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 termino-ontologique 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, n° 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, n° 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, n° 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.