BioSTransMatch Results @ OAEI 2024

Safaa Menad[1,*], Saïd Abdeddaïm [1] and Lina F. Soualmia[1]

[1]Univ Rouen Normandie, Normandie Univ, LITIS UR 4108 F-76000 Rouen, France

Abstract

This paper aims at presenting the results obtained by BioSTransMatch at the OAEI 2024 competition, marking its first participation in this event. In this context, we applied the model BioSTransformers we developed to the equivalence matching task in the Bio-ML track of the OAEI challenge. The model is founded on sentence transformers, that have recently achieved remarkable results in ontology matching tasks. Here, we leverage a transformer-based language model to identify similarities between ontology concepts. BioSTransformers is a siamese neural model that we developed and trained using biomedical scientific articles from PubMed. It embeds texts into a vector space to compare and identify similarities. The model optimizes a self-supervised contrastive learning objective using articles from the MEDLINE bibliographic database and their associated MeSH (Medical Subject Headings) keywords.

Keywords

Biomedical Ontologies, OAEI 2024, Siamese Transformers, Ontology Matching

1. Related Work

Ontology matching (OM) tasks have evolved significantly, incorporating a wide range of approaches, from traditional methods [1, 2] to cutting-edge transformer-based techniques [3, 4, 5].

Transformer-based approaches have successfully overcome a fundamental limitation of the existing traditional methods: the inability to effectively take into account account contexts and synonyms when comparing entities. These transformer-based methods can be categorized into two main types. The first category includes unsupervised learning methods, such as [4, 3], which utilize embeddings to capture semantic similarities. The second category involves supervised methods, which fine-tune models to improve performance, as seen in works like [5, 6].

More recently, large language model (LLM)-based methods have emerged as a promising direction. For instance, OLaLa [7] utilized an SBERT model to extract top-k matches from target ontologies, followed by an LLM to generate the alignments. LLMs4OM [8] is an end-to-end framework that employs a RAG approach to retrieve additional context from a knowledge base. The LLM is then queried using the retrieved information.

2. Presentation of the System

2.1. State, Purpose, General statement

In recent years, ontology matching has garnered significant attention within various representation learning systems. Particularly with advancements in machine learning, language model-based approaches, such as [6], have been increasingly applied to this task. Although these systems have been successfully used in ontology matching within general domains, the biomedical one requires special consideration due to the caracteristics of the clinical and medical language.

OM2024: The 19th International Workshop on Ontology Matching collocated with the 23rd International Semantic Web Conference (ISWC-2024), November 11th, Baltimore, USA

*Corresponding author.

safaa.menad1@univ-rouen.fr (S. Menad); said.abdeddaim@univ-rouen.fr (S. A.); fatima.soualmia@univ-rouen.fr (L. F. Soualmia)

0009-0009-2204-7786 (S. Menad); 0000-0002-7521-7955 (S. A.); 0000-0001-7668-2819 (L. F. Soualmia)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3897/oaei2024_paper6.pdf

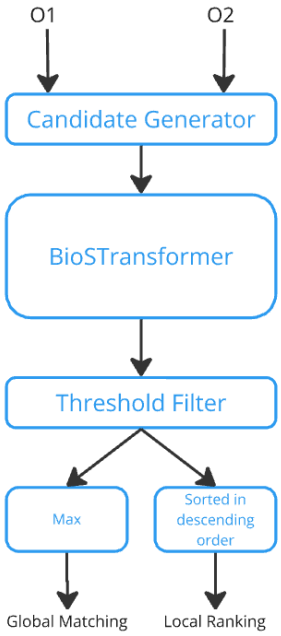

Figure 1: General view of BioSTransMatch.

BioSTransformers is a transformer-based model built on a biomedical model [9]. It constructs embeddings using the pretrained biomedical transformer, *Bio_ClinicalBERT*. This method has demonstrated that training on biomedical data extracted from the PubMed database can lead to better vector representations compared to other embedding models trained on general data. The purpose of our system is to leverage these embeddings to create an ontology matcher that offers a simple, fast, and generalizable approach.

Figure 1 presents the overall architecture of BioSTransMatch.

2.2. Candidate Generation

The first step of the matching process consists in determining the concepts to match. In the Bio-ML track, some classes have the *use_in_alignment* property, which indicates whether they should be used in the matching process. We have excluded all candidate concepts that have this property.

For the local ranking sub-task, candidate mappings are suggested from a test.cands file. Therefore, we utilize candidates from this list.

We utilized DeepOnto [10], a Python library designed to facilitate ontology engineering with deep learning techniques. DeepOnto encapsulates basic ontology processing functions from the OWL API and implements several essential components such as reasoning, verbalization, pruning, and taxonomy, etc.

2.3. BioSTranformers Embeddings

The goal of BioSTransformers is to obtain rich biomedical embeddings. Siamese transformers were originally designed to transform similarly sized sentences into vectors.

Sentence-BERT [11] is a BERT-based bi-encoder designed to generate semantically meaningful sentence embeddings for use in textual similarity comparisons. For each input, the model produces a fixed-size vector (u and v). The objective function is designed such that the angle between the two vectors u and v is smaller when the inputs are similar.

In our approach, we propose transforming MeSH terms, titles, and abstracts of PubMed articles into the same vector space by training a siamese transformer model on these data. To ensure compatibility

between short and long texts within this vector space, we trained our models using pairs of inputs such as (title, MeSH term) and (abstract, MeSH term).

We use a self-supervised contrastive learning objective function based on the Multiple Negative Ranking Loss (MNRL) function. The MNRL only needs positive pairs as input (the title (or abstract) and a MeSH term associated with the article in our case). For a positive pair (title i or abstract i, MeSH i), MNRL considers that each pair (title i or abstract i, MeSH j) with $i = j$ in the same batch is negative. Since an article can be associated with several MeSH terms, we ensured in the batch generation that an abstract (or title) associated with a MeSH term in PubMed is never taken as a negative pair.

2.4. Compute Similarities

Using our embeddings, we construct a similarity matrix founded on the cosine similarity measure as shown in Equation 1.

$$cos(u, v) = \frac{u.v}{|u||||v||} \tag{1}$$

In the matrix, rows represent concepts from the source ontology, and columns represent concepts from the target ontology. The matrix values are filled with the cosine similarities computed by BioS-Transformers.

2.5. Threshold filter

In the next step, we filter out all scores below a threshold, effectively eliminating all correspondences with low confidence. All correspondences with a score below are excluded. We set the threshold at 0.75, which yielded good results for the evaluated tasks.

3. Results

3.1. Bio-ML Track

The Bio-ML track [1] consists of five different pairs of datasets and includes both an equivalence matching task and a subsumption matching task. BioSTransformers participates in the equivalence matching task only.

The ontologies of this track are the OMIM (Online Mendelian Inheritance in Man), ORDO (Orphanet Rare Disease Ontology), NCIT (National Cancer Institute Thesaurus), DOID (Human Disease Ontology), FMA (Foundational Model of Anatomy), and SNOMED CT (Systematized Nomenclature of Medecine - Clinical Terms).

- OMIM desribes genes, genetic phenotypes, and gene-phenotype relations, generated through manual curation based on biomedical literature [12];
- ORDO is a classification of rare diseases and relationships between diseases, genes, and epidemiologic features [13];
- NCIT is an ontology on cancer-related concepts [14];
- DOID describes human diseases [15];
- FMA represents a coherent body of explicit declarative knowledge about the human anatomy [16].
- SNOMED CT is a structured clinical terminology that includes a vast collection of medical concepts, relationships, and terms to accurately represent clinical findings, procedures, and medications [17].

The equivalence matching task is further divided into two categories: an unsupervised setting, and a semi-supervised setting, where 30% of the reference alignments are provided in the training set.

[1] https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/

3.2. Unsupervised Setting

In this setting, we have used our model following the steps of the overall architecture in Figure 1. Table 1 shows the results on the different tasks.

Global Matching To find the mappings between concepts for the global matching task, we select the element in each row that represents the maximum similarity among the column elements. Specifically, we choose the first element that has the highest similarity.

Local Ranking In this step, we take all candidates and sort them in a decreasing order.

Table 1
Bio-ML track evaluation results for the equivalence matching task for the unsupervised setting.

Task	P	R	F1	MRR	H@1
OMIM-ORDO	0.312	0.586	0.407	0.7408	0.6825
NCIT-DOID	0.657	0.833	0.735	0.9001	0.8646
SNOMED-FMA	0.128	0.384	0.192	0.6326	0.5125
SNOMED-NCIT (Pharm)	0.584	0.844	0.690	0.9431	0.9182
SNOMED-NCIT (Neoplas)	0.289	0.663	0.402	0.8463	0.7893

3.3. Semi-supervised Setting

The supervised model requires a dataset of both positive and negative alignments during training. This dataset consists of reference alignments as well as generated positive and negative alignments.

Global Matching First, we trained our model on the training data from all tasks to fine-tune the model across the entire dataset. The training was conducted using cosine similarity, with reference alignments as positive samples and generating negative ones from the list of references.

The second step involves training a neural classifier to select the best alignment from a list of candidates. To achieve this, we used the fine-tuned model to generate similarity scores for each candidate from the source data, and the trained classifier to choose the optimal alignment from the list of candidates.

Local Ranking As in the local ranking process for the unsupervised setting, all candidates are considered and ranked. However, in this case, the ranking is based on the scores generated by the fine-tuned model. Table 2 presents the results for the different tasks.

Table 2
Bio-ML track evaluation results for the equivalence matching task for the semi-supervised setting

Task	P	R	F1	MRR	H@1	H@5
OMIM-ORDO	0.973	0.278	0.432	0.737	0.672	0.811
NCIT-DOID	0.698	0.741	0.719	0.906	0.872	0.943
SNOMED-FMA	0.357	0.661	0.464	0.855	0.798	0.924
SNOMED-NCIT (Pharm)	0.845	0.860	0.852	0.957	0.943	0.974
SNOMED-NCIT (Neoplas)	0.700	0.607	0.650	0.855	0.795	0.933

4. General Comments

BioSTransMatch performs well in the local ranking task within the unsupervised setting, as demonstrated in Table 1. The model achieves results comparable to state-of-the-art approaches in local ranking, even under zero-shot conditions.

However, for datasets such as SNOMED-FMA, SNOMED-NCIT (Neoplas) and OMIM-ORDO, BioSTransMatch yielded lower-than-expected outcomes. This suggests that our current strategy—selecting the alignment with the highest score—may not be optimal when identifying the correct match. To address this, we extended the model to a semi-supervised setting, allowing it to learn to select the best alignment from a list of candidates. This adjustment resulted in significantly improved performance compared to the unsupervised approach.

As potential avenues for optimizing the system, BioSTransMatch can be optimized in several ways. One key enhancement involves incorporating additional information into the alignment process, as our current method relies exclusively on label comparisons, which may not always provide sufficient detail for accurate matching.

Furthermore, investigating alternative strategies for alignment selection in the global matching task within the unsupervised setting could be advantageous. Rather than relying solely on the maximum score, exploring methods such as greedy algorithms or the Max Weight Bipartite Extractor could yield improved results. Additionally, experimenting with different threshold values for filtering could further refine the system's accuracy.

5. Conclusion

In this paper, we present the BioSTransMatch system and discuss its performance results in the BioML track of the OAEI challenge. Our evaluation showed that the model achieved moderate performance in the unsupervised mode. While BioSTransMatch delivered lower overall results in the global matching task, it performed well in the local ranking task.

In the semi-supervised setting, the system achieved significantly better results, highlighting the potential of supervised learning in enhancing the alignment process. Additionally, our findings suggest that leveraging textual information can be beneficial when generating correspondences between entities.

For this first OAEI participation of BioSTransMatch, the reported results motivate further research in the area of transformer-based ontology matching. The system can be greatly improved in the future, by improving the candidate generation pipeline or using more information in the entity comparison. There remains opportunities for enhancing its capabilities, we seek to a deeper analysis of ontology structures to better leverage structural context.

Future participations in OAEI also allows ongoing evaluation versus state-of-the-art matchers. Overall, the results provide motivation to push the boundaries of adaptive ontology matching.

References

[1] E. Jiménez-Ruiz, B. Cuenca Grau, Logmap: Logic-based and scalable ontology matching, in: The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10, Springer, 2011, pp. 273–288.

[2] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The agreementmakerlight ontology matching system, in: On the Move to Meaningful Internet Systems: OTM 2013 Conferences: Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings, Springer, 2013, pp. 527–541.

[3] Z. Wang, Amd results for oaei 2022., in: OM@ ISWC, 2022, pp. 145–152.

[4] D. Faria, M. C. Silva, P. Cotovio, L. Ferraz, L. Balbi, C. Pesquita, Results for matcha and matcha-dl in oaei 2023., in: OM@ ISWC, 2023, pp. 164–169.

[5] F. Gosselin, A. Zouaq, Sorbet: A siamese network for ontology embeddings using a distance-based regression loss and bert, in: International Semantic Web Conference, Springer, 2023, pp. 561–578.
[6] Y. He, J. Chen, D. Antonyrajah, I. Horrocks, Bertmap: a bert-based ontology alignment system, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 5684–5691.
[7] S. Hertling, H. Paulheim, Olala: Ontology matching with large language models, in: Proceedings of the 12th Knowledge Capture Conference 2023, 2023, pp. 131–139.
[8] H. Babaei Giglou, J. D'Souza, S. Auer, Llms4om: Matching ontologies with large language models, arXiv e-prints (2024) arXiv–2404.
[9] S. Menad, W. Laddada, S. Abdeddaïm, L. F. Soualmia, Biostransformers for biomedical ontologies alignment, in: Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2023, Volume 2: KEOD, SCITEPRESS, 2023, pp. 73–84. doi:10.5220/0012188600003598.
[10] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, B. Sapkota, Deeponto: A python package for ontology engineering with deep learning, arXiv preprint arXiv:2307.03067 (2023).
[11] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992.
[12] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders, Nucleic acids research 33 (2005) D514–D517.
[13] D. Vasant, L. Chanas, J. Malone, M. Hanauer, A. Olry, S. Jupp, P. N. Robinson, H. Parkinson, A. Rath, Ordo: an ontology connecting rare disease, epidemiology and genetic data, in: Proceedings of ISMB, volume 30, researchgate. net, 2014.
[14] N. Sioutos, S. de Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, L. W. Wright, Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information, Journal of biomedical informatics 40 (2007) 30–43.
[15] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, et al., Human disease ontology 2018 update: classification, content and workflow expansion, Nucleic acids research 47 (2019) D955–D962.
[16] C. Rosse, J. L. Mejino Jr, A reference ontology for biomedical informatics: the foundational model of anatomy, Journal of biomedical informatics 36 (2003) 478–500.
[17] K. Donnelly, et al., Snomed-ct: The advanced terminology and coding system for ehealth, Studies in health technology and informatics 121 (2006) 279.