Reconnaissance des entités nommées pour l'analyse des pharmacopées médiévales

Karim El Haff[*,***] Wissam Antoun[**], Florence Le Ber[*] Véronique Pitchon[***]

[*] Université de Strasbourg, ENGEES, CNRS, UMR 7357 ICube, F 67000 Strasbourg
kelhaff@unistra.fr, florence.le-ber@unistra.fr
[**] Inria-Paris, 75012 Paris, France
wissam.antoun@inria.fr
[***] Université de Strasbourg, CNRS, UMR 7044 Archimède, F 67000 Strasbourg
pitchon@unistra.fr

Résumé. Aujourd'hui, de nombreux projets se focalisent sur l'application des technologies linguistiques sur des corpus de médecine moderne surtout en matière de reconnaissance des entités nommées. Par ailleurs, les pharmacopées anciennes sont explorées avec une saisie manuelle des données par des spécialistes d'histoire et de biologie pour en retirer des connaissances. Ces analyses sont réalisées sans nécessairement passer par la reconnaissance des entités nommées, ce qui pourrait pourtant accélérer l'exploration des manuscrits. Par conséquent, nous proposons ici un mariage entre les deux pratiques par : (1) la création d'un ensemble de données de reconnaissance d'entités nommées pour les traductions anglaises de pharmacopées arabes médiévales et (2) l'entraînement et l'évaluation de modèles de langue pré-entraînés sur plusieurs domaines.

1 Introduction

Les progrès réalisés dans le traitement automatique du langage naturel (TAL) ou *Natural Language Processing* (NLP), une branche de l'intelligence artificielle qui permet aux ordinateurs d'analyser des textes écrits, parlés ou imagés, permettent d'extraire et de traiter les informations d'un corpus dans une langue humaine. Ce type de technologie peut être utilisé pour effectuer diverses tâches telles que la traduction automatique, l'exploration de textes, la reconnaissance d'entités nommées, la synthèse automatique de textes, la simplification automatique de textes, l'analyse de sentiments, les chatbots intelligents et d'autres applications qui pourront répondre aux besoins d'exploration de corpus. Les technologies du TAL s'appliquent dans de nombreux domaines d'intérêt majeur, tel que la médecine, où de nouveaux médicaments sont sans cesse recherchés.

Dans ce projet, nous nous focalisons sur une application du TAL dans le monde médical et historique, et plus précisément, la reconnaissance des entités nommées dans les pharmacopées de la civilisation arabe médiévale.

La période médiévale, surtout en Europe, est considérée comme une période sombre de l'histoire. De ce fait, la médecine médiévale est souvent négligée, étant perçue comme pleine

de superstitions et sans rigueur académique. En contraste avec ceci, d'autres cultures médicales comme celle du monde arabe ont connu leur âge d'or pendant cette même époque médiévale. Cela conduit à poser l'hypothèse que l'exploration de ces ouvrages anciens pourrait mener à des découvertes intéressantes. Un exemple qui illustre ce potentiel est le résultat obtenu par Tu Youyou, lauréate du prix Nobel de physiologie/médecine en 2015, qui a découvert en 1972 l'artemisinine, utilisée pour combattre le paludisme, en explorant manuellement un texte chinois taoïste de l'année 340 écrit par Ge Hong et intitulé « Un manuel de prescriptions d'urgence à garder dans sa manche ».

En utilisant les outils et méthodes de traitement automatique de langues à notre disposition, nous pourrons potentiellement gagner en temps et en ressources pour explorer les connaissances du passé et obtenir de nouvelles connaissances. Pour cela, ce projet se focalise sur la création de données pour la reconnaissance des entités nommées avec des étiquettes adaptées à l'exploration des traductions anglaises de pharmacopées arabes médiévales. Nos contributions sont les suivantes :

— la création et la diffusion d'un ensemble de données de reconnaissance des entités nommées (Named Entity Recognition ou NER) pour les traductions anglaises de pharmacopées arabes médiévales ;
— l'entraînement et l'évaluation de modèles linguistiques pré-entraînés qui couvrent une grande variété de domaines.

L'article est organisé comme suit. Après un bref état de l'art (section 2), la section 3 présente les données et l'architecture du modèle, tandis que la section 4 décrit les expérimentations et les résultats.

2 État de l'art

L'un des principaux modèles utilisés pour l'application de la reconnaissance des entités nommées sur des corpus de médecine moderne est le modèle BioBERT (Lee et al., 2019). C'est un modèle de langage spécifique au domaine biomédical, pré-entraîné sur des corpus biomédicaux à grande échelle. BioBERT surpasse BERT et d'autres modèles de pointe précédents dans une variété de tâches de fouille de textes biomédicaux du fait de son pré-entraînement sur des corpus adaptés au domaine. Cependant, BioBERT ne serait pas efficace dans la reconnaissance d'entités nommées lorsqu'il est utilisé directement pour reconnaître les entités des corpus de pharmacopées médiévales du fait de leur carence en termes scientifiques du monde biomédical moderne : les remèdes sont décrits principalement par des noms de plantes et d'ingrédients à base animale ou minérale ; les symptômes sont dénommés par des appellations issues de la culture orale et souvent par des termes archaïques. D'autre part, bien que des travaux sur la médecine moderne soient représentés dans les projets NER (Wang et al., 2021), nous constatons un manque de jeux de données provenant de pharmacopées anciennes. Donc, il serait intéressant d'explorer cette voie technologique pour ce qui relève de ces connaissances anciennes.

Par ailleurs, un des travaux importants d'exploration manuelle de pharmacopées anciennes est celui de l'équipe interdisciplinaire de Harrison et al. (2015) qui a estimé que les sociétés médiévales utilisaient une série de substances naturelles pour traiter des symptômes identifiables aujourd'hui comme des infections microbiennes ; ceci serait donc reflété dans leurs pharmacopées. En effet, ils ont identifié et reconstitué un remède potentiel pour l'infection du

Staphylococcus aureus à partir d'un livre médical anglo-saxon du Xe siècle. Le remède a tué à plusieurs reprises les biofilms établis de *Staphylococcus aureus* dans un modèle *in vitro* d'infection des tissus mous. Il a également tué le *Staphylococcus aureus* résistant à la méticilline dans un modèle de plaie chronique chez une souris. L'efficacité du remède est liée à l'action combinée de plusieurs de ses ingrédients, ce qui démontre le potentiel des anciennes pharmacopées comme source de connaissances médicales. Plus récemment, le travail d'analyse des données d'une pharmacopée médiévale britannique par Connelly et al. (2020), avec une saisie manuelle des données, a permis de découvrir des motifs intéressants dans les corpus explorés.

Finalement, ces travaux montrent l'intérêt des pharmacopées anciennes pour la découverte de médicaments utiles. Dans la continuité de ce processus de réflexion, nous estimons que l'utilisation de la reconnaissance des entités nommées serait une étape nécessaire à l'exploration en masse de nos connaissances anciennes qui pourraient encore cacher des perspectives médicales valables de nos jours.

3 Méthodologie

3.1 Création des données

La méthode de reconnaissance des entités nommées (NER) consiste à transformer les *tokens* du corpus en des vecteurs qu'un modèle de langage va parcourir afin de détecter automatiquement les entités du même type dans un texte inconnu. Pour entraîner un modèle, il est nécessaire de fournir des données d'entraînement convenables pour ce domaine. Pour cela, nous avons fait le choix d'annoter l'intégralité des remèdes d'une pharmacopée : il s'agit de la traduction anglaise par Oliver Kahl de l'ouvrage « Dispensatory in the Recension of the ʻAḍudī Hospital » écrit par Sābūr ibn Sahl au IXe siècle (Kahl, 2009). Le corpus est un manuscrit médical qui décrit 292 remèdes ou préparations. Le corpus est issu du PDF de l'ouvrage traduit et a été converti en format texte à l'aide de l'outil PdfToText [1]. Ensuite, la tokenisation a été effectuée à l'aide de NLTK (Bird et al., 2009) pour préparer le corpus à l'annotation. Ce corpus a été nettoyé et annoté manuellement par le premier auteur pendant un mois et revu en profondeur par une experte historienne en médecine arabe médiévale. L'ensemble du corpus est constitué de 36 961 tokens qui sont ensuite annotés avec leurs étiquettes respectives (voir tableau 1 pour la répartition des entités trouvées dans le corpus).

Pour effectuer l'annotation, 4 types d'étiquettes ont été utilisés :

— Type : B-Type I-Type, la forme du remède (pastille, pilule, etc.) ;
— Sym : B-Sym I-Sym, un symptôme de maladie ;
— Ing : B-Ing I-Ing, un ingrédient utilisé ;
— Org : B-org I-org, un organe mentionné ;
— O : le token n'est pas du domaine.

Les données sont annotées dans le format IOB2 (abréviation de « inside, outside, beginning ») qui est un format commun pour le marquage de tokens dans une tâche de *chunking* en traitement automatique des langues. Le préfixe B- devant une étiquette indique que c'est le début d'une entité. Le préfixe I- devant une étiquette indique que le token annoté se trouve à l'intérieur d'une entité. Une étiquette O indique que le token n'appartient à aucune entité.

1. https://github.com/jalan/pdftotext

	Entités uniques	Quantité Total
Ingrédients	1252	3089
Organes	61	172
Symptômes	420	782
Types	55	396

TAB. 1 – *Quantité d'entités selon leur type*

3.2 Architecture du modèle

Nous exploitons les avancées récentes dans les architectures d'apprentissage profond pour le TAL en utilisant des modèles de l'état de l'art fondés sur des *transformers* (Vaswani et al., 2017) pré-entraînés sur des corpus massifs de textes. Nous suivons l'approche standard selon laquelle une tâche de NER est considérée comme une tâche de classification de tokens, où les tokens sont introduits dans un modèle qui produit un espace vectoriel où chaque token en entrée est représenté par un vecteur. Ces représentations sont ensuite passées par un classificateur linéaire pour prédire l'étiquette IOB2. Ensuite, nous effectuons un *fine-tuning* pour l'ensemble du modèle en intégrant la couche ajoutée à l'aide de l'ensemble de données NER que nous avons créé.

4 Expérimentations et résultats

4.1 Choix de modèle

Nous avons mené des expérimentations avec une grande variété de modèles de *transformers* pré-entraînés. Notre objectif est de tester l'influence de la nature des données et de la méthode de pré-entraînement sur les résultats du *fine-tuning*. Nous avons donc testé les modèles suivants :

— Le modèle original BERT [2] (Devlin et al., 2019) développé par Google, entraîné sur des corpus anglais provenant de Wikipedia et BookCorpus (Zhu et al., 2015)
— RoBERTa [3] (Liu et al., 2019), une version optimisée de BERT ayant plus de données d'entraînement et un nombre plus grand d'étapes d'entraînement.
— XLM-R [4] (Conneau et al., 2019), une version multilingue de RoBERTa.
— DeBERTaV3 [5] (He et al., 2021b,a), un modèle BERT à l'architecture modifiée qui a récemment atteint l'état de l'art sur la tâche SuperGlue (Wang et al., 2019)
— BioBERT [6] (Lee et al., 2019), un modèle BERT pour le domaine biomédical qui ajoute, au corpus de pré-entraînement, des résumés PubMed et des articles en texte intégral de PubMed Central (PMC).

2. https ://huggingface.co/bert-base-cased
3. https ://huggingface.co/roberta-base
4. https ://huggingface.co/xlm-roberta-base
5. https ://huggingface.co/microsoft/deberta-v3-base
6. https ://huggingface.co/dmis-lab/biobert-base-cased-v1.2

Hyper-Paramètres	Valeurs
Longueur de Séquence Max.	256
Taille de Lot (Batch Size)	32
Taux d'Apprentissage	{4,5,6,7}e-5
Ratio d'Échauffement	{0,0.1}
Planificateur (Scheduler)	{linéaire, cosinus}
Précision	FP16

TAB. 2 – *Hyper-paramètres utilisés pendant le fine-tuning*

4.2 Configuration de l'entraînement

Pour déterminer le meilleur modèle et le meilleur ensemble d'hyper-paramètres, nous avons effectué toutes nos expériences avec un ensemble de validation croisée à 5 *splits* sans aucun remaniement, car le remaniement au niveau des phrases entraîne une fuite de données entre les données d'entraînement et les données de validation. Le tableau 2 montre les différentes valeurs de l'ensemble des hyper-paramètres que nous avons utilisés ; chaque modèle est entraîné sur la combinaison des hyper-paramètres pour un maximum de 10 *epochs*, soit un total de 480 exécutions sur une durée de 33 heures, et seuls les résultats de la meilleure *epoch* sont pris en compte. Les expérimentations ont été exécutées en utilisant une RTX 3080Ti avec 12GB de RAM avec la bibliothèque HuggingFace Transformers (Wolf et al., 2020).

4.3 Résultats

Le tableau 3 rapporte le score F1 moyen, la précision et le rappel sur les 5 splits à partir du meilleur ensemble d'hyper-paramètres. DeBERTaV3 obtient la meilleure performance de tous les modèles mais aussi la moins variable, surpassant BioBERT qui a été entraîné sur un corpus du domaine médical. Nous remarquons également que le modèle multilingue XLM-R a obtenu de moins bonnes performances que son homologue anglais RoBERTa. Enfin, le modèle BERT original a donné les pires résultats. Lors de l'ajustement des hyper-paramètres, DeBERTa a constamment surpassé tous les autres modèles, quel que soit le jeu de valeurs des hyper-paramètres, ce qui montre l'avantage du *disentangled attention* amélioré et de la méthode de pré-entraînement de DeBERTaV3. Le tableau 4 rapporte le score F1 détaillé de DeBERTaV3 pour chaque étiquette.

Modèle	Précision	Rappel	F1
XLM-R	83.36 ± 2.53	84.97 ± 4.35	84.12 ± 2.92
BERT	83.09 ± 1.92	86.19 ± 5.40	84.26 ± 3.33
BioBERTv1.2	83.47 ± 1.52	85.93 ± 4.15	84.66 ± 2.46
RoBERTa	84.78 ± 2.34	86.39 ± 3.63	85.56 ± 2.14
DeBERTaV3	$\mathbf{85.78 \pm 1.15}$	$\mathbf{87.09 \pm 2.46}$	$\mathbf{86.03 \pm 1.55}$

TAB. 3 – *Valeur moyenne et écart-type des 5 répétitions pour le meilleur jeu d'hyper-paramètres*

Type	Précision	Rappel	F1	Support
Ingrédient	89.95	91.49	90.41	600
Organe	80.04	72.97	75.04	43
Symptôme	84.47	84.01	85.58	153
Type	89.66	89.51	89.75	82

TAB. 4 – *Les scores détaillés pour chaque étiquette du meilleur modèle (DeBERTaV3).*

4.4 Analyse des erreurs

Une analyse des résultats de DeBERTa montre que notre meilleur modèle a encore de mauvaises performances sur les classes Organe et Type. Ceci est dû à la nature déséquilibrée de notre jeu de données, où les classe Ingrédient et Symptôme sont beaucoup plus représentées (voir tableau 1). En effet, en explorant les résultats sur des données test pour faire cette analyse qualitative d'erreurs, nous remarquons qu'une des erreurs les plus fréquentes est la non-détection des entités de la classe Organe pour certaines occurrences, ou bien le remplacement de l'étiquette Organe par une autre étiquette.

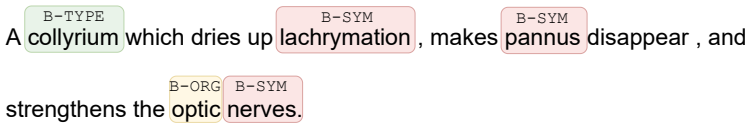

FIG. 1 – *Exemple d'erreur au niveau de la classe Organe*

Dans l'exemple de la figure 1, nous remarquons la division en deux parties de l'entité *optic nerves* qui devrait être de la classe Organe et la précision que le mot *nerves* est de la classe Symptôme. Ceci est probablement lié au fait que les entités de classe Symptôme sont plus représentées que les entités Organe formées de plusieurs mots. Dans d'autres cas, pour certaines occurrences, les entités de classe Type sont parfois non détectées. Ceci pourrait être dû au même problème que pour la classe Organe.

Pour diminuer ce type de problème dans cette tâche de reconnaissance d'entités nommées, nous pensons annoter plus de données et étendre ainsi le jeu de données d'entraînement.

5 Conclusion

Dans ce travail, nous avons procédé à la création d'un ensemble de données de reconnaissance d'entités nommées pour les traductions anglaises de pharmacopées arabes médiévales et nous avons entraîné et évalué des modèles de langue pré-entraînés pour explorer la possibilité d'obtenir des résultats satisfaisants sur les étiquettes customisées du domaine (Ing, Org, Sym et Type). Les expériences conduites démontrent que cette tâche a atteint le niveau de l'état de l'art selon les mesures usuelles avec tous les modèles testés, surtout avec DeBERTaV3 qui surpasse les autres modèles. Enfin, le modèle et le corpus seront mis à la disposition du public sur demande.

Cette conclusion nous incite à réfléchir aux nombreuses portes qui restent à ouvrir dans le domaine de l'exploration des pharmacopées médiévales, en particulier les manuscrits arabes traduits en langue anglaise. Nos objectifs futurs sont nombreux. En premier nous étendrons le jeux de données anglaises en exploitant d'autres manuscrits afin d'obtenir des résultats d'entraînement plus satisfaisants. Ensuite, nous pourrons étudier la faisabilité de travailler directement sur des manuscrits arabes, ce qui nécessiterait la mobilisation de modèles de la langue arabe tel que ARABERT (Antoun et al., 2020) et la comparaison des performances de ces modèles à ceux obtenus par les modèles anglais, pour un jeu de données de taille similaire.

Références

Antoun, W., F. Baly, et H. Hajj (2020). Arabert : Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, pp. 9.

Bird, S., E. Klein, et E. Loper (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, et V. Stoyanov (2019). Unsupervised cross-lingual representation learning at scale. *CoRR abs/1911.02116*.
en

Connelly, E., C. I. del Genio, et F. Harrison (2020). Data Mining a Medieval Medical Text Reveals Patterns in Ingredient Choice That Reflect Biological Activity against Infectious Agents. *mBio 11*(1), e03136–19.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Harrison, F., A. E. L. Roberts, R. Gabrilska, K. P. Rumbaugh, C. Lee, et S. P. Diggle (2015). A 1,000-Year-Old Antimicrobial Remedy with Antistaphylococcal Activity. *mBio 6*(4), e01129–15.

He, P., J. Gao, et W. Chen (2021a). DeBERTaV3 : Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

He, P., X. Liu, J. Gao, et W. Chen (2021b). DEBERTA : Decoding-Enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.
ar

Kahl, O. (2009). *Sabur Ibn Sahl's Dispensatory in the Recension of the Adudi Hospital*. BRILL. Google-Books-ID : RyhPkshiaLQC.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, et J. Kang (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, btz682. arXiv :1901.08746 [cs].

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, et V. Stoyanov (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv*

preprint arXiv :1907.11692.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems 30*, 5998–6008.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, et S. R. Bowman (2019). Superglue : A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv :1905.00537*.

Wang, B., Q. Xie, J. Pei, P. Tiwari, Z. Li, et J. Fu (2021). Pre-trained Language Models in Biomedical Domain : A Systematic Survey. *CoRR abs/2110.05006*.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, et A. M. Rush (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, Online, pp. 38–45. Association for Computational Linguistics.

Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, et S. Fidler (2015). Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Summary

Today, many projects focus on the application of linguistic technologies on modern medical corpora, especially in the field of Named Entity Recognition. Besides, ancient pharmacopoeias are being explored with manual data entry by specialists in history and biology in order to extract knowledge. These analyses are carried out without necessarily going through the automatic recognition of named entities which could accelerate the exploration of the manuscripts. Therefore, we propose here a link between the two practices by: (1) creating a named entity recognition dataset for English translations of medieval Arabic pharmacopoeias and (2) training and evaluating language models that are pre-trained on multiple domains.