Results for BioGITOM in OAEI 2024

Samira Oulefki[1,*], Lamia Berkani[1], Ladjel Bellatreche[2], Nassim Boudjenah[1] and Aicha Mokhtari[1]

[1] *Dep. of Artificial Intelligence and Data Sciences, Faculty of Informatics, USTHB, Bab Ezzouar 16111, Algiers, Algeria*
[2] *LIAS/ISAE-ENSMA, Poitiers, France*

Abstract

BioGITOM is an advanced ontology matching (OM) system developed for the biomedical domain, designed to meet the increasing need for precise ontology alignment and data interoperability. By integrating Graph Isomorphism Networks and Graph Transformers, BioGITOM produces enriched concept embeddings that combine both structural and semantic information. This hybrid model enables the system to accurately identify correspondences between concepts across various ontologies, effectively addressing the challenges presented by the complexity and diversity of biomedical data. BioGITOM demonstrated outstanding performance in the Bio-ML benchmark tasks, ranking as the top system in all three tasks and outperforming eight competing methods.

Keywords

Ontology matching, deep learning, GNN, graph transformer, graph isomorphism transformer.

1. Presentation of the System

The biomedical field has seen a tremendous growth in data repositories, each containing valuable information essential for research and healthcare. However, these repositories are often semantically heterogeneous, making their integration and interoperability a significant challenge. To address this issue, ontology matching (OM) has emerged as a crucial solution, aiming to identify semantic correspondences between entities in different ontologies [1]. This process ensures that data from diverse sources can be aligned, understood, and effectively utilized for research and clinical purposes.

Traditional OM methods often rely on external lexicons, rule-based systems, or predefined heuristics to establish mappings [2, 3]. While these approaches can be useful, they are limited in handling the complexity and scale of modern biomedical ontologies.

OM-2024: The 19th International Workshop on Ontology Matching collocated with the 23rd International Semantic Web Conference (ISWC 2024), November 11th, Baltimore, USA.
* Corresponding author.
soulefki@usthb.dz (S. Oulefki) ; lberkani@usthb.dz (L. Berkani) ; bellatreche@ensma.fr (L. Bellatreche) ; boudjenah36@gmail.com (N. Boudjenah) ; mokhtariaicha33@gmail.com (A. Mokhtari)

Recent advances in learning methods, particularly those utilizing Deep Learning (DL) and Graph Neural Networks (GNNs), offer a more powerful means of extracting meaningful entity representations for OM [4, 5, 6].

In this paper, we present BioGITOM, a novel OM system specifically designed for the biomedical domain. BioGITOM enhances concept matching by integrating both semantic and structural information. It leverages BioBERT to extract semantic features and employs a Graph Isomorphism Transformer (GIT) model, which combines Graph Isomorphism Networks (GINs) [7] and Graph Transformers (GTs) [8], to capture structural relationships within ontologies. This hybrid approach allows BioGITOM to deliver highly accurate correspondences between complex biomedical concepts, meeting the growing demand for more precise and scalable OM in biomedical research and applications.

1.1. State, Purpose, General Statement

BioGITOM is a specialized OM system developed to address the increasing complexity and heterogeneity in biomedical ontologies. Its core purpose is to ensure effective integration and alignment of disparate ontologies, which is essential for improving data interoperability in biomedical research. BioGITOM is particularly designed to manage the unique challenges posed by the biomedical field, where ontologies often differ in structure and semantics. By combining advanced graph-based techniques, BioGITOM is able to produce more accurate mappings between concepts from different ontologies, thereby supporting enhanced data sharing and collaboration across systems.

1.2. Specific Techniques Used

BioGITOM leverages a sophisticated set of techniques to deliver high-precision OM by integrating both structural and semantic features of biomedical concepts. The system employs a hybrid Graph Neural Network (GNN) model, combining the strengths of Graph Isomorphism Networks (GINs) and Graph Transformers (GTs) to handle complex biomedical data. Below is a detailed breakdown of the specific techniques used:

1. *Preprocessing:* This module prepares the raw ontology data for processing. It reads input files in OWL (Ontology Web Language) format, creates RDF (Resource Description Framework) graphs, and extracts concept labels and synonyms. By doing so, it generates a rich set of terms and relationships for further processing.
2. *Concept Name Encoder:* BioGITOM leverages BioBERT [9], a pre-trained language model specialized for biomedical text, to encode the names and synonyms of ontology concepts. BioBERT captures the semantic nuances of biomedical terms, providing rich embeddings for each concept.
3. *Graph Isomorphism Transformer (GIT):* The core of BioGITOM is the Graph Isomorphism Transformer (GIT) model. This hybrid approach combines the structural expressiveness of Graph Isomorphism Networks (GINs) with the ability of Graph Transformers to capture long- range dependencies in graphs. GINs ensure that the local graph structure of the ontology is accurately captured, while GTs excel at identifying more global, non-local relationships between concepts. This combination allows the system to create rich

structural embeddings for each concept, capturing both fine-grained and broad context information about how concepts relate to each other in the ontology graph.

4. *Gating Aggregator:* The Gating Aggregator is responsible for merging the semantic and structural embeddings generated by the Concept Name Encoder and GIT, respectively. This is done through a gated mechanism [10] that dynamically balances the importance of semantic and structural information for each concept. The gating function, controlled by a learnable weight matrix and bias, determines how much semantic information versus structural information should be reflected in the final embedding. This step ensures that the final embeddings used for matching are an optimal blend of both types of information, tailored to the specific characteristics of the ontologies being compared.
5. *Mappings Selector:* The final step in BioGITOM's architecture is the Mappings Selector, which compares the merged embeddings to identify correspondences between concepts from different ontologies. A similarity measure, such as cosine similarity, is applied to determine how closely the embeddings match. The output is a set of mappings between concepts, along with confidence scores indicating the strength of each match.

1.3. Adaptations Made for the Evaluation

For this evaluation, BioGITOM was applied in its standard configuration without any task-specific modifications. This approach demonstrates the system's inherent versatility and robustness, as it was capable of achieving high performance without the need for additional customization.

The results underscore BioGITOM's effectiveness and generalizability across different OM tasks within the biomedical domain, highlighting its potential as a reliable tool for diverse applications.

1.4. Link to the System and Parameters File

BioGITOM is currently in the development phase and has not yet been released to the public. A public release is planned once the core development is finalized, ensuring that the system is fully functional and ready for broader use in OM tasks.

2. Results

BioGITOM's results for OAEI 2024 are summarized in the following sub-sections:

2.1. Performance Evaluation of BioGITOM Using OMIM-ORDO Dataset

Table 1 demonstrates that BioGITOM excels on the OMIM-ORDO dataset. The system achieves an impressive precision of 0.951, reflecting its strong capability to generate highly accurate mappings while minimizing false positives. Additionally, BioGITOM delivers a solid recall rate of 0.773, indicating its effectiveness in identifying a significant number of relevant matches. This is further supported by a well-balanced F1 score of 0.853, underscoring the system's overall accuracy and reliability.

Table 1
Results of BioGITOM on the OMIM-ORDO dataset.

Tool	P	R	F1
BioGITOM	0.951	0.773	0.853

2.2. Performance evaluation of BioGITOM using DOID-NCIT dataset

Table 2 shows that BioGITOM performs exceptionally well on the DOID-NCIT dataset, achieving a precision of 0.944 and an F1 score of 0.913. While BioGITOM's recall value (0.884) is slightly lower than the highest recall of 0.959 achieved by LogMapBio, its overall performance remains highly competitive, demonstrating a strong balance between accuracy and recall.

Table 2
Results of BioGITOM on the DOID-NCIT dataset.

Tool	P	R	F1
BioGITOM	0.944	0.884	0.913

2.3. Performance evaluation of BioGITOM using SNOMED-FMA (Body) dataset

As shown in Table 3, BioGITOM delivers outstanding performance on the SNOMED-FMA (Body) dataset, achieving the highest precision (0.962), recall (0.886), and F1 score (0.923) among all competing methods.

Table 3
Results of BioGITOM on the SNOMED-FMA (Body) dataset.

Tool	P	R	F1
BioGITOM	0.962	0.886	0.923

2.4. Performance evaluation of BioGITOM using SNOMED-NCIT (Pharm) dataset

As shown in Table 4, BioGITOM demonstrates strong performance on the SNOMED-NCIT (Pharm) dataset, achieving the highest precision (0.983). However, its recall is relatively lower at 0.713, leading to an F1 score of 0.827. Despite this, BioGITOM secures the second position overall in this dataset, reflecting its high accuracy in producing correct mappings while acknowledging room for improvement in capturing a greater number of relevant matches.

Table 4
Results of BioGITOM on the SNOMED-NCIT (Pharm) dataset.

Tool	P	R	F1
BioGITOM	0.983	0.713	0.827

3. General Comments

3.1. Comments on the Results (Strengths and Weaknesses)

The experimental results highlight the significant advantages of BioGITOM compared to other highly ranked systems. A key strength of our approach lies in the Graph Isomorphism Transformer (GIT) model, which effectively generates contextually relevant representations, enabling the system to handle the complexities of biomedical ontologies. This capability is especially valuable when working with intricate and heterogeneous ontological structures.

However, a limitation of our current approach is its focus solely on generating equivalent mappings. This narrow focus does not fully address other types of semantic relationships, such as subsumption, which may be crucial in certain OM tasks.

3.2. Discussion on Improvements for the Proposed System

To enhance the system's versatility and performance, we are actively investigating ways to expand the range of matching relationships that BioGITOM can handle, moving beyond simple equivalences to include subsumption, and other relevant relationships.

Additionally, we are exploring the transfer of concept representations into a hyperbolic integration space. This shift is motivated by the limitations of Euclidean space for hierarchical ontologies, where distortions can occur. Hyperbolic space is better suited for preserving hierarchical structures, and we believe that this transformation will significantly improve BioGITOM's accuracy and representation of complex ontological relationships [11, 12].

4. Conclusion

BioGITOM is a novel approach for biomedical OM that leverages a hybrid Graph Neural Network model, GIT, integrating the strengths of Graph Transformers (GTs) and Graph Isomorphism Networks (GINs).

Experimental results show that BioGITOM consistently outperforms competitive methods across most of the evaluated datasets, underscoring its strong ability to produce highly accurate mappings. However, the system currently focuses exclusively on generating equivalent mappings. To address this limitation, we are actively working on extending BioGITOM to handle a broader range of matching relationships, such as subsumption, which will enhance the system's versatility and applicability in more complex ontology matching scenarios.

5. Acknowledgement

We would like to extend our heartfelt thanks to Jérôme Euzenat for his efforts and support in enabling the submission of our system to the OAEI. His contributions to the initiative are greatly appreciated, and we are grateful for the opportunity to participate.

References

[1] J. Euzenat, P. Shvaiko. 2013. Ontology Matching. Springer-Verlag, Heidelberg (DE).

[2] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I.F. Cruz, F.M. Couto. 2013. The AgreementMakerLight ontology matching system. In: On the Move to Meaningful Internet Systems: OTM 2013 Conferences, Springer Berlin Heidelberg, pp. 527–541.

[3] E. Jiménez-Ruiz, B. Cuenca Grau. 2011. LogMap: Logic-based and scalable ontology matching. In: Proceedings of the International Conference on Semantic Web (ISWC 2011), Springer Berlin Heidelberg, pp. 273–288.

[4] C. Xiang, T. Jiang, B. Chang, Z. Sui. 2015. ERSOM: A structural ontology matching approach using automatically learned entity representation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 2419–2429, http://dx.doi.org/10.18653/v1/d15-1289.

[5] P. Kolyvakis, A. Kalousis, D. Kiritsis. 2018. DeepAlignment: Unsupervised ontology matching with refined word vectors. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Association for Computational Linguistics, pp. 787–798, http://dx.doi.org/10.18653/v1/n18-1072.

[6] J. Wu, J. Lv, H. Guo, S. Ma. 2020. DAEOM: A Deep Attentional Embedding Approach for Biomedical Ontology Matching. Applied Sciences, 10(21), 7909. https://doi.org/10.3390/app10217909.

[7] K. Xu, W. Hu, J. Leskovec, S. Jegelka. 2019. How powerful are graph neural networks? In: Proceedings of the 7th International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arXiv.1810.00826.

[8] E. Min, R. Chen, Y. Bian, T. Xu, K. Zhao, W. Huang, P. Zhao, J. Huang, S. Ananiadou, Y. Rong. 2022. Transformer for graphs: An overview from architecture perspective. CoRR abs/2202.08455, arXiv preprint arXiv:2202.08455.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp. 1234–1240.

[10] Y. Gu, X. Qu, Z. Wang, Y. Zheng, B. Huai, N.J. Yuan. 2022. Delving deep into regularity: A simple but effective method for Chinese named entity recognition. In: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, pp. 1863–1873, http://dx.doi.org/10.18653/v1/2022.findings-naacl.143.

[11] J. Hao, C. Lei, V. Efthymiou, A. Quamar, F. Özcan, Y. Sun, W. Wang. 2021. MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), Virtual Event, Singapore, ACM, New York, NY, USA, pp. 2946–2954, https://doi.org/10.1145/3447548.3467138.

[12] P. Wang, Y. Hu. 2022. Matching Biomedical Ontologies via a Hybrid Graph Attention Network. Front Genet. 13:893409, https://doi.org/10.3389/fgene.2022.893409.