Aligning Network of Ontologies using Graph AI

Sevinj Teymurova

City, University of London, UK

Abstract

Ontology alignment is a crucial task in the Semantic Web community, yet current techniques still have drawbacks.The research of my PhD aims at building an ontology alignment system that incorporates novel Graph AI based features to align multiple ontologies. The approach intends to combine Knowledge Graph Embeddings with Graph Neural Networks to capture the rich semantics of OWL ontologies from various real-world domains like Life Science, Ecotoxicology, Oil and Gas etc. Current experiments have focused on the extension of an Ontology Embedding System to incorporate confidence in the graph edges represented by (uncertain) available mappings between the input ontologies.

1. Introduction

This research focuses on Ontology alignment, an important problem in the field of Semantic Web that involves identifying and reconciling semantic mappings between multiple ontologies. Existing approaches have made significant progress in ontology alignment, but they are still facing challenges to be fully adopted in real-world solutions [1, 2]. My PhD research aims at designing and implementing an ontology alignment system that captures the semantic information about entities and their relations relying on semantically-enriched deep learning by combining Knowledge Graph Embeddings with Graph Neural Networks.

Ontology Alignment. The motivation for performing ontology alignment is well supported by the FAIR [3] principles which are intended to act as a guide to enable digital resources to become more Findable, Accessible, Interoperable and Reusable for machines and thus also for humans. To put it succinctly, an “ontology” in AI is a formal and explicit specification of a shared domain conceptualization, expressed in a formal language, that provides a common understanding of a domain for a community of users and enables sharing and reuse of domain knowledge across applications and systems, serving as a conceptual schema that structures the graph data model and provides a framework for querying and reasoning over the graph data in knowledge graphs [4]. On the other hand, ”Ontology matching” (or alignment) is the process of finding relationships or correspondences between two or more entities in two or more independent ontologies. For example, as shown in Fig.1, `HeLiS:Fructose` in HeLiS ontology is equivalent (*e.g.*, owl:equivalentClass) to the concept `obo:FOODON_03301305` (fructose) in FoodOn.

As a consequence of a combination of numerous techniques and advanced tools, along with significant human effort in curation and complex auditing procedures, it has become feasible to create mappings between real-world ontologies. [6, 7, 8] Despite the effectiveness of current

ISWC 2023 Doctoral Consortium (DC), November, 2023, Athens, Greece
sevinj.teymurova@city.ac.uk (S. Teymurova)

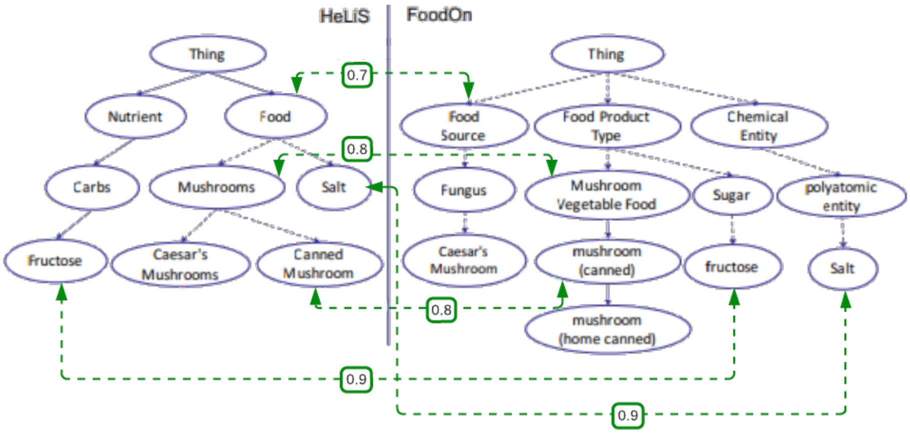

Figure 1: Fragment of an alignment of two ontologies: HeLiS and FoodOn (Figure adapted from [5]). The dash arrow means some intermediate classes are hidden. The green dash arrow denotes mappings with confidence values ranging from [0,1]. Confidence values represent the degree of certainty associated with a correspondence between entities.

matching tools in handling moderately sized ontologies, they still lack the capability to process large-scale bio-medical ontologies like NCI, FMA, or SNOMED CT [9]. The Ontology Alignment Evaluation Initiative (OAEI) began in 2004 [10] as a organized international forum for the collection of benchmark datasets for ontology matching systems, as well as the yearly evaluation of those matching systems supported by industry (e.g., IBM research, Pistoia Alliance, SIRIUS) [11]. The OAEI aims to provide an open and consistent platform for comparing the performance of different techniques and methodologies in ontology matching. It provides matching assessments to participants, and their outcomes are evaluated using measures influenced by information retrieval, including *Precision* (a measure of accuracy), *Recall* (a measure of comprehensiveness), and $F - measure$, which combines these measures.

Knowledge Graph Embeddings. The existing state-of-the-art ontology matching methods include embedding-based techniques, which have demonstrated promising results in enhancing alignment accuracy. [5, 12] These methods focus on learning feature representations that capture the semantics of ontologies and mappings, and use them to predict alignments. One of those methods is called *Knowledge Graph Embeddings.* Knowledge Graph Embedding or Ontology Embedding is the name given to a group of representation learning (or feature learning) approaches that turn data semantics, such as those found in sequences and graphs, into vectors that can be used for statistical analysis and machine learning prediction tasks later on [13].

A knowledge graph embedding is a mapping function:

$$f : V \to \mathbb{R}^d$$

that assigns a d-dimensional vector representation to each entity v in the knowledge graph. This mapping function aims at capturing the semantic information of the entities and their relationships in the vector space. When two entity vectors v_1 and v_2 are close in the embedding space, as quantified by a distance metric such as the Euclidean distance or cosine similarity, it suggests that the entities share common characteristics, exhibit similar behavior, or participate in similar relationships within the knowledge graph.

In this research, we propose to extend a state-of-the-art Knowledge Graph Embedding system based on random walks (*i.e.*, OWL2Vec*) with a new technique to improve alignment accuracy. Our approach incorporates edge confidence information from an input set of (incomplete and potentially inaccurate) mappings between the ontologies being aligned. By incorporating this information, we introduce bias in the random walks, which leads to improved alignment results. The ontology embeddings, along with a Graph Neural Network, are then combined to score candidate mappings. This extension allows us to leverage the power of random walks and edge confidence to achieve more accurate ontology alignments. The benefit of the current extension with respect to previous works [5] is the creation of entity vectors for all the ontologies within the same embedding space. Our hypothesis is that this will lead to an improved alignment accuracy. We also aim to go beyond pairwise alignment of ontologies to include cases where a network of ontologies needs to be aligned [14, 8].

2. State of the Art

A variety of solutions have been proposed to address the problem of ontology alignment, which has been a focus of active research in the Semantic Web community.(*e.g.*, [15, 11]). These state-of-the-art ontology matching methods include embedding-based techniques, which have demonstrated promising results in enhancing alignment accuracy [5, 12].

However, despite the progress made in handling large ontologies, the main challenge now lies in achieving better quality alignments and reducing dependency on threshold values that require constant adaptation. Modern techniques such as Machine Learning (ML) and Embedding hold the potential to address these challenges. As a result, even ontologies with millions of potential mappings or tens of thousands of classes, which encompass billions of possible mappings, can be more effectively handled by ontology matching tools.

One popular embedding technique for ontologies is OWL2Vec*, which has been shown to outperform other embedding methods in some downstream tasks [13]. However, existing approaches using OWL2Vec* or other embedding techniques (*e.g.*, [16, 17, 18]) typically consider only a single ontology and do not incorporate mappings between the ontologies at the embedding phase, which may negatively impact the subsequent alignment process. These mapppings can be partial and not necessarily accurate (*e.g.*, mapping computed by an alignment system), thus providing means to incorporate their confidence within the embedding process is paramount. In the case of OWL2Vec* we aim at exploiting the confidence of the mappings to bias the walk in the ontology graph.

Wei [19] proposed a novel approach for sampling from large-scale networks by exploiting random walk strategies. The proposed method extends the classic Metropolis-Hastings algorithm for random walk sampling by introducing several novel strategies, such as early rejection, importance sampling, and biased sampling, but it may not be suitable for networks with highly skewed or imbalanced degree distributions, as the sampling strategy may bias towards high-degree nodes and miss low-degree nodes. RDF2vec [16] presented a novel approach for learning node embeddings in RDF knowledge graphs. Steenwinckel et al. [20] extended RDF2vec by introducing several new walk extraction strategies that aim to capture different aspects of the semantic structure in RDF data (*e.g.*, hierarchical relations, property chains).

Prior research has proposed various ontology matching techniques, including logic-based methods (*e.g.*, [10, 15]), machine learning approaches (*e.g.*, [21, 22])., and embedding-based methods. Logic-based methods, such as ontology alignment based on Description Logics (DLs) (*e.g.*, [15]) and formal concept analysis, leverage logical reasoning to establish correspondences between ontologies. Logic-based methods use formal logic to represent the alignments and filter our the ones leading to logical errors, while Machine learning approaches, such as instance-based methods and supervised learning, utilize machine learning algorithms to learn matching rules from labeled data. (*e.g.*, [23])

The combination of Knowledge Graph Embeddings (KGE) with Graph Neural Networks (GNN) has been extensively studied for various tasks, such as link prediction and node classification in knowledge graphs. KGE techniques, such as TransE (*e.g.*, [17]) , DistMult (*e.g.*, [24]), and ComplEx (*e.g.*, [25]), have been successful in learning continuous vector representations of entities and relationships in knowledge graphs. GNN models, such as Graph Convolutional Networks (GCN) (*e.g.*, [26]), GraphSAGE (*e.g.*, [27]), and GAT (*e.g.*, [28]), have demonstrated their ability to capture structural information and propagate node features in graph-structured data. While the combination of KGE and GNN has been explored for tasks like link prediction and node classification, its application to ontology matching hasn't been explored yet. In particular, incorporating edge confidence information and considering mappings between multiple ontologies in embedding-based methods are important challenges that need to be addressed. Most methods are also limited to aligning only a pair of ontologies at a time, and they do not consider larger-scale integration of multiple ontologies or knowledge graphs. Some recent approaches like [8] are bringing the need of holistic ontology alignment to solve real-world problems.

3. Problem Statement and Contributions

In today's information-driven world, the internet is teeming with a vast amount of data and knowledge. However, a significant challenge arises when attempting to extract meaningful insights from this abundance of information. Various ontologies, each representing a specific domain, exist in different formats and structures. Consequently, multiple datasets may describe the same concept but differ in their representation, posing a significant obstacle to seamless data integration and knowledge sharing. Addressing this challenge requires the development of effective techniques for ontology alignment and mapping to establish meaningful connections between disparate ontologies and facilitate accurate information retrieval and analysis. This has been recognised as a key issues, and considerable recent research has been conducted on ontology alignment. My PhD aims at contributing in the following areas:

1. Multi-ontology alignment with the goal to enable interoperability between different systems or applications that use different ontologies, by creating a common understanding of the entities and relationships across the different ontologies
2. Conducting the alignment of a network of ontologies using Graph AI techniques.
3. Expanding upon existing state-of-the-art Knowledge Graph Embedding System by incorporating techniques to improve alignment accuracy and handle multiple ontologies.

4. Introducing Biased Random Walk by incorporating edge confidence information obtained from an input set of ontologies and their mappings to reduce the search space.
5. Leveraging the power of Graph Neural Networks to combine ontology embeddings and effectively score candidate mappings, ensuring high-quality alignments.
6. Combining extended techniques in a novel way to achieve improved performance in ontology alignment between multiple ontologies and mappings.
7. Providing a framework for evaluating and comparing different ontology matching systems using the OAEI platform.

Aligning network of ontologies involves discovering relationships or correpondences between entities in multiple independent ontologies (such as classes and properties), with the objective of creating a single shared ontology that can be utilized across applications and systems [29]

$$M = (e_1, e_2, n, r) \mid e_1 \in O_1, e_2 \in O_2, n \in [0, 1]$$

Here, M is the simples unit to identify a mapping, M is the set of tuples (e_1, e_2, n, r) that satisfy the conditions stated in the definition. The variables e_1 and e_2 represent entities, such as classes and properties, in the input ontologies O_1 and O_2, respectively. The variable n represents a confidence value between 0 and 1, indicating the level of certainty in the alignment. The variable r represents the semantic relationship between e_1 and e_2, which can be one of the three possible relationships: subsumption, equivalence, or disjointness.

Research Hypothesis:

H1. Incorporating (incomplete and possibly imperfect) mappings and their edge confidence between multiple ontologies in embedding-based methods (*e.g.*, OWL2Vec*) will lead to improved ontology alignment accuracy compared to existing embedding-based methods.

H2. Extending OWL2Vec* embedding technique to incorporate new walking strategies using edge confidence information to bias the random walks will lead to better ontology embeddings.

H3. Using a Graph Neural Networks will improve the scoring of candidate mapping in self-supervised and semi-supervised settings.

H4. Incorporating sampling techniques to avoid the exponential growth of walks will enable more efficient processing of larger knowledge graphs within an alignment setting without sacrificing accuracy.

4. Research Methods

We have designed an ontology alignment pipeline, where we adapt an embedding technique to integrate edge confidence information into the walking strategy combined with the similarity score computed by a graph neural network model to forecast the final mappings. This approach enables us to factor in edge confidence information and account for available mappings to attain enhanced results in aligning multiple OWL ontologies and mappings. Fig. 2 provides an overview of the alignment module, while Fig. 3 depicts the main steps of the KG embedding

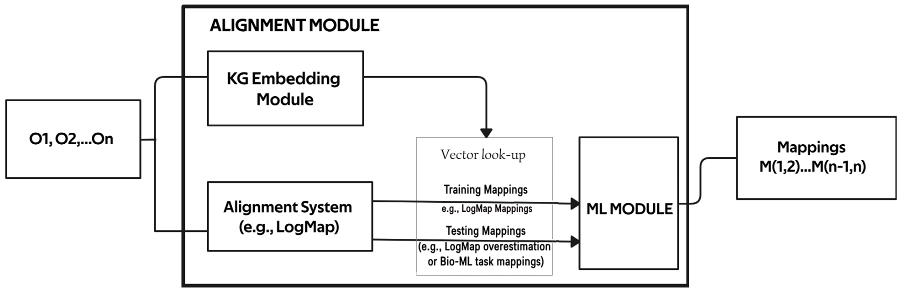

Figure 2: Ontology Alignment Module

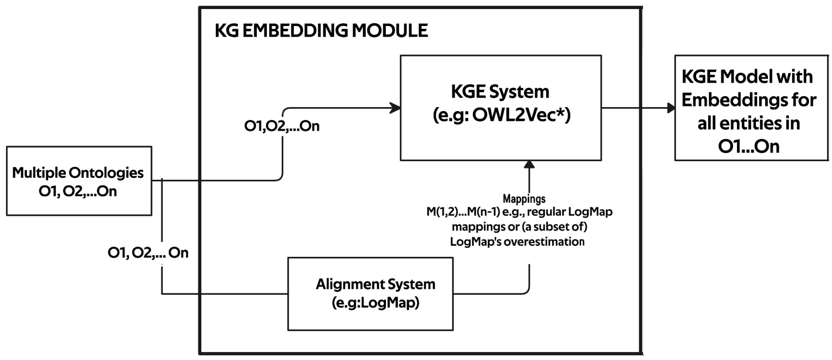

Figure 3: KG Embedding Module

module. The Ontology alignment module receives multiple ontologies $O1, O2..ON$ as input. Based on these ontologies, it computes final mappings $M(1,2)..M(n-1,n)$.

As shown in Fig 3, several ontologies and mappings are input into the KG Embedding system in order to produce the corresponding embeddings. The KG Embedding system uses a bias-based walk method and sampling approach to complete this objective. The sampling approach is used to condense an ontology's search space, increasing system effectiveness

To implement of an embedding system using edge confidence value and to experiment our approach, we extend one popular embedding technique for ontologies - OWL2Vec*[13], which has been shown to outperform other embedding methods in some tasks. OWL2Vec* relies on random walks and word embedding to encode the semantics of OWL ontologies by taking into account its graph structure, lexical information and logical information. However, OWL2Vec* like other embedding systems typically consider only a single ontology and do not incorporate mappings between the ontologies at the embedding phase, which may negatively impact the subsequent alignment process. These mappings can be partial and not necessarily accurate (*e.g.*, system-computed mappings), thus providing means to incorporate their confidence with the embedding process is paramount. In the case of OWL2Vec* we aim at exploiting the confidence of the mappings to bias the walk in the ontology graph. More specifically, edges originated from ontology axioms are assigned a confidence value of 1.0, indicating high confidence. On the other hand, mappings between ontologies are assigned confidence values ranging from 0 to 1, reflecting the level of uncertainty in the alignment. This ensures that the biased random walks will prioritize more reliable edges between the ontologies during the random walks.

The prototype system can be deployed on a cloud platform or a local machine, depending on the size and complexity of the datasets used for evaluation. For instance, if the dataset is large and complex, a cloud platform with high processing power and storage capacity can be used for the prototype system to achieve optimal performance.

The proposed approach would be compared with existing ontology alignment methods, such as LogMap [15] and AML [30], and conducting experiments on benchmark datasets within the OAEI [11] like Bio-ML [31] , Food-On, Helis and OAEI 2022 Conference data.

5. Conclusion

In conclusion, this research proposes an ontology alignment network that integrates new features and Graph AI-based methods to align multiple ontologies and mappings and capture richer semantic information about entities and their relations. The proposed system combines Knowledge Graph Embeddings with Graph Neural Networks to achieve a semantic-aware approach. The approach is evaluated using benchmark datasets and compared to existing ontology alignment systems, demonstrating improved accuracy and reduced computational complexity and memory requirements. The research has the potential to provide greater performance in ontology alignment and be widely useful in a variety of knowledge management software applications.

Acknowledgments

Funded by Ministry of Science and Education of AR. PhD thesis supervised by Ernesto Jiménez-Ruiz and Tillman Weyde at City, University of London, UK.

References

[1] I. Harrow, et al., Ontology mapping for semantically enabled applications., Drug discovery today (2019).

[2] H. Li, et al., User validation in ontology alignment: functional assessment and impact, Knowl. Eng. Rev. 34 (2019).

[3] G. Cota, et al., Best practices for implementing fair vocabularies and ontologies on the web, Applications and practices in ontology design, extraction, and reasoning 49 (2020) 39.

[4] T. R. Gruber, Toward principles for the design of ontologies used for knowledge sharing?, International journal of human-computer studies 43 (1995) 907–928.

[5] J. Chen, et al., Augmenting ontology alignment by semantic embedding and distant supervision, in: The Semantic Web: ESWC, Springer, 2021, pp. 392–408.

[6] P. Kolyvakis, et al., Biomedical ontology alignment: an approach based on representation learning, Journal of biomedical semantics 9 (2018) 1–20.

[7] K. Xu, et al., Coordinated reasoning for cross-lingual knowledge graph alignment, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 9354–9361.

[8] M. C. Silva, et al., Matching multiple ontologies to build a knowledge graph for personalized medicine, in: The Semantic Web - ESWC, volume 13261, 2022.
[9] D. Faria, et al., Tackling the challenges of matching biomedical ontologies, Journal of biomedical semantics 9 (2018) 1–19.
[10] J. Euzenat, et al., Ontology alignment evaluation initiative: six years of experience, Journal on data semantics XV (2011) 158–192.
[11] M. Pour, et al., Results of the ontology alignment evaluation initiative 2022, in: CEUR Workshop Proceedings 2022, CEUR, 2022.
[12] Y. He, J. Chen, D. Antonyrajah, I. Horrocks, BERTMap: A BERT-Based Ontology Alignment System, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, 2022.
[13] J. Chen, et al., OWL2Vec*: Embedding of owl ontologies, Machine Learning 110 (2021).
[14] J. Euzenat, The category of networks of ontologies, CoRR abs/1412.3279 (2014). URL: http://arxiv.org/abs/1412.3279. arXiv:1412.3279.
[15] E. Jimenez-Ruiz, et al., LogMap: Logic-based and scalable ontology matching, ISWC (2011).
[16] Ristoski, et al., RDF2vec: Rdf graph embeddings for data mining, in: ISWC, 2016.
[17] Bordes, et al., Translating embeddings for modeling multi-relational data, NIPS 26 (2013).
[18] Vandewiele, et al., pyrdf2vec: A python implementation and extension of RDF2vec, arXiv preprint arXiv:2205.02283 (2022).
[19] W. Wei, et al., Towards efficient sampling: Exploiting random walk strategies, in: AAAI, volume 4, Citeseer, 2004, pp. 670–676.
[20] Steenwinckel, et al., Walk extraction strategies for node embeddings with rdf2vec in knowledge graphs, in: DEXA, 2021.
[21] Iyer, et al., Veealign: a supervised deep learning approach to ontology alignment., in: OM@ ISWC, 2020, pp. 216–224.
[22] L. Chen, et al., A deep learning based method for extracting semantic information from patent documents, Scientometrics 125 (2020) 289–312.
[23] E. Rahm, P. A. Bernstein, A survey of approaches to automatic schema matching, the VLDB Journal 10 (2001) 334–350.
[24] B. Yang, et al., Embedding entities and relations for learning and inference in knowledge bases, arXiv preprint arXiv:1412.6575 (2014).
[25] T. Trouillon, et al., Complex embeddings for simple link prediction, in: International conference on machine learning, PMLR, 2016, pp. 2071–2080.
[26] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907 (2016).
[27] W. Hamilton, et al., Inductive representation learning on large graphs, Advances in neural information processing systems 30 (2017).
[28] Casanova, et al., Graph attention networks, ICLR. Petar Velickovic Guillem Cucurull Arantxa Casanova Adriana Romero Pietro Liò and Yoshua Bengio (2018).
[29] P. Shvaiko, J. Euzenat, Ontology matching: state of the art and future challenges, IEEE Transactions on knowledge and data engineering 25 (2011) 158–176.
[30] D. Faria, et al., The agreementmakerlight ontology matching system, in: OTM conferences, 2013.
[31] Y. He, et al., Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching, in: ISWC, 2022, pp. 575–591.