Nouvelles applications du TAL

Géraldine Damnati* — **Diana Inkpen****

* *Orange Innovation, DATA & AI, Lannion*
** *Université d'Ottawa, Canada*

RÉSUMÉ. *Les récentes avancées en traitement automatique des langues ainsi que la démocratisation de l'accès à des modèles de langue très performants et aux librairies logicielles permettant de les mettre en œuvre ont permis aux champs d'applications du TAL de s'étendre notablement. Cet article introduit le numéro spécial* « Nouvelles applications du TAL » *en tentant de dresser un panorama des domaines applicatifs du TAL, par le prisme des ateliers internationaux organisés autour de ces nouvelles thématiques. Parmi ces champs applicatifs, l'analyse des interactions dans les réseaux sociaux occupe une place significative tant du point de vue scientifique que du point de vue des enjeux sociétaux sous-jacents. C'est le domaine de l'article publié dans ce numéro, qui traite de la détection de comportements abusifs en ligne.*

ABSTRACT. *Recent advances in Natural Language Processing, along with easier access to very powerful Language Models and libraries to manipulate them have led to a notable expansion of the NLP application field. This article introduces the special issue "New Applications in NLP" with an overview of applicative domains through the prism of the various international* workshops *organized around these new topics. Among these applicative fields, social media interaction analysis plays a significant role, both from scientific and societal points of view. It is the applicative domain of the article published in this special issue, on abusive online behaviour detection.*

MOTS-CLÉS : *applications du traitement automatique des langues, champs applicatifs, cas d'usage.*

KEYWORDS: *Natural Language Processing applications, applicative fields, use cases.*

1. Introduction

Les récentes avancées en traitement automatique des langues (TAL) ainsi que la démocratisation de l'accès à des modèles de langue très performants et aux librairies logicielles permettant de les mettre en œuvre ont permis aux champs d'applications du TAL de s'étendre notablement. Du point de vue des compétences, le métier de *data scientist* se développe dans nombre d'entreprises et d'institutions publiques, s'accompagnant souvent désormais de compétences en TAL qui sont devenues incontournables dans de nombreux domaines d'activité. On assiste ainsi à la fois à une diversification de ces domaines applicatifs et à une augmentation des cas d'usage au sein de ces domaines.

Cette évolution s'accompagne de nouvelles problématiques de recherche. S'il n'est pas nécessairement question de nouvelles tâches, à proprement parler, les tâches traditionnelles comme la classification, l'extraction d'information, le *parsing*, le résumé ou la traduction s'orientent vers le traitement de phénomènes plus complexes, à partir de données textuelles hétérogènes, voire en combinaison avec de nouvelles modalités. Les domaines de spécialité, comme le juridique ou le médical, font appel au TAL sur des types de documents de nature de plus en plus variée. Concernant l'intermodalité, le domaine de la musique, par exemple, intègre désormais le TAL et envisage la combinaison des modalités textuelles et musicales. Le domaine des réseaux sociaux, qui à lui seul intègre de nombreux cas d'usage, peut combiner, par exemple, l'analyse textuelle des messages avec l'analyse des graphes d'interaction sous-jacents.

Signe de ce dynamisme, de plus en plus de *workshops* dédiés à des domaines applicatifs spécialisés ont vu le jour ces dernières années, accolés aux principales conférences internationales du domaine. Ainsi, il ne s'agit pas ici de dresser un état de l'art au sens habituel du terme, mais plutôt de proposer un recensement des domaines qui voient se structurer autour d'eux des communautés dédiées, sous le prisme des *workshops* que ces communautés organisent.

2. Approfondissement des domaines applicatifs historiquement abordés

Si certains champs applicatifs sont déjà abordés depuis plusieurs années comme le journalisme, l'analyse des réseaux sociaux, les humanités numériques ou le domaine biomédical, ils le sont maintenant sous l'angle de nouvelles problématiques, signe d'un dynamisme notable des activités dans ces domaines.

On trouve ainsi, au-delà du *workshop* généraliste *NLP meets journalism* (NLPJ, 2018) créé en 2015, des *workshops* dédiés à l'exploitation des graphes de connaissances pour le journalisme (SEMANTICJOURNALISM, 2020) ou à l'extraction d'événements sociopolitiques (AESPEN, 2020).

Un *workshop* dédié aux langues anciennes (LT4HALA, 2020) ainsi qu'un autre consacré à l'évolution des langues à travers le temps (LChange, 2021) accompagnent

désormais ceux dédiés plus largement aux humanités numériques (SIGHUM, 2020) ou (NLP4DH, 2021).

L'analyse des réseaux sociaux et plus largement des « données bruitées générées par les utilisateurs » (W-NUT, 2021) est menée sous différents angles comme la détection de la censure, de la propagande et de la désinformation (NLP4IF, 2020), les menaces dans les conversations en ligne (STOC, 2020), la modélisation des opinions (PEOPLES, 2020), la vérification de faits (FEVER, 2021), ou encore le harcèlement (TRAC, 2020). L'analyse d'opinions dans les commentaires utilisateurs pour le domaine de la relation client et du marketing est également un domaine traité depuis plusieurs années, et qui est appréhendé avec des tâches de plus en plus sophistiquées où les différentes dimensions de l'opinion sont abordées conjointement.

Le domaine biomédical, quant à lui, connaît un nombre croissant de *workshops* spécialisés. Si le *workshop* BioNLP (BioNLP, 2020) en est en 2021 à sa vingtième édition, d'autres événements ont enrichi le domaine comme LOUHI (LOUHI, 2021) créé en 2008 et consacré à la fouille de données médicales au sens large et ClinicalNLP (ClinicalNLP, 2020) dédié à l'analyse des données cliniques créé en 2016 ou MEDA (MEDA, 2020) dédié à l'exploitation des masses de données médicales avec les problématiques sous-jacentes d'agrégation dans des bases de connaissances. Les plus récemment créés concernent les domaines du fonctionnement et du handicap (AI4Function, 2020), la problématique de l'extraction d'information à partir de données hétérogènes (SIIRH, 2020) et, bien entendu, les recherches autour des documents liés au Covid-19 (NLPCovid19, 2020).

3. Diversification des domaines applicatifs

Parallèlement à ces domaines en expansion, de nouveaux domaines applicatifs ont fait leur apparition ces dernières années, ouvrant la voie à l'utilisation du TAL sur des données de nature très variée. Une communauté s'est structurée autour du TAL pour le domaine de l'économie et des finances (FIN-ECO, 2019) et des *workshops* spécialisés ont vu le jour autour de l'extraction de connaissances (KDF, 2021) ou de l'analyse des rapports descriptifs (*Financial Narrative Processing*) (FNP-FNS, 2020) où l'on trouve, par exemple, des jeux de questions-réponses dédiés à ce type de textes. Des journées sont organisées en 2021 autour du langage des affaires (PLIN, 2021) (communication des organisations et *Business Language*). Le e-commerce fait également appel au TAL avec différentes finalités applicatives (recommandation, classification pour campagnes marketing, chatbots…) (ECOMNLP, 2020) (ECNLP, 2021).

Le TAL est également présent autour de questions sociétales. Les textes de loi, les textes juridiques (IberLegal, 2019) et les documents liés aux administrations publiques (LT4Gov, 2020) font ainsi l'objet d'une attention particulière pour leur analyse. Le domaine de la justice prédictive est, bien entendu, concerné par le TAL. Dans un autre cadre, des travaux sont également menés en lien avec la sécurité (RISS, 2020).

Enfin, la musique fait désormais partie des nouveaux domaines où le TAL a fait son apparition, avec l'organisation du premier *workshop* NLP4Musa (NLP4MUSA, 2020) où sont abordés les thèmes de la recherche d'informations ou les agents conversationnels spécifiques à la musique. La musique y est également traitée comme donnée à part entière avec, par exemple, l'analyse de paroles de chansons ou l'introduction en 2020 d'un premier modèle de représentation multimodal incluant la musique (MusicBERT).

4. Contenu du numéro spécial

L'article publié dans le cadre de ce numéro spécial de la revue TAL s'inscrit dans le domaine applicatif de l'analyse des réseaux sociaux. Le cas d'usage abordé est celui de la détection de messages abusifs dans les conversations en ligne et l'article propose une analyse conjointe des dimensions langagières et interactives de ces conversations. Dans *Approche multimodale par plongement de texte et de graphes pour la détection de messages abusifs*, Noé Cécilion, Richard Dufour et Vincent Labatut abordent ainsi la problématique de la modération des plateformes d'échanges en ligne. L'approche proposée pour la détection de contenus abusifs consiste à combiner des plongements de mots et des plongements de graphes de façon à exploiter conjointement l'analyse du contenu textuel des messages et le graphe des interactions entre les participants.

Remerciements

Nous remercions le comité éditorial et scientifique de la revue TAL, ainsi que le comité scientifique invité, en particulier les relecteurs, qui ont contribué par leur temps et leurs efforts à la qualité de ce numéro : Muhammad Abdul-Mageed (University of British Columbia), Lina Rojas Barahona (Orange Labs), Frédéric Béchet (Aix-Marseille Université), Caroline Brun (Navelabs Europe), Nathalie Camelin (Université du Mans), Elena Epure (Deezer), Olivier Ferret (CEA), Mathias Gallé (Naverlabs Europe), Natalia Grabar (Université de Lille), Gaël Guibon (Télécom Paris, SNCF), Catherine Kobus (Airbus), Claudia Peersman (Universty of Bristol), Giuseppe Riccardi (Università di Trento), Mathieu Roche (CIRAD), Antonio Moreno Sandoval (Universidad Autónoma de Madrid), Xavier Tannier (Sorbonne Université).

5. Bibliographie

AESPEN, « Workshop Automatic Extraction of Socio-Political Events from News », `http://www.lrec-conf.org/proceedings/lrec2020/workshops/AESPEN2020/index.html`, 2020.

AI4Function, « First Workshop on Artificial Intelligence for Function, Disability, and Health », `https://slate.cse.ohio-state.edu/AI4Function2020/`, 2020.

BioNLP, « Workshop on Biomedical Natural Language Processing », https://aclweb.org/aclwiki/BioNLP_Workshop, 2020.

ClinicalNLP, « 3rd Clinical Natural Language Processing Workshop », https://clinical-nlp.github.io, 2020.

ECNLP, « The 4th Workshop on e-Commerce and NLP », https://sites.google.com/view/ecnlp/, 2021.

ECOMNLP, « Workshop on NLP in E-Commerce », https://ecomnlp.github.io, 2020.

FEVER, « Fourth Fact Extraction and VERification workshop », https://fever.ai/, 2021.

FIN-ECO, « Financial NLP Group », http://wp.lancs.ac.uk/cfie/fin-eco-nlp/, 2019.

FNP-FNS, « 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation », http://wp.lancs.ac.uk/cfie/fnp2020/, 2020.

IberLegal, « NLP for Legal Domain », https://jurix2019.oeg-upm.net/iberlegal.html, 2019.

KDF, « Workshop on Knowledge Discovery from Unstructured Data in Financial Services », https://aaai-kdf.github.io/kdf2021/, 2021.

LChange, « 2nd International Workshop on Computational Approaches to Historical Language Change », https://languagechange.org/events/2021-acl-lchange/, 2021.

LOUHI, « 12th International Workshop on Health Text Mining and Information Analysis », https://louhi2021.fbk.eu/, 2021.

LT4Gov, « 1st Workshop on Language Technologies for Government and Public Administration », https://www.plantl.gob.es/tecnologias-lenguaje/comunicacion-formacion/eventos/Paginas/lt4gov.aspx, 2020.

LT4HALA, « 1st Workshop on Language Technologies for Historical and Ancient LAnguages », https://circse.github.io/LT4HALA/, 2020.

MEDA, « Workshop on Curative Power of MEdical DAta », https://jcdl2020bionlp.wordpress.com/, 2020.

NLP4DH, « Workshop on Natural Language Processing for Digital Humanities », https://rootroo.com/en/nlp4dh-workshop/, 2021.

NLP4IF, « Third Workshop on NLP for Internet Freedom : Censorship, Disinformation, and Propaganda », http://www.netcopia.net/nlp4if/, 2020.

NLP4MUSA, « First Workshop on NLP for Music and Audio », https://sites.google.com/view/nlp4musa, 2020.

NLPCovid19, « NLP COVID-19 Workshop », https://www.nlpcovid19workshop.org/, 2020.

NLPJ, « Natural Language Processing meets Journalism », http://nlpj2018.fbk.eu/, 2018.

PEOPLES, « Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media », https://peopleswksh.github.io/, 2020.

PLIN, « PLIN Linguistic Days : Linguistics applied to Business Language in a multilingual and multicultural world », https://uclouvain.be/fr/instituts-recherche/ilc/plin/plinday2021.html, 2021.

RISS, « 3rd International Workshop on Research Innovation for Secure Societies », http://campus.pub.ro/RISS2020/, 2020.

SEMANTICJOURNALISM, « Semantic and knowledge graph advances for journalism », https://almoslmi.github.io/SemanticJournalism/, 2020.

SIGHUM, « Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature », https://sighum.wordpress.com/events/latech-clfl-2020/, 2020.

SIIRH, « 1st Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages », https://sites.google.com/view/siirh2020/, 2020.

STOC, « First International Workshop on Social Threats in Online Conversations : Understanding and Management », https://social-threats.github.io/, 2020.

TRAC, « Second Workshop on Trolling, Aggression and Cyberbullying », https://sites.google.com/view/trac2/home, 2020.

W-NUT, « Workshop on Noisy User-generated Text », http://noisy-text.github.io, 2021.