Extraction et analyse de concepts médicaux dans un corpus de spécialité en orthophonie

Tiphaine Le Clercq de Lannoy, Romaric Besancon, Olivier Ferret, Julien Tourille, Frédérique Brin-Henry, Bianca Vieru

Extraction et analyse de concepts médicaux dans un corpus de spécialité en orthophonie

Tiphaine Le Clercq de Lannoy[1] Romaric Besançon[1] Olivier Ferret[1]
Julien Tourille[1] Frédérique Brin-Henry[2] Bianca Vieru[1]
(1) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(2) CH Bar le Duc, ATILF UMR7118 CNRS-Université de Lorraine, 54063 Nancy, France
{prénom.nom}@cea.fr, fhenry@atilf.fr

1 Introduction

L'émergence de gros modèles de langue pré-entraînés tels que BERT (Devlin *et al.*, 2019) a développé la définition et l'application de stratégies d'apprentissage par transfert (*transfer learning*), en particulier par le biais de la notion d'affinage (*fine-tuning*). Bien que ce développement facilite l'apprentissage de modèles pour des domaines spécialisés à partir de modèles plus généraux, cet apprentissage souffre toujours de l'absence de données annotées en quantités suffisantes. Dans cet article, nous nous focalisons plus spécifiquement sur le domaine de la santé et sur la tâche de reconnaissance d'entités nommées en français. Nous explorons plus précisément deux voies pour faciliter l'adaptation aux domaines spécialisés. La première reprend l'idée, explorée initialement par Gururangan *et al.* (2020), qu'utiliser un corpus non annoté du domaine cible et l'utiliser afin de poursuivre l'entraînement d'un modèle pré-entraîné sur sa tâche de modélisation du langage permet de spécialiser ce modèle pour ce domaine et d'améliorer les résultats de l'affinage sur la tâche finale visée. Cette approche a été appliquée en particulier par Copara *et al.* (2020) pour la reconnaissance d'entités nommées médicales en français.

La seconde voie exploite quant à elle les connaissances existant pour le domaine cible, connaissances qui sont particulièrement riches dans le cas du domaine médical et de la santé. Plus précisément, parmi les nombreux travaux réalisés pour utiliser conjointement les modèles de langue neuronaux et des connaissances données a priori (Yin *et al.*, 2022; Wei *et al.*, 2021; Yang *et al.*, 2022), se distinguent les approches que l'on peut qualifier de précoces, visant à injecter les connaissances directement au sein des modèles, soit lors de leur construction, soit a posteriori, des approches dites tardives dans lesquelles modèles de langue et connaissances sont fusionnés au niveau des résultats liés à la tâche. Nous nous situons dans cette seconde perspective en nous distinguant néanmoins des approches de type auto-apprentissage (Gao *et al.*, 2021) dans lesquelles les connaissances sont utilisées pour réaliser une forme d'augmentation de données.

De plus, nous appliquons les techniques étudiées à un corpus d'orthophonie, OrthoCorpus (2019), afin d'analyser les extractions d'entités nommées sur des cas concrets, du point de vue de l'intérêt clinique de la démarche et de sa faisabilité pour les experts du domaine. D'un point de vue disciplinaire, cela permet en effet de questionner le classement conceptuel en santé dans un sous-domaine spécifique au carrefour des sciences biomédicales et des sciences humaines et sociales. L'examen des formes et du statut des candidats-termes [1] nous renseigne sur la langue de spécialité (L'Homme, 2011).

1. ISO 1087-1:2000 : un terme est une désignation représentant un concept général d'un domaine spécifique ou d'un sujet.

Plus précisément, au travers des contributions de cet article, nous montrons, pour la reconnaissance d'entités nommées dans le domaine de la santé, que :

— l'utilisation de corpus spécialisés pour l'adaptation de modèles de langue pré-entraînés peut être intéressante, même pour des corpus que l'on peut qualifier de petits vis-à-vis des expérimentations de Gururangan *et al.* (2020) ;
— différents modèles neuronaux et une approche à base de connaissances présentent des profils complémentaires qu'une combinaison tardive permet de valoriser.

2 Approche

Pour entraîner, malgré des données annotées en quantité limitée, un modèle de langue spécialisé pour la reconnaissance d'entités nommées médicales en français, nous nous appuyons principalement sur deux éléments : l'exploitation de connaissances structurées dans le domaine de la santé, sous forme principalement de thésaurus (cf. section 2.1), et l'adaptation d'un modèle de langue au domaine de la santé (cf. section 2.2). Comme les deux approches présentées précédemment reposent sur des techniques très différentes, les résultats obtenus par chacune d'entre elles peuvent se compléter efficacement (cf. section 2.3).

2.1 Exploitation de connaissances

Pour notre approche à base de connaissances, nous avons retenu une méthode comparable à QuickUMLS (Soldaini & Goharian, 2016), fondée sur la projection dans le corpus cible d'une terminologie de référence, structurée selon les types d'entités visés. Une dimension essentielle de cette approche est donc la constitution de cette terminologie, structurée dans notre cas selon les dix groupes de types sémantiques de l'UMLS [2] (Unified Medical Language System (Lindberg *et al.*, 1993)) retenus pour annoter le corpus QUAERO (Névéol *et al.*, 2014), utilisé dans nos évaluations. L'orthophonie étant une profession de santé, l'application de ces mêmes groupes de types sémantiques permet la comparaison des résultats dans ce domaine de spécialité avec ceux obtenus plus généralement dans le domaine médical. Il s'agit plus précisément des groupes : *Anatomy*, *Chemicals & Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology et Procedures*.

Cette terminologie est issue de plusieurs sources, à commencer bien entendu par l'UMLS lui-même puisque ce dernier contient un ensemble significatif de termes en français pour les dix groupes considérés, en particulier issus des terminologies MeSH, MedDRA et LOINC. Nous utilisons également les ressources constituées par Embarek & Ferret (2008) pour les types d'entités *Anatomy*, *Chemicals & Drugs*, *Disorders* et *Procedures*. Nous avons par ailleurs exploité les données du site de la base de données publique des médicaments (ANSM, 2021), qui référence tous les médicaments en vente sur le marché français ou en arrêt de commercialisation depuis moins de trois ans. Cette base nous a ainsi permis d'étoffer le type *Chemicals & Drugs* avec des médicaments et le type *Disorders* avec certaines pathologies. Pour finir, le site PasseportSante [3], recommandé par RESEAU CHU [4], nous a permis d'obtenir des termes plus grand public pour les types *Anatomy*, *Disorders* et *Procedures*.

Pour identifier dans les textes les types d'entités considérés à partir de ces ressources terminologiques,

2. https://www.nlm.nih.gov/research/umls/index.html
3. https://www.passeportsante.net/
4. https://www.reseau-chu.org/

nous avons défini et implémenté l'outil QuickMatching, fondé sur l'algorithme SimString (Okazaki & Tsujii, 2010), à l'instar de QuickUMLS. Cet outil calcule la similarité entre des termes de référence et les mots des textes sur la base d'un découpage en n-grammes. Cette mesure de similarité permet d'apparier un terme de référence avec un mot du texte en faisant abstraction de différences minimes, comme celles résultant de variations morphologiques mineures ou de fautes de frappe.

2.2 Adaptation de modèles de langue

Nous traitons la tâche d'identification des types d'entités médicales visés comme une tâche d'annotation de séquences au format BIO. Pour cela, nous reprenons l'architecture de Devlin *et al.* (2019) pour la tâche de reconnaissance d'entités nommées mais en utilisant CamemBERT (Martin *et al.*, 2020) comme modèle de langue initial. Pour l'adaptation de ce dernier au domaine médical, nous poursuivons sa tâche de modélisation du langage sur un corpus de textes du domaine de la santé. Nous présentons dans le tableau 1 les corpus utilisés pour cette adaptation, sélectionnés à la fois pour la possibilité de les obtenir facilement et l'absence de difficulté vis-à-vis de la problématique des données personnelles. Pour étudier à la fois l'influence de la taille des corpus et de leur nature sur une telle adaptation, nous avons entraîné un modèle spécifique pour chaque corpus ainsi qu'un modèle s'appuyant sur l'ensemble de ces corpus, ce qui représente un peu plus de 136 millions de mots.

Corpus	Description	Taille
OrthoCorpus (2019)	Articles de la revue spécialisée Rééducation Orthophonique (Brin-Henry, 2018)	6,7M
ISTEX	Articles de revues médicales indexées par ISTEX (Inist)	42,6M
EQueR	Articles scientifiques et de recommandations de bonne pratique médicale (CISMeF)	16,8M
PMC OA	Articles de revues médicales (PubMed Central Open Access)	3,8M
Cochrane	Résumés d'articles de l'organisation Cochrane	5,0M
EMA	Notices de l'Agence Européenne des Médicaments	21,2M
CRTT	Articles de revues, extraits de Science Direct	21,7M
E3C-Corpus	Résumés d'articles, articles de revues, cas cliniques	12,1M
Wikipédia	Articles Wikipédia dans le domaine médical	6,6M

TABLE 1 – Collections de textes du domaine de la santé utilisées. Les tailles sont exprimées en millions de mots

2.3 Combinaison des approches

L'exploitation de connaissances permet d'extraire des termes avec une bonne précision mais repose sur la qualité et la mise à jour de la terminologie de référence qui la sous-tend. L'utilisation des modèles de langue permet en revanche de généraliser l'annotation des termes vus durant l'entraînement mais nécessite des corpus annotés pour cette tâche. Pour combiner plusieurs approches, nous unissons les entités prédites par ces approches avec une gestion minimale des conflits au niveau des types. Plus précisément, si deux approches identifient une entité de même empan mais avec un type différent, la priorité est donnée au type trouvé par le modèle neuronal. Nous avons en effet constaté que donner la priorité à QuickMatching se traduisait par une dégradation des performances de l'ordre d'un point de F1-mesure.

Une version améliorée de cette combinaison d'approches, le vote, utilisée par Copara *et al.* (2020), consiste à utiliser plus de deux modèles et à procéder à un vote pour chaque entité prédite. Si deux modèles ou plus prédisent une entité, elle est alors conservée dans la version finale, ce qui

permet de réduire le bruit par rapport à la méthode précédente. Pour ce vote, nous considérons l'outil QuickMatching, le modèle CamemBERT pré-entraîné sur tous les corpus ainsi que le modèle Pyramid (Wang *et al.*, 2020), entraîné sur le corpus QUAERO et utilisé pour générer des entités imbriquées.

3 Expérimentations et résultats

3.1 Cadre expérimental

Données Nous évaluons les méthodes proposées sur le corpus QUAERO, annoté en entités médicales pour le français et utilisé dans le challenge CLEF eHealth 2016 (Névéol *et al.*, 2016). Ce corpus est composé de dix documents sur des médicaments issus de l'European Medicines Agency (EMEA) ainsi que de 2 498 titres d'articles de recherche disponibles dans la base de données de MEDLINE. Les types d'entités utilisés pour annoter le corpus correspondent aux dix groupes de types sémantiques de l'UMLS évoqués à la section 2.1. Il est à noter que cette annotation comporte des entités imbriquées, le nombre de niveaux d'imbrication pouvant aller jusqu'à quatre. Il n'y a pas de restrictions sur les types utilisés dans les entités imbriquées. Pour l'évaluation, nous considérons toutes les entités au niveau de la référence. En revanche, nos deux méthodes de base se comportent de façon différente : tandis que QuickMatching peut identifier des entités imbriquées, notre modèle neuronal est entraîné pour identifier seulement les entités de plus large extension, ce qui le désavantage nécessairement du point de vue du rappel. Compte tenu de la méthode de fusion, son résultat comporte les entités imbriquées issues de QuickMatching.

Entraînement des modèles Pour la manipulation des modèles pré-entraînés, nous nous sommes appuyés sur la bibliothèque Transformers de HuggingFace (Wolf *et al.*, 2020). Concernant l'adaptation du modèle de langue CamemBERT, nous avons appliqué la tâche de Masked Language Modeling (MLM) en masquant des mots entiers, à l'instar de Martin *et al.* (2020), et non les seuls WordPieces. Nous avons utilisé l'optimiseur Adam (Kingma & Ba, 2015), avec $\beta_1 = 0,9$ et $\beta_2 = 0,98$. Le taux d'apprentissage (*learning rate*) était égal à 2.10^{-5}. Pour chaque corpus, nous avons réalisé 15 époques (*epochs*) de MLM en sauvegardant le modèle à la fin de chaque époque et avons sélectionné la version du modèle obtenant les meilleurs résultats sur le jeu de validation du corpus QUAERO.

Concernant l'affinage sur la tâche de reconnaissance d'entités médicales, nous avons utilisé l'outil Optuna (Akiba *et al.*, 2019) pour la recherche des meilleures valeurs d'hyperparamètres en prenant en compte la taille des lots, le nombre d'époques, le taux d'apprentissage et le ratio d'échauffement (*warm-up*). Nous avons ainsi obtenu la combinaison : taille de lot $= 9$, taux d'apprentissage $= 8.10^{-5}$ et ratio d'échauffement $= 0,224$. L'ensemble du jeu d'entraînement de QUAERO (EMEA et MEDLINE) a été utilisé pour réaliser chaque affinage de modèle pré-entraîné.

L'entraînement du modèle Pyramid a été réalisé grâce au code fourni par les auteurs [5] en adoptant le modèle de base associé à une pyramide inverse, le tout d'une profondeur 9. Les plongements utilisés sont ceux de fastText (Bojanowski *et al.*, 2017) en français [6] et le taux d'apprentissage était de 0,01.

Métriques d'évaluation Pour évaluer les résultats des modèles, nous avons utilisé l'outil BRAT-Eval ehealth fourni avec le corpus QUAERO. Cet outil a été développé par (Verspoor *et al.*, 2013) et modifié par (Névéol *et al.*, 2014). Les métriques considérées sont la précision (P), le rappel (R) et la F1-mesure (F1). Ce sont des micro-mesures calculées en mode strict.

5. https ://github.com/LorrinWWW/Pyramid
6. https ://fasttext.cc/docs/en/crawl-vectors.html

Modèle	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
QuickMatching	62,6	**66,4**	64,4	60,9	**61,7**	**61,3**
Pyramid (Wang *et al.*, 2020)	67,0	58,5	62,4	59,0	53,7	56,2
CamemBERT	72,6 ± 1,5	60,1 ± 1,8	65,8 ± 1,5	62,4 ± 1,0	48,9 ± 0,9	54,8 ± 0,8
OrthoCorpus	73,4 ± 1,6	59,4 ± 2,0	65,7 ± 1,2	**63,1 ± 1,9**	47,6 ± 1,3	54,2 ± 0,6
PMC OA	71,5 ± 1,4	59,9 ± 1,4	65,2 ± 1,0	61,1 ± 0,7	48,1 ± 1,8	53,8 ± 0,9
Cochrane	72,1 ± 1,3	59,8 ± 1,5	65,3 ± 0,9	61,4 ± 0,8	47,8 ± 0,8	53,7 ± 0,5
EQueR	72,3 ± 1,1	60,5 ± 0,8	65,9 ± 0,4	61,9 ± 0,9	49,0 ± 1,1	54,7 ± 1,0
ISTEX	72,4 ± 1,4	60,0 ± 1,3	65,6 ± 1,2	63,0 ± 0,8	49,0 ± 0,8	55,1 ± 0,7
CRTT	73,4 ± 0,6	60,4 ± 1,7	66,3 ± 1,1	62,5 ± 1,2	48,6 ± 1,0	54,7 ± 0,6
E3C-Corpus	75,1 ± 1,3	61,8 ± 1,4	67,8 ± 1,0	61,7 ± 1,0	47,9 ± 0,7	53,9 ± 0,3
Wikipédia	72,9 ± 2,0	60,4 ± 1,7	66,1 ± 1,8	62,1 ± 1,5	48,6 ± 0,3	54,5 ± 0,6
EMA	**75,4 ± 0,8**	61,8 ± 1,1	**67,9 ± 0,9**	61,7 ± 2,1	47,8 ± 2,0	53,8 ± 2,0
Tous les corpus	73,4 ± 0,4	62,2 ± 0,6	67,4 ± 0,4	62,2 ± 1,3	49,7 ± 0,9	55,3 ± 1,0

TABLE 2 – Comparaison des références (QuickMatching, Pyramid et modèle CamemBERT entraîné sur QUAERO sans pré-entraînement et avec affinage) et des modèles pré-entraînés avec différents corpus. Les résultats sont donnés sous la forme de moyennes et écarts-types obtenus en utilisant cinq graines aléatoires

3.2 Résultats et discussion

Résultats des approches de base Le tableau 2 présente les résultats sur le test du corpus QUAERO de nos deux approches de base (lignes QuickMatching et CamemBERT) ainsi que des expériences d'adaptation de notre modèle neuronal par pré-entraînement sur différents corpus du domaine de la santé. Dans le cas de notre modèle neuronal, la condition de base correspond à un affinage à partir du modèle CamemBERT, sans pré-entraînement complémentaire. Les résultats pour tous les modèles neuronaux correspondent à des moyennes pour cinq graines aléatoires.

Nous constatons en premier lieu que le modèle présentant en moyenne les meilleurs résultats à la fois sur EMEA et sur MEDLINE est le modèle pré-entraîné sur tous les corpus. S'il n'est pas le meilleur sur le corpus EMEA, il est tout de même gratifié du meilleur rappel. Le meilleur modèle sur EMEA est obtenu en pré-entraînant CamemBERT avec EMA. Or, ces deux corpus proviennent tous deux de l'Agence Européenne des Médicaments et comportent donc de fortes similarités au niveau des types de documents ainsi que de leurs sujets. Ces résultats confirment ainsi deux tendances de fond : les performances en affinage bénéficient d'autant mieux des effets d'un pré-entraînement en MLM que celui-ci se fait sur un gros corpus. Néanmoins, la spécificité de ce corpus par rapport aux données de test a aussi son importance et un corpus plus petit mais plus spécialisé peut s'avérer plus efficace.

Concernant QuickMatching, nous remarquons que les résultats sont assez constants entre EMEA et MEDLINE, contrairement aux approches fondées sur CamemBERT. Comme les mêmes ressources sont utilisées pour obtenir les résultats sur les deux corpus, nous pouvons supposer qu'elles couvrent de manière équivalente les deux corpus.

Résultats des combinaisons La table 3 compare les différentes méthodes de combinaison de modèles décrites dans la section 2.3, l'union et le vote, toutes les deux réalisées sur les modèles CamemBERT pré-entraîné avec tous les corpus, QuickMatching et Pyramid.

Nous pouvons constater que, quelle que soit la combinaison utilisée, le rappel augmente significative-

Combinaison	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
Union	56,3 ± 0,2	**82,4 ± 0,4**	66,9 ± 0,2	51,6 ± 0,6	**76,6 ± 0,7**	61,7 ± 0,6
Vote	**80,3 ± 0,1**	65,6 ± 0,4	**72,2 ± 0,2**	**77,2 ± 0,2**	58,7 ± 0,3	**66,7 ± 0,2**

TABLE 3 – Comparaison des combinaisons de CamemBERT pré-entraîné sur tous les corpus, QuickMatching et Pyramid. Les résultats sont donnés sous la même forme que dans le tableau 2

Modèle	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
(Afzal *et al.*, 2016)	75,1	76,1	75.6	71,1	62,5	66,5
(van Mulligen *et al.*, 2016)	71,6	78,5	74.9	68,0	71,6	69,8
EDS-fine-tuned (Dura *et al.*, 2022)	-	-	72,9	-	-	59,7
(Chernyshevich & Stankevitch, 2015)	85,8	59,7	70.4	76,1	40,1	52,6

TABLE 4 – État de l'art sur le jeu de données QUAERO (métrique SeqEval pour Dura *et al.* (2022), BRAT-Eval ehealth pour les autres)

ment (jusqu'à 15 points sur MEDLINE pour l'union). Nous faisons l'hypothèse que l'amélioration de la couverture grâce à la combinaison des entités issues des trois approches en est la cause. En revanche, les deux méthodes diffèrent sur la précision : elle diminue fortement pour l'union, tandis qu'elle augmente pour le vote, de 5 et 14 points pour EMEA et MEDLINE respectivement. Cela résulte de la sélection des entités, présente dans le vote mais pas dans l'union.

Pour les experts et d'éventuelles applications concrètes dans le domaine de la santé, la difficulté réside dans le juste équilibre entre la précision et le rappel pour les cas d'usage précités.

Comparaison avec l'état de l'art Pour finir, nous comparons nos résultats avec les résultats de l'état de l'art, obtenus principalement lors des campagnes d'évaluation CLEF eHealth. Si nos méthodes peuvent rivaliser avec certains systèmes, comme celui de Chernyshevich & Stankevitch (2015) pour les deux corpus, il faut remarquer qu'une approche très fortement fondée sur des dictionnaires complétés par une traduction de termes en anglais, en l'occurrence (Afzal *et al.*, 2016), obtient de bien meilleurs résultats que QuickMatching, laissant à penser que la couverture de nos terminologies est insuffisante. Sur un autre plan, les performances de Chernyshevich & Stankevitch (2015), qui mettent en œuvre la fusion d'un grand nombre de modèles de type Champs Aléatoires Conditionnels (CRF), nous poussent à considérer des stratégies de fusion plus élaborées que celle que nous avons expérimentée. Finalement, nous pouvons constater que l'utilisation de documents cliniques pour continuer le pré-entraînement d'un modèle CamemBERT (Dura *et al.*, 2022) permet un gain significatif par rapport à nos méthodes.

4 Analyse des concepts extraits pour l'orthophonie

Le vote (cf. section 2.3) a ensuite été appliqué sur OrthoCorpus (2019), un corpus contenant des articles de la revue spécialisée *Rééducation Orthophonique*. Afin d'analyser plus facilement les concepts ainsi extraits (cf. section 4.2), nous proposons un nouveau type de qualification, les patrons syntaxico-sémantiques, que nous présentons dans la section 4.1, associés à une évaluation qualitative

de leurs ambiguïtés sur le corpus d'entraînement de QUAERO.

4.1 Patrons syntaxico-sémantiques des concepts médicaux

Nous proposons dans un premier temps une exploration de la structure linguistique de formation des termes médicaux, sous la forme de l'extraction de patrons syntaxico-sémantiques caractérisant ces concepts.

Plus précisément, les noms et adjectifs de l'UMLS sont d'abord extraits, associés à leurs types sémantiques. D'autres classes d'adjectifs sont également ajoutées, comme les adjectifs de quantification. Les termes complexes sont alors représentés comme des patrons syntaxico-sémantiques associant les catégories grammaticales et les types sémantiques de l'UMLS : par exemple, on identifie le patron *NOUN_diso ADJ_anat* indiquant un nom de pathologie suivi d'un adjectif anatomique. Ces patrons sont alors associés à des types sémantiques spécifiques en les projetant dans une annotation de référence. La figure 1 montre ainsi que certains patrons sont peu ambigus alors que d'autres, comme *NOUN_proc ADJ_anat*, se retrouvent en pratique dans des concepts de types différents.

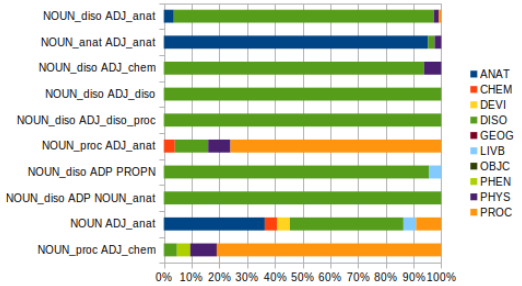

FIGURE 1 – Statistiques de projection de patrons syntaxico-sémantiques sur le corpus d'entraînement de QUAERO

En plus de fournir une analyse linguistique plus fine de la structuration syntaxico-sémantique des concepts médicaux, ces patrons sont également exploitables pour l'extraction automatique de concepts en santé. Ils offrent la possibilité de supprimer des annotations incohérentes avec les patrons, d'étendre des annotations trouvées selon les patrons ou de modifier les annotations selon les fréquences d'association des patrons avec les classes dans l'ensemble d'apprentissage. Dans nos expérimentations sur le corpus QUAERO, ces stratégies n'améliorent pas les résultats sur EMEA mais permettent de gagner jusqu'à deux points de f-score sur MEDLINE.

4.2 Analyse qualitative des candidats-termes classés

Un ensemble de plus de 11 000 candidats-termes a été collecté, dont plus de la moitié d'expressions polylexicales (Candito *et al.*, 2020), organisés selon dix des types sémantiques de l'UMLS. Une première analyse a montré que seuls certains de ces types ont une pertinence réelle du point de vue de la pratique de l'orthophonie (Anatomie, Pathologies, Physiologie, Procédures, Dispositifs, Êtres vivants).

De plus, on identifie bien certaines des spécificités du domaine : par exemple, on trouve *langue*, *cérébral*, *cordes vocales* et *système nerveux* parmi les termes d'Anatomie les plus fréquents, *trouble*

du langage et *trouble de la déglutition* parmi les Pathologies les plus citées, même si cette catégorie est très générale et mêle étiologie (*TC*, *lésion cérébrale*), symptômes (*modifications posturales*, *refus alimentaire*) et résultats (*échec scolaire*).

Par ailleurs, le problème de l'homonymie des termes est particulièrement présent, pour des raisons sémantiques mais également parce qu'aucune annotation syntaxique n'a été ajoutée. Ainsi le terme « marqueurs », identifié comme Nom est classé dans les produits chimiques, alors qu'il fait référence à des notions linguistiques (« des marqueurs du nom ») ou cliniques (« la présence de dysarthrie est un marqueur de mauvais pronostic »).

Cette analyse montre également qu'il reste beaucoup de bruit et d'imprécision dans l'extraction des termes (avec des termes manquants ou partiels) et qu'un travail manuel de sélection resterait nécessaire pour répondre aux besoins identifiés. En effet, les professionnels de santé (dont les orthophonistes) ont besoin d'explorer des documents non structurés de spécialité afin de réfléchir à leurs pratiques cliniques ainsi que de sélectionner et inclure des patients dans les études cliniques. Pour ce faire, la sélection fiable de termes pertinents est indispensable. La combinaison de méthodes automatiques et manuelles reste primordiale pour assurer une validité de la démarche.

5 Conclusions et perspectives

Nous avons présenté une approche hybride pour annoter des entités nommées dans le domaine médical. Cette approche combine un annotateur fondé sur des dictionnaires, un modèle de langue neuronal adapté au domaine avec des corpus de taille réduite et un modèle neuronal permettant de générer des entités imbriquées. Par ailleurs, pour ce qui est des modèles de langue neuronaux, nous avons montré qu'il n'est pas nécessaire d'avoir des corpus de grande taille pour observer une amélioration des résultats par rapport à un modèle dont le domaine n'a pas été adapté. Des études plus approfondies sur la similarité entre les corpus de test et les corpus utilisés pour l'adaptation permettront d'analyser plus finement ces résultats.

À plus long terme, nous continuerons à améliorer les modèles neuronaux par pré-entraînement sur des corpus non annotés ainsi qu'à enrichir les ressources de notre approche par dictionnaire. L'accès à des ressources de spécialité (Dictionnaire d'orthophonie) permettra également d'enrichir les tests sur ce domaine. Nous souhaitons également adapter le modèle CamemBERT aux entités imbriquées afin d'en améliorer la couverture. Enfin, le corpus DEFT 2020 (Cardon *et al.*, 2020) étant constitué de cas cliniques en français, nous souhaiterions l'utiliser pour tester les méthodes présentées afin d'évaluer leur potentiel d'adaptabilité à un type de corpus différent. Cela nous permettrait également de comparer les résultats ainsi obtenus à ceux de Copara *et al.* (2020), qui utilisent des méthodes similaires.

Remerciements

Ces travaux ont bénéficié d'un financement dans le cadre du programme e-Meuse Santé, porté par le Département de la Meuse et soutenu par les Départements de la Haute-Marne et de la Meurthe et Moselle, les GIP Objectif Meuse et Haute-Marne, la Région Grand Est, l'Agence Régionale de Santé Grand Est, et la Banque des Territoires au titre du programme France 2030. Ils ont été réalisés grâce au supercalculateur Factory-IA financé par le Conseil Régional d'Ile-de-France.

Références

AFZAL Z., AKHONDI S., HAAGEN H., VAN MULLIGEN E. M. & KORS J. (2016). Concept Recognition in French Biomedical Text Using Automatic Translation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9822, p. 162–173. DOI : 10.1007/978-3-319-44564-9_13.

AKIBA T., SANO S., YANASE T., OHTA T. & KOYAMA M. (2019). Optuna : A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

ANSM (2021). Base de données publique des médicaments.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BRIN-HENRY F. (2018). Pour une harmonisation de la terminologie orthophonique : contribution du projet OrthoCorpus (2015- 2017). In *Terminologica. TOTh 2018*.

CANDITO M., CONSTANT M., RAMISCH C., SAVARY A., GUILLAUME B., PARMENTIER Y. & CORDEIRO S. R. (2020). A French corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, **8**(2), 415–479. DOI : 10.15398/jlm.v8i2.265, HAL : hal-03016721.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In R. CARDON, N. GRABAR, C. GROUIN & T. HAMON, Éds., *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). Atelier DÉfi Fouille de Textes*, p. 1–13, Nancy, France : ATALA.

CHERNYSHEVICH M. & STANKEVITCH V. (2015). IHS-RD-BELARUS : Clinical Named Entities Identification in French Medical Texts. In *CLEF*.

COPARA J., KNAFOU J., NADERI N., MORO C., RUCH P. & TEODORO D. (2020). Contextualized French language models for biomedical named entity recognition. In *Actes de la 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). Atelier DÉfi Fouille de Textes*, p. 36–48, Nancy, France : ATALA et AFCP.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*.

DURA B., JEAN C., TANNIER X., CALLIGER A., BEY R., NEURAZ A. & FLICOTEAUX R. (2022). Learning structures of the french clinical language : development and validation of word embedding models using 21 million clinical reports from electronic health records. *ArXiv*, **abs/2207.12940**.

EMBAREK M. & FERRET O. (2008). Learning patterns for building resources about semantic relations in the medical domain. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA).

GAO S., KOTEVSKA O., SOROKINE A. & CHRISTIAN J. (2021). A pre-training and self-training approach for biomedical named entity recognition. *PLOS ONE*, **16**.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (2020). Don't Stop Pretraining : Adapt Language Models to Domains and Tasks. In *ACL 2020*.

KINGMA D. P. & BA J. (2015). Adam : A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

L'HOMME M.-C. (2011). Y a-t-il une langue de spécialité ? points de vue pratique et théorique. *Langues et linguistique*, p. 26–33.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The unified medical language system. *Methods of information in medicine*, **32**(4), 281–291.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *ACL 2020*.

NÉVÉOL A., COHEN K. B., GROUIN C., HAMON T., LAVERGNE T., KELLY L., GOEURIOT L., REY G., ROBERT A., TANNIER X. & ZWEIGENBAUM P. (2016). Clinical information extraction at the clef ehealth evaluation lab 2016. *CEUR workshop proceedings*, **1609**, 28—42.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *LREC Workshop BioTxtM2014*, p. 24–30.

OKAZAKI N. & TSUJII J. (2010). Simple and Efficient Algorithm for Approximate Dictionary Matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, p. 851–859, Beijing, China : Coling 2010 Organizing Committee.

ORTHOCORPUS (2019). ATILF. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

SOLDAINI L. & GOHARIAN N. (2016). QuickUMLS : a fast, unsupervised approach for medical concept extraction. In *SIGIR MedIR workshop*, p. 1–4.

VAN MULLIGEN E. M., AFZAL Z., AKHONDI S., VO-HAI D. & KORS J. A. (2016). Erasmus MC at CLEF eHealth 2016 : Concept recognition and coding in French texts. In *CEUR Workshop Proceedings*, p. 171–178 : CLEF.

VERSPOOR K., JIMENO YEPES A., CAVEDON L., MCINTOSH T., HERTEN-CRABB A., THOMAS Z. & PLAZZER J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, **2013**.

WANG J., SHOU L., CHEN K. & CHEN G. (2020). Pyramid : A layered model for nested named entity recognition. In *ACL 2020* : Association for Computational Linguistics.

WEI X., WANG S., ZHANG D., BHATIA P. & ARNOLD A. (2021). Knowledge Enhanced Pretrained Language Models : A Comprehensive Survey. *arXiv :2110.08455 [cs]*.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

YANG J., XIAO G., SHEN Y., JIANG W., HU X., ZHANG Y. & PENG J. (2022). A Survey of Knowledge Enhanced Pre-trained Models. *arXiv :2110.00269 [cs]*.

YIN D., DONG L., CHENG H., LIU X., CHANG K.-W., WEI F. & GAO J. (2022). A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models. *arXiv :2202.08772 [cs]*.