hence, we evaluate four ways: (a) steps 1–2, (b) steps 1–5, (c) steps 1–3 followed by 6 and 7, and (d) all steps 1–8. Table 4 reveals that combination (c) steps 1–3, 6 and 7 obtains higher precision while combination (d), applying all steps, increases the overall number of matches. So this order is applied in the final system architecture.

5.3.3. *Edit distance threshold value*

The maximum edit distance threshold value also affects precision. We test edit distance thresholds from 1 to 4 across the five development sets after contain match

Threshold_max	Dev1	Dev2	Dev3	Dev4	Dev5	AVG
1	4 (75.0%)	5 (60.0%)	5 (40.0%)	1 (100%)	1 (100%)	3 (75.0%)
2	7 (42.86%)	9 (55.56%)	8 (25.0%)	7 (42.86%)	4 (25.0%)	7 (38.26%)
3	7 (42.86%)	11 (45.45%)	9 (22.22%)	9 (33.33%)	6 (33.33%)	8 (35.44%)
4	10 (30.0%)	12 (41.67%)	11 (18.18%)	10 (30.0%)	11 (18.18%)	10 (28.85%)

Table 5. *Experiments of choosing optimal maximum edit distance threshold in approximate match.* 3 *is selected considering both matched mention amount and precision, thus* $[1, 2, 3]$ *is used sequentially during matching. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

Target language	Dev1	Dev2	Dev3	Dev4	Dev5	AVG
Chinese	9 (33.33%)	9 (66.67%)	11 (63.64%)	11 (81.82%)	9 (55.56%)	9 (60.20%)
French	3 (66.67%)	2 (0.0%)	5 (80.0%)	4 (100.0%)	2 (100.0%)	3 (69.33%)
German	4 (50.0%)	2 (50.0%)	3 (66.67%)	4 (75.0%)	2 (100.0%)	3 (68.33%)

Table 6. *Experiments with exact match back-translated mentions from three target languages: Chinese, French and German using three sources dictionaries. Each cell is the number of matched mentions (precision). AVG denotes to the averaged value of the five randomly sampled dev sets.*

of AM over three dictionaries, and find that there is minimal impact on the number of matched mentions, even when the threshold is set to 4 (Table 5). To balance precision and the number of matches, we set 3 as the maximum threshold value for edit distance.

5.3.4. *Back-translation impact*

There are still unmatched mentions after applying exact and approximate match with the eight basic pre-processing steps. To assess the potential benefits of back-translation to clinical concept normalization, we perform exact match on the mentions that are back-translated from three languages using three dictionaries. As shown in Table 6, back-translated results from the three target languages all have a positive effect, with an average precision in range of 60%–70%. The number of matched mentions back-translated from Chinese is larger than French and German, leading to more accurate matched mentions. This may be attributed to the fact that Chinese is linguistically distant from English, and that a mature commercial translation solution is available from Baidu Translate. Therefore, we first match the result from Chinese, then French and German in our experiments.

We conduct the whole match process with and without back-translation, and show that across the five dev sets, the average absolute improvement in accuracy is 0.69% (Table 7).

BT	Dev1	Dev2	Dev3	Dev4	Dev5	AVG
no	1,250 (85.34%)	1,248 (83.40%)	1,244 (84.29%)	1,258 (84.29%)	1,245 (84.59%)	1,249 (84.38%)
yes	1,263 (85.79%)	1,263 (84.22%)	1,259 (84.89%)	1,274 (85.42%)	1,256 (85.04%)	1,263 (85.07%)
+	13 (0.45%)	15 (0.82%)	15 (0.60%)	16 (1.13%)	11 (0.45%)	14 (0.69%)

Table 7. *The number of matched mentions and final accuracy with (yes) / without (no) back-translation (BT). The bottom line indicates the improvement in accuracy with BT.*

To further evaluate the effectiveness of back-translation (BT), we consider its application in two additional clinical concept normalization datasets, specifically ShARe/CLEF eHealth 2013 Task 1 (Suominen *et al.*, 2013) and ShARe/CLEF eHealth 2014 Task 2a (Mowery *et al.*, 2014). Both of these data sets normalize mentions to concepts in SNOMED-CT. As RxNorm is excluded in the annotation, SNOMED-CT dictionaries with minimal and moderate pre-processing are constructed (see Section 4). In these experiments, we report a baseline that uses exact match, and then a variant which continues to match back-translated mentions to synonyms in the dictionary (again via exact match).

There are 5,816 and 11,554 (mention, CUI) pairs in ShARe/CLEF eHealth 2013 Task 1 and ShARe/CLEF eHealth 2014 Task 2a training data, respectively, where 1,639 (28.2%) and 3,478 (30.1%) pairs fall into the "CUI-less" category. Exact match is performed with each concept mention text as input against the SNOMED-CT dictionary. Without BT, all unmatched mentions are labeled as CUI-less. With BT, each unmatched mention is augmented with back-translation to match synonyms in the SNOMED-CT dictionary. As in our previous experiments, Chinese, French and German are applied sequentially as the pivot language for BT.

Table 8 shows that BT also improves the accuracy of these two clinical concept normalization datasets by 0.62% and 0.36%, respectively, in line with the results on the MCN corpus. However, back-translation using German after the other two languages hurts the performance on both datasets, although it increases the overall number of matched mentions. Error analysis reveals that this is primarily due to mentions that have a gold-standard label of "CUI-less" rather than a valid SNOMED-CT CUI. As discussed in Section 3, compared with these other datasets, the MCN corpus has only 2.7% "CUI-less" terms, therefore we expect that some of these apparent errors are in fact valid normalizations not available in the gold standard.

5.4. *Held-out Evaluation*

In Section 5.3, we described the exploration of several design choices over the sample development sets, to determine the optimal matching procedure for our rule-based method. In this section, we present the evaluation of the final process on the held-out test data set of the MCN corpus (6,925 mentions), described in Section 3. Then we analyze the mentions predicted correctly by the neural classifier.

BT	ShARe/CLEF eHealth 2013 Task 1	ShARe/CLEF eHealth 2014 2014 Task 2a
no	3,711 (61.78%)	7,412 (61.58%)
bt_zh	4,076 (62.38%)	8,080 (61.98%)
bt_fr	4,129 **(62.43%)**	8,217 **(62.06%)**
bt_ge	4,158 (62.40%)	8,285 (61.94%)
+	0.62%	0.36%

Table 8. *Experiment with/without ("no") BT on ShARe/CLEF eHealth 2013 Task 1 and ShARe/CLEF eHealth 2014 Task 2a training data sets using exact match over* SNOMED-CT. *Each cell reports the number of matched mentions and system accuracy in this stage. The last row presents the improvement in accuracy after BT with all three languages (zh, fr, and de).*

We evaluate the performance of each sieve step using the number of matched mentions, the percentage of correctly matched mentions (precision), and the final accuracy after this sieve. We note that the final accuracy is calculated by first summing the number of correctly matched mentions and the correctly-assigned CUI-less mentions among all unmatched mentions, then dividing by the total number of mentions $(6,925)$. For example, considering the first row of Table 9, $57.94\% = (3,898 + 114)/6,925$, where 114 is the number of correctly-assigned CUI-less mentions among unmatched mentions after the first sieve.

Note that sieves that do not gain any matched mentions (MMs equal to 0) are omitted in Table 9, such as approximate match (cm) using training data with both minimal and moderate pre-processing, approximate match (ed) using moderate training data and SNOMED-CT & RxNorm, as well as UMLS Other Sources with minimal and moderate pre-processing. Moreover, approximate match (ed) using Training Data (minimal) does not contribute to matched mentions with a threshold of 3, and similarly, no mentions are matched using SNOMED-CT & RxNorm (minimal) with a threshold setting of 2, for example.

5.4.1. *Rule-based Method*

As shown in Table 9, exact match predicts more accurately than approximate match. Specifically, exact match obtains more correctly-matching mentions over the same number of matched mentions, resulting in higher precision. In terms of dictionaries, the training dictionary is the most accurate but provides limited variations of concept names (see Section 5.3.1). The dictionary built on SNOMED-CT & RxNorm vocabularies has higher accuracy than UMLS Other Sources, while including many more concept names. Therefore we employed the training dictionary first, then SNOMED-CT & RxNorm, and lastly UMLS Other Sources in matching. Importantly, back-translation increased absolute accuracy by 0.46%, with 33 mentions correct out of 43 matched mentions.

Match type	Dictionary	Ignore order	MMs	AMMs (%)	Cum-AMMs+Cor CUI-less	Cum-Accuracy
Exact	Training Data (minimal)	no	4,025	3,898 (96.84%)	3,898+114	57.94%
Exact	Training Data (moderate)	no	197	138 (70.05%)	4,036+106	59.81%
Exact	Training Data (minimal)	yes	7	6 (85.71%)	4,042+106	59.90%
Exact	Training Data (moderate)	yes	17	16 (94.12%)	4,058+106	60.13%
Exact	SNOMED-CT & RxNorm (minimal)	no	1,103	969 (87.85%)	5,027+100	74.04%
Exact	SNOMED-CT & RxNorm (moderate)	no	192	125 (65.10%)	5,152+96	75.78%
Exact	SNOMED-CT & RxNorm (minimal)	yes	2	2 (100.00%)	5,154+96	75.81%
Exact	SNOMED-CT & RxNorm (moderate)	yes	57	53 (92.98)	5,207+96	76.58%
Exact	UMLS Other Sources (prep 1,2,3,6,7)	no	279	193 (69.18%)	5,400+86	79.22%
Exact	UMLS Other Sources (minimal)	no	2	2 (100.00%)	5,402+86	79.25%
Exact	UMLS Other Sources (moderate)	no	14	11 (78.57%)	5,413+85	79.39%
Exact	UMLS Other Sources (minimal)	yes	3	2 (66.67%)	5,415+85	79.42%
Exact	UMLS Other Sources (moderate)	yes	3	2 (66.67%)	5,417+85	79.45%
Exact	UMLS Other Sources (advanced)	no	78	21 (26.92%)	5,438+82	79.71%
Approximate (cm)	SNOMED-CT & RxNorm (minimal)	NA	251	76 (30.28%)	5,514+62	80.52%
Approximate (cm)	SNOMED-CT & RxNorm (moderate)	NA	120	47 (39.17%)	5,561+58	81.14%
Approximate (cm)	UMLS Other Sources (minimal)	NA	12	9 (75.00%)	5,570+57	81.26%
Approximate (cm)	UMLS Other Sources (moderate)	NA	7	5 (71.43%)	5,575+57	81.33%
Approximate (ed:1,2)	Training Data (minimal)	NA	18	11 (61.11%)	5,586+54	81.44%
Approximate (ed:1,3)	SNOMED-CT & RxNorm (minimal)	NA	33	10 (30.30%)	5,596+52	81.56%
Exact (Chinese)	All Dicts (minimal/moderate)	no	24	19 (79.17%)	5,615+51	81.82%
Exact (French)	Two UMLS Dicts (minimal/moderate)	no	13	11 (84.62%)	5,626+51	81.98%
Exact (German)	Two UMLS Dicts (minimal/moderate)	no	6	3 (50.00%)	5,629+51	82.02%

Table 9. *Evaluation result of each sieve in the multiple passes. "MMs", "AMMs", "Cum-AMMs", "Cor CUI-less" and "Cum-Accuracy" refer to Matched mentions, Accurate matched mentions, Cumulative accurate matched mentions, Correctly-assigned CUI-less mentions and final Cumulative Accuracy, respectively. Note that Cum-Accuracy = (Cum-AMMs+Cor CUI-less) /* $6,925$. *Sieves that do not gain any matched mentions (MMs equal to* 0*) are omitted.*

	Rule-based	Neural	Hybrid
True	5,629	70	5,699
False	834	392	1,226
	6,463	462	6,925

Table 10. *The number of correct and incorrect CUIs predicted by the rule-based, neural and hybrid systems.*

5.4.2. *Combined Classifier*

We apply the neural classifier on the 462 mentions which are assigned CUI-less by the rule-based method, resulting in 70 additional correct mentions, as presented in Table 10.

We observe that among the 70 mentions, 11 are CUI-less. Following the features and changes of terms in Cohen *et al.* (2010), the remaining 59 mentions are grouped into seven types related to variations in the concept strings, as listed below. Most cases involve more than one such source of variation.

1) British/American English spelling differences (*-*sation* vs. *-*zation*)

2) singular/plural variants

N2C2 Team/Method Name	Accuracy
EM-UMLS (Luo *et al.*, 2019)	69.52%
EM-UMLS (removing common word tokens)	76.35%
EM-Train	51.75%
EM-Train (removing common word tokens)	76.27%
MetaMap	75.65%
MetaMap (removing common word tokens)	76.35%
Toyota Technological Institute (deep learning)	**85.26%**
Kaiser Permanente (rule-based)	81.94%
University of Arizona (rule-based and deep learning)	81.66%
Med Data Quest, Inc. (rule-based)	81.01%
Lucene (rule-based) (Xu *et al.*, 2020)	79.25%
Lucene+BERT-rank (rule-based and deep learning)	82.75%
Lucene+BERT-rank+ST-reg (rule-based and deep learning)	83.56%
Neural multi-class classifier (deep learning)	62.35%
Multi-pass sieve incorporating back translation (rule-based)	82.02%
Hybrid system of rule-based and neural classifier (rule-based and deep learning)	82.30%

Table 11. *Accuracy of our methods (bottom half) compared with top systems participating in N2C2 Track 3 Shared Task and recent SOTA hybrid system (Xu* et al.*, 2020) (middle half) and baselines (upper half) presented in MCN (Luo* et al.*, 2019). ST-reg refers to Semantic Type Regularization. The bold number is the best accuracy on MCN.*

3) reordering

4) inserted words (such as *blood*, *injection*, and *visual*)

5) removed words and hyphens (removed words include *body*, *screen*, *placement*, *measurement*, *arrest*, *activity*, *study*, and *cause*)

6) alternative expression of numerals (*30%* vs. *partial*)

7) synonym replacement consisting of morphological conversion from the same root and completely different words

5.5. *Comparison with Other Systems*

We compare our method with the top systems that participated in the N2C2 Track 3 shared task and baselines of MCN in Table 11. Toyota Technological Institute (TTI) attained the best result in the shared task, peaking at 85.26% accuracy with an ensemble model of five individually-trained BERT-based models. Our purely rule-based method achieves 82.02%, outperforming all the rule-based systems that participated in the shared task, and slightly better than the hybrid method from the University of Arizona which achieved an accuracy of 81.66%.

Due to the resource-hungry nature of deep learning algorithms, TTI requires a significant amount of computational power, and a huge memory footprint due to the incorporation of BERT. In addition to the dependency on large scale corpora, training BERT is time consuming, taking days to converge even with the support of multiple GPUs. These factors severely limit its applicability. In comparison to this complicated system, our rule-based method is solely reliant on the vocabularies of UMLS Metathesaurus, and therefore much more efficient in terms of time and computational resources.

Three major differences exist between the Kaiser Permanente (KP) rule-based method (the top-performing rule-based method in the N2C2 evaluation) and ours:

1. we utilize distinct strategies for approximate match. In our approach, we look for the corresponding concept name by judging whether the component tokens of the mention are contained in concept names or the edit distance is within the pre-defined threshold, while the KP system searches similar concept names based on character 3-grams;
2. the KP system used only SNOMED-CT and RxNorm, while we additionally incorporated other sources from UMLS;
3. the incorporation of back-translation in our approach.

The upper half of Table 11 includes the results of six baseline systems prepared by the organizers of the MCN shared task.These include two methods without access to the training data: exact match based on UMLS (EM-UMLS) and MetaMap in two settings, with and without removing common word tokens from the original mentions. An additional two baseline systems leverage the training data only to infer a dictionary, and are matched using exact match. All of these baselines have lower performance than both the other N2C2 submissions and our reported sieve-based methods.

Xu *et al.* (2020) also proposed a hybrid system, which differs from our ensemble approach in that they combine the rule-based candidate generator and neural ranker together internally, as components of an integrated normalization system rather than independent methods.

6. Error Analysis

We first compare the rule-based method and the neural method, investigating their respective strengths and weaknesses, and their common failures. The results of the rule-based method are then further analyzed, dividing cases into matched CUIs, and mentions not normalized to a CUI (assigned "CUI-less").

6.1. *Rule-based vs. Neural Methods*

As shown in Table 12, 4,338 mentions are assigned the same CUIs by the two methods, of which 4,160 are correct. For the rule-based method, 96.27% $(4,005/4,160)$

	Agreement	Rule-based method	Neural method
True	4,160 (60.07%)	1,520 (21.95%)	158 (2.28%)
False	178 (2.57%)	1,067 (15.41%)	2,429 (35.08%)
	4,338 (62.64%)	2,587 (37.36%)	2,587 (37.36%)

Table 12. *The number and proportion of correct and incorrect predictions over test data by the rule-based method and neural method.*

mentions are identified through exact match in the training data dictionary only, showing that performance on the task benefits from the labelling of real-world usage of the terms. Of the 2,587 mentions on which both methods do not agree, 1,520 are identified correctly by the rule-based method while only 158 are found by the neural method. The overall accuracy is 82.02% and 62.35%, respectively.

We observe that the neural classifier has substantially lower accuracy than the rule-based method, largely due to the limitation that the neural model can only learn for the 2,331 CUIs instantiated in the training data. The rule-based method is able to generalize more readily to unseen cases, by incorporating the vocabularies of UMLS. 80.61% $(1,958/2,429)$ of the incorrect predictions from the neural classifier can be attributed to lacking relevant examples in the training data, involving 1,392 unseen CUIs. However, as illustrated in Section 5.4.2, due to the use of word embeddings, the neural classifier is less sensitive to simple variations, such as removing or adding a word, and changing from plural to singular form. To analyze the common flaws, we categorize the 178 erroneous mentions shared by the two approaches into five error types (see Figure 2). We find that ambiguity contributes to the majority of errors, and requires context to resolve. In detail:

1. **Semantic type ambiguity** (90 cases): the same concept name maps to multiple concepts with different semantic types, and is context dependent;
2. **Training data misalignment** (49 cases): the mention can be correctly matched via the SNOMED-CT & RxNorm dictionary, but prioritizing the training data-derived dictionary introduced error. For instance, *monitor* maps to `C1292786: Observation - action` in the training data, while `C0181904: Monitor` would be selected via the SNOMED-CT and RxNorm dictionary;
3. **Underspecification** (26 cases): some mentions offer insufficient information to identify the corresponding CUI. For example, the CUI for *Calcification of breast* cannot be assigned based on the mention *Calcification* without further information related to the location of the calcification;
4. **Abbreviation ambiguity** (8 cases): the same lexical abbreviation may correspond to multiple concepts. For example, *lh* can refer to either *light-headedness* or *Luteinizing hormone*. These cases require context to make a correct assignment;

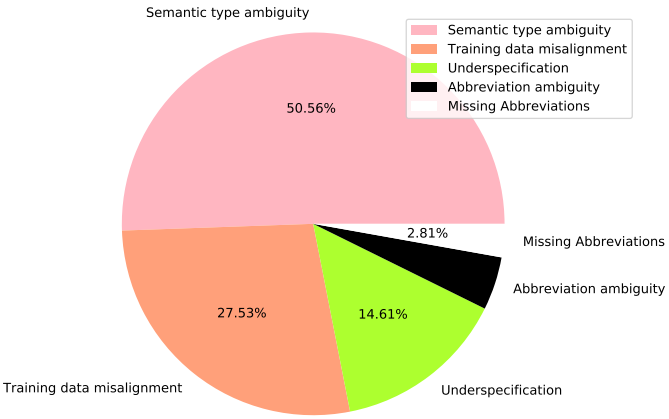


Figure 2. *Percentage of five major factors leading to the 178 commonly incorrect predictions of rule-based method and neural classifier.*

	Matched	Unmatched	Total
True	5,629	51	5,680
False	834	411	1,245
	6,463	462	6,925

Table 13. *The number of accurate and inaccurate assignments in mentions that are assigned CUI-less.*

5. **Missing abbreviations** (5 cases): for terms corresponding to abbreviations missing in the dictionary, the system fails to match a CUI. For instance, *Procan SR* corresponds to *Procainamide Extended Release Oral Tablet*, and *GI* in *further gi testing* (which stands for *gastrointestinal*) is missing. We expect that detection and expansion of abbreviations within mentions will help in such cases.

6.2. *Error Analysis of the Rule-based Method*

We perform error analysis of the rule-based method, considering two cases: mentions incorrectly assigned CUIs through matching (False Positives, 834 cases), and mentions unmatched to a CUI (False Negatives, 411 cases); see Table 13.

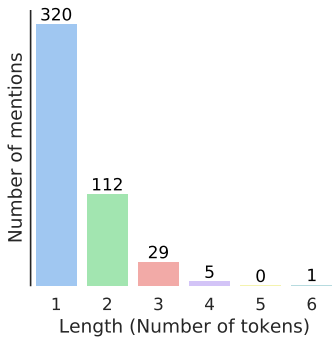


Figure 3. *Length distribution of 467 cases of incorrect normalizations not due to disambiguation errors (number of tokens).*

For the 834 cases erroneously assigned CUIs, the mentions closely match mentions in the training data or concept names in a source dictionary. Most errors are due to inadequate disambiguation, either of semantic type or abbreviation resolution. By examining mentions where the predicted and gold standard CUIs have lexically similar concepts, but have different semantic types, we identify 44% (367/834) mentions which lack appropriate disambiguation. For instance, the mention *Q-waves* was matched to `C1287077:Q-wave finding` with semantic type of `T033:Finding`, rather than the correct `C1305738:Q-wave feature` with `T201:Clinical Attribute`.

In the remaining 56% (467/834) of incorrect normalizations, as presented in Figure 3, there are 320 one-token, 112 two-token, and 35 multiple-token mentions. We find that 139/320, 52/112, and 5/35 of these are due to abbreviation ambiguity. Length impacts variability: limited variation of shorter strings facilitates lexical matching, but a simple disambiguation strategy leads to incorrect assignments. Considering the other 30/35 multiple-token mentions, errors result from: (1) matching to an overly specific concept, such as matching *injury to eyes* to `C0339055:Injury of globe of eye` (17 cases); and (2) matching to an overly general concept (13 cases).

Analyzing the 411 unmatched mentions by length, in contrast to the matched mentions above, single-token mentions are in the minority with only 13 cases, while there are 374 mentions 2–5 tokens in length, and the maximum length is 12 tokens (see Figure 4). Longer mentions are associated with significant variability, such that a large proportion of mentions are substantially lexically distinct from any synonym of a corresponding CUI. Presence of punctuation (61 cases: 40 mentions contain dash (-), and 21 mentions involve punctuation marks in the set {. % , # / () ' & ;+ }.),

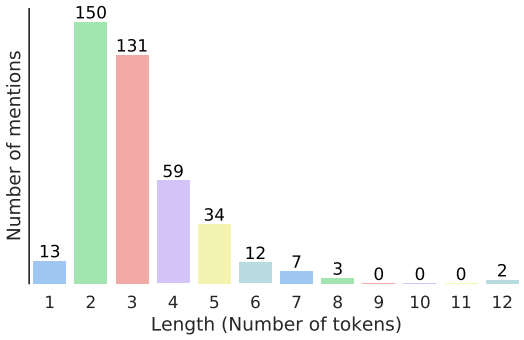


Figure 4. *Length distribution of 411 unmatched mentions (number of tokens).*

stopwords [8] (110 cases), and numeral mismatches (24 cases where numbers were in a different form from the concept – words, or Arabic or Roman numerals) further contribute to mismatches.

Term variations in mentions with more than six tokens involve a mixture of word reordering, synonym replacement, abbreviation expansion, stopword removal, numeral conversion and summation (*IV plus V* to *9*), and even summarization (*diminution of light touch, pinprick, position, and vibration sense* to `C0020580:Hypesthesia`). To resolve such cases, more sophisticated methods are required.

7. Conclusion

In this study, we presented a multi-pass sieve approach based on UMLS Metathesaurus with various preprocessing strategies. Our method achieves a new benchmark among rule-based methods on the Medical Concept Normalization corpus, with 82.02% accuracy. In addition, we empirically investigated the use of back-translation for the clinical concept normalization task, and achieved promising results. Our final system integrated a neural classifier to gain a modest 0.28% improvement in accuracy. Error analysis reveals that more consideration of context is required to distinguish ambiguous concept names, corresponding to multiple semantic types; we will consider this in future work.

8. Stopwords from `https://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/DATASET/`.

Acknowledgements

This work was supported by China Scholarship Council (CSC) and the University of Melbourne.We are grateful to the anonymous reviewers for their insightful comments.

8. References

Aamer H., Ofoghi B., Verspoor K., "Syndromic Surveillance through Measuring Lexical Shift in Emergency Department Chief Complaint Texts", *Proceedings of the Australasian Language Technology Association Workshop 2016*, Melbourne, Australia, p. 45-53, December, 2016.

Alsentzer E., Murphy J., Boag W., Weng W.-H., Jindi D., Naumann T., McDermott M., "Publicly Available Clinical BERT Embeddings", *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA, p. 72-78, 2019.

Aronson A. R., "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program", *Proceedings of the AMIA Symposium*, p. 17, 2001.

Bodenreider O., "The Unified Medical Language System (UMLS): integrating biomedical terminology", *Nucleic Acids Research*, vol. 32, p. D267-D270, 2004.

Cohen K. B., Roeder C., Baumgartner Jr. W. A., Hunter L. E., Verspoor K., "Test Suite Design for Biomedical Ontology Concept Recognition Systems", *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May, 2010.

Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *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)*, Minneapolis, Minnesota, p. 4171-4186, 2019.

D'Souza J., Ng V., "Sieve-Based Entity Linking for the Biomedical Domain", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, p. 297-302, July, 2015.

Elhadad N., Pradhan S., Gorman S., Manandhar S., Chapman W., Savova G., "SemEval-2015 task 14: Analysis of clinical text", *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 303-310, 2015.

Funk C., Baumgartner W., Garcia B., Roeder C., Bada M., Cohen K. B., Hunter L. E., Verspoor K., "Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters", *BMC Bioinformatics*, vol. 15, nº 1, p. 59, 2014.

Gong C., He D., Tan X., Qin T., Wang L., Liu T.-Y., "Frage: Frequency-agnostic word representation", *Advances in Neural Information Processing Systems*, p. 1334-1345, 2018.

Groza T., Verspoor K., "Automated generation of test suites for error analysis of concept recognition systems", *Proceedings of the Australasian Language Technology Association Workshop 2014*, p. 23-31, 2014.

Ji Z., Wei Q., Xu H., "Bert-based ranking for biomedical entity normalization", *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 269, 2020.

Kang N., Singh B., Afzal Z., van Mulligen E. M., Kors J. A., "Using rule-based natural language processing to improve disease normalization in biomedical text", *Journal of the American Medical Informatics Association*, vol. 20, n° 5, p. 876-881, 2013.

Leaman R., Islamaj Doğan R., Lu Z., "DNorm: disease name normalization with pairwise learning to rank", *Bioinformatics*, vol. 29, n° 22, p. 2909-2917, 2013.

Leaman R., Lu Z., "Automated disease normalization with low rank approximations", *Proceedings of BioNLP 2014*, p. 24-28, 2014.

Leaman R., Lu Z., "TaggerOne: joint named entity recognition and normalization with semi-Markov Models", *Bioinformatics*, vol. 32, n° 18, p. 2839-2846, 2016.

LePendu P., Liu Y., Iyer S., Udell M. R., Shah N. H., "Analyzing patterns of drug use in clinical notes for patient safety", *AMIA Summits on Translational Science Proceedings*, vol. 2012, p. 63, 2012.

Li H., Chen Q., Tang B., Wang X., Xu H., Wang B., Huang D., "CNN-based ranking for biomedical entity normalization", *BMC Bioinformatics*, vol. 18, n° 11, p. 79-86, 2017.

Limsopatham N., Collier N., "Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, p. 1014-1023, August, 2016.

Liu S., Ma W., Moore R., Ganesan V., Nelson S., "RxNorm: prescription for electronic drug information exchange", *IT Professional*, vol. 7, n° 5, p. 17-23, 2005.

Logeswaran L., Chang M.-W., Lee K., Toutanova K., Devlin J., Lee H., "Zero-Shot Entity Linking by Reading Entity Descriptions", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, p. 3449-3460, July, 2019.

Luo Y.-F., Sun W., Rumshisky A., "MCN: A comprehensive corpus for medical concept normalization", *Journal of Biomedical Informatics*, vol. 92, p. 103-132, 2019.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems*, p. 3111-3119, 2013.

Mowery D. L., Velupillai S., South B. R., Christensen L., Martinez D., Kelly L., Goeuriot L., Elhadad N., Pradhan S., Savova G. *et al.*, "Task 2: ShARe/CLEF eHealth evaluation lab 2014", *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2014.

Perez N., Cuadros M., Rigau G., "Biomedical term normalization of EHRs with UMLS", *arXiv preprint arXiv:1802.02870*, 2018.

Pradhan S., Chapman W., Man S., Savova G., "Semeval-2014 task 7: Analysis of clinical text", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014.

Pradhan S., Elhadad N., South B. R., Martinez D., Christensen L. M., Vogel A., Suominen H., Chapman W. W., Savova G. K., "Task 1: ShARe/CLEF eHealth Evaluation Lab 2013", *CLEF (Working Notes)*, 2013.

Reimers N., Gurevych I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, p. 3982-3992, November, 2019.

Roller R., Kittner M., Weissenborn D., Leser U., "Cross-lingual candidate search for biomedical concept normalization", *arXiv preprint arXiv:1805.01646*, 2018.

Savova G. K., Masanz J. J., Ogren P. V., Zheng J., Sohn S., Kipper-Schuler K. C., Chute C. G., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications", *Journal of the American Medical Informatics Association*, vol. 17, nº 5, p. 507-513, 2010.

Sennrich R., Haddow B., Birch A., "Improving Neural Machine Translation Models with Monolingual Data", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, p. 86-96, August, 2016.

Shah N. H., Bhatia N., Jonquet C., Rubin D., Chiang A. P., Musen M. A., "Comparison of concept recognizers for building the Open Biomedical Annotator", *BMC Bioinformatics*, vol. 10-S9, p. S14, 2009.

Spackman K. A., Campbell K. E., Côté R. A., "SNOMED RT: a reference terminology for health care", *Proceedings of the AMIA Annual Fall Symposium*, p. 640, 1997.

Suominen H., Salanterä S., Velupillai S., Chapman W. W., Savova G., Elhadad N., Pradhan S., South B. R., Mowery D. L., Jones G. J. *et al.*, "Overview of the ShARe/CLEF eHealth evaluation lab 2013", *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, p. 212-231, 2013.

Topaz M., Lai K., Dowding D., Lei V. J., Zisberg A., Bowles K. H., Zhou L., "Automated identification of wound information in clinical notes of patients with heart diseases: Developing and validating a natural language processing application", *International Journal of Nursing Studies*, vol. 64, p. 25-31, 2016.

Wang Y., Liu F., Verspoor K., Baldwin T., "Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity", *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Online, p. 105-111, July, 2020.

Weissman G. E., Hubbard R. A., Ungar L. H., Harhay M. O., Greene C. S., Himes B. E., Halpern S. D., "Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay", *Critical Care Medicine*, vol. 46, nº 7, p. 1125, 2018.

Xu D., Zhang Z., Bethard S., "A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 8452-8464, July, 2020.

Détection de la somnolence dans la voix : nouveaux marqueurs et nouvelles stratégies

Vincent P. Martin* — **Jean-Luc Rouas*** — **Pierre Philip****

* *{vincent.martin, rouas}@labri.fr*
LaBRI – Univ. Bordeaux – Bordeaux INP – CNRS – UMR 5800
** *pierre.philip@u-bordeaux.fr*
SANPSY – Univ. Bordeaux – CHU Pellegrin – CNRS – USR 3413

RÉSUMÉ. *Cet article traite de la détection automatique de la somnolence dans la voix en vue de l'amélioration du suivi des patients souffrant de maladies neuropsychiatriques chroniques. Notre première approche s'inspire des systèmes état de l'art mais en les appliquant au cas particulier des patients atteints de somnolence diurne excessive (SDE). Nous basons notre étude sur le corpus TILE, qui diffère des autres corpus existants par le fait que les sujets enregistrés souffrent de SDE et que leur niveau de somnolence est mesuré de manière subjective mais aussi objective. Le système proposé permet de détecter la somnolence objective grâce à des paramètres vocaux simples et explicables à des non-spécialistes. Par la suite, nous avons développé une approche originale basée sur les erreurs de lecture que nous avons confrontées aux différentes mesures de somnolence du corpus. Nous montrons ici que relever ces erreurs peut être utile pour élaborer des marqueurs robustes de la somnolence objective.*

ABSTRACT. *This paper deals with automatic sleepiness state estimation using speech applied to the follow-up of patients suffering from chronical neuropsychiatric diseases. Our first approach draws from state-of-the-art systems to estimate sleepiness level from voice, for the specific case of patients suffering from Excessive Daytime Sleepiness (EDS). We base our study on the MSLT corpus, that differs from other existing corpus by the fact that recorded subjects suffer from EDS and that their sleepiness level is measured by both subjective and objective means. The proposed system allows to detect objective sleepiness with simple vocal markers that are explainable to non-specialists. Furthermore, we devised a new method based on the reading errors and investigate their links with sleepiness measurements. We show that evaluating these reading errors can be useful to elaborate robust markers of objective sleepiness.*

MOTS-CLÉS : *détection de la somnolence, marqueurs vocaux, erreurs de lecture.*

KEYWORDS: *sleepiness detection, vocal markers, reading mistakes.*

1. Introduction

Dans un contexte de désertification médicale et d'augmentation de la demande médicale dans le domaine des pathologies neuropsychiatriques, la capacité de service offerte par les structures spécialisées n'est plus suffisante pour le suivi correct des patients. En effet, de nombreuses pathologies neuropsychiatriques chroniques nécessitent un suivi continu des patients afin de quantifier les symptômes et prévenir les rechutes précoces. Cependant, les nombreux entretiens nécessaires à une bonne prise en charge sont souvent irréguliers et ne permettent pas de mesurer les variations des différents symptômes en réponse au traitement lorsque les patients sont à leur domicile, dans des conditions écologiques. Dans ces conditions écologiques, différentes caractéristiques physiques peuvent toutefois être mesurées grâce à des dispositifs médicaux connectés tels que le poids, la pression sanguine ou encore l'activité physique. En revanche, des informations cruciales pour le suivi de ces patients comme la fatigue, la somnolence ou l'humeur des patients ne peuvent être mesurées par ces dispositifs.

Ainsi est née l'idée de créer un dispositif de suivi des patients à leur domicile sous la forme d'un agent conversationnel, élaboré en priorisant son acceptabilité lors d'entretiens réguliers et répétés (Philip *et al.*, 2020 ; Philip *et al.*, 2017). Cet agent, implémenté dans un smartphone que le patient ramène à son domicile, permet de régulariser la mesure des symptômes par le biais de questions posés par l'agent, tout en mesurant leur manifestation dans des conditions écologiques. Durant l'entretien avec l'agent virtuel, le patient a la possibilité de répondre sous la forme d'une interaction vocale. Notre but est de compléter l'analyse des réponses aux différents questionnaires proposés par le médecin virtuel en y ajoutant l'analyse de paramètres vocaux : il est désormais possible de détecter des indices permettant d'évaluer l'état du locuteur directement dans sa voix (Cummins *et al.*, 2018). Les avantages de cette méthode sont nombreux : elle ne nécessite pas de matériel ou de calibration complexe, ne repose pas sur des capteurs spécifiques et elle peut être mise en place dans des environnements variés, permettant ainsi un suivi régulier et non invasif des patients. La détection de la somnolence dans la voix en particulier est un sujet ayant déjà fait l'objet de nombreuses études, avec un pic d'intérêt lors de la compétition ajointe à la conférence Interspeech 2011 sur la détection d'état de locuteur (Schuller *et al.*, 2011) ou plus récemment celle associée à la conférence Interspeech 2019 sur l'estimation continue de la somnolence (Schuller *et al.*, 2019).

Les présents travaux, basés sur l'exploitation du corpus TILE (Martin *et al.*, 2020) enregistré au centre hospitalier universitaire de Bordeaux, se distinguent des précédents par trois aspects principaux.

Tout d'abord, les sujets enregistrés dans les bases de données associées aux compétitions (resp. le *Sleepy Language Corpus* et le corpus SLEEP pour Interspeech 2011 et Interspeech 2019) sont tous des sujets sains placés en privation de sommeil. Le médecin virtuel est destiné à une population se plaignant de somnolence diurne excessive (SDE) ayant potentiellement pour origine une maladie du sommeil : non seulement ces patients ont une perception de leur somnolence qui est différente de celle des

sujets sains, mais ils souffrent généralement de facteurs de comorbidité tels que la fatigue ou des troubles de l'humeur, susceptibles de venir interférer avec la mesure de la somnolence dans la voix.

Ensuite, l'élaboration du médecin virtuel nécessite une collaboration étroite avec le milieu médical, dont le but n'est pas tant l'implémentation d'un classificateur que la compréhension des phénomènes liés à la somnolence qui s'expriment dans la voix. La plupart des systèmes de l'état de l'art utilisent les marqueurs vocaux fournis lors des compétitions (calculés grâce à la boîte à outils openSMILE (Eyben et Schuller, 2015)), qui sont non seulement très nombreux (4 368) mais ne sont interprétables que pour des spécialistes en traitement du signal vocal. Nous souhaitons donc proposer un ensemble de marqueurs vocaux et une stratégie de classification permettant de conserver le sens des marqueurs vocaux pour pouvoir les lier, dans le cas de performances de classification acceptables, à des processus neuromoteurs ou cognitifs.

Enfin, nous désirons suivre la somnolence des patients lors de leur utilisation du médecin virtuel. Cela peut se faire selon deux modalités temporelles : soit une estimation de la somnolence à court terme, c'est-à-dire l'état du sujet sur des courtes périodes de temps, soit un suivi de la somnolence à plus long terme, qui dénote l'état habituel de somnolence du locuteur sur des échelles temporelles plus grandes et qui est un marqueur de maladies neuropsychiatriques (facteur « trait » du patient). Les deux échelles temporelles de somnolence peuvent être mesurées selon deux modalités : de manière objective, par électroencéphalogramme (EEG), ou de manière subjective, par le biais de questionnaires que remplissent les patients. À notre connaissance toutes les études menées jusqu'alors – à de rares exceptions près – se sont concentrées sur la détection de la somnolence subjective à court terme, souvent mesurée au moyen d'un questionnaire médical subjectif de somnolence comme le questionnaire de somnolence de Karolinska – KSS (Åkerstedt et Gillberg, 1990). Cette mesure de la somnolence est à la fois subjective et instantanée : l'échelle mesure l'état ressenti de somnolence du locuteur sur une période d'une dizaine de minutes. Pour un suivi à plus long terme, au contraire, nous cherchons à estimer un marqueur « trait » du locuteur, valable sur une longue durée. Pour cela, le seul corpus proposant de telles mesures est la base TILE, décrite à la section 2.

Notre objectif est donc double : d'une part, proposer une approche basée sur des marqueurs vocaux interprétables permettant la détection de la somnolence subjective à court terme dans la voix, pour des sujets sains (SLC) et des sujets atteints de SDE (base TILE). D'autre part, mettre au point une méthodologie pour la détection de la somnolence à long terme dans la voix, chez les patients souffrant de SDE (base TILE). Pour cela, nous proposons deux approches : l'une basée sur les marqueurs vocaux liés à la qualité de la voix du locuteur, l'autre sur les erreurs de lecture.

Cet article est organisé de la façon suivante. Les corpus utilisés dans cette étude sont présentés dans la section 2 tandis que les marqueurs vocaux sont introduits dans la section 3. La section 4 propose une taille minimale des échantillons audio pour la détection de la somnolence avec ces marqueurs audio. Les sections 5 et 6 présentent les méthodologies, résultats et discussions sur la détection de la somnolence à court

terme et à long terme grâce à des marqueurs basés sur la qualité de la voix, tandis que la section 7 introduit un nouveau paradigme de détection de la somnolence à long terme grâce aux erreurs de lecture des locuteurs. La conclusion et nos futurs travaux sur le sujet sont évoqués dans la section 8.

2. Corpus

2.1. *Corpus TILE*

Le corpus TILE (*Multiple Sleep Latency Test – MSLT – database* en anglais) est un corpus contenant les enregistrements de 106 patients ayant des plaintes concernant leur sommeil et venant passer un examen médical pour un suivi ou un diagnostic à la clinique du sommeil du centre hospitalier universitaire de Bordeaux. Durant cet examen, le test itératif de latence d'endormissement – TILE, il est demandé aux patients de faire une sieste de maximum 35 minutes toutes les 2 heures à partir de 9 heures du matin (Littner *et al.*, 2005). Une fois le test lancé, les patients ont 20 minutes pour s'endormir. S'ils y parviennent, le test continue et ils sont réveillés 15 minutes plus tard. Dans le cas contraire, l'itération du test est arrêtée. Les valeurs de TILE, qui correspondent aux latences d'endormissement des patients à chaque sieste, sont donc comprises entre 0 et 20 minutes.

Entre les siestes, les patients sont libres d'effectuer l'activité de leur choix (hors activité physique et consommation de stimulants tels que café ou thé), mais doivent arrêter de fumer au moins 30 minutes avant chaque sieste. Les enregistrements vocaux sont effectués une dizaine de minutes avant chaque sieste, lors de la lecture de textes d'environ 200 mots issus du *Petit Prince* d'Antoine de Saint-Exupéry (de Saint-Exupéry, 1943). Les textes sont différents pour chaque sieste mais identiques pour tous les patients à sieste identique. Tous les enregistrements ont une durée supérieure à 50 secondes. Ce texte a été choisi pour sa simplicité de vocabulaire et de grammaire, tout en ayant un contenu qui ne soit ni stimulant, ni relaxant, afin de ne pas interférer avec le test en cours.

Le corpus TILE permet l'étude de deux types de mesures de somnolence à des échelles temporelles différentes. Les échantillons audio sont étiquetés à la fois avec un questionnaire médical subjectif de somnolence instantanée et une valeur de somnolence objective rendant compte d'un facteur trait du locuteur.

2.1.1. *Somnolence subjective instantanée*

La somnolence subjective instantanée est mesurée dans la base TILE grâce à la version française du questionnaire médical KSS (*Karolinska Sleepiness Scale* (Åkerstedt et Gillberg, 1990)). Son échelle de notation va de 1 – « très éveillé » – à 9 – « très somnolent, avec de grands efforts pour rester éveillé, luttant contre le sommeil ». Il a une précision temporelle de l'ordre d'une dizaine de minutes (la consigne en français du questionnaire précise « au cours des dix dernières minutes »). Dans l'optique d'une classification binaire, il faut définir une limite sur le KSS permettant de distinguer des

enregistrements de voix somnolentes de celles non somnolentes. Si la limite la plus utilisée dans l'état de l'art est de 7,5 (Schuller *et al.*, 2011), nous préférons prendre celle de 7, pour deux raisons : non seulement cette limite correspond alors à un intitulé du questionnaire (« somnolent, mais sans effort pour rester éveillé »), mais une précédente étude sur le SLC a également montré qu'elle permet de meilleurs scores de classification (Martin *et al.*, 2019). Les échantillons associés à un KSS inférieur à 7 seront considérés comme produit par un locuteur non somnolent tandis que ceux associés à un KSS supérieur ou égal à 7 seront considérés comme produits par un locuteur somnolent.

2.1.2. *Détection de facteurs traits dans la voix*

En plus de permettre l'association entre des échantillons isolés et la somnolence subjective instantanée, ce corpus fournit des informations permettant l'estimation de traits propres aux locuteurs. En effet, la somnolence objective est mesurée par le temps d'endormissement à chaque sieste, mesure médicale liée aux variations de performances (Carskadon *et al.*, 1981) et appelée ici « valeur de TILE ». Le TILE est un test médical pour la détection de la narcolepsie lorsque la latence moyenne d'endormissement est inférieure à 8 minutes (Aldrich *et al.*, 1997). Cette mesure donne des informations sur un trait du locuteur, sur sa propension à la somnolence durant une longue période de temps. Nous réutilisons cette limite de 8 minutes pour notre tâche de classification de la somnolence. Par ailleurs, même si les seules mesures utilisées ici sont le KSS et la valeur de TILE, les patients de ce corpus sont largement phénotypés sur leur pathologie à travers de nombreux questionnaires subjectifs de somnolence, fatigue, anxiété, dépression, insomnie, addiction, etc. Ces mesures sont complétées de données physiques telles que la taille, le poids ou encore le tour de cou.

Un bref aperçu de ce corpus est présenté dans le tableau 1. Pour plus d'informations sur ce corpus, sa méthodologie de conception et les différentes mesures collectées, nous redirigeons le lecteur vers l'article le présentant (Martin *et al.*, 2020).

2.2. Sleepy Language Corpus (SLC)

Le SLC (*Sleepy Language Corpus*) est le corpus le plus utilisé dans l'état de l'art pour l'élaboration de systèmes de détection de la somnolence dans la voix (Cummins *et al.*, 2018). Élaboré pour la compétition Interspeech 2011 portant sur la détection de l'état instantané du locuteur dans la voix (Schuller *et al.*, 2011), ce corpus est constitué d'enregistrements de participants effectuant différentes tâches vocales, elles-mêmes conduites en parallèle d'autres études médicales induisant une privation de sommeil des sujets. Les sujets sont des volontaires germanophones et tous les échantillons sont soit en allemand, soit en anglais.

Les informations associées aux enregistrements vocaux sont la tâche effectuée lors de la lecture, le sexe du locuteur, l'affectation de l'échantillon dans la base d'entraînement, de développement ou de test, et la valeur de somnolence correspondante.

Donnée	Femmes	Hommes	Total
Nombre de sujets	63	43	106
Nombre d'échantillons	315	215	530
Âge moyen (écart-type)	33,9 (11,5)	38,7 (16,9)	35,9 (14,1)
Niveau social moyen (écart-type)	6,0 (2,5)	4,6 (2,3)	5,4 (2,5)
KSS moyen (écart-type)	4,6 (1,3)	4,3 (1,2)	4,4 (1,3)
Nombre d'échantillons S (KSS)	72	36	108
Durée totale S (KSS)	1 h 36 m 4 s	49 m 57 s	2 h 26 m
Nombre d'échantillons NS (KSS)	243	179	422
Durée totale NS (KSS)	4 h 58 m 22 s	4 h 31 s	8 h 58 m 53 s
TILE moyenne (écart-type) en minutes	11,8 (4,6)	10,4 (5,1)	11,2 (4,8)
Nombre de sujets S (TILE)	13	15	28
Durée totale d'enregistrement S (TILE)	1 h 20 m 55 s	1 h 39 m 37 s	3 h 32 s
Nombre de sujets NS (TILE)	50	28	78
Durée totale d'enregistrement NS (TILE)	5 h 13 m 30 s	3 h 10 m 52 s	8 h 24 m 22 s

Tableau 1. *Statistiques du corpus TILE. S : somnolent ; NS : non-somnolent*

La valeur de somnolence est la moyenne de trois KSS, un rempli par le patient lui-même, et deux remplis par des annotateurs externes. Nous redirigeons le lecteur vers (Krajewski *et al.*, 2009 ; Golz *et al.*, 2007) pour un descriptif détaillé des conditions expérimentales d'enregistrement et vers (Schuller *et al.*, 2013) pour la liste exhaustive des différents sous-corpus agrégés pour former le SLC.

Afin d'assurer une comparaison valide entre le corpus TILE et le SLC, nous sélectionnons uniquement les tâches de lecture de ce dernier. De plus, après l'étude menée dans la section 4, nous sélectionnons seulement les tâches de lecture dont la taille moyenne des échantillons est supérieure à 8 secondes : la lecture de la fable *Nordwind und Sonne* (version en allemand de la fable *La bise et le soleil*) dont la durée moyenne est de 36,5 secondes ; la lecture de deux simulations de communication de trafic aérien (« flight1 » et « flight2 » de durées moyennes respectives de 9,7 secondes et 13,8 secondes) et la lecture d'une simulation de discours d'un contrôleur de trafic aérien « roger1 » (durée moyenne : 8,5 secondes). Les statistiques de ce sous-corpus sont présentées dans le tableau 2.

3. Marqueurs vocaux

La grande majorité des systèmes de l'état de l'art ayant pour but de détecter la somnolence subjective dans la voix utilisent des marqueurs vocaux calculés avec la boîte à outils openSMILE (Eyben et Schuller, 2015). Les 4 368 marqueurs correspondant à la compétition Interspeech 2011 sur l'état de somnolence du locuteur ne sont malheureusement pas tous interprétables par des non-spécialistes de la voix. Or, l'élaboration d'un outil de détection de la somnolence dans la voix nécessite une collaboration étroite avec des médecins, qui ont besoin de pouvoir relier les marqueurs vocaux à des mécanismes neuromoteurs ou de performances cognitives. Nous avons

Sexe	Classe	**Ent.**	**Dev.**	**Test**	**Total**
Femmes	NS	10	8	9	27
		109 éch.	88 éch.	73 éch.	270 éch.
		4,15 (1,4)	4,0 (1,6)	4,18 (1,4)	4,11 (1,5)
		24 min 23 s	18 m 43 s	17 m 48 s	1 h 54 s
	S	5	6	6	17
		106 éch.	76 éch.	86 éch.	268 éch.
		8,12 (0,5)	8,13 (0,7)	8,21 (0,9)	8,15 (0,7)
		18 m 23 s	15 m 18 s	17 m 59 s	51 m 40 s
Hommes	NS	10	7	9	26
		54 éch.	27 éch.	52 éch.	133 éch.
		4,8 (1,2)	3,9 (1,6)	3,5 (1,8)	4,1 (1,6)
		17min 4s	8 m 17s	15 m 31 s	40 m 53s
	S	4	7	3	14
		33 éch.	56 éch.	30 éch.	119 éch.
		8,7 (0,9)	8,7 (1,0)	8,14(0,9)	8,6 (1,0)
		7 m 8 s	14 m 4s	6 m 41 s	27 m 53 s
Total	NS	20	15	18	53
		164 éch.	115 éch.	125 éch.	404 éch.
		4,3 (1,4)	4,0 (1,6)	3,9 (1,6)	4,1 (1,5)
		41 m 38 s	26 m 59 s	33 m 19 s	1 h 41 m 57s
	S	9	13	9	31
		139 éch.	132 éch.	116 éch.	387 éch.
		8,3 (0,7)	8,4 (0,9)	8,2 (0,9)	8,3 (0,8)
		25 m 31 s	29 m 21 s	24 m 40 s	1 h 19 m 33s

Tableau 2. *Nombre de locuteurs, nombre d'échantillons, KSS moyen (écart-type) et durée cumulée d'enregistrements du sous-corpus de la base SLC contenant uniquement des tâches de lecture. S : somnolent ; NS : non-somnolent ; Ent. : entraînement ; Dev. : développement*

donc élaboré notre propre ensemble de marqueurs vocaux, contenant exclusivement des marqueurs dont l'explicabilité a été mise à l'épreuve avec des médecins et qui peuvent être reliés à des mécanismes physiologiques.

3.1. *Statistiques concernant les parties voisées*

Les marqueurs vocaux sont calculés en deux temps. Tout d'abord, nous extrayons les segments voisés grâce à l'extraction de la fréquence fondamentale par l'algorithme ESPS (Sjölander, 2004), ainsi que l'extraction automatique de segments vocaliques (Pellegrino et Andre-Obrecht, 2000). Le premier sous-groupe de marqueurs est composé de statistiques sur ces segments, tandis que le second sous-groupe contient des marqueurs caractérisant la régularité de la production d'harmoniques sur les seg-

ments voisés. L'ensemble de ces marqueurs est ensuite moyenné pour obtenir un seul groupe de marqueurs audio par échantillon.

Les statistiques obtenues sur les parties voisées et les parties vocaliques reflètent le comportement global du locuteur et sont les suivantes :

– la durée totale des parties voisées (en secondes) ;

– le pourcentage en durée des parties voisées ;

– la durée totale des segments vocaliques (en secondes) ;

– le pourcentage en durée des segments vocaliques.

3.2. *Régularité de la production d'harmoniques sur les segments voisés*

Une fois les parties voisées et les parties vocaliques extraites, nous mesurons la régularité de la production d'harmoniques sur ces segments grâce à des mesures de fréquence fondamentale et de courbes d'intensité :

– F_0MEAN : la moyenne de la fréquence fondamentale sur les segments voisés ;

– F_0VAR : la variance de la fréquence fondamentale sur les segments voisés ;

– F_0SLOPE : le coefficient directeur de l'approximation linéaire de la fréquence fondamentale sur un segment voisé ;

– F_0MAX : le maximum de la fréquence fondamentale sur un segment voisé ;

– F_0MIN : le minimum de la fréquence fondamentale sur un segment voisé ;

– F_0EXTEND : l'amplitude de la fréquence fondamentale sur un segment voisé.

Les mêmes paramètres sont calculés sur les courbes d'intensité (NRJMEAN, NRJVAR, NRJMAX, NRJMIN, NRJEXTEND). Il en résulte 12 paramètres vocaux supplémentaires (6 sur la fréquence fondamentale F_0, 6 sur l'intensité). Nous avons également calculé les équivalents de F_0MEAN, F_0VAR, NRJMEAN et NRJVAR sur les segments vocaliques, ajoutant ainsi 4 paramètres vocaux.

Cet ensemble de paramètres est complété par des paramètres qui ont notamment été utilisés pour caractériser la classification d'attitudes sociales (Rouas *et al.*, 2019) et que nous avons calculés avec la boîte à outils Matlab Covarep (Degottex *et al.*, 2014) que nous avons modifiée pour les calculer seulement sur les segments voisés. Nous complétons ainsi notre ensemble de paramètres avec l'amplitude des harmoniques (H1, H2, H4), l'amplitude des formants (A1, A2, A3), leur fréquence (F1, F2, F3, F4) et leur bande passante (B1, B2, B3, B4), la différence entre les amplitudes des harmoniques (H1-H2, H2-H4), la différence d'amplitude entre les harmoniques et les formants (H1-A1, H1-A2, H1-A3), la *Cepstral Peak Prominence* (CPP) et les rapports harmoniques sur bruit dans différentes plages de fréquences (HNR05, HNR15, HNR25, HNR35). Tous ces paramètres sont moyennés sur chaque enregistrement, ce qui ajoute un total de 24 paramètres à notre ensemble de paramètres vocaux.

Cet ensemble de marqueurs contient ainsi un total de 44 paramètres vocaux.

4. Longueur minimale des échantillons pour la détection de la somnolence grâce à des marqueurs vocaux

Lors de la sélection d'un sous-corpus du SLC, une question jamais soulevée à notre connaissance dans l'état de l'art s'est imposée : quelle est la longueur d'enregistrement audio nécessaire pour permettre la détection de la somnolence dans la voix ?

Pour répondre à cette question, nous avons découpé tous les échantillons audio des deux corpus en tronçons contenant uniquement la première seconde de l'échantillon, uniquement les deux premières secondes de l'échantillon, uniquement les trois premières secondes de l'échantillon, etc. On obtient ainsi des échantillons de taille croissante, sur lesquels on calcule les marqueurs vocaux présentés dans la section 3. Pour éviter un biais qui serait propre à nos marqueurs, nous extrayons également les marqueurs de la conférence Interspeech 2011 grâce à la boîte à outils openSMILE (Eyben et Schuller, 2015) pour comparaison.

Ensuite, nous calculons pour chaque échantillon la similarité cosinus entre le marqueur correspondant au tronçon de taille i secondes et celui correspondant au tronçon de taille $i+1$ secondes, issus du même fichier audio :

$$s_{i,i+1} = \frac{|X_i| \cdot |X_{i+1}|}{||X_i|| \cdot ||X_{i+1}||}$$

Ainsi, quand $s_{i,i+1}$ est proche de 1, X_i est proche de X_{i+1} : l'information supplémentaire apportée par la seconde supplémentaire entre les échantillons i et $i+1$ est faible. Nous calculons la moyenne et l'écart-type des $s_{i,i+1}$ pour tous les échantillons, et nous obtenons le graphe présenté dans la figure 1. Il représente l'information supplémentaire apportée par chaque seconde supplémentaire dans l'échantillon, à partir d'un échantillon vide.

Une première remarque concerne la différence de valeurs entre l'évolution des marqueurs personnalisés et ceux extraits avec openSMILE. En effet, les valeurs de moyenne et d'écart-type de $s_{i,i+1}$ sont très proches respectivement de 1 et de 0 pour les marqueurs extraits avec openSMILE, et ce, quel que soit i. Nous faisons l'hypothèse que cela provient de la différence de taille des ensembles de marqueurs. En effet, dans le cas des marqueurs IS11, une différence franche sur un nombre réduit de marqueurs aura peu d'impact sur la similarité cosinus calculée sur les 4 368 marqueurs, contrairement aux marqueurs personnalisés, au nombre de 44. Cependant, cette différence ne change pas l'interprétation faite de l'évolution de $s_{i,i+1}$. En effet, quel que soit le set de marqueurs ou le corpus, pour une durée d'environ 8 secondes, la moyenne et l'écart-type de $s_{i,i+1}$ commencent à devenir stationnaires : toute information audio supplémentaire n'apporte plus d'information vis-à-vis des marqueurs audio calculés sur une durée plus courte. Nous prenons ainsi cette limite comme limite minimale requise pour la détection de la somnolence dans la voix grâce aux marqueurs vocaux.

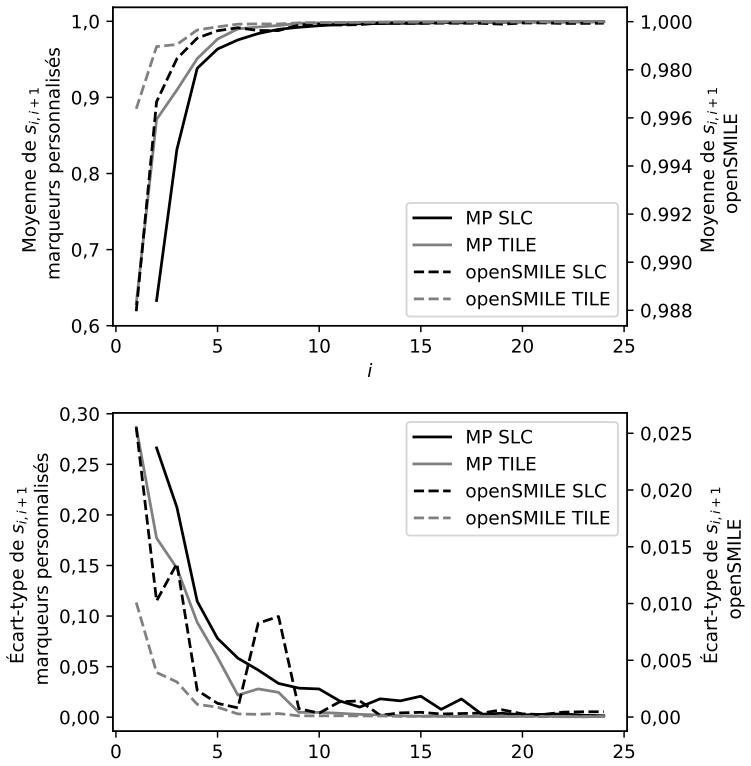


Figure 1. *Moyenne et écart-type de $s_{i,i+1}$ en fonction de i. MP : marqueurs personnalisés*

5. Détection de la somnolence subjective grâce à des marqueurs vocaux

Nous proposons dans cette partie d'estimer la somnolence subjective à court terme grâce à des marqueurs vocaux. La détection de la somnolence à court terme est utile pour anticiper les baisses de performance à court terme, qui peuvent avoir des conséquences dramatiques, par exemple lors de la conduite d'une voiture, le pilotage d'un avion ou à un poste dans lequel l'attention est critique (aiguilleur du ciel ou responsable d'une centrale nucléaire, ar exemple). À la fois dans le SLC et dans le corpus TILE, cette somnolence à court terme est mesurée grâce au questionnaire subjectif KSS. Il existe cependant une différence notable entre l'annotation des deux corpus : si dans le corpus TILE le KSS est rempli uniquement par le sujet avant l'enregistrement audio, celui du SLC est la moyenne entre un questionnaire rempli par le sujet lui-même et de deux annotateurs externes (assistants médicaux) entraînés auparavant à évaluer la somnolence.

Ce problème a été introduit lors de la compétition proposée au sein de la conférence Interspeech 2011 sur la classification d'état du locuteur, dont le meilleur système achevait une performance de 76,4 % (Huang *et al.*, 2014).

5.1. *Méthodologie*

La méthodologie employée pour calculer les performances sur les deux corpus est représentée dans la figure 2 et se décompose de la manière suivante :

– centrage des paramètres vocaux par locuteur. En soustrayant la moyenne des marqueurs vocaux d'un locuteur à tous les marqueurs de ce sujet, on élimine les facteurs propres au locuteur (sexe, âge, physionomie des voies respiratoires...) et on garde uniquement les variations instantanées des paramètres vocaux, qui ne sont plus pollués par des marqueurs traits s'exprimant dans la voix. Cette méthodologie semble d'autant plus pertinente du fait que l'on cherche à estimer la somnolence subjective à court terme et non un état général de somnolence sur le long terme du locuteur ;

– calcul pour chaque marqueur vocal de la corrélation (ρ de Spearman) entre le marqueur et la mesure de somnolence (KSS). Cela permet d'ordonner les marqueurs vocaux du plus corrélé au moins corrélé avec la mesure de somnolence. Par ailleurs, travailler avec des méthodes statistiques permet, contrairement à des techniques de réduction « classiques », comme l'analyse en composantes principales ou l'analyse en composantes indépendantes, de conserver le sens associé aux marqueurs vocaux et ainsi, postérieurement, de lier somnolence et manifestions physiologiques par l'intermédiaire de ces marqueurs. Ce calcul se fait sur l'ensemble entraînement et développement ;

– sélection du nombre de marqueurs et des paramètres optimaux du classificateur. Pour cela, nous calculons les performances du système (sur la base d'entraînement *vs* la base de développement) pour les 1, 2, ..., 44 marqueurs vocaux précédemment triés, et nous conservons le nombre de marqueurs vocaux et les paramètres du classificateur fournissant les meilleures performances. Le classificateur utilisé est un séparateur à vastes marges (SVM), dont les paramètres sont le type de noyau (linéaire ou gaussien), et les paramètres C et γ. Durant cette phase, les performances sur le corpus TILE sont mesurées avec le score F1 (moyenne géométrique de la précision et du rappel) en raison de la validation croisée qui laisse trop peu d'échantillons dans la base de développement pour que le score de rappel non biaisé (SRN), utilisé pour calculer les performances dans la suite, soit pertinent ;

– les paramètres C et γ obtenus lors de l'étape 3 sont utilisés pour entraîner le SVM sur le sous-corpus entraînement et développement et nous obtenons ainsi les classes de somnolence estimées de chaque échantillon du sous-corpus de test.

Le SLC est déjà divisé en sous-corpus d'entraînement, de développement et de test, mais ce n'est pas le cas pour la base TILE. Nous utilisons donc une validation croisée qui exclut à chaque itération un locuteur qui servira de test, puis les locuteurs restants sont divisés en bases d'entraînement et de développement (respectivement

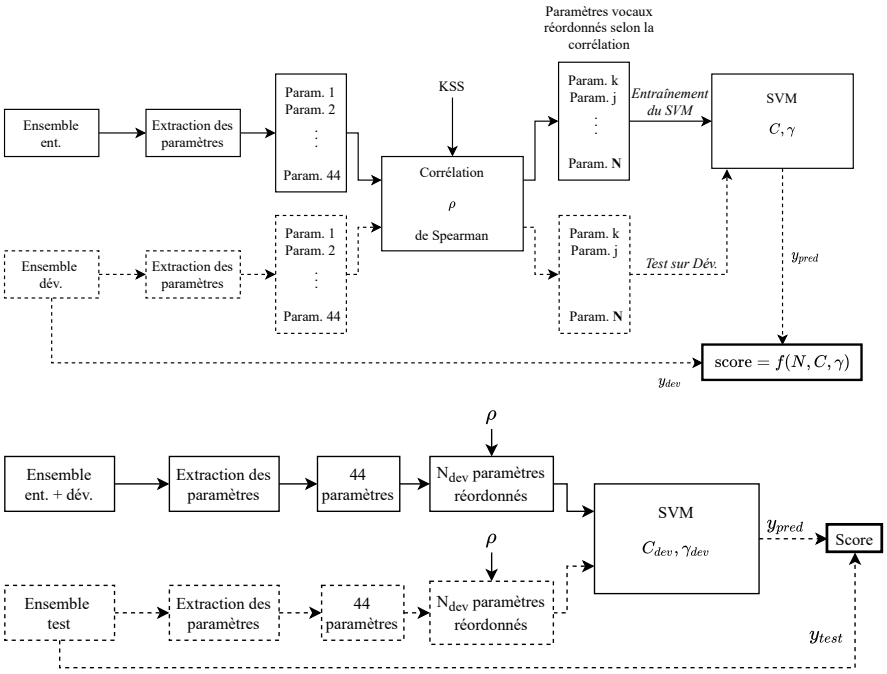


Figure 2. *Schéma du système proposé. Ent. : entraînement ; Dev. : développement*

quatre cinquièmes et un cinquième des locuteurs restants), ayant les mêmes distributions en termes de sexe, d'âge et de somnolence.

À chaque itération de la validation croisée, les classes estimées des échantillons du locuteur précédemment exclus pour le test sont ajoutées dans une matrice de confusion globale, sur laquelle le score global est calculé. Par ailleurs, en raison du faible nombre d'échantillons et de leur déséquilibre entre les deux classes S et NS, nous appliquons un suréchantillonnage de la classe minoritaire grâce au *Synthetic Minority Over-sampling Technique*, SMOTE (Chawla *et al.*, 2002) implémenté dans la boîte à outils Python Sklearn (Pedregosa *et al.*, 2011). Les résultats sont présentés dans le tableau 3.

5.2. *Résultats et discussion*

Une première analyse de ces résultats montre que les performances calculées sur la base TILE sont largement inférieures à celles calculées sur le SLC, avec ou sans centrage des marqueurs par locuteur. Par ailleurs, il est intéressant de remarquer que les performances sur la base TILE sont largement inférieures lorsque les marqueurs sont centrés que lorsqu'ils ne le sont pas (avec une différence de presque 8 %). Le

Réf.	Corpus	Système	SRN
(Huang *et al.*, 2014)	SLC	-	76,4 %
(1a)	SLC	avec centrage	**77,6 %**
(1b)	SLC	sans centrage	66,8 %
(1c)	SLC (locuteurs filtrés)	avec centrage	72,4 %
(1d)	base TILE	avec centrage	48,7 %
(1e)	base TILE	sans centrage	55,6 %
(1f)	base TILE	limite KSS = 5	56,3 %

Tableau 3. *Résultats des systèmes de classification pour la classification de somnolence subjective instantanée*

phénomène inverse est observé sur la base SLC : le système avec centrage par locuteur donne des performances supérieures de plus de 10 % au système sans centrage.

Nous formulons trois hypothèses pour expliquer ces résultats. Premièrement, chaque locuteur du SLC produit en moyenne 13 enregistrements correspondant aux tâches de lecture précédemment sélectionnées (nombre moyen d'enregistrements de tâches de lecture par patient : 13,4 ; écart-type : 19,4), ce qui est plus que ceux du corpus TILE qui sont limités à 5 enregistrements par les conditions expérimentales. Il est intéressant de noter que 5 patients sur les 94 du corpus ont produit à eux seuls 359 des 791 échantillons. Deux de ces locuteurs sont dans la base d'entraînement (n° 38 et n° 39), comptant respectivement 56 et 95 échantillons. Deux autres comptant 36 et 75 échantillons sont dans la base de développement (les n° 40 et n° 41) tandis que le dernier locuteur (n° 42) est dans la base de test et compte 96 échantillons. Nous pensons que le très grand nombre d'échantillons par locuteur dans la base SLC permet une meilleure estimation des traits vocaux lors du centrage des marqueurs par locuteur, ce qui induit un meilleur centrage et donc de meilleures performances. Pour vérifier cette hypothèse, nous avons réappliqué la procédure décrite à la section 5.1, mais sans les locuteurs précédemment ciblés. Cela conduit à un score de 72,4 % (1c), ce qui est à peine plus que 5 % de moins que le système (1a). De plus, ce score reste très supérieur à celui obtenu sur la base TILE avec la même méthodologie : si elle introduit un biais, la présence de locuteurs produisant de très nombreux échantillons dans la base SLC n'explique pas toutes les différences entre les deux systèmes.

Ces observations conduisent à une deuxième hypothèse qui concerne le ratio entre somnolents et non somnolents dans la base TILE, qui est très faible (à peine un cinquième des échantillons correspondent à un KSS supérieur à 7). En effet, un fort déséquilibre entre les classes, malgré l'augmentation de données par suréchantillonnage de la classe minoritaire, empêche une généralisation correcte des classificateurs. En abaissant la limite pour séparer les deux classes à 5 (« ni éveillé, ni somnolent »), on obtient une répartition plus équilibrée de 251 échantillons S contre 279 NS. En réappliquant la même méthodologie que précédemment avec ce nouvel étiquetage, les performances du système augmentent de manière anecdotique (1f) : à peine 0,7 de plus

que le score avec une limite pour le KSS de 7 (1d). Le déséquilibre entre les classes du KSS ne semble donc pas être la source majoritaire des erreurs du classificateur.

Nous formulons donc une troisième hypothèse qui concerne la validité de la mesure de somnolence dans la base TILE. En effet, si le score au KSS est corrélé à l'activité électroencéphalographique des sujets sains (Kaida *et al.*, 2006) comme c'est le cas dans le SLC, les patients souffrant de maladie du sommeil ont une mauvaise perception de leur somnolence subjective (Sangal, 1999). Une observation semblable avait été faite dans l'article présentant le corpus (Martin *et al.*, 2020). Ainsi, le KSS relevé dans la base TILE ne mesure pas les mêmes phénomènes sur les patients de la base TILE que sur les sujets sains du corpus SLC. Les bonnes performances des marqueurs sur le SLC tendent donc à confirmer que ceux-ci sont pertinents pour la détection de la somnolence subjective, mais aussi que ces marqueurs ne sont pas adaptés pour la détection du phénomène mesuré par le KSS sur les patients de la base TILE.

6. Estimation de la somnolence objective sur le long terme grâce à des marqueurs vocaux

Le suivi médical des patients souffrant de SDE peut tirer bénéfice de la détection de la somnolence à court terme, mais aussi à long terme, pour permettre aux médecins de suivre sur de longues plages de temps les variations des marqueurs de traits de somnolence des locuteurs. La détection d'une telle somnolence s'appuie sur le fait que dans le corpus TILE, chaque locuteur est enregistré cinq fois, à des moments différents de la journée. Cette partie a donc pour objectif de classer non plus les échantillons indépendamment les uns des autres mais les locuteurs entre somnolents (TILE moyenne inférieure ou égale à 8 minutes) et non-somnolents (TILE moyenne supérieure à 8 minutes) grâce aux enregistrements de leurs cinq siestes. La limite de 8 minutes sur la moyenne des latences d'endormissement est une limite médicale utilisée dans le diagnostic de nombreuses maladies telles que la narcolepsie par exemple (Aldrich *et al.*, 1997).

6.1. *Méthodologie et résultats*

La première intuition pour estimer la classe de somnolence des locuteurs déterminée par leur valeur moyenne des latences d'endormissement au TILE est de faire la moyenne des cinq jeux de marqueurs de chaque locuteur et d'effectuer la classification directement grâce à un unique ensemble de marqueurs moyens par locuteur. En utilisant la même validation croisée isolant un locuteur pour le test et la même procédure de sélection des marqueurs grâce à la corrélation de Spearman, on obtient un score final d'à peine 50 % (2a). Ce paradigme divisant le nombre d'échantillons par 5, la réduction drastique du nombre d'échantillons pourrait être la cause de ce faible résultat.

Une autre méthode s'appuie sur le fait que l'on a cinq enregistrements pour chaque locuteur et reprend la méthodologie précédemment détaillée dans la section 5.1. Pour chaque itération de la validation croisée, une fois le classificateur entraîné sur les itérations prises de manière indépendante, nous calculons les probabilités d'appartenance à chaque classe de somnolence des cinq enregistrements du locuteur de test. Nous moyennons ensuite ces cinq probabilités pour estimer la classe de somnolence du locuteur de test, que nous rajoutons dans une matrice de confusion globale.

Réf.	Sélection des marqueurs	limite TILE	SRN
(2a)	Moyenne des paramètres (Spearman)	8	50,2 %
(2b)	Moyenne des paramètres (Mann-Whitney)	8	54,8 %
(2c)	Spearman	8	45,6 %
(2d)	Mann-Whitney	8	53,6 %
(2e)	Mann-Whitney	13	**63,8 %**

Tableau 4. *Résultats des systèmes de classification pour la détection de la somnolence à long terme objective sur la base TILE*

L'application de cette méthodologie conduit à un SRN de 45,6 % (2c), ce qui est en dessous des performances qui seraient obtenues en tirant la classe de somnolence au hasard. Pour tenter d'améliorer ces résultats, nous conservons la méthodologie précédente et nous testons une autre méthode de sélection des marqueurs, basée sur le test statistique de Mann-Whitney au lieu de la corrélation de Spearman : au lieu de classer les marqueurs par leur corrélation à la mesure de somnolence, nous les classons par leur pouvoir discriminant entre les deux classes. En effet, plus le U du test de Mann-Whitney est faible, plus les distributions S et NS du marqueur étudié sont différentes (2d). Cette nouvelle approche permet un score de classification atteignant 53,6 %, ce qui représente une augmentation de plus de 8 % du score de classification.

De même que dans la partie 5, nous retestons notre système avec une autre limite pour séparer les deux classes de somnolence selon la valeur de TILE moyenne des patients. Afin d'avoir un meilleur équilibre entre les classes, nous proposons la limite de 13 minutes. En réappliquant les systèmes précédents avec cette nouvelle limite, nous obtenons un score de presque 64 % (2e), ce qui représente une amélioration de presque 10 % par rapport au système précédent.

Ce score reste malgré tout trop faible pour une utilisation en situation réelle, qui nécessiterait 80 % ou 85 % de performances pour une utilisation clinique. Nous faisons deux hypothèses pour expliquer ces résultats. D'une part, de même que le stress ou les émotions peuvent influencer l'expression de la somnolence immédiate dans la voix, l'anxiété, la dépression, et une multitude d'autres facteurs propres au locuteur peuvent également polluer les marqueurs vocaux utilisés pour la détection de la somnolence. Le corpus étant composé d'enregistrements de patients souffrant de SDE, la plupart ont des facteurs de comorbidité qui pourraient influencer leur voix.

D'autre part, du point de vue de la détection de la somnolence du locuteur, la base de données se réduit à 106 patients, ce qui est relativement faible, à la fois pour

l'entraînement et le calcul des performances. Par ailleurs, le fait de moyenner les probabilités des cinq échantillons de manière égale masque l'éventuelle importance que pourraient avoir certaines siestes par rapport à d'autres. Une étude plus approfondie de cette question semble nécessaire pour permettre l'estimation de la somnolence du locuteur grâce aux marqueurs vocaux de manière plus fine, ce qui pourrait mener à une meilleure compréhension des phénomènes mis en jeu et de meilleures performances.

7. Estimation de marqueurs de traits grâce aux erreurs de lecture

Une nouvelle approche pour la détection de la somnolence concerne l'utilisation des erreurs effectuées lors de la lecture des textes de la base TILE. En effet, si les marqueurs vocaux peuvent être liés à des processus neuromusculaires (Krajewski et Kroger, 2007), nous faisons l'hypothèse que les erreurs de lecture sont des marqueurs pertinents de l'influence de la somnolence sur les performances cognitives nécessaires à la lecture. Cette partie traitant de la détection de marqueurs de traits des locuteurs sur le corpus TILE, les patients seront dits « somnolents » si la valeur moyenne de leur latence d'endormissement au TILE est inférieure ou égale à 8 minutes.

7.1. *Liste des erreurs de lecture*

Afin de différencier différents comportements de lecture nous avons retenu quatre catégories d'erreurs, que nous avons voulues relativement générales afin d'obtenir un nombre suffisant d'observations dans chaque catégorie. Les erreurs prises en considération sont les suivantes :

– les achoppements (Ach) : « hésitation, coupure, dans le rythme de la parole » (Brin *et al.*, 2018). Ces erreurs sont un reflet de la capacité d'assemblage du lecteur, c'est-à-dire sa capacité de mettre bout à bout des syllabes pour former un mot. Ainsi, lorsque le lecteur commence la lecture d'un mot, s'arrête, et se reprend, le processus d'assemblage a été interrompu, causant un achoppement. Nous n'avons pas pris en compte les arrêts entre les mots mais seulement les arrêts qui se produisent au milieu d'un mot, ou les allongements artificiels de certaines voyelles, qui témoignent d'une hésitation. Dans le cas de la reprise d'une phrase ou d'un bout de phrase, un seul achoppement est compté, quelle que soit la longueur de la reprise ;

– les paralexies (Plx) : « erreur d'identification de mots écrits consistant à oraliser un mot écrit à la place d'un autre » (Brin *et al.*, 2018). Contrairement aux achoppements, les paralexies reflètent les erreurs d'adressage du lecteur. La capacité d'adressage est le fait de lire un mot dans sa globalité, sans le découper en syllabes ou le déchiffrer, dont les paralexies sont des erreurs symptomatiques. Nous avons généralisé cette catégorie à toute prononciation d'un mot, existant ou non, qui est lu à la place du mot correct. Les télescopages (oublis d'une ou plusieurs syllabes dans un mot) sont donc inclus dans cette catégorie ;

– les oublis de mots (O) : cette erreur est comptée lorsque le lecteur oublie de lire un mot et passe directement au début du mot suivant ;

– les additions de mots (Add) : cette erreur est comptée lorsque le lecteur ajoute un mot qui n'était pas dans le texte original.

Si un locuteur se reprend après une paralexie, un oubli ou une addition, aucun achoppement supplémentaire n'est compté, sauf s'il se trompe lors de la reprise.

7.2. *Sensibilité des erreurs de lecture à la somnolence*

Afin de mesurer si les erreurs élaborées précédemment varient avec la somnolence, les distributions du nombre total de chaque type d'erreur par locuteur chez les patients somnolents et non somnolents sont représentées dans la figure 3 (moyenne $\pm$ SEM – erreur standard de la moyenne). Sur tous les types d'erreurs, les patients somnolents font plus d'erreurs que leurs homologues non somnolents (tests de Mann-Whitney. Ach : $U = 873, p = 8,1 \times 10^{-2}$; O : $U = 738, p = 7,8 \times 10^{-3}$; Add : $U = 847, p = 5,0 \times 10^{-2}$; Plx : $U = 759, p = 1,2 \times 10^{-2}$; total : $U = 765, p = 1,4 \times 10^{-2}$).

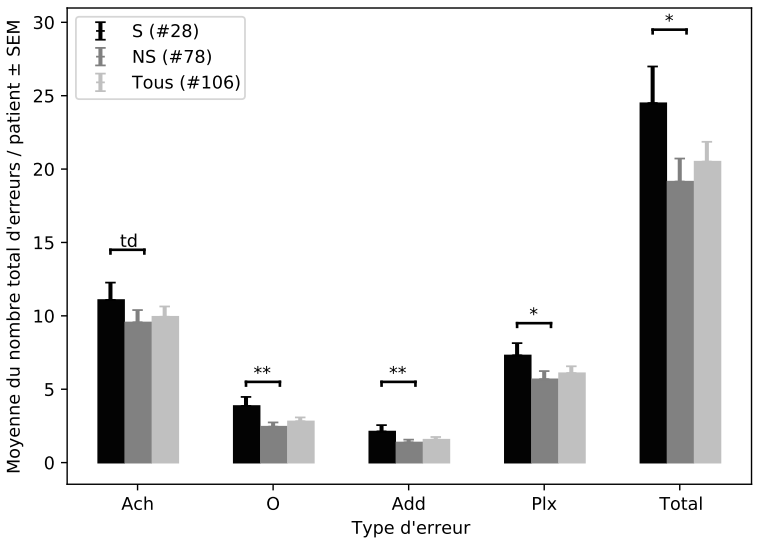


Figure 3. *Distribution du nombre total d'erreurs par locuteur (moyenne $\pm$ SEM). Ach : achoppements, O : oublis, Add : additions, Plx : paralexies. Tests de Mann-Whitney (td : $p < 10^{-1}$, * : $p < 5 \times 10^{-2}$, ** : $p < 10^{-2}$)*

7.3. *Étude des sources d'influence de production d'erreurs*

Il est nécessaire de pouvoir séparer l'influence de la somnolence des facteurs extérieurs pouvant provoquer ces erreurs. Ces facteurs peuvent être les différences entre les textes (différence de taille, quantité de dialogues, difficulté) ou les différents facteurs temporels tels que la prise de repas ou la fatigue accumulée de la journée. Dans la suite, « influence de l'itération » désignera l'influence de tels facteurs sur les erreurs produites par le locuteur. Afin de séparer la contribution de la somnolence de celle de l'itération, nous avons appliqué à nos données une ANOVA multivariée à mesures répétées avec R (R Core Team, 2017).

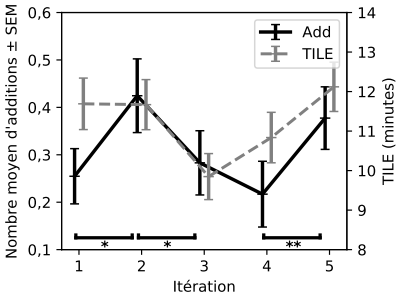
Les additions, oublis et paralexies sont représentés avec les valeurs de TILE et le KSS en fonction des itérations du TILE dans la figure 4. Les achoppements avec les valeurs de TILE et le KSS sont représentés dans la figure 5.

7.3.1. *Additions de mots*

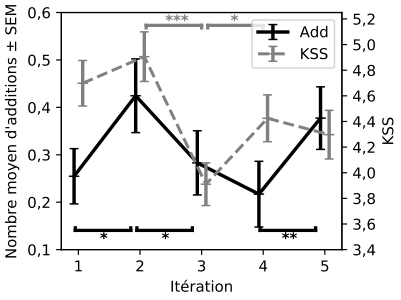
Une ANOVA prenant en compte le nombre d'additions, les différentes échelles de somnolence, et l'influence de l'itération montre que la somnolence objective a une influence quasiment significative sur les variations inter-sujets du nombre d'additions (influence de la valeur de TILE sur les variations inter-sujets : $F = 3,5 ; p = 6,6 \times 10^{-2}$) et que la somnolence subjective a un effet quasiment significatif sur les variations inter-sujets du nombre d'additions (influence du KSS sur les variations inter-sujets : $F = 3,8 ; p = 5,2 \times 10^{-2}$). Cela signifie que les différences observées entre les sujets indépendamment du temps sont principalement expliquées par leurs différences de TILE (ce qui confirme le lien entre TILE et additions) tandis que celles observées sur chaque sujet au cours du temps (influences conjointes de la session et du locuteur) sont principalement expliquées par les différences de variation de KSS au cours des itérations du test. La session n'a aucun effet significatif sur la production des additions. Nous faisons donc l'hypothèse que les variations du nombre d'additions sont principalement dues à celles des somnolences objectives et subjectives, et qu'elles sont donc indépendantes du texte et des autres effets d'itération.

7.3.2. *Oublis de mots*

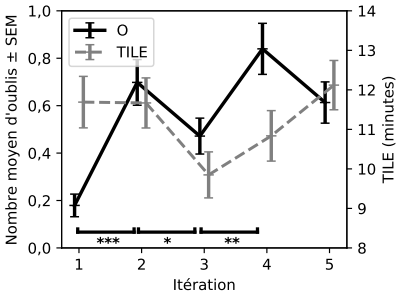
De même, la somnolence objective a une influence sur les variations de nombre d'oublis de mots (influence de la valeur de TILE sur les variations inter-sujets : $F = 3,2 ; p = 7,5 \times 10^{-2}$) tandis que la somnolence subjective a une influence sur les variations du nombre d'oublis de mots (influence du KSS sur les variations intra-sujets : $F = 3,1 ; p = 8,1 \times 10^{-2}$). En revanche, contrairement aux additions, ces erreurs subissent également les effets de l'itération (effet de l'itération sur les variations intra-sujets : $F = 12,0 ; p = 3,0 \times 10^{-9}$).



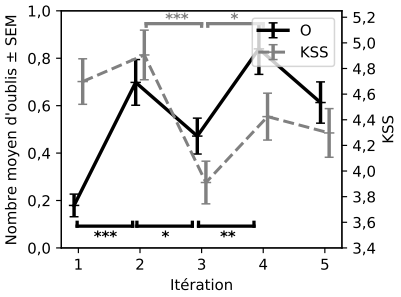
(a) *Additions et TILE en fonction des itérations*



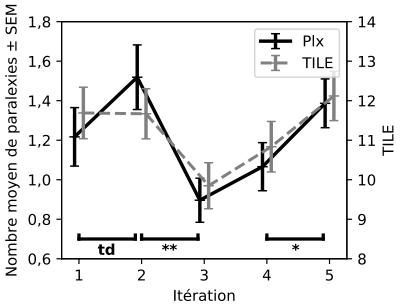
(b) *Additions et KSS en fonction des itérations*



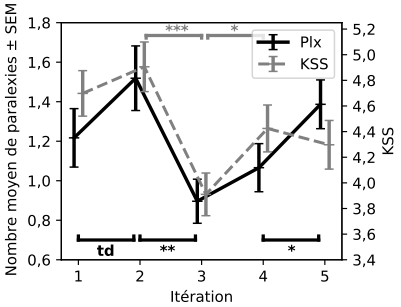
(c) *Oublis et TILE en fonction des itérations*



(d) *Oublis et KSS en fonction des itérations*



(e) *Paralexies et TILE en fonction des itérations*

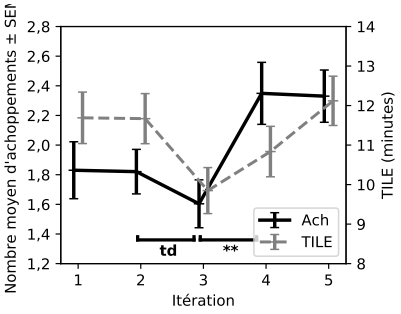


(f) *Paralexies et KSS en fonction des itérations*

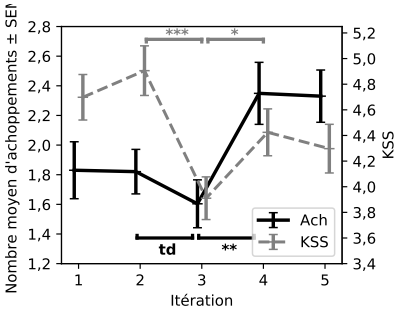
Figure 4. *Additions (a, b), oublis (c, d) et paralexies (e, f) comparées au TILE et au KSS (moyenne ± SEM). Tests de Mann-Whitney*
(td : $p < 1 \times 10^{-1}$*, * :* $p < 5 \times 10^{-2}$*,** :* $p < 10^{-2}$*, *** :* $p < 10^{-3}$*)*

7.3.3. *Achoppements*

De même que pour les oublis de mots, l'étude des divers effets ayant une influence sur les variations du nombre d'achoppements permet de mettre en évidence



(a) *Achoppements et TILE en fonction des itérations*



(b) *Achoppements et KSS en fonction des itérations*

Figure 5. *Achoppements comparés au TILE et au KSS (moyenne ± SEM). Tests de Mann-Whitney*
*(td : $p < 1 \times 10^{-1}$, * : $p < 5 \times 10^{-2}$, ** : $p < 10^{-2}$)*

une influence significative du KSS ($F = 4,2$; $p = 4,2 \times 10^{-2}$) et de l'itération ($F = 7,4$; $p = 9,4 \times 10^{-6}$) sur les variations intra-sujets de ce type d'erreurs. En revanche, la valeur de TILE ne semble avoir aucune influence sur ce type d'erreurs.

7.3.4. *Paralexies*

Les variations de paralexies au cours des itérations ne semblent être influencées que par les facteurs d'itération (effet de l'itération sur les variations intra-sujets : $F = 6,0$; $p = 1,1 \times 10^{-4}$).

7.4. *Estimation de la somnolence du locuteur grâce aux erreurs de lecture*

Nous utilisons les erreurs de lecture précédemment élaborées ayant comme facteur d'influence la somnolence comme marqueurs pour un système de classification d'état du locuteur. Pour cela, nous concaténons les erreurs de lecture des cinq siestes pour chaque locuteur, dont nous nous servons comme entrée à un classificateur (SVM) selon la même procédure que dans la section 5.1 pour estimer le niveau de somnolence objective. Les matrices de confusion correspondantes sont représentées dans le tableau 5 (gauche) dont le score de rappel non pondéré atteint 78,7 %. Par ailleurs, les paralexies n'étant pas liées à la somnolence mais uniquement à des effets d'itération, la même procédure sans considérer ce type d'erreurs conduit aux matrices de confusion présentées dans le tableau 5 (droite). Le SRN associé est de 82,6 %.

TILE (8)	S_{pred}	NS_{pred}
S_{th}	23	4
NS_{th}	22	57

SRN = 78,7 %

TILE (8)	S_{pred}	NS_{pred}
S_{th}	20	7
NS_{th}	7	72

SRN = 82,6 %

Tableau 5. *Matrices de confusion et scores de rappel non pondérés des classificateurs utilisant les erreurs de lecture comme marqueurs de la somnolence*

7.5. *Discussion*

Les erreurs de lecture semblent de bons marqueurs de l'état de somnolence objective des locuteurs. Cependant, en raison de leur définition ou du texte incitant ou non ces erreurs, elles n'ont pas toutes la même importance dans la détection de la somnolence. En effet, en moyennant les coefficients attribués par le SVM aux différentes erreurs selon les différentes itérations de la validation croisée, on obtient les quatre marqueurs suivant les plus importants dans la classification : les additions des premières et cinquièmes siestes (ayant des coefficients respectifs $c = 8,0 \times 10^{-2}$ et $c = -1,5 \times 10^{-1}$), les achoppements de la troisième sieste ($c = 9,7 \times 10^{-2}$) et les oublis de la quatrième sieste ($c = -1,3 \times 10^{-1}$).

Ces coefficients sont cohérents avec les résultats de la partie précédente. En effet, les additions, qui ont ici le plus de poids dans la prise de décision du niveau de somnolence, avaient été identifiées comme ne dépendant que de la somnolence objective concernant les variations inter-sujets et ne dépendant pas des effets d'itération. De même, le deuxième marqueur le plus important dans la prise de décision est les oublis, qui malgré les effets d'itération variaient avec la somnolence objective. Enfin, même si l'étude des paralexies n'avait pas mis en valeur d'influence de la valeur de TILE sur la production de ce type d'erreurs, leur contribution dans la prise de décision n'est pas négligeable.

La répartition des coefficients de manière inégale sur les différentes siestes du test pose la question de l'importance relative des itérations pour l'estimation du niveau global de somnolence des locuteurs. En effet, si les additions sont les marqueurs ayant le plus de poids sur la première et la dernière sieste, leur contribution pour la détection de l'état du locuteur lors de la troisième sieste est très faible ($c = 8,8 \times 10^{-3}$), alors que celle des achoppements est la plus importante. Une cause probable de ces disparités est l'inégalité de contenu des textes. En effet, de nombreuses erreurs du corpus se répètent et certains mots sont systématiquement la cible d'une erreur spécifique. Par exemple, « méditatif » est très souvent prononcé « médiatif », causant de nombreuses paralexies à la cinquième sieste, ou encore « Il me répéta alors » est souvent lu à la place de « Et il me répéta alors », causant de nombreux oublis à la troisième sieste. Cela souligne l'aspect capital du choix des textes lus pour l'utilisation des erreurs de lecture en tant que marqueurs de la somnolence.

Par ailleurs, la définition des erreurs a également une influence sur leur robustesse. Nous faisons effectivement l'hypothèse que notre définition des achoppements ne prenant en compte que les interruptions au sein des mots et non entre les mots induit un biais qui empêche le marqueur de refléter l'état de somnolence du locuteur. De même, la fusion des paralexies et des télescopages dans la même catégorie pourrait induire des biais qui réduisent leur intérêt comme marqueurs de la somnolence.

8. Conclusion et perspectives

Pour conclure, après avoir étudié la question de la longueur minimale des échantillons pour la détection de la somnolence, nous avons proposé trois systèmes pour répondre à trois problématiques différentes en relation avec la détection de la somnolence dans la voix. La détection de la somnolence subjective à court terme grâce à des marqueurs vocaux simples donne des résultats satisfaisants lorsqu'il s'agit de sujets sains (sous-corpus de lecture du SLC) mais ne donne pas de bonnes performances lorsque la même méthodologie est appliquée pour des patients souffrant de SDE (base TILE). Nous supposons que cela vient du manque de validité de l'auto-évaluation effectuée par les patients composant le corpus TILE. De même, la méthodologie pour estimer la somnolence à long terme avec des marqueurs vocaux semble souffrir des biais apportés par les comorbidités de la SDE, empêchant le système d'atteindre des performances satisfaisantes. L'estimation de la somnolence à long terme de ces mêmes patients grâce à la valeur de TILE est en revanche très efficace lorsque l'on utilise les erreurs de lecture comme marqueurs de la somnolence : ces marqueurs semblent en effet robustes aux effets parasites qui pourraient s'exprimer dans la voix.

Nos futurs travaux comprendront l'étude approfondie de l'importance relative des siestes pour la détection de la somnolence à long terme. Par ailleurs, les erreurs de lecture étant actuellement annotées manuellement, nous travaillons à l'élaboration d'une détection automatique de ces erreurs de lecture grâce à un système de transcription automatique de la parole utilisant les caractères comme unités de reconnaissance.

Remerciements

Cette étude a été réalisée dans le cadre du projet IS-OSA, financé par la région Nouvelle-Aquitaine, et du projet SOMVOICE, financé par le Labex BRAIN (université de Bordeaux). Nous remercions également le Pr Jarek Krajewski pour nous avoir donné accès au *Sleepy Language Corpus*.

9. Bibliographie

Åkerstedt T., Gillberg M., « Subjective and objective sleepiness in the active individual. », *Int J Neurosci*, vol. 52, p. 29-37, 1990.

Aldrich M. S., Chervin R. D., Malow B. A., « Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy », *Sleep*, vol. 20, n° 8, p. 620-629, 1997.

Brin F., Courrier C., Lederle E., Masy V., *Dictionnaire d'orthophonie - 4ème édition*, orthoedition edn, September, 2018.

Carskadon M. A., Harvey K., Dement W. C., « Sleep Loss in Young Adolescents », *Sleep*, vol. 4, n° 3, p. 299-312, September, 1981.

Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., « SMOTE : Synthetic Minority Over-sampling Technique », *Journal of Artificial Intelligence Research*, vol. 16, p. 321-357, June, 2002.

Cummins N., Baird A., Schuller B., « Speech analysis for health : Current state-of-the-art and the increasing impact of deep learning », *Health Informatics and Translational Data Analytics*, vol. 151, p. 1-54, 2018.

de Saint-Exupéry A., *Le Petit Prince*, gallimard edn, 1943.

Degottex G., Kane J., Drugman T., Raitio T., Scherer S., « COVAREP — A collaborative voice analysis repository for speech technologies », *IEEE - ICASSP*, p. 960-964, 2014.

Eyben F., Schuller B., « Opensmile », *ACM SIGMultimedia Records*, vol. 6, p. 4-13, 2015.

Golz M., Sommer D., Chen M., Mandic D., Trutschel U., « Feature Fusion for the Detection of Microsleep Events », *Journal of VLSI Signal Processing*, vol. 49, p. 329-342, 2007.

Huang D.-Y., Zhang Z., Ge S. S., « Speaker State Classification Based on Fusion of Asymmetric Simple Partial Least Squares (SIMPLS) and Support Vector Machines », *Comput. Speech Lang.*, vol. 28, n° 2, p. 392-419, 2014.

Kaida K., Takahashi M., Åkerstedt T., Nakata A., Otsuka Y., Haratani T., Fukasawa K., « Validation of the Karolinska sleepiness scale against performance and EEG variables », *Clinical Neurophysiology*, vol. 117, n° 7, p. 1574-1581, 2006.

Krajewski J., Batliner A., Golz M., « Acoustic sleepiness detection : Framework and validation of a speech-adapted pattern recognition approach », *Behavior Research Methods*, vol. 41, n° 3, p. 795-804, 2009.

Krajewski J., Kroger B., « Using prosodic and spectral characteristics for sleepiness detection », *Interspeech 2007*, p. 1841-1845, 2007.

Littner M. R., Kushida C., Wise M., Davila D. G., Morgenthaler T., Lee-Chiong T., Hirshkowitz M., Loube D. L., Bailey D., Berry R. B., Kapen S., Kramer M., « Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test », *Sleep*, vol. 28, n° 1, p. 113-121, 2005.

Martin V. P., Rouas J.-L., Micoulaud-Franchi J.-A., Philip P., « The Objective and Subjective Sleepiness Voice corpora », *12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 6525-6533, 2020.

Martin V. P., Rouas J.-L., Thivel P., Krajewski J., « Sleepiness detection on read speech using simple features », *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, 2019.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, 2011.

Pellegrino F., Andre-Obrecht R., « Automatic language identification : an alternative approach to phonetic modelling », *Signal Processing*, vol. 80, n° 7, p. 1231-1244, 2000.

Philip P., Dupuy L., Auriacombe M., Serre F., de Sevin E., Sauteraud A., Micoulaud-Franchi J.-A., « Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients », *npj Digital Medicine*, vol. 3, n° 1, p. 2, 2020.

Philip P., Micoulaud-Franchi J.-A., Sagaspe P., De Sevin E., Olive J., Bioulac S., Sauteraud A., « Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders », *Scientific Reports*, vol. 7, n° 1, p. 426-456, 2017.

R Core Team, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017.

Rouas J.-L., Shochi T., Guerry M., Rilliard A., « Categorisation of spoken social affects in Japanese : human vs. machine », *ICPhS*, 2019.

Sangal R., « Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy », *Clinical Neurophysiology*, vol. 110, n° 12, p. 2131-2135, 1999.

Schuller B., Batliner A., Bergler C., Pokorny F. B., Krajewski J., Cychocz M., Vollman R., Roelen S.-D., Schnieder S., Bergelson E., Cristia A., Seidl A., Warlaumont A., Yankowitz L., Nöth E., Amiriparian S., Hantke S., Schmitt M., « The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity », *Interspeech 2019*, 2019.

Schuller B., Steidl S., Batliner A., Schiel F., Krajewski J., « The INTERSPEECH 2011 Speaker State Challenge », *Interspeech 2011*, p. 3201-3204, 2011.

Schuller B., Steidl S., Batliner A., Schiel F., Krajewski J., Weninger F., Eyben F., « Medium-term speaker states-A review on intoxication, sleepiness and the first challenge », *Comput. Speech Lang.*, vol. 28, n° 2, p. 346-374, 2013.

Sjölander K., The Snack Sound Toolkit, Technical report, 2004.

Note de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Frédéric Landragin. Comment parle un robot ? Les machines à langage dans la science-fiction. *Le Bélial éditions*. 2020. 256 pages. ISBN 978-2-84344-965-9.

Lu par **Jan GOES**

Université d'Artois – Centre de recherche Grammatica (UR 4521)

Dans Comment parle un robot ?, *Frédéric Landragin se propose d'apporter des connaissances élémentaires en traitement automatique des langues (TAL) à un public de non-initiés. Pour ce faire il s'appuie – comme dans son volume précédent,* Comment parler à un Alien ? *– sur sa grande connaissance du domaine de la science-fiction, qui se prête très bien à ce type d'ouvrage*.

Le livre comporte un avant-propos, une introduction, cinq grands chapitres et une petite coda qui fait non seulement office de conclusion, mais nous présente également des perspectives d'avenir (*Anticipons !*) ; l'abondante bibliographie de fin de volume est thématique (études scientifiques, romans et nouvelles de SF, films), un index des notions clôt le volume. Les différentes facettes de l'intelligence artificielle (IA) et du TAL sont abondamment illustrées par des exemples de la science-fiction qui vont des grands classiques à des œuvres plus récentes et qui nous montrent la grande distance entre le rêve (la machine presque humaine) et la réalité.

L'*avant-propos* illustre parfaitement cette dichotomie : si le *Terminator T-800* de James Cameron choisit toujours la bonne réplique parmi celles qui défilent sur son écran, et que C-3PO (*La guerre des étoiles*) est capable de pratiquer six millions de formes de communication, on peut légitimement se demander par quels moyens techniques on pourrait obtenir de tels résultats. T-800 et C-3PO existent-ils ? Sont-ils envisageables ? Ce livre propose une réponse en explorant le monde de l'IA et du TAL.

Dans l'introduction, Landragin constate que les machines de la SF parlent bien mieux que les machines réelles ; en fait ce sont des intelligences artificielles douées de langage. Cette constatation constitue le point de départ d'une introduction aux notions fondamentales du TAL, telles que la linguistique computationnelle, la programmation à base de règles (symbolique) ou sur la base de probabilités (approche numérique). Cette dernière permet aux machines d'apprendre par le biais d'un réseau neuronal artificiel (approche connexionniste). Ainsi, le système construit ses propres règles, mais celles-ci sont illisibles par l'humain (c'est une

vraie « boîte noire »). L'IA peut faire des miracles, mais dans des domaines bien délimités ; elle ne peut simuler l'intelligence humaine dans sa globalité. Le TAL, quant à lui, fonctionne avec des corpus annotés, d'abord par des linguistes, ensuite en auto-apprentissage. Ses applications historiques sont la traduction automatique et le dialogue homme-machine, dont l'auteur décrit brièvement l'histoire. Le TAL est fondamental dans la société numérique d'aujourd'hui, mais il est beaucoup moins valorisé par la SF, qui lui préfère les IA « presque humaines ».

Cette constatation ouvre la voie au premier chapitre, qui traite des différentes facettes de l'IA parlante, dont l'un des meilleurs exemples est HAL (L'Odyssée de l'espace). F. Landragin se demande si l'IA mise en scène dans la SF a quelque chose à voir avec la science et si la machine peut dépasser l'humain. Pour lui, les contours de l'IA sont mouvants : là où la machine dépasse l'humain (échecs, reconnaissance optique de caractères imprimés), on sort de l'IA ; n'y reste que ce qui pose des énigmes, c'est-à-dire les domaines où la machine n'a pas dépassé l'homme : l'apprentissage, le langage… Ce dernier reste une barrière : le test de Turing montre qu'aucune machine parlante ne saurait se faire passer pour un humain pendant ne fût-ce que cinq minutes. Ce qui manque à une IA, c'est le sens commun propre aux êtres humains : comme on ne saurait programmer le monde, la machine ne peut apprendre ce qui nous entoure, même si l'apprentissage (*machine learning*) est désormais une facette classique de l'IA, surtout l'apprentissage profond (*deep learning*). Ce dernier se déroule en une phase d'apprentissage et une phase d'application. La première permet à la machine de construire son propre modèle d'apprentissage (différent de l'humain) grâce à ses réseaux neuronaux, la deuxième de l'appliquer. Même si les machines peuvent ainsi dépasser l'humain (dans le jeu de GO par exemple), elles restent dépendantes des données d'apprentissage fournies et des statistiques qui en résultent ; elles ne reconnaîtront donc pas une vache sur la plage et ne pourront donc pas rivaliser avec l'humain, ni soutenir une conversation (même si, depuis les années quatre-vingt-dix, il existe des agents conversationnels animés ou ACA que l'informatique affective tente de rendre plus sociaux). Le chapitre est abondamment illustré d'exemples de la SF (entre autres Tron de Steven Lisberger pour les ACA, Une logique nommée Joe de Murray Leinster pour les systèmes experts, « Demande infos » dans *Short circuit* de John Badham, pour l'apprentissage).

Le chapitre 2 traite du TAL (utilisé entre autres par les ACA) et est annoncé comme le chapitre le plus ardu. Fidèle à son option didactique, F. Landragin propose au lecteur de sauter le chapitre et d'y revenir après (tous les chapitres pouvant se lire indépendamment). Le chapitre aborde la terminologie du TAL et les multiples problèmes qu'il rencontre ; il parle donc indirectement de la linguistique, qui en est indissociable (linguistique théorique, linguistique formelle). Le chapitre est organisé en fonction des grandes thématiques de la linguistique car « *le TAL comporte autant de volets qu'il y a de domaines linguistiques* » (p. 75) : l'analyse lexicale et morphosyntaxique, l'analyse syntaxique, l'analyse sémantique, la détection des entités nommées, les actes de langage, la fouille de textes, avec la difficulté de reconnaissance des chaînes (co-)référentielles. Des exemples très clairs illustrent les problèmes que rencontre le TAL, tandis que chaque thème est mis en relation avec

des œuvres emblématiques de la SF. Comme l'approche numérique, associée à l'apprentissage profond, domine le TAL aux dépens de l'approche symbolique et finalement aussi de la linguistique, F. Landragin ne peut que constater l'éloignement grandissant entre ces deux disciplines. Il en résulte que la machine peut détecter, extraire de l'information, traduire avec *un succès certain*, mais elle ne verra que difficilement la différence avec *un certain succès* et finalement, ne comprendra rien, ni les ambiguïtés (*On a demandé au robot de sortir*), ni la polysémie (*L'alien prend un avocat*), car les statistiques ne s'occupent pas du *sens profond*.

On est donc assez loin de *La Machine qui comprend tout ce qu'on lui dit*, (chapitre 3). La reconnaissance de la parole constitue encore un autre défi pour le TAL. Actuellement, nous pouvons parler à notre GPS, à condition que ces commandes vocales soient bien articulées (enrhumés s'abstenir…), or la machine transcrit la parole sous une forme écrite pour traiter les mots-clés (*aller à*, *supprimer*, etc.). Ici encore, les résultats les plus encourageants s'obtiennent en soumettant de gros corpus de parole à l'apprentissage profond. Il n'en reste pas moins que, pour qu'elle nous comprenne vraiment, la machine devrait être capable de détecter nos émotions (informatique affective, analyse faciale), de s'aligner sur notre style et d'inférer ce qui est pertinent. Actuellement, la science ne produit que des robots qui simulent des émotions, contrairement à ceux de la SF, qui en éprouvent et sont capables de cognition et de langage.

Les amateurs de séries SF connaissent le *Traducteur automatique universel* (chapitre 4) : ce dernier est souvent implicitement présent (dans *Star Trek*, par exemple) pour éviter les difficultés de scénario, et les incompréhensions. En fait, trois possibilités s'offrent à nous pour communiquer malgré la multiplicité des langues : le traducteur universel, le recours à une langue universelle (le *globish* ?) ou l'augmentation de nos facultés cognitives, trois solutions explorées par la SF. Mais, quelle est la plausibilité scientifique du *traducteur* ? La traduction automatique c'est encore et toujours de la linguistique, ce qui implique que les problèmes déjà évoqués (ambiguïtés, polysémie) restent d'actualité ; en d'autres termes : pour bien traduire il faut comprendre le sens. Ceci est illusoire, car il faudrait encoder toutes les connaissances du monde, le bon sens humain, et l'imaginaire ! Dans ce domaine également, les chercheurs sont passés de la linguistique à la statistique et à l'apprentissage artificiel, ce qui donne des résultats légèrement meilleurs que la programmation basée sur des règles. En fin de chapitre, nous découvrons les réalités et enjeux actuels de la traduction automatique : la réintégration de connaissances linguistiques, la traduction assistée par la technologie, *Skype* avec traduction simultanée. On ne pourra néanmoins jamais se passer de l'homme pour les traductions littéraires, journalistiques, ou encore juridiques. Si le traducteur automatique universel restera « de la magie », il apparaît néanmoins que Google Neural Machine Translation a construit une représentation interne dans ses réseaux neuronaux profonds qui lui permet d'être multilingue.

Pour ce qui concerne le dialogue entre humains et machines (chapitre 5), F. Landragin se pose les questions suivantes : parle-t-on différemment à un robot et

à un humain ? Les machines à langage qui intègrent peu à peu la société, ont-elles des répercussions sur notre comportement ? Il se concentre sur la première question, car le dialogue H-M est un sous-domaine du TAL qui recouvre de nombreux domaines (déjà mentionnés, c'est pourquoi le chapitre se trouve en fin de volume) : la reconnaissance de la parole, la syntaxe, la sémantique, la pragmatique ; il a donné lieu à un nouveau domaine de recherche, la robotique sociale. On distingue le dialogue fermé, centré sur un domaine de spécialité et basé sur des arbres de décision (les réservations par exemple) du dialogue ouvert. Dans le premier cas un robot taxi ne comprendra pas « *Conduis-moi n'importe où* » (Verhoeven, *Total Recall*), mais le dialogue ouvert (HAL, Samantha, dans le film *Her* de Spike Jonze, 2013) reste difficile pour toutes les raisons langagières et de connaissance du monde déjà évoquées (cf. le test de Turing). On peut néanmoins améliorer les performances en exploitant des schémas Winograd dans l'apprentissage profond. Pour que le robot parle, il lui faut également une acoustique humaine (prosodie, intonation) ; le problème reste néanmoins le même : pour une synthèse vocale réussie, il faut comprendre le sens. Malgré les progrès (les approches statistiques sur corpus ; les premiers produits grand public), la gestion du dialogue H-M pose encore de nombreux problèmes techniques, que l'auteur analyse à l'aide d'exemples de la SF, évidemment !

Coda : *anticipons !* Actuellement, les machines se débrouillent avec des formes de surface et des statistiques ; on commence néanmoins à (re)mettre la morphosyntaxe et la syntaxe dans les programmes. Reste le plus dur : la sémantique et la pragmatique. Le TAL a un bel avenir devant lui, car outre les avancées dans la traduction automatique, la reconnaissance de la parole, le dialogue H-M, d'autres applications voient le jour, comme le résumé automatique des textes. On opérera probablement un retour vers la linguistique, avec une collaboration moins poussée, car le TAL se fait de plus en plus aider par l'IA. Cette dernière évolue vers des robots situés et ancrés, en interaction avec leur environnement (robotique comportementale). Nous aurons des robots utiles (médecins, traders, avocats), nourris d'immenses corpus. La SF, quant à elle, explore les limites de l'IA et pose les questions qui inspirent la science.

Tout comme *Comment parler à un alien* (Landragin, 2018), centré sur la linguistique, ce livre constitue une excellente introduction au TAL, très didactique, avec pléthore d'exemples classiques et récents de la SF à l'appui. À la fois les étudiants en sciences du langage, en traductologie, ou en TAL, et les amateurs éclairés de SF y trouveront amplement leur compte. Lecture recommandée, que nous avons également commandée pour notre bibliothèque universitaire.

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Lucie GIANOLA : lucie.gianola@yahoo.fr

Titre : Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique

Mots-clés : reconnaissance d'entités nommées, genre textuel, analyse criminelle, linguistique de corpus.

Title: *Textual Aspects of Judicial Proceedings Files and Perspectives for its Automatic Processing*

Keywords: *named entity recognition, textual genre, criminal analysis, corpus linguistics.*

Thèse de doctorat en sciences du langage, Agora, Université de Cergy-Pontoise, sous la direction de Julien Longhi (Pr, Université de Cergy-Pontoise). Thèse soutenue le 28/02/2020.

Jury : M. Julien Longhi (Pr, Université de Cergy-Pontoise, directeur), M. Patrick Paroubek (IR HDR, LIMSI, CNRS, rapporteur), Mme Sylvie Monjean-Decaudin (Pr, Université Paris IV, rapporteuse et présidente), M. Laurent Chartier (Colonel de gendarmerie, Gendarmerie nationale, examinateur), Mme Bénédicte Pincemin (CR, IHRIM, examinatrice), M. Olivier Ribaux (Pr, Université de Lausanne, examinateur).

Résumé : *L'analyse criminelle est une discipline d'appui aux enquêtes pratiquée au sein de la Gendarmerie nationale. Elle repose sur l'exploitation des documents compilés dans le dossier de procédure judiciaire (auditions, perquisitions, rapports d'expertise, données téléphoniques et bancaires, etc.) afin de synthétiser les informations collectées et de proposer un regard neuf sur les faits examinés. Si l'analyse criminelle a recours à des logiciels de visualisation de données (p. ex.* Analyst's Notebook

d'IBM) pour la mise en forme des hypothèses formulées, la gestion informatique et textuelle des documents de la procédure est à l'heure actuelle entièrement manuelle. D'autre part, l'analyse criminelle s'appuie entre autres sur la conceptualisation des informations du dossier en entités criminelles pour formaliser son travail. La présentation du contexte de recherche détaille la pratique de l'analyse criminelle ainsi que la constitution du dossier de procédure judiciaire en tant que corpus textuel. Nous proposons des perspectives pour l'adaptation des méthodes de traitement automatique de la langue et d'extraction d'information au cas d'étude, notamment la mise en parallèle des concepts d'entité en analyse criminelle et d'entité nommée. Cette comparaison est réalisée sur les plans conceptuels et linguistiques. À l'aide d'un corpus d'exemples constitué d'un dossier de procédure consacré à un homicide, une première approche minimale de détection des entités dans les auditions de témoins basée sur des grammaires locales est présentée. Enfin, le genre textuel étant un paramètre à prendre en compte lors de l'application de traitements automatiques à du texte, nous construisons une structuration du genre textuel « légal » en discours, genres et sous-genres par le biais d'une étude textométrique visant à caractériser différents types de textes (dont les auditions de témoins) produits par le domaine de la justice.

L'objectif de ce travail de thèse est de proposer une compréhension approfondie des concepts des trois domaines concernés (analyse criminelle, extraction d'information et linguistique textuelle) afin de poser les bases méthodologiques et épistémologiques de l'application des méthodes automatiques au cas de l'analyse criminelle.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02522680

Fadila TALEB : talebfadila@gmail.com

Titre : L'argumentation judiciaire à travers le prisme des scénarios modaux

Mots-clés : scénario modal, modalité, zone modale, sémantique des modalités, discours judiciaire, argumentation, rhétorique, genre textuel, textométrie, linguistique de corpus, droit des transports.

Title: *Judicial Argumentation through the Prism of Modal Scenarios. Application for Assistance in the Interpretation of Court Decisions*

Keywords: *modal scenario, modality, modal zone, semantics of modalities, legal discourse, argumentation, rhetoric, kind of text, textometry, corpus linguistics, transportation law.*

Thèse de doctorat en sciences du langage, linguistique, DYnamique du Langage In Situ (DYLIS), département sciences du langage et de la communication, UFR lettres et sciences humaines, Université de Rouen Normandie, sous la direction de Laurent Gosselin (Pr, Université de Rouen Normandie) et Maryvonne Holzem (MC HDR émérite, Université de Rouen Normandie). Thèse soutenue le 8/11/2019.

Jury : M. Laurent Gosselin (Pr, Université de Rouen Normandie, codirecteur), Mme Maryvonne Holzem (MC HDR émérite, Université de Rouen Normandie, codirectrice), M. Laurent Gautier (Pr, Université de Bourgogne, rapporteur), M. Doninique Legallois (Pr, Université Sorbonne Nouvelle-Paris 3, rapporteur), M. Alain Rabatel (Pr, Université Claude Bernard – Lyon 1, président).

Résumé : *Le travail de recherche présenté dans cette thèse s'inscrit dans le cadre général des travaux sur les humanités numériques qui cherchent, entre autres, à contribuer à l'amélioration des interactions homme-machine. L'objectif de l'étude est double. Dans un premier temps, il s'agit d'étudier un corpus de décisions de justice contenues dans la base de données de l'Institut du Droit International des Transports (IDIT) afin de déterminer les contraintes linguistiques du genre judiciaire. Dans un second temps, il est question de proposer des parcours interprétatifs pouvant aider les utilisateurs dans leur accès à l'information juridique recherchée. La problématique de l'aide à l'interprétation est appréhendée à travers l'étude des modalités et des scénarios modaux.*

Le parti pris de cette recherche est de considérer la pluridisciplinarité comme un atout théorique et méthodologique qui contribue à mieux éclairer un objet d'étude. De ce fait, plusieurs approches (sémantique des modalités, sémantique textuelle, argumentation rhétorique, textométrie) sont convoquées et articulées pour œuvrer ensemble vers les objectifs fixés. L'analyse du corpus a été menée à deux niveaux et selon deux approches.

Dans la première partie, l'analyse empirique proposée est quantitative et contrastive. Elle est menée au niveau microtextuel et mésotextuel dans la mesure où elle se focalise sur l'étude du lexique. Aidée de l'outil TXM, cette première investigation a permis une caractérisation linguistique globale du corpus et un premier aperçu de son profil modal grâce notamment à l'introduction de la notion de zone modale. Elle a également mis en exergue des expressions modales, constructions concessives, routines discursives, etc. qui focalisent sur des moments clés dans le déroulement argumentatif et peuvent donc servir dans le cadre de l'aide à l'interprétation.

Dans la seconde partie, l'étude empirique porte sur des analyses modales menées sur des textes complets. Elle est donc abordée dans une approche qualitative et au niveau macrotextuel. Cette analyse aboutit à la formulation d'un modèle de scénario modal minutieusement décrit pour trois sous-genres judiciaires : jugement du tribunal de commerce, arrêt de la cour d'appel et arrêt de la Cour de cassation. Pour chacun des sous-genres, le scénario modal a été décomposé en plusieurs niveaux : scénario modal apparent et scénario modal sous-jacent (selon les modalités qui l'ont construit : modalités de premier plan et modalités d'arrière-plan), et selon qu'il caractérise un texte complet (scénario modal global) ou une zone spécifique de ce texte (sous-scénario modal). Par ailleurs, la présentation schématique (semblable à un algorithme) proposée

pour les scénarios modaux a mis en évidence le rôle que représenterait chaque zone modale dans la perspective d'une aide à l'interprétation.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02995083

Tian TIAN : tian.tian@live.cn

Titre : Adaptation au domaine et combinaison de modèles pour l'annotation de textes multisources et multidomaines

Mots-clés : adaptation au domaine, reconnaissance des entités nommées, apprentissage automatique, champs aléatoires conditionnels, réseaux de neurones.

Title: *Domain Adaptation and Model Combination for the Annotation of Multi-source, Multi-domain Texts*

Keywords: *domain adaptation, named entity recognition, machine learning, conditional random fields, neural networks.*

Thèse de doctorat en sciences du langage, LaTTiCe, UMR 8094, Université Sorbonne Nouvelle – Paris 3, sous la direction de Isabelle Tellier (Pr, Université Sorbonne Nouvelle – Paris 3), Thierry Poibeau (DR, CNRS, LaTTiCe, UMR 8094) et Marco Dinarelli (CR, CNRS, Laboratoire d'Informatique de Genoble, UMR 5217). Thèse soutenue le 16/10/2019.

Jury : Mme Isabelle Tellier (Pr, Université Sorbonne Nouvelle – Paris 3, codirectrice), M. Thierry Poibeau (DR, CNRS, LaTTiCe, UMR 8094, codirecteur), M. Marco Dinarelli (CR, CNRS, Laboratoire d'Informatique de Genoble, UMR 5217, codirecteur), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, rapporteuse), Mme Anne-Laure Ligozat (MC HDR, ENSIIE, LIMSI, rapporteuse),, Mme Sophie Prévost (DR, CNRS, LaTTiCe, UMR 8094, présidente), M. Patrick Marty (IR, Fnac, examinateur).

Résumé : *Aujourd'hui, de nombreux services en ligne proposent aux utilisateurs de commenter, éditer et partager leurs points de vue sur différents sujets de discussion. Ce type de contenu, ou « contenu généré par utilisateur (*user generated content*, UGC) », est maintenant devenu la ressource principale pour les analyses d'opinions sur Internet. Les sujets de ces opinions varient du personnage politique, au produit sur un marché quelconque, au climat, aux sites touristiques et à la vie privée personnelle. L'analyse de cette masse de données permet de suivre l'évolution des opinions au fil du temps, ce qui pourrait être utile pour les changements de stratégies ou le choix de décision, et permettre l'évaluation de ces changements et des prises de décisions. Néanmoins, à cause des abréviations, du bruit, des fautes d'orthographe et de toute autre sorte de problèmes, les outils de classifications des textes et de traitements automatiques des langues ont des performances plus faibles que sur les textes bien formés.*

Cette thèse a pour objet la reconnaissance d'entités nommées sur les contenus générés par les utilisateurs sur Internet, les données cibles viennent essentiellement de forums spécialisés, Facebook et Twitter. Nous avons établi un corpus d'évaluation avec des textes multisources et multidomaines : nous avons distingué les textes de forums, de Facebook et des tweets *comme différentes sources et nous avons traité les domaines de discussions sur des produits qui varient entre fast-food, automobile, musique en* streaming *et jouets pour les enfants. Ensuite, nous avons développé un modèle de champs aléatoires conditionnels, entraîné sur un corpus annoté disponible, provenant des contenus générés par les utilisateurs (uniquement extrait de Twitter).*

Dans le but d'améliorer les résultats de la reconnaissance d'entités nommées, nous avons d'abord développé un étiqueteur morphosyntaxique pour les contenus générés par les utilisateurs. Les étiqueteurs morphosyntaxiques appris sur les textes bien formés ne fonctionnent pas aussi bien sur les données générées par les utilisateurs. Nous avons donc testé deux méthodes d'adaptation au domaine en mélangeant les textes bien formés et les textes générés par les utilisateurs dans les données d'apprentissage pour améliorer la performance de l'étiqueteur morphosyntaxique. Ensuite, nous avons utilisé les étiquettes prédites par cet étiqueteur morphosyntaxique comme un attribut du modèle des champs aléatoires conditionnels pour le modèle d'extraction d'entités nommées. Enfin, pour transformer les contenus générés par les utilisateurs en textes bien formés, nous avons développé un modèle de normalisation lexicale basé sur des réseaux de neurones pour détecter les mots non standards et proposer une forme correcte pour les remplacer.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02473489/
