Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Stratégies d'adaptation pour la reconnaissance d'entités médicales en français

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

RÉSUMÉ

Dans un contexte où peu de corpus annotés pour l'extraction d'entités médicales sont disponibles, nous étudions dans cet article une approche hybride combinant utilisation de connaissances spécialisées et adaptation de modèles de langues en mettant l'accent sur l'effet du pré-entraînement d'un modèle de langue généraliste (CamemBERT) sur différents corpus. Les résultats sont obtenus sur le corpus QUAERO. Nous montrons que pré-entraîner un modèle avec un corpus spécialisé, même de taille réduite, permet d'observer une amélioration des résultats. La combinaison de plusieurs approches permet de gagner un à sept points de F1-mesure selon le corpus de test et la méthode.

ABSTRACT

Adaptation strategies for biomedical named entity recognition in French

In a context where few annotated corpora for medical entity extraction are available, we study in this paper a hybrid approach combining the use of specialized knowledge and language model adaptation ; furthermore, we study the effect of pretraining a general language model (CamemBERT) with different biomedical corpora. The methods are tested on the QUAERO corpus. We show that, even with a small corpus, pretrain a model with a specialized corpus can improve the results. The combination of several approaches allows to gain one to seven points on the F1-score depending on the test corpus and the method.

MOTS-CLÉS : Extraction d'information, Reconnaissance d'entités nommées, UMLS, BERT.

KEYWORDS: Information Extraction, Named Entity Recognition, UMLS, BERT.

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 suffisamment importantes. Dans cet article, nous nous focalisons plus spécifiquement sur le domaine médical 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 qui sont particulièrement riches dans le cas du domaine médical. Plus précisément, parmi les nombreux travaux réalisés pour utiliser conjointement les modèles de langue et des connaissances données a priori (Yin *et al.*, 2022; Wei *et al.*, 2021), se distinguent les approches que l'on peut qualifier de précoces, visant à injecter les connaissances directement au sein des modèles (lors de leur construction ou a posteriori), des approches dites tardives dans lesquelles modèles de langue et connaissances sont fusionnés au niveau des résultats. 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.

Plus précisément, au travers des contributions de cet article, nous montrons pour la reconnaissance d'entités nommées dans le domaine médical 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) ;
— un modèle neuronal et une approche à base de connaissances présentent des profils complémentaires qu'une fusion 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 sur deux éléments : l'exploitation de connaissances structurées dans le domaine médical, sous forme principalement de thésaurus (cf. section 2.1), et l'adaptation d'un modèle de langue au domaine médical (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 [1] (Unified Medical Language System (Lindberg *et al.*, 1993)) retenus pour annoter le corpus QUAERO (Névéol *et al.*, 2014), utilisé dans nos évaluations. 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 consi-

1. https://www.nlm.nih.gov/research/umls/index.html

dérés, 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 PasseportSanté [2], recommandé par RESEAU CHU [3], 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, 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 en reprenant 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 médicaux. Nous présentons dans la table 1 les corpus utilisés pour cette adaptation, sélectionnés pour leur facilité d'accès et l'absence de difficultés 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 ($\text{CamemBERT}_{\text{all}}$), ce qui représente un peu plus de 136 millions de mots. Comme l'ajout d'une couche de type Conditional Random Fields (CRF) « en sortie » de CamemBERT n'améliorait pas systématiquement les résultats, nous avons réalisé les expériences avec une simple couche linéaire et un softmax pour la classification des tokens.

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) (Ayache *et al.*, 2006)	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	Résumés d'articles, articles de revues, cas cliniques (Magnini *et al.*, 2021)	12,1M
Wikipédia	Articles Wikipédia dans le domaine médical (Bawden *et al.*, 2020)	6,6M

TABLE 1 – Collections de textes médicaux utilisées. Les tailles sont exprimées en millions de mots

2. https://www.passeportsante.net/
3. https://www.reseau-chu.org/

2.3 Fusion des deux 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 fusionner les deux approches, nous regroupons les entités qu'elles produisent avec une gestion minimale des conflits au niveau des types : si les 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 (environ un point de F1-mesure).

3 Expérimentations et résultats

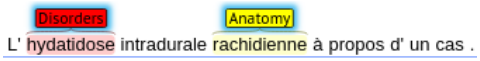
3.1 Cadre expérimental

Données Nous évaluons les méthodes proposées sur le corpus QUAERO, annoté en entités médicales pour le français. 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 (cf. figure 1a). La table 2 donne les statistiques caractéristiques de ce corpus. Les étiquettes utilisées pour annoter le corpus correspondent aux dix groupes sémantiques de l'UMLS évoqués à la section 2.1. 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 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.

	EMEA			MEDLINE		
	Train	Dev	Test	Train	Dev	Test
Documents	3	3	4	833	832	833
Tokens	14 944	13 271	12 042	10 552	10 503	10 871
Entities	2 695	2 260	2 204	2 994	2 977	3 103
Unique Entities	923	756	658	2 296	2 288	2 390
Unique CUIs	648	523	474	1 860	1 848	1 909

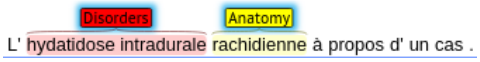
TABLE 2 – Description statistique du corpus QUAERO (Névéol *et al.*, 2016)

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 et non les seuls WordPieces (Martin *et al.*, 2020). 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 et sélectionné la version du modèle obtenant les meilleurs résultats sur le jeu de validation du corpus QUAERO sur la base des époques 5, 10 et 15.

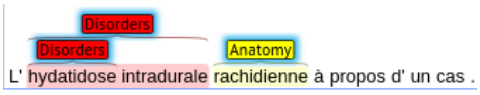


(a) Texte de référence du corpus de test de MEDLINE

(b) Texte annoté avec l'outil QuickMatching



(c) Texte annoté avec le modèle CamemBERT$_{all}$



(d) Texte annoté avec la fusion de QuickMatching et de CamemBERT$_{all}$

FIGURE 1 – Exemples de résultat des principales méthodes d'annotation (visualisation brat)

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 $= 0,224$.

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.

3.2 Résultats et discussion

Modèle	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
QuickMatching	65,8	46,4	54,4	62,1	**49,9**	**55,3**
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	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
all–{EMA,E3C}	72,1 ± 1,3	59,6 ± 1,1	65,2 ± 0,9	62,6 ± 1,3	48,1 ± 1,7	54,4 ± 0,9
all	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 3 – Comparaison des références (QuickMatching 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. Les meilleurs résultats sont en gras et les deuxièmes meilleurs sont soulignés.

Résultats des approches de base La table 3 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 du modèle neuronal par pré-entraînement sur différents corpus médicaux. Dans le cas de notre modèle neuronal, la condition de base correspond à un affinage à partir du modèle CamemBERT, sans pré-entraînement complémentaire. Comme QUAERO est composé de deux corpus différents, EMEA et MEDLINE, les résultats sur le test sont différenciés afin d'observer les spécificités de chacun, comme dans la campagne originale. Des exemples d'annotation sont présentés à la figure 1.

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 CamemBERT$_{all}$. 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.

Afin de mieux étudier l'effet de la spécialisation et de la taille du corpus de pré-entraînement, nous avons comparé l'effet d'un pré-entraînement utilisant uniquement EMA à celui d'un pré-entraînement utilisant tous les corpus à l'exception de EMA et E3C (ce dernier contenant des documents de EMA), totalisant environ 103 millions de mots (all–{EMA,E3C}). Nous pouvons constater qu'il obtient pour toutes les mesures des scores sur le corpus EMEA nettement inférieurs à ceux du modèle pré-entraîné avec EMA : il y a en particulier plus de 2 points de différence entre les deux F1-mesures. Dans ce cas, la spécialisation du corpus de pré-entraînement est préférable à la taille.

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.

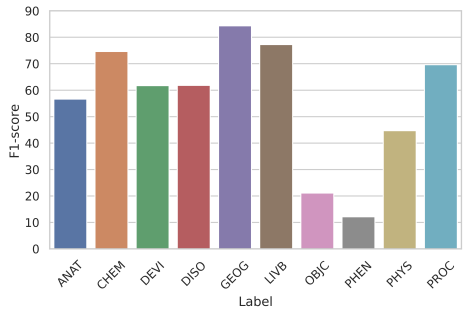
Fusion de QuickMatching avec	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
CamemBERT	63,3 ± 0,9	70,2 ± 1,1	66,6 ± 1,0	56,7 ± 0,6	67,4 ± 0,8	61,6 ± 0,6
CamemBERT$_{all}$	65,0 ± 0,3	72,9 ± 0,5	68,7 ± 0,3	56,8 ± 0,8	68,1 ± 0,8	62,0 ± 0,8

TABLE 4 – Comparaison des fusions. Les résultats sont sous la même forme que dans la table 3

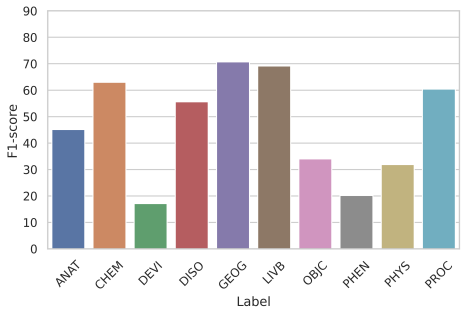
Résultats de la fusion La table 4 compare la fusion des résultats obtenus avec QuickMatching et des résultats du modèle CamemBERT de base et du modèle CamemBERT$_{all}$. Nous pouvons constater que la fusion améliore la mesure F1 des modèles CamemBERT d'environ un point sur le corpus EMEA et d'un peu moins de sept points sur le corpus MEDLINE en conservant la différence initiale entre les deux versions de CamemBERT. Le gain de performance est donc nettement plus notable pour le corpus MEDLINE, ce qui s'avère en fait tout à fait logique. Il s'agit en effet du corpus que les modèles CamemBERT ont le plus de mal à annoter mais sur lequel l'algorithme QuickMatching obtient les meilleures performances.

Quelle que soit la méthode, la fusion se traduit par une augmentation très significative du rappel par rapport aux méthodes initiales, jusqu'à 10 points, ce qui conduit à un rappel dépassant significativement la précision. Nous faisons l'hypothèse que l'amélioration de la couverture grâce à la combinaison des entités issues des deux approches en est la cause. En revanche, la précision diminue, ce qui est

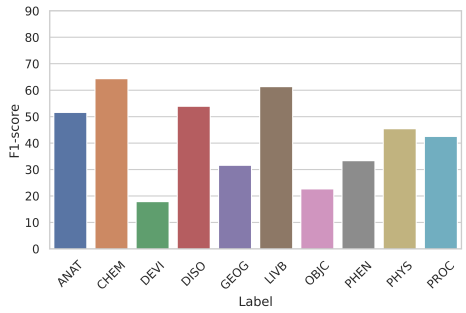
probablement dû au cumul du bruit des deux méthodes combinées. La combinaison d'annotations venant de deux méthodes différentes, ici l'exploitation de connaissances et l'adaptation de modèles de langue, permet donc bien d'améliorer significativement les résultats, même avec une fusion simple.



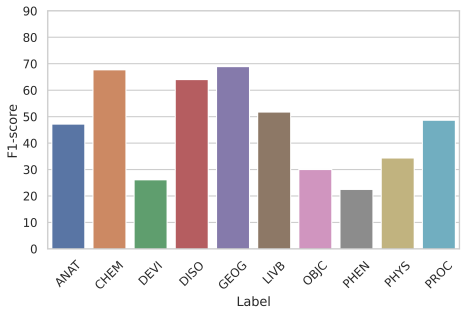
(a) Performance de CamemBERT$_{all}$ sur le test de EMEA



(b) Performance de CamemBERT$_{all}$ sur le test de MEDLINE



(c) Performance de QuickMatching sur le test de EMEA



(d) Performance de QuickMatching sur le test de MEDLINE

FIGURE 2 – Comparaison des performances par classe de QuickMatching et de CamemBERT$_{all}$ sur le test de EMEA et MEDLINE. Les résultats de CamemBERT sont des moyennes sur cinq essais.

Détail des performances des modèles sur les différentes classes La figure 2 détaille, pour les dix classes, les résultats des modèles CamemBERT$_{all}$ et QuickMatching sur EMEA et sur MEDLINE. Nous remarquons, sur le test de EMEA, que les performances des deux modèles sont très distinctes, souvent à l'avantage de CamemBERT$_{all}$. Le test de EMEA est composé de 4 documents divisés en 15 sous-documents. Cela signifie qu'un mot, par exemple le nom du médicament étudié, peut être présent de façon récurrente dans les sous-documents associés (soit un quart environ du corpus EMEA). Or, il suffit que ce mot ne soit pas présent dans les ressources de QuickMatching pour diminuer notablement les performances de la classe de ce mot. Par exemple, pour la classe Device, un tiers des faux négatifs contiennent le mot « dispositif », non présent dans les ressources de QuickMatching. Cela conduit à un rappel de 10 % et une F1-mesure de 17,86 %. Nous pouvons en conclure que la couverture des termes de QuickMatching n'est pas encore suffisante et doit être améliorée.

Concernant les modèles neuronaux, il faut noter que les textes de EMEA sont des notices de médicaments et suivent un patron défini. Par exemple, la phrase « Comment [médicament] est-il utilisé ? » est trouvée de façon récurrente dans les documents. Cela peut faciliter la généralisation de CamemBERT$_{all}$ face à des mots encore jamais rencontrés, en particulier pour les entités Chemicals & Drugs. Contrairement à QuickMatching en revanche, nous notons une baisse notable des résultats de CamemBERT$_{all}$ entre EMEA et MEDLINE pour toutes les classes (sauf pour Phenomena et Objects). Il serait donc

intéressant de réaliser par la suite une étude comme présentée dans (Kim & Kang, 2022) pour étudier la part de mémorisation et de généralisation que CamemBERT$_{all}$ parvient à réaliser.

Modèle	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
Cabot *et al.* (2016)-run1	53,8	37,8	44,4	54,0	47,6	50,6
Cabot *et al.* (2016)-run2	60,0	32,9	42,5	64,1	43,8	52,0
Ho-Dac *et al.* (2016)-run1	**78,4**	39,9	52,9	64,2	32,2	42,9
Ho-Dac *et al.* (2016)-run2	76,7	39,3	52,0	63,8	31,9	42,5
Mottin *et al.* (2016)-run1	52,3	18,4	27,2	57,1	44,2	49,8
Vivaldi *et al.* (2016)-run1	12,9	21,8	16,2	12,7	23,7	16,6
Vivaldi *et al.* (2016)-run2.unofficial	9,5	18,8	12,6	16,1	31,2	21,2
van Mulligen *et al.* (2016)-run1	62,3	**79,7**	69,9	61,7	69,0	65,1
van Mulligen *et al.* (2016)-run2	63,4	78,6	70,2	62,3	67,8	64,9
van Mulligen *et al.* (2016)-run3.unofficial	71,6	78,5	**74,9**	**68,0**	**71,6**	**69,8**

TABLE 5 – Résultats des participants de QUAERO (Névéol *et al.*, 2016), en précision (P), rappel (R) et F1-mesure (F1). Les meilleurs scores sont en gras et les deuxièmes meilleurs sont soulignés.

Comparaison avec l'état de l'art Pour finir, nous comparons les résultats obtenus 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 Ho-Dac *et al.* (2016) ou celui de Cabot *et al.* (2016), il faut remarquer qu'une approche très fortement fondée sur des dictionnaires complétés par traduction de termes en anglais, en l'occurrence (van Mulligen *et al.*, 2016), obtient de bien meilleurs résultats que QuickMatching, laissant de nouveau à penser que la couverture de nos terminologies est insuffisante. Finalement, il faut constater que l'avantage de modèles de type BERT observé pour l'anglais ne se retrouve pas nécessairement pour le français du fait de la faiblesse des corpus disponibles pour cette dernière langue.

4 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 et des modèles de langue neuronaux adaptés au domaine avec des corpus de taille réduite. Elle permet d'obtenir des entités imbriquées en combinant les deux modèles. Par ailleurs, pour ce qui est des modèles 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. 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

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.

AYACHE C., GRAU B. & VILNAT A. (2006). EQueR : the French evaluation campaign of question-answering systems. In *Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy : European Language Resources Association (ELRA).

BAWDEN R., DI NUNZIO G. M., GROZEA C., JAUREGI UNANUE I., JIMENO YEPES A., MAH N., MARTINEZ D., NÉVÉOL A., NEVES M., ORONOZ M., PEREZ-DE VIÑASPRE O., PICCARDI M., ROLLER R., SIU A., THOMAS P., VEZZANI F., VICENTE NAVARRO M., WIEMANN D. & YEGANOVA L. (2020). Findings of the WMT 2020 biomedical translation shared task : Basque, Italian and Russian as new additional languages. In *Fifth Conference on Machine Translation*, p. 660–687, Online : Association for Computational Linguistics.

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

CABOT C., SOUALMIA L. F. & DARMONI S. (2016). SIBM at CLEF eHealth Evaluation Lab 2016 : Extracting Concepts in French Medical Texts with ECMT and CIMIND. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, Évora, Portugal : CEUR-WS.org.

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.

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 *2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT 2019)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics.

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 *58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360, Online : Association for Computational Linguistics.

HO-DAC L.-M., TANGUY L., GRAUBY C., HEU MBY A., MALOSSE J., RIVIÈRE L., VELTZ-MAUCLAIR A. & WAUQUIER M. (2016). LITL at CLEF eHealth2016 : recognizing entities in French biomedical documents. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, Évora, Portugal : CEUR-WS.org.

KIM H. & KANG J. (2022). How do your biomedical named entity recognition models generalize to novel entities ? *IEEE Access*, **10**, 31513–31523.

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.

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

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2021). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. In J. MONTI, F. TAMBURINI & F. DELL'ORLETTA, Éds., *Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Collana dell'Associazione Italiana di Linguistica Computazionale, p. 258–264, Bologna, Italy : Accademia University Press.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online.

MOTTIN L., GOBEILL J., MOTTAZ A., PASCHE E., GAUDINAT A. & RUCH P. (2016). Bitem at clef ehealth evaluation lab 2016 task 2 : Multilingual information extraction. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, Évora, Portugal : CEUR-WS.org.

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 *Proceedings of the Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing*, 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**.

VIVALDI J., RODRIGUEZ H. & COTIK V. (2016). Semantic tagging and normalization of french medical entities. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, p. 179–192, Évora, Portugal : CEUR-WS.org.

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.

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]*.