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

Identification de profil clinique du patient: Une approche de classification de séquences utilisant des modèles de langage français contextualisés

Aidan Mannion[1,2] Thierry Chevalier[3] Didier Schwab[1] Lorraine Goeuriot[1]
(1) Univ. Grenoble Alpes, CNRS, LIG, 38000 Grenoble, France
(2) EPOS, 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

RÉSUMÉ

Cet article présente un résumé de notre soumission pour Tâche 1 de DEFT 2021. Cette tâche consiste à identifier le profil clinique d'un patient à partir d'une description textuelle de son cas clinique en identifiant les types de pathologie mentionnés dans le texte. Ce travail étudie des approches de classification de texte utilisant des plongements de mots contextualisés en français. À partir d'une base de référence d'un modèle constitué pour la compréhension générale de la langue française, nous utilisons des modèles pré-entraînés avec *masked language modelling* et affinés à la tâche d'identification, en utilisant un corpus externe de textes cliniques fourni par SOS Médecins, pour développer des ensembles de classifieurs binaires associant les textes cliniques à des catégories de pathologies.

ABSTRACT

Identification of patient clinical profiles : A sequence classification approach using contextualised French language models

This article summarises our submission to Task 1 of the text mining challenge DEFT 2021. This task involved the identification of the clinical profile of a patient from a textual description of their clinical case by identifying all the types of pathology mentioned in the text. This work investigates the utility of text classification approaches using contextualised French-language vector embeddings. Beginning from a baseline of a model trained for general French-language understanding, we employ both masked-language pre-training and fine-tuning on the DEFT task, using an external corpus of clinical text provided by SOS Médecins, to develop ensembles of binary classifiers to associate pathology types with a given segment of clinical text.

MOTS-CLÉS : TALN biomédicale, Classification des séquences, FlauBERT, plongements de mots contextualisé.

KEYWORDS: Biomedical NLP, Sequence classification, FlauBERT, contextualised word embeddings.

Atelier DÉfi Fouille de Textes (DEFT), pages 54–62.

1 Introduction

La tâche d'identification de profils cliniques à partir de données biomédicales textuelles est d'un grand intérêt pour les institutions, les entreprises et les praticiens du domaine médical. Le problème peut se montrer assez difficile, principalement à cause de la forme non-structurée des données textuelles ainsi que la complexité et la spécificité du domaine. Des développements récents dans le domaine de traitement du langage naturel, et plus particulièrement les plongements de mots contextualisés basés sur l'entraînement de réseaux de neurones avec l'architecture *transformer* (notamment (Devlin *et al.*, 2018)), montrent le potentiel de modélisation des dépendances complexes et à longue portée. Ces nouveaux modèles donnent à la communauté TALN biomédicale des pistes d'expérimentation pour l'amélioration d'extraction d'informations complexes à partir de dossiers de santé textuels.

Certaines études ont montré l'utilité de ces méthodes pour développer des outils en TALN biomédical en anglais (Huang *et al.*, 2019; Alsentzer *et al.*, 2019; Lee *et al.*, 2020) et leur application à diverses tâches du domaine (Yoon *et al.*, 2019; Peng *et al.*, 2019; Blinov *et al.*, 2020). À notre connaissance, il existe peu de travaux sur les applications des modèles de langue neuronaux contextualisés sur des tâches biomédicales sur le français comme celles de DEFT 2021.

Le modèle neuronal FlauBERT (Le *et al.*, 2020), un modèle de type *transformer* (voir la section 2.1) entraîné de manière auto-supervisée sur un corpus de textes de langue générale en français, est utilisé dans ce travail. FlauBERT est un modèle transformer bidirectionnel avec la même architecture que BERT (Devlin *et al.*, 2018) qui a été entraîné sur un corpus français hétérogène extrait de divers sources.

Plus spécifiquement, la tâche 1 du Défi Fouilles de Texte 2021 (Grouin *et al.*, 2021) vise à identifier le profil clinique d'un patient par le type de maladie de toutes les pathologies présente dans le texte associé avec un cas clinique. Nous formulons cette tâche comme un problème de classification des vecteurs représentant des séquences de texte. Parce qu'il peut y avoir plusieurs catégories associées à un document, il n'est pas possible d'utiliser un classificateur multi-label standard (dans plusieurs cas, plusieurs annotations sont mises en association avec le même mot dans le document source). Nous entraînons un classificateur binaire pour chacune des catégories de sortie. De cette manière, les modèles apprennent indépendamment les corrélations entre les plongements vectoriels et les variables cibles, à partir d'une base de réference qui utilise un modèle entraîné sur une tâche non-supervisée, détaillée dans la section 2.3. Les modèles utilisés pour les expériences sont pre-entraînés sur le corpus de SOS Médecins détaillés en section 2.2, et adapté pour la tâche d'identification de profils cliniques de deux manières ; en s'appuyant sur la version étiquetée du corpus de SOS (aussi détaillé en section 2.2), et finalement sur le corpus d'entraînement DEFT fourni pour cette tâche, sur lequel les résultats d'entraînement sont montrés dans la section 3.1, ainsi que les résultats sur le corpus d'évaluation.

2 Entraînement des modèles de langage

Divers types de vecteurs ont été expérimentés pour représenter du texte clinique (Khattak *et al.*, 2019), mais étant donné l'énorme complexité des relations possibles entre les entités biomédicales, et leur importance potentielle pour l'identification des profils cliniques, il est pertinent de penser que les modèles contextualisés peuvent apporter des améliorations aux tâches de classification automatique.

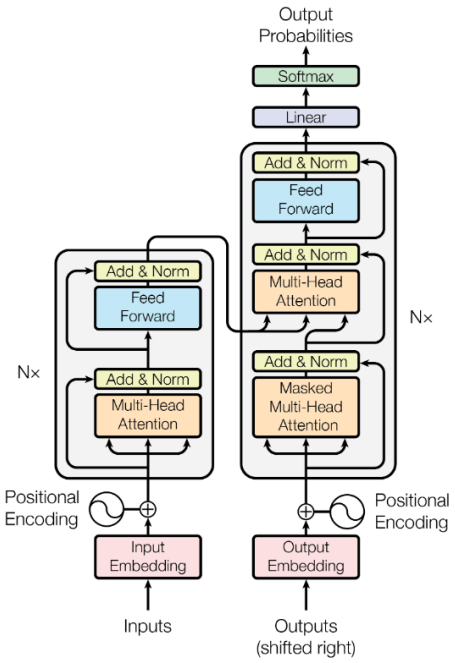

FIGURE 1 – L'architecture d'un transformer : diagramme issu de (Vaswani *et al.*, 2017).

2.1 Les réseaux de neurones *transformer*

Les réseaux transformer (Vaswani *et al.*, 2017) impliquent un modèle encodeur-décodeur neuronal mettant en œuvre une technique appelée auto-attention. Le principal avantage de ce méchanisme est de modèliser les dépendances à long terme dans le texte. Il utilise une opération appelée attention de produit scalaire qui crée une *matrice d'attention* pour chaque séquence de jetons d'entrée dans laquelle chaque composant a une probabilité égale afin d'être associés sémantiquement avec tout autre composant, quel que soit le nombre de composants entre eux. Un diagramme classique de l'architecture est montré en Figure 1.

2.2 Corpus SOS Médecins

Pour adapter les plongements FlauBERT (appris sur du vocabulaire général) au domaine clinique, nous utilisons un corpus de textes cliniques qui est composé de commentaires écrits par des médecins en consultation avec des patients dans le cadre des appels au service SOS Médecins.

L'objectif de l'entraînement supplémentaire avec ce corpus est d'améliorer les performances sur les séquences de texte du corpus DEFT. Étant donné que l'utilisation des réseaux transformer présente plus d'avantage dans le traitement des phrases longues avec plus de complexité en termes de dépendances linguistiques, nous filtrons les documents du corpus pour ne garder que ceux ayant un nombre de mots supérieur à un seuil (28). Cela nous donne un corpus d'apprentissage de 324 753 documents cliniques, avec un longueur (no. mots) moyenne de 41.3 et un longueur maximale de 390.

Nous disposons dans ce corpus de codes de diagnostics ajoutés aux données de la consultation manuellement par le médecin à l'issue de la visite au patient. Le format de ces codes est spécifique à SOS Médecins, mais nous avons bénéficié de l'expertise du deuxième auteur, également médecin, pour les associer aux catégories MeSH correspondantes. Cette conversion impliquait l'effacement de

certaines catégories des codes SOS, parce que ces codes sont plus spécifiquement des classifications de "résultats de la consultation", donc pas forcément toujours un diagnostic de pathologie qui peut être associé avec une chapitre de MeSH pertinente pour la tâche DEFT.

Dans ce travail, nous entraînons et comparons des classificateurs de profil MeSH à partir de trois différentes variantes de l'entraînement supplémentaire avec le corpus SOS Médecins ; un modèle "pre-entraîné" (section 2.3) un modèle adapté directement pour une variante de la tâche DEFT (section 2.4) et un avec les deux ensembles (le pre-entraînement en premier). Les étapes d'optimisation des réseaux de neurones modélisation de mots masqués (section 2.3) ainsi que l'adaptation aux tâches de classification (section 2.4) était fait en utilisant l'algorithme Adam (Kingma & Ba, 2015).

2.3 Pré-entraînement : modélisation des mots masqués

La tâche de pré-entraînement consiste à cacher aléatoirement des jetons dans le corpus d'entrée avec des jetons spéciaux appelés "masques" ; l'objectif de l'entraînement du réseau est alors de prédire le jeton caché.

Nous décrivons ici les expériences faites avec le modèle $\text{FlauBERT}_{\text{BASE}}$, qui consiste en 12 couches, 12 têtes d'attention et qui a une dimension maximale de plongements de 768, pour un total de 138M de paramètres.

Pour suivre la performance du modèle lors de l'entraînement, nous utilisons la perplexité, une métrique largement utilisée pour l'évaluation des modèles prédictifs non-supervisés ; c'est une technique avec ses origines provenant de la théorie d'information qui sert à comparer deux distributions de probabilité en prenant un moyen géométrique pondéré des inverses des probabilités sorties par un modèle pour un certain ensemble de données. Le corpus d'entraînement est divisé en deux parties, "train" et "eval", et à la fin de chaque époque d'entraînement, où le modèle s'entraîne sur le sous-ensemble *train*, la perplexité du modèle est évaluée sur le sous-ensemble *eval*. Le sous-ensemble *train* consiste à 80% du corpus d'entrée, choisi aléatoirement.

Pour notre pré-entraînement avec le corpus SOS Médecins, nous utilisons un seuil de convergence de 0.05 pour la perplexité, c'est-à-dire que l'entraînement s'est arrêté après avoir atteint une époque pour laquelle la perplexité a diminué de moins de 0.05. Le modèle utilisé dans les expériences (nommé "Base + MLM" dans la section 3.1) a complété 8 époques avant d'atteindre ce seuil. Bien qu'il s'agisse d'un nombre d'époques beaucoup plus faible que ce qui est généralement accepté comme raisonnable pour ce type d'entraînement, les contraintes de temps et de ressources ont nécessité l'utilisation ce seuil. La diminution de perplexité à travers des époques d'entraînement est montrée en Figure 2.

Les hyperparamètres utilisés pour l'entraînement étaient les suivants ;

- Une *probabilité de masquage*, c'est-à-dire la proportion des tokens d'entrée qui ont été cachés avec le token spécial `[MASK]`, de 0.15, comme c'est la norme spécifiée par les développeurs de BERT,
- Un taux d'apprentissage de 5×10^{-5},
- Une longueur de séquence maximale de 256 tokens.

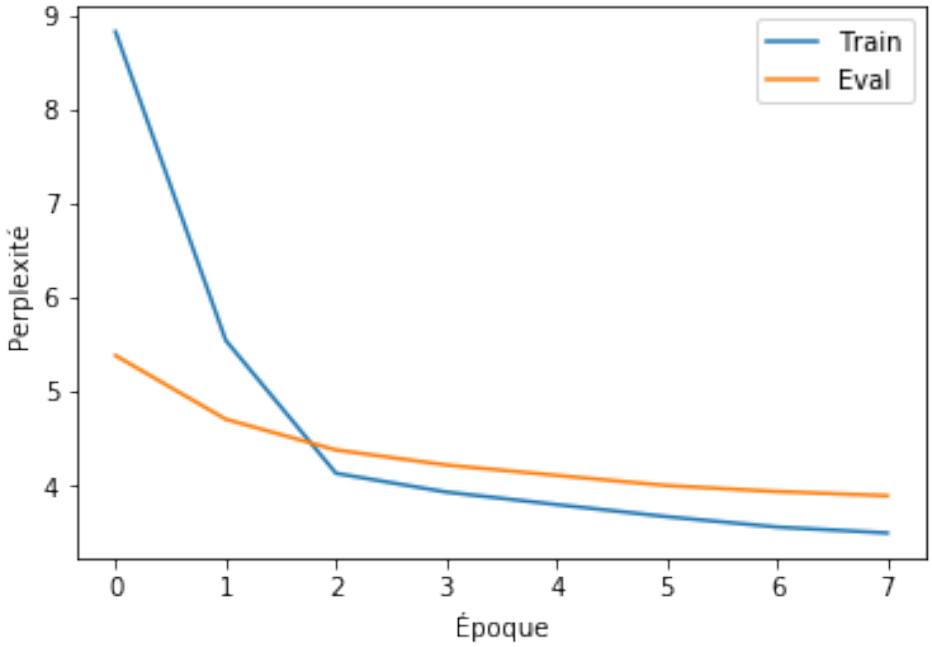

FIGURE 2 – La perplexité de FlauBERT$_{\text{BASE}}$ en s'entraînant sur les sous-ensembles *train* (80% du corpus) et *eval* (20%) du corpus SOS Médecins.

2.4 *Fine-tuning* : adaptation des modèles de langage à la tâche de classification

Pour l'étape de classification, c'est-à-dire l'adaptation des plongements entraînés sur la tâche de pré-entraînement à la tâche d'identification de profils cliniques qui nous intéresse, nous utilisons la méthode de classification intégrée aux plongements de style BERT : un jeton d'agrégation, étiqueté `[CLS]`, le premier jeton d'une séquence, qui est utilisé pour propager la perte à travers le modèle. Comme mentionné précédemment, pour pouvoir associer plusieurs catégories de maladie avec un document, il est nécessaire d'avoir un classifieur par catégorie. Au cas où un document dépasse la limite de longueur de séquences pour le modèle, il est divisé en plusieurs séquences.Dans la sortie finale pour évaluation, les documents sont associés aux chapitres MeSH correspondant à toutes les prédictions positives des classifieurs sur le document (ou au moins un de ses sous-séquences).

Pour construire un corpus d'entraînement pour des classificateurs avec notre corpus SOS Médecins, nous utilisons des codes d'identification diagnostic décrits dans la section 2.2.

En plus de la correspondance approximative entre les catégories de ce corpus et celles de la tâche DEFT, l'apport de ce corpus à l'apprentissage est limité par le déséquilibre entre les deux corpus, pas seulement en terme de taille mais en terme de prévalence des variables cibles. Dans la figure 3 on voit que les catégories de maladies les plus communes dans le corpus DEFT n'apparaissent pas dans le corpus de SOS. Cela veut dire que les distributions de probabilité modélisées par les classifieurs pourraient être assez différentes et un classifieur qui montre des bonnes performances sur la tâche de classification avec le corpus de SOS Médecins pourra bien avoir estimé une frontière de décision qui ne s'appliquera pas très bien à la tâche DEFT.

Après avoir supprimé les documents qui n'ont pas d'étiquette pertinente pour la tâche de classification, comme détaillé dans la section 2.2, il reste un corpus de taille 258 378 avec un longueur moyen de

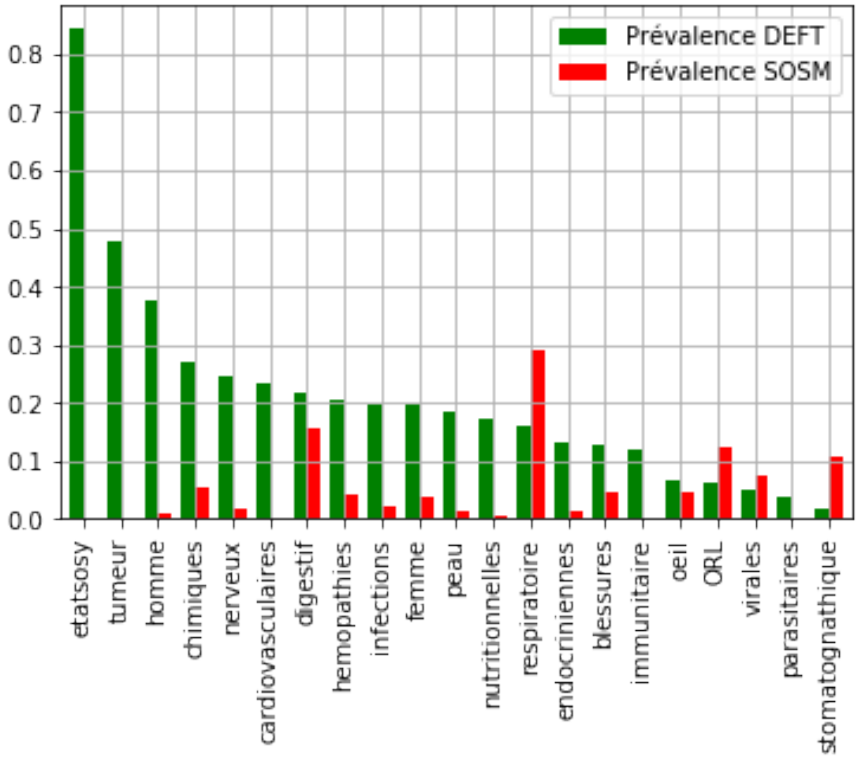

FIGURE 3 – Comparaison de la prévalence des variables cibles dans le corpus d'entraînement DEFT et le corpus de SOS Médecins étiqueté avec les chapitres MeSH pertinents.

41.77 et un longueur maximal de 390.

Les hyperparamètres utilisés pour l'entraînement des classificeurs sur le corpus SOS étaient les suivantes ;

— 4 époques,
— taux d'apprentissage : 2×10^{-5},
— longueur de séquence maximal : 256 tokens

Pour l'entraînement final avec les 167 documents donnés pour la tâche, les hyperparamètres étaient les mêmes sauf que le nombre d'époques était élevé jusqu'à 8.

3 Expériences

3.1 Expériences FlauBERT$_{\text{BASE}}$

Nous présentons l'évaluation de quatre différentes variantes des classifieurs sur la tâche d'identification de profil clinique ;

1. Le modèle FlauBERT$_{\text{BASE}}$ adapté directement à la tâche
2. Un modèle entraîné de manière non-supervisée sur le corpus SOS Médecins non-étiqueté comme décrit en section 2.3 ("Base + MLM" dans la figure 4)
3. FlauBERT$_{\text{BASE}}$ adapté à la classification des documents du corpus SOS Médecins étiqueté ("Base + clf").

	Précision		Rappel		F1	
	Train	**Eval**	**Train**	**Eval**	**Train**	**Eval**
Base	0.422	0.518	0.370	0.402	0.394	0.453
Base + MLM	0.475	0.528	0.457	0.435	0.466	0.477
Base + clf	0.452	0.542	0.477	0.451	0.464	0.492
Base + MLM + clf	0.487	**0.550**	0.520	**0.463**	0.503	**0.503**

FIGURE 4 – Les résultats d'entraînement et d'évaluation des différentes variantes de FlauBERT$_{\text{BASE}}$ sur la tâche. Les scores "Train" correspondent à la performance sur les données d'entraînement de 167 documents, et les scores "Eval" à la performance sur les 108 documents d'évaluation.

	Précision		Rappel		F1	
	Train	**Eval**	**Train**	**Eval**	**Train**	**Eval**
Base	0.366	0.390	0.353	0.444	0.359	0.416
Base + MLM	0.368	0.423	0.360	0.496	0.364	0.457
Base + MLM + clf	0.377	0.398	0.368	0.439	0.372	0.417

FIGURE 5 – Les résultats d'entraînement et d'évaluation des différentes variantes de FlauBERT$_{\text{SMALL}}$ sur la tâche.

4. La combinaison des deux approches précédentes ("Base + MLM + clf").

Les résultats sont présentés dans la figure 4. Comme prévu, il semble que l'addition de l'entraînement supplémentaire sur le corpus SOS Médecins augmente les mesures de performance. Il est intéressant de noter que les performances sur les données de test sont souvent meilleures que celles obtenues sur les données d'entraînement (principalement au niveau de la précision), ce qui suggère que le pouvoir prédictif des classifieurs vient des connaissances apprises pendant l'entraînement sur les corpus externes plutôt que pendant l'adaptation avec le corpus DEFT lui-même.

3.2 Soumission

À cause de certaines contraintes temporelles et de ressources de calculs, nous avons dû soumettre au défi des résultats des modèles entraînés avec FlauBERT$_{\text{SMALL}}$, une version de FlauBERT pas entièrement entraînés sur son corpus de base, donc les résultats pour la compétition n'était pas les meilleurs obtenus avec les modèles abordés dans cet article. Les résultats de ces expériences sont présentés en figure 5.

4 Discussion & Conclusion

Alors que les résultats n'étaient pas satisfaisants, nous suggérons que l'introduction d'un apprentissage supplémentaire sur l'ensemble de données SOS montre des améliorations encourageantes dans les performances des classifieurs qui pourraient être portées à un niveau acceptable avec plus de données d'entraînement de haute qualité, en particulier compte tenu du fait que nous n'avons pas introduit des corrections pour le déséquilibre dans les étiquettes d'entraînement.

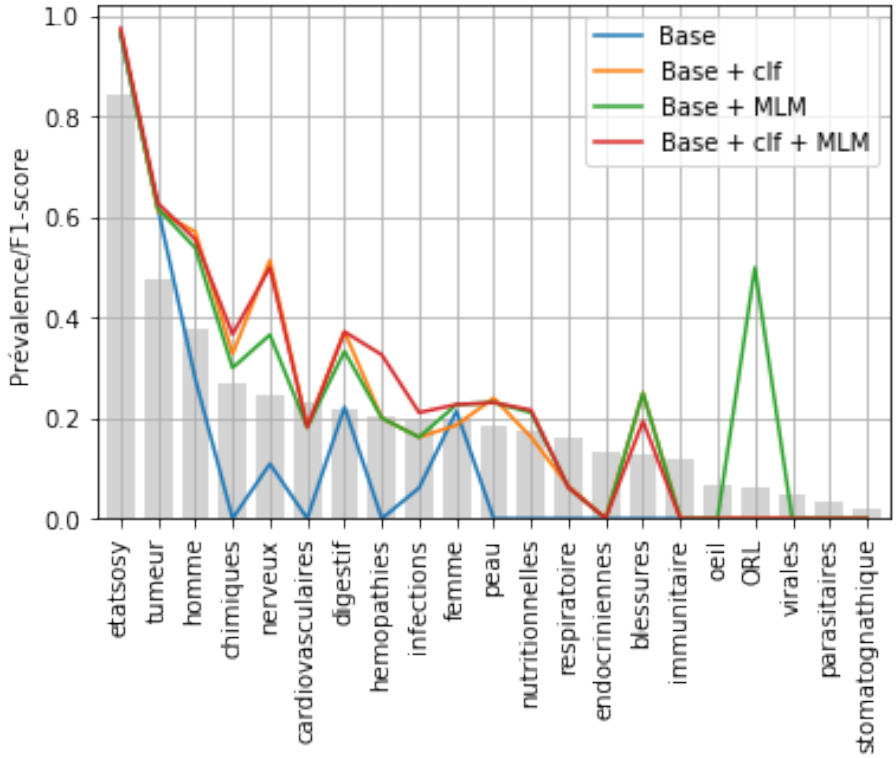

FIGURE 6 – Comparaison entre les prévalences des variables cibles dans le corpus d'entraînement DEFT et la performance des classifieurs entraînés sur le corpus d'évaluation (les barres grises représentent la prévalence de chaque catégorie de maladie).

4.1 Limitations du modélisation

Il y a plusieurs limites à l'efficacité des expériences réalisées dans ce travail qui pourraient être améliorées et constituer la base des futures expériences. Premièrement, en raison de contraintes de calcul et de temps, nous n'avons effectué aucun réglage important des hyperparamètres, en dehors du planificateur du taux d'apprentissage pour l'algorithme d'optimisation Adam dans l'entraînement non-supervisé. Il est généralement considéré comme une meilleure pratique de faire de la validation croisée en entraînant des classificateurs de l'apprentissage automatique, pour réduire l'impact de l'aléatoire dans la séparation des jeux d'apprentissage et de validation, mais ce n'était pas réalisable pour ces expériences, à cause du coût de calcul élevé de l'exécution. Il s'agit d'un amélioration possible pour les futurs expériences de retourner les entraînements avec des différents séparations afin de générer des estimations moins biaisées de la performance générale de ces techniques.

Dans la figure 6, on observe une corrélation entre le nombre d'exemplaires d'une variable cible dans le corpus d'entraînement et le score F1 d'un classificateur sur le corpus d'évaluation, ce qui suggère que la performance de notre système serait améliorée avec plus de données et d'exemples des différents types de maladies. Cette observation n'est surprenante, car il bien connu que les réseaux de neurones comme FlauBERT ont normalement besoin des énormes jeux de données pour sortir des bons résultats.

En plus, nous n'utilisons aucune connaissance externe explicitement encodée, comme les annotations supplémentaires ou des graphes de connaissance. Les graphes de connaissance externe sont souvent appliqués à des cas similaires, lorsque les jeux de données sont relativement petits (Costa *et al.*, 2021;

Chang *et al.*, 2020). Il s'agit d'une piste d'étude assez intéressante pour nous d'essayer de combiner des approches de ce type avec les techniques TALN discutées dans cet article.

Références

ALSENTZER E., MURPHY J. R., BOAG W., WENG W., JIN D., NEUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

BLINOV P., AVETISIAN M., KOKH V., UMERENKOV D. & TUZHILIN A. (2020). Predicting clinical diagnosis from patients' electronic health records using BERT-based neural networks. *arXiv :2007.07562*.

CHANG D., BALA˘ZEVI I., ALLEN C., CHAWLA D., BRANDT C. & TAYLOR R. A. (2020). Benchmark and best practices for biomedical knowledge graph embeddings. *arXiv :2006.13774*.

COSTA J. P., STOPAR L., REI L., MASSRI B. & GROBELNIK M. (2021). Exploring biomedical records through text mining-driven complex data visualisation. *medRxiv*. DOI : 10.1101/2021.03.27.21250248.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *arXiv :1810.04805v2*.

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.*

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). ClinicalBERT : Modeling clinical notes and predicting hospital readmission. *arXiv :1904.05342*.

KHATTAK F., JEBLEE S., POU-PROM C., ABDALLA M., MEANEY C. & RUDZICZ F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics X*.

KINGMA D. & BA J. (2015). Adam : A method for stochastic optimisation. *3rd International Conferenct on Learning Representations*.

LE H., VIAL L., FREJ J., SEGONNE V., COUAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for French. *arXiv :1912.05372v4*.

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*.

PENG Y., YAN S. & LU Z. (2019). Transfer learning in biomedical natural language processing : and evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv :1906.05474*.

VASWANI A., N.SHAZEER, PARMAR N., USZKOREIT J., JONES L., GOMEZ A., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. *Advance in Neural Information Processing Systems*.

YOON W., LEE J., KIM D., JEONG M. & KANG J. (2019). Pre-trained language model for biomedical question answering. *arXiv :1909.08229v1*.