Intégration de connaissances structurées par synthèse de texte spécialisé

Guilhem Piat, Ellington Kirby, Julien Tourille, Nasredine Semmar, Alexandre Allauzen, Hassane Essafi

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Intégration de connaissances structurées par synthèse de texte spécialisé

Guilhem Piat[1,2] Ellington Kirby[1] Julien Tourille[2]
Nasredine Semmar[2] Alexandre Allauzen[1] Hassane Essafi[2]
(1) Université Paris Dauphine, F-75775, Paris Cedex 16, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{prenom}.{nom}@dauphine.psl.eu, {prenom}.{nom}@cea.fr

RÉSUMÉ

Les modèles de langue de type Transformer peinent à incorporer les modifications ayant pour but d'intégrer des formats de données structurés non-textuels tels que les graphes de connaissances. Les exemples où cette intégration est faite avec succès requièrent généralement que le problème de désambiguïsation d'entités nommées soit résolu en amont, ou bien l'ajout d'une quantité importante de texte d'entraînement, généralement annotée. Ces contraintes rendent l'exploitation de connaissances structurées comme source de données difficile et parfois même contre-productive. Nous cherchons à adapter un modèle de langage au domaine biomédical en l'entraînant sur du texte de synthèse issu d'un graphe de connaissances, de manière à exploiter ces informations dans le cadre d'une modalité maîtrisée par le modèle de langage.

ABSTRACT

Knowledge Integration by In-domain Text Generation

Transformer-based language models have trouble integrating modifications whose purpose is to incorporate knowledge from structured, non-textual data such as knowledge graphs. Instances where this integration is successful generally require the problem of Entity Linking to be solved upstream, or the addition of a significant amount of (generally annotated) text to the training set. These constraints often make leveraging structured data difficult and/or counterproductive. We seek to adapt a language model to the biomedical domain through training on synthetic text derived from a knowledge graph, such that the information therein can be effectively leveraged in a format which the model can handle natively.

MOTS-CLÉS : Intégration de connaissances ; Génération de texte ; Adaptation au domaine ; Modèle de langage biomédical.

KEYWORDS: Knowledge integration ; Text generation ; Domain adaptation ; Biomedical language model.

1 Introduction et travaux connexes

En raison du vocabulaire spécialisé et de la spécificité des concepts traités dans certains domaines de spécialité, les performances des modèles de langage pré-entraînés sur du texte général, tels que BERT (Devlin *et al.*, 2019), ont tendance à souffrir dans ces domaines. Diverses approches de spécialisation existent, la plus notable et évidente d'entre elles étant la spécialisation par pré-

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 275–284.

entraînement sur du texte issu du domaine ciblé. La performance de cette approche a été étudiée par Gururangan *et al.* (2020) et El Boukkouri (2021), et de nombreux modèles ont été couronnés de succès en suivant cette méthode, comme par exemple LegalBERT (Chalkidis *et al.*, 2020) dans le domaine légal et PatentBERT (Lee & Hsiang, 2020) dans le domaine des brevets. Dans le domaine biomédical, sur lequel nous décidons de nous focaliser (bien que les concepts dont nous traitons soient *a priori* applicables à tout domaine de spécialité), une variété de modèles appliquent la même méthode générale, comme BioBERT (Lee *et al.*, 2019), BlueBERT (Peng *et al.*, 2019) ou PubMedBERT (Gu *et al.*, 2021). Cette approche n'est cependant pas sans défauts. En particulier, elle est susceptible à l'oubli catastrophique (Xu *et al.*, 2020), et requiert une quantité importante de texte issu du domaine, ce qui n'est pas une option disponible dans tous les domaines de spécialité. De surcroît, comme observé par Zipf (1935), la distribution des mots dans langage naturel est fortement biaisée, ce qui a pour conséquence qu'accroître le nombre de concepts couverts par le corpus nécessite une augmentation exponentielle de la quantité de texte.

Une classe concurrente d'approches consiste à intégrer des informations issues d'une base de connaissance spécialisée. Diverses stratégies ont été mises en œuvre, comme l'utilisation du mécanisme d'attention des modèles *Transformer* pour combiner les informations issues des mots avec celles issues des entités (par ERNIE (Zhang *et al.*, 2019), KnowBert (Peters *et al.*, 2019), K-Adapter (Wang *et al.*, 2021a) et DRAGON (Yasunaga *et al.*, 2022) en particulier), l'alignement de plongements de mots et de plongements d'entités (KEPLER (Wang *et al.*, 2021b), CODER (Yuan *et al.*, 2022)), ou encore le changement de la fonction de coût en entraînement de manière à modéliser explicitement la synonymie des termes spécialisés (UmlsBERT (Michalopoulos *et al.*, 2021)).

Bien que ces approches puissent donner des résultats remarquables, leur champ d'applicabilité est limité par le fait qu'elles ne permettent pas (ou peu) de réduction de la quantité de texte nécessaire à la spécialisation, et souvent même requièrent du texte aux entités annotées ; or ces ressources ne sont pas disponibles pour tous les domaines. De plus, la majorité des approches requièrent, en entraînement comme, généralement, en déploiement, que les entités de la base de connaissance soient identifiées dans le texte pour fonctionner. Or, ce problème n'est pas résolu pour tous les domaines, et les meilleurs outils actuels sont gourmands en temps de calcul, limitant l'échelle à laquelle ils peuvent être déployées. A l'heure d'écriture, l'état de l'art en désambiguïsation d'entités nommées biomédicales (Bhowmik *et al.*, 2021) atteint une F-mesure de $0,564$ sur le corpus MedMentions (Mohan & Li, 2019).

Nous cherchons donc une méthode d'intégration de connaissances telle qu'aucun texte du domaine cible et aucune désambiguïsation — en entraînement comme en inférence — ne soient nécessaires. La combinaison de modalités différentes débouchant invariablement sur un besoin de données d'entraînement supplémentaires, nous cherchons à exploiter notre base de connaissance sous forme textuelle. Nous proposons une procédure permettant de générer un corpus contenant une phrase par paire de concepts liés par une relation dans une base de connaissance. Il est attendu qu'un corpus produit ainsi soit plus dense en informations factuelles liées au domaine que du texte naturel, et couvrira une plus grande variété de concepts pour le même volume de texte, la fréquence d'apparition des concepts étant dictée par la topologie du graphe de connaissances et non l'usage du langage naturel.

2 Méthode

2.1 Graphe de connaissances

Nous nous plaçons dans le cas où les nœuds de notre graphe de connaissances représentent des concepts et les arêtes des relations. Nous pouvons donc extraire des triplets (c_i, r, c_j) tels qu'il existe un lien sémantique entre les concepts c_i et c_j qui peut être formulé en langage naturel. Spécifiquement, nous cherchons à associer des groupes nominaux N_i à c_i et N_j à c_j, ainsi qu'un syntagme verbal V_r à r de sorte que la phrase $N_i V_r N_j$ soit une phrase grammaticalement correcte qui capture le sens de la relation dans le graphe. La base de connaissance peut ensuite être représentée par un ensemble de phrases factuelles, exploitable comme corpus d'apprentissage par un modèle de langue.

Nous utilisons le graphe de connaissances UMLS, dans sa version 2022AB, avec toutes et seulement les ressources anglophones. Il existe dans notre version de la base de connaissance 1008 types de relations et environ $4,6$ millions d'entités distinctes pour un total d'environ $39,7$ millions de triplets. Chaque concept dispose d'un identifiant unique c et d'un ou plusieurs « *noms* » $N_c^1, \ldots, N_c^{k_c}$, groupes nominaux qui correspondent à des formulations récurrentes du concept en langage naturel, et dont un est considéré *préféré*. Chaque type de relation est représenté par une courte chaîne de caractères r proche du langage naturel, décrivant grossièrement la relation. Le graphe UMLS est représenté sous forme de base de données relationnelle, avec une table contenant les triplets d'intérêt, représentés de la manière suivante : c_i, r, c_j.

2.2 Génération de texte

Bien qu'il y ait un bénéfice potentiel évident à inclure, le cas échéant, plusieurs formulations pour chaque entité, plusieurs facteurs nous ont menés à conserver seulement le *nom préféré* de chaque concept :

- le niveau de redondance entre les noms est élevé, contenant principalement des doublons, des variations de casse et des variations d'accord au pluriel.
  ex. : Le concept pour l'ADN a 88 noms, dont 20 occurrences de *DNA*, 15 variations de casse et d'accord en nombre sur *Deoxyribonucleic Acid*, 4 variations de casse et d'accord sur *DNA molecule*, et 10 combinaisons des termes susmentionnés avec des ordres et ponctuations différents.
- Chaque triplet étant constitué de deux concepts, le temps de calcul augmente quadratiquement avec le nombre total de noms lors de la résolution des triplets sous la forme (c_j, r, c_i) à la forme $(N_{c_i}^{1,\ldots,k_{c_i}}, r, N_{c_j}^{1,\ldots,k_{c_j}})$.
- Inclure les noms non-préférés réduit l'homogénéité grammaticale des groupes nominaux (par ex., la grande majorité des noms préférés sont au singulier), complexifiant la tâche de générer des phrases grammaticalement correctes.
- Multiplier les occurrences des concepts associés à de nombreux noms induit un biais dans l'apprentissage qui n'est pas forcément désirable.

Étant donné le nombre important de triplets, la résolution des noms n'est pas traitable sur l'ensemble de la base de connaissances. De plus, tous les types de relation ne sont pas utiles ou documentés. Nous effectuons donc une sélection sur les types de relations. Nous commençons par éliminer les 925 relations les plus rares, qui constituent environ 15% des triplets. Ensuite, nous éliminons une relation de chaque paire de relations symétriques, c'est-à-dire que lorsqu'il existe deux relations r_1

et r_2 telles que (c_i, r_1, c_j) représente la même information que (c_j, r_2, c_i), nous ne conservons que la relation r_1. Environ 95% des relations font partie d'une paire symétrique, ce qui nous laisse 43 relations. Nous avons ensuite sélectionné les 28 relations qui avaient un sens apparent ou documenté et qui n'impliquaient pas principalement des concepts aux noms complexes comme des molécules ou protéines.

Une fois nos triplets (N_{c_i}, r, N_{c_j}) extraits de la base de connaissance, nous avons formulé une phrase-type par relation dans laquelle nous avons automatiquement inséré les noms d'entités correspondant aux triplets. Un échantillon des séquences ainsi générées depuis la base de connaissance UMLS se trouve en annexe A. Le corpus résultant est constitué d'environ 6 millions de phrases simples, ou 100 millions de mots. Nous appelons *CSGU* ce Corpus Synthétique Généré depuis UMLS.

2.3 Modèles de langue

Nous entraînons le modèle BERT (Devlin *et al.*, 2019) pré-entraîné, spécifiquement dans sa version $\text{BERT}_{\text{BASE}}$, sur trois corpus différents [1]. Le premier est CSGU, et nous appelons le modèle résultant $\text{BERT}_{\text{CSGU}}$. De manière à évaluer l'efficacité de notre corpus synthétique vis-à-vis d'un corpus naturel, nous avons assemblé un second corpus de texte biomédical que nous appelons PMC, de volume similaire (environ 94 millions de mots) constitué d'une collection d'articles scientifiques en accès libre recueillis depuis PubMed Central. Nous appelons le modèle entraîné sur ce corpus BERT_{PMC}. De manière à étudier l'effet du texte de synthèse comme option d'augmentation de données, nous entraînons un modèle sur un corpus hybride constitué de la concaténation du corpus de synthèse et du corpus naturel. Cependant, l'ajout de données d'entraînement se faisant rarement au détriment du processus d'apprentissage, nous entraînons également un modèle sur la moitié du corpus hybride, de manière à ce que la quantité de données d'entraînement soit comparable aux autres modèles. Nous appelons ces modèles $\text{BERT}^{100\%}_{\text{Hybr}}$ et $\text{BERT}^{50\%}_{\text{Hybr}}$ respectivement. Les informations sur les hyperparamètres des modèles sont disponibles en annexe B.

Nous effectuons une batterie de tests sur chacun des modèles. Nous détaillons ces tests dans la section 3 et présentons les résultats sur les tâches biomédicales et générales dans les Tables 1 et 2 respectivement [2].

3 Résultats expérimentaux

NB : Les tâches marquées par un obèle (†) sont altérées vis-à-vis du standard et les résultats ne sont pas nécessairement comparables à ceux de la littérature.

3.1 Complétion de phrases biomédicales à trous (MLM)

Le but de l'approche par intégration de connaissances étant de permettre au modèle de mieux modéliser les co-occurrences de concepts biomédicaux, nous cherchons à évaluer la capacité de

1. Le code pour l'entraînement des modèles et les informations nécessaires à la création des corpus est disponible sur notre dépôt GitHub.

2. En supplément des résultats présentés ici, nous avons abordé la tâche de reconnaissance d'entités nommées i2b2 2010/n2c2 (Uzuner *et al.*, 2011), mais ne rapportons pas les résultats car les différences de performance n'étaient pas statistiquement significatives.

notre modèle à prédire les concepts masqués dans des contextes de phrases biomédicales. Les textes biomédicaux aux concepts annotés étant cependant rares, nous évaluons nos modèles sur la tâche de modélisation de langue par masquage de mots. Nous appliquons cette tâche à un corpus biomédical d'environ $12,6$ millions de mots recueilli de la même manière que le corpus d'entraînement du modèle BERT_{PMC}. Puisque tous les termes masqués ne seront pas des termes biomédicaux, une bonne performance sur cette tâche est indicative à la fois d'une bonne maîtrise du langage général et de la terminologie biomédicale.

Pour évaluer les performances de nos modèles sur cette tâche, nous utilisons un critère apparenté d'un point de vue théorique à la *perplexité*, définie dans les modèles de langue autorégressifs comme l'exponentielle de la moyenne des log-vraisemblances de la séquence, et équivalente à l'exponentielle de l'entropie croisée entre les données et les prédictions :

$$\exp\left(-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y_i}\cdot\log(\hat{\mathbf{y}}_\mathbf{i})\right) \tag{1}$$

où $\mathbf{y_i}$ et $\hat{\mathbf{y}}_\mathbf{i}$ sont respectivement le label (en encodage *one-hot*) et la distribution de probabilité prédite pour le jeton i, avec N le nombre total de jetons dans la séquence.

BERT n'étant pas un modèle autorégressif, la perplexité n'est pas strictement définie. D'un point de vue pratique cependant, l'équation (1) peut être calculée, et nous donne une mesure de performance cohérente. Par souci de simplicité, nous désignerons donc cette mesure « perplexité », ou « *ppl.* ». Une perplexité *basse* indique une meilleure performance.

Nos résultats rapportés dans la table 1 indiquent que le texte synthétique confère à $\text{BERT}_{\text{CSGU}}$ une capacité prédictive accrue par rapport à $\text{BERT}_{\text{BASE}}$. Il semblerait cependant que le manque d'exposition au langage naturel dans cette phase d'entraînement lui porte préjudice puisque BERT_{PMC} dépasse $\text{BERT}_{\text{CSGU}}$. Les modèles $\text{BERT}_{\text{Hybr}}$, cependant, dépassent le niveau de performance de BERT_{PMC}, indiquant que les défaillances de CSGU sont facilement atténuées par l'ajout de texte naturel dans le corpus.

3.2 Extraction de relations biomédicales (ChemProt†)

La tâche ChemProt (Krallinger *et al.*, 2017) consiste à prédire, étant donné une séquence biomédicale contenant deux entités E_1 et E_2, la nature de leur relation, parmi six options. Nos résultats révèlent que le corpus CSGU contient des informations précieuses, mais les connaissances acquises au cours du processus d'apprentissage sont fragiles et l'inclusion de texte naturel peut être contreproductive dans leur assimilation.

NB : Plusieurs versions du corpus ChemProt existent avec des types de pré-traitement différents qui peuvent influencer les performances des modèles. Tous nos modèles sont entraînés sur la même version du corpus et sont donc comparables entre eux.

3.3 Questions-réponses biomédicales (PubmedQA)

Le corpus PubmedQA (Jin *et al.*, 2019) est associé à plusieurs tâches de questions-réponses. Chaque instance contient une question sous forme de problématique issue d'un article scientifique, plusieurs

séquences dites « de contexte » apportant des éléments de réponse, et une phrase dite de « réponse longue » contenant la conclusion nuancée de l'article. L'objectif est de classer les questions selon la nature affirmative, négative, ou indécise de leur conclusion.

Nous abordons la tâche dans sa version la plus simple, à savoir la version dite « sans raisonnement ». Dans ce cadre, notre processus d'entraînement inclut la question et la réponse longue, mais ignore les séquences de contexte.

Nous constatons d'après nos résultats expérimentaux (Table 1) que $\text{BERT}_{\text{BASE}}$ a des performances bien plus faibles que les modèles spécialisés, ce qui est attendu puisqu'il n'est pas familiarisé avec le vocabulaire biomédical. Les performances supérieures de $\text{BERT}_{\text{CSGU}}$ laissent penser que la maîtrise des relations entre concepts biomédicaux et l'exposition à une grande variété de concepts sont particulièrement utiles pour interpréter des problématiques et conclusions scientifiques, et la dominance de $\text{BERT}^{100\%}_{\text{Hybr}}$ indique que les connaissances apportées par les corpus CSGU et PMC sont complémentaires dans le contexte de cette tâche.

Model	MLM (ppl.)	ChemProt (F_1)	PubmedQA (F_1)
$\text{BERT}_{\text{BASE}}$	16,64	88,91	70,20
BERT_{PMC}	13,66	88,91	74,83
$\text{BERT}_{\text{CSGU}}$	14,67	<u>89,87</u>	75,67
$\text{BERT}^{50\%}_{\text{Hybr}}$	11,38	88,88	72,50
$\text{BERT}^{100\%}_{\text{Hybr}}$	<u>10,37</u>	88,89	<u>78,00</u>

TABLE 1 – Tableau comparatif des résultats des modèles incorporant du texte synthétique dans leur corpus d'entraînement vis-à-vis des modèles entraînés exclusivement sur du langage naturel sur les tâches biomédicales de complétion de phrases à trous (MLM), d'extraction de relations (ChemProt), et de questions-réponses (PubMedQA). Résultats moyens sur 4 expériences. Le résultat souligné est le meilleur pour chaque tâche.

3.4 Tâches non biomédicales (CoLA†, SNLI)

Nous évaluons également nos modèles sur des tâches non biomédicales de manière à évaluer la baisse de performances dans le domaine général encourue par les modèles spécialisés.

La tâche d'acceptabilité linguistique CoLA consiste à classer différentes séquences selon leur qualité grammaticale comme étant « acceptables » ou non. Le critère d'évaluation pour cette tâche est le coefficient de corrélation de Matthews. Les labels de la partition de test n'étant pas publics, et en raison de diverses restrictions de soumission, nous avons re-partitionné le jeu de données. Nous avons utilisé la partition de validation comme partition de test, et de manière à rendre ce partitionnement reproductible, nous avons utilisé les 500 dernières instances de la partition d'entraînement comme partition de validation.

La tâche d'inférence linguistique SNLI est une tâche de classification de paires de séquences selon l'existence d'une relation d'*implication*, de *contradiction*, ou l'*absence* d'une telle relation entre elles.

Nos résultats dans la Table 2 indiquent que l'apprentissage sur le texte de synthèse porte préjudice à la capacité du modèle à évaluer la qualité grammaticale d'une séquence, sans pour autant avoir une incidence majeure sur sa capacité à détecter les relations d'implication. L'affaiblissement en grammaire peut cependant être compensé en associant le texte de synthèse à du langage naturel.

Modèle	CoLA (Corr. M.)	SNLI (F_1)
$\text{BERT}_{\text{BASE}}$	63,17	<u>90,58</u>
BERT_{PMC}	<u>64,11</u>	90,38
$\text{BERT}_{\text{CSGU}}$	61,89	90,44
$\text{BERT}^{50\%}_{\text{Hybr}}$	61,62	90,28
$\text{BERT}^{100\%}_{\text{Hybr}}$	63,86	90,20

TABLE 2 – Tableau comparatif des résultats des modèles incorporant du texte synthétique dans leur corpus d'entraînement vis-à-vis des modèles entraînés exclusivement sur du langage naturel sur les tâches d'acceptabilité linguistique (CoLA), et d'inférence (SNLI). Résultats moyens sur 4 expériences. Le résultat souligné est le meilleur pour chaque tâche.

4 Conclusions et futurs travaux

Nous proposons une procédure d'intégration de connaissances pour l'adaptation des modèles de langage aux domaines de spécialité simple à mettre en œuvre, et qui ne dépend ni de texte issu du domaine, ni d'outils d'annotation d'entités, ni d'une architecture de modèle spécialisée. Nous démontrons que, malgré la qualité dégradée du texte généré par rapport à du texte naturel, il peut être exploité par un modèle de langue pour l'adaptation au domaine avec, pour une quantité de texte fixe, plus de succès que du texte naturel. Enfin, les faiblesses exhibées par les modèles entraînés sur du texte synthétique peuvent être minimisées par l'incorporation, dans le corpus de spécialisation, de texte issu du domaine lorsqu'il est disponible.

Outre l'application de cette méthode à d'autres bases de connaissance comme YAGO (Suchanek *et al.*, 2007) ou WorldKG (Dsouza *et al.*, 2021), nous aimerions à l'avenir intégrer à cette méthode un post-traitement intelligent capable d'identifier les séquences grammaticalement incorrectes, excessivement complexes ou autrement problématiques, et de les supprimer ou les corriger. Par ailleurs, le manque de variété linguistique constitue une faiblesse importante de notre approche. S'il serait sans doute difficile d'automatiser des variations sur la formulation des relations, une amélioration envisageable serait d'intégrer des informations concernant la variété des formulations de concepts, soit en sélectionnant aléatoirement, étant donné un concept c, une formulation parmi $N_c^1, \ldots, N_c^{K_c}$, soit en ajoutant au corpus de synthèse des séquences dédiées à expliciter les synonymes (par exemple : « "N_c^2" is another name for "N_c^1". »)

Enfin, cette méthode d'intégration de connaissances étant applicable à tout modèle de langue, la combiner avec d'autres méthodes comme KnowBert, KEPLER ou DRAGON pourrait être une manière peu coûteuse d'accroître leurs performances, et pourrait établir un nouvel état de l'art dans ce domaine de recherche.

Remerciements

Cette publication a été rendue possible grâce à l'utilisation du supercalculateur FactoryIA, soutenu financièrement par le Conseil Régional d'Ile-De-France.

Références

BHOWMIK R., STRATOS K. & DE MELO G. (2021). Fast and effective biomedical entity linking using a dual encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, p. 28–37.

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.

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.

DSOUZA A., TEMPELMEIER N., YU R., GOTTSCHALK S. & DEMIDOVA E. (2021). Worldkg : A world-scale geographic knowledge graph. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, p. 4475–4484.

EL BOUKKOURI H. (2021). *Domain adaptation of word embeddings through the exploitation of in-domain corpora and knowledge bases*. Thèse de doctorat, Université Paris-Saclay.

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 (HEALTH)*, **3**(1), 1–23.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (2020). Don't stop pretraining : Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.740.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). Pubmedqa : A dataset for biomedical research question answering. 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. 2567–2577.

KRALLINGER M., RABAL O., AKHONDI S. A., PÉREZ M. P., SANTAMARÍA J., RODRÍGUEZ G. P., TSATSARONIS G., INTXAURRONDO A., LÓPEZ J. A., NANDAL U. *et al.* (2017). Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, p. 141–146.

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

LEE J.-S. & HSIANG J. (2020). Patent classification by fine-tuning bert language model. *World Patent Information*, **61**, 101965.

MICHALOPOULOS G., WANG Y., KAKA H., CHEN H. & WONG A. (2021). UmlsBERT : Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1744–1753.

MOHAN S. & LI D. (2019). MedMentions : A Large Biomedical Corpus Annotated with UMLS Concepts. *arXiv :1902.09476 [cs]*.

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 18th BioNLP Workshop and Shared Task*, p. 58–65.

PETERS M. E., NEUMANN M., LOGAN R. L., SCHWARTZ R., JOSHI V., SINGH S. & SMITH N. A. (2019). Knowledge enhanced contextual word representations. In *EMNLP*.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2007). Yago : A Core of Semantic Knowledge. In *16th International Conference on the World Wide Web*, p. 697–706.

UZUNER Ö., SOUTH B. R., SHEN S. & DUVALL S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, **18**(5), 552–556.

WANG R., TANG D., DUAN N., WEI Z., HUANG X.-J., JI J., CAO G., JIANG D. & ZHOU M. (2021a). K-Adapter : Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1405–1418.

WANG X., GAO T., ZHU Z., ZHANG Z., LIU Z., LI J. & TANG J. (2021b). Kepler : A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, **9**, 176–194.

XU Y., ZHONG X., YEPES A. J. J. & LAU J. H. (2020). Forget me not : Reducing catastrophic forgetting for domain adaptation in reading comprehension. In *2020 International Joint Conference on Neural Networks (IJCNN)*, p. 1–8 : IEEE.

YASUNAGA M., BOSSELUT A., REN H., ZHANG X., MANNING C. D., LIANG P. & LESKOVEC J. (2022). Deep bidirectional language-knowledge graph pretraining. *arXiv preprint arXiv :2210.09338*.

YUAN Z., ZHAO Z., SUN H., LI J., WANG F. & YU S. (2022). CODER : Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, **126**, 103983.

ZHANG Z., HAN X., LIU Z., JIANG X., SUN M. & LIU Q. (2019). ERNIE : Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1441–1451.

ZIPF G. K. (1935). *The Psycho-Biology of language*, volume 21. Psychology Press.

Annexes

A Échantillon aléatoire du corpus généré

1. *Skull Fractures is a type of Fracture of bone of head.*
2. *Procedures on breast is a type of Procedure on trunk.*
3. *Tarsum is an ingredient in Coal Tar 20 MG/ML / Salicylic Acid 50 MG/ML Medicated Shampoo [Tarsum].*
4. *The concept of "MK-5108" is part of the CTRP Agent Terminology.*
5. *Multi vessel coronary artery disease is an example of Coronary Arteriosclerosis.*
6. *amitriptyline hydrochloride 25 MG / perphenazine 2 MG [Etrafon] is a brand name for perphenazine 2 MG.*
7. *Trunk of parotid branch of left superficial temporal artery is a type of Trunk of parotid branch of superficial temporal artery.*
8. *Biopsy of lesion of internal nose can be used to identify biopsy of nasal cavity : carcinoid tumor.*
9. *Uterine fibroid embolization is a type of Embolization of artery.*
10. *Meperidine analog-containing product is a type of Piperidine derivative.*

B Hyperparamètres

Nos expériences ont révélé que les hyperparamètres de pré-entraînement optimaux étaient les mêmes d'un modèle à l'autre, ils ne sont donc pas différenciés dans la Table 3. Les hyper-paramètres des tâches d'évaluation ont été optimisés vis-à-vis de $\text{BERT}_{\text{BASE}}$. Il n'y a pas d'entrée pour la tâche MLM car les modèles sont déjà optimisés pour cette tâche en pré-entraînement.

Tâche	Pas d'apprentissage	Weight Decay	Taille de batch	Époques
Pré-entraînement	2e-6	0,01	4096	1
ChemProt	2e-5	0,01	24	30
PubMedQA	2e-5	0,01	32	5
SNLI	2e-4	0,01	160	10
CoLA	2e-5	0,01	32	10
SNLI	2e-5	0,01	32	5

TABLE 3 – Valeurs des hyperparamètres pour les diverses tâches.