Exploitation de plongements de graphes pour l'extraction de relations biomédicales

Anfu TANG[1, 2] Robert Bossy [1] Louise Deléger[1]
Claire Nédellec[1] Pierre Zweigenbaum[2]
(1) Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France
(2) Université Paris-Saclay, CNRS, LISN, Orsay, France
anfu.tang@inrae.fr, robert.bossy@inrae.fr, louise.deleger@inrae.fr,
claire.nedellec@inrae.fr, pz@lisn.fr

RÉSUMÉ

L'intégration de connaissances externes dans les modèles neuronaux est très étudiée pour améliorer les performances des modèles de langue pré-entraînés, notamment en domaine biomédical. Dans cet article, nous explorons la contribution de plongements de bases de connaissances à une tâche d'extraction de relations. Pour deux mentions d'entités candidates dans un texte, nous faisons l'hypothèse que la connaissance de relations entre elles, issue d'une base de connaissances (BC) externe, aide à prédire l'existence d'une relation dans le texte, y compris lorsque les relations de BC sont différentes de celles du texte. Notre approche consiste à calculer des plongements du graphe de BC et à estimer la possibilité pour chaque paire d'entité du texte qu'elle soit reliée par une relation de BC. Les expériences menées sur trois tâches d'extraction de relations en domaine biomédical montrent que notre méthode surpasse le modèle PubMedBERT de base et obtient des performances comparables aux méthodes de l'état de l'art.

ABSTRACT

Exploiting Graph Embeddings from Knowledge Bases for Neural Biomedical Relation Extraction

Integrating external knowledge into neural models has been extensively studied to improve the performance of pre-trained Language Models, especially in the biomedical domain. In this paper, we explore the contribution of knowledge base embeddings to relation extraction (RE) tasks. Given two candidate entity mentions in a text, we hypothesize that knowing which relations exist between them in an external knowledge base (KB) helps to predict the existence of a relation in the text, even when the KB relations are different from those of the RE task. Our approach consists in computing KB graph embeddings and in estimating the possibility that each pair of entities is linked by a KB relation. Experiments conducted on three biomedical RE tasks show that our method outperforms the baseline PubMedBERT model and yields comparable performance to state-of-the-art methods.

MOTS-CLÉS : Extraction de relations, BERT, plongements de graphes, base de connaissances.

KEYWORDS: Relation Extraction, BERT, Graph Embedding, Knowledge Base.

1 Introduction

L'extraction de relations est une tâche importante du traitement automatique des langues sur laquelle de nombreuses études existent. Elle consiste à identifier le type de relation entre une paire d'entités étant donné une phrase entière. Un exemple d'extraction de relations est montré dans la figure 1 : l'objectif est de déterminer quelle relation existe entre *Argatroban* et *thrombin*, par exemple *CPR:4*.

FIGURE 1 – Un exemple venant du corpus d'extraction de relations ChemProt (Krallinger *et al.*, 2017). Une relation "CPR:4" est annotée dans la phrase entre *Argatroban* et *thrombin* ; on trouve une relation "decrease^activity" entre ces entités dans la BC externe CTD (Davis *et al.*, 2023).

Les modèles de langue pré-entraînés fondés sur les architectures de type Transformer (Vaswani *et al.*, 2017) tels que BERT (Devlin *et al.*, 2019) obtiennent des performances à l'état de l'art dans diverses tâches de Traitement Automatique des Langues (TAL). Le modèle de langue BERT est pré-entraîné sur un corpus du domaine général, ce qui entraîne des limitations lorsqu'il est appliqué à un domaine spécifique. Pour adapter BERT à des domaines particuliers, des variantes de BERT (Lee *et al.*, 2020; Beltagy *et al.*, 2019; Gu *et al.*, 2021) ont été pré-entraînées sur des corpus de ces domaines, soit en partant de la version pré-entraînée dans le domaine général (SciBERT (Beltagy *et al.*, 2019), BioBERT (Lee *et al.*, 2020)), soit en partant de zéro (PubMedBERT (Gu *et al.*, 2021), CharacterBERT (El Boukkouri *et al.*, 2020)). Dans ces domaines, ces modèles spécifiques surpassent le modèle BERT pré-entraîné sur un corpus du domaine général (Lee *et al.*, 2020).

Le pré-entraînement de BERT sur un corpus de textes n'a pas d'objectif explicite visant à acquérir des connaissances factuelles qui pourraient contribuer à réaliser notre objectif d'extraction de relations. Certains travaux (Wang *et al.*, 2021; Hao *et al.*, 2020) ont proposé d'ajouter un objectif de pré-entraînement lié à une base de connaissances (BC) pour injecter des connaissances factuelles. D'autres (Iinuma *et al.*, 2022; Hao *et al.*, 2020) se concentrent sur l'utilisation de bases de connaissances pour obtenir des données supplémentaires. Dans la communauté du Web sémantique, des méthodes de plongement de graphes telles que (Ribeiro *et al.*, 2017; Bordes *et al.*, 2013; Sun *et al.*, 2019) ont été développées. Elles fournissent des outils pour intégrer des informations provenant de graphes de connaissance.

Cependant, dans un contexte d'extraction de relations à partir de textes, les bases de connaissances de domaines spécifiques contiennent souvent des relations différentes de celles qui sont recherchées dans les textes. Prenons l'exemple de *Comparative Toxicogenomics Database* (CTD) (Davis *et al.*, 2023) et du corpus ChemProt (Krallinger *et al.*, 2017) fréquemment exploités ensemble pour l'extraction de relations d'interactions entre produits chimiques et gènes. Dans CTD il existe des relations pour 134 interactions, telles que "*affects_stability*", "*increase_reaction*", etc., mais dans le corpus ChemProt, seulement 6 interactions sont recherchées, par exemple "*inhibitor*", "*downregulator*", etc. Les relations de CTD sont plus précises que celles de ChemProt et elles ne sont pas directement alignables, ce qui rend leur exploitation plus complexe. Cependant, nous supposons que les relations

de CTD peuvent être utiles ; par exemple, dans la figure 1, la relation *decrease_activity* trouvée dans CTD peut suggérer qu'il est plus probable que la relation à prédire appartienne à la classe *CPR:4*.

Dans cet article, nous proposons KB-PubMedBERT, un modèle neuronal spécifiquement conçu pour l'extraction de relations biomédicales capable d'exploiter des relations de bases externes, notamment lorsqu'elles sont différentes de celles qui sont à extraire des textes. Notre architecture se compose du modèle PubMedBERT pré-entraîné dans le domaine biomédical (Gu *et al.*, 2021) et d'un composant de plongement de graphes basé sur la méthode RotatE (Sun *et al.*, 2019). Nous partons de graphes de connaissances du domaine qui contiennent des relations possiblement liées aux relations cibles des tâches d'extraction de relations : nous faisons l'hypothèse qu'elles peuvent aider à améliorer la classification des relations cibles. Nous utilisons des plongements de ces graphes de connaissances pour estimer la possibilité que des relations de la base de connaissances existent entre deux mentions d'entités. En ajoutant ce profil de possibilités à la sortie du modèle de langue pré-entraîné, nous faisons en sorte que notre modèle encode à la fois les informations textuelles et les informations de la base de connaissances. Nous supposons qu'il devrait ainsi être plus performant dans les tâches d'extraction de relations. À notre connaissance, nous sommes les premiers à étudier comment l'utilisation de relations d'un graphe de connaissances qui ne sont pas les relations cibles peut aider à améliorer l'extraction de relations à l'aide d'un modèle de langue pré-entraîné.

Notre article est organisé comme suit. Nous présentons d'abord RotatE, la méthode de plongement de graphes utilisée comme composant de notre modèle, puis des études antérieures sur l'intégration de connaissances de BC dans des modèles de langues pré-entraînés (section 2). Nous présentons ensuite l'architecture de notre modèle et l'hypothèse sous-jacente (section 3). Nous détaillons enfin les expériences menées et les résultats obtenus (section 4), puis concluons (section 5).

2 Travaux connexes

Dans cette section, nous présentons d'abord des études antérieures sur un composant important de notre modèle : les méthodes de plongement de graphe. Nous présentons ensuite d'autres méthodes visant à intégrer les informations d'une base de connaissances dans les modèles neuronaux. Ces recherches ont en commun un objectif de performance en domaine de spécialité.

2.1 Plongement de graphe

Les méthodes de plongement de graphe apprennent des représentations vectorielles pour les sommets (concepts) et les arêtes (relations) de graphes. Inspirées de word2vec (Mikolov *et al.*, 2013), les méthodes basées sur le contexte telles que node2vec (Grover & Leskovec, 2016) et Struc2vec (Ribeiro *et al.*, 2017) consistent d'abord à échantillonner aléatoirement des séquences de sommets à partir du graphe, puis à utiliser des sommets voisins comme contexte pour apprendre des plongements de sommets. D'autres méthodes telles que TransE (Bordes *et al.*, 2013) et RotatE (Sun *et al.*, 2019) consistent à modéliser une arête possédant une étiquette donnée comme une transformation entre les vecteurs des sommets qu'elle relie. Par exemple, étant donné une arête r entre deux sommets h et t, les vecteurs correspondants sont notés $(\mathbf{h}, \mathbf{r}, \mathbf{t})$, TransE modélise r comme une translation telle que $\mathbf{t} = \mathbf{h} + \mathbf{r}$. Suivant le même principe mais au lieu d'une translation, la méthode RotatE modélise les arêtes comme des rotations dans l'espace vectoriel complexe : $\mathbf{t} = \mathbf{h} \circ \mathbf{r}$, où $\circ$ est le produit de Hadamard (élément par élément). Dans notre travail, nous choisissons la méthode RotatE car elle

surpasse systématiquement TransE sur plusieurs jeux de données, comme indiqué dans (Sun *et al.*, 2019).

2.2 Intégration des informations d'une base de connaissances dans BERT

Nous classons les méthodes existantes en deux types : 1) utilisation de la base de connaissances pour obtenir des données d'entraînement supplémentaires ; 2) modification des objectifs de pré-entraînement de BERT. Les deux types de méthodes peuvent être utilisées ensemble ou séparément.

2.2.1 Ajout de données d'entraînement à l'aide d'une base de connaissances

Les méthodes suivantes sont représentatives de la diversité des approches. Wang *et al.* (2021) proposent de figer les poids d'un BERT pré-entraîné et de pré-entraîner un encodeur supplémentaire à base de Transformer, dit 'adaptateur factuel', sur des phrases collectées à partir de Wikipédia. Cet adaptateur est entraîné à la classification de relations sur un jeu de données obtenu par alignement entre des triplets de Wikidata et des phrases de Wikipédia (Elsahar *et al.*, 2018). Hao *et al.* (2020) proposent plutôt de générer directement des phrases au format "[CLS] concept_1 [relation] concept_2 [SEP]" où (concept_1, relation, concept_2) est un triplet de la base de connaissances ou un exemple négatif, et [CLS] et [SEP] sont des sous-mots utilisés respectivement pour marquer le début et la fin des phrases, comme indiqué dans (Devlin *et al.*, 2019). Ces phrases artificielles sont ensuite utilisées pour pré-entraîner BERT à une tâche de classification de relations. Weber *et al.* (2022) proposent d'ajouter directement les définitions de produits chimiques obtenues à partir d'une base de connaissances comme données supplémentaires pendant l'affinage du modèle.

2.2.2 Mise à jour des objectifs de pré-entraînement

Les objectifs originaux du pré-entraînement de BERT sont la prédiction de mot masqué et la prédiction de la phrase suivante. Dans UmlsBERT, au lieu de masquer aléatoirement un mot et de demander au modèle de langue de le prédire, Michalopoulos *et al.* (2021) proposent de prédire en plus les entités qui partagent le même identifiant unique de concept (CUI) que le mot masqué si ce mot fait partie d'une entité reconnue dans le Metathesaurus de l'UMLS (Bodenreider, 2004). Dans ERNIE, Zhang *et al.* (2019b) ajoutent un troisième objectif de pré-entraînement associé à un liage référentiel entre tokens et concepts de BC : ils masquent aléatoirement le liage d'un token à un concept et demandent au modèle de prédire le concept masqué. Dans KeBioLM, Yuan *et al.* (2021) ajoutent deux tâches dans le pré-entraînement : la détection d'entités et le liage référentiel. Selon les résultats expérimentales, ces méthodes montrent des performances plus élevées pour des tâches de TAL biomédicales que celles sans modification du pré-entraînement de BERT.

3 Méthode proposée

Dans cette partie, nous présentons notre modèle KB-PubMedBERT qui exploite une base de connaissances. Ce modèle contient deux composants : PubMedBERT pré-entraîné, et un composant de plongement de graphe qui inclut une couche de plongement de concepts et une couche de plongement

de relations. Nous choisissons PubMedBERT comme base car il crée son propre vocabulaire de sous-mots contenant des termes biomédicaux, et ses performances dépassent des variantes de BERT spécifiques précédentes telles que BioBERT et SciBERT sur plusieurs tâches biomédicales (Gu *et al.*, 2021).

3.1 Hypothèse

La plupart des modèles antérieurs qui utilisent une base de connaissances (Zhang *et al.*, 2019b; Yuan *et al.*, 2021) se concentrent sur l'intégration d'informations sur les entités de la base de connaissances (BC) dans des modèles de langue pré-entraînés. Cependant, nous soutenons que l'incorporation d'informations sur les *relations* de la BC est également importante, en particulier pour les tâches d'extraction de relations dans des textes. Un défi auquel nous sommes confrontés en utilisant les relations d'une BC est que dans la plupart des cas, ces relations sont différentes de celles des tâches d'extraction de relations. Cependant, dans une BC liée à un domaine spécifique, les relations de la BC ont des chances d'être sémantiquement liées aux relations des textes, même si ce n'est que faiblement. Par conséquent, trouver un moyen d'exprimer la proximité sémantique entre les relations de la BC et les relations des textes est un point crucial dans la construction d'une architecture d'extraction de relations exploitant une BC. Par exemple, Iinuma *et al.* (2022) créent manuellement un alignement entre les relations de la BC et les relations cibles, puis s'en servent pour créer des données de supervision distante. Nous proposons d'aller plus loin en supprimant cette opération coûteuse d'alignement manuel et en faisant en sorte que le modèle neuronal apprenne cette correspondance automatiquement. Nous émettons l'hypothèse que notre modèle neuronal est capable de construire un alignement souple entre les relations de la BC et les relations du texte, et que l'ajout de ces suggestions de relations hypothétiques en sus de l'encodage du texte par PubMedBERT peut améliorer les performances du modèle en extraction de relations.

3.2 Architecture du modèle

La figure 2 présente une vue d'ensemble de l'architecture de notre modèle. Ce modèle prend deux entrées : la phrase s, et les identifiants des concepts des entités source $subj$ et cible obj. Les plongements de concepts et de relations sont respectivement initialisés avec des plongements de concepts et de relations de la BC entraînés à l'aide de RotatE, et les plongements des concepts présents dans le texte mais absents de la BC sont initialisés aléatoirement. Une fois initialisés, les plongements de concepts et de relations sont ajustés pendant l'entraînement du modèle. Le flux de données dans notre modèle est le suivant : pour deux entités candidates du texte source et cible, dont la relation est à prédire, et dont nous disposons du concept associé, nous obtenons d'abord les plongements $\mathbf{e_{subj}}$, $\mathbf{e_{obj}}$ pour la source et la cible, nous calculons ensuite les M scores de la formule (1) :

$$score(r_i) = (\gamma - ||\mathbf{e_{subj}} \circ \mathbf{r_i} - \mathbf{e_{obj}}||)^{\top}_{i=1,2,...,M} \quad (1)$$

où γ est une marge fixe, un hyper-paramètre dont la valeur est fixée pendant l'entraînement de RotatE. Cette définition du score vient de la fonction de coût que l'on utilise pour entraîner les plongements RotatE, comme indiqué dans (Sun *et al.*, 2019). Selon la définition de RotatE, un $score(r_i)$ élevé reflète la possibilité que la relation r_i soit valide entre $subj$ et obj pour la BC. Soit $\mathbf{h}_s$ l'encodage de s par PubMedBERT et $\mathbf{h}_{score}$ un vecteur de dimension M contenant les scores de toutes les relations

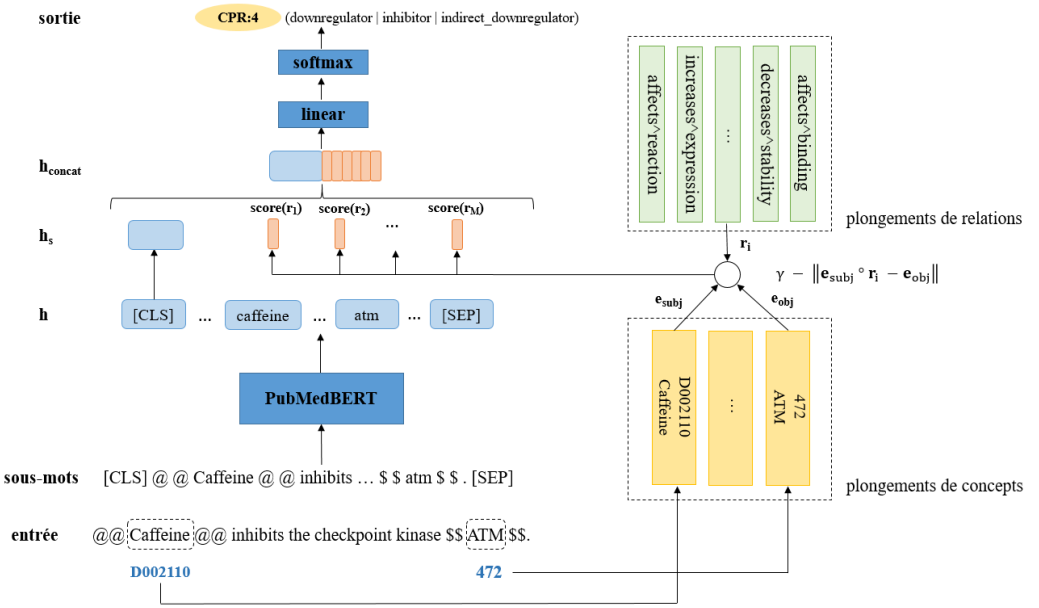

FIGURE 2 – Architecture globale de KB-PubMedBERT.

de BC, nous calculons la représentation combinée selon la formule (2) :

$$\mathbf{h}_{concat} = [\mathbf{h}_s; \mathbf{h}_{score}] \tag{2}$$

où $[.;.]$ désigne la concaténation vectorielle. La représentation combinée est ensuite transmise à une couche entièrement connectée suivie d'une couche softmax qui calcule la possibilité que les relations cibles soient exprimées entre les entités considérées. L'ensemble du modèle est affiné en utilisant l'entropie croisée comme fonction de coût avec des poids de PubMedBERT non-figés.

4 Expérimentations et Résultats

4.1 Jeux de données

Nous évaluons notre modèle sur trois tâches d'extraction de relations biomédicales dans des textes en anglais choisies pour la diversité de leurs caractéristiques. Les statistiques des corpus correspondants sont résumées dans le tableau 1.

1. La tâche ChemProt (Krallinger *et al.*, 2017) porte sur cinq types d'interactions de haut niveau entre produits chimiques et gènes, soit au total 6 relations dont la relation nulle (pas de relation).
2. La tâche DrugProt (Miranda *et al.*, 2021) porte sur 14 types d'interactions entre produits chimiques et gènes, dont la relation nulle.

	ChemProt	DrugProt	BB-Rel$_p$
nb. classes de relations	6	14	2
nb. exemples : entraînement	13 110	64 745	3 016
nb. exemples : développement	8 329	13 399	2 000
nb. exemples : test	10 990	238 694	2 473

TABLE 1 – Extraction de relations biomédicales : jeux de données ChemProt, DrugProt et BB-Rel$_p$.

3. La tâche BB-Rel de Bacteria Biotope (Bossy *et al.*, 2019) porte sur deux relations : "*lives_in*" entre micro-organismes et habitats ou zones géographiques, et "*exhibits*" entre micro-organismes et phénotypes. Comme nous ne disposons que de bases de connaissances portant sur les relations entre micro-organismes et habitats mais pas entre micro-organismes et zone géographique ou phénotypes, nous extrayons du jeu de données complet BB-Rel le sous-ensemble portant sur notre relation cible entre paires d'entités (micro-organisme, habitat). Nous désignons ce sous-ensemble par BB-Rel$_p$ (p pour *partiel*).

Nous choisissons des bases de connaissances adaptées à ces différentes tâches. Pour les tâches ChemProt et DrugProt, nous choisissons CTD (Davis *et al.*, 2023) qui recense 134 types d'interactions entre produits chimiques et gènes, comme "*affects^reaction*" ou "*increases^stability*". Pour obtenir une base de connaissances pour la tâche BB-rel$_p$, nous extrayons de la base de connaissances Omnicrobe (Dérozier *et al.*, 2023) les entités normalisées et la relation "*lives_in*" entre microorganismes et habitats provenant des sources de référence, BacDive, CIRM et GenBank.

4.2 Cadre expérimental

Pré-traitement Pour chacun des trois jeux de données, nous considérons uniquement les relations intra-phrase [1]. Comme la plupart des études précédentes sur l'extraction des relations telles que (Lee *et al.*, 2020; Gu *et al.*, 2021), nous utilisons deux types de marqueurs pour ajouter des informations de position sur les arguments d'une relation candidate : "@@" au début et à la fin de la mention de l'entité source ; "$$" au début et à la fin de la mention de l'entité cible. L'objectif de ces marqueurs est de fournir au modèle l'information de position des entités ciblées.

Liage référentiel. Pour une tâche d'extraction de relations, l'alignement entre les entités des textes et les concepts de la BC n'est pas toujours donné. Dans nos expériences, cet alignement est soit fourni comme référence dans les corpus annotés, soit obtenu par des outils existants pré-entraînés pour le liage référentiel (normalisation d'entités). La façon dont nous obtenons les normalisations des entités est résumée dans le tableau 2. Pour ChemProt et DrugProt, nous effectuons le liage référentiel des mentions d'entités vers des concepts CTD à l'aide de la méthode BioSyn (Sung *et al.*, 2020) [2]. Pour BB-Rel$_p$, les entités de type microbe sont normalisées par la taxonomie des espèces du NCBI à l'aide du modèle proposé par le meilleur participant (Mao & Liu, 2019) [3] à la tâche BB-Norm (Bossy *et al.*, 2019), et les entités de type habitat sont normalisées par l'ontologie OntoBiotope à

1. L'évaluation classe en faux-négatifs toutes les relations inter-phrases.
2. Nous utilisons deux modèles BioSyn entraînés par Sung *et al.* (2020) : biosyn-sapbert-bc5cdr-chemical pour les produits chimiques, et biosyn-sapbert-bc2gn pour les gènes. Les deux modèles obtiennent respectivement 96,6 et 91,3 comme acc@1 sur les tâches de normalisation d'entités correspondantes.
3. Ce modèle obtient 0,78 comme précision pour la normalisation des microbes de la tâche BB-Norm.

	ChemProt & DrugProt	BB-Rel$_p$
jeu d'entraînement	BioSyn	*gold*
jeu de validation	BioSyn	*gold*
jeu de test	BioSyn	C-Norm, *regression*

TABLE 2 – Sources de normalisation des entités de ChemProt, DrugProt et BB-rel$_p$: respectivement sur le jeu d'entraînement, de validation et de test. "*gold*" représente les annotations manuelles fournies dans BB-Norm ; "*regression*" réfère au modèle de (Mao & Liu, 2019) qui est un modèle de régression.

l'aide de la méthode état de l'art C-Norm (Ferré *et al.*, 2020) [4]. Soulignons que même si sur chaque tâche, les entités textuelles sont normalisées avec les concepts des référentiels utilisés par la BC choisie, il se peut que certains concepts d'entités n'aient pas de plongements pré-calculés par RotatE. En effet, il existe dans les BC des concepts isolés, i.e., des concepts n'ayant pas de relations avec d'autres concepts, ils ne sont alors pas utilisés pour entraîner RotatE. Pour les entités correspondant à des concepts isolés, nous utilisons un vecteur aléatoirement initialisé au début de l'affinage de KB-PubMedBERT.

Base de comparaison. Nous utilisons le modèle pré-entraîné PubMedBERT comme base de comparaison, car c'est le modèle dont est dérivé notre méthode. Sur chaque jeu de données, PubMedBERT est affiné pour classifier les relations cibles de chaque tâche. La comparaison des performances de notre modèle KB-PubMedBERT à celles de cette architecture de base permet de montrer directement si les informations intégrées à partir de la BC sont utiles.

De manière classique, le vecteur du token [CLS] encode la phrase. On le fait passer à travers une couche linéaire, puis une fonction SoftMax pour obtenir les probabilités des relations à prédire.

Hyperparamètres. Nous utilisons l'implémentation officielle [5] de RotatE (Sun *et al.*, 2019) pour calculer les plongements de concepts et de relations de la BC. Nous fixons expérimentalement la dimension des plongements à 200, γ à 24,0 et le taux d'apprentissage à $1e^{-4}$. Pour chaque jeu de données d'extraction de relations, nous utilisons l'ensemble de développement pour optimiser les hyperparamètres. Nous effectuons une recherche en grille pour deux hyperparamètres : le taux d'apprentissage $(1e^{-5}, 2e^{-5}, 5e^{-5})$ et la taille de lot $(8, 16)$. Nous maintenons le taux d'apprentissage constant pendant l'affinage et fixons le nombre d'époques à 15. Pour la base de comparaison, comme pour KB-PubMedBERT, chaque expérience est répétée avec 5 amorces différentes.

Évaluation. Le score Micro F1 excluant la relation nulle [6] est la métrique d'évaluation standard pour les trois jeux de données : concrètement, pour calculer le score F1, les vrais positifs pris en compte sont uniquement les prédictions correctes des relations non-nulles. Pour évaluer les performances de notre modèle, selon les cas, nous soumettons nos prédictions au service d'évaluation en ligne officiel (cas de BB-Rel) ou utilisons le kit d'évaluation officiel [7].

4. Ce modèle obtient 0,60 comme score strict et 0.78 comme score de Wang pour les habitats pour la tâche BB-Norm.
5. https://github.com/DeepGraphLearning/KnowledgeGraphEmbedding
6. Inclure la relation nulle, qui est généralement majoritaire, donne des résultats optimistes.
7. ChemProt : https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-5/ ; DrugProt : https://codalab.lisn.upsaclay.fr/competitions/8293#participate ; BB-Rel : http://bibliome.jouy.inra.fr/demo/BioNLP-OST-2019-Evaluation/index.html.

	ChemProt	DrugProt	BB-Rel$_p$
PubMedBERT	77,5 $\pm$ 0,7 / 79,6	75,1 $\pm$ 0,4 / 77,7	61,4 $\pm$ 1,2 / 64,3
KB-PubMedBERT	**78,4** $\pm$ **0,9** * / 80,4	75,5 $\pm$ 0,8 / 77,9	**63,3** $\pm$ **2,1** * / **65,7**
État de l'art	78,0 / —	— / **79,7**	— / 64,8

TABLE 3 – Scores micro F1 sur les tâches d'extraction de relation. Nous rapportons a/b où a représente le score moyen de 5 exécutions avec différentes initialisations aléatoires et b représente le score d'un ensemble à vote majoritaire. Nous rapportons les deux scores pour mieux comparer nos résultats aux résultats de l'état de l'art. * indique qu'un test T unilatéral émettant l'hypothèse que la performance moyenne de KB-PubMedBERT est meilleure que la base de comparaison obtient une p-valeur $p < 0{,}1$ (respectivement $0{,}07$ pour ChemProt et $0{,}08$ pour BB-Rel$_p$), ce qui est interprété comme une faible présomption contre l'hypothèse nulle. Le système état de l'art considéré est indiqué en début de section 4.3.

4.3 Résultats

Nous comparons KB-PubMedBERT à PubMedBERT et à l'état de l'art suivant sur chaque corpus :

1. Pour ChemProt : nous prenons comme état de l'art le modèle qui a obtenu le score le plus élevé en utilisant le même kit d'évaluation que nous. Il s'agit de SciFive (Phan *et al.*, 2021), un modèle T5 pré-entraîné sur la littérature biomédicale ;
2. Pour DrugProt : un ensemble de 10 RoBERTa-large-PM-M3-Voc (Lewis *et al.*, 2020) avec des définitions de produits chimiques sélectionnées à partir de CTD (Weber *et al.*, 2022) ;
3. Pour BB-Rel$_p$: un modèle Transformer à 12 couches pré-entraîné sur les corpus Wikipédia anglais, BooksCorpus, PubMed et PubMed Central (PMC) (Zhang *et al.*, 2019a).

Les résultats expérimentaux sont résumés dans le tableau 3. Nous observons que KB-PubMedBERT surpasse systématiquement la base de comparaison, ce qui prouve l'efficacité de notre méthode d'injection d'informations de BC. Cependant, la différence entre KB-PubMedBERT et la base de comparaison n'est pas significative sur DrugProt. Notre modèle surpasse l'état de l'art antérieur sur ChemProt et BB-Rel$_p$, mais reste environ 2 % derrière le meilleur score sur DrugProt. L'absence d'amélioration de KB-PubMedBERT sur DrugProt peut s'expliquer par le fait que près de 1 % des occurrences des entités de DrugProt sont absentes de la BC, alors que ce nombre est de $0{,}1$ % pour ChemProt et de 0 % pour BB-Rel$_p$. Comme les plongements pour ces entités absentes de la BC sont initialisés aléatoirement, un pourcentage plus élevé d'occurrences d'entités absentes signifie que moins d'informations venant de la BC sont intégrées.

4.4 Étude d'ablation complémentaire

L'examen de PubMedBERT seul ci-dessus constitue une première étude d'ablation dans laquelle le composant de plongement de graphes de notre modèle n'est pas utilisé. Inversement, pour vérifier la contribution intrinsèque du composant de plongement de graphes à la performance finale de l'extraction de relations, nous menons des expériences où nous supprimons complètement PubMedBERT de notre modèle : nous n'utilisons que la paire d'entités (source, cible) pour déduire le type d'interaction via les plongements de graphes RotatE. Les résultats sont dans le tableau 4. Nous observons que même sans aucun contexte, la base de connaissances obtient dans notre modèle un score F1 supérieur

	ChemProt	DrugProt	BB-Rel$_p$
KB-PubMedBERT$^-$	23,8 $\pm$ 1,6	19.5 $\pm$ 1,0	26,6 $\pm$ 0,3
majority	17,3	12,3	38,3

TABLE 4 – Ablation : KB-PubMedBERT$^-$ désigne notre modèle sans PubMedBERT. Nous rapportons le score moyen de 5 exécutions.

à 0,20 (la relation nulle est exclue de l'évaluation). Nous établissons par ailleurs un modèle simple qui prédit toujours la classe majoritaire. Ce modèle est nommé "*majority*" et ses résultats sont également montrés dans le tableau. On constate que les résultats de KB-PubMedBERT$^-$ sont significativement meilleurs que ceux du modèle aléatoire sur ChemProt et DrugProt, cela confirme que le profil de scores $\mathbf{h}_{score}$ représentant des suggestions de relations à grain fin de la base de connaissances entre deux entités, obtenues à partir des plongements de graphes RotatE, est utile pour l'extraction de relations biomédicales. Le fait que *majority* donne un meilleur résultat que KB-PubMedBERT$^-$ dans le cas de BB-Rel$_p$ peut être expliqué par le fait qu'il n'existe dans ce cas qu'une seule relation positive.

5 Conclusion

Dans cet article, nous proposons l'architecture KB-PubMedBERT qui injecte dans PubMedBERT les informations d'une base de connaissances pour améliorer ses performances en extraction de relations biomédicales. À la différence des modèles antérieurs utilisant des bases de connaissances, nous calculons d'abord à l'aide de la méthode de plongement de graphe RotatE les possibilités de relations de la BC entre les entités considérées, puis nous utilisons ces possibilités pour déduire les relations cibles de l'extraction de relation. Nous menons des expériences sur trois tâches d'extraction de relations biomédicales : notre modèle y surpasse systématiquement PubMedBERT et obtient des performances proches de l'état de l'art antérieur ou meilleures que lui. Une étude d'ablation confirme de plus la pertinence des relations de la base de connaissances indépendamment du modèle de langue PubMedBERT. À l'avenir, nous complèterons nos expériences sur d'autres jeux de données pour étendre l'étude de l'applicabilité de notre méthode.

Remerciements

Nous remercions la plateforme Saclay-IA de l'Université Paris-Saclay pour les ressources de calcul et de stockage du cluster GPU Lab-IA.

Ce travail a été financé par le Labex DigiCosme (projet ANR-11-LABEX-0045-DIGICOSME) opéré par l'ANR dans le cadre du programme "Investissement d'Avenir" Idex Paris-Saclay (ANR-11-IDEX-0003-02).

Références

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3615–3620, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

BODENREIDER O. (2004). The Unified Medical Language System (UMLS) : Integrating biomedical terminology. *Nucleic Acids Research*, **32**(Database issue), D267–270. DOI : 10.1093/nar/gkh061.

BORDES A., USUNIER N., GARCIA-DURÁN A., WESTON J. & YAKHNENKO O. (2013). Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 2787–2795, Red Hook, NY, USA : Curran Associates, Inc.

BOSSY R., DELÉGER L., CHAIX E., BA M. & NÉDELLEC C. (2019). Bacteria Biotope at BioNLP Open Shared Tasks 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, p. 121–131, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5719.

DAVIS A. P., GRONDIN C. J., JOHNSON R. J., SCIAKY D., WIEGERS J., WIEGERS T. C. & MATTINGLY C. J. (2023). Comparative Toxicogenomics Database (CTD) : update 2023. *Nucleic acids research*, **51**(D1), D1257–D1262. DOI : 10.1093/nar/gkac833.

DÉROZIER S., BOSSY R., DELÉGER L., BA M., CHAIX E., HARLÉ O., LOUX V., FALENTIN H. & NÉDELLEC C. (2023). Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. *PloS one*, **18**(1), e0272473. DOI : 10.1371/journal.pone.0272473.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EL BOUKKOURI H., FERRET O., LAVERGNE T., NOJI H., ZWEIGENBAUM P. & TSUJII J. (2020). CharacterBERT : Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6903–6915, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.609.

ELSAHAR H., VOUGIOUKLIS P., REMACI A., GRAVIER C., HARE J., LAFOREST F. & SIMPERL E. (2018). T-REx : A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

FERRÉ A., DELÉGER L., BOSSY R., ZWEIGENBAUM P. & NÉDELLEC C. (2020). C-Norm : a neural approach to few-shot entity normalization. *BMC bioinformatics*, **21**(23), 579. DOI : 10.1186/s12859-020-03886-8.

GROVER A. & LESKOVEC J. (2016). node2vec : Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 855–864, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2939672.2939754.

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. DOI : 10.1145/3458754.

HAO B., ZHU H. & PASCHALIDIS I. (2020). Enhancing clinical BERT embedding using a biomedical knowledge base. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 657–661, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.57.

IINUMA N., MIWA M. & SASAKI Y. (2022). Improving supervised drug-protein relation extraction with distantly supervised models. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, p. 161–170, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bionlp-1.16.

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. (2020). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. DOI : 10.1093/bioinformatics/btz682.

LEWIS P., OTT M., DU J. & STOYANOV V. (2020). Pretrained language models for biomedical and clinical tasks : Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, p. 146–157, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.clinicalnlp-1.17.

MAO J. & LIU W. (2019). Integration of deep learning and traditional machine learning for knowledge extraction from biomedical literature. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, p. 168–173, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5724.

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, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.139.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In Y. BENGIO & Y. LECUN, Éds., *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

MIRANDA A., MEHRYARY F., LUOMA J., PYYSALO S., VALENCIA A. & KRALLINGER M. (2021). Overview of DrugProt BioCreative VII track : quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*, p. 11–21.

PHAN L. N., ANIBAL J. T., TRAN H., CHANANA S., BAHADROGLU E., PELTEKIAN A. & ALTAN-BONNET G. (2021). SciFive : a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv :2106.03598*.

RIBEIRO L. F., SAVERESE P. H. & FIGUEIREDO D. R. (2017). struc2vec : Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, p. 385–394.

SUN Z., DENG Z., NIE J. & TANG J. (2019). RotatE : Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* : OpenReview.net.

SUNG M., JEON H., LEE J. & KANG J. (2020). Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3641–3650, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.335.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30 de *NIPS'17*, p. 6000–6010, Red Hook, NY, USA : Curran Associates, Inc.

WANG R., TANG D., DUAN N., WEI Z., HUANG X., JI J., CAO G., JIANG D. & ZHOU M. (2021). K-Adapter : Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1405–1418, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.121.

WEBER L., SÄNGER M., GARDA S., BARTH F., ALT C. & LESER U. (2022). Chemical–protein relation extraction with ensembles of carefully tuned pretrained language models. *Database*, **2022**. baac098, DOI : 10.1093/database/baac098.

YUAN Z., LIU Y., TAN C., HUANG S. & HUANG F. (2021). Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 180–190, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.20.

ZHANG Q., LIU C., CHI Y., XIE X. & HUA X. (2019a). A multi-task learning framework for extracting bacteria biotope information. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, p. 105–109, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-5716.

ZHANG Z., HAN X., LIU Z., JIANG X., SUN M. & LIU Q. (2019b). ERNIE : Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1441–1451, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1139.