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

Yanis Labrak, Mickaël Rouvier, Richard Dufour

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 Université (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

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).

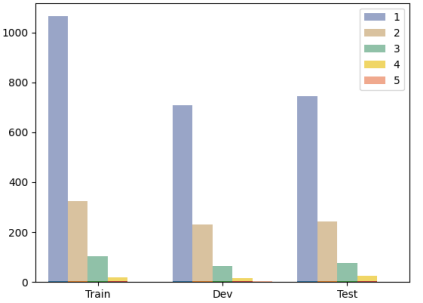
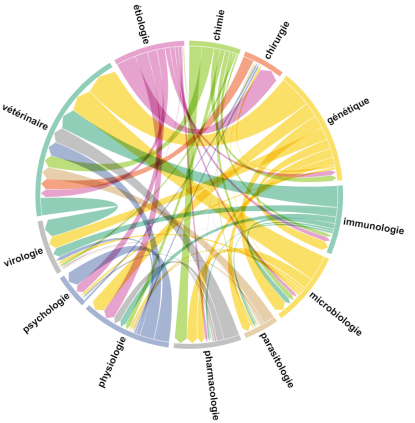


FIGURE 1 – Distribution du nombre de spécialités par article et par sous-ensemble de données : Apprentissage (Train), Développement (Dev) et Test (Test).

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.



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