SORBETMatcher Results for OAEI 2023

Francis Gosselin*[1], Amal Zouaq[1]

[1]LAMA-WeST Lab, Departement of Computer Engineering and Software Engineering, Polytechnique Montreal, 2500 Chem. de Polytechnique, Montréal, QC H3T 1J4, Canada

Abstract

This paper presents the results of SORBETMatcher in the OAEI 2023 competition. SORBETMatcher is a schema matching system for both equivalence matching and subsumption matching. SORBETMatcher is largely based on SORBET Embeddings, a novel ontology embedding method that leverages large language models, random walks, and a regression loss to construct a latent space that encapsulates ontology structures. Despite recognizing certain limitations inherent in SORBET Embeddings, SORBETMatcher performed well in the OAEI competition. It emerged as the leading system in three out of the five subsumption matching challenges within the Bio-ML track, as well as in the equivalence matching problem involving ORDO-DOID.

Keywords

Ontology alignment, Schema matching, Representation Learning, ISWC-2023

1. Presentation of the system

1.1. State, purpose, general statement

The ontology alignment task has seen the emergence of many representation learning systems in recent years. Particularly, systems that are based on Large Language Models (LLM) such as DAEOM [9], BERTMAP[5], SEBMatcher[3] and BERTSUBS [2] have seen positive results. While these systems have used LLMs for interpreting or encoding ontology classes and properties, there is still some work to be done on how to properly leverage LLM in Ontology Alignment. SORBETMatcher is a LLM-based Ontology matching system that explores the usage of SORBET Embeddings [4] for the task. SORBET is a novel Ontology Embedding method that aims to construct embeddings starting from pre-trained SentenceBERT embeddings [8]. The method has demonstrated that infusing structural knowledge through a regression distance-based loss can lead to better vector representation compared to other ontology embedding techniques when it comes to equivalence and subsumption matching. The purpose of SORBETMatcher is to leverage those embeddings to create an ontology matcher that offers a simple, yet powerful and generalizable approach. [1]

Figure 1 presents the overall architecture of SORBETMatcher.

OM 2023: The 18th International Workshop on Ontology Matching collocated with the 22nd International Semantic Web Conference ISWC-2023 November 7th, 2023, Athens, Greece

*Corresponding author.

francis.gosselin@polymtl.ca (F. Gosselin*); amal.zouaq@polymtl.ca (A. Zouaq)

[1]Code available at: https://github.com/Lama-West/SORBETMatcher

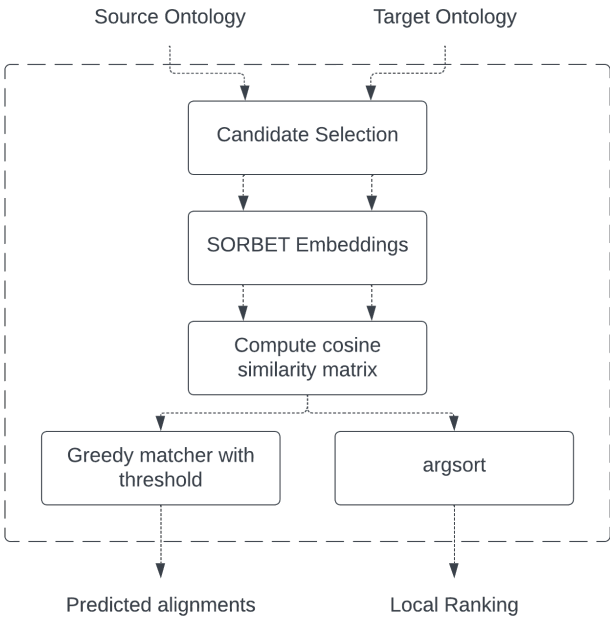

Figure 1: General architecture of SORBETMatcher

1.2. Specific techniques used

1.2.1. Candidate Selection

The first step of the matching process is to determine which concepts are likely to be matched. Since fetching SORBET Embeddings can be a long process for large ontologies, reducing the number of candidate concepts can greatly improve the runtime. There are three strategies to obtain a smaller set of candidate classes. Firstly, we employ a string matcher that identifies pairs of concepts with matching labels or synonyms as alignments. Concepts originating from these high-precision alignments are pruned from the set of considered classes. Secondly, some classes in the Bio-ML track has the *use_in_alignment* tag indicating whether they should be used or not. Finally, in the local ranking of the Bio-ML track, candidates mappings are suggested from a *test.cands* file. We identify each unique class in the candidates, and consider them as the sole relevant classes.

1.2.2. SORBET Embeddings

SORBET is an Ontology Embedding method that has the goal of obtaining rich BERT embeddings while rearranging the latent space based on the ontology's structure. To achieve this, SORBET fine-tunes SentenceBERT, a pre-trained siamese BERT model, with a regression loss based on the distance between classes:

$$L = \frac{1}{|M|} \sum_{(c_i,c_j)\in M} [(sim_\theta(c_i, c_j) - \frac{A - min(d(c_i, c_j), A)}{A})^2] \quad (1)$$

Where M, is a training dataset containing pairs of classes, c_i, c_j. sim_θ is a predicted similarity, and A is a hyperparameter representing the distance between 2 classes. Intuitively the parameter A will control the sparsity of the ontology's classes in the latent space. The bigger the value of A the larger the distance between neighbor classes. The distance d is defined by the number of *subClassOf* relationships between c_i and c_j.

To obtain SORBET embeddings representing classes, the input of the SentenceBERT model is a random walk describing each class, providing context to classes given their neighbor subclasses, parent classes, and classes related by object properties. Both at training and inference time, a new random walk is created to describe a concept. The fine-tuning of SentenceBERT is achieved with pairs of concepts composed of positive samples, semi-negative samples and negative samples.

By sampling a class and its neighbors, then computing a similarity score relative to their distance, SORBET Embeddings attempts to replicate the structure of the ontology in the latent space. Therefore, similar classes from different ontologies get restricted into the same region of the latent space, making embeddings well-suited for the alignment or matching task.

1.2.3. Compute cosine similarity matrix

Using the embedding of all relevant classes, a similarity matrix is constructed using the cosine similarity measure as highlighted by equation 2.

$$S_{i,j} = \frac{\Omega(c_i) \cdot \Omega(c_j)}{||\Omega(c_i)||\ ||\Omega(c_j)||} \quad (2)$$

where Ω is a function that transforms a concept into its SORBET Embedding and c_i, c_j represent the source and target concept respectively.

The i-th row represents the i-th concept from the source ontology and the j-th column represents the j-th concept from the target ontology. This matrix is initialized with a few values. The similarity of alignments outputted by the string matcher (described in the candidate selection in section 1.2.1) are set to 1.0 while the columns and rows of the concepts whose *use_in_alignment* property is False are set to 0. For local rankings, the similarity between pairs of classes that are not in the candidates are set to 0. All the remaining cells are filled with the cosine similarity values.

1.2.4. Greedy Matcher with threshold

To determine which mappings from the similarity matrix will be retained, we utilize a straightforward greedy algorithm, akin to approaches in related works such as [7] and [1]. This simple algorithm sorts the similarity values and then iterates through each element S_{ij} in descending order of scores and selecting mappings provided that neither its source nor target concepts have already been chosen. The algorithm goes on until the value of S_{ij} goes below the threshold

value of 0.75. Even though the neighbors of equivalent classes also have a high similarities, the goal of the greedy matching algorithm is to reduce these false positive alignments and to produce 1:1 alignments.

1.2.5. Local Ranking

The local ranking evaluation method requires only the target candidates for a concept to be sorted in descending order. The algorithm then iterates through each non-null row i, and applies an index sort to indicate to which j-th concept (from the j-th column) the source concept is most likely to be matched with.

1.3. Specific settings and Hyperparameters

For the OAEI competition, two models with different hyperparameters were used, the MELT[6] submission model and the semi-supervised Bio-ML model. For the both models, SORBET Embeddings were trained starting from the pre-trained SentenceBERT *sentence-transformers/all-MiniLM-L6-v2*. The MELT model was trained simultaneously on the conference and anatomy track with a value of A equal to 5 while no other changes were made to the original hyper-parameters used in SORBET [4]. This model was also used for the evaluation of the unsupervised equivalence matching of the Bio-ML track. This was done to show how SORBETMatcher performs in a zero-shot learning context.

For the remaining results of the Bio-ML track, SORBET was individually fine-tuned on the sub-tracks, using the train reference alignments as positive samples in the SORBET training. The hyper-parameters of SORBET's semi-supervised version on Bio-ML were the following: For the A value, our experiments hinted that shallow ontologies are better embedded with a low A value. Therefore, the OMIM-ORDO and NCIT-DOID had a value of A kept at 4, while for the rest of the sub-tracks A was reduced to 3. Other experiments have also shown that the generation of negative samples during training lead to worse results, this caused us to remove them completely. This may be due to the fact that negative samples are normally used to increase the precision but at the cost of reducing recall. However, since the precision is high in most sub-tracks, the trade-off can be counter-productive.

2. Results

Full results for all SORBETMatcher's alignments are shown in Table 1.

2.1. Anatomy

The anatomy track involves aligning the Adult Mouse Anatomy (MA) with the NCI Thesaurus, which describes Human Anatomy (NCI). SEBMatcher achieved an F1-score of 0.909, with a precision of 0.923 and a recall of 0.895. In comparison to other systems in this year's competition, SORBET obtained the 3rd position out of 9 based on the F1 score. However, it is worth noting that SEBMatcher's performance lagged in terms of runtime, having a total time of 4032 seconds.

Table 1
SORBETMatcher's results in the anatomy and conference tracks

Reference alignments	rank (F1-score)	F1-score	precision	recall
ra1-m1	2/11	0.76	0.78	0.75
anatomy	3/9	0.909	0.923	0.895

Table 2
SORBETMatcher results in the Bio-ML equivalence matching track

Subtrack	rank (F1-score)	F1-score	precision	recall
OMIM-ORDO Unsupervised	2/9	0.663	0.693	0.635
OMIM-ORDO Semi-Supervised	5/9	0.607	0.568	0.652
NCIT-DOID Unsupervised	1/9	0.913	0.920	0.907
NCIT-DOITD Semi-Supervised	1/9	0.903	0.925	0.882
Body Unsupervised	6/9	0.677	0.618	0.749
Body Semi-Supervised	4/9	0.746	0.794	0.704
Pharm Unsupervised	3/8 (tie)	0.748	0.973	0.607
Pharm Semi-Supervised	8/8	0.715	0.876	0.604
Neoplas Unsupervised	7/9	0.634	0.626	0.642
Neoplas Semi-Supervised	4/9 (tie)	0.662	0.731	0.605

2.2. Conference

The conference track involves aligning a set of ontologies that describe the domain of conference organization. This track encompasses multiple reference alignment sets, with M1 alignments focusing solely on classes, M2 on properties, and M3 containing both classes and properties. Given that SORBET is presently only able to embed classes exclusively, its performance is less robust when applied to the M3 reference alignments.

2.3. Bio-ML

The Bio-ML consists of 5 different reference alignments across multiple ontologies. It is separated into equivalence matching and subsumption matching. SORBETMatcher participated to both sub-tasks. The equivalence matching is also decomposed into 2 categories, one with the unsupervised test set (100% of reference alignments) and one with the semi-supervised test set (70% of reference alignments).

Table 3
SORBETMatcher results in the Bio-ML subsumption matching track

Subtrack	rank (MRR)	MRR	Hits@1	Hits@5	Hits@10
OMIM-ORDO	6/6	0.272	0.181	0.347	0.431
NCIT-DOID	1/6	0.802	0.695	0.941	0.977
Body	2/6	0.516	0.311	0.821	0.941
Pharm	1/6	0.760	0.659	0.880	0.912
Neoplas	1/6	0.685	0.557	0.859	0.899

3. General comments and Conclusion

Overall, SORBETMatcher achieved a top performance in some of the tracks while still having some improvements to be made on others.

The Bio-ML subsumption track is the task where SORBETMatcher scored the strongest, with three first places and one second place. However, SORBETMatcher scored last in the OMIM-ORDO sub-track by a a large margin, especially for higher Hits@K. This may indicate a flaw in the SORBET Embeddings obtained on the OMIM or ORDO ontologies. The nature of this problem is still to be further investigated, but our initial hypothesis is that it might be due to the restriction axioms (which are numerous in these ontologies) and which are not considered by SORBET in its semi-negative sampling.

The results of the Bio-ML equivalence matching track are mixed. SORBETMatcher scored the best in the NCIT-DOID subtrack where it achieved first place in both unsupervised and supervised test sets. Considering the subsumption results for the NCIT-DOID subtrack, where SORBETMatcher largely outperformed other systems, we hypothesize that SORBET Embeddings are much more representative of ontologies with higher depths such as DOID. Another conclusion we can draw from these results is the capability of SORBET Embeddings to work in zero-shot learning tasks. Indeed, the unsupervised results all come from the MELT packaging of the system, in which SORBET is frozen after being trained on the conference and anatomy tracks. Therefore, at inference time, the BERT model has never seen the concept to embed, hence our conclusion about its zero-shot capability. As for datasets like Pharm and Neoplas, SORBETMatcher has yielded disappointing results. The problem may be of the same nature as the OMIM-ORDO dataset in subsumption matching, but it could also be because of the lack of hyper-parameters tuning, which can be very sensitive.

As for the performance of SORBETMatcher on the conference and anatomy tracks, SORBETMatcher was able to obtain good results by reaching the second and third place respectivelly.

4. Acknowledgements

This research has been funded by Canada's NSERC Discovery Research Program.

References

[1] Alexandre Bento, Amal Zouaq, and Michel Gagnon. "Ontology Matching Using Convolutional Neural Networks". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 5648–5653. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.693.

[2] Jiaoyan Chen et al. "Contextual semantic embeddings for ontology subsumption prediction". In: *World Wide Web* 26.5 (Sept. 2023), pp. 2569–2591. ISSN: 1573-1413. DOI: 10.1007/s11280-023-01169-9. URL: https://doi.org/10.1007/s11280-023-01169-9.

[3] Francis Gosselin and Amal Zouaq. "SEBMatcher Results for OAEI 2022". In: *Ontology Matching 2022 : Proceedings of the 17th International Workshop on Ontology Matching (OM 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022), Hangzhou, China, virtual conference, October 23, 2022*. Vol. 3324. CEUR Workshops Proceedings. CEUR-WS.org, 2022, pp. 202–209.

[4] Francis Gosselin and Amal 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.

[5] Yuan He et al. "Bertmap: A bert-based ontology alignment system". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 5. 2022, pp. 5684–5691.

[6] Sven Hertling, Jan Portisch, and Heiko Paulheim. "MELT - Matching EvaLuation Toolkit". In: *Semantic Systems. The Power of AI and Knowledge Graphs - 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings*. 2019, pp. 231–245. DOI: 10.1007/978-3-030-33220-4_17. URL: https://doi.org/10.1007/978-3-030-33220-4%5C_17.

[7] Vivek Iyer, Arvind Agarwal, and Harshit Kumar. "VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10780–10792. DOI: 10.18653/v1/2021.emnlp-main.842. URL: https://aclanthology.org/2021.emnlp-main.842.

[8] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Conference on Empirical Methods in Natural Language Processing*. 2019.

[9] Jifang Wu et al. "Daeom: A deep attentional embedding approach for biomedical ontology matching". In: *Applied Sciences* 10.21 (2020), p. 7909.