Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Adaptation au domaine de modèles de langue à l'aide de réseaux à base de graphes

Merieme Bouhandi[1] Emmanuel Morin[1] Thierry Hamon[2]
(1) LS2N, UMR CNRS 6004, Nantes Université, Nantes, France
(2) LISN, Université Paris-Saclay & Université Sorbonne Paris Nord, France
{merieme.bouhandi, emmanuel.morin}@ls2n.fr
thierry.hamon@limsi.fr

RÉSUMÉ

Les modèles de langue prodonds encodent les propriétés linguistiques et sont utilisés comme entrée pour des modèles plus spécifiques. Utiliser leurs représentations de mots telles quelles pour des domaines peu dotés se révèle être moins efficace. De plus, ces modèles négligent souvent les informations globales sur le vocabulaire au profit d'une plus forte dépendance à *l'attention*. Nous considérons que ces informations influent sur les résultats des tâches en aval. Leur combinaison avec les représentations contextuelles est effectuée à l'aide de réseaux de neurones à base de graphes. Nous montrons que l'utilité de cette combinaison qui surpassent les performances de *baselines*.

ABSTRACT

Graph Neural Networks for Adapting General Domain Language Modèles Specialised Corpora

Language Modèles encode linguistic proprieties and are used as input for more specific Modèles. Using their word representations as-is for low-resource domains might be less efficient. Methods of adapting them exist ; but these Modèles often overlook global information due to their strong reliance on attention. We considers that global information can influence the downstream tasks results, and combination with contextual information is performed using graph neural networks. By outperforming baselines, we show that this architecture is profitable for a range of domain-specific tasks.

MOTS-CLÉS : modèles de langue, modèles neuronaux à base de graphes, plongements de mots, domaine spécialisé.

KEYWORDS: language Modèles, grape neural networks, word embeddings, specialised domains.

1 Introduction

L'analyse distributionnelle repose sur l'hypothèse harrissienne (Harris, 1954) selon laquelle le degré de similarité des contextes de deux mots définit leur proximité sémantique, en traitant le sens et les concepts comme des dimensions sous-jacentes du texte. Traditionnellement, les modèles de sémantique distributionnelle prennent deux formes : la première consiste à créer une matrice pondérée de mots cibles et de leurs contextes de distribution associés. La seconde, l'approche prédictive, consiste à utiliser des méthodes neuronales pour créer des représentations denses des mots : (Mikolov *et al.*, 2013; Pennington *et al.*, 2014; Bojanowski *et al.*, 2017). Les sorties de ces modèles sont utilisées, en aval, comme moyen de caractériser les unités sémantiques dans des tâches de traitement automatique des langues (TAL). Grâce aux récents progrès en TAL, de nouvelles architectures

telles que BERT (Devlin *et al.*, 2019), ont vu le jour. Ces modèles de langues contextuels occupent désormais une place centrale dans un large éventail de tâches classiques du TAL et sont construits à l'aide de gigantesques corpus de langue générale (Graff *et al.*, 2003; Zhu *et al.*, 2015).

La qualité des représentations issues de ces méthodes est souvent corrélée au volume de données disponibles. Dans le cas des domaines spécialisés, les corpus sont généralement de taille plus modeste, et ces méthodes se révèlent moins efficaces [1]. Les modèles utilisant des mécanismes d'attention, tels que BERT, capturent efficacement les informations contextuelles et les nuances locales au sein d'une phrase ainsi que bon nombre d'informations sémantiques et syntaxiques, mais leur capacité à capturer des informations globales sur le vocabulaire est plus limitée (Lu *et al.*, 2020). L'utilisation de l'auto-attention ou *self-attention* contribue à la construction de représentations plus riches, en considérant la relation d'un mot (ou d'un *"token"*) et de ses voisins dans la phrase, mais cela peut ne pas être suffisant : pour la plupart des tâches spécialisées, des informations supplémentaires sur la façon dont les mots et les termes sont liés de façon globale sont souvent essentielles pour obtenir de bons résultats sur des tâches en aval. Pertinent pour générer une représentation contextualisée, le contexte seul pourrait ne pas être suffisamment discriminant pour générer une bonne représentation. Des approches ont été récemment développées pour explorer et exploiter la façon dont les mots sont liés les uns aux autres dans le corpus. Les méthodes à base de graphes s'y prêtent bien : on peut citer les réseaux convolutifs à base de graphes (GCN) (Kipf & Welling, 2017) et ses variants, Text-GCN (Yao *et al.*, 2019) et VGCN (Lu *et al.*, 2020). À la manière de ce dernier, nous proposons de combiner les principales forces des deux modèles (GCN et BERT) dans la même architecture, limitant ainsi l'influence du contexte sur la représentation des mots rares ou du domaine en y ajoutant des informations plus conceptuelles.

2 Tâches et Ressources

TermEval 2020 Les méthodes existantes pour l'extraction terminologique ne répondent souvent pas aux attentes des spécialistes du domaine, et sont considérés comme pouvant être considérablement améliorées. Nous utilisons les données de TermEval 2020 [2], qui consiste en quatre corpus traitant de corruption, dressage, énergie éolienne et d'insuffisance cardiaque, en anglais, français et néerlandais.

I2B2 2010 and BC5CDR I2B2 (Uzuner *et al.*, 2011) porte sur l'extraction automatique de concepts médicaux (problèmes, tests et traitements) à partir de rapports cliniques. Cette tâche a été proposée par l'édition 2010 du I2B2/VA *Natural Language Processing Challenges for Clinical Records* [3].

BioCreative V CDR (BC5CDR) est une collection de 1,500 titres et résumés PubMed sélectionnés à partir du corpus CTD-Pfizer et a été utilisée dans une tâche d'extraction de relations entre produits chimiques et maladies. Ici, nous nous contentons d'extraire les entités.

MedSTS and BIOSSES Créé par (Soğancıoğlu *et al.*, 2017), BIOSSES comprend 100 paires de phrases en langue anglaise, tirées du jeu de données d'entraînement de la TAC (Text Analysis

1. Il faut noter que si cela n'est pas toujours valable pour tous les domaines, en particulier pour les domaines spécialisés en langue anglaise (on peut penser aux corpus biomédicaux assez volumineux tels que MEDLINE ou MIMIC3 (Johnson *et al.*, 2016)), elle est souvent vraie pour d'autres langues bien moins dotées (Eisenschlos *et al.*, 2019)
2. https ://termeval.ugent.be/acter-dataset/
3. https ://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Conference) Biomedical Summarization Track, contenant des articles du domaine biomédical.

Medical Semantic Textual Similarity (MedSTS) (Wang *et al.*, 2020) est une tâche de similarité sémantique entre phrases du domaine clinique. Le jeu de données est composé de 1642 paires de phrases, provenant de la tâche partagée de 2018 et de 2 dossiers médicaux électroniques, GE et Epic.

3 Méthodologie

Adaptation au domaine Les plongements de mots pré-entraînés sont un élément essentiel dans de nombreuses architectures TAL. Tirés de modèles de langue tout-venant, ils sont intégrées dans un modèle spécifique à la tâche, ce qui améliore souvent les résultats. L'adaptation de ces modèles prend généralement deux formes (Peters *et al.*, 2019) : l'extraction de descripteurs, où les poids du modèle sont utilisés tels quels en entrée d'un autre système, et le *"fine-tuning"*, où les poids du modèle continuent à être entraînés sur les nouvelles données pour une tâche spécifique.

Réseaux Convolutifs à base de Graphes Les GNN sont connus pour traiter des problèmes de représentations hiérarchisées : on peut les utiliser pour créer des graphes de vocabulaire et des embeddings conservant les relations hiérarchiques entre les mots dans un texte. La version convolutive de ce modèle, le GCN, permet d'effectuer des opérations de convolution sur les nœuds voisins dans le graphe, ce qui permet d'obtenir une représentation d'un mot qui intègre des informations de son voisinage, ce qui permet d'intégrer des informations sur le contexte global de mot. Puisque nous utilisons le vocabulaire pour construire le graphe, nous utilisons le VGCN (Lu *et al.*, 2020). Le VGCN prend en compte les informations plus globales sur le vocabulaire, mais échoue à capturer les informations locales, c'est pourquoi nous l'utilisons conjointement avec BERT. Le graphe de vocabulaire est construit en utilisant à la fois les *tokens* (Schuster & Nakajima, 2012) et les co-occurrences des mots dans les documents. Nous le combinons ensuite cette représentation au niveau de la couche d'*embedding* de BERT. Ces embeddings sont alors passés directement au premier encodeur à auto-attention (*self-attention*) de BERT. Pour chaque *token*, l'*embedding* du graphe et l'*embedding* du mot interagissent au travers du mécanisme d'auto-attention tout en adaptant le modèle pour la tâche.

Assez traditionnellement, notre GCN comporte deux couches (Kipf & Welling, 2017; Lu *et al.*, 2020). Avec un GCN à une couche, chaque nœud ne peut obtenir les informations que de ses voisins immédiats. En ajoutant une autre couche convolutive par-dessus, nous répétons cette "agrégation" du voisinage, mais cette fois, les voisins ont déjà des informations sur leurs voisins, qui leur viennent de la convolution précédente. Le nombre de couches est en réalité le nombre maximal de sauts que chaque nœud peut atteindre. Cependant, nous ne voulons généralement pas aller trop loin dans le graphe au risque de le lisser, d'en effacer des informations importantes, et de rendre la représentation générée moins significative, ce qui peut entraîner une baisse des performances (Kipf & Welling, 2017) dans la tâche en aval. Les nœuds du GCN sont des "entités" de la tâche, telles que des mots, des *"tokens"* ou des documents. Cette architecture exige donc que toutes les entités, c'est-à-dire celles de l'ensemble d'apprentissage, de validation et de test, soient présentes dans le graphe pendant l'entraînement.

Inspiré de (Lu *et al.*, 2020), notre modèle combine les plongements du GCN et les plongements de BERT et les transmets au premier encodeur de BERT. Cela permettra de conserver l'information

liée à l'ordre des mots dans la phrase et d'utiliser les informations locales et contextuelles de la représentation BERT, tout en y incorporant les informations plus globales obtenues par GCN. Ces dernières interagiront avec la représentation BERT au travers des 12 couches d'encodeurs du modèle profond. Nous testons également deux autres méthodes de combinaison : pour la première, au lieu d'intégrer le GCN directement au niveau des couches d'encodeurs de BERT, nous l'ajoutons simplement au plongement de BERT avant de le transmettre à l'encodeur. La seconde consiste à produire deux sorties, une du GCN et une du BERT, à les concaténer juste avant d'appliquer un RELU et à les transmettre à la couche de classification *"fully-connected"*.

Pour résumer, dans ce travail nous testons plusieurs modèles, avec quelques *baselines* et plusieurs combinaisons BERT et GCN :

— **GCN** : Un GCN à deux couches avec plongement BERT utilisés en tant que descripteurs des nœuds. Ce modèle n'utilise donc que des informations globales sur le vocabulaire pour réaliser les tâches de classification.

— **BERT** : BERT préentraîné pour la classification de *tokens*, avec une couche de classification *"fully-connected"* en sortie. Pour l'anglais, on utilise *bert-base-cased*. Pour le français, c'est Camembert (Martin *et al.*, 2019) qui est utilisé. Pour le néerlandais, on utilise *bert-base-multilingual-cased*. Pour Termeval, l'implémentation de nos modèles suit celle de (Hazem *et al.*, 2020).

— **BERT+GCN$_{\text{add}}$** ou **BERT avec plongements de graphe ajoutés** : Deux représentations sont générées ici : l'une avec le GCN et l'autre avec BERT. Elles sont ensuite combinées (additionnées) et passées à la couche d'encodeurs de BERT.

— **BERT+GCN$_{\text{vanilla}}$** ou **BERT avec plongements de graphe concaténés** : Deux représentations sont générées ici : une avec le GCN et l'autre avec BERT. Elles sont ensuite combinées (concaténées) et passées à la couche d'encodeurs de BERT.

— **BERT+GCN$_{\text{embedding}}$** ou **BERT avec plongements de graphe intégré** : Au lieu d'utiliser la couche *BERT Embedding* classique de l'architecture de (Devlin *et al.*, 2019), on y combine y ajoute une couche GCN. C'est ce nouvel embedding qui sera passé à la couche d'encodeurs de BERT.

4 Résultats

	I2B2			BC5CDR			MedSTS	BIOSSES
Modèle	P	R	F1	P	R	F1	Pearson	Pearson
GCN	71.4 ± 0.5	52.1 ± 0.2	63.7 ± 0.1	79.9 ± 0.9	77.2 ± 1.0	78.6 ± 0.9	70.29 ± 0.9	72.3 ± 0.8
BERT	87.4 ± 0.3	87.0 ± 0.1	87.2 ± 0.8	86.0 ± 0.7	85.0 ± 0.8	85.5 ± 0.1	83.52 ± 0.7	83.7 ± 1.2
+ GCN$_{\text{vanilla}}$	87.7 ± 1.2	86.5 ± 0.2	87.4 ± 0.2	86.3 ± 0.6	**86.1 ± 0.7**	86.2 ± 0.3	84.41 ± 0.9	84.8 ± 0.7
+ GCN$_{\text{add}}$	**89.7 ± 0.4**	86.0 ± 0.7	87.7 ± 0.1	87.7 ± 0.4	85.7 ± 0.3	86.3 ± 0.2	84.22 ± 0.3	85.0 ± 0.5
+ GCN$_{\text{embed}}$	89.0 ± 0.2	**88.8 ± 1.0**	**88.9 ± 0.2**	**87.9 ± 0.5**	85.7 ± 0.1	**86.7 ± 0.4**	**84.65 ± 0.3**	**86.1 ± 0.6**

TABLE 1 – Analyse (en % F1 mesure et coefficient de corrélation de Pearson) des sorties des différents modèles pour les tâches d'étiquetage de séquences et de similarité sémantique des phrases.

Dans l'ensemble, tous les modèles enrichis avec des informations globales tirées du GCN sont plus performants que les modèles de base, à savoir. Cela confirme nos intuitions et montre qu'il est bénéfique de fusionner les informations locales et globales, et que les représentations qui en résultent

	Anglais			Français			Néerlandais		
Modèle	P	R	F1	P	R	F1	P	R	F1
GCN	24,3 ± 15.3	15.8 ± 0.4	18.0 ± 0.5	19.0 ± 0.8	13.9 ± 0.4	15.7 ± 0.8	27.8 ± 0.1	16.0 ± 0.6	21.8 ± 0.3
BERT	36,1 ± 0,5	71.6 ± 1.0	48.8 ± 0.9	40.4 ± 0.7	65.2 ± 0.6	49.9 ± 0.2	32.2 ± 0.4	75.3 ± 0.2	45.0 ± 0.3
+ $GCN_{vanilla}$	38,6 ± 0,2	**69.7 ± 0.7**	**49.7 ± 0.6**	41.4 ± 0.3	65.0 ± 0.3	**50.4 ± 0.1**	33.5 ± 0.2	74.4 ± 0.8	46.2 ± 0.3
+ GCN_{add}	38,6 ± 0,3	69.3 ± 1.0	49.1 ± 0.1	**42.8 ± 0.9**	61.3 ± 0.7	49.8 ± 0.7	**34.2 ± 0.8**	**75.4 ± 1.0**	47.0 ± 1.0
+ GCN_{embed}	**40,0 ± 0,7**	68,7 ± 0,5	**49,7 ± 0,4**	42,2 ± 0,6	**62,8 ± 0,6**	50,3 ± 0,9	34,1 ± 0,1	72,2 ± 0,9	**47,6 ± 0,1**

TABLE 2 – Analyse (en % F1 mesure) des sorties des différents modèles pour TermEval 2020.

sont plus utiles pour les tâches en aval. Une tendance intéressante semble se dessiner, à la fois pour les tâches présentées table 1 et table 2 : l'augmentation du score global vient souvent d'une meilleure précision pour les modèles augmentés. Nous observons parfois une hausse simultanée de la précision et une diminution du rappel, indiquant une baisse du taux de faux positifs. Puisque ces tendances sont similaires pour presque toutes les tâches et configurations, nous effectuons une étude de cas sur l'une d'elles (i2b2) pour comprendre ce que l'ajout des plongements GCN apporte réellement à BERT.

3. Echocardiogram on **DATE[Nov 6 2007] , showed ejection fraction of 55% , mild mitral insufficiency , and 1+ tricuspid insufficiency with mild pulmonary hypertension .
DERMOPLAST TOPICAL TP Q12H PRN Pain DOCUSATE SODIUM 100 MG PO BID PRN Constipation IBUPROFEN 400-600 MG PO Q6H PRN Pain
The patient had headache that was relieved only with oxycodone . A CT scan of the head showed microvascular ischemic changes . A followup MRI which also showed similar changes . This was most likely due to her multiple myeloma with hyperviscosity .

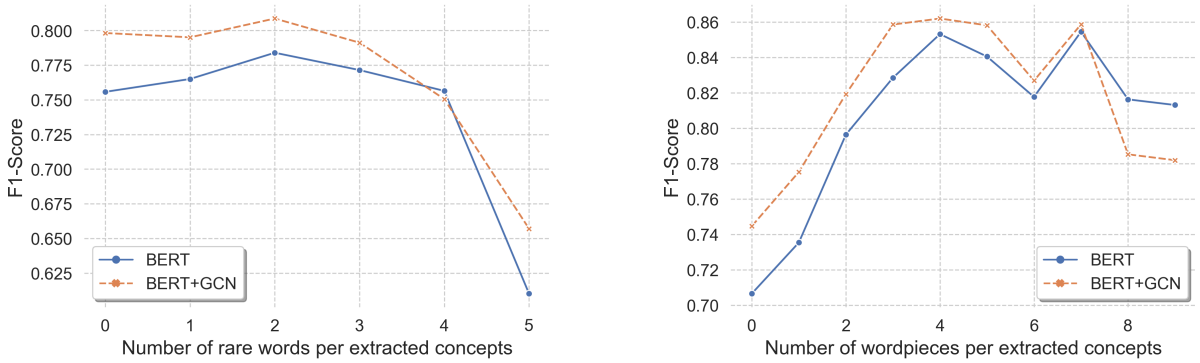
TABLE 3 – Exemples de concepts (Problème , Traitement , and Test) extraits du corpus i2b2 2010 (table from (Roberts, 2016))

5 Étude de cas : I2B2 2010

La table 3 montre des exemples de différents types de concepts issus du corpus i2b2 2010. Nous étudions d'abord la répartition des erreurs de classification pour évaluer la différence entre les sorties deux modèles. Nous analysons ensuite quelques faux négatifs et faux positifs du BERT classique et du BERT augmenté pour évaluer qualitativement les gains du modèle. Enfin, nous nous penchons sur le problème des mots rares et des *wordpieces*.

Type d'erreurs L'étude des résultats négatifs peut permettre de comprendre les distinctions entre nos systèmes. Nous rapportons ici les pourcentages de cinq types d'erreurs prises en compte lors de l'évaluation (Nejadgholi *et al.*, 2020). Pour chaque entité prédite e_p et notre ensemble de test E_T :

— **Type-1** (Faux positif) : e_p n'est pas présente dans E_T.
— **Type-2** (Faux négatif) : une entité de E_T n'est pas prédite.
— **Type-3** : e_p et une entité de E_T ont les mêmes empans, mais des étiquettes différentes.
— **Type-4** : e_p et une entité de E_T ont des empans qui se chevauchent et des étiquettes différentes.



(a) F1 mesure (en %) pour les concepts contenant un certain nombre de constituants rares.

(b) F1 mesure (en %) pour les concepts contenant un certain nombre de *wordpieces*.

FIGURE 1 – F1 mesure sur les constituants

— **Type-5** : e_p et une entité de E_T ont des empans qui se chevauchent et les mêmes étiquettes.

	Modèles	
Type d'erreur	BERT	BERT+GCN$_{\text{embedding}}$
Type-1	1,844 (24,31%)	1,540 (24,42%)
Type-2	1,771 (23,34%)	1,557 (24,69%)
Type-3	874 (11,52%)	729 (11,56%)
Type-4	394 (5,19%)	302 (4,78%)
Type-5	2,703 (35,63%)	2,177 (34,52%)

TABLE 4 – Analyse des erreurs pour la tâche i2b2 2010

Une baisse des cinq types d'erreurs peut être observée avec le BERT augmenté par rapport au BERT classique, mais avec une distribution des erreurs similaire dans l'ensemble (voir table 4) pour les deux modèles. Comme l'indique (Nejadgholi *et al.*, 2020), l'erreur de type 5, à savoir une entité prédite avec les bonnes étiquettes, mais un empan partiellement incorrect, est l'erreur la plus courante pour les modèles d'étiquetage de séquences. Un exemple serait d'avoir *"hypertension pulmonaire légère"* (problème) comme entité véritable et de n'extraire que *"hypertension pulmonaire"* (problème).

Mots rares et Wordpieces Les modèles de langue ont du mal à représenter les mots peu fréquents (Schick & Schütze, 2020). Ce problème est quasiment inévitable lorsque nous disposons de corpus d'entraînement de petite taille : l'exploitation des informations globales et locales permet alors d'obtenir des représentations plus riches (1a). En effet, pour les entités multi-mots contenant des constituants rares, les résultats sont souvent meilleurs pour BERT augmenté.

Par ailleurs, BERT décompose les mots en *wordpieces* : ainsi, *"adenocarcinoma"* sera décomposé en *"aden"*, *"oca"*, *"rc"*, *"ino"*, *"ma"* et *"carcinoma"* sera décomposé en *"car"*, *"cino"*, *"ma"*, faisant de *"ma"* le seul *wordpieces* qu'ils partagent, même si les deux mots sont sémantiquement très proches. Pour ce problème, des méthodes existent, notamment pour la génération de *wordpieces* adaptés aux textes cliniques (Nguyen *et al.*, 2019), mais leur utilisation signifie que nous devons ré-entraîner du modèle BERT. Pour tenter de contrer le problème des *wordpieces*, nous utilisons à la fois des mots complets et les *wordpieces* pour construire nos graphes de vocabulaire. Cela permet au modèle

d'associer les informations des *wordpieces* à des contextes plus riches. Nous pouvons voir (figure 1b) que pour les entités multi-mots contenant un certain nombre de *wordpieces*, la F1 mesure de BERT augmenté avec le GCN est plus élevé, suggérant que la fusion des informations sur les mots complets et les *wordpieces* dans le graphe est effectivement bénéfique. Pour plus de sept mots, la performance chute. Cela peut être imputé à des problèmes d'empans, comme les erreurs de type 5 (table 4) : plus l'entité extraite est longue, plus la probabilité d'avoir une correspondance partielle augmente.

6 Discussion

Globalement, les résultats que nous reportons suggèrent qu'il est effectivement bénéfique de fusionner les informations locales et globales. Cela se constate lors de la lecture du tableau 4, où le nombre d'erreurs diminue lorsque des informations globales sont ajoutées. Cette baisse des erreurs, en particulier de faux positifs, est visible dans le tableau 1, où une augmentation de la précision est notée dans tous les cas. Cependant, les scores globaux et le pourcentage élevé d'erreurs de *span* indiquent que des améliorations sont encore possibles. Une analyse plus approfondie peut également être menée sur les *wordpieces*. BERT décompose les mots *wordpieces*. Le problème pour les domaines spécialisés est qu'une méthode de tokenisation des *wordpieces* plus orientée vers le domaine est probablement plus appropriée. Par exemple, avec le tokeniseur de mots de BERT, *"adenocarcinoma"* sera décomposé en *"aden"*, *"oca"*, *"rc"*, *"ino"*, *"ma"* et, étonnamment, *"carcinoma"* sera décomposé en *"car"*, *"cino"*, *"ma"*, faisant de *"ma"* le seul *wordpieces* qu'ils partagent, même si les deux mots sont sémantiquement très proches. Certaines méthodes ont été développées, notamment pour les *wordpieces* adaptés aux textes cliniques (Nguyen *et al.*, 2019). Cependant, leur utilisation signifie que nous devons réentraîner l'ensemble du modèle BERT (comme pour ClinicalBERT, par exemple), ce qui va à l'encontre de la notion de l'adaptation ou de *transfer learning*. Nous comptons donc sur le GCN pour récupérer ces éléments d'information manquants et les réintégrer dans le modèle BERT. D'après les résultats de la figure 1b, pour les mots contenant des *wordpieces*, le F1-score est plus élevé avec le BERT augmenté qu'avec le BERT classique. Les mêmes tendances peuvent être observées pour les mots rares : même si tous les modèles sont moins performants sur ces mots rares, dans l'ensemble, les résultats sont plus élevés pour BERT avec des informations globales supplémentaires. Cependant, l'une des principales limites de ce travail — et des travaux sur les réseaux de neurones à base de graphes en général — est la nécessité d'utiliser toutes les entités, y compris celles de l'ensemble d'apprentissage, de l'ensemble de validation et de l'ensemble de test, pour construire le graphe de vocabulaire. Une autre chose à mentionner est que des travaux antérieurs (Nejadgholi *et al.*, 2020) ont souligné qu'un certain nombre de ces erreurs de type 5, comme vu dans le tableau 4 sont en fait des entités valides, et les catégoriser comme faux positifs pourrait en fait être un défaut des mesures d'évaluation existantes et spécifiques à la tâche, en particulier dans le cas de la reconnaissance d'entités nommées et de l'étiquetage de séquences.

Dans ce travail, nous avons voulu examiner s'il était possible d'améliorer les modèles tout-venant pour des langues et des domaines pour lesquels il n'existe pas de BioBERT et de Clinical BERT. Par exemple, il n'existe pas de WindBERT ou de PoliticalBERT pour la tâche TermEval, et donc cette étape d'adaptation au domaine a du sens et est même nécessaire, si l'on souhaite utiliser BERT. L'utilisation d'un réseau neuronal à base de graphes construit sur le vocabulaire du corpus exclusivement, sans ajouter de ressources provenant de terminologies ou de graphes de connaissances externes, suit la même logique et découle de notre décision de trouver un moyen d'améliorer le modèle de manière intrinsèque sans utiliser de ressources externes. Comme nous l'avons dit précédemment,

même si de nombreuses ressources existent pour la langue anglaise, il est toujours important d'explorer divers moyens d'adapter les modèles existants à nos tâches spécialisées en aval où le volume de données est souvent insuffisant. Le point le plus important à retenir de ce travail est le fait d'obtenir ces résultats sans aucun corpus supplémentaire, de façon tout à fait endogène. Cela en dit long sur la puissance de ces architectures neuronales et sur la façon dont les représentations contextuelles et non contextuelles peuvent être améliorées en modifiant simplement l'architecture des modèles et en réarrangeant astucieusement les données d'entraînement.

7 Conclusion

La qualité des représentations vectorielles de mots obtenues à partir des modèles de langue est souvent corrélée au volume de données disponibles pour l'entraînement de ces derniers. Les tâches liées aux domaines spécialisés s'en trouvent désavantagées : l'utilisation de ces méthodes implique une phase d'adaptation gourmande en ressources textuelles alors que les corpus spécialisés sont généralement de taille modeste. Dans ce travail, nous tentons de comprendre comment ces méthodes peuvent être adaptées en utilisant des descripteurs supplémentaires pour une série de tâches spécialisées. Les résultats obtenus avec nos méthodes surpassent les *baselines* sur nos tâches sans jeu de données supplémentaires, montrant qu'il est bénéfique de fusionner les informations locales (BERT) et globales (GCN). Nos travaux futurs porteront sur une analyse plus approfondie des couches d'attention après l'ajout d'informations globales, ainsi que sur la prise en compte des relations plus complexes pour construire le GCN, avec notamment l'exploitation de graphes orientés pour coder les relations de dépendances, la synonymie, l'hypo- et l'hyperonymie.

Références

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

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 (NAACL-HLT'18)*, p. 4171–4186, Minneapolis, MN, USA. DOI : 10.18653/v1/N19-1423.

EISENSCHLOS J., RUDER S., CZAPLA P., KADRAS M., GUGGER S. & HOWARD J. (2019). MultiFiT : Efficient multi-lingual language model fine-tuning. 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. 5702–5707, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1572.

GRAFF D., KONG J., CHEN K. & MAEDA K. (2003). English gigaword. *Linguistic Data Consortium, Philadelphia*, **4**(1), 34.

HARRIS Z. S. (1954). Distributional structure. *WORD*, **10**(2-3), 146–162. DOI : 10.1080/00437956.1954.11659520.

HAZEM A., BOUHANDI M., BOUDIN F. & DAILLE B. (2020). TermEval 2020 : TALN-LS2N system for automatic term extraction. In *Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM'20)*, p. 95–100, Marseille, France.

JOHNSON A. E. W., POLLARD T. J., SHEN L., WEI H. LEHMAN L., FENG M., GHASSEMI M. M., MOODY B., SZOLOVITS P., CELI L. A. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, **3**.

KIPF T. N. & WELLING M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

LU Z., DU P. & NIE J.-Y. (2020). Vgcn-bert : Augmenting bert with graph embedding for text classification. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 de *Lecture Notes in Computer Science*, p. 369–382 : Springer.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2019). CamemBERT : a Tasty French Language Model. *arXiv e-prints*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

NEJADGHOLI I., FRASER K. C. & DE BRUIJN B. (2020). Extensive error analysis and a learning-based evaluation of medical entity recognition systems to approximate user experience. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, p. 177–186, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.bionlp-1.19.

NGUYEN V., KARIMI S. & XING Z. (2019). Investigating the effect of lexical segmentation in transformer-based models on medical datasets. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, p. 165–171, Sydney, Australia : Australasian Language Technology Association.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, p. 1532–1543, Doha, Qatar.

PETERS M. E., RUDER S. & SMITH N. A. (2019). To Tune or Not to Tune ? Adapting Pretrained Representations to Diverse Tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP'19)*, volume abs/1903.05987, p. 7–14, Florence, Italy.

ROBERTS K. (2016). Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP'16)*, p. 54–63, Osaka, Japan.

SCHICK T. & SCHÜTZE H. (2020). Rare Words : A Major Problem for Contextualized Embeddings And How to Fix it by Attentive Mimicking. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, p. 8766–8774 : AAAI Press.

SCHUSTER M. & NAKAJIMA K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5149–5152. DOI : 10.1109/ICASSP.2012.6289079.

SOĞANCIOĞLU G., ÖZTÜRK H. & ÖZGÜR A. (2017). BIOSSES : a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, **33**(14), i49–i58. DOI : 10.1093/bioinformatics/btx238.

UZUNER O., SOUTH B., SHEN S. & DUVALL S. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, **18**, 552–6. DOI : 10.1136/amiajnl-2011-000203.

WANG Y., FU S., SHEN F., HENRY S., UZUNER O. & LIU H. (2020). The 2019 n2c2/ohnlp track on clinical semantic textual similarity : Overview. *JMIR Med Inform*, **8**(11), e23375. DOI : 10.2196/23375.

YAO L., MAO C. & LUO Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 7370–7377. DOI : 10.1609/aaai.v33i01.33017370.

ZHU Y., KIROS R., ZEMEL R., SALAKHUTDINOV R., URTASUN R., TORRALBA A. & FIDLER S. (2015). Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, p. 19–27, USA : IEEE Computer Society. DOI : 10.1109/ICCV.2015.11.