DrBERT: Un modèle robuste pré-entraîné en français pour les domaines biomédical et clinique

Yanis Labrak[1,4], Adrien Bazoge[2,3], Richard Dufour[2], Mickael Rouvier[1], Emmanuel Morin[2], Béatrice Daille[2] and Pierre-Antoine Gourraud[3]
(1) LIA - Avignon Univserité (2) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France (3) Nantes Université, Clinique des données, CHU de Nantes (4) Zenidoc
{prenom.nom}@univ-avignon.fr, {prenom.nom}@univ-nantes.fr

RÉSUMÉ

Ces dernières années, les modèles de langage pré-entraînés ont obtenu les meilleures performances sur un large éventail de tâches de traitement automatique du langage naturel (TALN). Alors que les premiers modèles ont été entraînés sur des données issues de domaines généraux, des modèles spécialisés sont apparus pour traiter plus efficacement des domaines spécifiques. Dans cet article, nous proposons une étude originale de modèles de langue dans le domaine médical en français. Nous comparons pour la première fois les performances de modèles entraînés sur des données publiques issues du web et sur des données privées issues d'établissements de santé. Nous évaluons également différentes stratégies d'apprentissage sur un ensemble de tâches biomédicales. Enfin, nous publions les premiers modèles spécialisés pour le domaine biomédical en français, appelés DrBERT, ainsi que le plus grand corpus de données médicales sous licence libre sur lequel ces modèles sont entraînés.

ABSTRACT

DrBERT : A Robust Pre-trained Model in French for Biomedical and Clinical domains.

In recent years, pre-trained language models (PLMs) achieve the best performance on a wide range of natural language processing tasks. While the first models were trained on general domain data, specialized ones have emerged to more effectively treat specific domains. In this paper, we propose an original study of PLMs in the medical domain on French language. We compare, for the first time, the performance of PLMs trained on both public data from the web and private data from healthcare establishments. We also evaluate different learning strategies on a set of biomedical tasks. Finally, we release the first specialized PLMs for the biomedical field in French, called DrBERT, as well as the largest corpus of medical data under free license on which these models are trained.

MOTS-CLÉS : BERT ; RoBERTa ; Transformers ; Biomédical ; Clinique ; Modèle de langue.

KEYWORDS: BERT ; RoBERTa ; Transformers ; Biomedical ; Clinical ; Language Model.

1 Introduction

Au cours des dernières années, les modèles de langage pré-entraînés ont permis d'améliorer considérablement les performances de nombreuses tâches du traitement automatique du langage naturel (TALN). Les modèles récents, tels que BERT (Devlin *et al.*, 2019) ou RoBERTa (Liu *et al.*, 2019), tirent profit des énormes quantités de données non étiquetées grâce à des approches non-supervisées fondées sur l'architecture Transformers (Vaswani *et al.*, 2017). La plupart de ces modèles sont souvent pré-entraînés sur des corpus de domaines généraux et une étape supplémentaire de réglage fin

(*fine-tuning*) peut être appliquée pour utiliser ces modèles sur une tâche cible (Devlin *et al.*, 2019).

L'adaptation des modèles de langue à un domaine de spécialité suit généralement deux stratégies. La première consiste à entraîner à partir de zéro (*from scratch*) un nouveau modèle en utilisant uniquement les données de la spécialité cible. La seconde approche, appelée pré-entraînement continué (Howard & Ruder, 2018), poursuit l'entraînement de modèles déjà pré-entraînés, permettant de passer d'un modèle générique à un modèle spécialisé. Bien que des études aient montré que la première stratégie offre généralement de meilleures performances (Lee *et al.*, 2019), la seconde nécessite un nombre de ressources moins important (Chalkidis *et al.*, 2020; El Boukkouri *et al.*, 2022) que ce soit en termes de ressources informatiques ou de quantité de données.

Récemment, de multiples modèles de langue ont été développés pour les domaines biomédical et clinique, principalement pour l'anglais. Les modèles BioBERT (Lee *et al.*, 2019), BlueBERT (Peng *et al.*, 2019) et ClinicalBERT (Huang *et al.*, 2019) ont utilisé le pré-entraînement continué à partir d'un modèle BERT générique. D'autres modèles, comme SciBERT (Beltagy *et al.*, 2019) et PubMedBERT (Gu *et al.*, 2021), ont été pré-entraînés à partir de zéro. Dans les langues autres que l'anglais, les modèles s'appuyant sur BERT sont plus rares et reposent principalement sur un pré-entraînement continué. Pour le français, il n'existe, à notre connaissance, aucun modèle disponible publiquement pour le domaine biomédical.

Dans cet article, nous décrivons et diffusons librement DrBERT, les premiers modèles spécialisés dans le domaine biomédical pour le français, ainsi que le corpus qui a permis leur entraînement. Nous proposons également une étude originale évaluant les différentes stratégies de pré-entraînement de modèles de langue pour le domaine médical, en comparant les modèles DrBERT avec un modèle pré-entrainé sur des données cliniques privées, nommé ChuBERT. Nos contributions sont :

— Un nouveau *benchmark* agrégeant un ensemble de tâches de TALN dans le domaine médical en français, permettant d'évaluer les modèles de langue au niveau syntaxique et sémantique.
— Un corpus de données textuelles, nommé NACHOS, rassemblant plusieurs sources biomédicales en ligne.
— La construction et l'évaluation des premiers modèles de langues libres en français pour le domaine biomédical fondés sur l'architecture RoBERTa, appelés DrBERT, comprenant l'analyse de différentes stratégies de pré-entraînement.
— Un ensemble de modèles utilisant des données publiques et privées entraînés sur des tailles de données comparables. Ces modèles ont ensuite été comparés en évaluant leurs performances sur un large éventail de tâches, tant publiques que privées.
— La distribution des modèles de langue publics sous la licence open-source MIT ainsi que le corpus NACHOS sous la licence CC0 1.0.

2 Corpus de pré-entraînement

Les travaux précédents dans le domaine biomédical (Gu *et al.*, 2021) sur les modèles de langue ont souligné l'importance de faire correspondre les sources de données utilisées lors de l'entraînement avec les tâches cibles. En raison de leur nature sensible, les données cliniques sont extrêmement difficiles à obtenir. La collecte massive de données web relatives à ce domaine apparaît comme une solution permettant de pallier ce manque.

Le tableau 1 donne un aperçu général des deux corpus collectés. Les données publiques issues du web ont permis la constitution d'un corpus, appelé NACHOS$_{large}$, à partir de sources disponibles ouvertement en ligne contenant 7,4 Go de données. Le jeu de données privées, appelé XBDW$_{small}$,

Corpus	Taille	#mots	#phrases
NACHOS$_{large}$ (public)	7,4 Go	1,1 G	54,2 M
NACHOS$_{small}$ (public)	4 Go	646 M	25.3 M
XBDW$_{small}$ [1] (privé)	4 Go	655 M	43.1 M
XBDW$_{mixed}$ (public+privé)	4+4 Go	1,3 G	68,4 M

TABLE 1 – Aperçu des corpus de données publiques (NACHOS) et privées (XBDW) collectées.

contient 4 Go de données provenant du Centre Hospitalier Universitaire (CHU) de XXX [2]. Afin d'effectuer des expériences comparables, nous avons extrait un sous-corpus NACHOS (NACHOS$_{small}$) de la même taille que les données privées.

Corpus public - NACHOS opeN crAwled frenCh Healthcare cOrpuS (NACHOS) est un ensemble de données textuelles médicales françaises distribué de façon libre et obtenu à partir de la collecte massive d'une variété de sources textuelles autour du sujet médical sur le web. Il se compose de plus d'un milliard de mots, tirés de 24 sites web francophones de haute qualité. Le corpus comprend un large éventail d'informations médicales : descriptions de maladies, fiches d'information sur les traitements et les médicaments, conseils généraux sur la santé, rapports de réunions scientifiques officielles, cas cliniques anonymisés, littérature scientifique, thèses, cours de médecine universitaire, ainsi que de nombreuses données obtenues à partir de sources textuelles brutes, de *scrapping web* et de Reconnaissance Optique de Caractères (ROC) comme présenté dans le Tableau 6 en annexe. Nous avons utilisé des heuristiques pour découper les textes en phrases et filtrer les phrases courtes ou de mauvaise qualité comme celles obtenues par ROC. Ensuite, nous les classons par langues en utilisant notre propre classificateur entraîné sur les corpus multilingues Opus EMEA (Tiedemann & Nygaard, 2004) et MASSIVE (FitzGerald *et al.*, 2022) pour ne garder que les phrases en français. Pour la version 4 Go de NACHOS (NACHOS$_{small}$), nous avons mélangé l'ensemble du corpus et sélectionné au hasard 25,3 millions de phrases afin de maximiser l'homogénéité des sources de données.

Corpus privé - XBDW Le corpus privé, appelé XXX Biomedical Data Warehouse (XBDW), a été obtenu en utilisant l'entrepôt de données du CHU de XXX. Cet entrepôt de données comprend différentes dimensions de données relatives aux patients : socio-démographiques, prescriptions médicamenteuses et autres informations associées au séjour hospitalier (diagnostic, biologie, imagerie, etc.). L'autorisation de mise en œuvre et d'exploitation de l'entrepôt de données XBDW a été accordée en XXXX [3] par la CNIL (Commission Nationale de l'Informatique et des Libertés) ; autorisation N° XXX [4]. Pour ce travail, un échantillon de 1,7 millions de comptes-rendus hospitaliers désidentifiés a été sélectionné aléatoirement et extrait de l'entrepôt de données. Les comptes-rendus proviennent de différents services hospitaliers, tels que les urgences, la gynécologie et la cardiologie. Ce corpus contient 655 millions de mots, issus de 43,1 millions de phrases, pour une taille totale d'environ 4 Go.

3 Pré-entraînement des modèles

3.1 Impact des données

L'un des problèmes consiste à identifier la quantité de données nécessaires pour créer un modèle performant et capable de rivaliser avec les modèles pré-entraînés sur des domaines généraux. Des études récentes, comme celles de Martin *et al.* (2020) et Zhang *et al.* (2021), discutent de l'impact de

2. Nom de la ville masqué à des fins d'anonymisation lors de la relecture par les pairs.
3. Année cachée pour le double aveugle.
4. Numéro d'autorisation caché pour le double aveugle.

la taille des données de pré-entraînement sur la performance des modèles et montrent que certaines tâches bénéficient d'une quantité moindre de données. Dans le domaine médical, aucune étude n'a été menée pour comparer l'impact de la variation de la quantité de données sur le pré-entraînement, ou pour évaluer l'impact de la qualité des données en fonction de leur source de collecte.

Nous proposons donc d'évaluer l'impact des différentes sources de données en comparant NACHOS$_{small}$ et XBDW$_{small}$ entre elles comme décrit dans la section 2. De plus, nous proposons de comparer les résultats obtenus avec ceux d'un modèle pré-entraîné sur une quantité de données plus grande (NACHOS$_{large}$) afin d'étudier si le fait de disposer de presque deux fois plus de données permet d'améliorer les performances. Pour finir, nous avons évalué une combinaison des corpus public (NACHOS$_{small}$) et privé (XBDW$_{small}$) pour un total de 8 Go (XBDW$_{mixed}$), afin de démontrer si la combinaison de données de qualité variable permettent des représentations complémentaires.

3.2 Stratégies de pré-entraînement

En plus de l'analyse de la taille et des sources de données, nous cherchons également à évaluer trois stratégies de pré-entraînement des modèles de langue pour le domaine médical :

— Pré-entraînement du modèle à partir de zéro, incluant une tokenization spécifique des mots à partir de nos données d'apprentissage.
— Poursuivre le pré-entraînement d'un modèle de langue général pour le français, ici CamemBERT, sur nos données du domaine médical, tout en conservant le tokenizer initial (*i.e.* celui de CamemBERT).
— Poursuivre le pré-entraînement d'un modèle de langage spécifique au domaine médical anglais, appelé PubMedBERT, sur nos données en français, en conservant le tokenizer initial.

Concernant la dernière stratégie, l'objectif est de comparer les performances d'un modèle médical anglais pré-entraîné sur nos données médicales françaises, à un autre modèle basé sur un modèle générique français. En effet, ces deux langues partagent une terminologie de spécialité commune.

Modèle	Architecture	Stratégie de pré-entraînement	Corpus
DrBERT	RoBERTa	À partir de zéro	NACHOS$_{large}$
DrBERT	RoBERTa	À partir de zéro	NACHOS$_{small}$
ChuBERT	RoBERTa	À partir de zéro	XBDW$_{small}$
ChuBERT	RoBERTa	À partir de zéro	XBDW$_{mixed}$
CamemBERT	RoBERTa	Pré-entraînement continué	NACHOS$_{small}$
PubMedBERT	BERT	Pré-entraînement continué	NACHOS$_{small}$
CamemBERT	RoBERTa	Pré-entraînement continué	XBDW$_{small}$

TABLE 2 – Liste des configurations des modèles pré-entraînés.

Le tableau 2 résume toutes les configurations évaluées dans cet article, intégrant à la fois l'étude de la taille des données et les stratégies de pré-entraînement.

4 Jeux de données et tâches d'évaluation

Pour évaluer les différentes configurations de pré-entraînement de nos modèles, un ensemble de tâches biomédicales est nécessaire. Si un tel *benchmark* spécifique au domaine existe pour l'anglais (BLURB (Gu *et al.*, 2021)), il n'en existe aucun pour le français. Dans cette section, nous décrivons un *benchmark* original, résumé dans le tableau 3, intégrant diverses tâches biomédicales de TALN pour le français. Parmi celles-ci, certaines proviennent de corpus publics, permettant la réplication de nos expériences. D'autres tâches proviennent de corpus privés et ne peuvent être partagées. Cependant,

Thématique / Corpus	Tâche	Métrique	Entraînement	Validation	Test
Corpus Publics					
ESSAIS (Dalloux *et al.*, 2021)	POS Tagging	Macro F1	9 693	2 077	2 078
CAS : Corpus de cas cliniques (Grabar *et al.*, 2018)	POS Tagging	Macro F1	5 306	1 137	1 137
MUSCA-DET - Déterminants sociaux de santé (Task 1)	REN imbriquée	Macro F1	19 861	2 207	5 518
MUSCA-DET - Déterminants sociaux de santé (Task 2)	Classification Multi-label	Macro F1	19 861	2 207	5 518
QUAERO French Medical Corpus - EMEA (Névéol *et al.*, 2014)	REN imbriquée	Weighted F1	11	12	15
QUAERO French Medical Corpus - MEDLINE (Névéol *et al.*, 2014)	REN imbriquée	Weighted F1	833	832	833
FrenchMedMCQA (Labrak *et al.*, 2022)	MCQA	EMR / Hamming Score	2 171	312	622
Corpus Privés					
Structuration de l'insuffisance cardiaque aiguë dans les comptes-rendus	REN	Macro F1	2 527	281	703
Classification de l'insuffisance cardiaque aiguë	Classification Binaire	Macro F1	1 179	132	328
Tri des comptes-rendus par spécialité	Classification Multi-classe	Macro F1	4 413	1 470	1 473
Structuration des prescriptions dans les comptes-rendus	REN	Macro F1	61	15	26

TABLE 3 – Corpus, tâches et métrique utilisés pour évaluer les modèles de langue.

elles sont utiles pour évaluer nos modèles avec plus de précision et ainsi observer leurs capacités de généralisation.

5 Résultats et discussions

Comme décrit précédemment, nous évaluons les performances de nos modèles de langue pré-entraînés sur un ensemble de tâches publiques et privées liées au domaine biomédical. Nous proposons d'analyser les résultats en fonction des différentes stratégies de pré-entraînement (section 5.1) puis de discuter de l'impact des données de pré-entraînement, que ce soit en termes de taille ou de nature (section 5.2).

	aHF NER			aHF classification			NER Medical Report			Specialities Classification		
	P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
CamemBERT OSCAR 138 GB	40,89	35,22	35,13	81,90	79,12	80,13	87,98	91,66	89,35	99,32	99,09	99,20
CamemBERT OSCAR 4 GB	46,32	43,17	42,66	81,49	81,42	81,41	87,79	90,74	88,78	99,53	99,69	99,61
CamemBERT CCNET 4 GB	47,25	42,2	43,11	82,02	79,30	79,98	87,61	92,28	89,34	99,54	99,55	99,55
DrBERT NACHOS$_{large}$	55,29	46,66	48,22	81,33	81,25	81,25	87,99	**92,80**	**89,83**	99,82	**99,90**	**99,86**
DrBERT NACHOS$_{small}$	54,55	43,39	45,93	79,85	80,10	79,87	87,57	92,76	89,44	**99,85**	99,85	99,85
ChuBERT XBDW$_{small}$	**56,92**	47,46	49,01	81,03	**82,67**	81,56	87,76	92,63	89,58	99,76	**99,90**	99,83
ChuBERT XBDW$_{mixed}$	54,62	47,81	**49,14**	82,23	81,71	**81,98**	87,42	92,36	89,30	99,81	99,82	99,81
CamemBERT NACHOS$_{small}$	22,02	16,67	16,08	74,86	69,82	69,80	65,72	68,49	66,74	99,44	99,67	99,54
PubMedBERT NACHOS$_{small}$	53,44	**48,21**	48,72	**83,06**	80,39	81,40	87,35	92,69	89,36	99,52	99,58	99,55
CamemBERT XBDW$_{small}$	25,44	19,33	19,12	79,50	74,74	76,02	68,80	71,23	69,64	99,60	99,57	99,58

TABLE 4 – Performance sur nos tâches biomédicales privées. Le meilleur modèle est en gras et le second est souligné. Pour les modèles CamemBERT NACHOS$_{small}$ et CamemBERT XBDW$_{small}$, le modèle CamemBERT OSCAR 138 GB a été utilisé comme abse initiale pour le pré-entraînement continué.

Tous les modèles ont été réglés finement (*fine-tuned*) de la même façon pour toutes les tâches. Tous les résultats rapportés sont la moyenne des scores de quatre exécutions. Les performances obtenues sur les tâches biomédicales sont présentées dans les tableaux 4 et 5 pour les tâches privées et publiques respectivement. Pour des raisons de lisibilité, la première partie de chaque tableau présente les résultats des modèles déjà existants, la deuxième partie nos modèles spécialisés entraînés à partir de zéro, et la dernière partie nos modèles utilisant un pré-entraînement continué.

5.1 Impact des stratégies de pré-entraînement

Comme montré dans les tableaux 4 et 5, les modèles pré-entraînés à partir de zéro (DrBERT NACHOS et ChuBERT XBDW) obtiennent les meilleurs scores F1 sur toutes les tâches privées et sur la majorité

	MUSCA-DET T1			MUSCA-DET T2			ESSAI POS			CAS POS			FrenchMedMCQA		QUAERO-EMEA			QUAERO-MEDLINE		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*Hamming*	*EMR*	*P*	*R*	*F1*	*P*	*R*	*F1*
CamemBERT OSCAR 138 GB	89,04	88,59	88,54	89,87	87,12	88,20	81,57	81,01	81,10	96,37	94,53	95,22	36,24	**16,55**	90,57	91,06	90,71	76,58	78,67	77,41
CamemBERT OSCAR 4 GB	86,09	85,45	85,43	92,68	90,34	91,27	84,01	83,51	83,69	98,15	95,34	96,42	35,75	15,37	90,75	91,16	90,83	78,55	79,33	78,76
CamemBERT CCNET 4 GB	91,12	89,91	90,33	93,10	90,42	91,38	85,60	85,63	85,42	**98,19**	**96,75**	**97,33**	34,71	14,41	90,31	90,59	90,33	78,06	78,11	77,61
DrBERT NACHOS$_{large}$	92,10	90,27	91,04	**94,97**	90,41	92,24	**90,96**	**89,19**	**89,75**	97,37	94,49	95,65	36,66	15,32	**91,93**	**92,52**	**92,09**	77,85	78,54	77,88
DrBERT NACHOS$_{small}$	93,35	90,62	91,77	91,31	86,60	88,57	90,12	88,37	88,76	97,04	94,88	95,70	**37,37**	13,34	91,54	92,00	91,66	77,91	79,34	78,18
ChuBERT XBDW$_{small}$	**94,88**	90,79	92,23	94,77	90,27	92,17	88,53	87,73	87,71	97,00	94,65	95,61	35,16	14,79	88,11	88,78	88,15	75,05	76,57	74,94
ChuBERT XBDW$_{mixed}$	94,39	**91,93**	**92,73**	94,22	90,02	91,71	86,36	85,50	85,73	97,77	95,30	96,35	34,58	12,21	90,36	90,94	90,52	78,61	79,32	78,63
CamemBERT NACHOS$_{small}$	81,44	81,39	80,96	79,74	78,08	78,70	80,59	79,88	80,04	95,64	91,57	92,46	32,87	13,76	67,56	77,48	71,10	55,45	62,34	57,43
PubMedBERT NACHOS$_{small}$	92,51	91,49	91,53	94,95	**92,55**	**93,62**	84,73	83,80	83,85	97,82	96,12	96,81	35,88	15,21	90,97	91,27	91,03	**82,03**	**81,71**	**81,73**
CamemBERT XBDW$_{small}$	82,35	81,59	81,57	78,14	76,38	77,12	79,44	79,79	79,25	95,98	92,11	93,18	27,73	11,89	53,44	73,11	61,75	48,71	61,33	53,05

TABLE 5 – Performance sur les tâches biomédicales publiques. Le meilleur modèle est en gras et le second est souligné. Pour les modèles CamemBERT NACHOS$_{small}$ et CamemBERT XBDW$_{small}$, le modèle CamemBERT OSCAR 138 GB a été utilisé comme abse initiale pour le pré-entraînement continué.

des tâches publiques (5 tâches sur 7). Les deux tâches publiques restantes (MUSCA-DET T2 et Quaero-Medline) sont mieux traitées avec PubMedBERT NACHOS$_{small}$, un modèle pré-entraîné deux fois, une première fois sur des données biomédicales anglaises, puis sur nos données biomédicales françaises (NACHOS$_{small}$). Nous observons également que le pré-entraînement continué à partir des modèles génériques français (CamemBERT NACHOS$_{small}$ ou CamemBERT XBDW$_{small}$) est moins performant que le pré-entraînement à partir de zéro. Enfin, les modèles état de l'art pré-entraînés sur des données génériques (CamemBERT OSCAR) restent compétitifs dans un certain nombre de tâches publiques biomédicales (CAS POS, FrenchMCQA ou MUSCA-DET T2), mais rencontrent plus de difficultés sur les tâches privées. Cela met en évidence la difficulté des tâches privées lorsque les données de pré-entraînement sont moins spécialisées.

5.2 Impact des données

En ce qui concerne la quantité de données utilisées pour le pré-entraînement des modèles ($small$ vs. $large$ ou $mixed$), les résultats montrent que plus les quantités de données sont grandes, plus les modèles sont performants, quelle que soit la stratégie de pré-entraînement ou la source de données (privée ou publique). Nous remarquons une nette domination des modèles pré-entraînés sur des sources web, notamment OSCAR et NACHOS, lorsqu'ils sont appliqués à des tâches publiques. En effet, les modèles reposant sur des données privées XBDW n'obtiennent les meilleures performances (en termes de score F1) que pour la tâche MUSCA-DET T1. Cette tendance n'est pas tout à fait observée sur les tâches privées, où les modèles fondés sur XBDW obtiennent des performances similaires, voire meilleures, lorsqu'ils sont mélangés avec des données biomédicales publiques (ChuBERT XBDW$_{mixed}$), comme le montre le tableau 4. Nous pensons que cette divergence est principalement due à la nature différente des données traitées. Enfin, nous observons que les modèles issus d'un pré-entraînement continué en partant d'un modèle spécialisé anglais (ici PubMedBERT) ont des performances supérieures à celles des modèles fondés sur CamemBERT, corroborant en partie notre hypothèse sur l'efficacité du transfert de connaissances inter-langues.

6 Conclusion

Dans ce travail, nous avons introduit les premiers modèles de langage français pour le domaine biomédical et clinique fondés sur l'architecture RoBERTa. Nous proposons aussi une étude comparative sur une collection de tâches biomédicales diverses provenant de sources privées et publiques. Nos modèles open-source DrBERT ont établi des performances à l'état de l'art dans la quasi-totalité des tâches biomédicales, surpassant le modèle généraliste français (CamemBERT). De plus, nous avons démontré que les pré-entraînements sur des ressources spécialisées de taille limitées (4 Go) obtenues sur le web permettent de très souvent dépasser les modèles entraînés avec des données spécialisées provenant de comptes-rendus médicaux.

Les modèles pré-entraînés ainsi que les scripts de pré-apprentissage [5] ont été publiés publiquement en ligne sous une licence open source MIT. Pour ce qui est du corpus NACHOS, l'objectif principal est de promouvoir le développement d'outils de TALN robustes par la communauté, nous avons donc décidé de rendre les corpus accessibles pour la recherche académique seulement.

Remerciements

Ce travail a été réalisé grâce aux ressources de GENCI-IDRIS (Grant 2022-AD011013061R1 and 2022-AD011013715) et du CCIPL (Centre de Calcul Intensif des Pays de la Loire). Ce travail a été soutenu financièrement par l'ANR AIBy4 (ANR-20-THIA-0011) and Zenidoc.

Références

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.

CHALKIDIS I., FERGADIOTIS M., MALAKASIOTIS P., ALETRAS N. & ANDROUTSOPOULOS I. (2020). LEGAL-BERT : The muppets straight out of law school. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2898–2904, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.261.

DALLOUX C., CLAVEAU V., GRABAR N., OLIVEIRA L. E. S., MORO C. M. C., GUMIEL Y. B. & CARVALHO D. R. (2021). Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, **27**(2), 181–201. DOI : 10.1017/S1351324920000352.

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.

5. https://drbert.univ-avignon.fr/

EL BOUKKOURI H., FERRET O., LAVERGNE T. & ZWEIGENBAUM P. (2022). Re-train or Train from Scratch ? Comparing Pre-training Strategies of BERT in the Medical Domain. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, p. 2626–2633, Marseille, France : European Language Resources Association.

FITZGERALD J., HENCH C., PERIS C., MACKIE S., ROTTMANN K., SANCHEZ A., NASH A., URBACH L., KAKARALA V., SINGH R., RANGANATH S., CRIST L., BRITAN M., LEEUWIS W., TUR G. & NATARAJAN P. (2022). Massive : A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. DOI : 10.48550/ARXIV.2204.08582.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French Corpus with Clinical Cases. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, p. 1–7, Brussels, Belgium. HAL : hal-01937096.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, **3**(1), 1–23. DOI : 10.1145/3458754.

HOWARD J. & RUDER S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 328–339, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1031.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). Clinicalbert : Modeling clinical notes and predicting hospital readmission. DOI : 10.48550/ARXIV.1904.05342.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French Multiple-Choice Question Answering Dataset for Medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, Abou Dhabi, United Arab Emirates. HAL : hal-03824241.

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*. DOI : 10.1093/bioinformatics/btz682.

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. *CoRR*, **abs/1907.11692**.

LUCCIONI A. S., VIGUIER S. & LIGOZAT A.-L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model.

MARTIN L., MULLER B., SUÁ REZ P. J. O., 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 (ACL)*, p. 7203—-7219 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

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.

PENG Y., YAN S. & LU Z. (2019). Transfer learning in biomedical natural language processing : An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, p. 58–65.

TIEDEMANN J. & NYGAARD L. (2004). The OPUS corpus - parallel and free : http://logos.uio.no/opus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & 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 (NIPS)*, volume 30 : Curran Associates, Inc.

ZHANG Y., WARSTADT A., LI X. & BOWMAN S. R. (2021). When do you need billions of words of pretraining data ? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1112–1125, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.90.

A Annexes

A.1 Sources NACHOS

Ressource	# mots
HAL	638 508 261
Haute Autorité de Santé (HAS)	113 394 539
Drug leaflets	74 770 229
Scrapping Web Médical	64 904 334
ANSES SAISINE	51 372 932
Base de données publique des médicaments (BDPM)	48 302 695
ISTEX	44 124 422
CRTT	26 210 756
WMT-16	10 282 494
EMEA-V3	6 601 617
Wikipedia Science de la vie	4 671 944
ANSES RCP	2 953 045
Cerimes	1 717 552
LiSSa	235 838
DEFT-2020	231 396
CLEAR	225 898
CNEDiMTS	175 416
QUAERO French Medical Corpus	72 031
ANSM Registre des essais cliniques	47 678
ECDC	44 482
QualiScope	12 718
WMT-18-Medline	7 673
Total	**1,088,867,950**

TABLE 6 – Sources de données comprises dans le corpus NACHOS.

A.2 Impact écologique

Une quantité considérable de ressources de calcul a été utilisée pour mener cette étude, puisqu'environ 18 000 heures de calcul GPU ont été utilisées pour créer les 7 modèles présentés ici, ainsi qu'environ 7 500 heures de GPU pour le débogage en raison de problèmes techniques liés aux configurations des modèles et à la mauvaise performance, pour un total de 25 500 heures. Le coût environnemental total, selon la documentation du supercalculateur Jean Zay [6] équivaut à 6 604 500 Wh soit 376,45 kg CO2eq

6. http://www.idris.fr/media/jean-zay/jean-zay-conso-heure-calcul.pdf

basé sur l'intensité carbone du réseau énergétique mentionné par l'étude de coût environnemental de BLOOM, qui elle aussi a été réalisée sur Jean Zay (Luccioni *et al.*, 2022).

A.3 Pré-traitement des corpus d'évaluation

L'extraction d'entités imbriquées n'est pas directement réalisable avec un modèle BERT sans adapter le corpus. Parmi les corpus d'évaluation, deux corpus traitent la tâche de reconnaissance d'entités nommées imbriquées : QUAERO et MUSCA-DET. Pour simplifier le processus d'évaluation de ces corpus, nous trions les étiquettes imbriquées par ordre alphabétique et les concaténons en une seule pour transformer la tâche en un format utilisable pour la classification de tokens avec les architectures basées sur BERT.

A.4 Évaluation généraliste sur le français

Le tableau 7 donne les résultats obtenus par tous les modèles sur les tâches généralistes. Ces tâches proviennent de Martin *et al.* (2020) et ont été utilisées pour évaluer les différents modèles CamemBERT. Les quatre premières sont des tâches d'étiquetage morphosyntaxique POS (GSD, SEQUOIA, SPOKEN et PARTUT) et la dernière est une tâche d'inférence en langage naturel (XNLI).

Tous les résultats de nos modèles diminuent les performances sur toutes les tâches. La baisse la plus importante concerne la tâche d'inférence en langage naturel, avec une performance de ChuBERT NBDW$_{small}$ presque 13% inférieure à celle de CamemBERT 138 Go. Nous observons également que les modèles spécialisés en anglais sont aussi performants que nos modèles biomédicaux en français. Il semble assez clair d'après les observations précédentes que les modèles spécialisés sont difficiles à généraliser à d'autres tâches, mais que des informations spécialisées capturées dans une langue pourraient être transférées dans une autre langue.

	GSD	SEQUOIA	SPOKEN	PARTUT	XNLI
CamemBERT OSCAR 138 Go	**98.28**	98.68	97.26	97.70	**81.94**
CamemBERT OSCAR 4 Go	98.14	**99.18**	**97.57**	97.86	81.76
CamemBERT CCNET 4 Go	98.18	98.92	97.20	**97.92**	81.26
PubMedBERT	96.48	96.49	90.00	93.97	73.79
BERT clinique	96.49	96.31	89.60	93.17	70.57
BioBERT v1.1	97.32	96.54	91.81	94.52	71.54
DrBERT NACHOS$_{large}$	96.94	98.05	95.92	96.54	72.18
DrBERT NACHOS$_{small}$	97.17	98.21	96.38	96.45	72.86
ChuBERT NBDW$_{small}$	96.45	97.38	94.90	95.83	69.00
ChuBERT NBDW$_{mixed}$	97.18	98.10	96.43	96.33	72.32
CamemBERT NACHOS$_{small}$	97.63	96.90	91.12	94.00	71.26
PubMedBERT NACHOS$_{small}$	97.41	98.71	95.54	97.01	77.35
CamemBERT NBDW$_{small}$	97.55	96.26	89.17	91.34	72.73

TABLE 7 – Performance sur les tâches généralistes. Le meilleur modèle est en gras et le deuxième est souligné.

A.5 Intersection des vocabulaires

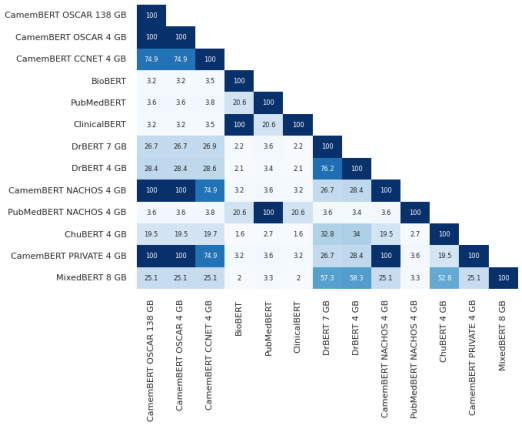
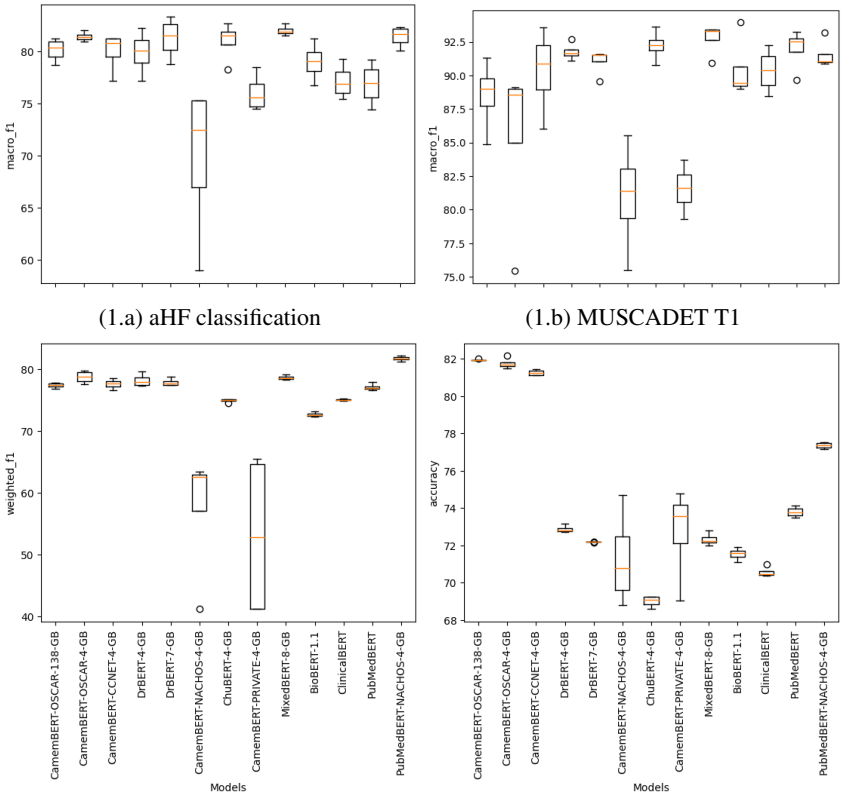


FIGURE 1 – Matrice d'intersection des vocabulaires.

Comme nous pouvons le voir sur la figure 1, malgré des performances similaires, certains modèles ne partagent pas beaucoup de vocabulaire mutuel.

A.6 Stabilité des modèles

Nous observons lors de la phase d'évaluation que la plupart des modèles basés sur la stratégie de pré-formation continue de CamemBERT OSCAR 138 Go souffrent d'une mauvaise stabilité lors de l'affinage, ce qui se traduit par une fluctuation des performances entre les exécutions.



(1.a) aHF classification

(1.b) MUSCADET T1

(2.c) QUAERO MEDLINE

(2.d) XNLI

FIGURE 2 – Boîte à moustaches pour chaque modèle.