Nouveaux réseaux neuronaux profonds pour l'alignement d'ontologies

Ingénierie des Connaissances PFIA 2023

Safaa MENAD, Wissame LADDADA, Saïd ABDEDDAIM, Lina F. SOUALMIA

04 Juillet 2023

Plan

1 Introduction

- Introduction
- Méthode proposée
- Alignement d'ontologies
- Conclusion

Contexte

1 Introduction

- L'**alignement d'ontologies** s'appuie souvent sur des approches lexicales (occurrences des termes, etc.).
  Exemple: disease $\neq$ illness

Contexte

1 Introduction

- L'**alignement d'ontologies** s'appuie souvent sur des approches lexicales (occurrences des termes, etc.).
  Exemple: disease $\neq$ illness
- Comparaison contextuelle plutôt que caractéristiques lexicales grâce aux **modèles de langage** basés sur les transformeurs.

Contexte

1 Introduction

- L'**alignement d'ontologies** joue un rôle crucial dans l'intégration des connaissances.

Contexte

1 Introduction

- L'**alignement d'ontologies** joue un rôle crucial dans l'intégration des connaissances.
- **Correspondance d'ontologies (Ontology Matching - OM)** : identifier les **similarités** entre les ontologies.

Contexte

1 Introduction

- L'**alignement d'ontologies** joue un rôle crucial dans l'intégration des connaissances.
- **Correspondance d'ontologies (Ontology Matching - OM)** : identifier les **similarités** entre les ontologies.
- **Correspondance lexicale** comme base, combinée avec la correspondance structurelle.

Contexte

- L'**alignement d'ontologies** joue un rôle crucial dans l'intégration des connaissances.
- **Correspondance d'ontologies (Ontology Matching - OM)** : identifier les **similarités** entre les ontologies.
- **Correspondance lexicale** comme base, combinée avec la correspondance structurelle.
- **L'apprentissage automatique** est une alternative (Fine-TOM (Hertling et al., 2021), DAEOM (Wu et al., 2020), et CIDER-LM (Vela et al., 2022)).

Objectif

- Les **modèles de langage basés sur des transformeurs pré-entraînés** sont efficaces pour prédire la **similarité sémantique**.

Objectif

- Les **modèles de langage basés sur des transformeurs pré-entraînés** sont efficaces pour prédire la **similarité sémantique**.
- L'**abondance** des données biomédicales a rendu possible l'entraînement de ces modèles sur des corpus biomédicaux.

Objectif

- Les **modèles de langage basés sur des transformeurs pré-entraînés** sont efficaces pour prédire la **similarité sémantique**.
- L'**abondance** des données biomédicales a rendu possible l'entraînement de ces modèles sur des corpus biomédicaux.
- Un **affinage** sur des données **précises, supervisées** et **rarement disponibles**.

Objectif

- Les **modèles de langage basés sur des transformeurs pré-entraînés** sont efficaces pour prédire la **similarité sémantique**.
- L'**abondance** des données biomédicales a rendu possible l'entraînement de ces modèles sur des corpus biomédicaux.
- Un **affinage** sur des données **précises, supervisées** et **rarement disponibles**.
- Nouveaux modèles neuronaux siamois qui optimisent une fonction d'apprentissage contrastif auto-supervisé sur des articles biomédicaux.

Plan

2 Méthode proposée

- Introduction
- Méthode proposée
- Alignement d'ontologies
- Conclusion

Modèles siamois

2 Méthode proposée

- Les transformeurs de **paires** de phrases **(sentence transformers)** ont été développés pour la tâche de calcul de similarité entre deux phrases
  - score (SARS-CoV-2, covid) =0.51
  - score(SARS-CoV-2, insomnie) = 0.05

Modèles siamois

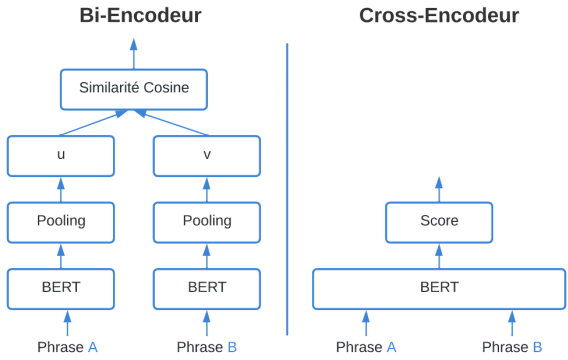
- Les transformeurs de **paires** de phrases **(sentence transformers)** ont été développés pour la tâche de calcul de similarité entre deux phrases
  - score (SARS-CoV-2, covid) =0.51
  - score(SARS-CoV-2, insomnie) = 0.05
- Ils sont utilisés pour la recherche d'informations, la reformulation de phrases etc.

Modèles siamois

- Les transformeurs de **paires** de phrases **(sentence transformers)** ont été développés pour la tâche de calcul de similarité entre deux phrases
  - score (SARS-CoV-2, covid) =0.51
  - score(SARS-CoV-2, insomnie) = 0.05
- Ils sont utilisés pour la recherche d'informations, la reformulation de phrases etc.
- **Sentence-BERT** (Reimers et Gurevych, 2019) est un exemple de transformeurs de paires de phrases

Modèles siamois

2 Méthode proposée



Il existe deux types d'architectures :

- Modèles **siamois bi-encodeurs**
- Modèles cross-encodeurs

Modèles proposés

Nouveaux modèles **siamois pré-entraînés** (Menad et al. 2023) sur le corpus PubMed:

- Les transformeurs siamois encodent des phrases de **tailles similaires** en vecteurs.

Modèles proposés

Nouveaux modèles **siamois pré-entraînés** (Menad et al. 2023) sur le corpus PubMed:

- Les transformeurs siamois encodent des phrases de **tailles similaires** en vecteurs.
- Encoder des textes et des termes de **tailles différentes** en vecteurs:
  - textes : les **titres** et les **résumés** des articles PubMed
  - termes : les termes **MeSH**

Modèles proposés

Nouveaux modèles **siamois pré-entraînés** (Menad et al. 2023) sur le corpus PubMed:

- Les transformeurs siamois encodent des phrases de **tailles similaires** en vecteurs.
- Encoder des textes et des termes de **tailles différentes** en vecteurs:
  - textes : les **titres** et les **résumés** des articles PubMed
  - termes : les termes **MeSH**
- Articles du corpus PubMed avec leurs termes clés MeSH associées comme paires **(texte, terme)** en entrée de nos modèles.

Modèles proposés

- Dans cette étude, nous présentons une nouvelle variante de nos modèles BioSTransformers (Menad et al. 2023), appelée **SBio_ClinicalBERT**.
- Nous exploitons ce modèle dans un scénario pratique qui implique l'alignement d'ontologies biomédicales.

Plan

3 Alignement d'ontologies

- Introduction
- Méthode proposée
- Alignement d'ontologies
- Conclusion

Notions théoriques

3 Alignement d'ontologies

- Une ontologie O_i est un ensemble de **vocabulaires** définis au moyen de taxonomies pour décrire un **domaine** d'intérêt donné.

Notions théoriques

3 Alignement d'ontologies

- Une ontologie O_i est un ensemble de **vocabulaires** définis au moyen de taxonomies pour décrire un **domaine** d'intérêt donné.
- Un vocabulaire est considéré comme un ensemble d'éléments $e_i =< C_i, R_i, I_i >$; avec :
  - C_i étant l'ensemble de **concepts**,
  - R_i rassemblant les **relations** pour relier les concepts,
  - I_i regroupant l'ensemble des **instances** pour interpréter les concepts et les relier avec R_i.

Processus d'alignement

- Une **correspondance** décrit la **relation** entre deux ontologies.

Processus d'alignement

- Une **correspondance** décrit la **relation** entre deux ontologies.
- Une correspondance A est un ensemble de **triplets**.

Processus d'alignement

- Une **correspondance** décrit la **relation** entre deux ontologies.
- Une correspondance A est un ensemble de **triplets**.
- Chaque triplet est spécifié par la **relation binaire** $r(e_1, e_2)$; où r représente la relation entre les deux éléments $e_1 \in O_1$ et $e_2 \in O_2$.

Processus d'alignement

- Une **correspondance** décrit la **relation** entre deux ontologies.
- Une correspondance A est un ensemble de **triplets**.
- Chaque triplet est spécifié par la **relation binaire** $r(e_1, e_2)$; où r représente la relation entre les deux éléments $e_1 \in O_1$ et $e_2 \in O_2$.
- Un **score de confiance** c évalue la similarité entre e_1 et e_2 (par exemple, la valeur de $c \in [0,1]$).

Processus d'alignement

3 Alignement d'ontologies

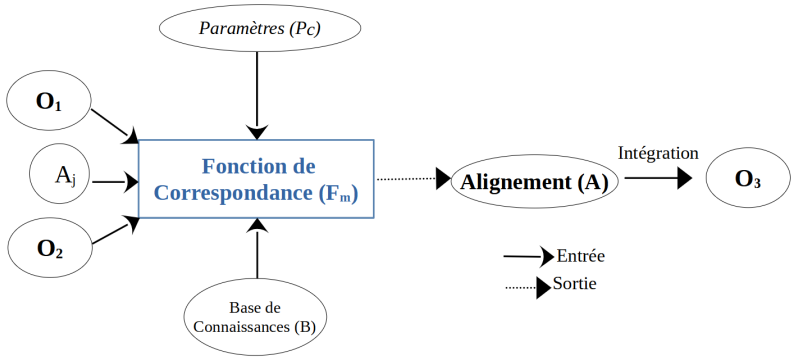


Figure: Processus d'alignement d'ontologies.

Processus d'alignement

3 Alignement d'ontologies

- Le **processus de mise en correspondance** comme un **problème de similarité**

[1] https://bioportal.bioontology.org/ontologies/DOID
[2] https://bioportal.bioontology.org/ontologies/DRON

Processus d'alignement

- Le **processus de mise en correspondance** comme un **problème de similarité**
- Mise en correspondance des éléments de deux ontologies biomédicales : **DOID (Human Disease Ontology)**[1] et **DrOn (Drug Ontology)**[2]

[1] `https://bioportal.bioontology.org/ontologies/DOID`
[2] `https://bioportal.bioontology.org/ontologies/DRON`

Phases d'alignement

La phase de prétraitement :

- (i) les **Classes** en tant qu'éléments de DOID qui **définissent** une maladie (Espace des noms de maladies[3])

[3] `http://purl.obolibrary.org/obo/`

Phases d'alignement

La phase de prétraitement :

- (i) les **Classes** en tant qu'éléments de DOID qui **définissent** une maladie (Espace des noms de maladies[3])
- (ii) les **Métadonnées** des Entités Chimiques d'Intérêt Biologique (ChEBI) à partir desquelles les médicaments sont principalement composés dans DrOn.

[3] `http://purl.obolibrary.org/obo/`

Phases d'alignement

La phase de mise en correspondance :

- Le modèle BioSTransformers est une **fonction de correspondance**
- Les connaissances de base représentent les **données** sur lesquelles le modèle est **entraîné** : d'abord sur **PubMed**, puis sur **MIMIC III**

Phases d'alignement

La phase de mise en correspondance :

- Le modèle BioSTransformers est une **fonction de correspondance**
- Les connaissances de base représentent les **données** sur lesquelles le modèle est **entraîné** : d'abord sur **PubMed**, puis sur **MIMIC III**
- CHoix du modèle **SBio_ClinicalBERT**.

Phases d'alignement

- Seulement les **noms de maladies** à partir de l'ontologie DOID (rdfs : label)
- **Autres éléments** de DOID :
  - **"multi-label"** : concaténer plusieurs éléments de l'ontologie DOID.
    - nom de la maladie (rdf : label),
    - à sa définition (obo : IAO0000115)
    - et à plusieurs noms de maladies connexes (oboInOwl : hasExactSynonym).
  - **"max-label"** : exploiter un seul élément à la fois à partir de DOID.
    - nom de la maladie (rdf : label)
    - ou la définition de la maladie (obo : IAO0000115)
    - ou encore un seul nom de maladie connexe (oboInOwl : hasExactSynonym).

Phases d'alignement

3 Alignement d'ontologies

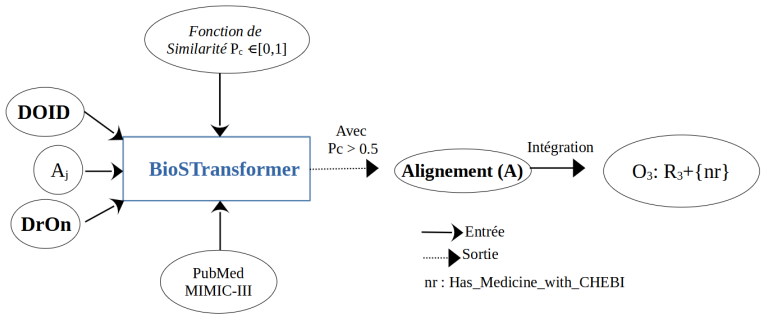


Figure: Alignement de DOID et de DRON en utilisant BioSTransformers.

Phases d'alignement

La phase d'alignement :

- Les correspondances générées sont des **correspondances un-à-un**

Phases d'alignement

La phase d'alignement :

- Les correspondances générées sont des **correspondances un-à-un**
- Le type de correspondance est une ***inclusion*** ($\sqsubseteq$)

Phases d'alignement

La phase d'alignement :

- Les correspondances générées sont des **correspondances un-à-un**
- Le type de correspondance est une ***inclusion*** **(⊑)**
- Le **score de confiance** (le score de similarité) est **supérieur** au seuil de 0,5

Phases d'alignement

La phase d'alignement :

- Les correspondances générées sont des **correspondances un-à-un**
- Le type de correspondance est une ***inclusion*** **(⊑)**
- Le **score de confiance** (le score de similarité) est **supérieur** au seuil de 0,5
- Une nouvelle relation ***Has_Medicine_with_CHEBI*** est définie

Résultats

3 Alignement d'ontologies

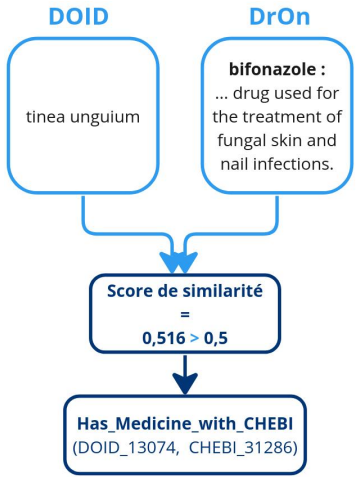
Nombre d'alignements générés par les trois approches :

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

Figure: Nombre d'alignements générés pour chaque mode de mise en correspondance.

Le nom de la maladie n'est pas aussi représentatif que les autres métadonnées.

615 correspondances avec un score de confiance supérieur à 0,5



Évaluation

- Méthodes **structurelles** basées sur les hiérarchies de concepts existantes dans chacune des deux ontologies utilisées.

[4] `https://uts-ws.nlm.nih.gov/rest/content/current/CUI/code/relations?includeAdditionalRelationLabels=may_be_treated_by&apiKey`

Évaluation

- Méthodes **structurelles** basées sur les hiérarchies de concepts existantes dans chacune des deux ontologies utilisées.
- **Métathésaurus UMLS (Unified Medical Language System)** en utilisant l'**API** UMLS [4].

[4] `https://uts-ws.nlm.nih.gov/rest/content/current/CUI/code/relations?includeAdditionalRelationLabels=may_be_treated_by&apiKey`

Plan

4 Conclusion

- Introduction
- Méthode proposée
- Alignement d'ontologies
- Conclusion

Conclusion et perspectives

4 Conclusion

- Nouveaux modèles **siamois** BioSTransformers et BioS-MiniLM capables de résoudre plusieurs **tâches biomédicales** sans exemple **(zero-shot)**

Conclusion et perspectives

4 Conclusion

- Nouveaux modèles **siamois** BioSTransformers et BioS-MiniLM capables de résoudre plusieurs **tâches biomédicales** sans exemple **(zero-shot)**
- Exploiter nos modèles dans la **mise en correspondance d'entités** de deux ontologies biomédicales distinctes

Conclusion et perspectives

4 Conclusion

- Nouveaux modèles **siamois** BioSTransformers et BioS-MiniLM capables de résoudre plusieurs **tâches biomédicales** sans exemple **(zero-shot)**
- Exploiter nos modèles dans la **mise en correspondance d'entités** de deux ontologies biomédicales distinctes

Perspectives :

- Une évaluation plus poussée des correspondances obtenues
- L'intégration d'autres ontologies (par exemple, les événements indésirables liés aux médicaments)

Références

4 Conclusion

- Menad, S., Abdeddaim, S., Soualmia, L.F.: BioSTransformers: Modèles de langage pour l'apprentissage sans exemple dans des textes biomédicaux. In: Faron, C., Loudcher, S. (eds.) Extraction et Gestion des Connaissances, EGC 2023, Lyon, France, 16 - 20 janvier 2023. RNTI, vol. E-39, pp. 409–416. Éditions RNTI (2023), http://editions-rnti.fr/?inprocid=1002844
- Hertling, S., Portisch, J., Paulheim, H.: Matching with transformers in melt (09 2021)
- Wu, J., Lv, J., Guo, H., Ma, S.: Daeom: A deep attentional embedding approach for biomedical ontology matching. Applied Sciences 10(21) (2020)
- Vela, J., Gracia, J.: Cross-lingual ontology matching with cider-lm: results for oaei 2022 (2022)
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)
- Peng, Y., Yan, S., Lu, Z.: 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. pp. 58–65 (2019)
- Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of (EMNLP-IJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
- Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910 (2021)
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in Neural Information Processing Systems 33, 5776–5788 (2020)
- Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316 (2022)
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare 3(1), 1–23(jan 2022)

Q&A

Merci de votre attention !
Des questions ?

safaa.menad1@univ-rouen.fr