18e Conférence en Recherche d'Information et Applications
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues
(CORIA-TALN) [1]

Actes de CORIA-TALN 2023.

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023

Florian Boudin, Béatrice Daille, Richard Dufour, Oumaima El Khettari, Maël Houbre, Léane Jourdan, Nihel Kooli (Éds.)

Paris, France, 5 au 9 juin 2023

1. https://coria-taln-2023.sciencesconf.org/

Avec le soutien de

Préface

Le nombre d'articles scientifiques produits chaque année ne cesse d'augmenter. Rien que dans l'archive ouverte `arXiv`, le nombre d'articles scientifiques déposés en 2022 s'élève à plus de 185 000, soit près de 500 dépôts chaque jour [2]. Face à cette explosion du volume de littérature scientifique, des solutions intelligentes sont nécessaires pour faciliter la recherche et la lecture des articles scientifiques et pour en analyser le contenu et y extraire des informations utiles aux chercheurs et aux applications qui les utilisent. De plus, l'avènement de la science ouverte et la disponibilité croissante des textes intégraux soulèvent de nouveaux enjeux pour le traitement automatisé des articles scientifiques et interroge sur l'utilisabilité des modèles de langues actuels. Comment analyser et rendre accessible les informations contenues dans les tables, les équations ou les figures sont autant de questions qui doivent être explorées.

L'atelier sur l'Analyse et la Recherche de Textes Scientifiques (ARTS) [3], qui se déroule le 5 juin 2023 pendant la conférence CORIA-TALN à Paris, se veut un lieu de rencontre et d'échange pour les chercheurs en Recherche d'Information (RI) et en Traitement Automatique des Langues (TAL) qui s'intéressent aux textes scientifiques. Douze communications écrites ont été acceptées, puis présentées sous la forme d'un poster pendant l'atelier.
Les travaux présentés portent sur une diversité de problématiques, allant de l'annotation et de la collecte de corpus, à la classification de documents, à la traduction automatique, ou encore la simplification de textes scientifiques.

Nous adressons des remerciements particuliers au conférencier invité, Mathieu Constant (ATILF, Université de Lorraine), qui nous a fait le plaisir de présenter ses travaux sur la *construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX*.

Le comité d'organisation remercie également l'ensemble des contributeurs à l'atelier, la diversité des travaux de recherche présentés montrant l'intérêt important, et croissant, que suscite ce domaine de recherche ouvert.

Nous souhaitons enfin remercier chaleureusement l'ensemble des membres du comité de programme scientifique pour leur aide importante quant à la relecture et à la sélection des papiers.

L'atelier ARTS est soutenu par le projet DGA-CNRS NaviTerm [4] (convention 2022 65 0079 CNRS Occitanie Ouest) ayant pour objectif d'accélérer la montée en compétence des chercheurs par la création automatisée de représentations navigables des connaissances scientifiques.

2. https://arxiv.org/stats/main
3. https://arts2023.sciencesconf.org
4. https://cnrs-naviterm.github.io/

Comités

Comité scientifique

— Sabine Barreaux (INIST, CNRS)
— Guillaume Cabanac (IRIT, Université Toulouse 3)
— Florian Boudin (LS2N, Nantes Université)
— Mathieu Constant (ATILF, Université de Lorraine)
— Béatrice Daille (LS2N, Nantes Université)
— Richard Dufour (LS2N, Nantes Université)
— Natalia Grabar (STL, Université de Lille)
— Thierry Hamon (LISN, Université Sorbonne Paris Nord)
— Evelyne Jacquey (ATILF, CNRS)
— Cyril Labbé (LIG, Université Grenoble Alpes)
— François Yvon (LISN, CNRS)

Comité d'organisation

— Florian Boudin (LS2N, Nantes Université)
— Béatrice Daille (LS2N, Nantes Université)
— Richard Dufour (LS2N, Nantes Université)
— Oumaima El Khettari (LS2N, Nantes Université)
— Maël Houbre (LS2N, Nantes Université)
— Léane Jourdan (LS2N, Nantes Université)
— Nihel Kooli (DGA)

Présentation invitée

Mathieu Constant (ATILF, Université de Lorraine)

Titre : Construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX

Résumé : La plateforme ISTEX (https://www.istex.fr/) permet d'accéder à une large base d'archives scientifiques comptant plus de 25 millions de documents de tous les grands domaines scientifiques. Les documents incluent non seulement les métadonnées mais aussi le texte plein, et ont été prétraités de manière homogène pour faciliter leur traitement automatique. Dans cet exposé, nous présenterons une initiative pour créer une dynamique de recherche en TAL et TDM autour de ces données. En particulier, nous présenterons les travaux en cours pour la construction d'un jeu de données dédié au TAL et la fouille de textes.

Table des matières

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

La pré-annotation automatique de textes cliniques comme support au dialogue avec les experts du domaine lors de la mise au point d'un schéma d'annotation

Virgile Barthet[1] Marie José Aroulanda[2] Laura Monceaux-Cachard[3]
Christine Jacquin[3] Cyril Grouin[1] Johann Gutton[2] Guillaume Hoquet[2]
Pascal de Groote[4] Michel Komajda[2] Emmanuel Morin[3] Pierre Zweigenbaum[1]
(1) Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
(2) Hôpital Saint Joseph, DIMID et Service de Cardiologie, Paris, France
(3) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
(4) CHU de Lille, Service de Cardiologie, Lille, France
{prenom.nom}@lisn.fr {laura.monceaux,jacquin-c,emmanuel.morin}@ls2n.fr
{maroulanda,jgutton,ghoquet,mkomajda@ghpsj.fr} pascal.degroote@chu-lille.fr

RÉSUMÉ

La pré-annotation automatique de textes est une tâche essentielle qui peut faciliter l'annotation d'un corpus de textes. Dans le contexte de la cardiologie, l'annotation est une tâche complexe qui nécessite des connaissances approfondies dans le domaine et une expérience pratique dans le métier. Pré-annoter les textes vise à diminuer le temps de sollicitation des experts, facilitant leur concentration sur les aspects plus critiques de l'annotation. Nous rapportons ici une expérience de pré-annotation de textes cliniques en cardiologie : nous présentons ses modalités et les observations que nous en retirons sur l'interaction avec les experts du domaine et la mise au point du schéma d'annotation.

ABSTRACT

Automatic pre-annotation of clinical texts to support the dialogue with domain experts during the design of an annotation schema

Automatic text pre-annotation is an essential task that can facilitate the annotation of a text corpus. In the context of cardiology, manual text annotation is a complex task that requires in-depth domain knowledge and practical professional experience. Pre-annotating texts aims to reduce the time spent by experts on manual annotation and to focus their intervention on more critical aspects of annotation. We report here a pre-annotation experiment for clinical texts in cardiology : we present its modalities and the lessons learnt about our interaction with domain experts and on annotation schema design.

MOTS-CLÉS : TAL, Médical, Cardiologie, Annotation, Pré-annotation, Schéma d'annotation.

KEYWORDS: NLP, Medical, Cardiology, Annotation, Pre-annotation, Annotation schema.

1 Introduction

Selon une étude de l'OMS datant de 2020 (World Health Organization, 2020), les maladies cardiovasculaires sont l'une des principales causes de décès dans le monde et l'analyse des données cliniques joue un rôle crucial dans l'amélioration de la prise en charge des patients. Dans ce contexte,

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 1–7.

nous cherchons à analyser les textes de dossiers de patients en cardiologie dans le but de déterminer précocément si un patient insuffisant cardiaque présente un risque de décès dans les trois mois suivant son hospitalisation. Ces textes ont en commun avec les articles scientifiques de traiter d'un domaine spécialisé, et mettent en jeu des entités de même nature. Les méthodes à l'état de l'art pour la détection automatique de ces entités reposent sur l'entraînement de classifieurs supervisés à partir de corpus annotés manuellement. (Patel *et al.*, 2018)

Cependant, l'annotation manuelle de grands volumes de données textuelles cliniques dans le contexte de la cardiologie peut être une tâche complexe et chronophage, nécessitant une connaissance approfondie du domaine et une expérience pratique dans le métier. C'est pourquoi une pré-annotation automatique des textes est souvent mise en place pour faciliter l'annotation humaine d'un corpus de textes et aider à réduire le temps nécessaire à l'annotation, tout en améliorant potentiellement la cohérence de ces annotations (Lingren *et al.*, 2012).

Dans ce contexte, la présente étude se focalise sur les méthodes et ressources utilisées pour pré-annoter des textes cliniques en cardiologie, ainsi que sur l'interaction avec des experts du domaine pour définir et redéfinir les types d'entité, les points d'arbitrage et les méthodes de représentation de l'information. L'objectif est de mettre en évidence l'importance de ces éléments pour l'annotation efficace et précise de données textuelles en cardiologie. L'annotation requiert également d'établir un bon schéma d'annotation, cohérent et qui reflète correctement la manière de penser des experts (Shinohara *et al.*, 2022). Cette étude constitue un retour d'expérience sur la pré-annotation de textes cliniques en cardiologie et les échanges avec des experts non-initiés au domaine du traitement automatique des langues, mettant en évidence les défis spécifiques liés à la communication entre des experts d'un autre domaine et des informaticiens, ainsi que des pratiques pour surmonter ces obstacles. Nous présentons nos méthodes de pré-annotation et les types d'entités de notre schéma d'annotation (section 2), puis les résultats obtenus et les enseignements que nous en tirons (section 4).

2 Pré-annotation automatique des entités

Pour la pré-annotation automatique des entités, nous utilisons des méthodes traditionnelles de détection d'entités. Ces méthodes ont l'avantage d'être agiles, dans la mesure où il est facile et rapide de prendre en compte de nouvelles expressions. Nous les avons utilisées sur un échantillon de documents pendant la mise au point du schéma d'annotation. Leur capacité de généralisation reste néanmoins limitée, d'où l'intérêt de passer à des méthodes par apprentissage supervisé dans une seconde phase une fois le corpus entier annoté et corrigé manuellement.

1. Appariement exact de termes. Nous avons pour cela construit des lexiques à partir de terminologies médicales en français, notamment présentes dans le Metathesaurus de l'UMLS (Bodenreider, 2004), ou proposées par des agences françaises comme la BDPM [1] pour les traitements médicaux. Nous avons aussi créé des listes de termes ne provenant pas de thésaurus préexistants, par exemple dans le cas des entités de type Entourage pour détecter les termes du champ lexical de la famille.
2. Détection de 20 préfixes et 7 suffixes révélateurs. Par exemple, *postero-* et *antero-* suivis de termes ayant le type Anatomie ajoutent un degré de précision supplémentaire sur la localisation anatomique. Selon le contexte, les préfixes *hypo-* et *hyper-* peuvent indiquer des pathologies ou des signes ; le suffixe *-pathie* indique une pathologie, *-émie* indique un paramètre mesurable, et

1. https://base-donnees-publique.medicaments.gouv.fr/

le suffixe *-graphie* peut indiquer un examen d'imagerie. Des exceptions peuvent être ajoutées dans certains cas, par exemple le terme anémie désigne une pathologie et non un paramètre mesurable, malgré la présence du suffixe *-émie*.

3. Plus de 100 mots-clefs, par exemple *non*, *ni* ou *sans* indiquent généralement des négations, et *Hôpital* et *en* des lieux. Certains mots-clefs nécessitent un traitement supplémentaire pour comprendre le contexte de leur utilisation. Par exemple le mot *majoration* aura un sens différent lorsqu'il est utilisé avant un signe, auquel cas il s'agit d'une Dégradation de l'état du patient, ou avant un traitement médical, où il s'agira d'une Augmentation du traitement — ne signifiant pas nécessairement que l'état général du patient change.
4. Environ 20 mots-déclencheurs indiquent le type d'entité du texte qui suit. Par exemple, *Greffe de* ou *Thérapie* indiquent un traitement non médical, alors que *Maladie de* ou *Syndrome de* indiquent des pathologies, et *Tranfert en* ou *Orienté en* expriment des changements de lieu. La différence principale entre un mot-clef et un mot déclencheur est que le mot-clef se suffit généralement à lui même (le mot *non* suffit à indiquer une négation par exemple), alors que le mot déclencheur annonce d'autres mots qui vont venir compléter le terme. Par exemple *Syndrome* à lui tout seul ne donne pas d'information pertinente sur l'état du patient, alors que *Syndrome coronarien aigu* si.
5. 25 expressions régulières sont utiles pour traiter les diverses valeurs numériques comme les Dates ou les Valeurs composées de chiffres (par exemple : 42cm², 2020-06-12 etc.).

3 Types d'entités et évolutions

Nous décrivons maintenant les principaux types d'entités de notre schéma d'annotation. Ces types d'entités sont le résultat de discussions entre TAListes et cliniciens. La table 2 de l'annexe A donne la liste complète de ces types d'entités.

Signes et symptômes, Pathologies Les signes et symptômes sont utilisés pour apprécier l'état de santé du patient et déterminer ses pathologies. Dans notre projet, la pathologie clé est l'insuffisance cardiaque qui est responsable de symptômes tels que l'essoufflement ou la fatigue anormale et qui entraîne l'apparition de signes cliniques comme des crépitants. Les pathologies incluent également les comorbidités, qui peuvent être observées simultanément à l'insuffisance cardiaque et apporter un risque supplémentaire.

Initialement, ces entités formaient une seule classe. Après examen de l'annotation résultante, les cliniciens ont fait valoir l'intérêt de distinguer les trois sous-classes Signes, Symptômes, et Pathologies. La mise à jour des ressources de pré-annotation a montré la difficulté à distinguer clairement Signes et Symptômes, pointant des exemples où une même observation apparaissait comme l'un ou l'autre selon le point de vue. Cela a abouti a un accord sur deux classes : Signes et symptômes, et Pathologies.

Examens et traitements Les examens représentent les différents tests, procédures et analyses médicales effectuées sur le patient pour diagnostiquer ou évaluer sa condition médicale, tels que l'échocardiographie ou l'électrocardiogramme. Les traitements représentent les différentes options de prise en charge thérapeutique, comme les diurétiques ou les béta-bloquants, ainsi que les interventions chirurgicales ou la pose de dispositifs médicaux tels que les stents.

Le schéma initial prévoyait des paramètres mesurables comme le poids ou la FEVG (fraction d'éjection du ventricule gauche). La pré-annotation de ces entités a fait ressortir la non-annotation des examens eux-mêmes, comme l'ECG (électrocardiogramme). Après discussion, ils ont été ajoutés au schéma d'annotation et les ressources de pré-annotation ont été mises à jour.

Vie du patient Il s'agit d'un ensemble de variables qui décrivent les aspects socio-démographiques et comportementaux du patient, ainsi que son environnement social. Les comportements du patient comprennent des informations sur son mode de vie et ses habitudes, comme l'activité physique, le régime alimentaire, la consommation de tabac et d'alcool, etc. L'autonomie du patient fait référence à sa capacité à réaliser les activités quotidiennes de manière indépendante, ainsi qu'à ses besoins en matière de soutien et d'assistance. L'entourage du patient comprend des informations sur les personnes qui gravitent autour du patient, comme les membres de sa famille ou les aidants, et leur rôle dans la prise en charge du patient. Les entités Vie du patient peuvent fournir des informations importantes pour comprendre le contexte dans lequel se déroule la prise en charge de l'insuffisance cardiaque, et permettent d'évaluer les facteurs de risque et les déterminants sociaux de la santé qui peuvent influencer l'évolution de la maladie.

À ce stade, la pré-annotation reste incomplète et montre une difficulté particulière à cerner les éléments qui concourent à l'évaluation de l'autonomie et de l'entourage. Cela nous a amenés à démarrer un sous-projet spécifique pour mieux les déterminer.

Entités auxiliaires permettant d'annoter des informations, généralement pour préciser des entités de type Signe, Examen, Traitement, etc. :

- temporelles (date, âge du patient, durée d'un traitement...) ;
- localisation du patient (lieu de séjour) et ses éventuels changements (provenance, destination) ;
- localisation anatomique et valeurs précisant les signes ou traitements ;
- évolution ou absence de signes ou de traitements.

Ici encore, la pré-annotation nous a donné un support concret pour des allers-retours avec les experts du domaine. Cela a mené notamment à étendre le champ d'application des valeurs, initialement uniquement associées aux paramètres mesurables, aux signes et symptômes.

La plupart des types d'entités présentés ici se rencontrent dans les articles scientifiques du même domaine. Par exemple, des types d'entités tels que *pathologie*, *anatomie*, *paramètres mesurables*, *traitement (non) médical* et *examens* peuvent être pertinents pour l'annotation des articles scientifiques dans des domaines tels que la médecine et la biologie. Étant donné que la même expertise serait mobilisée, le travail présenté ici devrait être représentatif au moins d'une partie du travail qui devrait être fait pour l'extraction d'informations de textes scientifiques.

4 Résultats et enseignements tirés

Nous avons évalué la qualité de la pré-annotation sur 3 textes entièrement annotés par un expert du domaine. Nous avons pour cela utilisé l'outil brat-eval [2] qui évalue la correction des types et des frontières d'entités. Le mode d'évaluation utilisé pour les types est *EXACT* : brat-eval indique que le type est correct uniquement si le type d'entité annoté par l'expert et celui produit par le programme

2. https://github.com/READ-BioMed/brateval

sont identiques. Le mode de respect des frontières est *OVERLAP*, ce qui donne une flexibilité dans la délimitation des entités : par exemple, étant donné le terme *un carcinome canalaire*, si l'expert annote l'expression entière comme une Pathologie alors que le programme annote seulement *carcinome canalaire* comme Pathologie, brat-eval considère quand même que l'annotation est correcte. Les résultats figurent dans la table 1.

Type	Précision	Rappel	Nb	Type	Précision	Rappel	Nb
Pathologie	0,70	1,00	76	Autonomie	1,00	1,00	4
Signe_Symptome	0,87	0,68	111	Entourage	0,68	1,00	11
Gravité	0,43	0,29	35	Date	0,90	0,29	38
Hypothétique	0,80	0,67	8	Fréquence	1,00	0,94	34
Examen	0,94	0,88	26	Âge	1,00	1,00	7
Param_mesurable	0,89	0,78	78	Lieu	1,00	0,57	22
Param_physiologique	1,00	1,00	3	Anatomie	0,58	0,74	86
Trt_non_médical	0,91	0,83	25	Valeur	0,81	0,75	151
Trt_médical	1,00	0,77	107	Négation	0,57	0,87	76
Comportement	1,00	1,00	6	Évolution	1,00	0,31	46
Heure	0,63	1,00	7	Mode	1,00	0,86	14
Concentration	0,94	0,76	42	Chgt_lieu	0,50	0,55	11

TABLE 1 – Évaluation de la pré-annotation selon brat-eval, en précisant le mode d'évaluation *EXACT* et le mode de respect des frontières *OVERLAP*.

Le travail de pré-annotation a servi de support pédagogique auprès des experts non informaticiens. En effet, certains concepts de traitement automatique des langues nécessitent du temps pour être assimilés correctement, tandis que certains choix peuvent paraître obscurs et nécessiter des explications. Montrer une annotation automatique sur l'ensemble des textes en cours d'examen a permis d'illustrer les points discutés et de faire réagir les cliniciens sur l'état courant de la modélisation. Cela s'est avéré utile lors des échanges réguliers avec les cliniciens pour mieux comprendre leur point de vue sur la manière dont certaines choses sont annotées.

Nous avons constaté que la pré-annotation a été globalement bien reçue par les cliniciens, qui ont compris très clairement l'intérêt de ce travail. Les avantages en termes de gain de temps et de réduction de la charge mentale ont été ressentis. Les résultats de la table 1 montrent capacité de la pré-annotation à produire des résultats de qualité suffisamment proche de l'annotation manuelle des experts pour que l'on puisse estimer que leur temps de correction reste inférieur à une annotation manuelle ex nihilo.

La pré-annotation a également été un outil précieux pour tester et valider le contenu du schéma d'annotation en cours de développement. Les discussions autour des choix de pré-annotation ont permis d'identifier des incohérences ou des manques dans le schéma d'annotation initial, qui ont été corrigés en conséquence. De plus, la pré-annotation a permis de mettre à l'épreuve le schéma d'annotation en identifiant les cas d'utilisation courants et les cas plus complexes, ce qui a permis de renforcer et d'affiner ce schéma de manière itérative. Nous notons enfin que des méthodes et outils ont été proposés pour opérationnaliser les opérations de pré-annotation que nous avons décrites plus haut (Lison *et al.*, 2021), que nous envisageons d'utiliser dans le futur.

Remerciements

Ce travail a été soutenu par l'Agence Nationale pour la Recherche (ANR) dans le cadre du projet PREDHIC (ANR-21-CE23-0039).

Références

BODENREIDER O. (2004). The Unified Medical Language System (UMLS) : Integrating biomedical terminology. *Nucleic Acids Research*, **32**(Database issue), D267–270.

LINGREN T., DELEGER L., MOLNAR K., ZHAI H., MEINZEN-DERR J., KAISER M., STOUTENBOROUGH L., LI Q. & SOLTI I. (2012). Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development : Evaluating the impact on annotation speed and potential bias. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, p. 108–108. DOI : 10.1109/HISB.2012.33.

LISON P., BARNES J. & HUBIN A. (2021). skweak : Weak supervision made easy for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 337–346, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-demo.40.

PATEL P., DAVEY D., PANCHAL V. & PATHAK P. (2018). Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2033–2042, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1228.

SHINOHARA E., SHIBATA D. & KAWAZOE Y. (2022). Development of comprehensive annotation criteria for patients' states from clinical texts. *J Biomed Inform*, **134**, 104200.

WORLD HEALTH ORGANIZATION (2020). Leading causes of death and disability — a visual summary of global and regional trends 2000-2019. consulté le 31/3/2023.

A Types d'entités

Type	Description	Type	Description
Pathologie	Maladie responsable de symptômes et signes cliniques, y compris comorbidité.	Signe et Symptôme	Signe : manifestation d'une maladie, constatée par un observateur. Symptôme : signe dont le malade se plaint.
Gravité	Sévérité d'un signe ou d'une pathologie.	Hypothétique	Incertitude sur la présence d'un signe ou d'une pathologie.
Examen	Examen (d'imagerie, etc.).	Paramètre mesurable	Variable mesurée par un examen.
Paramètre physiologique	Variable désignant une fonction normale (physiologique) du patient.	Traitement médical	Traitement médicamenteux, typiquement nom de médicament.
Traitement non médical	Traitement autre que médicament.	Mode d'administration	Mode d'administration d'un médicament ou lié à un paramètre mesurable.
Comportement	Comportement du patient.	Autonomie	Information sur le degré d'autonomie du patient.
Entourage	Information sur la présence de famille, etc. dans l'entourage du patient.	Date	Date d'un événement, absolue (chiffrée) ou relative (non chiffrée).
Heure	Heure d'un événement, absolue (chiffrée) ou relative (non chiffrée).	Durée	Durée d'un événement, absolue ou relative.
Fréquence	Fréquence d'un traitement ou d'un comportement.	Âge	Âge du patient.
Lieu	Lieux de provenance, de séjour et de sortie du patient.	Changement de lieu	Changements de lieux qui pourraient indiquer une évolution de l'état du patient.
Anatomie	Parties du corps.	Valeur	Valeur qualitative (textuelle) ou quantitative (numérique) d'une entité.
Concentration	Concentration d'un médicament.	Négation	Absence d'entité.
Évolution	Évolution d'un signe, un traitement ou de l'état général du patient.	Dose	Dosage d'un traitement médical.

TABLE 2 – Liste des types d'entités

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

MaTOS: traduction automatique pour la science ouverte

Maud Bénard[1] Natalie Kübler[1] Alexandra Mestivier[1] Lichao Zhu[1]
Rachel Bawden[2] Eric de la Clergerie[2] Laurent Romary[2]
Mathilde Huguin[3] Jean-Francois Nominé[3] Ziqian Peng[4] François Yvon[4]
(1) CLLILAC-ARP, Université Paris Cité, Paris, France
(2) Centre Inria Paris, France
(3) INIST, CNRS, Nancy, France
(4) ISIR, CNRS et Sorbonne Université, Paris, France

RÉSUMÉ

Cette contribution présente le projet MaTOS (Machine Translation for Open Science), qui vise à développer de nouvelles méthodes pour la traduction automatique (TA) intégrale de documents scientifiques entre le français et l'anglais, ainsi que des métriques automatiques pour évaluer la qualité des traductions produites. Pour ce faire, MaTOS s'intéresse (a) au recueil de ressources ouvertes pour la TA spécialisée ; (b) à la description des marqueurs de cohérence textuelle pour les articles scientifiques ; (c) au développement de nouvelles méthodes de traitement multilingue pour les documents ; (d) aux métriques mesurant les progrès de la traduction de documents complets.

ABSTRACT

MaTOS : Machine Translation for Open Science.

This contribution presents the MaTOS (Machine Translation for Open Science) project, which aims to develop new methods for the complete machine translation (MT) of scientific documents between English and French, as well as automatic metrics to evaluate the translation quality. To this end, MaTOS is interested in (a) the collection of open resources for specialised MT ; (b) the description of textual coherence markers for scientific articles ; (c) the development of new multilingual processing methods for documents ; and (d) metrics to measure progress in document-level machine translation.

MOTS-CLÉS : Traduction Automatique de Documents ; Analyse de Documents Scientifiques.

KEYWORDS: Document-Level Machine Translation ; Analysis of Scholarly Documents.

1 Introduction

Avec l'avènement d'architectures neuronales attentionnelles encodeur-décodeur (Cho *et al.*, 2014; Bahdanau *et al.*, 2015; Vaswani *et al.*, 2017) les systèmes de traduction automatique ont réalisé des progrès considérables, offrant des services de plus en plus utiles pour une vaste gamme d'utilisateurs et d'applications. La recherche continue de progresser à vive allure pour prendre en compte davantage de domaines et de paires de langues, en particulier en développant des méthodes massivement multilingues (Freitag & Firat, 2020; Fan *et al.*, 2021). L'essentiel de cet effort s'inscrit dans un cadre de traduction de segments isolés et évalue les progrès réalisés à l'aune de métriques simplistes, au premier rang desquelles BLEU (Papineni *et al.*, 2002) et METEOR (Banerjee & Lavie, 2005).

Par contraste, le projet MaTOS (*Machine Translation for Open Science*) s'intéresse à la traduction

automatique (TA) de documents scientifiques complets, qui pose des questions difficiles induites par les dépendances entre les segments ou entre parties d'un même document. Ces dépendances peuvent être locales, comme les co-références pronominales (Bawden *et al.*, 2018) et les ellipses (Voita *et al.*, 2019), ou plus globales, reflétant des contraintes de cohérence lexicale et terminologique (Pu *et al.*, 2017) ou de structuration du discours (Guzmán *et al.*, 2014). Dans les sections qui suivent, nous présentons en détail le contexte scientifique et les principaux objectifs du projet.

2 Contexte général

Le projet MaTOS s'inscrit dans un mouvement qui vise à développer le multilinguisme dans la communication scientifique et dont les objectifs principaux sont résumés dans l'initiative d'Helsinki [1]. En encourageant la production et la diffusion des savoirs scientifiques dans d'autres langues que l'anglais, cette initiative vise à gommer un certain nombre de biais attribuables au monolinguisme qui s'est progressivement imposé dans de nombreux champs disciplinaires (Gordin, 2015).

Une motivation seconde pour favoriser la diversité linguistique correspond à un objectif de diffusion des connaissances vers le plus large public : rendre les sources scientifiques accessibles dans un plus grand nombre de langues facilite leur circulation au-delà des cercles disposant de l'expertise scientifique : étudiants, journalistes, ingénieurs, décideurs politiques. Ce besoin a été mis en évidence durant la crise du Covid de 2020 et est rappelé dans le 2e plan national pour la science ouverte, qui souligne l'apport que la TA pourrait avoir pour y répondre (Fiorini *et al.*, 2020).

Les questions de la traduction automatique pour les documents ont fait l'objet de plusieurs projets récents, par exemple le projet ANR/COSMAT (Lambert *et al.*, 2012) ou encore le projet Européen *Health in My Language* [2], focalisé, comme la campagne d'évaluation WMT-Biomedical [3], sur la traduction de textes en domaine biomédical. Mentionnons également l'initiative ACL60-60 [4] qui développera des services de traduction d'articles et de sous-titrage d'exposés pour le Traitement Automatique des Langues (TAL).

3 Détail du programme scientifique

3.1 Ressources pour la traduction automatique

La première tâche du projet est de rassembler les ressources nécessaires à sa bonne exécution, en premier lieu les corpus et les données terminologiques disponibles pour trois domaines : les sciences biomédicales, le TAL et les sciences de la Terre. Le premier de ces domaines a déjà fait l'objet de nombreuses études en recherche et extraction d'information. Il existe également des terminologies bilingues [5], de grands corpus parallèles (développés dans le cadre des challenges WMT-Biomedical), y compris des corpus de résumés parallèles extraits de l'archive PubMed.

Les deux autres domaines sont comparativement moins bien pourvus, même si l'initiative ACL60-60

1. https://www.helsinki-initiative.org/fr
2. https://www.himl.eu/
3. Voir https://www.statmt.org/wmt22/biomedical-translation-task.html/ pour l'édition 2022.
4. Initiative promue par l'*Assocation for Computational Linguistics* : voir https://www.acl6060.org/
5. http://mesh.inserm.fr/FrenchMesh/index.htm

est susceptible de changer la situation pour le TAL. [6] Pour remédier à ce manque de ressources, nous comptons d'une part exploiter les données parallèles existantes dans les archives scientifiques en appliquant les méthodes du projet SciPar (Roussis *et al.*, 2022) à des archives telles que ISTEX [7] et HAL [8]. Nous comptons, d'autre part, utiliser des méthodes d'augmentation artificielle de données exploitant les grands corpus de textes monolingues qui existent en langue anglaise, et dans une moindre mesure, en langue française. Pour ce qui concerne la terminologie associée à ces domaines, MaTOS prévoit de combiner des méthodes automatiques de repérage de termes avec les ressources existantes telles que les bases Loterre [9] de l'INIST et base ARTES [10] (Pecman & Kübler, 2012).

3.2 Questions terminologiques

Une première difficulté de la traduction scientifique, ou plus généralement de la traduction en domaine de spécialité, porte sur la traduction correcte des termes du domaine. En exploitant des listes de termes bilingues, il était relativement simple pour les méthodes de TA statistiques de contraindre la traduction d'un terme connu. Le portage de ces techniques dans un cadre de TA neuronale, qui n'explicite pas les alignements de mots, n'est pas trivial. Il est toutefois possible d'aboutir à ce résultat soit en imposant des contraintes lexicales dans l'algorithme de décodage (Hokamp & Liu, 2017; Post & Vilar, 2018), soit en positionnant directement les termes dans la phrase source par des prétraitements (Dinu *et al.*, 2019; Susanto *et al.*, 2020; Xu & Carpuat, 2021).

Dans le cadre de MaTOS, l'objectif est, d'une part, d'aller plus loin dans la traduction des termes connus en s'intéressant à la prise en compte de la variation terminologique (Daille, 2017) au sein d'un document, en relation avec les objectifs pragmatiques associés à la selection de chaque variante (forme complète, compactée, étendue, siglée, etc). Il faudra pour cela, développer des outils de reconnaissance et de génération de variantes des termes et progresser dans leur caractérisation par des études en corpus. Nous souhaitons, d'autre part, nous intéresser également à la traduction *de termes inconnus*. Ce problème est relativement nouveau et nous envisageons deux scénarios : la production d'un terme cible à partir d'une définition en langue cible ou directement depuis la définition source.

3.3 Traduire des documents complets

Du point de vue formel, la TA au niveau document demande de mettre en œuvre un modèle de la forme [1] $P(\mathbf{e}_1 \dots \mathbf{e}_{l_d} | \mathbf{f}_1 \dots \mathbf{f}_{l_d}; \boldsymbol{\theta})$ avec $\mathbf{e}_t$ (resp. $\mathbf{f}_t$) les phrases cibles (resp. sources), l_d la longueur du document et $\boldsymbol{\theta}$ les paramètres. La plupart des architectures neuronales étudiées par (Maruf *et al.*, 2021), à l'exception peut-être de (Yu *et al.*, 2020), se limitent à des modèles de la forme $P(\mathbf{e}_t | \mathbf{f}_t, \mathbf{C}_t; \boldsymbol{\theta})$ avec $\mathbf{C}_t$ une représentation condensée du contexte discursif source et cible. Entrent dans ce cadre aussi bien les architectures par concaténation de (Scherrer *et al.*, 2019; Ma *et al.*, 2020) que les architectures multi-encodeurs de (Miculicich *et al.*, 2018; Bawden *et al.*, 2018; Li *et al.*, 2020). Ces propositions sont comparées par (Lopes *et al.*, 2020; Ma *et al.*, 2021).

L'alternative que nous souhaitons explorer est l'implémentation directe de [1]. Elle se heurte à

6. Des premières données, correspondant à des traductions d'exposés vers 10 langues, dont le français, sont déjà distribuées dans le cadre de la campagne IWSLT 2023 (https://iwslt.org/2023/multilingual).
7. https://www.inist.fr/services/analyser/istex-textes-corpus/
8. https://hal.science/
9. https://www.loterre.fr/
10. https://artes.app.univ-paris-diderot.fr/artes-symfony/web/app.php

la complexité algorithmique des calculs effectués au sein des modèles Transformers. Pour ces architectures, le calcul de l'attention est quadratique en temps et en espace en fonction de la longueur totale du texte ou du document en entrée. Cette difficulté a conduit à de multiples propositions (Tay *et al.*, 2022), qui soit simplifient le calcul de l'attention (Child *et al.*, 2019; Zaheer *et al.*, 2020; Beltagy *et al.*, 2020), soit remplacent le composant attentionnel par des alternatives récurrentes plus efficaces (Gu *et al.*, 2022). En dépit de ces efforts, la taille des séquences utilisées dans l'état de l'art se limite à quelques milliers de mots, ce qui est encore trop peu pour encoder un document complet.

L'objectif de MaTOS est d'étendre ces méthodes à des documents complets et d'étudier les améliorations en terme de cohérence qui en découlent. Nous envisageons également de nous pencher sur plusieurs problèmes essentiels pour notre application, à savoir la traduction (ou la localisation) des éléments non-textuels : titres, légendes, en-têtes de tableaux, formules et équations, appels de référence, etc. (Zhu *et al.*, 2021).

3.4 Mesurer les progrès

Pour progresser sur les questions évoquées ci-dessus, un préalable est de disposer des métriques idoines, car les métriques automatiques standard ne permettent pas de rendre compte des améliorations de la modélisation des questions terminologiques et discursives (Popescu-Belis, 2019). Face à ce constat, diverses métriques ont été proposées : pour la terminologie, citons en particulier (ibn Alam *et al.*, 2021), qui évalue grossièrement le degré d'accord entre les traductions de termes générées et celles qui figurent dans une liste de référence. Pour ce qui concerne les phénomènes discursifs locaux, deux approches sont communément employées. D'une part les évaluations *focalisées*, qui n'évaluent que la correction de quelques mots spécifiques (par exemple, les pronoms (Guillou & Hardmeier, 2016)), d'autre part les évaluations *contrastives*, qui comparent les scores attribués respectivement à des traductions qui ne diffèrent que dans leur traitement d'un mot ou d'un phénomène problématique (le genre d'un pronom, le temps d'un verbe, etc) (Sennrich, 2017; Bawden *et al.*, 2018). Ces méthodes s'appliquent également pour évaluer la cohérence des choix lexicaux. Pour l'évaluation du niveau global, les propositions sont très lacunaires et s'appuient principalement sur le repérage d'éléments de cohésion : chaines de co-références, similarités entre fragments textuels (Jiang *et al.*, 2022).

Dans le cadre de MaTOS, trois pistes seront explorées : la première essaiera de dépasser les apories des évaluations des choix terminologiques en s'intéressant à la question de la variation. Cette partie de l'étude reposera sur la catégorisation manuelle d'erreurs terminologiques, puis à leur repérage automatique. Une seconde piste d'exploration concerne l'évaluation globale du document, qui sera analysée du point de vue de sa cohérence/cohésion, en exploitant ici encore des modèles neuronaux à large contexte (Deng *et al.*, 2022; Abhishek *et al.*, 2022).

En complément de l'étude de méthodes automatiques, nous envisageons de réaliser une évaluation humaine participative, en incitant la communauté des utilisateurs de l'archive HAL à traduire (par post-édition) des résumés en français ou en anglais de leur propre production, afin que les deux langues soient toujours disponibles dans les méta-données qui décrivent un document. Cette évaluation vise plusieurs objectifs. Il s'agit, tout d'abord, de déterminer le niveau de qualité qui serait acceptable par les utilisateurs finaux en matière de traduction de documents scientifiques. Il s'agit également d'accroître la quantité de données parallèles disponibles pour mener à bien les autres expériences de notre projet. Un première expérience pilote sera présentée durant la conférence TALN 2023.

4 Conclusion

Le projet MaTOS s'inscrit fermement dans le mouvement de la science ouverte et la promotion du multilinguisme dans la communication scientifique. Nous visons à développer de nouvelles méthodes pour la traduction automatique intégrale, anglais-français, de documents scientifiques, ainsi que des métriques automatiques pour évaluer la qualité des traductions produites. Notre approche repose notamment sur l'étude des variations terminologiques et la description des marqueurs de cohérence textuelle pour les articles scientifiques.

5 Remerciements

Ce projet a reçu un soutien de l'Agence Nationale de la Recherche (convention ANR-22-CE23-0033).

Références

ABHISHEK T., RAWAT D., GUPTA M. & VARMA V. (2022). Transformer models for text coherence assessment.

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the first International Conference on Learning Representations*, San Diego, CA.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, p. 65–72, Ann Arbor, Michigan.

BAWDEN R., SENNRICH R., BIRCH A. & HADDOW B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1304–1313, New Orleans, Louisiana. DOI : 10.18653/v1/N18-1118.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The long-document transformer.

CHILD R., GRAY S., RADFORD A. & SUTSKEVER I. (2019). Generating long sequences with sparse transformers.

CHO K., VAN MERRIENBOER B., BAHDANAU D. & BENGIO Y. (2014). On the properties of neural machine translation : Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, p. 103–111, Doha, Qatar.

DAILLE B. (2017). *Term variation in specialised corpora : characterisation, automatic discovery and applications.*, volume 19 de *Terminology and Lexicography Research and Practice series*. John Benjamins Publishing Company. DOI : 10.1075/tlrp.19, HAL : hal-01693035.

DENG Y., KULESHOV V. & RUSH A. (2022). Model criticism for long-form text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11887–11912, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

DINU G., MATHUR P., FEDERICO M. & AL-ONAIZAN Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association*

for Computational Linguistics, p. 3063–3068, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1294.

FAN A., BHOSALE S., SCHWENK H., MA Z., EL-KISHKY A., GOYAL S., BAINES M., CELEBI O., WENZEK G., CHAUDHARY V., GOYAL N., BIRCH T., LIPTCHINSKY V., EDUNOV S., AULI M. & JOULIN A. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, **22**(107), 1–48.

FIORINI S., BARBIN F., GARNIER-RIZET M., MORIN K. H., HUMPHREYS F., JOSSELIN-LERAY A., KÜBLER N., LOOCK R., MARTIKAINEN H., NOMINÉ J.-F., PLAG C., ROSSI C. & YVON F. (2020). *Rapport du groupe de travail "Traductions et science ouverte"*. Technical report, Comité pour la science ouverte. DOI : 10.52949/20, HAL : hal-03640511.

FREITAG M. & FIRAT O. (2020). Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, p. 550–560, Online.

GORDIN M. D. (2015). *Scientific Babel How Science Was Done Before and After Global English Scientific Babel : How Science Was Done Before and After Global English*. University of Chicago Press.

GU A., GOEL K. & RÉ C. (2022). Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* : OpenReview.net.

GUILLOU L. & HARDMEIER C. (2016). PROTEST : A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 636–643, Portorož, Slovenia : European Language Resources Association (ELRA).

GUZMÁN F., JOTY S., MÀRQUEZ L. & NAKOV P. (2014). Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 687–698, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/P14-1065.

HOKAMP C. & LIU Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1535–1546 : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1141.

IBN ALAM M. M., ANASTASOPOULOS A., BESACIER L., CROSS J., GALLÉ M., KOEHN P. & NIKOULINA V. (2021). On the evaluation of machine translation for terminology consistency.

JIANG Y., LIU T., MA S., ZHANG D., YANG J., HUANG H., SENNRICH R., COTTERELL R., SACHAN M. & ZHOU M. (2022). BlonDe : An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1550–1565, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.111.

LAMBERT P., SENELLART J., ROMARY L., SCHWENK H., ZIPSER F., LOPEZ P. & BLAIN F. (2012). Collaborative machine translation service for scientific texts. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 11–15, Avignon, France : Association for Computational Linguistics.

LI B., LIU H., WANG Z., JIANG Y., XIAO T., ZHU J., LIU T. & LI C. (2020). Does multi-encoder help ? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3512–3518, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.322.

LOPES A., FARAJIAN M. A., BAWDEN R., ZHANG M. & MARTINS A. (2020). Document-level neural MT : A systematic comparison. In *22nd Annual Conference of the European Association for Machine Translation*, p. 225–234.

MA S., ZHANG D. & ZHOU M. (2020). A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3505–3511, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.321.

MA Z., EDUNOV S. & AULI M. (2021). A comparison of approaches to document-level machine translation. arXiv preprint arXiv :1910.07481.

MARUF S., SALEH F. & HAFFARI G. (2021). A survey on document-level neural machine translation : Methods and evaluation. *ACM Comput. Surv.*, **54**(2). DOI : 10.1145/3441691.

MICULICICH L., RAM D., PAPPAS N. & HENDERSON J. (2018). Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2947–2954, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1325.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PECMAN M. & KÜBLER N. (2012). The ARTES bilingual LSP dictionary : From collocation to higher order phraseology. In *Electronic Lexicography* : Oxford University Press. DOI : 10.1093/acprof :oso/9780199654864.003.0010.

POPESCU-BELIS A. (2019). Context in neural machine translation : A review of models and evaluations.

POST M. & VILAR D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1314–1324, New Orleans, Louisiana : Association for Computational Linguistics.

PU X., MASCARELL L. & POPESCU-BELIS A. (2017). Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 948–957, Valencia, Spain : Association for Computational Linguistics.

ROUSSIS D., PAPAVASSILIOU V., PROKOPIDIS P., PIPERIDIS S. & KATSOUROS V. (2022). SciPar : A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2652–2657, Marseille, France : European Language Resources Association.

SCHERRER Y., TIEDEMANN J. & LOÁICIGA S. (2019). Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, p. 51–61, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-6506.

SENNRICH R. (2017). How grammatical is character-level neural machine translation ? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 376–382, Valencia, Spain : Association for Computational Linguistics.

SUSANTO R. H., CHOLLAMPATT S. & TAN L. (2020). Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3536–3543, Online : Association for Computational Linguistics.

TAY Y., DEHGHANI M., BAHRI D. & METZLER D. (2022). Efficient transformers : A survey. *ACM Comput. Surv.* Just Accepted, DOI : 10.1145/3530811.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30*, p. 5998–6008 : Curran Associates, Inc.

VOITA E., SENNRICH R. & TITOV I. (2019). When a good translation is wrong in context : Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1198–1212, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1116.

XU W. & CARPUAT M. (2021). Rule-based morphological inflection improves neural terminology translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 5902–5914, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.477.

YU L., SARTRAN L., STOKOWIEC W., LING W., KONG L., BLUNSOM P. & DYER C. (2020). Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, **8**, 346–360. DOI : 10.1162/tacl_a_00319.

ZAHEER M., GURUGANESH G., DUBEY A., AINSLIE J., ALBERTI C., ONTANON S., PHAM P., RAVULA A., WANG Q., YANG L. & AHMED A. (2020). Big Bird : Transformers for longer sequences.

ZHU K., GAO Y., GUO J. & LOU J.-G. (2021). Translating headers of tabular data : A pilot study of schema translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 56–66, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.5.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Projet NaviTerm : navigation terminologique pour une montée en compétence rapide et personnalisée sur un domaine de recherche

Florian Boudin Richard Dufour Béatrice Daille
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
prenom.nom@univ-nantes.fr

RÉSUMÉ

Cet article présente le projet NaviTerm dont l'objectif est d'accélérer la montée en compétence des chercheurs sur un domaine de recherche par la création automatique de représentations terminologiques synthétiques et navigables des connaissances scientifiques.

ABSTRACT

The NaviTerm project : terminological navigation for assisting researchers in gaining expertise on a new research domain.

This article presents the NaviTerm project, whose objective is to assist scholars in gaining new skills in a field of research by automatically producing concise and navigable terminological representations of scientific knowledge.

MOTS-CLÉS : Ici une liste de mots-clés en français.

KEYWORDS: Here a list of keywords in English.

1 Introduction

L'explosion de la production scientifique mondiale amène les chercheurs à revoir en profondeur leur démarche de veille scientifique. Malgré l'émergence de moteurs de recherche dédiés (e.g. Google Scholar, Semantic Scholar, PubMed), parcourir la littérature pour monter en compétence sur un domaine de recherche est de plus en plus laborieux et chronophage. À cela s'ajoute l'opacité des algorithmes utilisés par ces moteurs de recherche qui impose de s'interroger sur la pertinence des résultats retournés (Martín-Martín *et al.*, 2018). Faciliter l'accès aux nouvelles connaissances scientifiques de manière efficace et transparente est donc, plus que jamais, un enjeu majeur pour la recherche scientifique. Cet article présente le projet NaviTerm, financé dans le cadre d'un accord entre le CNRS et l'Agence de l'Innovation et de la Défense (AID), qui apporte une réponse à cet enjeu sous l'angle de la terminologie et de la navigation documentaire.

Plus précisément, le projet NaviTerm porte sur le développement de méthodes automatisées pour extraire, structurer et ordonner par importance les termes d'une collection d'articles scientifiques relevant d'un domaine de recherche. Ces termes dressent une cartographie des connaissances scientifiques et constituent une interface d'accès naturel efficace au contenu des articles. Associés à une méthodologie de recherche à facettes ou à une interface de navigation, ils offrent un moyen intuitif et rapide de repérer les articles clés d'un domaine. Pour illustrer ce point, un exemple de navigation par termes-clés pour l'accès aux connaissances scientifiques est présenté dans la Figure 1. Cet exemple

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 16–20.

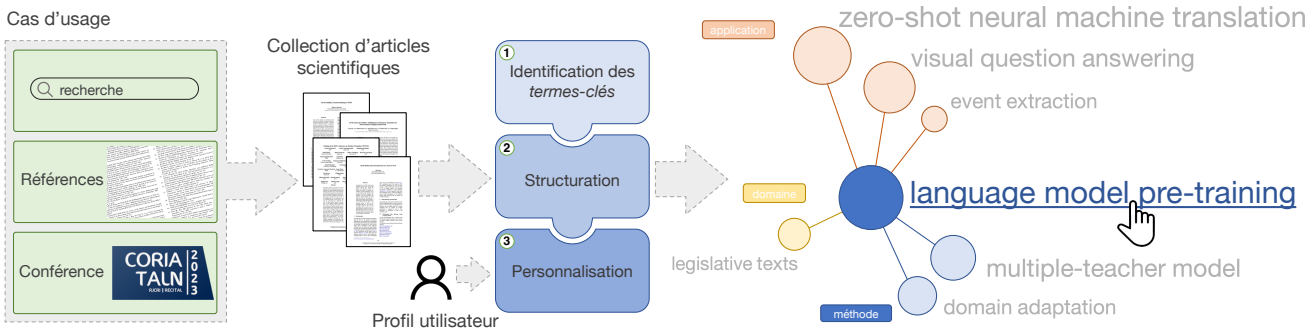


FIGURE 1 – Illustration du processus de création de représentation terminologique navigable à partir d'une collection d'articles scientifiques. Trois cas d'usage sont présentés : monter en compétence à partir des résultats d'une recherche, à partir des références citées dans un article, ou à partir des actes d'une conférence.

permet également de mettre en avant un élément important du projet NaviTerm qui est la structuration automatique des termes, ici en fonction des catégories [application , domaine , methode]. L'objectif de cette structuration est de faciliter la navigation dans la collection d'articles scientifiques et de proposer des listes filtrées et priorisées des articles clés.

L'objectif central du projet NaviTerm est le développement et la mise à disposition de nouveaux outils de recherche bibliographique qui intègrent les dernières avancées du TAL. La montée en compétence scientifique et technique est une problématique transversale et interdisciplinaire, dont les applications potentielles dans l'académique et l'industrie sont nombreuses. Le projet NaviTerm se distingue de l'existant sur deux aspects : d'abord d'un point de vue méthodologique en s'éloignant du paradigme dominant de l'apprentissage automatique supervisé de bout-en-bout au profit d'approches faiblement ou non supervisées de l'état-de-l'art, associées à des algorithmes facilement interprétables ; ensuite d'un point de vue applicatif avec la personnalisation des représentations terminologiques et des résultats de recherche pour accélérer la montée en compétences des utilisateurs. La section suivante rappelle les objectifs du projet et présente les trois verrous scientifiques que nous chercherons à lever.

2 Verrous scientifiques et solutions envisagées

Le projet NaviTerm s'attaque à la problématique de plus en plus prégnante de l'accès rapide aux connaissances scientifiques dans le contexte actuel d'explosion du nombre des publications. Ses objectifs sont doubles : 1) la création automatisée de représentations terminologiques navigables des connaissances scientifiques d'un domaine ; et 2) la personnalisation des représentations produites pour accélérer et maîtriser la montée en compétence des utilisateurs. Pour cela, nous tenterons de lever les trois verrous scientifiques suivants :

1. **Identifier les termes « clés » d'une collection de documents** ; les méthodes d'extraction terminologique existantes permettent d'identifier avec précision les termes spécialisés relevant d'un domaine mais n'évaluent pas leur importance vis-à-vis de ce dernier (Hazem *et al.*, 2020). Pour cela, nous proposons d'étendre ces méthodes avec des techniques non supervisées d'ordonnancement de texte utilisées pour l'extraction de mots-clés (Mihalcea & Tarau, 2004; Bougouin *et al.*, 2013). Il s'agira d'explorer comment ces techniques peuvent être étendues du

niveau de granularité du document à celui de la collection et d'appréhender les problématiques qui en découlent (e.g. passage à l'échelle, évaluation). Nous motivons ce choix méthodologique par une volonté de transparence et de confiance dans les algorithmes utilisés pour le projet.

2. **Structurer les connaissances terminologiques** ; les termes « clés » retenus devront être catégorisés (e.g. application, méthode, ensemble de données), référencés (i.e. identifiant (DOI) et position(s) d'occurrence dans l'article) et structurés (e.g. associer un ensemble de données avec une application). La solution envisagée pour cela est d'explorer comment les récentes méthodes de co-apprentissage (joint learning) fondées sur des réseaux convolutifs sur graphes (Sun *et al.*, 2019) peuvent être adaptées dans un cadre faiblement supervisé pour la détection de relations terminologiques entre termes « clés ». Dans un second temps, il s'agira d'étudier comment inférer les catégories et les relations entre les termes pour permettre la transférabilité à d'autres domaines et types de données. A cet égard, les travaux actuels sur l'extraction non-supervisée de relations (Tran *et al.*, 2020; Yuan & Eldardiry, 2021) constituent une piste de recherche intéressante.

3. **Personnaliser les connaissances d'un domaine** ; les représentations terminologiques construites à partir des termes « clés » retenus devront être adaptées pour permettre une montée en compétences plus efficace et rapide. Nous explorerons des stratégies de filtrage (ou de tri) pour mettre en exergue les connaissances importantes selon différents prismes : par rapport à un profil utilisateur (e.g. une bibliographie personnelle) ; par rapport à un événement scientifique (e.g. une conférence du domaine), par rapport à une temporalité (e.g. 5 dernières années) ; par rapport à la saillance d'un événement (e.g. information récente qui prend de l'ampleur, approche récurrente dans les documents). Pour cela, nous envisageons d'adapter le fonctionnement des méthodes neuronales de recommandation de tags (Hassan *et al.*, 2018) à notre problématique, puis dans une forme très exploratoire d'étudier comment personnaliser l'interface de navigation avec comme point de départ les travaux sur les interfaces de navigation exploratoire par mots-clés (Shukla & Hoeber, 2021) et les interfaces de visualisations selon une échelle temporelle, à la manière de graphes dynamiques.

3 Discussion et perspectives

Monter en compétence sur un domaine de recherche est une tâche de plus en plus complexe, la faute à un volume de littérature scientifique en pleine croissance. Accéder aux articles scientifiques les plus pertinents d'un domaine de recherche suppose une connaissance préalable de sa terminologie, ce qui n'est à l'évidence pas le cas des chercheurs qui souhaitent se former. L'objectif du projet NaviTerm est de lever cette difficulté en offrant un accès direct aux articles scientifiques au travers de représentations terminologiques construites automatiquement. L'accès au contenu des articles ne constitue évidemment qu'une première étape dans le chemin vers la montée en compétence et soulève de nouvelles questions de recherche. Par exemple, comment faciliter et accélérer la lecture des articles scientifiques (Head *et al.*, 2021; Fok *et al.*, 2022) ou comment construire un parcours pour acquérir des connaissances sont autant de questions qu'il conviendra d'examiner à l'avenir.

Remerciements

Ce travail est financé dans cadre du projet AID-CNRS NaviTerm (convention 2022 65 0079 CNRS Occitanie Ouest).

Références

BOUGOUIN A., BOUDIN F. & DAILLE B. (2013). TopicRank : Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 543–551, Nagoya, Japan : Asian Federation of Natural Language Processing.

FOK R., KAMBHAMETTU H., SOLDAINI L., BRAGG J., LO K., HEARST M. A., HEAD A. & WELD D. S. (2022). Scim : Intelligent skimming support for scientific papers. *arXiv preprint arXiv :2205.04561*.

HASSAN H. A. M., SANSONETTI G., GASPARETTI F. & MICARELLI A. (2018). Semantic-based tag recommendation in scientific bookmarking systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, p. 465–469, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3240323.3240409.

HAZEM A., BOUHANDI M., BOUDIN F. & DAILLE B. (2020). TermEval 2020 : TALN-LS2N system for automatic term extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*, p. 95–100, Marseille, France : European Language Resources Association.

HEAD A., LO K., KANG D., FOK R., SKJONSBERG S., WELD D. S. & HEARST M. A. (2021). Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, p. 1–18.

MARTÍN-MARTÍN A., ORDUNA-MALEA E., THELWALL M. & DELGADO LÓPEZ-CÓZAR E. (2018). Google scholar, web of science, and scopus : A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, **12**(4), 1160–1177. DOI : https ://doi.org/10.1016/j.joi.2018.09.002.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

SHUKLA S. & HOEBER O. (2021). Visually linked keywords to support exploratory browsing. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, p. 273–277, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3406522.3446037.

SUN C., GONG Y., WU Y., GONG M., JIANG D., LAN M., SUN S. & DUAN N. (2019). Joint type inference on entities and relations via graph convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1361–1370, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1131.

TRAN T. T., LE P. & ANANIADOU S. (2020). Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7498–7505, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.669.

YUAN C. & ELDARDIRY H. (2021). Unsupervised relation extraction : A variational autoencoder approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 1929–1938, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.147.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Annotation d'interactions hôte-microbiote dans des articles scientifiques par similarité sémantique avec une ontologie

Oumaima El Khettari[1] Solen Quiniou[1], Samuel Chaffron[1]
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000
oumaima.el-khettari@ls2n.fr, solen.quiniou@ls2n.fr,
samuel.chaffron@ls2n.fr

RÉSUMÉ

Nous nous intéressons à l'extraction de relations, dans des articles scientifiques, portant sur le microbiome humain. Afin de construire un corpus annoté, nous avons évalué l'utilisation de l'ontologie OHMI pour détecter les relations présentes dans les phrases des articles scientifiques, en calculant la similarité sémantique entre les relations définies dans l'ontologie et les phrases des articles. Le modèle BERT et trois variantes biomédicales sont utilisés pour obtenir les représentations des relations et des phrases. Ces modèles sont comparés sur un corpus construit à partir d'articles scientifiques complets issus de la plateforme ISTEX, dont une petite sous-partie a été annotée manuellement.

ABSTRACT

Annotating host-microbiota interactions in scientific articles using semantic similarity to an ontology

We are interested in the task of relation extraction in scientific articles on the subdomain of the human microbiome. In order to build an annotated corpus, we investigated the use of the OHMI ontology to detect the relations appearing in scientific article sentences by computing the semantic similarity between the relations defined in the ontology and the sentences. The BERT model and three biomedical variations are used to compute the representations of both relations and sentences. These models are compared on a corpus built from full-text scientific articles from the ISTEX platform, with a small subpart being manually annotated.

MOTS-CLÉS : Extraction de relations, ontologie, similarité sémantique, BERT, ISTEX.

KEYWORDS: Relation extraction, Ontology, Semantic similarity, BERT, ISTEX.

1 Introduction

Ces dernières années, le nombre de publications dans le domaine biomédical a énormément augmenté. L'extraction d'information devient une tâche indispensable pour la veille scientifique et l'intégration des connaissances dans ce domaine. Nous nous intéressons plus particulièrement à l'extraction de relations, dans les articles scientifiques, pour l'étude du microbiome humain. Cette tâche est généralement considérée comme une tâche de classification (Zhou *et al.*, 2014) sur un corpus annoté. Néanmoins, il n'existe aucun corpus annoté pour la tâche d'extraction de relations sur le microbiome humain et ses interactions. Il existe cependant des ontologies, dans le domaine biomédical et dans certains de ses sous-domaines : c'est le cas pour le microbiome humain, pour lequel il existe l'ontologie OHMI : Ontology of Host-Microbiome Interactions (He *et al.*, 2019).

Dans la suite, nous nous appuyons sur l'hypothèse distributionnelle (Harris, 1954), affirmant que la similarité des contextes dans lesquels apparaissent deux mots permet de mesurer leur proximité sémantique. Cela nous permet de calculer la similarité sémantique entre une phrase d'un article scientifique et les relations d'une ontologie, afin d'étudier la correspondance de ce score avec la présence ou l'absence d'une relation dans la phrase. Nous nous appuyons également sur l'utilisation de modèles adaptés de type BERT, pour obtenir les représentations sémantiques des relations de l'ontologie et des phrases des articles, ces modèles ayant permis d'obtenir d'importants progrès dans les domaines de spécialité (Pankaj & Gautam, 2022).

2 Ressources : ontologie, corpus et modèles

Ontologie des interactions hôte-microbiome (OHMI) Afin de définir les noms des interactions explicitement liées à l'étude du microbiome humain, nous nous appuyons sur l'ontologie des interactions hôte-microbiome OHMI (Ontology of Host-Microbiome) (He *et al.*, 2019). Celle-ci se définit comme une ontologie communautaire des interactions de l'être humain avec les éléments de son microbiome, le microbiote. OHMI contient un ensemble d'informations incluant le microbiote, avec une taxonomie microbienne des espèces hôtes, les entités anatomiques des hôtes, et les interactions hôte-microbiote dans différentes conditions. Dans la suite de notre étude, nous utilisons les 129 relations présentes dans l'ontologie. Ces relations sont plus ou moins longues et plus ou moins précises, comme l'illustrent les 2 exemples suivants : *'negatively regulated by'* et *'microbe susceptibly depleted in host with disease'*. Malgré le fait que certaines relations extraites aient des similitudes sémantiques et syntaxiques, elles ont un niveau de précision différent. Par conséquent, chaque relation est exprimée de manière unique. Pour illustrer cela, nous pouvons citer les deux relations *'causally upstream of'*, *'causally upstream of or within, negative effect'*. Lors de l'annotation, il est demandé d'annoter avec la relation la plus précise. Il est important de préciser que ce cas particulier des relations proches représente 19 relations, ce qui correspond à approximativement 15% de la liste des relations.

Corpus, pré-traitements et annotation La plateforme ISTEX (Cuxac & Thouvenin, 2017) permet d'accéder à plus de 25 millions de publications scientifiques. Nous l'utilisons pour construire un corpus de publications scientifiques, en anglais, portant sur le microbiome humain, à partir de la requête suivante : <"GUT MICROBIOTA" OR "GUT MICROBIOME" OR "INTESTINAL MICROBIOTA" OR "INTESTINAL MICROBIOME" AND LANGUAGE :ENG>. Nous obtenons, en résultat, un corpus de 8 657 publications scientifiques, dont les années de publication varient entre 2001 à 2019.

Le corpus obtenu est ensuite découpé en phrases, en utilisant la bibliothèque spaCy : on obtient ainsi 1 104 240 phrases. Afin de ne considérer que les phrases comportant potentiellement l'expression sémantique d'une relation sur le microbiome humain, nous partons de l'hypothèse que la présence d'une telle relation est positivement corrélée à la présence d'une ou plusieurs entités nommées liées au domaine biomédical. Nous faisons également l'hypothèse qu'une relation est contenue à l'intérieur d'une phrase, tout d'abord par simplicité, et aussi parce que c'est généralement le cas. Cela nous permet également de considérer les phrases, indépendamment les unes des autres, lors de l'annotation de la présence ou l'absence de relations, à l'intérieur de celles-ci.

La bibliothèque scispaCy (Neumann *et al.*, 2019) est utilisée pour identifier les entités nommées du domaine biomédical. Nous utilisons ainsi le modèle `en_ner_craft_md`, entraîné sur le corpus CRAFT (Bada *et al.*, 2012), et le modèle `en_ner_bc5cdr_md`, entraîné sur le corpus BC5CDR (Li

et al., 2016). Le premier modèle détecte les cellules (*Cell line*), les GGP (*Gene-or-Gene-Product*) et les taxons (*Taxon*) ainsi que les entités CHEBI présentes dans Chemical Entities of Biological Interest ontology (Degtyarenko *et al.*, 2007), les entités SO de Sequence Ontology (Eilbeck *et al.*, 2005) et les entités GO de Gene Ontology (Consortium, 2004). Quant au deuxième modèle, il détecte les noms de maladies (*Disease*) et les espèces chimiques (*Chemical*).

Afin d'étudier la distribution des scores de similarité sémantique, en fonction du nombre d'entités nommées biomédicales présentes dans les phrases, nous avons créé 4 sous-corpus, à partir de notre corpus initial de 1 104 240 phrases : le premier sous-corpus contient les 1 021 phrases avec une seule entité nommée, le deuxième contient les 762 phrases avec deux entités, le troisième contient les 454 phrases avec trois entités, et le dernier contient 618 phrases avec 4 entités nommées. Afin d'étudier l'utilisation du score de similarité sémantique pour la tâche d'extraction de relations, nous avons annoté une petite sous-partie de chacun des 4 sous-corpus : nous avons choisi aléatoirement 4 phrases, dans chaque sous-corpus, et les avons annotées avec la relation qui y était présente, en utilisant les 129 relations issues de l'ontologie OHMI. Cette tâche étant chronophage et le corpus actuellement construit étant de taille réduite, les premiers résultats obtenus sur celui-ci ne pourront donner que des indications sur l'utilisation du score de similarité sémantique pour la tâche d'extraction de relations.

Modèles de représentation du texte Compte tenu de la nature de la tâche de similarité sémantique, nous utilisons les modèles Sentence Transformers (Reimers & Gurevych, 2019) pour obtenir une représentation sémantique à la fois des relations extraites de l'ontologie et des phrases du corpus. Nous utilisons le modèle BERT (Devlin *et al.*, 2018) ainsi que trois variantes entraînées sur des données biomédicales, à savoir BioBERT (Lee *et al.*, 2019), SciBERT (Beltagy *et al.*, 2019) et PubMedBERT (Gu *et al.*, 2021). BioBERT est issu du pré-entraînement continu de BERT sur des résumés de PubMed. Quant à SciBERT, il est entièrement entraîné sur des articles complets de Semantic Scholar dont 82% s'inscrit dans le domaine biomédical. Enfin, PubMedBERT est entièrement entraîné sur des résumés de PubMed.

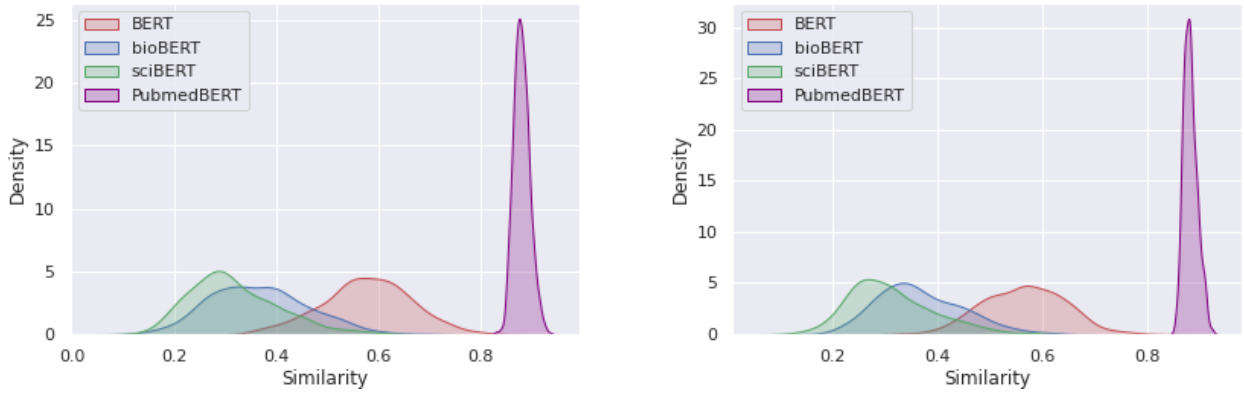
3 Méthodologie

Afin d'identifier si une relation est présente dans une phrase et la nature de cette relation, le cas échéant, la similarité cosinus est calculée entre la représentation sémantique de la phrase considérée (qui contient au moins une entité nommée) et la représentation sémantique de chacune des 129 relations issues de l'ontologie. La relation correspondant au score de similarité le plus élevé est ensuite attribuée à la phrase. Si la phrase ne contient aucune entité nommée du domaine biomédical, on considère qu'elle ne contient aucune relation de l'ontologie.

4 Expérimentations et résultats

Évaluation des scores de similarité sémantique Dans un premier temps, nous comparons la distribution des scores de similarité, sur chacun des 4 sous-corpus, selon la représentation sémantique des différents modèles considérés. La figure 1 montre la densité des scores de similarité sémantique, pour chaque modèle considéré, sur le sous-corpus contenant les phrases avec 1 entité nommée et sur celui contenant les phrases avec 4 entités nommées. En effet, les distribution des scores sont très

similaires, pour chaque modèle, sur chacun des 4 sous-corpus. Le nombre d'entités présentes dans une phrase affecte peu le score de similarité, ce qui induit que seule la présence des entités, sans prendre en considération leur nombre, peut être considérée comme un indicateur de la présence d'une relation dans une phrase. En effet, il est probable d'avoir des entités nommées dans une phrase sans qu'elles ne participent à la relation exprimée, comme pour la citation d'exemples dans une phrase. Le modèle PubMedBERT donne les scores les plus hauts, se situant entre 0,83 et 0,93. Ceci peut être expliqué par la sensibilité du modèle au vocabulaire de spécialité, puisqu'il est entièrement entraîné sur des textes du domaine biomédical. Les distributions de SciBERT et BioBERT restent assez proches, malgré les différences entre les deux modèles en termes de techniques de pré-entraînement et de corpus d'entraînement, et sont plus faibles que les scores obtenus avec le modèle BERT.



(a) Phrases avec 1 entité nommée

(b) Phrases avec 4 entités nommées

FIGURE 1 – Densité des scores de similarité cosinus par modèle et sur deux sous-corpus

Évaluation pour l'identification des relations Dans un second temps, nous donnons des premiers résultats obtenus sur un corpus annoté manuellement, afin de discuter de l'approche proposée pour identifier les relations en lien avec le microbiome humain. La table 1 présente les scores d'*accuracy*, pour chacun des 4 modèles, sur le petit sous-corpus annoté manuellement, en comparant la relation associée à la similarité cosinus la plus élevée avec la relation annotée manuellement.

Modèle	BERT	BioBERT	SciBERT	PubMedBERT
Accuracy	0,37	0,62	0,37	0,50

TABLE 1 – Accuracy des modèles sur les 16 phrases annotées manuellement

Sur ce petit corpus de 16 phrases, BioBERT obtient les meilleurs résultats. Alors que BioBERT et SciBERT obtenait des distributions de scores de similarité proches, SciBERT obtient des résultats nettement moins bons (similaires à ceux obtenus pour BERT, qui n'a pas été spécifiquement entraîné sur des données biomédicales). Le modèle PubMedBERT obtient des résultats entre ceux de BioBERT et ceux de SciBERT, dû a son entraînement exclusif sur des données biomédicales issues d'articles scientifiques. Le modèle BioBERT semble être celui qui allie le mieux l'encodage de l'information sémantique avec la connaissance du domaine de spécialité. Il convient de noter que les résultats des modèles sont souvent proches les uns des autres. Cela signifie que même si le modèle se trompe sur la relation attribuée à la phrase, la relation prédite reste cohérente avec ce qui est exprimé dans la phrase

et ne contredit pas la réalité. Ceci dit, dans certains cas, l'erreur peut être causée par la présence d'un terme de spécialité dans la phrase et dans la relation, ce qui pousse le modèle à les associer même si ce n'est pas la bonne relation. Nous en déduisons que l'amélioration de ces résultats consisterait à réduire le nombre élevé de relations extraites d'OHMI afin d'alléger le processus d'annotation et à opter pour des formulations plus simples pour améliorer la précision des résultats.

Compte tenu du grand nombre de relations à prendre en considération lors de l'annotation, à ce stade, un unique annotateur a effectué la tâche d'annotation, dans le but d'obtenir plus de visibilité sur la complexité de la tâche en l'état actuel. Quelques exemples sont fournis dans la table 2.

Phrases	**Annotations**
Moreover, airway microbiota composition and greater bacterial diversity were significantly correlated with bronchial hyperresponsiveness, including the relative abundance of specific microbiota belonging to bacterial families within the Proteobacteria.	microbe susceptibly expanded in respiratory airway of human with disease
Paradoxically, some degree of innate immune recognition of commensal bacteria is essential for normal development and function of the mucosal and peripheral immune system.	microbial population phenotype in host
Bronchoscopic studies indicate that the lungs of healthy people who smoke are inhabited by diverse types of bacteria in relatively small numbers and that this microbiome changes with disease.	microbe susceptibly depleted in respiratory airway of human with disease

TABLE 2 – Exemples de phrases avec leur relation manuellement annotée

5 Conclusion

Nous avons évalué l'utilisation de l'ontologie OHMI pour détecter les relations présentes dans les phrases des articles scientifiques, en calculant la similarité cosinus entre les relations définies dans l'ontologie et les phrases des articles. Le modèle BERT et trois variantes biomédicales ont été utilisés pour obtenir les représentations des relations et des phrases. Ces modèles ont été comparés sur un corpus construit à partir d'articles scientifiques complets issus de la plateforme ISTEX, dont une petite sous-partie a été annotée manuellement.

Cette étude a permis d'observer que le modèle BioBERT obtenait les meilleurs résultats et semblait être le plus adapté aux articles scientifiques considérés, en alliant connaissances biomédicales et informations sémantiques, pour identifier la relation présente dans une phrase. Pour pouvoir identifier plus précisément la relation de l'ontologie OHMI, présente dans une phrase donnée, il sera nécessaire de réduire le nombre de relations considérées. En effet, les 129 relations sélectionnées sont trop nombreuses, ce qui complexifie la tâche d'identification de la relation présente ainsi que la tâche d'annotation (l'annotation d'une phrase pouvait prendre jusqu'à 30 minutes).

Références

BADA M., ECKERT M., EVANS D., GARCIA K., SHIPLEY K., SITNIKOV D., BAUMGARTNER W. A., COHEN K. B., VERSPOOR K., BLAKE J. A. *et al.* (2012). Concept annotation in the craft corpus. *BMC bioinformatics*, **13**(1), 1–20.

BELTAGY I., LO K. & COHAN A. (2019). Scibert : A pretrained language model for scientific text. *arXiv preprint arXiv :1903.10676*.

CONSORTIUM G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, **32**(suppl_1), D258–D261.

CUXAC P. & THOUVENIN N. (2017). Archives numériques et fouille de textes : le projet istex. *Atelier TextMine, EGC 2017 (Extraction et Gestion des Connaissances), Grenoble, France, January 24*, **27**, 2017.

DEGTYARENKO K., DE MATOS P., ENNIS M., HASTINGS J., ZBINDEN M., MCNAUGHT A., ALCÁNTARA R., DARSOW M., GUEDJ M. & ASHBURNER M. (2007). Chebi : a database and ontology for chemical entities of biological interest. *Nucleic acids research*, **36**(suppl_1), D344–D350.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

EILBECK K., LEWIS S. E., MUNGALL C. J., YANDELL M., STEIN L., DURBIN R. & ASHBURNER M. (2005). The sequence ontology : a tool for the unification of genome annotations. *Genome biology*, **6**(5), 1–12.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, **3**(1), 1–23.

HARRIS Z. S. (1954). Distributional Structure. *WORD*, **10**(2-3), 146–162. DOI : 10.1080/00437956.1954.11659520.

HE Y., WANG H., ZHENG J., BEITING D. P., MASCI A. M., YU H., LIU K., WU J., CURTIS J. L., SMITH B. *et al.* (2019). Ohmi : the ontology of host-microbiome interactions. *Journal of biomedical semantics*, **10**, 1–14.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. DOI : 10.1093/bioinformatics/btz682.

LI J., SUN Y., JOHNSON R. J., SCIAKY D., WEI C.-H., LEAMAN R., DAVIS A. P., MATTINGLY C. J., WIEGERS T. C. & LU Z. (2016). Biocreative v cdr task corpus : a resource for chemical disease relation extraction. *Database*, **2016**.

NEUMANN M., KING D., BELTAGY I. & AMMAR W. (2019). Scispacy : Fast and robust models for biomedical natural language processing. *CoRR*, **abs/1902.07669**.

PANKAJ S. & GAUTAM A. (2022). Augmented bio-sbert : Improving performance for pairwise sentence tasks in bio-medical domain. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, p. 43–47.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

ZHOU D., ZHONG D. & HE Y. (2014). Biomedical relation extraction : from binary to complex. *Computational and mathematical methods in medicine*, **2014**.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Quand des Non-Experts Recherchent des Textes Scientifiques Rapport sur l'action CLEF 2023 SimpleText

Liana Ermakova[1] Stéphane Huet[2] Éric SanJuan[2]
Olivier Augereau[3] Hosein Azarbonyad[4] Jaap Kamps[5]

(1) Université de Bretagne Occidentale, HCTI, 29200 Brest, France
(2) Avignon Université, LIA, France
(3) ENIB, Lab-STICC UMR CNRS 6285, France
(4) Elsevier, The Netherlands
(5) University of Amsterdam, The Netherlands
liana.ermakova@univ-brest.fr, contact@simpletext-project.com

RÉSUMÉ

Le grand public a tendance à éviter les sources fiables telles que la littérature scientifique en raison de leur langage complexe et du manque de connaissances nécessaires. Au lieu de cela, il s'appuie sur des sources superficielles, trouvées sur internet ou dans les médias sociaux et qui sont pourtant souvent publiées pour des raisons commerciales ou politiques, plutôt que pour leur valeur informative. La simplification des textes peut-elle contribuer à supprimer certains de ces obstacles à l'accès ? Cet article présente l'action « CLEF 2023 SimpleText » qui aborde les défis techniques et d'évaluation de l'accès à l'information scientifique pour le grand public. Nous fournissons des données réutilisables et des critères de référence pour la simplification des textes scientifiques et encourageons les recherches visant à faciliter à la compréhension des textes complexes.

ABSTRACT

What Happens if Non-Experts Search Scientific Texts ? CLEF 2023 SimpleText track report

Users tend to stay away from credible sources such as scientific literature due to their intricate language or their lack of prior knowledge. Rather, they depend on shallow sources from the web or social media that can be published for economic or political motives rather than for their enlightening value. Text simplification might remove some of these barriers. This paper presents the report on the CLEF 2023 SimpleText track aiming to encourage the research on automatic simplification of scientific texts by providing reusable data and benchmarks.

MOTS-CLÉS : Textes scientifiques, simplification, recherche d'information.

KEYWORDS: Scientific texts, simplification, information retrieval.

1 Introduction

Les textes scientifiques tels que les articles de recherche sont difficiles à comprendre pour le grand public et même pour les scientifiques qui ne sont pas spécialisés dans le domaine en question. L'action CLEF 2023 SimpleText (Ermakova *et al.*, 2023) se concentre sur la simplification des textes scientifiques en combinant des aspects de la recherche d'information et du traitement automatique du

langage naturel (TALN).

La complexité des textes, les niveaux de lecture et la simplification des textes en général sont étudiés depuis longtemps dans les domaines de la linguistique, des sciences de l'éducation et du TALN. L'amélioration de la compréhensibilité des textes reste un défi, car il est difficile de définir le résultat souhaitable de la simplification (Grabar & Saggion, 2022). Les scores de lisibilité traditionnels se limitent à la longueur des mots ou des phrases, tandis que les mesures basées sur le chevauchement du vocabulaire ne tiennent pas compte de la distorsion de l'information.

Récemment, la simplification des textes a suscité un intérêt croissant. L'atelier sur le traitement des documents scientifiques [1] s'est adressé à un public spécialisé dans le TALN (Chandrasekaran *et al.*, 2020), en accueillant des tâches sur le résumé de documents scientifiques. Lors de l'EMNLP 2022, le projet TSAR (*Text Simplification, Accessibility, and Readability*) [2] a accueilli une tâche de simplification lexicale, et TermEval 2020 a exécuté une tâche partagée sur l'extraction automatique de termes (Rigouts Terryn *et al.*, 2020). L'action SimpleText que nous proposons ne se limite pas à la simplification lexicale et grammaticale.

SimpleText vise à améliorer l'accès à la connaissance scientifique pour le grand public, en fournissant des données réutilisables et des critères de référence pour la simplification des textes pour réduire les obstacles à la compréhension des textes complexes. Contrairement aux travaux précédents, nous nous concentrons sur (1) la sélection d'informations adaptées au grand public, (2) la recherche de concepts difficiles, notamment de mots, d'abréviations, etc. qui doivent être expliqués, et (3) l'évaluation de la distorsion de l'information susceptible de se produire au cours du processus de simplification.

SimpleText est basée sur la série suivante d'actions (Ermakova *et al.*, 2022c) : (1) sélectionner les informations à inclure dans un résumé simplifié ; (2) décider si les informations sélectionnées sont suffisantes et compréhensibles ou fournir des connaissances de base si ce n'est pas le cas ; et (3) améliorer la lisibilité du texte. Il en résulte trois tâches :

- ***Tâche 1: What is in, or out?*** *Selectionner des passages à inclure dans un résumé simplifié,*
- ***Tâche 2: What is unclear?*** *Identifier et expliquer les concepts complexes,*
- ***Tâche 3: Rewrite this!*** *Simplifier, réécrire un texte scientifique.*

Une quatrième tâche ouverte accepte également toute soumission qui utilise nos données d'une autre manière. Dans la suite de cet article, nous allons présenter chacune des trois premières tâches.

Pour cette nouvelle édition, nous enrichirons les données produites par les éditions précédentes (Ermakova *et al.*, 2022c) en y ajoutant des labels. Nous mettrons également l'accent sur l'évaluation automatique sur des données de test réutilisables.

2 Tâche 1 : Selectionner des passages à inclure dans un résumé simplifié

Cette tâche vise à extraire des passages qui peuvent aider à comprendre un article de vulgarisation scientifique à partir d'un large corpus de résumés académiques et de métadonnées bibliographiques. Les passages pertinents doivent se rapporter à l'un des sujets de l'article source.

Données. Nous utilisons les articles de vulgarisation scientifique comme sources pour les types de

1. https://sdproc.org/2022/sharedtasks.html
2. https://taln.upf.edu/pages/tsar2022-st/

sujets qui intéressent le grand public et comme validation du niveau de lecture qui leur convient. Le corpus principal est un vaste ensemble de résumés scientifiques et de métadonnées associées, couvrant le domaine de l'informatique et de l'ingénierie. Nous réutilisons la collection de résumés universitaires du Citation Network Dataset (12e version publiée en 2020) [3] (Tang *et al.*, 2008). Cette collection a été extraite de DBLP, ACM, MAG (Microsoft Academic Graph) et d'autres sources. Il contient : 4 894 083 références bibliographiques publiées avant 2020, 4 232 520 résumés en anglais, 3 058 315 auteurs avec leurs affiliations et 45 565 790 citations ACM. Nous fournissons un index ElasticSearch pour permettre aux participants de retrouver des passages ou des résumés à l'aide de BM25 (Robertson *et al.*, 2009). Grâce à une API, des requêtes peuvent être effectuées sur le contenu textuel des résumés ainsi que sur la paternité des auteurs. Ainsi, l'ensemble de données partagées fournit : a) le contenu des résumés des documents ; b) les auteurs des documents pour l'analyse des coauteurs ; c) la relation de citation entre les documents pour l'analyse des co-citations ; et d) les citations par auteur pour l'analyse du facteur d'impact des auteurs.

Les articles de presse utilisés s'adressent à un public général et proviennent de deux sources : *The Guardian*, un journal d'audience internationale destiné au grand public et comportant une section technologique, et *Tech Xplore*, [4] un site web qui participe au réseau Science X en se focalisant sur les progrès de l'ingénierie et de la technologie. Chacun de ces articles de vulgarisation scientifique représente un sujet général qui doit être analysé pour extraire les informations scientifiques pertinentes du corpus. Nous fournissons les URL des articles originaux, le titre et le contenu textuel de chaque article de vulgarisation scientifique en tant que sujet général. Chaque thème général a été également enrichi d'un ou plusieurs mots-clés spécifiques extraits manuellement de leur contenu, ce qui crée une tâche classique de recherche d'informations consistant à classer les passages ou les résumés en réponse à une requête. Dans l'édition précédente, 40 articles, 20 de chaque source, ont été mis à disposition (Ermakova *et al.*, 2022c). Nous prévoyons de l'étendre à 10 autres sujets utilisés comme ensemble de test. Les qrels 2022 couvrent de nombreux sujets (31) et requêtes (67), mais avec une profondeur limitée en documents annotés. En 2023, nous augmenterons la profondeur avec au moins 50 résumés jugés par requête.

Évaluation. La pertinence thématique n'a été évaluée l'année dernière qu'avec une note de 0 à 5 sur le degré de pertinence par rapport à l'article original. Si cette grande échelle permet de mesurer la proximité du résumé extrait avec le mot-clé, le titre ou le contenu textuel, d'autres facettes, pourtant importantes dans le contexte de la simplification des textes, manquaient à l'appel. En 2023, nous continuerons à évaluer la pertinence thématique, mais aussi la complexité du texte (à l'aide de mesures de lisibilité et d'une comparaison avec des scores attribués manuellement) et l'autorité de la source (à l'aide de mesures de l'impact académique).

La collection de test fournie sera simplifiée en trois notes sur une échelle de 0 à 2 :

- **Pertinence du sujet** : Non pertinent (0), pertinent (1), très pertinent (2) ;
- **Complexité du texte** : Facile (0), difficile (1), très difficile (2) ;
- **Crédibilité de la source** : Faible (0), moyenne (1), forte crédibilité (2).

Tout en continuant à utiliser un classement sur la pertinence à l'aide du NDCG, les deux autres critères permettront de comparer les systèmes sur d'autres aspects à prendre en compte.

3. https://www.aminer.cn/citation
4. https://techxplore.com/

3 Tâche 2 : Identifier et expliquer les concepts complexes

L'objectif de cette tâche est de déterminer quels concepts dans les résumés scientifiques nécessitent une explication et une mise en contexte afin d'aider le lecteur à comprendre le texte scientifique. L'identification des mots complexes et la simplification lexicale sont les approches les plus populaires pour évaluer et réduire la complexité (Cruz *et al.*, 2019; Yimam *et al.*, 2018; Kochmar *et al.*, 2020). Dans le cadre d'une requête, certains concepts clés doivent être contextualisés avec une définition, un exemple ou des cas d'utilisation plus faciles à comprendre pour le lecteur. Des recherches sont en cours à ce sujet, en générant des définitions d'une complexité contrôlable (August *et al.*, 2022).

Nous demandons aux participants d'identifier ces concepts et de fournir des explications utiles et compréhensibles. La tâche comporte deux étapes : (1) retrouver jusqu'à cinq termes difficiles dans un passage donné d'un résumé scientifique ; (2) fournir une explication de ces termes difficiles (par exemple, définition, déchiffrage d'abréviation, etc.).

Données. Le corpus de la Tâche 2 est basé sur les phrases des résumés les mieux classés pour les requêtes de la Tâche 1. Pour la première étape de la tâche, c'est-à-dire l'extraction des termes difficiles, nous utiliserons les données d'entraînement recueillies en 2022 (Ermakova *et al.*, 2022c). En ce qui concerne les données de test, nous fournirons des passages supplémentaires provenant des résumés DBLP comme dans la Tâche 1.

Pour la deuxième étape de la tâche, nous fournirons des données d'entraînement supplémentaires pour la génération de définitions, extraites d'un corpus beaucoup plus large d'articles en texte intégral. Ces données de formation contiennent des paires de $< phrase, concept >$ et un label par paire est fournie. Une paire indique si la phrase fournit une bonne définition du concept ou non. Les échantillons de cet ensemble de données sont extraits de livres et d'articles publiés dans ScienceDirect [5]. En plus de cet ensemble de données, les participants sont encouragés à utiliser des ensembles de données existants extraits d'autres ressources, tels que l'ensemble de données de la CMT (Navigli & Velardi, 2010) pour entraîner le modèle de génération de définitions.

Évaluation Comme en 2022, nous évaluerons la détection de concepts complexes en fonction de leur complexité et de la portée des concepts détectés (Ermakova *et al.*, 2022c). Pour l'explication des termes difficiles, l'ensemble d'évaluation contiendra 1 000 concepts et leurs définitions extraites par des experts en la matière. Nous évaluerons automatiquement les explications fournies en les comparant à des références (par exemple ROUGE, similarité cosinus, etc.). Nous évaluerons manuellement les explications fournies au niveau de l'utilité par rapport à une requête et de la complexité pour un public général. Les explications fournies peuvent prendre différentes formes : définition, décryptage d'abréviations, exemples, cas d'utilisation, etc.

4 Tâche 3 : Simplifier, réécrire un texte scientifique

L'objectif de cette tâche est de fournir une version simplifiée des phrases extraites des résumés scientifiques. Les participants recevront les articles et requêtes de vulgarisation scientifique ainsi que les résumés correspondants d'articles scientifiques, divisés en phrases individuelles.

Données. La Tâche 3 utilise un corpus constitué de phrases issues des résumés les mieux classés

5. https://www.sciencedirect.com/

pour les requêtes de la Tâche 1, complétées par des données d'entraînement supplémentaires dans le domaine de la santé. Nos données d'entraînement sont un corpus véritablement parallèle de phrases directement simplifiées (648 phrases pour l'instant) provenant de résumés scientifiques du DBLP Citation Network Dataset pour le sujet *Computer Science* et d'articles de Google Scholar et PubMed sur *la santé et la médecine* (Ermakova *et al.*, 2021, 2022a,c,b). Ces passages de texte ont été simplifiés soit par des étudiants en master de rédaction technique et de traduction, soit par un expert du domaine (un informaticien) et un traducteur professionnel (de langue maternelle anglaise) travaillant ensemble (Ermakova *et al.*, 2022a,c,b).

Évaluation. En 2023, nous mettrons l'accent sur les mesures d'évaluation automatique à grande échelle (SARI, ROUGE, compression, lisibilité) qui fournissent une collection réutilisable de tests. Elles seront complétées par une évaluation humaine d'autres aspects, essentielle pour une analyse plus approfondie. Comme en 2022, nous évaluerons la qualité des simplifications au niveau du vocabulaire et de la syntaxe ainsi que les erreurs (syntaxe incorrecte ; anaphore non résolue due à la simplification ; répétition/itération inutile ; erreurs d'orthographe, de typographie ou de ponctuation) (Ermakova *et al.*, 2022c). Plutôt que de nous concentrer uniquement sur cette évaluation, qui est similaire à celle utilisée dans des travaux antérieurs (Štajner *et al.*, 2022), nous examinerons également les résultats du point de vue de la distorsion de l'information qui peut survenir au cours du processus de simplification, avec un niveau de gravité comme suit : Style (1) ; Insertion de détails inutiles par rapport à une requête (1) ; Redondance (sans chevauchement lexical) (2) ; Insertion d'informations fausses ou non étayées (3) ; Omission de détails essentiels par rapport à une requête (4) ; Généralisation excessive (5) ; Simplification excessive (5) ; Changement de sujet (5) ; Contre-sens / contradiction (6) ; Ambiguïté (6) ; Absurdité (7).

5 Conclusions

Cet article décrit la mise en place de l'action CLEF 2023 SimpleText, qui contient trois tâches interconnectées sur la simplification des textes scientifiques. Dans le cadre du projet SimpleText, nous avons déjà publié des corpus de taille conséquente et des données étiquetées manuellement :

- un corpus de plus de 4 millions de résumés scientifiques utilisable pour la vulgarisation scientifique,
- des termes scientifiques utilisés dans des résumés scientifiques, avec des scores de difficulté attribués manuellement,
- un corpus parallèle de phrases simplifiées manuellement à partir de la littérature scientifique,
- un corpus parallèle de phrases avec différents types de distorsion de l'information et de niveau de simplification.

Pour plus de détails, il est possible de consulter le site (`http://simpletext-project.com`).

Remerciements

Ce travail n'aurait pas été possible sans le soutien de nombreuses personnes et le groupe de recherche MaDICS. Cette recherche a été financée en tout ou partie, par l'Agence Nationale de la Recherche (ANR) au titre du projet « ANR-22-CE23-0019-01 ».

Références

AUGUST T., REINECKE K. & SMITH N. A. (2022). Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8298–8317.

CHANDRASEKARAN M. K., FEIGENBLAT G., FREITAG D., GHOSAL T., HOVY E., MAYR P., SHMUELI-SCHEUER M. & DE WAARD A. (2020). Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, p. 1–6, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.sdp-1.1.

CRUZ F., COUSTATY M., AUGEREAU O., KISE K. & JOURNET N. (2019). An interactive recommendation system for 2nd language vocabulary learning-vocabulometer 2.0. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 3, p. 28–32 : IEEE.

ERMAKOVA L., BELLOT P., BRASLAVSKI P., KAMPS J., MOTHE J., NURBAKOVA D., OVCHINNIKOVA I. & SANJUAN E. (2021). Overview of SimpleText 2021 - CLEF Workshop on Text Simplification for Scientific Information Access. In K. S. CANDAN, B. IONESCU, L. GOEURIOT, B. LARSEN, H. MÜLLER, A. JOLY, M. MAISTRO, F. PIROI, G. FAGGIOLI & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, p. 432–449, Cham : Springer International Publishing.

ERMAKOVA L., BELLOT P., KAMPS J., NURBAKOVA D., OVCHINNIKOVA I., SANJUAN E., MATHURIN E., ARAÚJO S., HANNACHI R., HUET S. & POINSU N. (2022a). Automatic Simplification of Scientific Texts : SimpleText Lab at CLEF-2022. In M. HAGEN, S. VERBERNE, C. MACDONALD, C. SEIFERT, K. BALOG, K. NØRVÅG & V. SETTY, Éds., *Advances in Information Retrieval*, volume 13186, p. 364–373. Cham : Springer International Publishing.

ERMAKOVA L., OVCHINNIKOVA I., KAMPS J., NURBAKOVA D., ARAÚJO S. & HANNACHI R. (2022b). Overview of the CLEF 2022 SimpleText Task 3 : Query biased simplification of scientific texts. CEUR Workshop Proceedings.

ERMAKOVA L., SANJUAN E., HUET S., AUGEREAU O., AZARBONYAD H. & KAMPS J. (2023). CLEF 2023 SimpleText Track. In J. KAMPS, L. GOEURIOT, F. CRESTANI, M. MAISTRO, H. JOHO, B. DAVIS, C. GURRIN, U. KRUSCHWITZ & A. CAPUTO, Éds., *Advances in Information Retrieval*, p. 536–545, Cham : Springer Nature Switzerland.

ERMAKOVA L., SANJUAN E., KAMPS J., HUET S., OVCHINNIKOVA I., NURBAKOVA D., ARAÚJO S., HANNACHI R., MATHURIN É. & BELLOT P. (2022c). Overview of the CLEF 2022 SimpleText Lab : Automatic simplification of scientific texts. In A. BARRÓN-CEDEÑO, G. D. S. MARTINO, M. D. ESPOSTI, F. SEBASTIANI, C. MACDONALD, G. PASI, A. HANBURY, M. POTTHAST, G. FAGGIOLI & N. FERRO, Éds., *CLEF'22 : Proceedings of the Thirteenth International Conference of the CLEF Association*, Lecture Notes in Computer Science : Springer.

GRABAR N. & SAGGION H. (2022). Evaluation of automatic text simplification : Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 453–463.

KOCHMAR E., GOODING S. & SHARDLOW M. (2020). Detecting multiword expression type helps lexical complexity assessment. In *LREC 2020 : Proceedings of the 12th Conference on Language Resources and Evaluation*.

NAVIGLI R. & VELARDI P. (2010). Learning word-class lattices for definition and hypernym extraction. In *ACL*, p. 1318–1327.

RIGOUTS TERRYN A., HOSTE V., DROUIN P. & LEFEVER E. (2020). Termeval 2020 : Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, p. 85–94 : European Language Resources Association (ELRA).

ROBERTSON S., ZARAGOZA H. *et al.* (2009). The probabilistic relevance framework : Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, **3**(4), 333–389.

TANG J., ZHANG J., YAO L., LI J., ZHANG L. & SU Z. (2008). ArnetMiner : Extraction and mining of academic social networks. In *KDD'08*, p. 990–998.

YIMAM S. M., BIEMANN C., MALMASI S., PAETZOLD G., SPECIA L., ŠTAJNER S., TACK A. & ZAMPIERI M. (2018). A report on the complex word identification shared task 2018. In *The 13th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL2018 Workshops)*.

ŠTAJNER S., SHEANG K. C. & SAGGION H. (2022). Sentence Simplification Capabilities of Transfer-Based Models.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Apprentissage de dépendances entre labels pour la classification multi-labels à l'aide de transformeurs

Haytame Fallah[1, 3, *] Elisabeth Murisasco[2, *]
Emmanuel Bruno[2, *] Patrice Bellot[1, *]
(1) Aix-Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
(2) Université de Toulon, Aix-Marseille Université, CNRS, LIS, Toulon, France
(3) Hyperbios, Aix-en-Provence, France
(*) prénom.nom@lis-lab.fr

RÉSUMÉ

Dans cet article, nous proposons des approches pour améliorer les architectures basées sur des transformeurs pour la classification de documents multi-labels. Les dépendances entre les labels sont cruciales dans ce contexte. Notre méthode, appelée DepReg, ajoute un terme de régularisation à la fonction de perte pour encourager le modèle à prédire des labels susceptibles de coexister. Nous introduisons également un nouveau jeu de données nommé "arXiv-ACM", composé de résumés scientifiques de la bibliothèque numérique arXiv, étiquetés avec les mots-clés ACM correspondants.

ABSTRACT

Exploiting Label Dependencies for Multi-Label Document Classification Using Transformers.

In this paper, we propose approaches to improve transformer-based architectures for multi-label document classification. Dependencies between labels are crucial in this context. Our method, called DepReg, adds a regularization term to the loss function to encourage the model to predict labels that may coexist. We also introduce a new dataset named "arXiv-ACM", consisting of scientific abstracts from the arXiv digital library, labeled with the corresponding ACM keywords.

MOTS-CLÉS : Classification multi-labels, Transformeurs, Dépendances entre labels.

KEYWORDS: Multi-label Classification, Transformers, Label Dependencies.

1 Introduction

La classification multi-labels peut être considérée comme une généralisation de la classification binaire ou multi-classes traditionnelle. Le but est d'associer un ou plusieurs labels au texte d'entrée. C'est une tâche importante pour de nombreuses applications telles que la classification d'articles de recherche (Mustafa *et al.*, 2021; Sajid *et al.*, 2011) ou la réponse aux questions (Sahu *et al.*, 2019; Wu *et al.*, 2019).

Plusieurs méthodes ont été proposées pour résoudre la classification multi-labels. Elles peuvent être divisées en trois familles : transformation de problème (Luaces *et al.*, 2012; Tsoumakas *et al.*, 2010), adaptation de problème et méthodes d'ensemble (Saini & Ghosh, 2017; Tsoumakas & Vlahavas, 2007). Dans la configuration multi-labels, il est nécessaire de trouver des caractéristiques discriminantes pour identifier chacun d'entre eux dans le texte donné, mais dans de nombreuses applications, des

dépendances peuvent exister entre les labels cibles.

Nous proposons dans cet article une approche d'adaptation de problème qui consiste dans l'exploitation de la corrélation des labels pour la classification de documents multi-labels qui tire parti de leurs cooccurrences de manière simple mais efficace. Pour cette approche, nous proposons un terme de régularisation qui est ajouté à la fonction de perte de la tâche de classification, basé sur les labels prédits et la matrice de similarité de labels. Une matrice calculée en utilisant la similarité cosinus entre les coocurrences de chaque label par rapport aux autres.

Nous proposons un nouveau jeu de données multi-labels arXiv-ACM, qui diffère de ceux présents dans la littérature de part sa grande cardinalité ainsi qu'une meilleure distribution des échantillons par nombre de labels. Il est construit à partir de résumés d'articles scientifiques de la bibliothèque numérique arXiv et appariés avec les mots-clés ACM de niveau 2 fournis par les auteurs.

2 Travaux connexes

Diverses méthodes ont été proposées pour modéliser la dépendance entre les labels en utilisant des structures hiérarchiques (Alaydie *et al.*, 2012; Yang *et al.*, 2016), des graphes et réseaux conditionnels (Guo & Gu, 2011; Zhang & Zhang, 2010), des cooccurrences de labels, ou encore une combinaison de ces approches (Wu *et al.*, 2018). Toutefois, ces méthodes tendent à privilégier les relations verticales entre les labels plutôt que les dépendances latérales.

MAGNET (Pal *et al.*, 2020), un réseau de neurones en graphes qui utilise les plongements de mots de BERT, met en œuvre le mécanisme d'attention pour capturer la structure de dépendance entre les labels. LW-PT (Liu *et al.*, 2020) introduit une nouvelle tâche de pré-entraînement de classification de documents par labels et entraîne des encodeurs de documents par labels. Ces deux méthodes parviennent à avoir de bonnes performances en score F1 pour les ensembles de données AAPD et Reuters (cf. section 4) tout en utilisant des LSTMs pour l'extraction des caractéristiques textuelles.

Dans le domaine de la vision par ordinateur, et plus précisément dans la détection d'objets qui est essentiellement une tâche de classification multi-labels, (Cheng *et al.*, 2021) utilise un modèle transformeur et une attention au niveau des pixels pour capturer les dépendances spatiales dans une image, ainsi qu'un jeton spécifique à l'objet qui est ajouté à une étape ultérieure dans le modèle. Ce jeton est utilisé pour prédire le nombre de labels pour une instance donnée et peut aider le modèle à établir une correspondance plus robuste entre les caractéristiques d'entrée et les labels à prédire.

Une approche alternative consiste dans l'utilisatiion des fonctions de perte d'équilibrage pour améliorer la performance des labels sous-représentés (Huang *et al.*, 2021). Cependant, cela peut réduire l'information apportée par les cooccurrences de labels qui peuvent corriger les biais des modèles et améliorer la performance de la classification pour les labels moins fréquentes.

Dans Liu *et al.* (2022) (CNLE), les labels sont encodés sous forme de plongements de mots et alimentés avec la séquence de texte dans un réseau de co-attention (Seo *et al.*, 2016). Les représentations de texte en fonction des labels et les labels en fonction du texte sont ensuite utilisées pour la classification multi-labels en la considérant comme un problème de génération de séquences. Le mécanisme d'attention tient compte de la relation entre les labels, mais son efficacité est limitée dans les cas où les labels ne sont pas des mots complets mais plutôt des abréviations ou des codes, comme "cs.it" dans le jeu de données AAPD. Dans de tels cas, les labels sont initialisées avec une distribution

de probabilité plutôt que des plongements pré-entraînés, ce qui ne fournit pas une représentation optimale pour les labels. Dans la Section 4, nous montrons que notre approche ne dépend pas de la nature des labels et contribue à une amélioration des performances pour tous les ensembles de données étudiés.

3 Régularisation par Dépendance (DepReg)

Nous présentons ici notre méthode de Régularisation par Dépendance (DepReg), qui ajoute un terme de régularisation à la fonction de coût du modèle transformeur pour incorporer les informations de cooccurrence des labels.

Nous construisons une matrice de similarité S en utilisant la matrice de cooccurrence C.

$$S = CS_\theta(C, C) \tag{1}$$

Le terme de régularisation L_{DepReg} est calculé comme le produit scalaire entre le vecteur de dissimilarité D_{sim} et le vecteur de prédiction transposé $\hat{Y}^T$.

$$D_{sim} = 1 - CS_\theta(S, \hat{Y}) = 1 - \frac{S \cdot \hat{Y}}{|S| \cdot |\hat{Y}|} \tag{2}$$

$$L_{DepReg} = D_{sim} \cdot \hat{Y}^T \tag{3}$$

Le terme de régularisation est ajouté à la fonction de coût principale lors de l'apprentissage pour encourager le modèle à faire des prédictions cohérentes avec les dépendances des labels présentes dans le jeu de données :

$$L_{total} = L_{BCE} + \lambda_{reg}.L_{DepReg} \tag{4}$$

où L_{BCE} est la perte d'entropie croisée binaire (Binary Cross Entropy) et λ_{reg} est un hyperparamètre contrôlant le poids du terme de régularisation dans la perte totale.

Le terme de régularisation par dépendance (DepReg) aide le modèle à éviter de faire des prédictions qui vont à l'encontre des informations de cooccurrence tout en lui permettant de faire des prédictions basées sur les motifs appris dans les données d'entraînement.

4 Expériences et Résultats

Pour l'évaluation des méthodes proposées, nous utilisons l'implémentation de HuggingFace du modèle BERT (Devlin *et al.*, 2019) en version *uncased-base*. Nous ajoutons un réseau de neurones (FFNN) avec $L = 2$ couches et utilisons l'entropie croisée binaire (BCE) pour la version de base de BERT et pour notre approche DepReg. AdamW est l'optimiseur utilisé avec un λ_{reg} de 0.2.

Jeux de données :

— **AAPD** (arXiv Academic Paper Dataset) est une collection de "résumés" d'articles scientifiques de la bibliothèque numérique arXiv. Un article peut avoir une ou plusieurs classifications parmi 54 labels. Nous utilisons la même distribution d'entraînement (53840), de validation (1000)

et de test (1000) que (Yang *et al.*, 2018). Cet ensemble de données présente de nombreuses limites. La plus contraignante est le nombre d'instances par labels. En général, les instances avec un seul label sont beaucoup plus courantes que celles avec plusieurs.

— Pour remédier à ces limitations, nous introduisons dans cet article un nouveau jeu de données multi-labels, que nous appelons "arXiv-ACM", avec une grande cardinalité, une taille raisonnable et une meilleure distribution des échantillons selon le nombre de labels. **ArXiv-ACM** est composé de résumés d'articles en informatique extraits via l'API arXiv [1], publiés entre 1998 et 2021. Ces résumés ont ensuite été associés aux mots-clés ACM [2] fournis par les auteurs des articles lors de la soumission. Seuls les mots-clés de deuxième niveau ont été considérés, car le premier niveau est trop large et les niveaux suivants sont trop spécifiques. Nous avons ensuite filtré les labels qui ont moins de 20 instances pour obtenir 64 labels.

Baselines :

— **MAGNET** (Pal *et al.*, 2020) : classification de texte multi-labels utilisant un réseau de neurones en graphes avec mécanisme d'attention pour capturer les dépendances entre les labels,

— **CNLE** (Liu *et al.*, 2022) : un modèle de transformeur qui introduit les plongments des labels en plus de ceux du texte, liés par une co-attention pour obtenir une représentation contextualisée de la séquence d'entrée par les labels de classification,

Résultats : Comme le montre le tableau 1, l'utilisation des informations de dépendance contenues dans la matrice de cooccurrence des labels entraîne une augmentation du score micro-F1 pour les deux jeux de données avec notre approche.

SVM peut être considérée comme la meilleure approche non neuronale, mais elle est inférieure aux autres méthodes testées. La précision plus élevée a un coût de rappel plus faible, réduisant ainsi le score micro F1. La précision seule n'est pas un facteur fiable pour mesurer les performances dans les tâches de classification.

AAPD - En raison de la nature du vocabulaire utilisé dans les résumés scientifiques, ce jeu de données est complexe. Les scores de performance montrent que les modèles ont du mal à associer les sujets à leurs résumés correspondants. L'apprentissage de la dépendance obtient cette fois-ci une augmentation du rappel avec un score micro-F1 de 73,81.

arXiv-ACM - Ce jeu de données partage la nature scientifique des documents avec AAPD. Malgré une distribution plus équilibrée du nombre de labels par instance, les scores sont les plus bas pour les modèles testés. La méthode DepReg présente le gain le plus élevé en terme de précision avec un gain de 1,76 par rapport à la version base de BERT. Cette augmentation notable, associée à un gain de rappel, contribue au meilleur score micro-F1 pour cet ensemble de données (58,08), avec la plus forte augmentation par rapport à la version de base de BERT.

L'augmentation des performances obtenue par l'approche d'apprentissage de dépendance que nous proposons peut être expliquée par le fait que la prédiction d'un label est influencée par la prédiction de tous les autres labels en utilisant les co-occurrences. Dans certains cas, ces informations aident à prédire des labels qui n'auraient pas été prédites autrement (augmentation du rappel). D'autre part, les dépendances entre les labels peuvent réduire le nombre de faux positifs en réduisant le biais que le modèle peut avoir pour les labels fréquentes dans l'ensemble de données, contribuant ainsi à une amélioration de la précision.

1. https ://arxiv.org/help/api/
2. https ://www.acm.org/publications/computing-classification-system/1998/ccs98

Models	arXiv-ACM			AAPD		
	Pr.	R	F1	Pr.	R	F1
Baselines						
GradientBoost	57,99	29,87	39,43	79,73	46,8	58,98
SVM	**70,15**	39,79	50,78	**80,85**	59,98	68,86
MAGNET	57,31	53,24	55,2	72,88	66,79	69,7
CNLE	56,85	52,37	54,52	74,71	69,11	71,80
BERT	60,04	55,58	57,72	74,49	72,03	73,24
Our Dependency Learning Approaches						
BERT+ *DepReg*	61,80	**56,09**	**58,08**	75,53	**72,16**	**73,81**

TABLE 1 – Micro-précision (Pr.), micro-rappel (R) et scores micro-F1 (F1) pour les ensembles de tests arXiv-ACM et AAPD. Les meilleurs sont en bleu gras.

Models	arXiv-ACM				AAPD			
	Head	Med	Tail	3+ subset	Head	Med	Tail	3+ subset
BERT	58,95	52,11	**41,55**	54.32	73,97	69,29	**64,82**	66,47
BERT+*DepReg*	**59.80**	**56.62**	40.38	**55,39**	74,56	**69,75**	64,23	**67,18**

TABLE 2 – Scores Micro-F1 pour les labels de tête, de moyenne et de queue. Les meilleurs scores sont en bleu gras.

Head, Med, Tail - Pour étudier l'influence du déséquilibre des données, nous suggérons d'analyser les résultats d'inférence sur les labels tête (Head), moyenne (Med) et queue (Tail) selon la fréquence des labels pour arXiv-ACM (head > 600 instances, Med entre 50-600, Tail $\leq$ 50) et AAPD (head >3000 instances, Med entre 1000-3000, Tail $\leq$ 1000). Notre approche basée sur la cooccurrence de labels réalise une augmentation notable des performances, mais seulement dans les sous-ensembles tête et moyen, (cf. tableau 2).

3+ Subset - Pour une évaluation plus pratique des gains de performance en classification de documents, nous analysons également le sous-ensemble de chaque jeu de données contenant au moins 3 labels. Notre méthode de dépendance de labels obtient généralement les scores les plus élevés pour ces sous-ensembles. Ce qui montre que notre méthode capture efficacement les dépendances entre labels et conduit à une amélioration des performances.

5 Conclusion

La classification multi-label est une tâche pertinente, en particulier pour la gestion des bibliothèques numériques et l'étiquetage automatique de documents. Dans cet article, nous avons proposé une méthode simple mais efficace pour utiliser les informations de cooccurrence des labels pour permettre aux modèles à base de transformeurs d'apprendre les dépendances entre les labels. La méthode de régularisation de dépendance (DepReg) s'est avérée particulièrement efficace, améliorant les

performances du modèle BERT de base. Cependant, cette approche pénalise les labels moins fréquents en raison de leurs faibles probabilités de cooccurrence. Atténuer cet inconvénient peut conduire à un gain de performance plus notable. Enfin, nous avons introduit dans cet article un nouveau jeu de données multi-label, l'ensemble de données "arXiv-ACM", plus adapté pour tester les nouvelles approches multi-label. Notre jeu de données et tout le code de mise en œuvre seront disponibles au moment de la publication [3].

Références

ALAYDIE N., REDDY C. K. & FOTOUHI F. (2012). Exploiting Label Dependency for Hierarchical Multi-label Classification. In P.-N. TAN, S. CHAWLA, C. K. HO & J. BAILEY, Éds., *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, p. 294–305, Berlin, Heidelberg : Springer.

CHENG X., LIN H., WU X., YANG F., SHEN D., WANG Z., SHI N. & LIU H. (2021). Mltr : Multi-label classification with transformer. *CoRR*, **abs/2106.06195**.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

GUO Y. & GU S. (2011). Multi-label classification using conditional dependency networks. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, p. 1300–1305, Barcelona, Catalonia, Spain : AAAI Press.

HUANG Y., GILEDERELI B., KÖKSAL A., ÖZGÜR A. & OZKIRIMLI E. (2021). Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 8153–8161, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.643.

LIU H., YUAN C. & WANG X. (2020). Label-Wise Document Pre-training for Multi-label Text Classification. In X. ZHU, M. ZHANG, Y. HONG & R. HE, Éds., *Natural Language Processing and Chinese Computing*, Lecture Notes in Computer Science, p. 641–653, Cham : Springer International Publishing.

LIU M., LIU L., CAO J. & DU Q. (2022). Co-attention network with label embedding for text classification. *Neurocomputing*, **471**, 61–69. DOI : https ://doi.org/10.1016/j.neucom.2021.10.099.

LUACES O., DÍEZ J., BARRANQUERO J., DEL COZ J. J. & BAHAMONDE A. (2012). Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, **1**(4), 303–313.

MUSTAFA G., USMAN M., YU L., AFZAL M., SULAIMAN M. & SHAHID A. (2021). Multi-label classification of research articles using word2vec and identification of similarity threshold. *Scientific Reports*, **11**, 21900. DOI : 10.1038/s41598-021-01460-7.

PAL A., SELVAKUMAR M. & SANKARASUBBU M. (2020). Multi-Label Text Classification using Attention-based Graph Neural Network. In *ICAART*.

SAHU T. P., THUMMALAPUDI R. S. & NAGWANI N. K. (2019). Automatic Question Tagging Using Multi-label Classification in Community Question Answering Sites. In *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, p. 63–68. DOI : 10.1109/CSCloud/EdgeCom.2019.00-17.

3. https ://github.com/hf-lis/Coria-TALN-2023

SAINI R. & GHOSH S. (2017). Ensemble classifiers in remote sensing : A review. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, p. 1148–1152. DOI : 10.1109/CCAA.2017.8229969.

SAJID N. A., ALI T., AFZAL M. T., AHMAD M. & QADIR M. A. (2011). Exploiting reference section to classify paper's topics. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, MEDES '11, p. 220–225, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2077489.2077531.

SEO M. J., KEMBHAVI A., FARHADI A. & HAJISHIRZI H. (2016). Bidirectional attention flow for machine comprehension. *CoRR*, **abs/1611.01603**.

TSOUMAKAS G., KATAKIS I. & VLAHAVAS I. (2010). Mining Multi-label Data. In O. MAIMON & L. ROKACH, Éds., *Data Mining and Knowledge Discovery Handbook*, p. 667–685. Boston, MA : Springer US.

TSOUMAKAS G. & VLAHAVAS I. (2007). Random k -Labelsets : An Ensemble Method for Multilabel Classification. volume 4701, p. 406–417.

WU B., JIA F., LIU W., GHANEM B. & LYU S. (2018). Multi-label Learning with Missing Labels Using Mixed Dependency Graphs. *International Journal of Computer Vision*, **126**(8), 875–896. DOI : 10.1007/s11263-018-1085-3.

WU H., ZHANG S., WANG J., LIU M. & LI S. (2019). Multi-label Aspect Classification on Question-Answering Text with Contextualized Attention-Based Neural Network. In M. SUN, X. HUANG, H. JI, Z. LIU & Y. LIU, Éds., *Chinese Computational Linguistics*, Lecture Notes in Computer Science, p. 479–491, Cham : Springer International Publishing.

YANG P., SUN X., LI W., MA S., WU W. & WANG H. (2018). SGM : Sequence Generation Model for Multi-label Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 3915–3926, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

YANG Z., YANG D., DYER C., HE X., SMOLA A. & HOVY E. (2016). Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1480–1489, San Diego, California : Association for Computational Linguistics.

ZHANG M.-L. & ZHANG K. (2010). Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, p. 999–1008, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1835804.1835930.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Elaboration d'un corpus d'apprentissage à partir d'articles de recherche en chimie

Bénédicte Goujon
Thales R&T France, 1 avenue Fresnel, 91767 Palaiseau Cedex, France
benedicte.goujon@thalesgroup.com

RESUME

Dans le cadre d'un projet mené en 2021, un objectif consistait à extraire automatiquement des informations à partir d'articles de recherche en chimie des matériaux : des valeurs associées à des propriétés pour différents composants chimiques. Le travail présenté ici décrit les étapes de la construction du corpus textuel d'apprentissage, annoté manuellement par des experts du domaine selon les besoins identifiés dans le projet, pour une utilisation ultérieure par des outils d'extraction d'informations.

ABSTRACT

Here the title in English.
In a project conducted in 2021 for a chemistry consortium, an objective was dealing with the automatic extraction of the following specific information from research papers: numerical values associated to properties for various chemical components. The work presented here describes the steps for building the learning corpus manually annotated by domain experts according to the project specific needs, for a later use by extraction information tools.

MOTS-CLES : extraction d'informations, annotation manuelle, modélisation, articles de chimie
KEYWORDS: information extraction, manual annotation, modelling, chemistry papers

1 Introduction

Dans le cadre d'un projet mené en 2021 avec un consortium de chimie, un objectif consistait à extraire automatiquement les informations spécifiques suivantes à partir d'articles de recherche : des valeurs numériques associées à des propriétés pour différents composants chimiques. L'idée était de permettre ensuite des recherches dans la base de données regroupant les informations extraites, visant certains composants chimiques, certaines propriétés, ou avec des fourchettes sur des valeurs, pour repérer par exemple les valeurs obtenues sur une propriété ou pour identifier les composants chimiques ayant des valeurs recherchées sur une propriété, en proposant un lien avec le texte source afin de faciliter la validation des informations extraites.

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 41–46.

Une quinzaine d'expert.es en chimie, avec des compétences et expertises variées, étaient disponibles pour participer aux différentes discussions et procéder à l'annotation manuelle des corpus. La plateforme INCEpTION (Klie et al., 2018), qui intègre l'outil d'annotation WebAnno, a été choisie pour permettre l'annotation manuelle d'extraits textuels par plusieurs annotateurs et annotatrices, en lien avec un modèle d'annotation partagé non modifiable devant être défini.

Alors que de nombreux travaux ciblent la génétique ou le biomédical (Islamaj et al. 2022) (Wei et al., 2016), la chimie des matériaux est pauvre en support pour la construction de corpus d'apprentissage annotés manuellement. Concernant les corpus annotés existants, il est difficile d'en trouver qui contiennent des entités de type Composant Chimique (Chemical) et des entités de type Propriété (Property). Le corpus bc5cdr[1], disponible sur HuggingFace, propose le type d'entité « Chemical », ainsi que « Disease », mais il ne couvre pas des textes en lien avec la chimie des matériaux comme visés ici. Un autre corpus, CHEMDNER (Krallinger et al., 2015)[2], contient des annotations de noms de composants chimiques, mais ce n'est pas non plus un corpus lié à la chimie des matériaux et les types d'entité utilisés (abbreviation, family, formula, identifier, multiple, systematic et trivial), qui correspondent à différentes formes textuelles des noms de composants chimiques, ne sont pas directement pertinents dans notre approche.

Le travail présenté ici aborde des étapes et réflexions ayant permis de construire un corpus d'apprentissage, annoté manuellement par des experts du domaine selon les besoins identifiés, pour une utilisation ultérieure par des outils d'extraction d'informations utilisant l'apprentissage symbolique de patrons linguistiques tels que STRASS (Goujon, 2021) ou l'apprentissage à base de réseaux de neurones comme les Transformers, en vue d'alimenter une base de données regroupant des informations issues de travaux de recherches ciblés.

2 La construction d'un corpus d'extraits textuels à annoter

Le consortium de chimie s'est intéressé aux propriétés mécaniques telles que la résistance à la traction (« tensile strength ») de composants chimiques en lien avec le polymère acrylonitrile butadiène styrène ou ABS. Un premier corpus contenant 43 documents a été récupéré de bases d'articles de recherche en ligne par différents experts. Les documents obtenus, au format pdf, font en moyenne 15 pages avec un minimum de 9 pages et un maximum de 26 pages.

Afin de filtrer cet ensemble textuel pour ne fournir aux personnes devant annoter que des phrases potentiellement porteuses d'informations recherchées (valeurs sur des propriétés), une experte a parcouru chaque document initial pour en extraire des sous-paragraphes. Résultant de cette étape, un corpus de 89 extraits d'articles au format texte a été obtenu. Chaque extrait, contenant entre 2 et 10 phrases, a fait l'objet d'un nettoyage manuel afin d'obtenir des textes non bruités (sans sauts de ligne inutiles ou références bibliographiques). Les tableaux des textes sources, contenant de nombreuses valeurs recherchées, n'ont pas été retenus pour les annotations manuelles. Ils ont fait l'objet d'annotations automatiques, via des requêtes (proches de règles symboliques) portant sur le contenu textuel des en-têtes de lignes et colonnes, avec l'outil I2E[3] de Linguamatics utilisé en parallèle du travail présenté ici. Voici trois sous-extraits pour illustrer nos remarques, tirés de (Dul et al., 2018) (Aw et al. 2018) (Verbeteen et al., 2021) :

[1] https://huggingface.co/datasets/tner/bc5cdr
[2] https://huggingface.co/datasets/bigbio/chemdner
[3] https://www.linguamatics.com/products/i2e

- Extrait 1 : « … for ABS/graphene composites. Interestingly enough, these composites show elastic modulus of about 7362 MPa and tensile strength of about 44 MPa. »

- Extrait 2 : « The tensile strength for ABS/ZnO line samples were 23.3, 24.19, and 28.24 MPa for the infill density of 50%, 75%, and 100%, respectively. For CABS/ZnO line samples, the tensile strength improved 6.3% to 10.31 MPa when infill density changed from 50% to 100%. »

- Extrait 3 : « The set at v_p = 5 mm/s has the lowest elastic modulus of E = 1.9–2.0 GPa, while samples fabricated at v_p = 20 mm/s have values around 2.2 GPa. »

En analysant les extraits textuels obtenus, on a pu observer que peu de phrases portent explicitement toutes les informations recherchées, et que certaines phrases contiennent plusieurs informations simultanément, ce qui, dans les deux cas, pénalise l'efficacité des annotations manuelles. En effet, d'une part, le recours aux coréférences est assez fréquent dans les articles assez longs, notamment pour améliorer la lisibilité des textes, comme avec « these composites » dans l'extrait 1. Or, dans notre contexte, le recours à des extraits textuels et à des experts du domaine sans expertise en langage naturel nous ont amenés à ne pas gérer l'annotation des coréférences. D'autre part, certaines phrases comparent des valeurs obtenues, dont les détails sont présentés dans des tableaux, en se focalisant sur les améliorations obtenues (augmentations ou diminutions), comme dans l'extrait 2 avec « the tensile strength improved 6.3% », où la valeur initiale n'est pas explicitée. Au final, l'annotation manuelle et l'annotation automatique visée ont peu de chance d'être suffisantes pour l'extraction de toutes les informations visées, et devraient nécessiter des relectures et complétions.

3 La construction du modèle d'annotation

Des ontologies existent pour couvrir le domaine très large de la chimie, telles que ChEBI (Chemical Entities of Biological Interest) (Hastings, 2016), qui contient 46 000 entrées. Ces modèles très détaillés ne sont pas adaptés à l'annotation manuelle pour l'extraction d'informations spécifiques. D'une part, il n'est pas envisageable de proposer pour l'annotation manuelle des modèles contenant des milliers de concepts, dont la structure et le contenu devraient être maîtrisés pour la production d'annotations manuelles homogènes et de bonne qualité. De plus, les relations répertoriées dans ChEBI ne couvrent pas l'association de valeurs numériques à des propriétés de composants chimiques mais principalement l'organisation des concepts : « is a », « has part »… Enfin, de nombreux noms de composants chimiques cités dans les articles de recherche ne correspondent pas à des entités chimiques préexistantes mais font juste référence à des compositions spécifiques testées dans les expériences décrites, tels que « ABS/graphene » (extrait 1) ou « ABS/ZnO » (extrait 2).

Dans ce contexte, nous avons choisi de définir notre propre modèle d'annotation, le plus simple possible, centré sur le besoin visé. Initialement, nous avons proposé le modèle suivant : un type d'entité central « Chemical » et deux types complémentaires « Property » et « Value » complétés par des relations de type « has_property » et « has_value ». Cependant, la confrontation avec les premiers extraits textuels a rapidement fait remonter certaines limites. En effet, peu de phrases sont réellement centrées sur le nom explicite du composant chimique, avec une valeur et un nom de propriété. Ainsi dans la phrase 1, le nom du composant chimique n'est pas précisé. L'objectif étant l'extraction de valeurs sur les propriétés, il nous a ensuite semblé pertinent de centrer le modèle sur les entités de type Value, mais cette modélisation n'a pas été validée par certains experts, et nous avons pu observer que certaines valeurs n'étaient pas toujours explicitement données dans les textes (voir extrait 3). Enfin, les conditions d'obtention des valeurs, telles que les conditions de

températures par exemple, sont très importantes pour expliquer des différences de valeurs. Nous avons dû ajouter le nouveau type d'entité « Condition », associé à des valeurs numériques (extrait 2 : « infill density of 50% ») ou non (extrait 2 : « line » correspond à la mise en forme du matériau). Ce type d'entité regroupe des informations très variées, allant de températures à des noms d'appareils de mesure utilisés.

L'annotation manuelle de textes permet de tracer des relations binaires orientées entre deux entités, or dans le besoin visé ici plusieurs relations différentes peuvent être définies entre les quatre types d'entités, puisqu'une valeur correspond à une propriété pour un composant chimique avec une ou plusieurs conditions et valeurs associées. Afin de simplifier la tâche d'annotation, et pour éviter d'avoir différentes annotations acceptables pour une même phrase, nous avons retenu les quatre types de relations suivants pour l'annotation manuelle, plutôt centrés autour du concept « Property », avec entre parenthèses le ou les types d'entités source possibles suivis du type d'entité cible : Is_property_of(Property, Chemical), Has_value([Property, Condition], Value), Measured_at([Property,Value], Condition) et With_condition(Chemical, Condition).

4 La préparation à l'annotation manuelle

Les chimistes disponibles pour réaliser l'annotation manuelle du corpus n'ont jamais eu à effectuer ce type de tâche. Un tutoriel a été mis en place et leur a été présenté afin de diriger leur façon d'annoter les extraits d'articles avec l'outil INCEpTION. Voici quelques commentaires qui ont été exprimés pour la production des annotations manuelles et leur exploitation :

- Une mention d'entité se compose d'un ou plusieurs mots consécutifs d'une même phrase, et ne peut être recouverte par une autre. En chimie, un nom de matériau pouvant être formé des noms de ses composants, comme dans « ABS/ZnO », le choix a été fait d'annoter l'expression complète comme « Chemical », l'identification de ses sous-composants « ABS » et « ZnO » étant transférée à des post-traitements. Quand la phrase ne contient qu'une expression peu spécifique, comme « these composites » dans l'extrait 1, cet élément doit être annoté comme un composant chimique. Dans l'extrait 3, l'annotation de « the set at v_p = 5 mm/s » est délicate, l'expression désignant un échantillon de type « Chemical » tout en contenant des informations de type « Condition ». Concernant les valeurs, nous avons fait le choix de fusionner dans « Value » la valeur numérique et l'unité, comme « 44 MPa » ou « 50% », afin de simplifier l'annotation manuelle. Dans certains cas (extrait 2), l'unité n'est précisée qu'une fois pour plusieurs valeurs et doit être récupérée via des post-traitements. Par ailleurs, les valeurs complexes telles que « 1.9–2.0 GPa » (extrait 3) doivent être annotées comme « Value » et gérées plus finement avec des post-traitements.

- Une relation est binaire et relie deux entités qui appartiennent à une même phrase. Cela permet d'obtenir un premier ensemble d'informations. Des post-traitements de type fusion seront apportés d'une part pour relier des informations issues de différentes phrases, d'autre part pour réattribuer les bonnes valeurs aux informations multiples sous forme de listes comme dans la phrase 2 : « 24.19 » correspond à « 75% » de la condition « infill density ».

5 Les résultats obtenus

Après l'annotation de chaque extrait textuel par un ou plusieurs annotateur(s), 288 extraits annotés ont été obtenus, certains étant peu annotés et sans forcément d'homogénéité entre les annotateurs. Pour quantifier les résultats obtenus, dans un sous-ensemble de 58 extraits distincts bien annotés, contenant environ 270 phrases, on a obtenu environ 380 mentions de « Chemistry » et de « Property » et environ 600 mentions de « Condition » et de « Value ». Côté relations, on a observé environ 200 mentions de « has_value » et « with_condition », et environ 40 mentions de « is_property_of » et « measured_at ».

Des post-traitements ont été ajoutés afin d'enrichir les annotations manuelles avec les unités complétant les valeurs via une relation « has_unit » par exemples, et pour améliorer l'extraction d'informations suite à des annotations automatiques.

Références

Aw Y. Y., Yeoh C. K., Idris M. A., Teh P. L., Hamzah K. A., Sazali S. A. (2018). Effect of Printing Parameters on Tensile, Dynamic Mechanical, and Thermoelectric Properties of FDM 3D Printed CABS/ZnO Composites. Materials (Basel). 2018 Mar 22;11(4):466.

Dul S, Fambri L, Pegoretti A. (2018) Filaments Production and Fused Deposition Modelling of ABS/Carbon Nanotubes Composites. Nanomaterials (Basel). 2018 Jan 18;8(1):49. doi: 10.3390/nano8010049.

Goujon B. (2021). Extraction d'informations spécifiques à partir de textes avec peu de textes d'apprentissage, in *TextMine2021*, Montpellier.

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2015). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. DOI: 10.1093/nar/gkv1031

Islamaj R., Leaman R., Cissel D., Coss C., Denicola J., Fisher C., Guzman R., Gokal Kochar P., Miliaras N., Punske Z., Sekiya K., Trinh D., Whitman D., Schmidt S., Lu Z. (2022). NLM-Chem-BC7: manually annotated full-text resources for chemical entity annotation and indexing in biomedical articles, Database, Volume 2022, baac102. DOI: 10.1093/database/baac102

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R. and Gurevych, I. (2018): The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA.

Krallinger, M., Rabal, O., Leitner, F. et al. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 7 (Suppl 1), S2 (2015). DOI: 10.1186/1758-2946-7-S1-S2

Verbeeten, W.M.H.; Arnold-Bik, R.J.; Lorenzo-Bañuelos, M. (2021). Print Velocity Effects on Strain-Rate Sensitivity of Acrylonitrile-Butadiene-Styrene Using Material Extrusion Additive Manufacturing. Polymers 2021, 13, 149. DOI: 10.3390/polym13010149

Wei C.-H., Peng Y., Leaman R., Davis A. P., Mattingly C. J., Li J., Wiegers T. C. , Lu Z. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task, Database, Volume 2016, 2016, baw032. DOI: 10.1093/database/baw032

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Classification de relation pour la génération de mots-clés absents

Maël Houbre Florian Boudin Béatrice Daille
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
prénom.nom@univ-nantes.fr

RÉSUMÉ

Les modèles encodeur-décodeur constituent l'état de l'art en génération de mots-clés. Cependant, malgré de nombreuses adaptations de cette architecture, générer des mots-clés absents du texte du document est toujours une tâche difficile. Cette étude montre qu'entraîner au préalable un modèle sur une tâche de classification de relation entre un document et un mot-clé, permet d'améliorer la génération de mots-clés absents.

ABSTRACT

Relation classification for absent keyphrase generation

Encoder-decoder models are the current state of the art for keyphrase generation. However, despite numerous adaptations of this architecture, generating keyphrases that are absent from the source text is still a difficult task. This study shows that training a model on predicting the relation between a document and a keyphrase improves absent keyphrase generation.

MOTS-CLÉS : Génération de mots-clés absents, classification de séquence, modèle encodeur-décodeur, indexation.

KEYWORDS: Absent keyphrase generation, sequence classification, encoder-decoder model, indexing.

1 Introduction

La génération de mots-clés consiste à générer un ensemble de mots ou expressions (le terme "*mot-clé*" est utilisé dans les deux cas) représentant les points d'intérêt d'un document. Ceux-ci sont utilisés pour différentes tâches telles que le résumé automatique (Zha, 2002; Wan *et al.*, 2007; Qazvinian *et al.*, 2010; Pasunuru & Bansal, 2018) ou l'indexation (Harter, 1975; Barker *et al.*, 1972). Ces mots-clés peuvent être présents dans le texte source du document ou absents de celui-ci. Ils sont alors appelés mots-clés présents (respectivement absents). Les mots-clés absents apportent une plus-value en enrichissant l'indexation des documents scientifiques (Boudin & Gallina, 2021). Contrairement à son équivalent extractif, la génération de mots-clés a la particularité de permettre de générer ces mots-clés absents. La génération de mots-clés a été introduite avec l'utilisation de l'architecture encodeur-décodeur (Meng *et al.*, 2017). Cependant, les mots-clés absents résultant d'une abstraction, leur génération est particulièrement difficile, et ce malgré de nombreuses adaptations du modèle encodeur-décodeur (Yuan *et al.*, 2020; Meng *et al.*, 2021; Chen *et al.*, 2020; Bahuleyan & El Asri, 2020; Ye *et al.*, 2021).

Plutôt que d'essayer d'améliorer l'architecture, plusieurs travaux se sont concentrés sur l'utilisation

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 47–53.

de tâches support pour améliorer la représentation du texte avec des modèles pré-entraînés (Yasunaga *et al.*, 2022; Kulkarni *et al.*, 2022; Wu *et al.*, 2022). Cependant, aucun de ces travaux n'a travaillé sur la représentation de l'ensemble des mots-clés (présents et absents). Nous inspirant de ces travaux, nous introduisons une nouvelle tâche support visant à améliorer l'encodage des mots-clés et notamment des mots-clés absents.

Les contributions de cette étude sont comme suit :

- des travaux préliminaires sur une nouvelle tâche support à la génération de mots-clés : la prédiction de relation entre un mot-clé et un document.
- deux modèles pré-entraînés, utilisables avec la librairie transformers [1].

2 Méthodologie

Notre approche repose sur l'utilisation de deux affinages successifs d'un modèle pré-entraîné ; un premier affinage sur une tâche de classification suivi d'un affinage sur la génération de mots-clés. Le but de notre méthode est d'apprendre au modèle à mieux représenter la relation entre un document et ses mots-clés. En demandant au modèle de reconnaître les mots-clés auteur d'un document, nous avons pour objectif d'améliorer la représentation entre un mot-clé et son document par l'encodeur. De précédents travaux se sont appuyés sur des tâches support afin d'insister sur certains aspects importants du document tels que certains passages ciblés ou les mots-clés présents (Wu *et al.*, 2022; Kulkarni *et al.*, 2022). Cependant, aucun n'utilise l'ensemble complet des mots-clés (présents et absents) dans ces différentes tâches.

Les travaux à l'origine du modèle LinkBERT (Yasunaga *et al.*, 2022) améliorent significativement l'encodeur BERT en utilisant un graphe document-hyperlien. Ce graphe permet de récupérer des textes issus de documents liés dans le graphe pour enrichir les données d'entraînement. Le modèle est ensuite entraîné à déterminer si un passage est modifié avec du contenu provenant du document lui-même, d'un document lié ou d'un document aléatoire. Nos travaux s'inscrivent dans cette ligne en utilisant également une tâche de classification de séquence mais où le lien entre les documents est basé sur les mots-clés. Nous entraînons ensuite le modèle obtenu sur la génération de mots-clés.

Dans cet article, la tâche de classification consiste à déterminer si la séquence contient un mot-clé auteur, un mot-clé lié ou un mot-clé aléatoire. Les mots-clés auteurs d'un article sont les mot-clés attribués par les auteurs de celui-ci. Pour chaque mot-clé Y d'un document D_1, nous déterminons les documents D_k qui possèdent aussi ce mot-clé. Les mots-clés de D_k autres que Y sont appelés mots-clés liés. La figure 1 représente un exemple. Le mot-clé en bleu et rouge est un mot-clé partagé entre les documents D1 et D2. Les mots-clés bleus sont donc les mots-clés liés du document D1. Les mots-clés aléatoires (en vert sur la figure 1) sont pris aléatoirement dans l'ensemble restant des mots-clés du corpus (i.e sans les mots-clés auteur du document et ses mots-clés liés).

Pour la tâche de classification, la séquence à classifier est constituée de deux sous-séquences ; un mot-clé, puis la concaténation du titre et du résumé du document. Le mot clé est soit un mot-clé auteur du document, soit un mot-clé lié, soit un mot-clé aléatoire. Le modèle doit classifier la séquence selon trois étiquettes "auteur, lié, aléatoire". L'hypothèse est qu'à l'instar de LinkBERT, entraîner le modèle à faire cette distinction améliorera l'encodeur et permettra d'avoir une meilleure représentation entre les mots-clés et leurs documents associés.

1. https://huggingface.co/docs/transformers/index

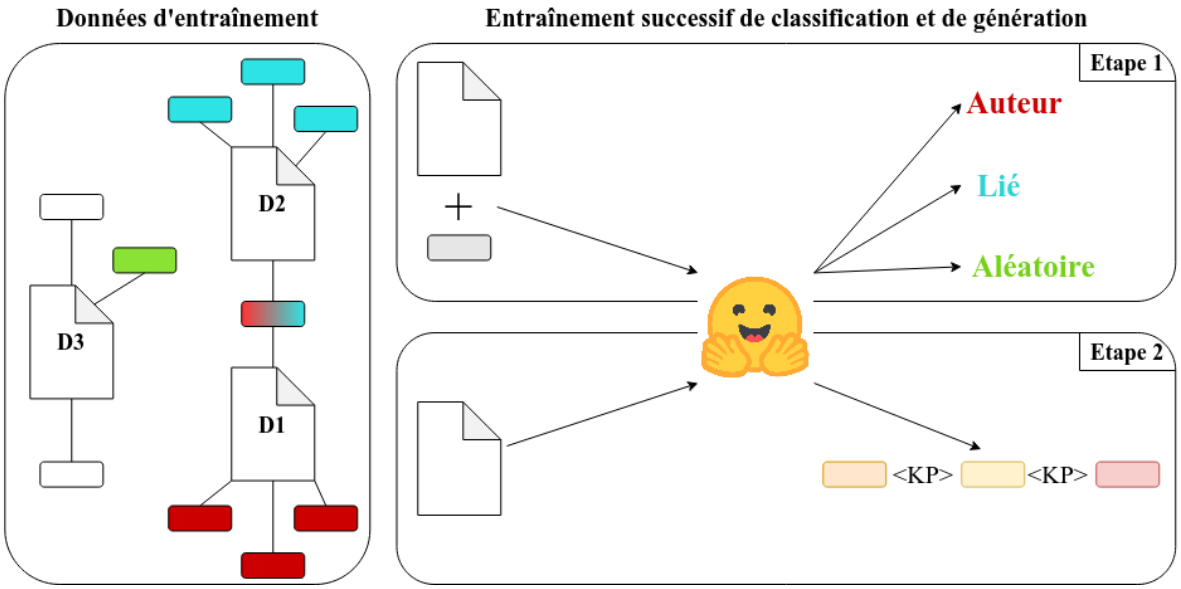


FIGURE 1 – Illustration de l'approche. Les mots-clés de D2 non communs avec D1 sont dits "liés".

Notre approche s'appuie sur le modèle BART (Lewis *et al.*, 2020). Ce modèle génératif avec une architecture encodeur-décodeur a déjà été utilisé pour la génération de mots-clés (Chowdhury *et al.*, 2022; Houbre *et al.*, 2022; Meng *et al.*, 2022) et notamment dans des travaux avec des tâches support (Kulkarni *et al.*, 2022; Wu *et al.*, 2022).

3 Expériences

Nous entraînons le modèle sur le dataset KP20k (Meng *et al.*, 2017) pour les deux tâches (classification et génération). Ce dataset contient 530 000 documents scientifiques en anglais dans le domaine des sciences informatiques issus de la bibliothèque numérique ACM Digital Library [2]. Chaque document est annoté avec en moyenne 5 mots-clés. En plus de ces mots-clés auteurs, nous ajoutons les mots-clés liés et les mots-clés aléatoires. Nous ajoutons au maximum 5 mots-clés liés et 5 mots-clés aléatoires à chaque document. Pour la classification, chaque séquence ne comporte qu'un seul mot-clé. L'ensemble d'entraînement est constitué de 8,1 millions de paires document/mot-clé. Le modèle est entraîné pendant 4 époques sur 4 cartes graphiques V100 32Go. L'entraînement de la tâche support prend 96 heures. Le modèle est ensuite entraîné sur la génération de mots-clés. Nous utilisons le paradigme One2Seq (Meng *et al.*, 2021) qui consiste pour un document en entrée, à générer l'ensemble des mots-clés dans une unique séquence. La séquence de mots-clés de référence est composée des mots-clés présents dans leur ordre d'apparition dans le texte, suivis des mots-clés absents dans l'ordre donnée par l'auteur puis des mots-clés liés et aléatoires. Le modèle est entraîné pendant 10 époques. L'entraînement de la génération de mots-clés prend 9 heures.

Conformément aux travaux à l'état de l'art, nous distinguons l'évaluation de la génération des mots-clés présents et absents. Pour les mots-clés présents, nous utilisons la F1@M et la F1@10. La F1@M (respectivement F1@10) est la f-mesure appliquée sur la première séquence de mots-clés générée par le modèle (respectivement les top 10 mots-clés générés). Pour la génération de mots-clés

2. https://dl.acm.org/

absents, nous utilisons le rappel appliqué au top 10 mots-clés générés (R@10). Pour obtenir 10 mots-clés, nous utilisons une recherche par faisceau et générons 20 séquences. Nous prenons ensuite les 10 premiers mots-clés en enlevant les répétitions. Si nous ne pouvons pas obtenir 10 mots-clés uniques, la séquence est complétée par autant d'unités "<UNK>" que nécessaire. Avant d'effectuer la comparaison, les mots-clés de la référence et les mots-clés prédits sont racinisés avec l'algorithme de Porter. La significativité statistique des résultats est vérifiée par un test de student avec $p < 0.05$.

4 Résultats et discussion

Le tableau 1 détaille les résultats des expériences sur l'ensemble de test de KP20k. Le symbole †représente une différence statistiquement significative entre les résultats des deux modèles.

Modèle	F1@M	F1@10	R@10
BART	31.4	28.5	4.7
BART classif (notre approche)	31.5	28.3	5.0†

TABLE 1 – Résultats de la génération de mots-clés présents (F1@10 et F1@M) et absents (R@10)

Les résultats du tableau 1 montrent que pour la génération de mots-clés présents, il n'y a pas de différence significative entre le modèle avec tâche support *BART classif* et celui entraîné uniquement sur la génération de mots-clés. Concernant la génération de mots-clés absents, les performances du modèle *BART classif* sont meilleures que celles du modèle BART uniquement entraîné sur la génération de mots-clés. Cependant, bien que la différence soit statistiquement significative, l'amélioration n'est que de 6.4% relatifs. La différence entre les résultats ne dépassant pas les 10%, nous qualifions plutôt cette amélioration de perceptible plutôt que de significative (Sparck Jones, 1974).

L'une des hypothèses pour lesquelles la génération de mots-clés présents n'est pas améliorée est que la tâche de classification implique un plus grand nombre de mots-clés absents que de présents. Avec 5 mots-clés liés et 5 mots-clés aléatoires supplémentaires par document, le ratio entre mots-clés présents et mots-clés absents est inversé. Affiner le choix et le nombre des mots-clés liés et aléatoires est une première voie d'amélioration. La forme de la séquence donnée en entrée de la classification peut également être une des raisons de ces faibles performances par rapport à l'état de l'art sur des architectures dédiées à la génération de mots-clés. En effet, les travaux de (Kulkarni *et al.*, 2022) et (Wu *et al.*, 2022) portaient sur le remplacement d'éléments du texte source. Ceci fournissait ainsi un contexte à l'encodeur pour distinguer le contenu "étranger" au document. Dans notre étude, le mot-clé est introduit en tout début de séquence sans contexte supplémentaire. Améliorer le paradigme pour la classification fera l'objet de prochains travaux. Une autre hypothèse est que l'utilisation de deux affinages successifs ne permet pas de tirer pleinement profit de la tâche support. L'utilisation d'un entraînement multi-tâche est une voie à explorer.

5 Conclusion

Dans cette étude, nous avons proposé une nouvelle tâche support visant à améliorer la génération de mots-clés. Nous entraînons un modèle à distinguer si un mot-clé est associé par l'auteur, extrait

de l'ensemble des mots-clés d'un document lié (i.e avec lequel le document partage un ou plusieurs mots-clés) ou aléatoire. Nous avons montré qu'un modèle BART entraîné sur cette tâche avant la génération de mots-clés, voyait ses performances significativement améliorées pour la génération de mots-clés absents. Cependant, la tâche support augmente drastiquement le temps d'entraînement pour des résultats qui ne dépassent pas l'état de l'art. De futurs travaux se concentreront sur l'amélioration de la définition de la tâche support, ainsi que son utilisation pour d'autres modèles génératifs.

6 Remerciements

Ce travail s'inscrit dans le cadre du projet ANR DELICES (ANR-19-CE38-0005) et a été effectué en utilisant les resources de calcul de GENCI-IDRIS (dossier 2022-[AD011013670]).

Références

BAHULEYAN H. & EL ASRI L. (2020). Diverse keyphrase generation with neural unlikelihood training. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5271–5287, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.462.

BARKER F., VEAL D. & WYATT B. (1972). COMPARATIVE EFFICIENCY OF SEARCHING TITLES, ABSTRACTS, AND INDEX TERMS IN A FREE-TEXT DATA BASE. *Journal of Documentation*, **28**(1), 22–36. DOI : 10.1108/eb026527.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BOUDIN F. & GALLINA Y. (2021). Redefining Absent Keyphrases and their Effect on Retrieval Effectiveness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4185–4193, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.330.

CHEN W., CHAN H. P., LI P. & KING I. (2020). Exclusive hierarchical decoding for deep keyphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1095–1105, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.103.

CHOWDHURY M. F. M., ROSSIELLO G., GLASS M., MIHINDUKULASOORIYA N. & GLIOZZO A. (2022). Applying a Generic Sequence-to-Sequence Model for Simple and Effective Keyphrase Generation. *arXiv :2201.05302 [cs]*. arXiv : 2201.05302.

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

HARTER S. P. (1975). A probabilistic approach to automatic keyword indexing. Part I. On the Distribution of Specialty Words in a Technical Literature. *Journal of the American Society for Information Science*, **26**(4), 197–206. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.4630260402, DOI : 10.1002/asi.4630260402.

HOUBRE M., BOUDIN F. & DAILLE B. (2022). A large-scale dataset for biomedical keyphrase generation. In *Proceedings of the 13th International Workshop on Health Text Mining and Infor-*

mation Analysis (LOUHI), p. 47–53, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

KULKARNI M., MAHATA D., ARORA R. & BHOWMIK R. (2022). Learning Rich Representation of Keyphrases from Text. In *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 891–906, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.67.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

MENG R., WANG T., YUAN X., ZHOU Y. & HE D. (2022). General-to-Specific Transfer Labeling for Domain Adaptable Keyphrase Generation. arXiv :2208.09606 [cs].

MENG R., YUAN X., WANG T., ZHAO S., TRISCHLER A. & HE D. (2021). An Empirical Study on Neural Keyphrase Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4985–5007, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.396.

MENG R., ZHAO S., HAN S., HE D., BRUSILOVSKY P. & CHI Y. (2017). Deep Keyphrase Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 582–592, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1054.

PASUNURU R. & BANSAL M. (2018). Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 646–653, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2102.

QAZVINIAN V., RADEV D. R. & ÖZGÜR A. (2010). Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 895–903, Beijing, China : Coling 2010 Organizing Committee.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SPARCK JONES K. (1974). AUTOMATIC INDEXING. *Journal of Documentation*, **30**(4), 393–432. Publisher : MCB UP Ltd, DOI : 10.1108/eb026588.

WAN X., YANG J. & XIAO J. (2007). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 552–559, Prague, Czech Republic : Association for Computational Linguistics.

WU D., AHMAD W., DEV S. & CHANG K.-W. (2022). Representation learning for resource-constrained keyphrase generation. In *Findings of the Association for Computational Linguistics :*

EMNLP 2022, p. 700–716, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

YASUNAGA M., LESKOVEC J. & LIANG P. (2022). LinkBERT : Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8003–8016, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.551.

YE J., GUI T., LUO Y., XU Y. & ZHANG Q. (2021). One2Set : Generating diverse keyphrases as a set. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 4598–4608, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.354.

YUAN X., WANG T., MENG R., THAKER K., BRUSILOVSKY P., HE D. & TRISCHLER A. (2020). One size does not fit all : Generating and evaluating variable number of keyphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7961–7975, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.710.

ZHA H. (2002). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, p. 113–120, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/564376.564398.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Le corpus *Machine Translation* Une exploration diachronique des (méta)données Istex

Mathilde Huguin[1] Sabine Barreaux[1]
(1) Inist - CNRS (UAR 76), 54 519 Vandœuvre-lès-Nancy, France
mathilde.huguin@inist.fr, sabine.barreaux@inist.fr

RÉSUMÉ

Le corpus *Machine Translation* se compose de publications scientifiques issues du réservoir Istex. Conçu comme un cas d'usage, il permet d'explorer l'histoire de la traduction automatique au travers des métadonnées et des textes intégraux disponibles pour chacun de ses documents. D'une part, les métadonnées permettent d'apporter un premier regard sur le paysage de la traduction automatique grâce à des tableaux de bord bibliométriques. D'autre part, l'utilisation d'outils de fouille de textes sur le texte intégral rend saillantes des informations inaccessibles sans une lecture approfondie des articles. L'exploration du corpus est réalisée grâce à Lodex, logiciel open source dédié à la valorisation de données structurées.

ABSTRACT

The *Machine Translation* Corpus. A diachronic exploration of Istex (meta)data

The *Machine Translation* corpus consists of scientific publications from the Istex repository. Conceived as a use case, it allows one to explore the history of machine translation through the metadata and full texts available for each of its documents. On the one hand, the metadata provide a first look at the machine translation landscape through bibliometric dashboards. On the other hand, the use of text mining tools on the full text brings out information that is inaccessible without a thorough reading of the articles. The exploration of the corpus is carried out using Lodex, an open source software dedicated to the valorisation of structured data.

MOTS-CLÉS : Traduction Automatique, Corpus, Istex, Bibliométrie, Fouille de textes.

KEYWORDS: Machine Translation, Corpus, Istex, Bibliometrics, Text Mining.

1 Introduction

L'atelier que nous proposons poursuit deux objectifs : **(i)** présenter la ressource Istex [1] (*Initiative d'excellence en Information Scientifique et Technique*), qui permet de construire des corpus de publications scientifiques, et **(ii)** montrer que les formats et outils accessibles depuis Istex offrent des solutions profitables pour la fouille de textes. Notre démonstration trouve son origine dans la dynamique nationale (ex. ANR-22-MaTOS-0033 [2] ; Fiorini *et al.*, 2020) et européenne actuelle (ex.

1. Istex (www.istex.fr) est né en 2011 dans le cadre des Programmes d'Investissement d'Avenir (ANR-10-IDEX-0004-02).
2. https://anr-matos.github.io

OPERAS [3] ; Helsinki-Initiative, 2019) autour des nouvelles méthodes de traduction automatique (désormais TA). Istex renfermant plus de sept siècles d'archives scientifiques, nous avons choisi d'explorer l'aspect diachronique de la TA à travers un corpus. Nous montrons, d'une part, comment l'analyse des métadonnées des publications donne accès à un panorama des travaux réalisés dans ce domaine et, d'autre part, en quoi l'analyse du texte intégral contribue à retracer l'histoire des méthodes et des approches mises en œuvre en TA.

Notre article s'organise comme suit. En 2, nous présentons le contenu d'Istex et les outils que nous manipulons. En 3, nous expliquons la méthodologie appliquée pour constituer le corpus et présentons quelques-uns des résultats obtenus.

2 L'écosystème Istex

Le réservoir Istex contient plus de 27 millions de publications scientifiques dans toutes les disciplines et dans plus de 50 langues. Il est alimenté en continu au travers des acquisitions pérennes des licences nationales [4] et de celles du GIS CollEx-Persée [5]. En 2023, il regroupe plus de 9 000 revues et 430 000 ebooks de 41 éditeurs différents, publiés entre le XIV^{e} siècle et aujourd'hui. Ces documents sont accessibles depuis l'interface Istex-DL [6] (*Istex Download*), connectée à une API [7], qui permet de télécharger les publications et de choisir les formats appropriés. En effet, si le réservoir constitue d'abord une ressource documentaire, Istex n'est pas exclusivement un outil de consultation. Il offre la possibilité de constituer des corpus à des fins de fouille de textes et d'analyse de contenu (ex. Bordignon & Maisonobe, 2022). Dans ce cas, il s'agit non seulement de rechercher des documents, mais aussi de vérifier leurs propriétés, de les télécharger massivement, pour finalement effectuer des traitements seuls ou les intégrer dans des outils. Tous les documents Istex sont disponibles à la fois sous forme de métadonnées et dans leur version en texte intégral, et ce, dans différents formats soit d'origine, soit convertis dans des standards (XML MODS et XML TEI) pour faciliter leur exploitation dans les outils de fouille ou de textométrie (ex. TXM Heiden *et al.*, 2010). Les documents Istex ont également été enrichis grâce à des outils, développés ou adaptés pour Istex (Cuxac & Thouvenin, 2017). Parmi ces enrichissements, nous exploiterons (§3.2) plus particulièrement la structuration des références bibliographiques obtenue avec l'outil GROBID [8] (*GeneRation Of Bibliographic Data*).

Lors du téléchargement, Istex-DL propose une passerelle vers l'outil Lodex [9] (*Linked Open Data EXperiment*), que nous utiliserons pour naviguer dans le contenu du corpus. Lodex est un logiciel open source [10] créé pour les besoins du projet Istex afin de valoriser ses données structurées (Gregorio *et al.*, 2019). Il permet de concevoir des sites web offrant des interfaces pour explorer visuellement un jeu de données (CSV, JSON, etc.) au travers de tableaux de bord dynamiques présentant des indicateurs bibliométriques. Lodex offre également la possibilité d'utiliser des services web [11] afin d'enrichir les documents à l'aide de programmes d'analyse, de curation, d'annotation et d'indexation.

3. https://operas-eu.org/projects/translations-and-open-science
4. https://www.licencesnationales.fr
5. https://www.collexpersee.eu
6. https://dl.istex.fr
7. https://api.istex.fr/document/?q=
8. https://github.com/kermitt2/grobid
9. https://lodex.inist.fr
10. https://github.com/Inist-CNRS/lodex
11. https://objectif-tdm.inist.fr/category/services

3 Le corpus *Machine Translation*

3.1 Méthodologie de constitution de corpus

L'élaboration du corpus *Machine Translation* suit une procédure itérative (de Salabert & Barreaux, 2020). La requête interrogeant Istex utilise la syntaxe Lucene [12] (pour les non-initiés, il est possible d'utiliser l'outil de recherche assistée). Notre requête initiale filtre les documents comportant des mots-clés relatifs à la TA dans tous les champs. Les documents sont téléchargés via Istex-DL, puis importés dans Lodex. Comme nous le montrerons lors de l'atelier, la visualisation et l'analyse de ces données aident à détecter le bruit et/ou le silence et amènent à une révision de la requête initiale. La requête corrigée cible les champs dans lesquels nous cherchons les mots-clés (ex. *title :"machine translation"*), restreint les langues cibles (ex. *language.raw :"eng" "fre"*) et exclut certains éditeurs provoquant du bruit (ex. *NOT corpusName :"nature"*). Dans sa version finale, le corpus *Machine Translation* se compose de 7 160 documents en anglais et en français, soit plus de 54 millions de mots. Il est accessible suivant ce lien : `http://traduction-machinetranslation.corpus.istex.fr`.

3.2 Indicateurs bibliométriques

Selon les choix de calculs, de curations ou de visualisations opérés, les métadonnées et les enrichissements apportés par Istex fournissent plus d'une quinzaine d'indicateurs bibliométriques. La Figure 1 présente, par exemple, les dix revues les plus fréquentes dans notre corpus. On y retrouve ainsi *Machine Translation*, la principale revue de TA.

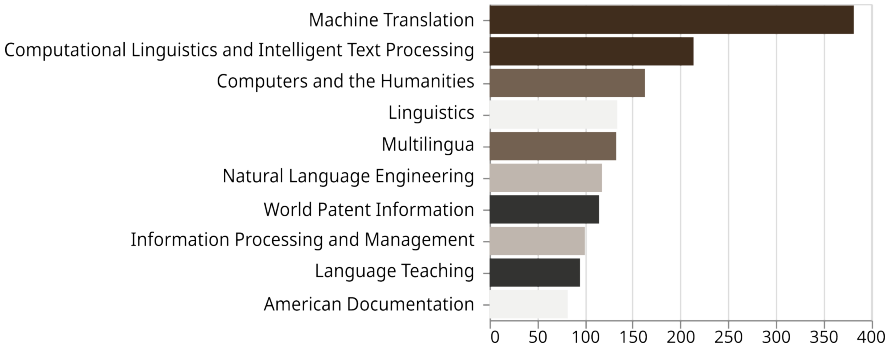


FIGURE 1 – Les dix revues majoritaires dans le corpus *Machine Translation*

Nous focaliserons notre présentation sur deux indicateurs bibliométriques obtenus grâce aux enrichissements et à l'utilisation de services web sur les métadonnées fournies par Istex.

(a) Les champs d'affiliations des auteurs sont restructurés et normalisés pour obtenir une cartographie des pays publiants, cf. Figure 2. Pour ce faire, nous utilisons deux services web développés par l'Inist et appelés depuis Lodex. Le service web de découpage d'adresses [13] retourne une adresse au format texte en tableau de champs. Le service web exploitant le thésaurus Loterre des noms de Pays [14] normalise les graphies des pays (regroupant par exemple *USA* et *États-Unis*). Comme attendu, les

12. `www.elastic.co`
13. `https://gitbucket.inist.fr/tdm/web-services`
14. `https://skosmos.loterre.fr/9SD/fr`

États-Unis sont le premier pays publiant sur la TA (en noir dans la figure), suivis de la Chine et du Japon qui sont les deux premiers co-publiants des États-Unis.

(b) La structuration des références bibliographiques permet notamment de détecter les auteurs et revues les plus cités, cf. Figure 2. Ces indications fournissent indirectement des éléments historiques sur la TA. L'auteur le plus cité est *Philipp Koehn*, considéré comme l'un des inventeurs de la méthode *phrase based* (Koehn *et al.*, 2003) et développeur de l'outil Moses (Koehn *et al.*, 2007). La détection renvoie également *Warren Weaver* et *Adrew Booth*, instigateurs des premières idées pour traduire les langues naturelles (Hutchins, 2007), ou encore *Peter Toma* développeur de Systran.

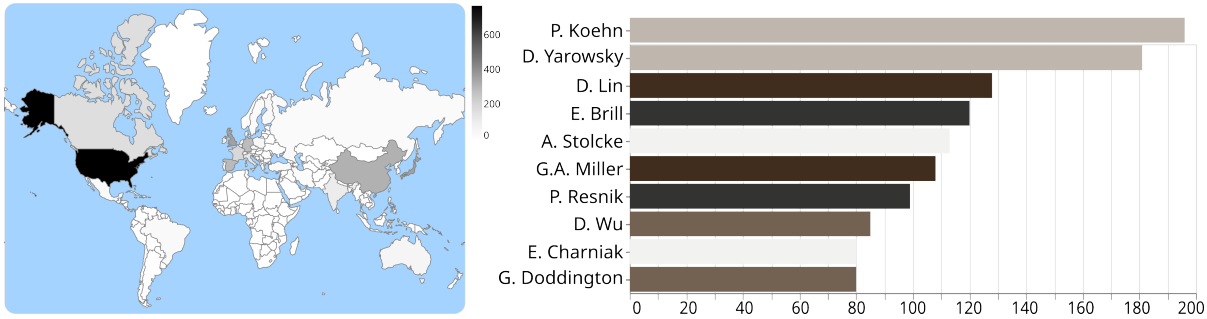


FIGURE 2 – Focus : **(a)** pays publiants et **(b)** auteurs citées

3.3 Diachronie de la TA

Nous nous servons de l'annotation du texte intégral pour retracer l'histoire de la TA et l'évolution des méthodes et des outils. Pour aboutir à ce résultat, nous avons initié une expérience en construisant une ressource terminologique bilingue d'environ 60 termes avec leurs variantes (sigles, alias). Ces termes désignent des modèles, outils, ou techniques de TA (ex. *lexical-functional grammar*, *Moses*). Ils sont regroupés selon leur appartenance à trois approches *Rule Based Machine Translation* (RBMT), *Statistical Machine Translation* (SMT) et *Neural Machine Translation* (NMT) (Hutchins, 2000, 2007).

La ressource terminologique est projetée à la fois sur les textes intégraux mais aussi sur les métadonnées (XML TEI) grâce à une feuille de style XSLT. Cette double annotation nous permet de vérifier l'apport de l'utilisation du texte intégral, cf. Tableau 1. Par rapport à une annotation des seules métadonnées, l'annotation du texte intégral permet d'obtenir 5 fois plus de documents dans lesquels des termes de la ressource sont détectés. Près de 4 200 documents sont ainsi catégorisés selon l'approche de TA utilisée [15].

	Métadonnées	Textes intégraux
Avec termes détectés	878	4 170
Sans termes détectés	6 282	2 990

TABLE 1 – Nombre de documents annotés selon l'input de l'annotation

Les occurrences des termes détectés dans les textes intégraux sont ensuite utilisées pour construire un graphique de flux montrant l'évolution temporelle des approches au sein des documents, cf. Figure 3.

15. Certains documents, hybrides, sont associés à plusieurs approches.

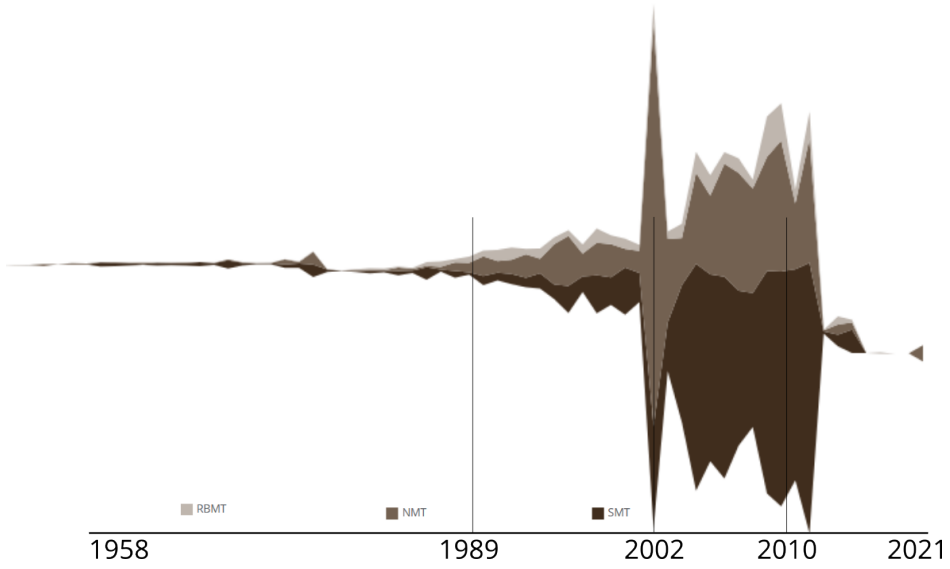


FIGURE 3 – Graphique de flux des approches en TA

La présence de certains termes dans plusieurs approches (ex. *Systran*) explique l'apparition de la NMT dès les années 80. Le graphique témoigne d'un essor de la TA dans les années 80 qui coïncide également avec l'invention de l'*example-based machine translation* (Song, 2022). La recherche en TA s'intensifie réellement depuis les années 90 (ce qui correspond à la création du modèle IBM de SMT). Deux valeurs sont particulièrement remarquables : le pic de 2002 coïncide peu ou prou avec la date de lancement du premier système de traduction sur internet et celui de 2010 à l'essor de la NMT.

4 Conclusion

À travers un cas d'usage, notre atelier montre qu'Istex est une ressource puissante pour créer des corpus de publications scientifiques et pour exploiter toute la richesse des métadonnées et du texte intégral des publications. L'écosystème Istex, dans sa globalité, offre une boîte à outils pour transformer, nettoyer, enrichir et visualiser ses données en vue de les analyser plus finement.

Pour aller plus loin, le corpus *Machine Translation* pourra (i) être enrichi avec des documents provenant d'autres sources, afin d'étoffer sa couverture chronologique, et (ii) être utilisé pour repérer automatiquement des définitions afin de compléter la ressource terminologique sur la TA.

Remerciements

Nous remercions chaleureusement la mastérante en traduction Manon Delorme pour son soutien dans la constitution de la ressource terminologique de TA et le responsable de ressources terminologiques Majid Khayari pour son aide technique (plus que précieuse) dans l'annotation du corpus *Machine Translation*.

Références

BORDIGNON F. & MAISONOBE M. (2022). Researchers and their data : A study based on the use of the word data in scholarly articles. *Quantitative Science Studies*, **3**(4), 1156–1178. DOI : 10.1162/qss_a_00220.

CUXAC P. & THOUVENIN N. (2017). Archives numériques et fouille de textes : le projet ISTEX. In P. CUXAC, V. LEMAIRE & J.-C. LAMIREL, Éds., *Atelier TextMine - EGC 17*, p. 43–51, Grenoble, France.

DE SALABERT C. & BARREAUX S. (2020). Vers un corpus optimal pour la fouille de textes : stratégie de constitution de corpus spécialisés à partir d'ISTEX. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*.

FIORINI S., BARBIN F., GARNIER-RIZET M., MORIN K. H., HUMPHREYS F., JOSSELIN-LERAY A., KÜBLER N., LOOCK R., MARTIKAINEN H., NOMINÉ J.-F., PLAG C., ROSSI C. & YVON F. (2020). *Rapport du groupe de travail "Traductions et science ouverte"*. report, Comité pour la science ouverte. Pages : 44 p., DOI : 10.52949/20.

GREGORIO S., COLLIGNON A., PARMENTIER F. & THOUVENIN N. (2019). LODEX : des données structurées au web sémantique. In *Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019)*, Metz, France.

HEIDEN S., MAGUÉ J.-P. & PINCEMIN B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *Statistical Analysis of Textual Data - Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, volume 2, p. 1021–1032, Rome, Italie : Edizioni Universitarie di Lettere Economia Diritto. Issue : 3.

HELSINKI-INITIATIVE (2019). *Helsinki Initiative on Multilingualism in Scholarly Communication*. Rapport interne, Federation of Finnish Learned Societies ; The Committee for Public Information ; Publishing, The Finnish Association for Scholarly ; Universities Norway ; European Network for Research Evaluation in the Social Sciences and the Humanities, Helsinki. Publisher : figshare.

HUTCHINS J. (2007). Machine translation : A concise history. *Computer aided translation : Theory and practice*, **13**(29-70), 11. Publisher : Chinese University of Hong Kong.

HUTCHINS J. W. (2000). Early years in machine translation. *Early Years in Machine Translation*, p. 1–411. Publisher : John Benjamins Publishing Company.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 177–180, Prague, Czech Republic : Association for Computational Linguistics.

KOEHN P., OCH F. J. & MARCU D. (2003). *Statistical phrase-based translation*. Rapport interne, University of Southern California Marina Del Rey Information Sciences Inst.

SONG R. (2022). Analysis on the Recent Trends in Machine Translation. *Highlights in Science, Engineering and Technology*, **16**, 40–47.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

CASIMIR : un Corpus d'Articles Scientifiques Intégrant les ModIfications et Révisions des auteurs

Léane Jourdan[1] Florian Boudin[1] Nicolas Hernandez[1] Richard Dufour[1]
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
prénom.nom@univ-nantes.fr

RÉSUMÉ

Écrire un article scientifique est une tâche difficile. L'écriture scientifique étant un genre très codifié, de bonnes compétences d'écriture sont essentielles pour transmettre ses idées et les résultats de ses recherches. Cet article décrit les motivations et les travaux préliminaires de la création du corpus CASIMIR dont l'objectif est d'offrir une ressource sur l'étape de révision du processus d'écriture d'un article scientifique. CASIMIR est un corpus des multiples versions de 26 355 articles scientifiques provenant d'OpenReview accompagné des relectures par les pairs.

ABSTRACT

CASIMIR : a Corpus of Scientific Articles Integrating the ModIfications et Revisions of authors

Writing a scientific article is a challenging task as it is a highly codified genre. Good writing skills are essential to convey ideas and the results of research work properly. This paper describes the motivations and first steps of the CASIMIR creation process. The objective is to offer a new resource for text revision on the revision step of scientific article writing process CASIMIR is a corpus of the multiple versions of 26 355 scientific papers from OpenReview with their associated peer reviews.

MOTS-CLÉS : corpus, jeu de données, articles scientifiques, openreview, relectures, révision de textes.

KEYWORDS: corpus, dataset, scientific articles, openreview, reviews, text revision.

1 Introduction

Le processus d'écriture d'un article scientifique est une tâche complexe et difficile, particulièrement pour les jeunes chercheurs qui doivent apprendre les conventions de l'écriture scientifique. C'est a fortiori vrai pour les chercheurs non natifs anglophones qui doivent également faire face à la barrière de la langue. Juniors ou seniors, tous doivent prêter une attention particulière à la qualité de leur écriture afin de transmettre efficacement leurs idées au lecteur.

L'écriture scientifique est un genre à part avec ses propres codes et spécificités : structure de l'article (format IMRaD : *Introduction, Methods, Results and Discussion* (Swales, 1990)), style concis et précis, usages des temps, des pronoms et de la terminologie (Kallestinova, 2011; Bourekkache, 2022).

Plusieurs propositions ont été faites à travers la littérature pour décrire le processus d'écriture d'un article scientifique (Silveira *et al.*, 2022; Laksmi, 2006; Bailey, 2014; Seow, 2002; Du *et al.*, 2022a). Toutes ces propositions partagent des étapes communes et peuvent être résumées comme : *Étape 1 : Pré-écriture*, *Étape 2 : Brouillon*, *Étape 3 : Révision* et *Étape 4 : Relecture de finitions*. Le

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 60–65.

présent corpus a vocation à être une ressource pour l'étape de *Révision* qui est réalisée en amont de la soumission par l'auteur lui-même, puis à la suite de la phase de relecture par les pairs, basée sur leurs suggestions. Elle consiste à faire des changements en profondeur dans le texte, sur le fond, la structure des phrases, la façon de connecter les idées. Cette étape est itérative (on la répète jusqu'à obtenir un résultat satisfaisant) (Du *et al.*, 2022a) et 1-vers-N (une section de texte peut avoir plusieurs révisions correctes (Ito *et al.*, 2019)). Apporter une aide automatique à cette étape du processus d'écriture permettrait aux auteurs d'améliorer plus rapidement et efficacement leurs textes.

Il existe actuellement peu de corpus pour cette tâche dont les plus semblables à notre travail sont donnés ci-après. 1- `PeerRead` (Kang *et al.*, 2018) est un corpus de 14 784 brouillons d'articles accompagnés de la décision de publication acceptée/rejetée. Ce corpus ne contient pas les versions finales des articles. 2- `IteraTeR` (Du *et al.*, 2022b) un corpus de 31 631 documents (toutes les versions incluses) dont 11 443 résumés provenant de ArXiv, le reste provenant de Wikipedia et Wikinews. Il contient toutes les révisions pour un même résumé, alignées phrase à phrase, mais il n'inclut pas les articles dans leur intégralité. 3- `arXivEdits` (Jiang *et al.*, 2022) composé de 751 articles provenant de ArXiv et leurs différentes versions, alignées phrase à phrase, pour un total de 1 790 documents.

Dans cet article, nous présentons la première étape de la création de `CASIMIR`, un corpus d'articles scientifiques en anglais accompagnés de leurs différentes versions et relectures faites par les pairs, collectés à partir de OpenReview [1]. Il aura une taille supérieure aux précédents corpus et proposera un alignement au niveau des paragraphes en plus de l'alignement au niveau phrase.Cette nouvelle ressource pourra par exemple être mise à profit dans l'entraînement de modèles répondant aux diverses tâches de l'assistance à l'écriture scientifique.

la révision de textes (Du *et al.*, 2022a), la correction orthographique, la prédiction d'acceptation/rejet d'un papier (Kang *et al.*, 2018), etc.

2 Création du corpus

Les articles sont collectés depuis OpenReview, une plateforme ouverte de relecture par les pairs qui permet d'héberger les différentes versions d'un même article au format PDF ainsi que ses relectures. Son avantage est de proposer des premières versions peu relues et donc d'avoir des révisions plus importantes au fil des re-soumissions (exemple en annexe A). De plus, le contenu des relectures par les pairs est un guide sur la qualité des articles associés et les intentions sous-jacentes aux révisions effectuées. Toutefois, OpenReview présente un inconvénient majeur : les articles ne sont disponibles qu'en version PDF et non en format LaTeX et devront être convertis vers un autre format, ici le XML.

On considèrera les termes suivants issus de la terminologie OpenReview :

— Un **forum** désigne l'espace de discussion et de dépôt attribué à un article.
— Une **relecture** désigne un commentaire écrit sur l'article : une relecture complète par un pair, une réponse de l'auteur ou du relecteur ou la décision finale sur la publication de l'article.

1. https ://openreview.net/

2.1 Méthodologie de collecte et première étape de filtration

L'objectif est de collecter exhaustivement l'ensemble des documents disponibles sur OpenReview au 10/03/2023. Pour y parvenir, nous utilisons l'API [2] fournie par la plateforme.

Le processus de collecte est présenté ci-dessous :

1. Collecte de la liste des événements (ateliers, conférences, etc) hébergés sur la plateforme.
2. Collecte, grâce à cette liste, de l'ensemble des identifiants de forums liés à chaque événement.
3. Collecte, grâce aux identifiants des forums, des métadonnées associées aux versions d'un même article et création d'un fichier de correspondance associant l'identifiant du forum (identifiant du papier final) à l'identifiant de ses versions antérieures.
4. Collecte, grâce aux identifiants des forums, des relectures (messages sur le forum) et création d'un fichier de correspondance associant l'identifiant du forum et celui de ses relectures.
5. Collecte, des PDF disponibles sur le site des différentes versions des articles.

390 Go de données sont collectées dont 730 invitations et 121 492 PDF pour 29 504 articles.

2.2 Conversion des PDF

Les fichiers PDF ne sont pas directement utilisables pour l'entraînement de modèles, il est nécessaire de les convertir vers un format approprié, ici le XML. Pour extraire le contenu des fichiers PDF tout en conservant leur structure, on utilise l'outil état de l'art Grobid(GRO, 2008 2023).
Après la conversion en XML, les citations, formules, figures et bibliographies seront retirées. Les fichiers PDF qui ne peuvent être convertis seront exclus du corpus et les articles qui n'ont plus qu'une seule version seront à nouveau filtrés.

Une première observation de la qualité de la conversion a été réalisée sur un sous-ensemble de documents déjà convertis. Ils comportent des erreurs telles que la mauvaise détection des tables et figures, la détection incorrecte de paragraphes en tant que figures, la suppression de portions de phrases, la mauvaise détection de sections, la retranscription des formules incluses dans les paragraphes, etc. Ces erreurs rendent l'alignement entre les différentes versions des articles plus complexe, car elles génèrent des différences supplémentaires entre les versions d'un article qui n'existent pas initialement.

3 Filtrage et description du corpus

Seuls les articles ayant au moins deux versions et dont les PDF et les métadonnées ont pu être collectés sont conservés. Le corpus résultant de cette première étape de filtrage comprend 26 355 articles et leurs versions antérieures (89.33% du nombre d'articles initial), pour un total de 118 415 documents (97.46% du nombre de documents initial). Il comprend également les métadonnées de chaque version

2. https ://openreview-py.readthedocs.io/en/latest/api.html

et les relectures associées. 37 conférences sont représentées dans les données (hors soumissions et challenges indépendants). Parmi les domaines les plus représentés, nous retrouvons l'apprentissage automatique (ICLR, ICML, NeurIPS), la robotique (RSS, CoRL), le traitement automatique des langues (ACL), la vision par ordinateur (ECCV), etc.

La distribution du nombre de versions antérieures et de relectures par article est présentée dans la Figure 1. La majorité des articles ont moins de 10 versions. Toutes les soumissions de l'auteur sur la plateforme sont comptabilisées comme versions, plusieurs peuvent donc être effectuées avant même les relectures ou présenter des différences mineures. Pour les relectures, tous les échanges sur le forum lié à un article sont comptabilisés, expliquant le nombre élevé de relectures pour certains articles. Cependant, il est possible de différencier les "réelles" relectures des autres messages en utilisant les attributs des métadonnées.

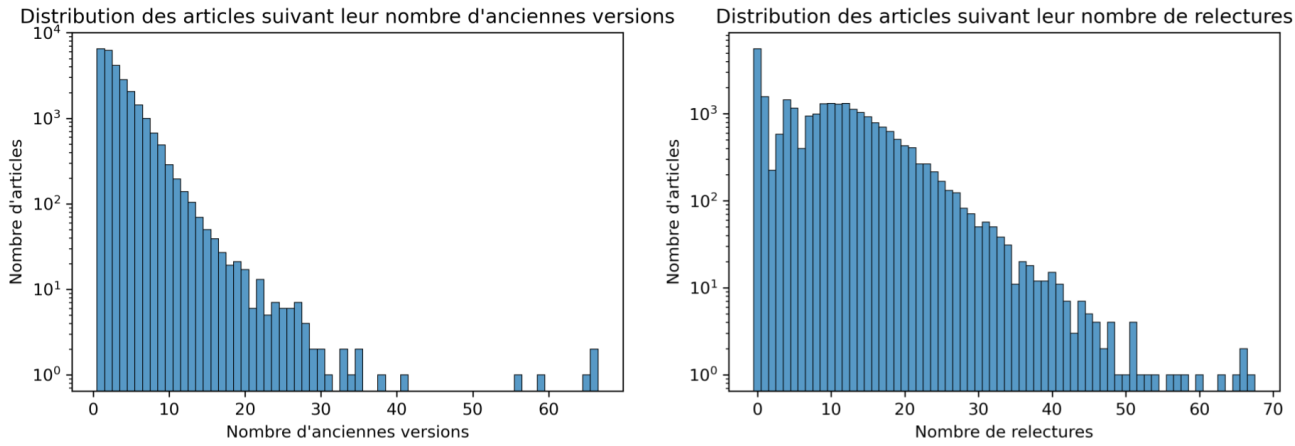


FIGURE 1 – Distribution du nombre de versions (gauche) et de relectures (droite) par article collectés

4 Discussion

Dans cet article, nous avons présenté la première étape de création du corpus de révisions d'article scientifiques CASIMIR. Lors de ces travaux préliminaires, plusieurs problèmes ont été rencontrés. Tout d'abord, des difficultés liées à OpenReview, certains fichiers PDF étaient manquants et certaines conférences ne contenaient aucun papier. Le corpus a donc été limité aux données disponibles au moment de la collecte.

Une autre difficulté rencontrée concerne la conversion des PDF évoquée en Section 2.2. Pour pallier ce problème, trois articles ont été sélectionnés aléatoirement, avec leur première et dernière version, soit six documents. Les versions XML générées par Grobid ont été corrigées manuellement pour ces six documents, avec un temps approximatif de 75 minutes par article. Ces articles annotés manuellement serviront de références pour évaluer la qualité de la conversion en XML générée par Grobid, ainsi que la dégradation de la qualité de l'alignement automatique sur les fichiers convertis automatiquement.

Pour poursuivre la création du corpus, l'ensemble des documents doit être converti en XML, puis les différentes versions des articles doivent être alignées paragraphe à paragraphe et phrase à phrase. Pour cela, nous pourrons nous reposer sur le modèle d'alignement de phrases proposé par (Jiang

et al., 2022) ainsi que leur algorithme d'alignement des paragraphes en l'améliorant pour tenir compte des fusions et divisions de paragraphes. Cela permettra d'extraire les révisions entre les différentes versions d'un article. Les trois articles de référence corrigés manuellement seront également alignés manuellement.

Enfin, les documents seront annotés selon une taxonomie de révisions à définir (exemples : clarté, grammaire, langage, style, etc) pour être utilisés pour l'entrainement de modèles de révision de texte.

Références

(2008–2023). Grobid. https://github.com/kermitt2/grobid.

AGLIONBY G. & TEUFEL S. (2022). Identifying relevant common sense information in knowledge graphs. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, p. 1–7, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.csrr-1.1.

BAILEY S. (2014). *Academic writing : A handbook for international students*. Routledge.

BOUREKKACHE S. (2022). English for specific purposes : writing scientific research papers. case study : Phd students in the computer science department. Mémoire de master, University of Biskra, Algeria.

DU W., KIM Z. M., RUNDERSTANDAHEJA V., KUMAR D. & KANG D. (2022a). Read, revise, repeat : A system demonstration for human-in-the-loop iterative text revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, p. 96–108, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.in2writing-1.14.

DU W., RAHEJA V., KUMAR D., KIM Z. M., LOPEZ M. & KANG D. (2022b). Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3573–3590, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.250.

ITO T., KURIBAYASHI T., KOBAYASHI H., BRASSARD A., HAGIWARA M., SUZUKI J. & INUI K. (2019). Diamonds in the rough : Generating fluent sentences from early-stage drafts for academic writing assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*, p. 40–53, Tokyo, Japan : Association for Computational Linguistics. DOI : 10.18653/v1/W19-8606.

JIANG C., XU W. & STEVENS S. (2022). arxivedits : Understanding the human revision process in scientific writing. *In Proceedings of EMNLP 2022*.

KALLESTINOVA E. D. (2011). How to write your first research paper. *The Yale journal of biology and medicine*, **84**(3), 181.

KANG D., AMMAR W., DALVI B., VAN ZUYLEN M., KOHLMEIER S., HOVY E. & SCHWARTZ R. (2018). A dataset of peer reviews (PeerRead) : Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1647–1661, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1149.

LAKSMI E. D. (2006). " scaffolding" students'writing in efl class : Implementing process approac. *TEFLIN Journal*, **17**(2), 144–156.

SEOW A. (2002). The writing process and process writing. *Methodology in language teaching : An anthology of current practice*, **315**, 320.

SILVEIRA E. A., DE SOUSA ROMEIRO A. M. & NOLL M. (2022). Guide for scientific writing : how to avoid common mistakes in a scientific article. *Journal of Human Growth and Development*, **32**(3), 341–352.

SWALES J. M. (1990). *Genre Analysis : English in academic and research settings*. The Cambridge applied linguistics series. The press syndicate of the University of Cambridge.

A Exemple des différences entre deux versions d'un même article issues de OpenReview

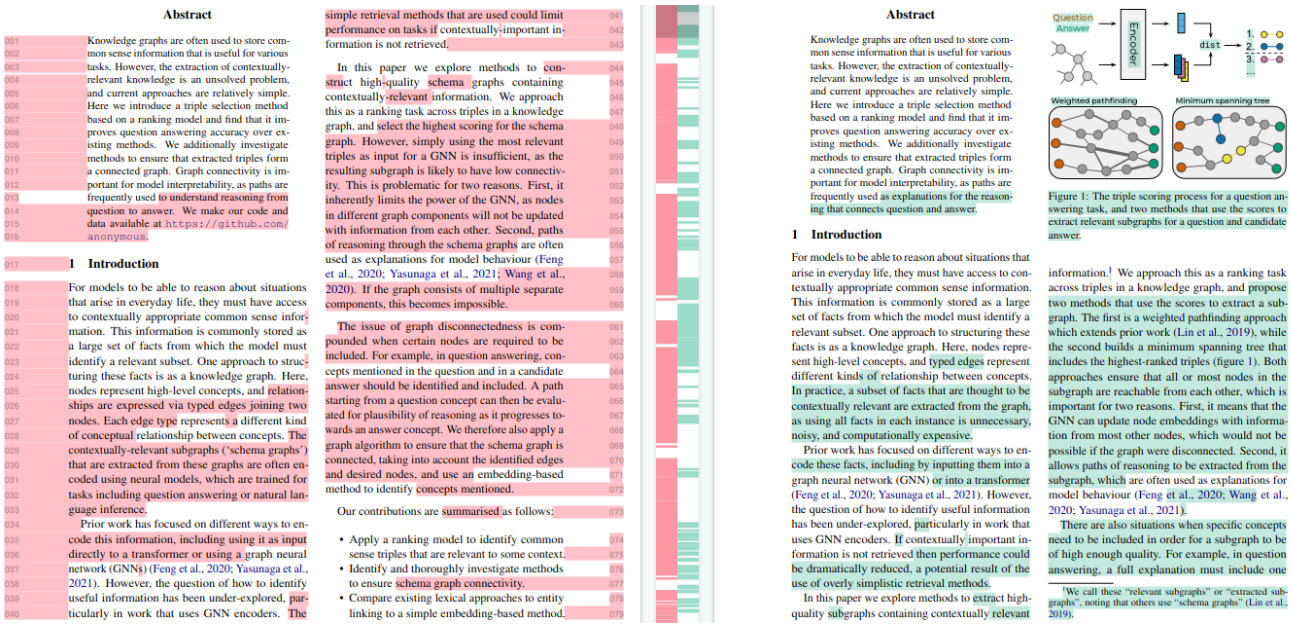


FIGURE 2 – Visualisation des différences entre la première, la dernière version de (Aglionby & Teufel, 2022)

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

MORFITT : Un corpus multi-labels d'articles scientifiques français dans le domaine biomédical

Yanis Labrak[1,3], Mickael Rouvier[1], Richard Dufour[2],
(1) LIA - Avignon Univserité (2) LS2N, UMR CNRS 6004, Nantes Université (3) Zenidoc
{yanis.labrak, mickael.rouvier}@univ-avignon.fr,
richard.dufour@univ-nantes.fr

RÉSUMÉ

Cet article présente MORFITT, le premier corpus multi-labels en français annoté en spécialités dans le domaine médical. MORFITT est composé de 3 624 résumés d'articles scientifiques issus de PubMed, annotés en 12 spécialités pour un total de 5 116 annotations. Nous détaillons le corpus, les expérimentations et les résultats préliminaires obtenus à l'aide d'un classifieur fondé sur le modèle de langage pré-entraîné CamemBERT. Ces résultats préliminaires démontrent la difficulté de la tâche, avec un score F1 moyen pondéré de 61,78 %.

ABSTRACT

MORFITT : A multi-label corpus of French scientific articles in the biomedical domain

This article presents MORFITT, the first multi-label corpus in French annotated in specialties in the medical field. MORFITT is composed of 3,624 abstracts of scientific articles from PubMed, annotated in 12 specialties for a total of 5,116 annotations. We detail the corpus, the experiments and the preliminary results obtained using a classifier based on the pre-trained language model CamemBERT. These preliminary results demonstrate the difficulty of the task, with a weighted average F1-score of 61.78%.

MOTS-CLÉS : BERT ; RoBERTa ; Transformers ; Biomédical ; Clinique ; Spécialités ; multi-labels.

KEYWORDS: BERT ; RoBERTa ; Transformers ; Biomedical ; Clinical ; Topics ; multi-labels.

1 Introduction

Depuis maintenant plusieurs années, le domaine médical suscite l'engouement des chercheurs de par les enjeux importants qui lui sont liés, avec par exemple une attente sociétale forte autour d'outils liés au traitement automatique du langage naturel (TALN) (Bazoge, 2021). La multiplication de ces travaux a conduit à une explosion des articles scientifiques disponibles, entraînant une surcharge d'informations et de connaissances à traiter par les scientifiques mais également les professionnels de la santé afin d'être en capacité de rester informé sur les avancées scientifiques (Chen *et al.*, 2022). Par exemple, dans le cadre de la pandémie de COVID-19, cela a pu avoir un impact sur la qualité des soins prodigués, entraînant des retards dans la prise de décision ou la prescription potentielle de traitements inadaptés (Riera *et al.*, 2021).

Une approche possible consiste à indexer automatiquement chaque document reçu afin d'aider les professionnels de santé à prioriser la lecture des documents, ou à accéder plus rapidement aux

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 66–70.

documents recherchés. Cette indexation peut être réalisée grâce à des méthodes de classification automatique multi-labels qui consistent, à partir du contenu textuel d'un document, à identifier automatiquement une ou plusieurs spécialités qui lui sont liées. Ces classifieurs automatiques reposent sur des modèles entraînés sur des ensembles de données préalablement étiquetées dans le but de pouvoir mettre en avant les caractéristiques importantes et identifier la ou les spécialités traitées dans le document.

Nous pouvons trouver, dans la littérature, plusieurs corpus multi-labels dans le domaine médical. Par exemple, LitCovid (Chen *et al.*, 2021) est un corpus de résumés d'articles scientifiques, extraits depuis PubMed, portant sur la COVID-19 et annotés en 8 spécialités. Nous pouvons également citer le corpus Hallmarks Of Cancer (HOC) (Baker *et al.*, 2016), un autre corpus de résumés d'articles scientifiques issus de PubMed et annotés avec 10 caractéristiques du cancer. Malheureusement, tous ces corpus sont en langue anglaise et, à notre connaissance, il n'existe actuellement aucun corpus multi-labels disponible librement en français dans le domaine médical.

Afin de palier ce manque, nous présentons dans cet article MORFITT, le premier corpus multi-labels en français composé de 3 624 résumés d'articles scientifiques dans le domaine médical extraits depuis PubMed, lesquels ont été annotés en 12 spécialités. Nous avons également évalué ce corpus au moyen d'un système état-de-l'art intégrant un classifieur multi-labels fondé sur le modèle de langue pré-entrainé sur le français (ici, CamemBERT (Martin *et al.*, 2020)). Les premiers résultats obtenus sur ce corpus sont rapportés dans cet article.

L'article est organisé comme suit. La Section 2 présente le corpus MORFITT, puis la Section 3 introduit les expériences et les résultats préliminaires que nous avons obtenu sur ce corpus. Enfin, la Section 4 conclut l'article.

2 Présentation du corpus MORFITT

Le corpus que nous proposons est constitué d'articles scientifiques dans le domaine médical provenant de PubMed [1], un moteur de recherche de données bibliographiques qui indexe l'ensemble des documents issus des domaines de spécialisation de la biologie et de la médecine. Nous avons, dans un premier temps, téléchargé l'ensemble des résumés des articles indexés par PubMed ainsi que les mots-clés MeSH [2] associés aux résumés d'articles en français à l'aide d'un script maison partant des 303 Go d'archives brutes. Les mots-clés MeSH principaux des articles sont utilisés pour définir les spécialités d'un article. Nous avons sélectionné une liste de mots-clés MeSH principaux, correspondant à 12 spécialités médicales ciblées (Tableau 1). Il est à noter que les mots-clés MeSH principaux associés aux articles sont sélectionnés manuellement par leurs auteurs à partir d'une liste de choix prédéfinis : bien que cette annotation manuelle soit réalisée par les auteurs eux-mêmes, nous avons cependant réalisé une vérification et correction manuelle afin de s'assurer de la qualité des spécialités assignées. Notons que nous n'avons pas corrigé les omissions de mots-clés par les auteurs vu l'ampleur de ce travail.

Le corpus a été découpé en trois sous-ensembles de données, à savoir le corpus d'entraînement, de développement et de test, contenant respectivement 1 514 (41,77 %), 1 022 (28,20 %) et 1 088 (30,02 %) documents. La distribution des spécialités dans le corpus est présentée dans le Tableau 1. Notons

1. https://pubmed.ncbi.nlm.nih.gov/
2. MeSH (Medical Subject Headings) est un thésaurus biomédical publié et mis-à-jour par la National Library of Medicine (US), et utilisé notamment pour l'indexation des références bibliographiques de MEDLINE/PubMed.

que chaque document peut être associé à plusieurs spécialités, ce qui explique le décalage entre nombre de documents et nombre de spécialités. De plus, nous avons porté une grande attention à la distribution des classes dans chacun des sous-ensembles, dans le but d'avoir des distributions de classes similaires, malgré la difficulté qui est liée à l'annotation multi-labels sur chaque document. Cette contrainte a ainsi entraîner une distribution générale du corpus assez peu commune, de 41.77 % pour le train, 28.20 % pour le dev et 30.02 % pour le test. De plus, comme nous pouvons le voir, les spécialités *Vétérinaire*, *Étiologie* et *Psychologie* sont les plus représentées, suivies de la *Chirurgie* et de la *Génétique*. De plus, 30,51 % des articles sont attribués à plus d'un sujet, comme présenté dans la Figure 1.

	Train	Dev	Test	Total
Vétérinaire	320	250	254	**824**
Étiologie	317	202	222	**741**
Psychologie	255	175	179	**609**
Chirurgie	223	169	157	**549**
Génétique	207	139	159	**505**
Physiologie	217	125	148	**490**
Pharmacologie	112	84	103	**299**
Microbiologie	115	72	86	**273**
Immunologie	106	86	70	**262**
Chimie	94	53	65	**212**
Virologie	76	57	67	**200**
Parasitologie	68	34	50	**152**
Total	**2 110**	**1 446**	**1 560**	**5 116**

TABLE 1 – Distribution des étiquettes au travers des sous-ensembles de données : Apprentissage (Train), Développement (Dev) et Test (Test).

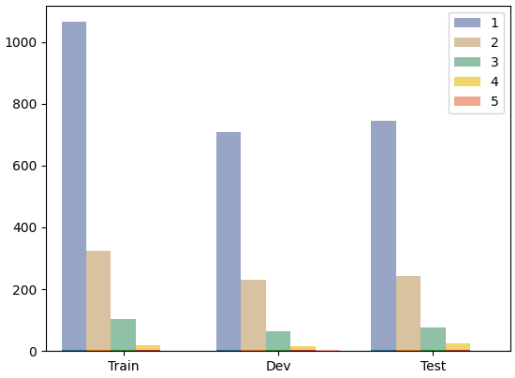


FIGURE 1 – Distribution du nombre de spécialités par article et par sous-ensemble de données : Apprentissage (Train), Développement (Dev) et Test (Test).

Nous avons observé trois principaux schémas de co-occurrences entre les spécialités, comme détaillé par le diagramme de Chord dans la Figure 2 : (i) génétique-vétérinaire avec 150 co-occurrences ; (ii) microbiologie-vétérinaire avec 117 co-occurrences ; (iii) immunologie-vétérinaire avec 98 co-occurrences.

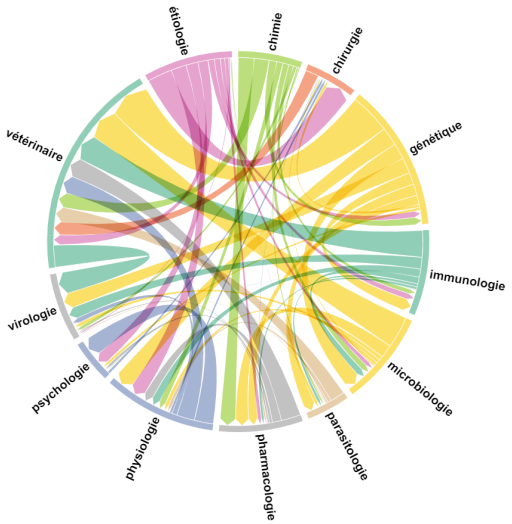


FIGURE 2 – Diagramme de Chord des co-occurrences des spécialités sur l'ensemble du corpus.

3 Expériences et résultats

Pour identifier automatiquement les différentes spécialités, nous utilisons le modèle pré-entraîné CamemBERT (Martin *et al.*, 2020). Afin d'améliorer les performances, nous avons affiné ce modèle sur le corpus d'apprentissage MORFITT. Nous avons également entraîné une couche de classification de dimension 12, qui correspond aux spécialités traitées, sur laquelle nous avons appliqué une fonction objective BCE pendant 32 itérations avec un taux d'apprentissage de $2e-5$. Nous avons fixé le seuil de sélection des classes à $0,70$ pour toutes les spécialités en utilisant un processus manuel d'essais et d'erreurs. Ces méta-paramètres ont été optimisés sur le corpus de développement.

Nous évaluons les performances du système en utilisant trois métriques : la précision moyenne pondérée et macro, le rappel ainsi que le score F1.

$$\textbf{Précision} = \frac{\text{vrai positif}}{\text{vrai positif + faux positif}} \quad (1) \qquad \textbf{Rappel} = \frac{\text{vrai positif}}{\text{vrai positif + faux négatif}} \quad (2) \qquad \textbf{Score F1} = \frac{\text{2 x précision x rappel}}{\text{précision + rappel}} \quad (3)$$

Le Tableau 2 liste les résultats du classifieur CamemBERT 138 GB OSCAR sur l'ensemble des spécialités. On observe que le score F1 moyen pondéré est de 61,78 %. Trois spécialités ont obtenu un score supérieur à 75 % tandis que pour les autres spécialités, et environ 40 % d'entre-eux ont obtenu un score inférieur à 50 %. La spécialité ayant obtenu les meilleurs résultats est *Psychologie*, avec un score de 86,98 %, tandis que la spécialité *Parasitologie* a obtenu les résultats les plus faibles, avec seulement 7,55 %.

Les spécialités ayant obtenu des scores F1 bas sont principalement dues au rappel plutôt qu'à la précision. Il semble que cela ne soit pas corrélé à la quantité de résumés d'articles présents dans le corpus. En effet, la spécialité *Parasitologie* est la troisième classe la plus représentée en termes de nombre de résumés présents dans l'ensemble d'apprentissage, mais elle obtient cependant de mauvais résultats. Il semble plutôt que cela soit dû à la difficulté intrinsèque des spécialités et des résumés que nous cherchons à identifier.

Spécialités	Précision (%)	Rappel (%)	Score F1 (%)
Vétérinaire	79,76	77,56	78,64
Étiologie	68,25	58,11	62,77
Psychologie	86,26	87,71	86,98
Chirurgie	80,38	80,89	80,63
Génétique	81,75	64,78	72,28
Physiologie	75,00	38,51	50,89
Pharmacologie	73,08	18,45	29,46
Microbiologie	70,91	45,35	55,32
Immunologie	65,52	27,14	38,38
Chimie	76,19	24,62	37,21
Virologie	75,00	17,91	28,92
Parasitologie	66,67	4,00	7,55
Moyenne Macro	**74,90**	**45,42**	**52,42**
Moyenne Pondérée	**76,34**	**56,22**	**61,78**

TABLE 2 – Résultats obtenus par CamemBERT sur le sous-ensemble de test.

4 Conclusions

Nous avons proposé le premier corpus multi-labels d'articles scientifiques biomédicaux annotés en spécialités. Les données ainsi que le modèle état-de-l'art s'appuyant sur CamemBERT sont disponibles librement. Dans de futurs travaux, nous expérimenterons des modèles BERT spécifiques au domaine médical comme le modèle français DrBERT (Labrak *et al.*, 2023) ou anglais BioBERT (Lee *et al.*,

2019; Wang *et al.*, 2022), ou encore des modèles capables de gérer des séquences plus longues comme les LongFormer (Beltagy *et al.*, 2020) et sa version biomédicale Clinical-Longformer (Li *et al.*, 2022; Liu *et al.*, 2022). Ces types de modèles ont été appliqués avec succès à des tâches similaires du domaine médical pour la langue anglaise et seraient susceptibles aussi d'améliorer les performances pour le français, car ils fournissent une représentation plus contextualisée des termes médicaux présents dans les résumés. Les données ainsi que les modèles sont disponibles librement sur Github [3] et HuggingFace [4].

Références

BAKER S., SILINS I., GUO Y., ALI I., HÖGBERG J., STENIUS U. & KORHONEN A. (2016). Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinform.*, **32**(3), 432–440. DOI : 10.1093/bioinformatics/btv585.

BAZOGE A. (2021). Revue de la littérature : entrepôts de données biomédicales et traitement automatique de la langue. *Traitement Automatique des Langues Naturelles*, p. 94–107.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The long-document transformer.

CHEN Q., ALLOT A., LEAMAN R., WEI C.-H., AGHAARABI E., GUERRERIO J., XU L. & LU Z. (2022). LitCovid in 2022 : an information resource for the COVID-19 literature. *Nucleic Acids Research*, **51**(D1), D1512–D1518. DOI : 10.1093/nar/gkac1005.

CHEN Q., ALLOT A. & LU Z. (2021). Litcovid : an open database of covid-19 literature. *Nucleic acids research*, **49**(D1), D1534–D1540.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). Drbert : A robust pre-trained model in french for biomedical and clinical domains. *The 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. DOI : 10.1093/bioinformatics/btz682.

LI Y., WEHBE R. M., AHMAD F. S., WANG H. & LUO Y. (2022). Clinical-longformer and clinical-bigbird : Transformers for long clinical sequences. DOI : 10.48550/ARXIV.2201.11838.

LIU L., PEREZ-CONCHA O., NGUYEN A., BENNETT V. & JORM L. (2022). Automated icd coding using extreme multi-label long text transformer-based models.

MARTIN L., MULLER B., SUÁ REZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 7203—-7219 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

RIERA R., BAGATTINI Â. M., PACHECO R. L., PACHITO D. V., ROITBERG F. & ILBAWI A. (2021). Delays and disruptions in cancer health care due to covid-19 pandemic : systematic review. *JCO Global Oncology*, **7**(1), 311–323.

WANG X., WANG J., TANG W. & ZHANG H. (2022). Multi-label topic classification for covid-19 literature annotation : A biobert-based feature enhancement approach. In *CIBDA 2022 ; 3rd International Conference on Computer Information and Big Data Applications*, p. 1–4.

3. https://huggingface.co/datasets/qanastek/MORFITT
4. https://github.com/qanastek/MORFITT

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

La détection de textes générés par des modèles de langue : une tâche complexe? Une étude sur des textes académiques

Vijini Liyanage Davide Buscaldi
Le Laboratoire d'Informatique de Paris-Nord (LIPN), France
liyanage@lipn.univ-paris13.fr , davide.buscaldi@lipn.univ-paris13.fr

RÉSUMÉ

L'émergence de modèles de langue très puissants tels que GPT-3 a sensibilisé les chercheurs à la problématique de la détection de textes académiques générés automatiquement, principalement dans un souci de prévention de plagiat. Plusieurs études ont montré que les modèles de détection actuels ont une précision élevée, en donnant l'impression est résolue. Cependant, nous avons observé que les ensembles de données utilisés pour ces expériences contiennent des textes générés automatiquement à partir de modèles pré-entraînés. Une utilisation plus réaliste des modèles de langue consisterait à effectuer un affinage sur un texte écrit par un humain pour compléter les parties manquantes. Ainsi, nous avons constitué un corpus de textes générés de manière plus réaliste et mené des expériences avec plusieurs modèles de classification. Nos résultats montrent que lorsque les ensembles de données sont générés de manière réaliste pour simuler l'utilisation de modèles de langue par les chercheurs, la détection de ces textes devient une tâche assez difficile.

ABSTRACT

How difficult is it to detect LLM-generated text ? A study on academic text

Prevalence of competent models such as GPT-3 has fostered many researches to focus on the detection of automatically generated academic text. Several studies have shown that current detection models have high accuracy, which may give the impression that the task is solved. However, we observed that the datasets used for these experiments often contain automatically generated texts from pre-trained models. A more realistic use of language models would be to fine-tune them on an original text to complete its missing parts (e.g. abstract or conclusions). Therefore, we built a corpus of texts generated in a more realistic way and conducted experiments with several classification models. Our results show that when datasets are generated realistically to simulate the use of language models by researchers, detecting these texts becomes more challenging.

MOTS-CLÉS : Modèles de langue, Génération automatique de texte, Détection, Classification.

KEYWORDS: LLMs, Automatic text generation, Detection, Classification.

1 Introduction

Depuis l'introduction de grands modèles de langue (ML) pour la génération de texte, les chercheurs ont essayé de déterminer si le texte généré peut être détecté et avec quelle efficacité. Par exemple, Giant Language Testing Room (GLTR) (Gehrmann *et al.*, 2019) est un modèle de visualisation qui aide les humains à détecter le texte généré artificiellement, en se basant sur le principe qu'une prévalence de tokens très probables (selon un modèle de langue pre-entrainé - dans leur cas GPT-

2) est un indicateur. (Rodriguez *et al.*, 2022) ont étudié la détectabilité de documents techniques falsifiés par des modèles de classification basés sur BERT, obtenant des précisions allant de 86 à 95%. DetectGPT (Mitchell *et al.*, 2023) évalue l'effet de perturbations aléatoires du texte sur les probabilités de génération des tokens par un modèle GPT, de façon similaire a GLTR. Ils affichent une AUROC de 0.95.

Récemment, certaines tâches partagées ont été proposées sur sujet de la détection de texte academique généré automatiquement. DAGPap22 a été accueilli dans le cadre du Third Workshop on Scholarly Document Processing (Cohan *et al.*, 2022). Les organisateurs ont rassemblé des articles originaux sur MICPRO (microprocesseurs et microsystèmes) et les objectifs de développement durable des Nations unies, et ont créé des résumés artificiels en utilisant diverses méthodes (modèles de résumé et modèles de type GPT). La tâche s'est avérée plus facile à résoudre que prévu, le meilleur modèle étant basé sur trois variantes de BERT : SciBERT(Beltagy *et al.*, 2019), RoBERTa(Liu *et al.*, 2019) et DeBERTa(He *et al.*, 2020). En utilisant un ensemble de ces trois modèles, il a obtenu un score F1 maximal de 99,24% (Glazkova & Glazkov, 2022). SynSciPass (Rosati, 2022) utilise le modèle SciBERT pour la détection et ont obtenu une précision de 98,3% pour DAGPap22.

Toutefois, ces excellents résultats peuvent être trompeurs. En effet, si l'on regarde plus attentivement aux jeux de données utilisés, nous trouverons que les sous-ensembles de données générés sont souvent assemblés en utilisant des ML pre-entrainés. Par exemple, (Mitchell *et al.*, 2023) utilisent 500 articles d'un jeu de donnés d'actualités pour les échantillons originaux, et 500 articles générés automatiquement par 4 différents ML avec les premiers 30 mots de chaque article original. Le jeu de données WikiGPT [1] a été créé avec le prompt "Introduction de 200 mots de style wikipedia sur {titre} {texte_intro}" où *titre* est le titre de la page wikipedia, et *texte_intro* est constitué par les sept premiers mots de l'introduction Wikipedia originale.

2 Un jeu de données plus réaliste

Nous avons commencé à construire un corpus avec l'objectif de simuler au mieux la façon dont un auteur humain utiliserait les LMs pour ses travaux : par exemple, pour combler les trous dans certaines parties de l'article ou pour compléter un résumé en donnant le corps de l'article. Après, notre objectif c'était aussi celui de vérifier si la détectabilité des textes générés automatiquement restait élévée aussi dans ce cas. Nous avons donc assemblé un jeu de données composé des sections suivantes :

- Un ensemble de données composé d'articles générés par un réglage fin du modèle GPT-2 avec un paramètre de température de 0,7. **Entièrement généré GPT-2 affiné (t = 0.7) (D1)** - 200 articles, 1250 mots par article en moyenne ;
- Un ensemble de données composé de résumés qui contiennent à la fois du contenu généré par une machine (en utilisant GPT-2 affiné sur ArXiv) et du contenu écrit par un humain. **Abstraites hybride ArXiv-NLP pré-entraîné (t = 0.7) (D2)** - 200 articles, 150 mots par article en moyenne ;
- Le jeu de données du concours DAGPap22 [2]. **DAGPap22(D3)** - 5350 articles, 160 mots par article en moyenne ;
- Un ensemble de données composé d'articles générés par un réglage fin du modèle GPT-2 avec un paramètre de température de 0,9. **Entièrement généré GPT-2 affiné (t = 0.9)(D4)** - 200

1. https://huggingface.co/datasets/aadityaubhat/GPT-wiki-intro
2. https://www.kaggle.com/competitions/detecting-generated-scientific-papers

articles, 1250 mots par article en moyenne ;

— Un jeu de données de résumés obtenu en exploitant le modèle GPT-2 pré-entraîné, sans réglage fin. **Abstraites GPT-2 pré-entraîné (t = 0.7) (D5)** - 200 articles, 125 mots par article en moyenne ;

Pour construire les ensembles de données D1 et D4, nous avons effectué un réglage d'un modèle GPT-2 sur les articles originaux. Pour s'assurer qu'il n'y a pas de chevauchement entre les données d'entraînement de GPT-2 et nos données, nous avons sélectionné pour notre jeu de données uniquement des articles récents (à partir de 2022). Afin d'obtenir des résultats plus fidèles à partir du modèle pré-entraîné, nous avons limité notre jeu de données au même domaine (NLP) et à la même source (ArXiv). Pour chaque article original, nous envoyons un prompt de 50 mots provenant de l'article lui-même au modèle afin de produire un nouvel article. La génération continue jusqu'à ce que la longueur du nouvel article soit similaire à la longueur de l'article original, afin d'éviter tout biais de longueur lors de la classification. La différence entre D1 et D4 réside uniquement dans les températures utilisées pour les générateurs, respectivement $0,7$ et $0,9$. Le modèle GPT-2 de base était la version avec 12 couches, 12 têtes d'attention et 124M de paramètres. Le modèle a été affiné par un seul article original à la fois et une moyenne de 45 minutes GPU a été consommée pour chaque article (avec un GPU Nvidia T4). Ainsi, pour l'ensemble du jeu de données, environ 75 heures de GPU ont été allouées.

Le jeu de données D2 a été créé avec une intervention humaine, en utilisant l'interface 'write with transformer' sur Huggingface [3]. Les auteurs, agissant en tant qu'experts du domaine, ont supprimé la partie conclusive des résumés, et les ont complétés avec des phrases proposées par le modèle GPT-2 mis au point sur ArXiv-NLP. Nous avons choisi l'une des trois meilleures complétions proposées par le modèle. Ce scénario est équivalent au scénario de 'tampering' proposé dans (Rodriguez *et al.*, 2022), avec des curateurs manuels.

Finalement, le jeu de données D5 réprésente le cas où le modèle est utilisé sans réglage, avec uniquement un prompt de 50 mots de l'article original. Comme on peut apprécier en Table 1, DetectGPT n'est pas capable de détecter correctement le fait que dans D1 et D2 le contenu des articles a été manipulé par GPT-2 (en Figure 1 c'est aussi possible de voir la sortie de GLTR pour un exemple dans les différents datasets, même si nous n'avons pas calculé les résultats de GLTR sur tous les exemples). Nous n'avons pas le résultat final pour D4 pour des contraintes de temps, mais le corpus étant très similaire à D1, nous attendons des résultats comparables.

Modèle	**D1**	**D2**	**D5**
Z-score	-0,351	0,239	0,911

TABLE 1 – Z-scores moyens obtenus par DetectGPT pour les jeux des données D1, D2 et D5. Z-scores supérieurs à 0.7 indiquent que DetectGPT considère les textes synthétiques.

3 Expériences

Pour comprendre le degré de similarité du texte généré avec l'original et donc estimer la difficulté inhérente des jeux de données, nous avons calculé les scores BLEU (Bilingual Evaluation Understudy) (Papineni *et al.*, 2002) et ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) (Tableau 2). Lorsque le paramètre de température d'un modèle de génération est augmenté, le caractère aléatoire du contenu généré est plus élevé. Par conséquent, la similarité n-gram entre le contenu généré et le contenu de référence (original) correspondant devient faible. Ainsi, comme le montre le

3. https://transformer.huggingface.co/

Tableau 2, BLEU et ROUGE, qui sont des scores basés sur les n-grammes, sont plus faibles pour le D4 plus “chaud" que les scores produits par le jeu de données D1 plus “froid".

Les scores BLEU et ROUGE du jeu de données D5, généré à l’aide du modèle GPT-2 sans réglage, sont très faibles par rapport aux corpus générés à l’aide du modèle GPT-2 affiné. Il s’agit d’un comportement attendu puisque lorsqu’un modèle n’est pas affiné, le contenu généré ne peut pas être très similaire au contenu original, ce qui rend la similarité n-grammes très faible.

Méthode de construction de la base de données	**UGL-BLEU**	**S-BLEU**	**Rouge-1**	**Rouge-2**	**Rouge-L**
(D1)GPT-2 affiné (t = 0.7)	0.867	0.809	0.853	0.810	0.853
(D4)GPT-2 affiné (t = 0.9)	0.858	0.766	0.830	0.772	0.834
(D5)GPT-2 pré-entraîné (t = 0.7)	0.467	0.356	0.549	0.431	0.533
(D3) ArXiv-NLP pré-entraîné (t = 0.7)	0.824	0.792	0.882	0.840	0.881

TABLE 2 – Scores moyens BLEU et ROUGE(Recall) , UGL-BLEU : Niveau d’unigramme BLEU, S-BLEU : Phrase BLEU

Pour la classification des textes en généré ou original, nous avons exploité plusières variantes de BERT telles que SciBERT, RoBERTa, DeBERTa and ELECTRA (Clark *et al.*, 2020). Le tableau 5 en annexe fournit les paramètres de chaque variante. Tous les jeux de données ont été divisés aléatoirement en 80 : 20 pour les tests.

Calculé pour chaque modèle le score F1 (voir Table 3). Tous ces scores sont la moyenne de trois expériences avec une repartition aléatoire des jeux d’entrainement et test. En général, les scores sont élevés pour les jeux de données D3 et D5, ce qui prouve qu’ils sont plus faciles à détecter. D3 n’utilise pas de méthodes neuronales pour la partie générée, tandis que D5 est généré sans réglage. D2 est un jeu de données altéré, ce qui signifie que le texte synthétique est produit en incluant de légères modifications au contenu original. La détection est donc difficile, comme le prouvent les scores de classification relativement faibles pour ce jeu de données.

Modèle	**D1**	**D2**	**D3**	**D4**	**D5**
BERT	33.33	56.04	53.29	60.11	89.90
SciBERT	84.65	52.23	95.02	79.16	94.99
RoBERTa	67.83	55.23	96.87	33.33	84.85
DeBERTa	67.03	48.85	97.17	48.13	95.00
Electra$_{small}$	60.11	51.00	93.95	56.36	92.50
Electra$_{base}$	56.16	41.30	96.64	48.13	95.00

TABLE 3 – F-1 scores obtenus par les variantes BERT

4 Conclusions

Cette étude a examiné la détection de textes synthétiques générés à l’aide de modèles de langue, en se concentrant sur les modèles entraînés sur des articles académiques. Les résultats ont montré que lorsque les modèles sont affinés à l’aide d’articles originaux, la détection des textes synthétiques devient plus difficile, car le texte généré peut contenir des éléments de l’article original. En revanche, textes générés par des modèles non affinés sont plus faciles à détecter. Les textes modifiés manuellement représentent un défi important pour les modèles, mais les résultats peuvent varier en fonction de

l'architecture du classificateur ou de la construction de l'ensemble de données. L'étude a également examiné l'effet de l'ajustement de la température du modèle de génération sur la détection, mais les résultats sont mitigés et nécessitent des recherches supplémentaires.

Références

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3615–3620, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

CLARK K., LUONG M., LE Q. V. & MANNING C. D. (2020). ELECTRA : pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

COHAN A., FEIGENBLAT G., FREITAG D., GHOSAL T., HERRMANNOVA D., KNOTH P., LO K., MAYR P., SHMUELI-SCHEUER M., DE WAARD A. & WANG L. L. (2022). Overview of the third workshop on scholarly document processing. In *Proceedings of the Third Workshop on Scholarly Document Processing*, p. 1–6, Gyeongju, Republic of Korea : Association for Computational Linguistics.

GEHRMANN S., STROBELT H. & RUSH A. (2019). GLTR : Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 111–116, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-3019.

GLAZKOVA A. & GLAZKOV M. (2022). Detecting generated scientific papers using an ensemble of transformer models. In *Proceedings of the Third Workshop on Scholarly Document Processing*, p. 223–228, Gyeongju, Republic of Korea : Association for Computational Linguistics.

HE P., LIU X., GAO J. & CHEN W. (2020). Deberta : Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv :2006.03654*.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Text summarization branches out*, p. 74–81.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MITCHELL E., LEE Y., KHAZATSKY A., MANNING C. D. & FINN C. (2023). Detectgpt : Zero-shot machine-generated text detection using probability curvature.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

RODRIGUEZ J., HAY T., GROS D., SHAMSI Z. & SRINIVASAN R. (2022). Cross-domain detection of GPT-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1213–1233, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.88.

ROSATI D. (2022). SynSciPass : detecting appropriate uses of scientific text generation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, p. 214–222, Gyeongju, Republic of Korea : Association for Computational Linguistics.

Annexe

Origine	Extrait
Document original	Moreover, when temporal word embeddings provide words' representations that are more inclined towards their literal core meaning (and not the metaphorical one), models exploiting end up correctly identifying metaphors more easily. **Furthermore, embeddings of words belonging to some language domains in specific time periods can be biased towards their metaphorical meaning, leading to words being used in metaphorical contexts way more than in literal ones. This would prevent neural models from correctly identifying the words as metaphors.**
D1	Moreover, when temporal word embeddings provide words' representations that are more inclined towards their literal core meaning (and not the metaphorical one), models exploiting end up correctly identifying metaphors more easily. **One way to investigate these hypotheses further is to explore the nearest neighbors of a word in the word embeddings used in a figurative way inside a sentence, both in a static (atemporal) word embedding space, e.g., obtained with GloVe [Jeffrey Pennington, 2014] and in a decade-specific temporal space, e.g., obtained from the CoHa1 corpus with Procrustes [Edouard Grave, 2018].**
D4	Moreover, when temporal word embeddings provide words' representations that are more inclined towards their literal core meaning (and not the metaphorical one), models exploiting end up correctly identifying metaphors more easily. **One way to investigate these hypotheses further is to explore the nearest neighbors of a word in the word embeddings used in a figurative way inside a sentence, both in a static (atemporal) word embedding, i.e., a decade, time period, and/or a temporal (osfringes) word embedding, i.e., a decade, time period.**
D5	Moreover, when temporal word embeddings provide words' representations that are more inclined towards their literal core meaning (and not the metaphorical one), models exploiting end up correctly identifying metaphors more easily. **For example, if a model is able to identify a metaphor that is more likely to be associated with the literal meaning of a word, then it will be more accurate in identifying the metaphor in the context of the word itself. In this paper, we present a novel approach to the identification of metaphors in language.**
D2	Moreover, when temporal word embeddings provide words' representations that are more inclined towards their literal core meaning (and not the metaphorical one), models exploiting end up correctly identifying metaphors more easily. **In contrast to sentence embeddings, the method of segmentation also exploits the concept of end-to-end (or equivalence) for metaphors and the corresponding temporal word embeddings. In this work, we conduct a multidisciplinary investigation of a well-known technique for identifying and segmenting metaphors.**

TABLE 4 – Extraits d'un article et ses versions générées à partir des différentes sections de l'ensemble de données (non applicable pour D3).

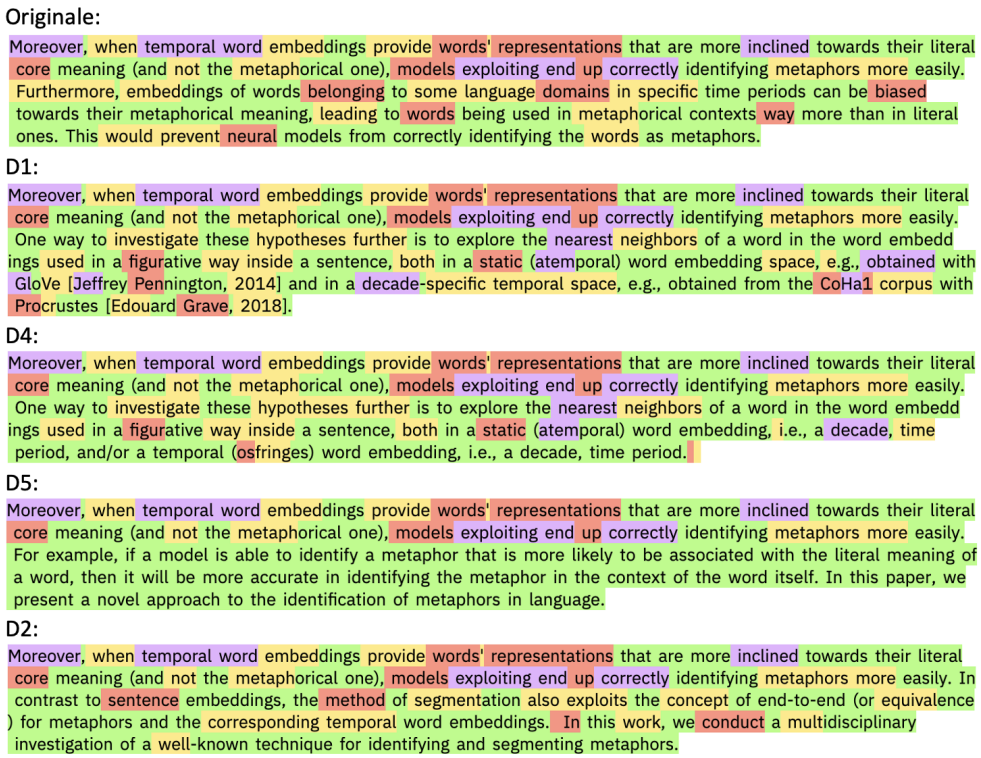


FIGURE 1 – Sortie du modèle GLTR pour les exemples du tableau 4. Vert : le token est dans le top 10 des tokens les plus probables pour le LM. Jaune : token dans le top 100. Rouge : token dans le top 1000. Violet : le token n'apparaît pas dans les 1000 meilleures options pour le LM.

Modèle	**Vocab (K)**	**Taille cachée**	**Couches**	**Taille du lot**	**Époques**	**Paramètres(M)**
BERT_{base}	30	762	12	64	3	110
DistilBERT	30	768	6	16	3	66
SciBERT_{base}	30	768	12	16	3	110
RoBERTa_{large}	50	1024	16	16	3	355
DeBERTa_{large}	50	1024	24	16	3	350
Electra_{small}	30	256	12	16	3	14
Electra_{base}	30	768	12	16	3	110
XLNet_{base}	32	768	12	16	3	110

TABLE 5 – Hyper Paramètres des variantes de BERT considérées

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Construction d'un jeu de données de publications scientifiques pour le TAL et la fouille de textes à partir d'ISTEX

Mathieu Constant[1]
(1) Université de Lorraine, CNRS, ATILF, 44 avenue de la Libération, 54000 Nancy, France
Mathieu.Constant@univ-lorraine.fr

RÉSUMÉ

La plateforme ISTEX (https://www.istex.fr) permet d'accéder à une large base d'archives scientifiques comptant plus de 25 millions de documents de tous les grands domaines scientifiques. Les documents incluent non seulement les métadonnées mais aussi le texte plein, et ont été prétraités de manière homogène pour faciliter leur traitement automatique. Dans cet exposé, nous présenterons une initiative pour favoriser les travaux de recherche en TAL et fouille de textes autour de ces données. En particulier, nous présenterons les travaux en cours pour la construction d'un jeu de données permettant d'apprendre et d'évaluer des modèles pour différentes tâches comme l'extraction de mots-clés, l'identification du domaine scientifique ou la génération de résumés. Nous avons circonscrit notre ensemble de travail aux publications d'ISTEX dont il existe une version open-access, grâce à la collaboration de l'INIST, avec l'objectif de constituer un jeu de données ouvert. Ce filtrage a permis d'identifier un ensemble multilingue de 3 millions de documents environ, dont une très large majorité en anglais mais couvrant une dizaine de langues différentes. Dans cet exposé, nous décrirons en particulier les traitements réalisés à l'ATILF pour éliminer les documents non pertinents ou bruités (ex. mauvaise qualité d'OCR), et ainsi ne garder que les documents de qualité.

ABSTRACT

Construction of a dataset of scientific publications for NLP and text mining from ISTEX.

The ISTEX platform (https://www.istex.fr) provides access to a large database of scientific archives with more than 25 million documents from all major scientific fields. The documents include not only the metadata but also the plain text, and have been pre-processed in a homogeneous way to facilitate their automatic processing. In this talk, we will present an initiative to promote research work in NLP and text mining around these data. In particular, we will present the work in progress for the construction of a dataset to learn and evaluate models for different tasks like keyword extraction, scientific domain identification, abstract generation. We have limited our working set to ISTEX publications having an open-access version, thanks to the collaboration of INIST, with the objective of constituting an open dataset. This filtering made it possible to identify a multilingual set of approximately 3 million documents, the vast majority of which are in English but covering around ten languages. In this talk, we will describe in particular the processing carried out at ATILF to eliminate irrelevant or noisy documents (e.g. poor OCR quality), and thus keep only documents of good quality.

MOTS-CLÉS : Publications scientifiques, construction de jeu données, traitement automatique des langues, fouille de textes.

KEYWORDS: Scientific publications, dataset construction, natural language processing, text mining.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

What shall we read: the article or the citations? - A case study on scientific language understanding

Aman Sinha[1, 2] Sam Bigeard[1] Marianne Clausel[1] Mathieu Constant[2]
(1) IECL, Université de Lorraine, Nancy, France
(2) ATILF, Université de Lorraine, Nancy, France
{firstname.lastname}@univ-lorraine.fr

RÉSUMÉ

The number of scientific articles is increasing tremendously across all domains to such an extent that it has become hard for researchers to remain up-to-date. Evidently, scientific language understanding systems and Information Extraction (IE) systems, with the advancement of Natural Language Processing (NLP) techniques, are benefiting the needs of users. Although the majority of the practices for building such systems are data-driven, advocating the idea of "*The more, the better*". In this work, we revisit the paradigm - questioning what type of data : text (title, abstract) or citations, can have more impact on the performance of scientific language understanding systems.

ABSTRACT

Lire l'article ou les citations ? - Une étude de cas sur la compréhension du langage scientifique

Le nombre d'articles scientifiques explose dans tous les domaines à tel point qu'il est devenu difficile pour les chercheurs de suivre l'évolution de la littérature scientifique. De toute évidence, intégrer les avancées dans le domaine du traitement du langage naturel (TAL), ne peut être que bénéfique aux utilisateurs cherchant à faire de la veille scientifique. La majorité des systèmes actuels est basée sur l'exploitation du caractère massif des données, partant de l'idée que "*Plus de données implique de meilleures performances*". Dans ce travail de recherche, nous revisitons le paradigme : se demander quel type de données : texte (titre, résumé) ou de citations, peut avoir plus d'impact sur la performance des systèmes scientifiques de compréhension du langage.

MOTS-CLÉS : Littérature scientifique, Recherche d'information, Compréhension du langage scientifique.

KEYWORDS: Scientific text, Information Extraction, Scientific Language Understanding.

1 Introduction

Scientific databases [1] are accumulating a large amount of literature to such an extent that it is getting overwhelming (Johnson *et al.*, 2018) and practically impossible to be up-to-date for researchers. Several recent works have looked into building intelligent systems for tasks such as ad-hoc based retrieval, conversational-agents, recommendation, summarization, document search, and re-ranking, as shown by the advances in the area of Neural Information Retrieval(IR) and Biomedical text mining (Zhang *et al.*, 2016; Gu *et al.*, 2020; Thakur *et al.*, 2021). Given the unstructured nature, the lengthiness and other metadata (such as citations information), in the scientific documents, the key question that arises when dealing with scientific information extraction is what type of data can be helpful to build high-performance models : text and/or citation information ?

1. Examples : Pubmed (https ://pubmed.ncbi.nlm.nih.gov/), ASCO(https ://www.asco.org/) ; ArXiv (https ://arxiv.org/), etc.

We investigate the informativeness of different type of features : *text-only* features, *graph-only* features and *text-graph* features via the document classification task. Additionally, we further explore impact of increasing the amount of text features. The main contribution of our work is to benchmark the effectiveness of the three type of features via various machine learning and deep learning based models for document classification task.

2 Related Work

This work investigates the impact of different text and graph-based features for document classification task. We briefly discuss the works which are related to our study by grouping them into the following categories :

Neural IE & Transfer Learning. The effectiveness of more text for text-ranking and text-retrieval tasks has been studied previously(Lin, 2009). Several works (Guo *et al.*, 2011; Ermakova *et al.*, 2018; Yeganova *et al.*, 2021) have also investigated at the importance of features from different sections in scientific articles using methods such as traditional BM25 scoring, deep-NN (Huang *et al.*, 2013) and weighted word-count (Yu *et al.*, 2014). In parallel, transfer learning (Peng *et al.*, 2019; Gu *et al.*, 2020; Kanakarajan *et al.*, 2021) has received a lot of attention with domain specific pre-trained language models (PLMs) for various downstream tasks such NER, relation extraction, question answering, document classification for scientific and biomedical text mining.

Citation graphs. Viewing data as a *graph* has been positively explored for various NLP tasks (Wu *et al.*, 2023) involving knowledge graphs (Mondal *et al.*, 2021; Jin *et al.*, 2019) and ontology-integrated models (Sinha *et al.*, 2022), showcasing the existing informativeness in graphs. We explore graph representation learning (Perozzi *et al.*, 2014; Hamilton *et al.*, 2017; Kipf & Welling, 2017) combining the structural information of graph with text-features in order to enhance the learning of information extraction models.

3 Experiments

Dataset We use the *PubMed-Diabetes* (Galileo Mark Namata & Huang, 2012) for our experiments. The dataset contains 19717 articles belonging to 3 classes of diabetes-mellitus, *Experimental*, *Type-1*, and *Type-2*. The dataset also provides citation information (eg. paper A $\xrightarrow{\text{cites}}$ paper B) and original PubMed-IDs. Using the graph terminology, PubMed-ID(s) denotes node(s) and the citation relation denotes edge(s) and therefore, the dataset can be also viewed as a citation graph with 19716 nodes [2] and 44338 edges.

Models We briefly describe the different models we use to perform our learning task. These models exploit standalone text features or graph features or combination of both the features.

— **Text-CNN** (Zhang & Wallace, 2015) : Besides the BERT (Devlin *et al.*, 2018) [CLS]-baseline, we used BERT-CNN architecture, which comprised of five 2d-CNN blocks applied on the token representations followed by max-pooling layer, to capture better quality text features.

— **DeepWALK** (Perozzi *et al.*, 2014) : Similar to Word2Vec algorithm (Mikolov *et al.*, 2013), this algorithm takes as input, a sequence of connected nodes and provides static embeddings for each node in the citation graph.

— **GraphSAGE** (GS) (Hamilton *et al.*, 2017) : This algorithm takes as input the citation graph and node-features (text features) to generate node embeddings, which contains both combina-

2. Note : One of the article-id ($pid.17874530$) was not anymore valid on PubMed, so we removed it from the nodes set ; no article was connected with the removed article-id and so the edge list was unaffected.

tion of graph-based and text-based information. The algorithm works under the concept of iterative message aggregation for every node from the neighborhood.

Features We briefly describe various features and PLMs that we used in our experiments.

— **random** : We use random noise feature in order to show the impact of addition of text/graph features for any model.

— **tfidf** : The dataset was provided with 500-length tf-idf features based on a curated list of keywords (Galileo Mark Namata & Huang, 2012) (tfidf-c). For our study on different text portions (title-only or abstract-only), we generated manual tf-idf features (tfidf-m) separately with each corpora.

— **BERT** (Devlin *et al.*, 2018) : We use the contextual embeddings ([CLS]-layer) from BERT model which is pretrained with English Wikipedia and BookCorpus. We use it as our baseline for domain-specific PLMs.

— **BioBERT** (Lee *et al.*, 2020) : This is a domain-specific language model obtained by training BERT on biomedical corpora including PubMed and PubMed Central (PMC).

— **PubMedBERT** (Gu *et al.*, 2020) : Contrary to continual pretraining as BioBERT, this model is pretrained from scratch using abstracts from PubMed. Its main advantage over BERT and BioBERT is its in-domain vocabulary, which helps it to identify medical terms (eg. *naloxone*, *acetyltransferase*) avoiding subword fragmentation [3].

— **BioELECTRA** (Kanakarajan *et al.*, 2021) : It is a biomedical domain-specific language encoder model that adapts ELECTRA (Clark *et al.*, 2020) by pretraining it from scratch with biomedical domain text from PubMed+PMC and uses domain specific vocabulary from PubMedBERT. It outperforms the previous models and achieves state-of-the-art (SOTA) in BLURB benchmark (Gu *et al.*, 2020).

Implementation Details In our experiments, we follow the random data splitting from the (Yang *et al.*, 2016) work by which we keep 20 random instances of each class for training, 500 random instances for development set, and 1000 constant instances for test. We report the average results for 5 seeded model runs. We keep the hyperparameter setting common across all the models. We use learning-rate (1e-06), epochs(500), maxlen (title=64, abst=512), and batch-size(title=64, abst=8).

4 Results & Discussion

Table 1 shows the results divided into 4 subsections. The first subsection shows *random* features to compare and show the impact of feature-based models. In the next subsection, we have *text-only* based features including *tfidf* and PLM-based models with two variants : [CLS]-finetuning and Text-CNN model. In the last two subsections, we report *graph-only* features and *text+graph*, which is a combination of text and graph features.

Random VS Features Comparing with *random* features, we notice the impact of text-only features (*tfidf-c/m*, PLM), graph-only features (*DW*) and combination of text+graph features (*tfidf+DW*, *GS*).

Text features We notice that *tfidf-c* performs better than *tfidf-m*, which can be attributed to the curated keyword list compared as automatic generated keyword list. We continue with *tfidf-m* to study **Title**- and **Abst**- text separately. We observed that *PLM*-based features perform predominately better than *tfidf* features with the exception of `BERT-base-case` (because of its general vocabulary) and further, all the **Abst**- models outperform the **Title**- based models showing the relevance of the Abstract section. We notice that surprisingly BioElectra performs lower than PubMedBERT for all cases (except for [CLS] baseline with title corpora) and its performance degrades with Text-CNN model.

3. Word shattering for out-of-vocab words eg. *acetyltransferase* → [ace, ##ty, ##lt, ##ran, ##sf, ##eras, ##e]

	Method	Title				Abstract			
		P	R	F1	Acc	P	R	F1	Acc
R	random	$0.36_{0.03}$	$0.33_{0.03}$	$0.33_{0.03}$	$0.32_{0.03}$	-	-	-	-
T	tfidf-c	$0.70_{0.02}$	$0.68_{0.03}$	$0.67_{0.03}$	$0.68_{0.03}$	-	-	-	-
	tfidf-m	$0.61_{0.03}$	$0.60_{0.02}$	$0.59_{0.03}$	$0.60_{0.02}$	$0.71_{0.02}$	$0.70_{0.03}$	$0.69_{0.03}$	$0.70_{0.03}$
	BERT	$0.41_{0.09}$	$0.43_{0.02}$	$0.36_{0.04}$	$0.43_{0.02}$	$0.45_{0.16}$	$0.50_{0.07}$	$0.43_{0.11}$	$0.50_{0.07}$
	BioBERT	$0.67_{0.04}$	$0.64_{0.06}$	$0.62_{0.09}$	$0.64_{0.06}$	$0.72_{0.08}$	$0.67_{0.14}$	$0.66_{0.17}$	$0.67_{0.14}$
	PubmedBERT	$0.71_{0.06}$	$0.68_{0.07}$	$0.67_{0.09}$	$0.68_{0.07}$	$0.86_{0.02}$	$0.86_{0.02}$	$0.86_{0.02}$	$0.85_{0.03}$
	BioElectra	$0.76_{0.03}$	$0.71_{0.04}$	$0.70_{0.05}$	$0.71_{0.04}$	$0.86_{0.02}$	$0.85_{0.02}$	$0.85_{0.02}$	$0.84_{0.03}$
	BERTCNN	$0.60_{0.07}$	$0.59_{0.08}$	$0.59_{0.08}$	$0.59_{0.08}$	$0.66_{0.05}$	$0.65_{0.05}$	$0.65_{0.05}$	$0.65_{0.05}$
	BioBERTCNN	$0.70_{0.06}$	$0.69_{0.06}$	$0.69_{0.06}$	$0.69_{0.06}$	$0.72_{0.03}$	$0.70_{0.03}$	$0.70_{0.04}$	$0.69_{0.05}$
	PubmedCNN	$0.83_{0.03}$	$0.82_{0.04}$	$0.82_{0.04}$	$0.82_{0.04}$	$0.87_{0.03}$	$0.87_{0.03}$	$0.87_{0.04}$	$0.87_{0.04}$
	BioElectraCNN	$0.65_{0.10}$	$0.62_{0.06}$	$0.58_{0.09}$	$0.62_{0.06}$	$0.85_{0.02}$	$0.84_{0.02}$	$0.84_{0.02}$	$0.84_{0.03}$
G	DeepWALK (DW)	$0.67_{0.03}$	$0.64_{0.03}$	$0.65_{0.03}$	$0.64_{0.03}$	-	-	-	-
	random+GraphSAGE(GS)	$0.17_{0.10}$	$0.36_{0.08}$	$0.21_{0.08}$	$0.36_{0.08}$	-	-	-	-
T+G	tfidf-c+DW	$0.68_{0.02}$	$0.65_{0.02}$	$0.65_{0.02}$	$0.65_{0.02}$	-	-	-	-
	tfidf-m+DW	$0.68_{0.02}$	$0.65_{0.02}$	$0.66_{0.02}$	$0.65_{0.02}$	$0.69_{0.02}$	$0.65_{0.02}$	$0.66_{0.02}$	$0.65_{0.02}$
	tfidf-c+GS	$0.78_{0.03}$	$0.76_{0.02}$	$0.76_{0.02}$	$0.76_{0.02}$	-	-	-	-
	tfidf-m+GS	$0.42_{0.02}$	$0.40_{0.04}$	$0.40_{0.04}$	$0.40_{0.04}$	$0.41_{0.01}$	$0.39_{0.02}$	$0.38_{0.01}$	$0.39_{0.02}$

R : *no-feature* ; T : *text-only* ; G : *graph-only* ; T+G : *text+graph*

TABLE 1 – Experiment results. Subscript denotes standard deviation across multiple runs

Graph Features We report *DeepWALK (DW)* model performance as we notice its informativeness is comparable to *tfidf* features. We also experiment with noise features as node-feature input to GraphSAGE(GS) model, and we notice drastic drop in performance compared to static graph features(DW). This indicates the impact of initial node-features for GS learning algorithm.

Text+Graph features Initially, we examine the concatenation of *DW* and *tfidf* and notice that *tfidf-m*(**Abst**)+*DW* performs comparable to *tfidf-m*(**title**)+*DW* but both the models perform lower than *tfidf-m*(**Abst**). Further, it is interesting to see that *DW* features perform better than *tfidf-m* trained on **Title**-text (when compared to *tfidf-m* only), but not for **Abstract**-text. Next, we experiment with *GS* with two settings, first with *tfidf-m* features we notice the performance decreases compared to *text-only* setting. This can be attributed to the iterative message passing during the process of node embedding generation can damage the original information. Lastly, with *tfidf-c* we notice an increased performance compared to *tfidf-m* model, which can be attributed to the curated list of keywords and spareness of *tfidf-m* features.

5 Conclusion

In this work, we provide a benchmark to study the informativeness of different features : text-based and citation information-based for document classification. We show that the models perform better with the Abstract text and we show that the information contained in the citation graph can be useful in addition to the text based features. In future, we would like to extend our study to investigate the relevance of other sections in scientific articles by combining graph models with PLMs. Additionally, the dynamic nature of the citation graphs can be interesting to understand the evolution and change points phenomena to study novelty and discovery in scientific citation graphs.

6 Acknowledgement

This work has been partly funded by the Project OLKI (Lorraine Université d'Excellence 2018-2021).

Références

CLARK K., LUONG M.-T., LE Q. V. & MANNING C. D. (2020). Electra : Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv :2003.10555*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

ERMAKOVA L., BORDIGNON F., TURENNE N. & NOEL M. (2018). Is the abstract a mere teaser ? evaluating generosity of article abstracts in the environmental sciences. *Frontiers Res. Metrics Anal.*, **3**, 16.

GALILEO MARK NAMATA, BEN LONDON L. G. & HUANG B. (2012). Query-driven active surveying for collective classification. In *International Workshop on Mining and Learning with Graphs*, Edinburgh, Scotland.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2020). Domain-specific language model pretraining for biomedical natural language processing.

GUO Y., KORHONEN A., LIAKATA M., SILINS I., HÖGBERG J. & STENIUS U. (2011). A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, **12**, 69 – 69.

HAMILTON W., YING Z. & LESKOVEC J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, **30**.

HUANG P.-S., HE X., GAO J., DENG L., ACERO A. & HECK L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, p. 2333–2338.

JIN W., ZHANG C., SZEKELY P. A. & REN X. (2019). Recurrent event network for reasoning over temporal knowledge graphs. *ArXiv*, **abs/1904.05530**.

JOHNSON R., WATKINSON A. & MABE M. (2018). The stm report : An overview of scientific and scholarly publishing.

KANAKARAJAN K. R., KUNDUMANI B. & SANKARASUBBU M. (2021). BioELECTRA :pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 143–154, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.16.

KIPF T. & WELLING M. (2017). Semi-supervised classification with graph convolutional networks. *ArXiv*, **abs/1609.02907**.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LIN J. J. (2009). Is searching full text more effective than searching abstracts ? *BMC Bioinformatics*, **10**, 46 – 46.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, **26**.

MONDAL I., HOU Y. & JOCHIM C. (2021). End-to-end construction of nlp knowledge graph. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1885–1895.

PENG Y., YAN S. & LU Z. (2019). Transfer learning in biomedical natural language processing : an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv :1906.05474*.

PEROZZI B., AL-RFOU R. & SKIENA S. (2014). Deepwalk : Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 701–710.

SINHA A., OLLINGER S. & CONSTANT M. (2022). Word sense disambiguation of french lexicographical examples using lexical networks. In *TextGraphs-16 : Graph-based Methods for Natural Language Processing*, p. 70–76.

THAKUR N., REIMERS N., RÜCKLÉ A., SRIVASTAVA A. & GUREVYCH I. (2021). Beir : A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv :2104.08663*.

WU L., CHEN Y., SHEN K., GUO X., GAO H., LI S., PEI J., LONG B. *et al.* (2023). Graph neural networks for natural language processing : A survey. *Foundations and Trends® in Machine Learning*, **16**(2), 119–328.

YANG Z., COHEN W. & SALAKHUDINOV R. (2016). Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, p. 40–48 : PMLR.

YEGANOVA L., KIM W. G., COMEAU D., WILBUR W. J. & LU Z. (2021). Measuring the relative importance of full text sections for information retrieval from scientific literature. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 247–256, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.27.

YU L., HERMANN K. M., BLUNSOM P. & PULMAN S. (2014). Deep learning for answer sentence selection. *arXiv preprint arXiv :1412.1632*.

ZHANG Y., RAHMAN M. M., BRAYLAN A., DANG B., CHANG H.-L., KIM H., MCNAMARA Q., ANGERT A., BANNER E., KHETAN V. *et al.* (2016). Neural information retrieval : A literature review. *arXiv preprint arXiv :1611.06792*.

ZHANG Y. & WALLACE B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv :1510.03820*.