Augmentation des modèles de langue français par graphes de connaissances pour la reconnaissance des entités biomédicales

Aidan Mannion[1,2] Didier Schwab[1] Lorraine Goeuriot[1] Thierry Chevalier[3]
(1) Laboratoire d'Informatique de Grenoble, Univ. Grenoble Alpes, CNRS, 38058 Grenoble, France
(2) EPOS SAS, 2-4 Boulevard Des Îles, 92130 Issy-les-Moulineaux, France
(3) UFR de Médecine Univ. Grenoble Alpes, Domaine de la Merci, 38700 La Tronche, France
prénom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Des travaux récents dans le domaine du traitement du langage naturel ont démontré l'efficacité des modèles de langage pré-entraînés pour une grande variété d'applications générales. Les modèles de langage à grande échelle acquièrent généralement ces capacités en modélisant la distribution statistique des mots par un apprentissage auto-supervisé sur de grandes quantités de texte. Toutefois, pour les domaines spécialisés à faibles ressources, tels que le traitement de documents cliniques, la nécessité d'intégrer des connaissances structurées reste d'une grande importance. Cette nécessité est d'autant plus grande pour les langues autres que l'anglais. Cet article se concentre sur l'une de ces applications spécialisées de la modélisation du langage à partir de ressources limitées : l'extraction d'informations à partir de documents biomédicaux et cliniques en français. En particulier, nous montrons qu'en complétant le pré-entraînement en mots masqués des réseaux neuronaux transformer par des objectifs de prédiction extraits d'une base de connaissances biomédicales, leurs performances sur deux tâches différentes de reconnaissance d'entités nommées en français peuvent être augmentées.

ABSTRACT

Unsupervised language model augmentation via knowledge graphs for French biomedical entity recognition

Recent work in natural language processing has demonstrated the effectiveness of large pre-trained language models for a wide variety of general applications. Large language models typically acquire these capabilities by modelling the statistical distribution of words via self-supervised training on large amounts of text. However, for low-resource settings, such as clinical document processing, particularly in languages other than English, the need to integrate structured domain knowledge remains of high importance. This paper focuses on one such low-resource specialised application of language modelling - information extraction from French-language clinical & biomedical documents. In particular, we show that supplementing the masked-language pre-training of transformer neural networks with prediction objectives extracted from structured biomedical knowledge graphs improves the downstream performance on two different named entity recognition tasks in French.

MOTS-CLÉS : TALN biomédical/clinique, extraction des informations, apprentissage automatique supervisé et auto-supervisé.

KEYWORDS: Biomedical/clinical NLP, information extraction, supervised & self-supervised machine learning.

1 Introduction

Les développements des dernières années en informatique et science des données représentent des pistes de recherche assez prometteuses pour le domaine de la santé, et l'utilisation généralisée des dossiers de santé électroniques (DSÉ) dans l'informatique médicale a beaucoup augmenté la disponibilité de dossiers cliniques en texte libre qui peuvent potentiellement être utilisés pour aider à la prise de décision par les professionel.les de santé (Wang *et al.*, 2019). Simultanément, les progrès dans le traitement automatique du langage naturel (TALN), notamment des approches basées sur l'apprentissage profond, ont rendu possible un traitement du texte beaucoup plus sophistiqué et efficace, et ont facilité l'amélioration de l'état de l'art dans plusieurs domaines d'application. En particulier, la modélisation de langage générale par des réseaux de neurones *transformer* (Vaswani *et al.*, 2017) s'est avérée très efficace pour de nombreuses applications. L'utilité de cette modélisation repose sur le paradigme d'apprentissage par transfert, où les représentations des données apprises sur une certaine tâche d'apprentissage, habituellement auto-supervisée, peuvent être utilisées d'une manière performante pour une autre tâche dans le même domaine d'application. Cette approche est d'autant plus utile pour les domaines à faibles ressources, tels que le domaine du traitement des textes cliniques, pour lesquels la quantité de données disponibles pour entraîner des algorithmes sur des tâches spécifiques n'est souvent pas suffisante (Dubois *et al.*, 2017). La communauté de la recherche biomédicale fait alors un usage de plus en plus intensif de ces techniques d'exploration de texte pour toute une série d'applications différentes (Ayala Solares *et al.*, 2020). Toutefois, la modélisation de langage faite uniquement à partir du texte libre se trouve souvent insuffisante pour bien intégrer dans ces modèles les connaissances du domaine clinique qui sont nécessaires pour un traitement efficace et fiable des documents cliniques (Sushil *et al.*, 2021). Dans ce travail, nous nous concentrons sur la tâche de la reconnaissance des entités nommées, l'une des tâches supervisées les plus étudiées dans le domaine du traitement des textes cliniques. Le travail d'adaptation des modèles de langage est divisé en trois parties :

— L'adaptation des modèles pour le domaine clinique par pré-entraînement sur un corpus de cas cliniques (section 2.2),
— L'intégration des connaissances du domaine par l'introduction des tâches supplémentaires - classification des triplets et prédiction des liens - basées sur le graphe de connaissances UMLS (Bodenreider (2004), section 2.3).
— L'adaptation des modèles pour les tâches de classification des mots par rapport à des catégories spécifiques au domaine (sections 2.4 et 3).

Nous étudions les performances de trois variantes de l'encodeur de texte BERT (Devlin *et al.*, 2019) sur trois tâches de reconnaissance des entités nommées. Les trois variantes pour lesquelles nous présentons des résultats dans ce papier sont CamemBERT (Martin *et al.*, 2020), FlauBERT (Le *et al.*, 2020), et une version de l'encodeur entraînée *from-scratch* sur un corpus de cas cliniques en français. Une description plus détaillée des corpus utilisés se trouve dans la section 1.3. Nous mettons à disposition le code et les modèles utilisés lors de nos expériences.

1.1 Encodeurs transformer pour le domaine clinique

Il existe plusieurs études qui adaptent l'architecture neuronale des transformers pour le domaine médical ; notamment Alrowili & Shanker (2021); Li *et al.* (2020); Lee *et al.* (2020) et Alsentzer *et al.* (2019). Ces travaux ont tous introduit des nouvelles variantes du modèle BERT et montrent des résultats au niveau de l'état de l'art sur des tâches telles que la prédiction de mortalité des patients, la

reconnaissance des entités médicales, la classification des documents, etc. Les travaux listés ci-dessus sont tous basés sur la langue anglaise et assez peu de projets ont été réalisés pour le français. Bien que le développement d'un modèle de langage français spécialisé pour le domaine biomédical soit un sujet de recherche très pertinent sur lequel quelques travaux ont déjà été menés (Berhe *et al.*, 2022; Dura *et al.*, 2022; El Boukkouri *et al.*, 2020), il n'existe pas de modèle de langage clinique en français librement disponible à la communauté de recherche, au moment de la rédaction de cet article. Il existe aussi quelques études sur l'adaptation des transformers français à des tâches de traitement de textes médicaux comme la reconnaissance des entités (Le Clercq de Lannoy *et al.*, 2022; Copara *et al.*, 2020) et la classification des documents (Chenais *et al.*, 2021), mais il s'agit d'un domaine qui n'a pas encore été pleinement exploré.

1.2 Travaux connexes : intégration des bases de connaissances biomédicales

Une grande partie des données liées au domaine de la santé et de la médecine est stockée sous une forme structurée, comme les ontologies ou les graphes de connaissances. L'importance d'exploiter ces données pour des applications d'apprentissage machine est donc largement reconnu dans ces domaines, particulièrement pour le traitement du texte (Chang *et al.*, 2020; Nicholson & Greene, 2020). Il a été démontré que les méthodes d'apprentissage combinant des représentations neuronales de texte avec des données biomédicales structurées améliorent les résultats obtenus dans une large gamme de tâches du traitement du texte médical en anglais (Naseem *et al.*, 2022; Meng *et al.*, 2021; Roy & Pan, 2021). Plusieurs stratégies pour intégrer les graphes de connaissances dans les modèles de langue de type BERT existent. Ces méthodes peuvent être divisées en deux types d'approches : **1)** des modifications de l'architecture neuronale du modèle pour intégrer des plongements de graphes de connaissance dans le même espace vectoriel que les plongements de mots (Peters *et al.* (2019), par exemple), et **2)** des modifications des données et de l'objectif de pré-entraînement pour que le modèle prenne en compte les informations contenues dans la base de connaissances en construisant ses plongements de mots, sans changer la structure interne du transformer. Dans ce travail, nous nous limitons aux approches du deuxième type, en nous positionnant dans un paradigme centré sur les données (Hamid, 2022). Autrement dit, nous exploitons uniquement la flexibilité des transformers de s'adapter à plusieurs modalités de données et plusieurs objectifs d'apprentissage. Nous pouvons conclure qu'avec des architectures qui sont généralisables de telle manière pour la modélisation de langage, il n'y a pas toujours besoin de changer leur structure pour les adapter aux types de données autres que le texte.

1.3 Données utilisées

Corpus de texte clinique Pour le pré-entraînement par mots masqués des modèles, nous utilisons la partie française du corpus E3C (European Clinical Case Corpus, Minard *et al.* (2021)), mis à disposition par l'organisation ELG (European Language Grid) [1]. Ce corpus consiste en des descriptions des cas cliniques tirés de revues médicales en français ainsi que des documents d'information sur les médicaments en provenance d'une base de données publique française. Nous nous limitons à ce corpus uniquement pour montrer l'efficacité de ces approches, même dans les situations pour lesquelles une quantité relativement faible de textes est disponible. Au total, il contient 25 740 documents. Pour nettoyer les documents de ce corpus, nous avons simplement enlevé les URLs et supprimé les phrases

1. https://live.european-language-grid.eu/catalogue/corpus/7618

contenant moins de quatre mots. Après ce nettoyage, le corpus entier de pré-entraînement se compose de 63,7 millions de mots et 2,7 millions de phrases.

Base de connaissances médicales Nous exploitons la base de connaissances UMLS (Système de langage médical unifié, (Bodenreider, 2004)) pour l'entraînement supplémentaire sur les données structurées. La partie française du métathésaurus UMLS (version 2022AB) contient 203 059 noms de concepts, catégorisés par 60 764 identifiants uniques (CUIs) et liés les uns aux autres par presque 1,5 millions de relations sémantiques structurées.

Corpus d'évaluation Nous utilisons deux tâches de classification des mots pour évaluer la capacité des modèles de reconnaître les entités médicales dans les documents cliniques, à partir des corpus médicaux annotés suivants :

1. **QUAERO :** (Névéol *et al.*, 2014) Pour cette évaluation, nous disposons du corpus de titres MEDLINE annotés dans le cadre de QUAERO, une ressource pour la reconnaissance et la normalisation des entités médicales. Ces titres sont annotés au niveau des mots avec des CUIs et des regroupements sémantiques d'UMLS.
2. **CAS-DEFT :** Ce corpus consiste en des annotations du corpus CAS (Grabar *et al.*, 2018) qui ont été mises à disposition dans le contexte du Défi Fouilles de Texte (DEFT) 2021 (Grouin *et al.*, 2021). Nous utilisons les annotations des mots en fonction des groupes sémantiques UMLS comme les étiquettes d'entraînement pour cette tâche.

Les tailles des partitions de ces corpus utilisées dans nos expériences sont détaillées dans le tableau 4.

TABLE 1 – Nombre de documents utilisés pour l'affinage des modèles de langue pour la reconnaissance des entités médicales.

	train	dev	test
QUAERO	788	790	787
CAS-NER	167	54	54

2 Méthodologie

2.1 Le Métathésaurus UMLS

Le métathésaurus UMLS est constitué de grandes quantités de concepts biomédicaux recoltés à partir de nombreuses sources d'informations biomédicales ontologiques. L'un de ses aspects les plus utiles est la manière dont il fournit des liens sémantiques entre des paires de concepts provenant de systèmes terminologiques différents. Cela permet notamment de regrouper sous le même identifiant de concept plusieurs termes différents provenant de vocabulaires médicaux différentes. Nous considérons le métathésaurus UMLS comme un graphe $\mathcal{G} = (C, R, E)$, où C est l'ensemble de concepts du métathésaurus, R est l'ensemble de types de rélations sémantiques qui peuvent exister entre les éléments de C, et E est l'ensemble de liens sémantiques qui existent dans le graphe. Nous générons alors nos données d'apprentissage à partir d'ensembles de triplets ordonnés $(h, r, t) \in E$, où $(h, r) \in C \times C$ et $r \in R$. Les valeurs possibles pour r, soit les types de regroupement sémantique du métathésaurus, sont les suivantes :

1. AQ : h peut qualifier t
2. QB : h peut être qualifié par t
3. PAR : h est un concept parent de t dans un des vocabulaires sources
4. CHD : t est un concept parent de h dans un des vocabulaires sources
5. RN : h a une définition plus étroite que t
6. RB : h a une définition plus large que t
7. SY : h et t sont synonymes

En utilisant les termes associés avec chaque concept $c \in C$, nous construisons des séquences textuelles à partir des triplets (h, r, t). Comme il peut y avoir plusieurs termes associés avec chaque concept, nous utilisons les *preferred terms* indiqués par le métathésaurus, sauf dans les cas des synonymes $r =$SY, auquel cas nous utilisons un autre terme associé avec le concept pour répresenter l'entité t.

2.2 Entraînement des modèles de langage

Pour effectuer l'entraînement par mots masqués des modèles de langue, nous séparons le corpus E3C décrit dans la section 1.3 en phrases, en coupant celles qui dépassent la longueur de séquence maximale. Pour les expériences détaillées dans ce travail, nous utilisons la configuration standard pour des modèles de type BERT ; 15% des mots masqués, taux d'apprentissage 2×10^{-5}, avec une longueur de séquence maximale de 256.

2.3 Entraînement avec une base de connaissance

À partir du métathésaurus UMLS, nous formulons deux objectifs de classification pour compléter le pré-entraînement des modèles de langue. Le premier pas pour faciliter l'intégration de ces tâches dans le processus d'entraînement est de redéfinir le vocabulaire et les tokens spéciaux utilisés par les modèles BERT. Pour les modèles que nous entraînons à partir de zéro, nous rajoutons la liste de tous les mots uniques apparaissant dans la partie française de l'UMLS (5 870 mots), en excluant ceux qui contiennent des caractères non alphabétiques, au vocabulaire initial tiré du corpus E3C. Ce vocabulaire supplementaire tiré du métathésaurus comprend les termes complexes (des concepts médicaux qui sont dénotés sous forme de groupes de mots plutôt que des mots uniques) ainsi que les termes simples. Cela permet aux transformers d'améliorer la modélisation des textes spécifiques au domaine médical et d'encoder plus directement les concepts du graphe de connaissance. Pour structurer les séquences d'entrées pour l'entraînement à base des relations sémantiques, nous rajoutons 8 tokens spéciaux aux vocabulaires des modèles ; un pour représenter chacun des types de relation qui apparaissent dans le graphe.

Classification des triplets En s'inspirant du travail de Hao *et al.* (2020), nous construisons un jeu de données pour la classification binaire des triplets (h, r, t) du graphe de connaissances comme étant vrai ou faux, où h et t sont les noms des concepts en question et r la relation entre eux. Nous tirons un échantillon des triplets du graphe comme exemples positifs. Ensuite, pour équilibrer le jeu de données avec des exemples négatifs, nous tirons des paires de concepts qui appartiennent au même groupe sémantique, mais pour lesquels une relation r n'existe pas. Afin de générer des exemples d'entraînement de faux triplets, nous utilisons deux stratégies différentes d'échantillonnage négatif.

Premièrement, pour former des exemples directement contrastés pour les relations existantes, nous échantillonnons les triples (h, r, t) où h et t appartiennent à différents groupes sémantiques et construisons des faux triplets correspondants avec le même type de relation et les mêmes catégories de groupe sémantique, c'est-à-dire $(\hat{h}, r, \hat{t}) \notin G$ où $\hat{h}$ et $\hat{t}$ appartiennent au même groupe sémantique que h et t, respectivement. Deuxièmement, afin de fournir des exemples contrastés pour les types de relation, nous échantillonnons des triplets pour lesquels h et t appartiennent au même groupe sémantique, et formons l'exemple d'entraînement négatif en changeant le type de relation r. Par souci de l'équilibre du jeu de données, les ensembles de données de classification des triples utilisés dans ce travail sont constitués de 50 % d'exemples positifs (triples réels du métathésaurus), de 25 % d'exemples générés par la première méthode d'échantillonnage négatif et du reste par la seconde. Les données d'entrée pour les transformers sont alors sous la forme `[CLS]` $w_1^h \cdots w_m^h$ `[REL]` $w_1^t \cdots w_n^t$ `[SEP]`, où `[CLS]` et `[SEP]` sont des tokens spéciaux standards pour l'encodage BERT, `[REL]` est le token spécial qui correspond à r, et les w_i^h et w_i^t sont les séquences de tokens émises par le tokenizer pour h et t respectivement. Le token `[CLS]` est transmis à une couche de classification linéaire qui produit une prédiction binaire, comme il est d'usage pour des tâches de classification avec des modèles BERT pré-entraînés.

Prédiction des entités Pour cette tâche de classification, nous complexifions le problème de classification des triplets décrit ci-dessus en le combinant avec l'objectif de prédiction des mots masqués. Les séquences d'entrées pour cette tâche ont la même structure que la précédente, mais au lieu de créer des exemples synthétiques des fausses relations et utiliser uniquement `[CLS]` pour prédire si la relation existe ou pas, nous masquons systématiquement les tokens w_i^t avec le même *mask token* utilisé pour l'entraînement à partir du texte libre. Le modèle aura alors comme objectif de prédire l'entité t d'une relation à partir de h et r. La formulation de ce dernier est partiellement basée sur celle de Meng *et al.* (2021).

Pour chacune des deux tâches décrites ci-dessus, nos expériences sont faites sur un jeu de données de 100K séquences, où les relations ont été tirées aléatoirement parmi les entrées françaises dans le tableau de relations `MRREL.RRF` du métathésaurus.

2.4 Affinage des modèles pour la reconnaissance des entités biomédicales

À partir des trois corpus d'évaluation mentionnés dans la section 1.3, nous effectuons l'affinage de bout-en-bout des modèles pour la classification des mots des cas cliniques en fonction des catégories spécifiques au domaine médical. Pour faciliter l'implémentation de ces deux tâches, nous avons pré-traité les documents pour que l'étiquetage des mots soit "plat", c'est-à-dire avoir une seule étiquette par mot. Pour effectuer ce prétraitement, nous avons utilisé directement l'ensemble des relations sémantiques UMLS pour remplacer des chevauchements par des concepts plus généraux.

Les annotations sont sous la forme BRAT (Stenetorp *et al.*, 2012), et peuvent alors être dénotées comme un ensemble $\mathcal{A} = \{a_i = (\text{CUI}^{a_i}, \text{SG}^{a_i}, s_1^{a_i}, s_2^{a_i})\} \subset C \times T \times \mathbb{N} \times \mathbb{N}$ où CUI et SG représentent l'identifiant du concept et le groupe sémantique respectivement. $s1$ et s_2 correspondent au *span* occupé par le mot dans le document en question (par souci de simplicité, nous nous limitons à la notation des spans continues, mais dans la pratique, nous traitons également des annotations discontinues). Pour aplatir les entités imbriquées, nous définissons un ensemble de concepts "de base" pour chaque groupe sémantique, en utilisant les types sémantiques, une catégorisation plus fine des concepts de l'UMLS. Ensuite, en utilisant les relations hiérarchiques (`PAR`/`CHD` et `RB`/`RN`), nous

pouvons calculer une mesure de généralité d'un concept par le nombre minimal d'étapes entre le concept et un des concepts de base. La procèdure de mise à plat des annotations $\mathcal{A}$ d'un document est alors la suivante :

1. Calculer l'ensemble (de taille n) des chevauchements entre mentions d'entités, effectivement un ensemble de combinaisons des éléments de $\mathcal{A}$:

$$\mathcal{O} = \{g_k = (a_1 \cdots a_m) : s_2^{a_1} \leq s_1^{a_2}, \ldots, s_2^{a_{m-1}} \leq s_1^{a_m}\}_{k=1}^{n}$$

2. Pour chaque élément g de $\mathcal{O}$, nous prenons ensuite l'ensemble $\phi(g) \subset C$ de concepts ayant une relation hiérarchique plus général (`PAR` ou `RB`) avec les concepts du chevauchement.
3. Les annotations imbriquées g sont alors remplacées avec l'élément de $\phi(g)$ le plus spécifique :

$$\text{argmax}_{\phi(g)} \left[\text{min}_{b \in B} d(g, b)\right]$$

où d est une fonction qui produit la "distance" hiérarchique entre deux concepts et B est l'ensemble de concepts de base. Si $\phi(g)$ est vide, nous ne gardons que l'annotation la plus courte parmi les éléments du chevauchement.

Tâche 1 : QUAERO-SG Dans cette tâche, l'objectif est de classifier les mots en fonction du groupe sémantique de l'UMLS auquel ils appartiennent. La granularité utilisée pour les annotations QUAERO nous donne dix regroupements, dont nous nous focalisons sur les suivants :

1. Désordre/trouble (`DISO`)
2. Procédure (`PROC`)
3. Structure anatomique (`ANAT`)
4. Substance chimique (`CHEM`)
5. Être vivant (`LIVB`)
6. Entité physiologique (`PHYS`)

En rajoutant une catégorie `none` pour les mots qui n'appartiennent à aucun de ces regroupements, nous obtenons alors un problème de classification avec une cardinalité de 7, pour lequel le nombre d'occurrences est affiché dans le tableau 2.

TABLE 2 – Nombre d'occurrences de chacune des catégories cibles dans la tâche QUAERO-SG.

Catégories QUAERO-SG	**train**	**dev**	**test**
DISO	1067	963	1144
PROC	741	748	719
ANAT	486	460	474
CHEM	377	395	363
LIVB	323	345	340
PHYS	177	173	157
Total	3171	2911	3040

Tâche 2 : CAS-CATEG Les cas cliniques de ce corpus sont annotés avec 14 types d'entités, 7 étant utilisées pour la tâche de reconnaissance des entités, présentées dans le tableau 3.

TABLE 3 – Nombre d'occurrences de chacune des catégories cibles dans la tâche CAS-CATEG.

Catégories CAS-CATEG	train	dev	test
sosy (signe ou symptôme)	13 977	4 940	4 297
examen	2 997	1 054	917
pathologie	2 064	460	238
traitement	1 824	859	884
moment	1 783	467	478
substance	1 517	542	346
dose	1 122	131	58
Total	25 284	8 453	7 218

3 Expériences

L'entraînement par mots masqués a été lancé à partir de deux modèles de langue pré-entraînés sur les corpus en langue française du domaine général ; les versions *base* de CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020). Nous comparons les performances de ces deux modèles avec une version du modèle DistilBERT (Sanh *et al.*, 2019) initialisé de zéro avec les mêmes paramètres d'architecture que `camembert-base`. Toutes les trois architectures ont le même nombre (12) de couches neuronales et de têtes d'attention, et la même taille de vecteurs de sortie ($d = 768$). La seule différence architecturale réside dans la valeur de la dimension intérieure : $H = 2048$ pour FlauBERT, contre $H = 3072$ pour les deux autres. Pour l'entraînement par mots masqués effectué sur le corpus E3C, nous avons entraîné chacune de ces variantes pendant 64 époques avec la méthode d'optimisation Adam avec pondération (Loshchilov & Hutter, 2017), ce que nous utilisons également pour le reste des entraînements et affinage.

Pour la tâche de classification des triplets, nous entraînons les modèles pendant 8 époques avec un taux d'apprentissage de 10^{-4}. La tâche de prédiction des entités étant plus complexe et ayant une dimension de sortie plus élevée, la décroissance de la valeur de la fonction de perte a atteint un palier autour de 24 époques, ce qui est effectivement la durée d'apprentissage pour laquelle les résultats sont présentés (le taux d'apprentissage utilisé est identique à celui de la classification des triplets).

Pour les étapes finales d'affinage pour la reconnaissance des entités, nous avons entraîné les modèles de bout-en-bout avec une couche linéaire de classification des mots pendant 4 époques sur le jeu *train*, avec les hyperparamètres suivants (choisis en fonction de la performance sur le jeu *dev*) : taux d'apprentissage 5×10^{-5}, longueur maximale des séquences 512 et une taille effective des batchs de 64. Les F-mesures (macro) sur le jeu de test pour les deux tâches de reconnaissance des entités sont indiquées dans le tableau 4, avec pour référence une comparaison avec un des modèles de langue anglaise biomédicale les plus largement utilisés, BioBERT de Lee *et al.* (2020). Dans le tableau, "MLM E3C" fait référence au pré-entraînement sur le corpus E3C, "ClfTriples" à la tâche de classification des triplets et "EntPred" à l'affinage par prédiction des entités.

Nous voyons qu'en général, l'entraînement sur le corpus E3C apporte le plus de bénéfices pour les modèles pré-entraînés. L'ajout des tâches basées sur l'UMLS apporte des améliorations variables en fonction des deux différentes tâches d'évaluation. Bien que les résultats obtenus dans ces expériences ne soient pas au niveau de l'état de l'art sur les tâches définies sur les corpus QUAERO et CAS (Hiot *et al.*, 2021; van Mulligen *et al.*, 2016), il est important de noter l'amélioration systématique de la F-mesure apportée par l'ajout des tâches d'entraînement sur les relations sémantiques UMLS,

TABLE 4 – Résultats sur les deux jeux de données test.

Modèle de base	Adaptation	QUAERO-SG	CAS-CATEG
BioBERT	-	51,8	46,1
`flaubert_base_cased`	-	55,3	46,0
	+ MLM E3C	59,4 (+4,1)	48,1 (+2,1)
	+ ClfTriples	59,7 (+0,3)	51,5 (+3,4)
	+ EntPred	64,5 (+4,8)	54,2 (+2,7)
`camembert-base`	-	57,4	46,2
	+ MLM E3C	61,8 (+4,4)	49,6 (+3,4)
	+ ClfTriples	63,1 (+1,3)	52,8 (+3,2)
	+ EntPred	**68,9 (+5,8)**	**56,6 (+3,8)**
DistilBERT *from-scratch*	+ MLM E3C	53,5	39,7
	+ ClfTriples	61,7 (+8,2)	48,0 (+8,3)
	+ EntPred	64,3 (+2,6)	53,1 (+5,1)

notamment l'objectif de classification des triplets dans le cas du modèle entraîné à partir de zéro.

4 Conclusion et Perspectives

Cet article présente une étude sur l'effet de l'augmentation des modèles de langage en utilisant une phase de pré-entraînement supplémentaire sur un graphe de connaissances médicales. Cette augmentation permet notamment d'améliorer les performances des modèles considérés sur deux tâches de reconnaissance d'entités nommées. Le code et les modèles utilisés seront mis à disposition pour permettre l'éventuelle reproduction des expériences et l'amélioration des approches.

Limitations et travaux futurs La mise en œuvre des approches proposées dans ce travail présente de nombreuses limitations qui peuvent entraver les performances. Celles-ci seront abordées dans des travaux ultérieurs. Premièrement, pour faciliter la comparaison directe des différentes phases d'entraînement, nous avons enchaîné de manière séquentielle les pré-entraînements MLM et KB, ce qui risque d'introduire un biais en faveur des objectifs les plus récents, étant donné la tendance de tels modèles de langue d'oublier des informations précédemment apprises (Korbak *et al.*, 2021). Dans les prochaines étapes de ce travail, nous comptons mélanger les séquences d'entraînement en provenance du métathésaurus avec celles provenant des corpus du texte libre, et faire apprendre les modèles avec une fonction objectif mixte (Hao *et al.*, 2020; Yao *et al.*, 2019). La continuation de ce travail comprendra également l'élargissement de l'éventail des objectifs supplémentaires ; nous étudierons l'impact de la prédiction des liens entre les concepts du graphe, ainsi que la classification des concepts par rapport à leurs catégories sémantiques dans l'ontologie UMLS. Enfin, il est important de noter la limitation des tâches d'évaluation en termes de taille des corpus et de granularité et d'exhaustivité des catégories d'entités considérées. Les travaux futurs impliqueront aussi des évaluations sur la reconnaissance des entités d'une granularité plus riche ainsi que d'autres types de tâches d'évaluation du domaine du traitement du langage clinique.

Références

ALROWILI S. & SHANKER V. (2021). BioM-Transformers : Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. p. 221–227, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.24.

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

AYALA SOLARES J. R., DILETTA RAIMONDI F. E., ZHU Y., RAHIMIAN F., CANOY D., TRAN J., PINHO GOMES A. C., PAYBERAH A. H., ZOTTOLI M., NAZARZADEH M., CONRAD N., RAHIMI K. & SALIMI-KHORSHIDI G. (2020). Deep learning for electronic health records : A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, **101**, 103337. DOI : 10.1016/j.jbi.2019.103337.

BERHE A., DRAZNIEKS G., MARTENOT V., MASDEU V., DAVY L. & ZUCKER J.-D. (2022). ALIBERT : A Pretrained language model for French biomedical text : a preprint. working paper or preprint.

BODENREIDER O. (2004). The unified medical language system (UMLS) : integrating biomedical terminology. PubMed PMID : 14681409; PubMed Central PMCID : PMC308795, DOI : doi : 10.1093/nar/gkh061.

CHANG D., BALAŽEVIĆ I., ALLEN C., CHAWLA D., BRANDT C. & TAYLOR A. (2020). Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, p. 167–176, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.bionlp-1.18.

CHENAIS G., TOUCHAIS H., AVALOS M., BOURDOIS L., REVEL P., GIL-JARDINÉ C. & LAGARDE E. (2021). Performance en classification de données textuelles des passages aux urgences des modèles BERT pour le français. In *PFIA 2021 - Journée Santé et I.A.*, Bordeaux / Virtual, France.

COPARA J., KNAFOU J., NADERI N., MORO C., RUCH P. & TEODORO D. (2020). Contextualized French Language Models for Biomedical Named Entity Recognition. 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. 36–48, Nancy, France : ATALA.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DUBOIS S., ROMANO N., JUNG K., SHAH N. & KALE A. D. C. (2017). The Effectiveness of Transfer Learning in Electronic Health Records Data.

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. DOI : 10.48550/ARXIV.2207.12940.

EL BOUKKOURI H., FERRET O., LAVERGNE T., NOJI H., ZWEIGENBAUM P. & TSUJII J. (2020). CharacterBERT : Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6903–6915, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.609.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French Corpus with Clinical Cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

GROUIN C., GRABAR N. & ILLOUZ G. (2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne deft 2021. *Actes de DEFT. Lille.*

HAMID O. H. (2022). From model-centric to data-centric AI : A paradigm shift or rather a complementary approach ? In *2022 8th International Conference on Information Technology Trends (ITT)*, p. 196–199. DOI : 10.1109/ITT56123.2022.9863935.

HAO B., ZHU H. & PASCHALIDIS I. (2020). Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base. p. 657–661, Barcelona, Spain (Online) : International Committee on Computational Linguistics.

HIOT N., MINARD A.-L. & BADIN F. (2021). Doing@deft : utilisation de lexiques pour une classification efficace de cas cliniques. In *Actes de l'atelier Défi Fouille de Textes@TALN 2020 Classification de cas cliniques et correction automatique de copies d'étudiants. Atelier DÉfi Fouille de Textes*, p. 41–53, Lille, France : Association pour le Traitement Automatique des Langues.

KORBAK T., ELSAHAR H., KRUSZEWSKI G. & DYMETMAN M. (2021). Controlling conditional language models with distributional policy gradients. *CtrlGen @ Neural Information Processing Systems*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

LE CLERCQ DE LANNOY T., BESANÇON R., FERRET O., TOURILLE J., BRIN-HENRY F. & VIERU B. (2022). Stratégies d'adaptation pour la reconnaissance d'entités médicales en français. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Traitement Automatique des Langues Naturelles(TALN 2022)*, p. 215–225, Avignon, France : ATALA. HAL : hal-03701500.

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

LI Y., RAO S., SOLARES J. R. A., HASSAINE A., RAMAKRISHNAN R., CANOY D., ZHU Y., RAHIMI K. & SALIMI-KHORSHIDI G. (2020). BEHRT : Transformer for Electronic Health Records. *Scientific Reports*, **10**, 7155. Number : 1 Publisher : Nature Publishing Group, DOI : 10.1038/s41598-020-62922-y.

LOSHCHILOV I. & HUTTER F. (2017). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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*, p. 7203–7219, Online : Association for Computational Linguistics.

MENG Z., LIU F., CLARK T., SHAREGHI E. & COLLIER N. (2021). Mixture-of-Partitions : Infusing Large Biomedical Knowledge Graphs into BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4672–4681, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.383.

MINARD A.-L., ZANOLI R., ALTUNA B., SPERANZA M., MAGNINI B. & LAVELLI A. (2021). European clinical case corpus. Bruno Kessler Foundation. DOI : 10.57771/DEY2-G751.

NASEEM U., BANDI A., RAZA S., RASHID J. & CHAKRAVARTHI B. R. (2022). Incorporating Medical Knowledge to Transformer-based Language Models for Medical Dialogue Generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, p. 110–115, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bionlp-1.10.

NICHOLSON D. N. & GREENE C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and structural biotechnology journal*, **18**, 1414–1428. Place : Netherlands, DOI : 10.1016/j.csbj.2020.05.017.

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 *Proc of BioTextMining Work*, p. 24–30.

PETERS M. E., NEUMANN M., LOGAN R., SCHWARTZ R., JOSHI V., SINGH S. & SMITH N. A. (2019). Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 43–54, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1005.

ROY A. & PAN S. (2021). Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 5357–5366, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.435.

SANH V., DEBUT L., CHAUMOND J. & WOLF T. (2019). Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. *ArXiv*, **abs/1910.01108**.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a web-based tool for nlp-assisted text annotation.

SUSHIL M., SUSTER S. & DAELEMANS W. (2021). Are we there yet ? Exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 41–53, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.5.

VAN MULLIGEN E. M., AFZAL Z., AKHONDI S. A., VO D. & KORS J. A. (2016). Erasmus mc at clef ehealth 2016 : Concept recognition and coding in french texts. In *Conference and Labs of the Evaluation Forum*.

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

WANG Y., TAFTI A., SOHN S. & ZHANG R. (2019). Applications of Natural Language Processing in Clinical Research and Practice. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Tutorials*, p. 22–25, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-5006.

YAO L., MAO C. & LUO Y. (2019). KG-BERT : BERT for knowledge graph completion. *ArXiv*, **abs/1909.03193**.