BioSTransMatch Results @ OAEI 2025

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 presents the results achieved by BioSTransMatch in the OAEI 2025 competition, marking its second participation in the Bio-ML track. Unlike the 2024 version, which combined supervised and unsupervised learning, the 2025 edition focuses exclusively on a graph-based unsupervised approach. The proposed method exploits the hierarchical structure of SNOMED CT by generating graph-based contexts through parent–child relations. These contexts are then used to train sentence transformer models using positive and negative pairs derived from the ontology graph. We trained three different transformer backbones and evaluated their performance on the equivalence and local ranking subtasks of the Bio-ML track. The obtained results highlight the potential of structural context learning for ontology alignment without explicit supervision.

Keywords

Biomedical Ontologies, OAEI 2025, Siamese Transformers, Ontology Matching

1. Related Work

Recent advances in ontology matching (OM) have increasingly relied on deep representation learning. Traditional systems such as LogMap [1] or AgreementMakerLight [2] mainly depend on lexical and logical reasoning, whereas transformer-based models [3, 4] introduced semantic embeddings to better capture contextual information. More recently, LLM-based frameworks such as OLaLa [5] and LLMs4OM [6] have shown that large language models can be effectively integrated to generate or validate alignments.

However, these methods typically require either labeled data or external textual corpora. In contrast, the 2025 version of BioSTransMatch adopts a fully unsupervised learning strategy, leveraging only the graph topology of biomedical ontologies to generate training data. This approach combines the strengths of transformer-based embeddings with the structural richness of ontological hierarchies.

2. Presentation of the System

2.1. State, Purpose, General statement

BioSTransMatch aims to explore representation learning directly from biomedical ontology graphs. The goal is to create concept embeddings that encode not only lexical semantics but also hierarchical relationships such as is-a links, without using manually aligned pairs. This allows the matcher to adapt to new ontologies with minimal preprocessing.

2.2. Graph-based Context Generation

To introduce structural knowledge, we transformed the SNOMED CT ontology into a directed graph, where nodes represent concepts and edges represent hierarchical relations. For each node, all direct parents and children were collected to build a contextual description as follows:

```
parent of <parent_label> ; child of <child_label>
```

OM2025: The 20th International Workshop on Ontology Matching collocated with the 24th International Semantic Web Conference (ISWC-2025), November 2nd, 2025, Nara, Japan

*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)

2.3. Unsupervised Training Strategy

Each concept–context pair is treated as a positive example, while a random concept from the graph serves as a negative pair. The model is trained using a Cosine Similarity Loss, which maximizes similarity between positive pairs and minimizes it between negatives. Formally, for embeddings u and v, the similarity is computed as:

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

Training follows the standard SentenceTransformer pipeline with a batch size of 16, one training epoch, and intermediate evaluation using an embedding similarity evaluator. We used our model BioSTransformers [7, 8].

2.4. Similarity Computation and Filtering

After encoding ontology concepts into embeddings, we construct a similarity matrix based on cosine similarity. Each row corresponds to a source concept and each column to a target concept. For the global matching task, the highest-scoring target is selected for each source. A similarity threshold of 0.75 is applied to remove low-confidence correspondences (See Figure 1).

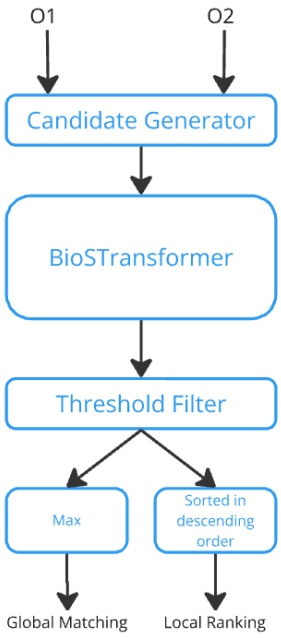

Figure 1: General view of BioSTransMatch.

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. BioSTransMatch 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),

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

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 [9];
- ORDO is a classification of rare diseases and relationships between diseases, genes, and epidemiologic features [10];
- NCIT is an ontology on cancer-related concepts [11];
- DOID describes human diseases [12];
- FMA represents a coherent body of explicit declarative knowledge about the human anatomy [13].
- 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 [14].

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.

3.2. Unsupervised Setting

In this experiment, the model was applied according to the workflow illustrated in Figure 1. Table 1 presents the results on the SNOMED–FMA task, both with and without context. The context-free (without context) configuration corresponds to the first participation of BioSTransMatch [15].

Table 1
Bio-ML track results for BioSTransMatch.

Model	Task	P	R	F1	MRR	H@1
context-free	SNOMED–FMA	0.128	0.384	0.192	0.633	0.513
With context	SNOMED–FMA	0.231(+80.47%)	0.613(+59.9%)	0.336(+75%)	0.801(+26.56%)	0.730(+42.1%)

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.

4. General Comments

Training the model with contextual information led to a slight improvement in the equivalence matching task and a notable improvement in the ranking task. This indicates that incorporating contextual information helps the model capture a richer semantic understanding between entities. While the model performs well in local ranking, it still struggles to identify exact equivalences. These results highlight the need for further work to enhance its ability to discriminate true matches among similar candidates, possibly by exploring alternative strategies for selecting the best candidate in equivalence matching rather than relying solely on the maximum score.

5. Conclusion

The 2025 edition of BioSTransMatch introduces a fully graph-driven unsupervised approach for ontology matching. By training on contexts generated from hierarchical walks within SNOMED CT, the system

learns representations that capture both textual and relational semantics. Despite the absence of supervision, the results are competitive, particularly in the ranking task.

Future work will focus on integrating multi-hop relational contexts, combining textual definitions with structural paths, and exploring contrastive graph objectives to enhance structural embeddings. This approach opens promising directions for scalable, domain-adaptive ontology matching.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

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] 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.

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

[5] S. Hertling, H. Paulheim, Olala: Ontology matching with large language models, in: Proceedings of the 12th Knowledge Capture Conference 2023, 2023, pp. 131–139.

[6] H. Babaei Giglou, J. D'Souza, S. Auer, Llms4om: Matching ontologies with large language models, arXiv e-prints (2024) arXiv–2404.

[7] 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.

[8] S. Menad, W. Laddada, S. Abdeddaïm, L. F. Soualmia, New siamese neural networks for text classification and ontologies alignment, in: International Conference on Complex Computational Ecosystems, Springer, 2023, pp. 16–29.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] K. Donnelly, et al., Snomed-ct: The advanced terminology and coding system for ehealth, Studies in health technology and informatics 121 (2006) 279.

[15] S. Menad, S. Abdeddaïm, L. F. Soualmia, Biostransmatch results in oaei 2024, in: Proceedings of

the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, volume 3897, 2024, pp. 132–137.