Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Classification multi-label de cas cliniques avec CamemBERT

Alexandre Bailly[1, *] Corentin Blanc[1, *] Thierry Guillotin [1]
(1) Everteam Software, 17 quai Joseph Gillet, 69004 Lyon, France
(*) Contributions égales
a.bailly@everteam.com, c.blanc@everteam.com, t.guillotin@everteam.com

RÉSUMÉ

La quantité de documents textuels médicaux allant grandissant, la nécessité d'en extraire automatiquement des informations concernant des patients devient de plus en plus grande. La prédiction du profil clinique permet de gagner du temps pour le praticien tout en extrayant l'essentiel de l'information concernant un patient. Avec l'explosion du nombre de documents (médicaux ou non), des modèles pré-entraînés tels que BERT pour l'anglais ou CamemBERT pour le français ont émergé. L'utilisation de ces modèles permet d'encoder contextuellement du texte afin de l'utiliser dans des réseaux neuronaux pour notamment prédire des profils cliniques. Cet article vise à comparer différentes méthodes de prédiction de profil clinique en se basant sur l'utilisation de CamemBERT. Dans un premier temps, uniquement du texte provenant de documents médicaux a été utilisé. Dans un second temps, des entités nommées ont été injectées en plus du texte par concaténation ou par sommation pondérée. Les résultats ont montré un succès limité et dépendant de la prévalence des chapitres à prédire dans le corpus ainsi qu'une dégradation des performances lors de l'ajout des entités nommées.

ABSTRACT

Multi-label classification of clinical cases with CamemBERT

As quantity of textual medical data is increasing, the necessity to extract automatically information about patients increases accordingly. Predicting the clinical profile of a patient record allows to save time for praticians by exhibiting essential information concerning the patient. Together with the explosion in the number of documents (medical or not), pretrained models such as BERT for english or CamemBERT for french has emerged. Using these models allows to encode contextually a text to this encoded representation in neural networks notably for NLP tasks such as predicting clinical profiles. This article aims to compare different methods of clinical profile prediction based on CamemBERT. In a first time, only the text from medical documents was used. In a second time, named entities were injected in addition to the text by concatenation or pondered sum. Results show a limited success depending on the prevalence of the chapters to predict in the corpus as well as a decrease of performances with the use of named entitie types.

MOTS-CLÉS : Classification multi-label ; Fouille de texte ; CamemBERT.

KEYWORDS: Multi-label classification ; Data mining ; CamemBERT.

1 Introduction

Avec l'augmentation du nombre de consultations médicales, la quantité de documents textuels concernant les patients a considérablement augmenté. Les divers compte-rendus de consultation ou de

Atelier DÉfi Fouille de Textes (DEFT), pages 14–20.

prise en charge hospitalière forment une masse de données importante et riche en informations sur le patient. Ces informations sont très intéressantes pour les différents praticiens et leur récupération est un enjeu important. L'extraction du profil clinique d'un patient (l'ensemble des pathologies associées à son cas) à partir d'un document textuel peut prendre un temps important, au détriment du temps accordé au patient. Cette étape reste néanmoins indispensable et une automatisation de l'extraction est une bonne alternative pour obtenir les informations nécessaires.

Cette explosion du nombre de documents médicaux a poussé la communauté scientifique à créer de nouveaux modèles de langue facilitant leur traitement. Ces modèles pré-entraînés sur des quantités colossales de données permettent d'encoder une phrase ainsi que les mots la constituant en tenant compte de leur contexte. Le plus connu est BERT (Bidirectionnal Encoder Representation from Transformers) qui a permis d'améliorer l'état de l'art sur une grande majorité des tâches de Traitement Automatique du Langage Naturel (TALN) en anglais (Devlin *et al.*, 2019). Suite à ce succès, de nombreux autres modèles dérivés de BERT ont vu le jour comme CamemBERT (Martin *et al.*, 2020) pour le français.

Ce papier vise à étudier la prédiction du profil clinique d'un patient en utilisant à la fois du texte brut provenant de documents médicaux mais aussi différentes entités nommées qui ont été préalablement mises en évidence. Trois approches seront étudiées : le traitement du texte brut par CamemBERT dans un premier temps puis l'injection par concaténation et sommation pondérée des entités nommées au texte brut dans un second temps.

2 Matériel et méthodes

2.1 Données

Ces travaux se situent dans le contexte de la compétition DEFT-21 (Grouin *et al.*, 2021). Le corpus de DEFT 2021 était constitué de 275 cas cliniques répartis en un jeu de d'entraînement (167) et un jeu de test (108). Chaque cas clinique était composé d'un texte brut accompagné d'un certain nombre d'entités nommées préalablement identifiées parmi 19 types distincts comme le montre l'exemple sur la Figure 1.

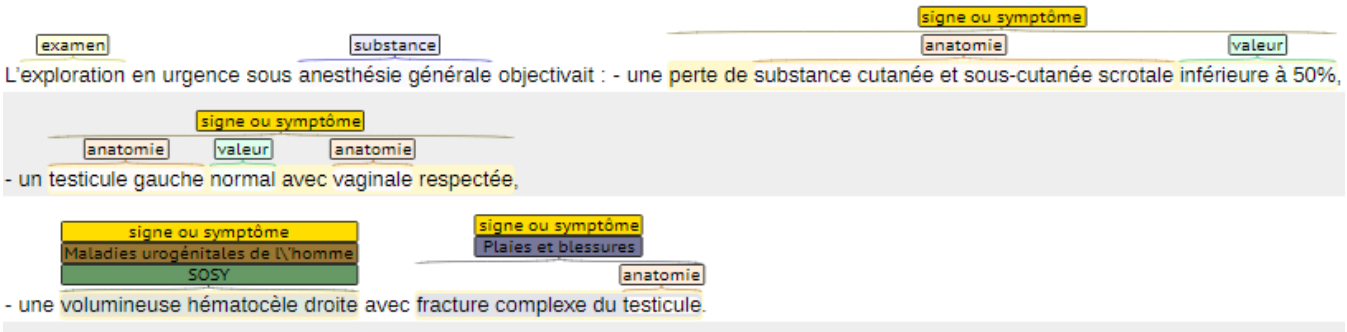

FIGURE 1 – Exemple de texte brut accompagné d'entités nommées d'un cas clinique

Un profil clinique constitué d'au moins un des 23 chapitres du MeSH a été attribué à tous les cas cliniques. Au sein du corpus, la distribution des chapitres était extrêmement déséquilibrée comme suit sur la Table 1.

Chapitre	Entraînement	Test	Total
Stomatognathique	3	3	6
Parasitaires	6	1	7
Virales	8	4	12
ORL	10	3	13
Oeil	11	9	20
Génétique	19	10	29
Immunitaire	20	11	31
Endocriniennes	22	14	36
Blessures	21	19	40
Osteomusculaires	21	22	43
Respiratoire	27	17	44
Peau	31	16	47
Nutritionnelles	29	23	52
Digestif	36	22	58
Hémopathies	34	25	59
Infections	33	27	60
Femme	33	32	65
Cardiovasculaires	39	27	66
Chimiques	45	22	67
Nerveux	41	46	87
Homme	63	36	99
Tumeur	80	51	131
Etatsosy	141	101	242

TABLE 1 – Distribution des chapitres du MeSH dans les données

Dans la suite, certains chapitres trop peu représentés ont été volontairement écartées des possibilités de prédiction afin d'améliorer celles des autres chapitres. Les chapitres écartés sont les suivants : *stomatognathiques* (3), *parasitaires* (6), *virales* (8), *ORL* (10) et *oeil* (11).

2.2 Approches proposées

CamemBERT est un modèle de langue pré-entraîné sur une énorme quantité de données permettant d'encoder le contexte d'une phrase. La meilleure manière d'appliquer un tel modèle à une tâche TALN est de connecter une couche de sortie pour réaliser la prédiction puis de régler finement tous les poids de bout en bout.

Dans la suite, chacune des méthodes introduites utilise le modèle de langue CamemBERT. Pour un cas clinique donné, toutes les phrases du texte brut ont été encodées par CamemBERT puis moyennées afin d'obtenir une unique représentation du cas clinique. Cette représentation a ensuite été utilisée afin d'effectuer la prédiction.

2.2.1 Texte brut

Dans un premier temps, une couche linéaire de 18 neurones avec une sigmoïde comme fonction d'activation a été connectée à CamemBERT afin de calculer une probabilité pour chacun des 18 chapitres précédemment retenus.

2.2.2 Injection des entités nommées au texte brut

Dans un second temps, les entités nommées extraites de chaque document ont été utilisées pour enrichir les entrées du modèle. Pour chaque cas clinique, les types d'entités nommées ont été encodées grâce à un vecteur binaire représentant la présence (1) ou l'absence (0) au sein du texte associé au cas clinique. Afin de les injecter au texte brut, deux méthodes ont été utilisées :

— Concaténation : l'encodage du texte brut était concaténé à celui des entités nommées.
— Somme pondérée : l'encodage du texte brut était sommé à une projection linéaire de l'encodage des entités nommées. La projection tout comme la pondération étaient apprises lors de la phase d'entraînement.

Ces deux méthodes ont permis de construire de nouvelles représentations du cas clinique qui ont ensuite été utilisées dans un couche linéaire de 18 neurones, couplée à une fonction sigmoïde, afin de calculer une probabilité pour chacun des chapitres retenus.

2.3 Paramètres d'entraînement

Pour toutes les approches, le nombre d'epochs a été fixé à 5 au vue de la convergence de la Binary Cross-Entropy Loss. L'optimisateur AdamW (Loshchilov & Hutter, 2019) a été utilisé avec un taux d'apprentissage fixé à 5e-3 sur la première epoch puis diminuant linéairement. Pour chacune des trois méthodes proposées, un seuil par chapitre a été recherché afin d'optimiser les résultats.

2.4 Évaluation

La recherche des paramètres d'entraînement a été effectuée par une méthode de bootstrap (Efron, 1979) à partir du corpus d'entraînement. Une fois les paramètres d'entraînement sélectionnés, les modèles ont été entraînés sur la globalité du corpus d'entraînement. Trois métriques ont été utilisées pour évaluer les performances sur chaque chapitre et de manière globale : le rappel, la précision et le f1-score. Les résultats présentés ci-après sont ceux obtenus sur le jeux de test.

3 Résultats

Évaluation par chapitre Pour chacun des différents modèles qui ont été entraînés, les performances obtenues pour les prédictions dépendent des chapitres et ne différent pas grandement d'une méthode à l'autre. En effet, comme il est visible dans la table 2, le f1-score pour les différents chapitres se situe entre 0.105 et 0.967. Les chapitres les moins représentés dans le corpus sont ceux qui présentent

	Rappel			Précision			F1		
	M1†	M2*	M3$	M1†	M2*	M3$	M1†	M2*	M3$
Stomatognathique	—	—	—	—	—	—	—	—	—
Parasitaires	—	—	—	—	—	—	—	—	—
Virales	—	—	—	—	—	—	—	—	—
ORL	—	—	—	—	—	—	—	—	—
Oeil	—	—	—	—	—	—	—	—	—
Génétique	**1.000**	0.700	0.400	**0.118**	0.091	0.103	**0.211**	0.161	0.163
Immunitaire	0.091	**0.364**	0.091	**0.250**	0.154	0.067	0.133	**0.216**	0.077
Endocriniennes	**1.000**	0.857	0.857	**0.140**	0.126	0.121	**0.246**	0.220	0.212
Blessures	0.158	**0.684**	0.158	0.188	0.171	**0.214**	0.171	**0.274**	0.182
Osteomusculaires	0.000	0.318	**0.409**	0.000	**0.250**	0.191	0.000	**0.280**	0.261
Respiratoire	0.235	**0.471**	0.176	**0.286**	0.167	0.075	**0.258**	0.246	0.105
Peau	0.312	0.688	**0.750**	**0.217**	0.125	0.124	**0.256**	0.212	0.212
Nutritionnelles	**1.000**	0.652	0.696	0.213	0.217	**0.229**	**0.351**	0.326	0.344
Digestif	**0.773**	0.591	0.545	**0.250**	0.197	0.200	**0.378**	0.295	0.293
Hémopathies	**0.520**	0.320	0.120	**0.371**	0.205	0.150	**0.433**	0.250	0.133
Infections	0.778	0.741	**0.926**	**0.292**	0.267	0.255	**0.424**	0.392	0.400
Femme	**0.844**	0.719	0.719	**0.386**	0.303	0.307	**0.529**	0.426	0.430
Cardiovasculaires	**1.000**	**1.000**	**1.000**	0.250	0.250	**0.260**	0.400	0.400	**0.412**
Chimiques	0.182	**0.273**	0.091	**0.571**	0.182	0.065	**0.276**	0.218	0.075
Nerveux	**0.457**	0.304	0.348	**0.636**	0.359	0.457	**0.532**	0.329	0.395
Homme	0.472	**0.722**	0.694	**0.500**	0.342	0.321	**0.486**	0.464	0.439
Tumeur	**0.941**	0.608	0.745	**0.623**	0.397	0.458	**0.750**	0.481	0.567
Etatsosy	**1.000**	0.931	**1.000**	**0.935**	0.931	**0.935**	0.935	0.931	**0.967**
Global	**0.683**	0.651	0.637	**0.370**	0.283	0.298	**0.480**	0.394	0.406

†Modèle textuel uniquement - * Modèle avec concaténation - $ Modèle avec pondération

TABLE 2 – Résultat obtenu pour chaque chapitre du MeSH

les f1-scores les plus faibles, et ce quel que soit le modèle. Le f1-score pour le chapitre *génétique*, qui est le moins représenté, est de 0.161 pour le modèle avec concaténation, de 0.163 pour celui avec pondération et de 0.211 pour le modèle n'utilisant que le texte brut. Au contraire, les chapitres les plus représentés dans le corpus présentent de bon résultats, notamment pour *etatsosy*, avec des f1-score de 0.935, 0.931 et 0.967 respectivement pour le modèle utilisant seulement le texte, le modèle avec la concaténation et le modèle avec la pondération. Le chapitre *chimiques* fait ici figure d'exception, avec des f1-scores de 0.276, 0.218 et seulement 0.075 respectivement, alors qu'il fait partie des chapitres avec les plus grandes prévalences. Les faibles performances en terme de f1-score pour les chapitres sous-représentés sont dues à une faible précision. En effet, pour le chapitre *endocriniennes* par exemple, le modèle avec pondération a obtenu un rappel de 0.857 mais une précision de seulement 0.121, ce qui explique alors le f1-score de 0.212.

Comparaison des modèles L'évaluation globale des modèles montre que quelque soit la métrique observée, le modèle n'utilisant que le texte obtient de meilleurs performances. En terme de f1-score, le modèle utilisant seulement le texte atteint 0.480 alors que les modèles utilisant la concaténation et

la pondération n'atteignent respectivement que 0.394 et 0.406. La comparaison de ces deux dernières valeurs semble indiquer que la pondération conduit à de meilleurs performances que la concaténation.

4 Discussion

La prédiction des différents chapitres est plus ou moins bonne selon leur prévalence dans le corpus d'entraînement. En effet, les chapitres les moins représentés ont tendance à être moins bien prédits. La faible prévalence de certains chapitres semble donc être un frein considérable pour tous les modèles comparés lors de l'apprentissage.

La moyenne des représentations des phrases d'un texte obtenues grâce à CamemBERT permet dans une certaine mesure d'attribuer les chapitres associés à ce même texte. Néanmoins, les performances de ce modèle restent limitées. Cela peut être notamment dû à la quantité limités de cas cliniques disponibles pour l'entraînement. En effet, l'utilisation de CamemBERT implique un grand nombre de poids à entraîner et donc nécessite beaucoup de données pour l'entraînement. Une autre explication pourrait être le fait que l'ensemble du texte a été considéré pour la prédiction, alors que l'information recherchée peut n'être présente que dans une partie seulement. Certaines phrases pourrait donc être à l'origine de bruit dans les données.

L'utilisation des entités nommées pour enrichir le texte était supposée apporter davantage d'informations et permettre d'améliorer la prédiction. Cependant, les modèles les incluant ont vu leurs performances se dégrader et ce peu importe la façon dont elles ont été injectées. L'information de la seule présence des entités dans le texte ne semble donc ne pas être suffisante pour améliorer les prédictions.

5 Conclusion

Dans une certaine mesure, l'utilisation du modèle pré-entraîné CamemBERT a permis de retrouver les chapitres du MeSH associés à différents cas cliniques à partir du texte. En revanche, la quantité de données présente n'a pas permis d'atteindre de bonne performances avec cette méthode. L'ajout de la présence de certaines entités nommées a eu pour seul effet de dégrader légèrement les performances initiales.

Références

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

EFRON B. (1979). Bootstrap Methods : Another Look at the Jackknife. *Annals of Statistics*, **7**, 1–26.

GROUIN C., GRABAR N. & ILLOUZ G., Éds. (2021). *Actes de TALN 2021 (Traitement automatique des langues naturelles)*, Lille.

LOSHCHILOV I. & HUTTER F. (2019). Decoupled Weight Decay Regularization. *arXiv :1711.05101 [cs, math]*.

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.