Nouveaux réseaux neuronaux profonds pour l'alignement d'ontologies

Menad Safaa, Wissame Laddada, Saïd Abdeddaim, Lina F. Soualmia

Nouveaux réseaux neuronaux profonds pour l'alignement d'ontologies

S. Menad[1], W. Laddada[1], S. Abdeddaïm[1], LF. Soualmia[1]

[1] Univ. Rouen Normandie, LITIS UR4108, 76000, Rouen

{safaa.menad1,wissame.laddada,said.abdeddaim,soualfat}@univ-rouen.fr

Résumé

L'alignement d'ontologies s'appuie souvent sur des approches lexicales. Cependant, avec le développement des modèles de langage basés sur les transformeurs, il est désormais possible de comparer des textes en se basant sur le contexte plutôt que sur les caractéristiques lexicales. Dans ce travail, nous proposons de nouveaux modèles neuronaux siamois qui optimisent une fonction d'apprentissage contrastif auto-supervisé sur des articles de la littérature scientifique. Les résultats obtenus sur plusieurs benchmarks montrent que les modèles proposés permettent d'améliorer différentes tâches biomédicales. Ensuite, nous exploitons ces modèles dans la tâche d'alignement d'ontologies biomédicales.

Mots-clés

Modèles de langage, Transformeurs, Modèles neuronaux siamois, Apprentissage sans exemple, Textes biomédicaux, Ontologies biomédicales, Alignement d'ontologies.

Abstract

Ontology alignment often relies on lexical approaches. However, with the development of transformer-based language models, it is now possible to compare texts based on context rather than lexical characteristics. In this work, we propose new siamese neural models that optimize a self-supervised contrastive learning function using scientific literature articles. The results obtained from multiple benchmarks demonstrate that the proposed models improve various biomedical tasks. Moreover, we apply these models to the task of biomedical ontology alignment.

Keywords

Language Models, Transformers, Siamese Neural Networks, Zero-shot Learning, Biomedical Texts, Biomedical Ontologies, Ontology Alignment.

1 Introduction

L'alignement d'ontologie joue un rôle important dans l'intégration de connaissances. Il permet de faire correspondre des entités sémantiquement liées provenant de différentes ontologies. Les ontologies de domaine contiennent souvent un grand nombre de classes, ce qui non seulement pose des problèmes d'évolutivité, mais rend également plus difficile la distinction entre des classes ayant des noms et/ou des contextes similaires mais représentant des objets différents. Les solutions d'alignement ontologique existantes reposent généralement sur la correspondance lexicale comme base et la combinent avec la correspondance structurelle et la réparation d'alignements basée sur la logique.

Récemment, l'apprentissage automatique a été proposé comme une alternative aux méthodes de correspondance lexicale et/ou structurelle. Par exemple, DeepAlignment [11] utilise des plongements de mots pour représenter les classes et calcule la similarité de deux classes en fonction de la distance euclidienne de leurs vecteurs de mots. Néanmoins, ces méthodes adoptent des modèles de plongement de mots généralistes non contextuels tels que Word2Vec.

Les modèles de langage basés sur des transformeurs préentraînés tels que BERT [4] peuvent apprendre des plongements de texte contextuels robustes. Bien que ces modèles donnent de bons résultats dans de nombreuses tâches de traitement automatique du langage naturel (TAL), ils n'ont pas encore été suffisamment étudiés dans la tâche d'alignement ontologique et de mapping de concepts. Dans cet article, nous introduisons nos modèles transformeurs que nous avons développés dans la tâche d'alignement d'ontologies en présentant comment ces modèles pourraient être utilisés pour mapper sémantiquement des entités provenant de différentes ontologies biomédicales.

Cet article se structure comme suit : dans la section 2 nous citons des travaux existants dans le domaine de l'alignement d'ontologies. Les section 3 et 4 sont dédiées à la description des modèles de langage que nous proposons, BioS-Transformers et BioS-MiniLM, ainsi que la fonction objectif d'apprentissage contrastif. Les sections 5 et 6 décrivent les premiers résultats d'alignement d'ontologie avec nos modèles siamois sur deux ontologies biomédicales. Nous concluons et ouvrons des perspectives en section 6.

2 Travaux existants

Plusieurs ontologies de domaine ou d'application sont utilisées pour un même objectif. Cependant, des redondances ou des relations manquantes entre les concepts issus de différentes ontologies peuvent exister. Dans la littérature, l'alignement d'ontologies constitue une solution pour remédier à cette hétérogénéité et permettre une inter-opérabilité

sémantique entre les applications qui reposent sur plusieurs ontologies. L'alignement d'ontologies peut être défini comme une amélioration sémantique entre les concepts, les rôles et les instances de plusieurs ontologies pour une application donnée.
Dans [23], les auteurs ont défini un système distribué comme un système reliant deux ontologies. En fonction de cette définition, trois sémantiques d'un système distribué sont distinguées : une sémantique distribuée simple où la représentation des connaissances est interprétée dans un seul domaine, une sémantique distribuée intégrée où chaque représentation locale des connaissances est interprétée dans son propre domaine, et une sémantique distribuée contextuelle où le domaine d'interprétation n'est pas global.
Dans notre travail, nous souhaitons aligner deux ontologies issues d'un seul domaine (ontologies biomédicales) à l'aide de transformeurs pré-entraînés. De ce fait, la sémantique déployée est une sémantique distribuée simple.
L'alignement d'ontologies résulte d'une tâche importante de mise en correspondance (Ontology Mathcing - OM) où un alignement est défini pour identifier les similarités entre les ontologies. En ce qui concerne la classification des systèmes d'alignement présentée dans [20], un alignement peut reposer sur des similarités terminologiques (par exemple, des labels, des commentaires, des attributs, etc.), structurelles (description d'ontologie), extensionnelles (instances) ou sémantiques (interprétation et raisonnement logique). De plus, en raison du faible niveau d'expressivité sémantique de certaines ontologies, des ressources externes peuvent être exploitées dans les approches d'alignement. Par exemple, c'était le cas dans l'étude [13] pour l'alignement de l'ontologie SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) avec BioTopLite2, une ontologie de haut niveau.
En ce qui concerne la mise en correspondance des éléments d'ontologies, une étude de la littérature approfondie est présentée dans [18] pour décrire les ressources externes et leur utilisation. Les auteurs distinguent quatre catégories d'approches d'alignement utilisant ces ressources : les requêtes factuelles, où les données stockées dans la ressource sont simplement récupérées ; les approches reposant sur la structure, où les éléments structurels de la ressource sont exploités ; les approches statistiques/neuronales (Fine-TOM [9], DAEOM [22]), où des approches statistiques ou d'apprentissage profond sont appliqués à la ressource ; et les approches orientées logique où le raisonnement est déployé sur la ressource externe. Par exemple, dans [2], des stratégies terminologiques et structurelles ainsi qu'une ressource externe ont été employées pour aligner des ontologies biomédicales.
Tout comme CIDER-LM [21], notre approche d'alignement dépend de similitudes terminologiques calculées avec des approches neuronales mais aussi une similitude contextuelle. Nous exploitons des ressources externes sur lesquelles un apprentissage profond est appliqué afin de propager un contexte de similitude entre les éléments (propriétés et classes) de deux ontologies biomédicales. La différence entre les deux approches réside dans le modèle de représentation utilisé. Dans [21] c'est le modèle S-BERT [19] qui est utilisé, tandis que dans notre travail, nous appliquons les modèles BioSTransformers que nous avons développés et que nous détaillons ci-après.

3 Modèles siamois

Les sentence-transformers sont des modèles de langage qui ont été développés pour la tâche de calcul d'un score de similarité entre deux phrases. Ils utilisent des transformeurs pour des tâches liées aux paires de phrases : calcul de similarité sémantique entre phrases, recherche d'informations, reformulation de phrases, etc.
Ces transformeurs reposent sur deux architectures : les cross-encodeurs qui traitent la concaténation de la paire et les modèles siamois bi-encodeurs qui encodent en vecteur chacun des éléments de la paire. Sentence-BERT [19] est un bi-encodeur basé sur BERT permettant de générer des plongements de phrases sémantiquement significatifs à utiliser dans des comparaisons de similarité textuelle.

4 Modèles de langage proposés

Les transformeurs siamois ont été initialement conçus pour transformer des phrases (de taille similaire) en vecteurs. Nous proposons dans notre approche de transformer dans le même espace vectoriel les termes MeSH, les titres et les résumés des articles PubMed en entraînant un modèle de transformeur siamois sur ces données. Nous voulons nous assurer qu'il y a une correspondance dans cet espace vectoriel entre le texte court et le texte long. Nous avons donc entraîné nos modèles avec des paires d'entrées (titre, terme MeSH) et (résumé, terme MeSH).
Sur ces données nous avons construit deux types de modèles [14] : le premier type est notre propre transformeur siamois (BioSTransformers) construit à partir d'un transformeur pré-entraîné sur des données biomédicales et le second est un transformeur siamois déjà pré-entraîné sur des données généralistes (BioS-MiniLM).

Dans cette étude, nous présentons une nouvelle variante de nos modèles BioSTransformers, appelée S-BioClinicalBERT. Ce nouveau modèle a été pré-entraîné sur les notes cliniques de la base de données MIMIC. Ensuite, nous l'avons entraîné sur nos données biomédicales et utilisé pour résoudre la tâche d'alignement d'ontologies. Le tableau 1 présente les résultats obtenus par ce modèle sur différents benchmarks selon le F1 score.
BioSTransformers. Pour construire les BioSTransformers, nous nous sommes inspirés du modèle Sentence-BERT [19] en remplaçant BERT par d'autres transformeurs entraînés sur des données biomédicales (bio-transformeurs). Nous avons sélectionné quatre bio-transformeurs BlueBERT [17], PubMed BERT [6], BioELECTRA [10] et BioClinicalBERT [1].
Pour l'entraînement, nous utilisons une fonction objectif d'apprentissage contrastif auto-supervisé basée sur la fonction de perte de classement négatif multiple [8] dite MNRL (Multiple Negative Ran-

king Loss) dans le package Sentence-Transformers (https ://www.sbert.net/docs/package_reference/losses.html /#multiplenegativesrankingloss). La MNRL n'a besoin que des paires positives en entrée (le titre ou le résumé et un terme MeSH associé à l'article dans notre cas). Pour une paire positive (titre_i ou résumé_i, MeSH_i), la MNRL considère que chaque paire (titre_i ou résumé_i, MeSH_j) avec $i \neq j$ dans le même batch est négative.

Modèle/Corpus	HoC	PubMedQA	BioASQ
S-BioClinicalBERT	0.457	0.652	0.714

TABLE 1 – Évaluation de notre modèle sur les différents benchmarks selon le F1 score.

5 Alignement d'ontologies

Pour mieux comprendre notre cas d'utilisation, qui illustre un alignement d'ontologies biomédicales, nous présentons dans cette section quelques définitions inspirées d'autres travaux [18, 5, 16] et adaptées à notre objectif. La Figure 1 résume le processus d'un alignement d'ontologies suivant les définitions présentées dans cette section.

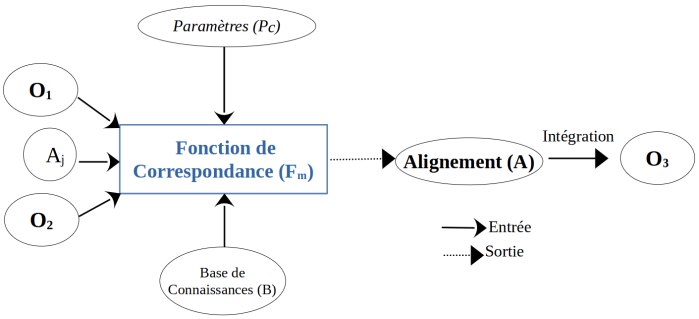

FIGURE 1 – Processus d'alignement d'ontologies (inspiré de [20]).

Définition d'ontologie : Nous considérons une ontologie O_i comme un ensemble de vocabulaire défini au moyen de taxonomies pour décrire un domaine d'application. Ce vocabulaire est considéré comme un ensemble d'éléments $e_i =< C_i, R_i, I_i >$; avec C_i pour décrire l'ensemble des concepts, R_i regroupe les relations pour relier les concepts et I_i compose l'ensemble des instances pour interpréter les concepts et les relier avec R_i. La sémantique d'une ontologie O_i peut être enrichie en définissant des axiomes (X_i) formalisés via la logique de description ou les logiques du premier ordre.

Alignement d'ontologies : un alignement décrit la correspondance entre deux ontologies. Étant donné deux ontologies O_1 et O_2, nous définissons l'alignement A comme un ensemble de triplets. Chaque triplet est spécifié par la terminologie de la relation binaire $r(e_1, e_2)$, où r représente la relation entre les deux éléments $e_1 \in O_1$ et $e_2 \in O_2$. L'alignement définit donc, le processus de recherche de ces ensembles de correspondance. Un score de confiance c peut également être ajouté au triplet de correspondance pour mesurer la similarité entre e_1 et e_2 (par exemple, la valeur de $c \in$ [0,1]).

Système de correspondance : il peut être défini comme une fonction de correspondance ayant plusieurs paramètres pour calculer la similarité entre les entités. Soit $F_m(O_1, O_2, A_j, P_c, B)$ avec P_c comme étant le paramètre qui précise la valeur de confiance de similarité et B comme étant l'ensemble de ressources externes utilisées pour trouver (ou pas) un alignement A_j entre l'élément e_1 et e_2.

Intégration d'ontologies : en considérant le travail présenté dans [16], nous définissons l'intégration d'ontologie comme un enrichissement sémantique d'une ontologie cible O_1 en exploitant des éléments d'une ontologie source O_2. Le résultat est une nouvelle ontologie O_3 définie par l'alignement $A =< r_j, e_{1,j}, e_{2,j}, c_j >$.

6 Modèles pour l'alignement

Dans cette section, nous décrivons notre approche pour aligner les éléments de différentes ontologies biomédicales en utilisant nos modèles siamois décrits précédemment. Ainsi, ce dernier est un système central du processus de correspondance. Étant donné que les transformeurs fonctionnent comme des modèles de langage, il est important que les éléments de l'ontologie soient définis par des labels (ou des commentaires) et enrichis par des relations (propriétés).
Nous considérons le processus de correspondance comme étant un problème de similarité où notre modèle (BioS-Transformers) reçoit des éléments extraits à partir des ontologies en entrée et calcule leur similarité. En fonction du score de sortie, nous concluons si une correspondance est possible entre les deux éléments. Plusieurs ressources ont été utilisées et sont décrites ci-après :
I) *RxNorm* [15] est un standard nomenclature développé dans le domaine de traitement médical, par la NLM (*United States National Library of Medicine*- Bibliothèque américaine de médecine). La création de ce standard est motivée par le besoin d'unifier la terminologie qui permet de représenter les médicaments, mais également de permettre de l'inter-opérabilité sémantique. De plus, ce standard fournit une normalisation pour les médicaments cliniques et les noms de médicaments connexes. Ces derniers sont reliés à des vocabulaires couramment utilisés dans ce même domaine.
II) *ChEBI* [3] est un dictionnaire d'entités moléculaires décrivant les "petits" composants chimiques (182 374 classes, 10 relations). Les entités moléculaires sont soit des produits naturels, soit des produits synthétiques. ChEBI contient aussi des groupes (fait-partie d'entités moléculaires) et des classes d'entités. Ce dictionnaire comprend une classification ontologique, dans laquelle les relations entre les entités moléculaires ou les classes d'entités et leurs parents et/ou enfants sont spécifiées.
III) *DRON* [7] a été développée à partir de l'alignement

d'entité de RxNorm et ChEBI. DRON est composée de 661 999 classes et 125 relations avec une profondeur de 27 niveaux.

IV) ***DOID*** [12] décrit des maladies et un vocabulaire médical via l'alignement de plusieurs ressources externes dans le but de lier les données biomédicales sur les gènes et les maladies. Elle est composée de 8 127 classes, 46 relations, avec une profondeur maximale de 13.

Notre cas d'utilisation décrit l'alignement d'éléments de deux ontologies biomédicales : DOID (*Human Disease Ontology* - ontologie des maladies humaines [1]) et DRON (*Drug Ontology* - ontologie des médicaments [2]. Le résultat de cet alignement représente une intégration d'ontologie dans laquelle chaque maladie est associée à une liste de médicaments potentiels. La finalité de ce processus d'alignement est d'intégrer l'ontologie résultante dans le système prédictif PrediBioOntoL. Afin de décrire la démarche du processus d'alignement, les phases listées dans [16] ont été adoptées.

6.1 La phase de prétraitement

Les données textuelles ont été extraites à partir des deux ontologies DOID et DRON via des requêtes SPARQL. Ces données concernent : (i) les classes (élément de DOID) qui décrivent une maladie [3]) et (ii) les métadonnées issues de ChEBI (Chemical Entities of Biological Interest - entités chimiques d'intérêt biologique) à partir desquelles l'ontologie DRON a été décrite. Ces métadonnées représentent des informations sur une maladie à travers une définition de propriété de données (métadonnées de ChEBI [4]). Cependant, aucune association n'est établie entre les ontologies DOID et DRON. Nous avons pu extraire un total de 13 678 maladies (DOID) et 3 295 métadonnées (DRON).

6.2 La phase de mise en correspondance

Le modèle BioSTransformers est utilisé comme fonction de correspondance, où les bases de connaissances externes représentent les données sur lesquelles le modèle est entraîné : PubMed dans un premier temps, puis sur MIMIC III (une base de données contenant les dossiers médicaux électroniques des patients). Pour cette étape, nous avons choisi le modèle SBio_ClinicalBERT. Par rapport à d'autres modèles, ce modèle fournit de bons résultats pour la comparaison des labels. Cela est dû au fait que ce modèle est entraîné sur les notes cliniques de MIMIC III.

6.3 Processus de mise en correspondance

Pour trouver les similarités entre les noms de maladies et les métadonnées, nous avons procédé de différentes manières. Dans un premier temps, nous avons pris seulement les noms de maladies à partir de l'ontologie DOID ($rdfs : label$) et avons calculé les similarités entre ces éléments et les métadonnées de l'ontologie DRON ($obo : IAO_0000115$).

Nous avons ensuite amélioré notre processus en considérant deux approches qui prennent en compte d'autres éléments

1. https ://bioportal.bioontology.org/ontologies/DOID
2. https ://bioportal.bioontology.org/ontologies/DRON)
3. http ://purl.obolibrary.org/obo/
4. http ://purl.obolibrary.org/obo/$IAO_0000115$

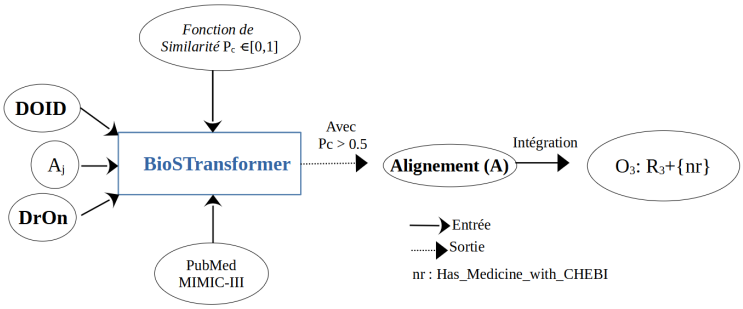

FIGURE 2 – Alignement de DOID et de DRON en utilisant BioSTransformers.

de DOID :

- La première consiste à *concaténer* plusieurs éléments de l'ontologie DOID. Ces éléments correspondent : au nom de la maladie ($rdf : label$), à sa définition ($obo : IAO_0000115$) et à plusieurs noms de maladies connexes ($oboInOwl : hasExactSynonym$). Nous appelons cette stratégie "multi-label". La concaténation est considérée comme une entrée pour BioSTransformers.
- La deuxième consiste à exploiter un seul élément à la fois à partir de DOID. Plus précisément, nous prenons en compte à chaque calcul d'une similarité soit le nom de la maladie ($rdf : label$) ou bien la définition de la maladie ($obo : IAO_0000115$) ou encore un seul nom de maladie connexe ($oboInOwl : hasExactSynonym$). Nous appelons cette stratégie "max-label". Ainsi, pour chaque élément de DRON pris en considération par BioSTransformers, la correspondance est établie avec un élément de DOID, en choisissant le score de similarité maximal résultant entre la métadonnée provenant de DRON ($obo : IAO_0000115$) et l'une des métadonnées de DOID ($rdf : label$ ou $oboInOwl : hasExactSynonym$ ou bien $obo : IAO_0000115$). Ce score doit être supérieur à 0,5.

6.4 La phase d'alignement

Les alignements générés sont des correspondances entre un seul concept de DOID et un seul concept de DRON (*alignement one-to-one*). Le type de correspondance est une *inclusion* entre les métadonnées qui définissent une classe ChEBI et celles qui définissent une maladie. Cet alignement est maintenu lorsque le score de confiance (le score de similarité) est supérieur au seuil de 0,5.

Si un alignement existe alors une nouvelle relation est définie entre la maladie et le concept ChEBI. Cette nouvelle relation permet de générer une troisième ontologie (ontologie d'intégration) enrichie par les ontologies DRON et DOID. Nous dénommons cette relation *Has_Medicine_with_CHEBI*. La Figure 2 illustre comment les BioSTransformers sont utilisés dans la tâche d'alignement d'ontologie.

Le nombre d'alignements générés par les trois approches est illustré dans la Table 2. Nous constatons que la troisième approche produit le plus grand nombre d'alignements. Ainsi, le nom de la maladie n'est pas aussi représentatif que les autres métadonnées.
Les résultats obtenus sont très encourageants lors de l'utilisation de BioSTransformers pour trouver une similarité. Par exemple, dans DRON, l'élément "*CHEBI_27779*", qui compose le médicament sous le nom de "*Griseofulvin*", est défini par la métadonnée "*An oxaspiro compound produced by Penicillium griseofulvum. It is used by mouth as an antifungal drug for infections involving the scalp, hair, nails and skin that do not respond to topical treatment"*–"Un composé oxaspiro produit par Penicillium griseofulvum. Il est utilisé par voie orale comme médicament antifongique pour les infections du cuir chevelu, des cheveux, des ongles et de la peau qui ne répondent pas au traitement topique". Dans DOID, la maladie "*DOID_3136*" est définie par la métadonnée *"scalp dermatosis"*–"dermatose du cuir chevelu". Le processus de correspondance donne un score de similarité de 0,5608. Étant donné que le score de confiance est supérieur à 0,5 (seuil défini sans expérimentation), nous créons une nouvelle relation *"Has_Medicine_with_CHEBI(DOID_3136, CHEBI_27779)"*. Toutes les nouvelles relations peuvent être récupérées via une simple requête SPARQL.
La prochaine étape nécessaire à la validation des alignements consistera à les évaluer en utilisant des méthodes structurelles fondées sur les hiérarchies de concepts existantes dans chacune des deux ontologies utilisées. Une autre approche possible est l'utilisation du Metathesaurus de l'UMLS [5] (Unified Medical Language System) en tant que ressource d'évaluation.
Initialement, nous avons utilisé l'approche "multi-label". Ensuite, pour chaque maladie, nous avons cherché son médicament correspondant dans l'UMLS en utilisant son identifiant (CUI) et l'API UMLS [6], où le code est le CUI et nous avons récupéré son traitement par la relation sémantique "may_be_treated_by". Seules 490 maladies étaient associées à des codes CUI, tandis que pour les autres maladies, des codes alternatifs étaient nécessaires mais nous ne les avons pas utilisés ici.
Pour les 490 maladies ayant des codes CUI, nous avons cherché leurs médicaments correspondants dans l'UMLS et nous avons constaté que seuls 131 d'entre eux avaient la relation sémantique "may_be_treated_by" dans le réseau sémantique de l'UMLS. Nous avons également constaté dans l'UMLS que plusieurs maladies avaient le même médicament suggéré par nos modèles.
En analysant les résultats de manière plus approfondie et en examinant les définitions de chaque médicament, nous avons découvert que les incohérences étaient dues à notre modèle qui suggérait une entité chimique faisant partie de la composition du médicament. Cela est dû au fait que nous avons effectué l'alignement en utilisant DRON, qui est basé sur ChEBI, une ontologie d'entités chimiques.
En conclusion, nous pouvons affirmer que nos alignements sémantiques sont corrects même si le médicament suggéré ne correspond pas à celui fourni par l'UMLS. Cependant, notre méthode peut aider à enrichir l'UMLS avec des relations sémantiques supplémentaires qui n'y figurent pas encore.

Méthode	Nom de la maladie	multi-label	max-label
Nombre d'alignements	615	770	**1035**

TABLE 2 – Nombre d'alignements générés pour chaque mode de mise en correspondance.

7 Conclusion

Dans cette étude, nous avons proposé de nouveaux modèles siamois [14] BioSTransformers et BioS-MiniLM qui permettent d'améliorer différentes tâches biomédicales dans une configuration sans exemples (zero shot). Ces modèles plongent des paires de textes dans le même espace de représentation et calculent la similarité sémantique entre des textes de différentes longueurs.
En outre, nous avons proposé d'exploiter nos modèles dans un scénario pratique qui consiste à aligner des entités de deux ontologies biomédicales distinctes afin d'établir de nouvelles relations.
L'approche a été instanciée dans un premier temps entre les deux ontologies DOID et DRON, dans le but de proposer un médicament potentiel pour une maladie donnée. Une évaluation des alignements obtenus est en cours et compte-tenu des premiers résultats, l'intégration d'autres ontologies (par exemple, les effets indésirables liés aux médicaments, ou d'autres ressources de médicaments comme DrugBank) est prévue. Enfin, la validation de l'approche proposée pour en démontrer sa généricité sur des ressources en Français est également envisagée.

Références

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[2] Watson Wei Khong Chua and Jung jae Kim. Boat : Automatic alignment of biomedical ontologies using term informativeness and candidate selection. *Journal of Biomedical Informatics*, 45(2) :337–349, 2012.

[3] Kirill Degtyarenko, Paula Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickael Guedj, and Michael Ashburner. Chebi : A database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36 :D344–50, 02 2008.

5. https ://www.nlm.nih.gov/research/umls/index.html
6. https ://uts-ws.nlm.nih.gov/rest/content/current/CUI/code/relations ?includeAdditionalRelationLabels=may_be_treated_by&apiKey

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[5] Jérôme Euzenat, Pavel Shvaiko, et al. *Ontology matching*, volume 18. Springer, 2007.

[6] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1) :1–23, jan 2022.

[7] Josh Hanna, Eric Joseph, Mathias Brochhausen, and William Hogan. Building a drug ontology based on rxnorm and other sources. *Journal of biomedical semantics*, 4 :44, 12 2013.

[8] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv :1705.00652*, 2017.

[9] Sven Hertling, Jan Portisch, and Heiko Paulheim. Matching with transformers in melt. 09 2021.

[10] Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. BioELECTRA : Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online, June 2021. Association for Computational Linguistics.

[11] Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. Deepalignment : Unsupervised ontology matching with refined word vectors. In *Proceedings of NAACL-HLT, 787–798.*, pages 787–798, 2018.

[12] Schriml Lynn, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Kibbe. Disease ontology : A backbone for disease semantic integration. *Nucleic acids research*, 40 :D940–6, 11 2011.

[13] Mélissa Mary, Lina Soualmia, Xavier Gansel, Stéfan Darmoni, Daniel Karlsson, and Stefan Schulz. Ontological representation of laboratory test observables : Challenges and perspectives in the snomed ct observable entity model adoption. pages 14–23, 05 2017.

[14] Safaa Menad, Saïd Abdeddaïm, and Lina Fatima Soualmia. Biostransformers : Modèles de langage pour l'apprentissage sans exemple dans des textes biomédicaux. In Catherine Faron and Sabine Loudcher, editors, *Extraction et Gestion des Connaissances, EGC 2023, Lyon, France, 16 - 20 janvier 2023*, volume E-39 of *RNTI*, pages 409–416. Editions RNTI, 2023.

[15] Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. Normalized names for clinical drugs : RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4) :441–448, 04 2011.

[16] Inès Osman, Sadok Ben Yahia, and Gayo Diallo. Ontology integration : Approaches and challenging issues. *Information Fusion*, 71 :38–63, Jul 2021.

[17] Yifan Peng, Shankai Yan, and Zhiyong Lu. 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*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics.

[18] Jan Portisch, Michael Hladik, and Heiko Paulheim. Background knowledge in ontology matching : A survey. *Semantic Web*, pages 1–55, 09 2022.

[19] Nils Reimers and Iryna Gurevych. Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *Proceedings of (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[20] Pavel Shvaiko and Jérôme Euzenat. Ontology matching : State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25 :158–176, 2013.

[21] Javier Vela and Jorge Gracia. Cross-lingual ontology matching with cider-lm : results for oaei 2022. 2022.

[22] Jifang Wu, Jianghua Lv, Haoming Guo, and Shilong Ma. Daeom : A deep attentional embedding approach for biomedical ontology matching. *Applied Sciences*, 10(21), 2020.

[23] Antoine Zimmermann and Jérôme Euzenat. Three semantics for distributed systems and their relations with alignment composition. In *The Semantic Web - ISWC 2006*, pages 16–29, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.