CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé

Rian Touchent Laurent Romary Éric Villemonte de la Clergerie
INRIA, 2 rue Simone IFF, 75012, France
{prénom.nom}@inria.fr

RÉSUMÉ

Les données cliniques dans les hôpitaux sont de plus en plus accessibles pour la recherche à travers les entrepôts de données de santé, cependant ces documents sont non-structurés. Il est donc nécessaire d'extraire les informations des comptes-rendus médicaux. L'utilisation du transfert d'apprentissage grâce à des modèles de type BERT comme CamemBERT ont permis des avancées majeures, notamment pour la reconnaissance d'entités nommées. Cependant, ces modèles sont entraînés pour le langage courant et sont moins performants sur des données biomédicales. C'est pourquoi nous proposons un nouveau jeu de données biomédical public français sur lequel nous avons poursuivi le pré-entraînement de CamemBERT. Ainsi, nous présentons une première version de CamemBERT-bio, un modèle public spécialisé pour le domaine biomédical français qui montre un gain de 2,54 points de F-mesure en moyenne sur différents jeux d'évaluations de reconnaissance d'entités nommées biomédicales.

ABSTRACT

a Tasty French Language Model Better for your Health

Clinical data in hospitals are increasingly accessible for research through clinical data warehouses, however these documents are unstructured. It is therefore necessary to extract information from medical reports to conduct clinical studies. Transfer learning with BERT-like models such as CamemBERT has allowed major advances, especially for named entity recognition. However, these models are trained for plain language and are less efficient on biomedical data. This is why we propose a new French public biomedical dataset on which we have continued the pre-training of CamemBERT. Thus, we introduce a first version of CamemBERT-bio, a specialized public model for the French biomedical domain that shows 2.54 points of F1 score improvement on average on different biomedical named entity recognition tasks.

MOTS-CLÉS : comptes-rendus médicaux, TAL clinique, CamemBERT, extraction d'information, biomédical, reconnaissance d'entités nommées.

KEYWORDS: EHR, clinical NLP, CamemBERT, information extraction, biomedical, named entity recognition.

1 Introduction

On observe ces dernières années un développement des entrepôts de données de santé (EDS) dans les hôpitaux. Ce sont des bases de données cliniques ayant pour but d'être plus accessibles pour la recherche. Ces documents représentent une opportunité pour des études cliniques massives sur des

données réelles. Elles peuvent prendre plusieurs formes, comme des comptes-rendus, des imageries médicales ou encore des prescriptions. Cependant c'est dans les comptes-rendus que la plupart des informations se trouvent. On estime que jusqu'à 80% des entités sont absentes des autres modalités (Raghavan *et al.*, 2014). Ces données, bien que très riches, sont non-structurées, ce qui implique un pré-traitement avant de pouvoir être utilisées dans une étude clinique.

Modèles de langue pour l'extraction d'information Les modèles de type BERT (Devlin *et al.*, 2019) montrent de manière consistante des résultats à l'état de l'art pour tout un ensemble de tâches de TAL. L'adaptation de BERT au français, avec notamment le modèle CamemBERT (Martin *et al.*, 2020) a permis de répliquer ces performances pour le TAL du français. CamemBERT est basé sur RoBERTa (Liu *et al.*, 2019), une version plus efficiente de BERT. Il est entraîné sur un corpus français extrait du web nommé OSCAR (Ortiz Suárez *et al.*, 2019).

Pour extraire les informations des comptes-rendus, il est nécessaire d'avoir des modèles de langue performants sur des données cliniques françaises, notamment pour de la reconnaissance d'entités nommées. Il est possible de simplement utiliser CamemBERT, cependant les résultats de ce modèle sur des données biomédicales sont décevants (Cardon *et al.*, 2020) car sur certains jeux d'évaluation il présente des performances inférieures à des modèles heuristiques. Ces résultats sont prévisibles car il s'agit d'un modèle entraîné pour du langage courant, souvent issu de pages web de type forum, or les données biomédicales et particulièrement les données cliniques sont bien différentes. Elles présentent des termes techniques, très rares voir absents du langage courant, et un style radicalement distinct, souvent télégraphique, présentant rarement des phrases complètes avec des abréviations qui peuvent varier.

Confidentialité des données cliniques Une des problématiques majeures avec les entrepôts de données de santé est la confidentialité des données. En effet ce sont des données réglementées, soumises à des régulations par la CNIL. C'est pourquoi les adaptations de CamemBERT au domaine biomédical réalisées au sein des infrastructures des hôpitaux (Dura *et al.*, 2022) ne peuvent être publiées. Leurs jeux de données d'entraînement sont soumis à des contraintes de publication. Ces contraintes s'appliquent également aux modèles résultants. Il n'est donc pas possible d'échanger ces modèles entre différents établissements de santé. Un modèle public n'aurait pas ses contraintes, et pourrait donc être utilisé dans différents établissements.

Au travers de cet article, nous présentons deux contributions principales [1] :

— La création d'un nouveau jeu de données français public spécialisé dans le domaine biomédical
— L'introduction d'une adaptation de CamemBERT publique pour le domaine biomédical, présentant un gain de performance sur des tâches de reconnaissance d'entités nommées.

2 État de l'art

Les travaux sur l'adaptation de modèles de langue à de nouveaux domaines sont nombreux. Gururangan *et al.* (2020) montrent qu'une seconde phase de pré-entraînement sur un domaine cible

1. Nos contributions sont disponibles sur le hub Huggingface : almanach/camembert-bio-base, rntc/biomed-fr

permet d'améliorer les performances sur différentes tâches par la suite, même lorsque le corpus du domaine cible est de taille restreinte. On observe jusqu'à 3 points de gain de F-mesure dans le biomédical par rapport au même modèle sans la seconde phase.

Cet article a inspiré la création de nouveaux modèles basés sur BERT, en utilisant une seconde phase de pré-entraînement sur différents domaines spécialisés. Lee *et al.* (2019) introduit BioBERT, un modèle de type BERT spécialisé pour le biomédical anglais. BioBERT montre un gain de performance sur de nombreuses tâches de TAL biomédicales, dont $0,62\%$ d'amélioration de F-mesure sur de la reconnaissance d'entités nommées, $2,80\%$ de F-mesure sur de l'extraction de relations, et $12,24\%$ de MMR sur des questions-réponses. La seconde phase se déroule sur un corpus extrait de PubMed et PMC, composé d'articles scientifiques biomédicaux d'environ 18 milliards de mots. C'est un corpus conséquent mais toutefois composé uniquement du style scientifique. On observe cependant des gains de performance dans tous les styles. La présence du vocabulaire médical dans le corpus permet probablement une grande amélioration par rapport au langage courant.

Il est également possible d'entraîner de nouveaux modèles *from scratch*. C'est l'approche explorée par SciBERT (Beltagy *et al.*, 2019) et PubMedBERT (Gu *et al.*, 2022), deux modèles spécialisés sur des articles scientifiques biomédicaux. PubMedBERT montre que cette méthode permet de meilleures performances que les modèles entraînés par une seconde phase de spécialisation. Cependant, les gains sont assez faibles, et cette approche est particulièrement plus coûteuse. Partir de zéro nécessite un entraînement plus long et un plus grand corpus pour obtenir ces performances.

Pour le français, les modèles de référence sont CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020), cependant il n'existe pas de version biomédicale publique et disponible à ce jour de ces modèles. Cela dit, de nombreux travaux ont tenté d'adapter CamemBERT à ce domaine : Copara *et al.* (2020) ont exploré une seconde phase de pré-entraînement sur 31 000 articles scientifiques français biomédicaux. Ils n'observent néanmoins pas, sur la version large de CamemBERT, une amélioration significative sur une tâche de reconnaissance d'entités nommées cliniques. La combinaison d'un corpus assez restreint (31k documents contre 18 milliards de mots pour BioBERT) et de la version large du modèle CamemBERT, en sont probablement la cause. Le Clercq de Lannoy *et al.* (2022) ont également adapté CamemBERT au domaine biomédical. Ils ont pour cela agrégé des documents de différentes sources, tels que PubMed, Cochrane, ISTEX ou encore Wikipédia. Cela forme ainsi un plus grand corpus, partiellement public, d'environ 136 millions de mots. Ils observent une amélioration de 2 points de F-mesure sur un jeu d'évaluation de reconnaissance d'entités nommées composé de notices de médicaments (EMEA), mais pas d'amélioration significative sur un jeu composé de titres d'articles scientifiques (MEDLINE). Enfin, Dura *et al.* (2022) ont continué le pré-entraînement de CamemBERT sur 21 millions de documents cliniques de l'entrepôt de données de santé de l'APHP. On observe une amélioration significative de 3% sur APMed, un jeu de reconnaissance d'entités nommées cliniques privé appartenant à l'APHP. On note également des scores similaires à CamemBERT sur EMEA et MEDLINE. Ainsi leur nouveau modèle est meilleur sur des données cliniques et obtient des scores similaires à ceux de CamemBERT sur le reste du biomédical. Aucun de ces modèles adaptés pour le biomédical n'a été rendu public [2].

2. Un article publié suite à la soumission de notre travail annonce la publication d'un nouveau modèle biomédical français public nommé DrBERT (Labrak *et al.*, 2023)

3 Méthodes

3.1 Corpus : biomed-fr

Tout d'abord, nous avons constitué un corpus français biomédical, composé uniquement de documents publics pour minimiser les contraintes d'usage précédemment évoquées. Les documents proviennent de trois sources différentes (cf. table 1), dont la principale est ISTEX. Ce nouveau corpus que nous nommons *biomed-fr* est composé de 413 millions de mots, soit 2,7 GB de données. Martin *et al.* (2020) ont montré qu'avec seulement 4 GB de données, il était possible de quasiment égaler les performances du modèle entraîné avec les 138 GB d'OSCAR (Ortiz Suárez *et al.*, 2019). Pour une adaptation de CamemBERT au biomédical, on peut estimer que c'est une quantité de données suffisante.

Corpus	Détails	Taille
ISTEX	Divers documents de la littérature scientifique indexés sur ISTEX	276 M
CLEAR	Notices de médicaments	73 M
E3C	Divers documents issus de journaux, de notices et de cas cliniques	64 M
Total		413 M

TABLE 1 – Composition du corpus biomed-fr (en millions de mots)

ISTEX La base de données ISTEX référence 27 millions de publications scientifiques. Nous avons extrait 108 183 documents français publiés dans une revue de biologie ou de médecine depuis 1990. Les articles dont l'année de parution est plus ancienne que cette date présentent de nombreuses erreurs typographiques. Ce sont souvent des articles scannés, pour lesquels il faut appliquer des algorithmes de reconnaissance de caractères, ce qui amène à un certain nombre d'erreurs. Ce genre d'erreur se retrouve plus marginalement dans les articles publiés après 1990. Certains documents, bien qu'en français, contiennent des passages en anglais. Il y a donc une quantité indéterminée d'anglais dans ce corpus. Il est cependant peu probable que cela impacte significativement le pré-entraînement. Les erreurs typographiques et la présence d'autres langues sont des points qui pourront être corrigés par de futures versions de biomed-fr.

CLEAR Le corpus CLEAR (Grabar & Cardon, 2018) est composé d'articles d'encyclopédies, de notices de médicaments, et de résumés d'articles scientifiques. Chaque document est présent en deux versions, l'une en langage technique et l'autre en langage simplifié. Nous avons récupéré l'ensemble de ces documents dans les deux versions. Concernant les notices de médicaments, nous avons retiré les phrases redondantes en début et fin de document. Il s'agit notamment de la barre de navigation du site web dont ont été extraits les documents, ou encore d'informations sur l'entreprise qui met en vente les documents.

E3C Ce corpus (Magnini *et al.*, 2021) est composé de 3 couches. Les deux premières sont annotées ou semi-annotées, et seront retenues pour l'évaluation. La dernière couche n'est pas annotée et c'est celle que nous avons récupérée. Elle est composée de concours d'admission en spécialité de médecine,

de notices de médicaments et de résumés de thèses de médecine. Il est possible que des notices soient des doublons de ceux présents dans CLEAR.

biomed-fr-small Nous avons créé un second corpus plus petit, nommé *biomed-fr-small*. Il est constitué de 10% du contenu de biomed-fr, avec une sélection aléatoire des documents. Il va nous permettre de mesurer l'impact de la taille du corpus.

3.2 Pré-entraînements

Pour l'adaptation de CamemBERT au domaine biomédical, nous avons réalisé une seconde phase de pré-entraînement sur les deux versions du corpus biomed-fr en partant des poids et de la configuration du modèle camembert-base. Ainsi, nous avons appliqué la tâche de *Masked Language Modeling* (MLM) avec un masquage de mots entiers, pour suivre la méthode de Martin *et al.* (2020). Nous avons utilisé l'optimisateur Adam (Kingma & Ba, 2017) avec $\beta_1 = 0.9$ et $\beta_2 = 0,98$ et un taux d'apprentissage de $5\mathrm{e}{-5}$. Nous avons effectué 50 000 pas (*steps*) pendant 39 heures avec deux Tesla V100. Nous avons utilisé une taille de lots (*batch size*) de 8 par GPU et de l'accumulation de gradient sur 16 pas pour obtenir une taille de lots effective de 256.

3.3 Affinages et évaluations

Concernant l'évaluation des modèles, nous avons récolté trois jeux de données d'évaluation de reconnaissance d'entités nommées. Les trois présentent des styles variés, ce qui permet d'évaluer la polyvalence du modèle sur les différents sous-domaines du biomédical.

QUAERO Le corpus QUAERO (Névéol *et al.*, 2014) est composé de deux jeux d'évaluation : EMEA, contenant des notices de médicaments et MEDLINE, contenant des titres d'articles scientifiques. Les entités sont annotées manuellement en suivant 10 groupes sémantiques de l'UMLS (Lindberg *et al.*, 1993). Certaines de ces entités étant imbriquées, nous avons simplement gardé les entités de plus large granularité. Les F-mesures sont calculées de la même manière.

E3C Pour l'évaluation, contrairement au corpus biomed-fr, nous utilisons les couches 1 et 2. Ces dernières présentent des documents de natures différentes. Il s'agit de cas cliniques extraits d'articles scientifiques. La couche 2 est semi-annotée. C'est celle-ci que nous utilisons comme jeu d'entraînement pour l'affinage, avec 10% dédié au jeu de validation. Nous évaluons sur la couche 1, qui est entièrement annotée à la main. Il n'y a qu'une seule classe, l'objectif est de trouver les entités cliniques dans le texte, quel que soit le type.

CAS Le corpus CAS (Grouin *et al.*, 2019) est également composé de cas cliniques issus d'articles scientifiques. Nous nous focalisons sur la tâche 3 de DEFT 2020 (Cardon *et al.*, 2020). C'est une tâche d'extraction d'information basée sur CAS. Elle comprend deux sous-tâches, et donc deux jeux d'annotations. Dans la première il faut identifier deux classes : pathologie et signe ou symptômes. La seconde concerne les informations associées : anatomie, dose, examen, mode, moment, substance, traitement et valeur. Ces deux tâches seront respectivement désignées par la suite par CAS1 et CAS2.

Affinage Concernant l'affinage, nous avons utilisé Optuna (Akiba *et al.*, 2019) pour la sélection des hyperparamètres. Ainsi nous avons un taux d'apprentissage de $5e{-}5$, un ratio d'échauffement (*warmup ratio*) de 0,224 et une taille de lots (*batch size*) de 16. Nous effectuons 2000 pas (*steps*). Les prédictions sont faites avec une simple couche linéaire en tête du modèle. Aucune des couches de CamemBERT n'est figée.

Evaluation Les scores sont mesurés avec l'outil seqeval (Nakayama, 2018) en mode strict avec micro-moyenne et le schéma "IOB2". Pour chaque évaluation, le meilleur modèle de l'affinage sur le jeu de validation est choisi pour mesurer le score final sur le jeu de test. Nous faisons la moyenne sur 10 évaluations avec différentes amorces (*seed*).

4 Résultats et discussions

				CamemBERT-bio	
Style	Dataset	Score	CamemBERT	biomed-fr-small	biomed-fr
Clinique	CAS1	F1	70,50 ± 1,75	72,94 ± 1,12	**73,03 ± 1,29**
		P	70,12 ± 1,93	**72,97 ± 0,84**	71,71 ± 1,61
		R	70,89 ± 1,78	72,92 ± 1,39	**74,42 ± 1,49**
	CAS2	F1	79,02 ± 0,92	80,00 ± 0,32	**81,66 ± 0,59**
		P	77,3 ± 1,36	78,29 ± 0,91	**80,96 ± 0,91**
		R	80,83 ± 0,96	81,80 ± 0,48	**82,37 ± 0,69**
	E3C	F1	67,63 ± 1,45	67,96 ± 1,85	**69,85 ± 1,58**
		P	78,19 ± 0,72	77,41 ± 1,01	**79,11 ± 0,42**
		R	59,61 ± 2,25	60,57 ± 2,32	**62,56 ± 2,50**
Notices	EMEA	F1	74,14 ± 1,95	75,93 ± 2,42	**76,71 ± 1,50**
		P	74,62 ± 1,97	76,23 ± 2,27	**76,92 ± 1,96**
		R	73,68 ± 2,22	75,63 ± 2,61	**76,52 ± 1,62**
Scientifique	MEDLINE	F1	65,73 ± 0,40	65,48 ± 0,31	**68,47 ± 0,54**
		P	64,94 ± 0,82	64,43 ± 0,50	**67,77 ± 0,88**
		R	66,56 ± 0,56	66,56 ± 0,16	**69,21 ± 1,32**

TABLE 2 – Moyennes sur 10 évaluations des F-mesures sur différents jeux biomédicaux de reconnaissance d'entités nommées

CamemBERT vs CamemBERT-bio On observe (cf. table 2) un gain significatif de performance sur tous les jeux d'évaluation avec notre nouveau modèle. Nous avons en moyenne 2,54 points d'amélioration de F-mesure. Ce gain s'observe dans tous les styles, ce qui montre la polyvalence du modèle pour les domaines cliniques et scientifiques du biomédical.

biomed-fr-small vs biomed-fr On note une baisse de performance avec le jeu biomed-fr-small, mais toujours un gain significatif sur certains jeux de données par rapport à CamemBERT. Cela

confirme que la taille du corpus influe positivement sur les performances même dans un domaine spécialisé comme le biomédical.

		CAS1	CAS2	EMEA			MEDLINE		
Évaluateur	Auteurs	F1	F1	F1	P	R	F1	P	R
seqeval	Dura *et al.* (2022)-fine-tuned	-	-	72,90	-	-	59,70	-	-
	Dura *et al.* (2022)-from-scratch	-	-	69,30	-	-	60,10	-	-
	Notre approche	**73,03**	**81,66**	**76,71**	**76,92**	**76,52**	**68,47**	**67,77**	**69,21**
BRATeval	Le Clercq de Lannoy *et al.* (2022)			67,4	73,4	62,2	55,3	62,2	49,7
	Mulligen *et al.* (2016)	-	-	74,9	71,6	**78,5**	**69,8**	68	**71,6**
	Copara *et al.* (2020)	61,53	73,7	-	-	-	-	-	-
	Notre approche	**84,97**	**83,25**	**77,80**	**79,77**	75,93	56,16	**75,33**	44,82

TABLE 3 – Comparaison de CamemBERT-bio avec différentes approches sur les 4 tâches de reconnaissance d'entités nommées. Dans la première partie de la table les scores sont mesurés avec seqeval (Nakayama, 2018), et la seconde avec BRATeval, qui est l'outil d'évaluation fourni pour la campagne CLEF eHealth Evaluation lab 2016 (Névéol *et al.*, 2016).

	Corpus d'adaptation	
Auteurs	Origine	Taille
Dura *et al.* (2022)	APHP	21 MD
Le Clercq de Lannoy *et al.* (2022)	divers	136 MW
Copara *et al.* (2020)	PubMed	31 KD
Notre approche	biomed-fr	413 MW

TABLE 4 – Corpus de pré-entraînement des différents approches (cf. table 3)

Comparaison avec l'état de l'art Nous avons comparé les performances de CamemBERT-bio avec les différentes approches précédemment évoquées (cf. table 3). CamemBERT-bio obtient pour presque tous les jeux d'évaluation les meilleurs résultats. Dura *et al.* (2022) n'ont pas observé d'amélioration sur EMEA et MEDLINE par rapport à CamemBERT car leur corpus de pré-entraînement (cf. table 4) est composé de documents provenant de l'APHP, ce qui en fait un corpus moins varié. Ils gagnent cependant plusieurs points sur leur jeu d'évaluation basé lui aussi sur les documents de l'APHP. Mulligen *et al.* (2016) présente le meilleur score sur MEDLINE, ainsi que le meilleur rappel sur EMEA. Leur approche est basée sur un modèle à base de connaissances, ce qui leur permet d'obtenir le meilleur rappel sur les deux jeux d'évaluations de QUAERO. Enfin, leur approche est la seule capable de gérer les entités imbriquées, ce qui leur donne un avantage.

Il est important de noter que ces différentes approches basées sur CamemBERT ont des cadres expérimentaux variés. La présence de couches CRF plutôt qu'une simple couche linéaire en sortie de CamemBERT, le gel des couches de CamemBERT ou encore les hyperparamètres sont des exemples de variation qu'on observe en plus des corpus de pré-entraînement, ce qui rend la comparaison plus difficile.

Analyse de la tokenisation CamemBERT-bio est un modèle adapté pour le biomédical de CamemBERT. Contrairement à un nouveau modèle entraîné *from scratch*, il partage le même vocabulaire. Le vocabulaire de CamemBERT a été construit en utilisant SentencePiece (Kudo & Richardson, 2018) sur un échantillon d'OSCAR. C'est donc un vocabulaire généraliste fait pour le langage courant. On peut faire l'hypothèse que le tokeniseur de CamemBERT va produire des sur-segmentations de termes techniques biomédicaux.

Nous avons alors exploré cette possibilité en entraînant un tokeniseur spécialisé sur biomed-fr-small, et en calculant l'intersection des deux vocabulaires.

Termes	généraliste	spécialisé
échocardiographie	écho-cardi-ographie	échocardiographi-e
transthoracique	trans-thorac-ique	trans-thoracique
glimépiride	g-lim-épi-ride	gli-m-épi-ride
cardiopathie	cardio-pathie	cardiopathie
diastoliques	dia-s-tol-iques	diastolique-s

TABLE 5 – Comparaison de la segmentation entre un tokeniseur généraliste et un tokeniseur spécialisé sur quelques termes techniques biomédicaux

On calcule une intersection de 45% entre les deux vocabulaires, ce qui est assez proche de l'intersection de 42% trouvé par Beltagy *et al.* (2019) entre le vocabulaire de BERT et celui de SciBERT. Il y a donc une différence significative des termes les plus fréquents.

5 Conclusion et perspectives

Nous avons introduit un nouveau corpus biomédical français nommé *biomed-fr* de 413 millions de mots composé de notices de médicaments et de documents de la littérature scientifique en médecine et en biologie. Ce nouveau corpus nous a permis d'adapter CamemBERT au domaine biomédical avec une seconde phase de pré-entraînement. On observe une amélioration des performances sur tous nos jeux d'évaluation de reconnaissance d'entités nommées. On note un gain de 2,54 points de F-mesure en moyenne.

Nous avons des pistes pour de nouvelles versions de biomed-fr. D'une part, réaliser un plus grand nettoyage des données en retirant les passages au sein des documents qui ne sont pas en français, ou en retirant les documents contenant un trop grand nombre d'erreurs typographiques. D'autre part, en augmentant la quantité de données. Cela pourrait passer par l'exploitation des documents archivés sur HAL concernant les sciences de la vie, notamment publiés par l'INSERM, ou la récupération de résumés d'articles français sur PubMed.

L'analyse de la tokenisation nous pousse à réfléchir à la création d'un nouveau vocabulaire pour CamemBERT-bio. Malgré le gain de performance assez faible de PubMedBERT par rapport à BioBERT malgré son vocabulaire spécialisé, la sur-segmentation des termes techniques et le faible taux d'intersection entre le vocabulaire généraliste et le vocabulaire spécialisé montrent l'intérêt de

l'expérience.

Enfin, ces derniers mois, de nombreux modèles génératifs, souvent de plusieurs milliards de paramètres, ont montré des performances remarquables sur des tâches biomédicales, outrepassant parfois les modèles spécialisés comme BioBERT (Agrawal *et al.*, 2022; Singhal *et al.*, 2022). C'est une piste de recherche prometteuse pour l'extraction d'information biomédicale. Cependant nous avons des raisons de penser qu'un modèle de type BERT a toujours de l'intérêt (Lehman *et al.*, 2023). D'une part, dans le contexte clinique, les modèles doivent souvent être utilisés au sein des infrastructures des établissements de santé, ce qui se traduit par des contraintes de ressources. Il est alors plus facile de déployer des petits modèles spécialisés que des grands modèles généralistes. D'autre part l'utilisation de ces modèles génératifs nécessite souvent de passer par des serveurs distants, souvent à travers des API, ce qui rend difficile leur utilisation compte tenu des contraintes de confidentialité auxquelles sont soumis les documents cliniques.

Références

AGRAWAL M., HEGSELMANN S., LANG H., KIM Y. & SONTAG D. (2022). Large Language Models are Few-Shot Clinical Information Extractors. arXiv :2205.12689 [cs], DOI : 10.48550/arXiv.2205.12689.

AKIBA T., SANO S., YANASE T., OHTA T. & KOYAMA M. (2019). Optuna : A Next-generation Hyperparameter Optimization Framework. arXiv :1907.10902 [cs, stat], DOI : 10.48550/arXiv.1907.10902.

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A Pretrained Language Model for Scientific Text. 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. 3615–3620, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

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 *DEFT 2020*, Nancy, 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., *Traitement Automatique des Langues Naturelles (TALN, 27e 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. arXiv :1810.04805 [cs], DOI : 10.48550/arXiv.1810.04805.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

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*. Rapport interne arXiv :2207.12940, arXiv. arXiv :2207.12940 [cs, stat] type : article.

GRABAR N. & CARDON R. (2018). CLEAR – Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, p. 3–9, Tilburg, the Netherlands : Association for Computational Linguistics. DOI : 10.18653/v1/W18-7002.

GROUIN C., GRABAR N., CLAVEAU V. & HAMON T. (2019). Clinical Case Reports for NLP. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 273–282, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5029.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, **3**(1), 1–23. arXiv :2007.15779 [cs], DOI : 10.1145/3458754.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.740.

KINGMA D. P. & BA J. (2017). Adam : A Method for Stochastic Optimization. arXiv :1412.6980 [cs], DOI : 10.48550/arXiv.1412.6980.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). Drbert : A robust pre-trained model in french for biomedical and clinical domains.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

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.

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 (Adaptation strategies for biomedical named entity recognition in French). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 215–225, Avignon, France : ATALA.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, p. btz682. arXiv :1901.08746 [cs], DOI : 10.1093/bioinformatics/btz682.

LEHMAN E., HERNANDEZ E., MAHAJAN D., WULFF J., SMITH M. J., ZIEGLER Z., NADLER D., SZOLOVITS P., JOHNSON A. & ALSENTZER E. (2023). Do We Still Need Clinical Language Models ? arXiv :2302.08091 [cs], DOI : 10.48550/arXiv.2302.08091.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, **32**(4), 281–291. DOI : 10.1055/s-0038-1634945.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv :1907.11692 [cs], DOI : 10.48550/arXiv.1907.11692.

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., *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021*, Collana dell'Associazione Italiana di Linguistica Computazionale, p. 258–264. Torino : Accademia University Press. Code : Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021.

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.

MULLIGEN E. M. V., AFZAL Z., AKHONDI S. A., VO D. & KORS J. A. (2016). Erasmus MC at CLEF eHealth 2016 : Concept Recognition and Coding in French Texts.

NAKAYAMA H. (2018). seqeval : A python framework for sequence labeling evaluation. Software available from https ://github.com/chakki-works/seqeval.

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

ORTIZ SUÁREZ P. J., SAGOT B. & ROMARY L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, p. 9 – 16, Mannheim : Leibniz-Institut für Deutsche Sprache. DOI : 10.14618/ids-pub-9021.

RAGHAVAN P., CHEN J. L., FOSLER-LUSSIER E. & LAI A. M. (2014). How essential are unstructured clinical narratives and information fusion to clinical trial recruitment ? *AMIA Summits on Translational Science Proceedings*, **2014**, 218–223.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SINGHAL K., AZIZI S., TU T., MAHDAVI S. S., *et al.* (2022). Large Language Models Encode Clinical Knowledge. arXiv :2212.13138 [cs], DOI : 10.48550/arXiv.2212.13138.