Entity Alignment for Knowledge Graphs with Multi-order Convolutional Networks (Extended Abstract)

Nguyen Thanh Tam[1], Huynh Thanh Trung[2], Hongzhi Yin[3], Tong Van Vinh[4], Darnbi Sakong[2], Bolong Zheng[5], Nguyen Quoc Viet Hung[2]

[1]*Ho Chi Minh City University of Technology (HUTECH),* [2]*Griffith University,* [3]*The University of Queensland,* [4]*Hanoi University of Science and Technology,* [5]*Huazhong University of Science and Technology*

***Abstract*—Knowledge graph (KG) entity alignment is the task of identifying corresponding entities across different KGs. Existing alignment techniques often require large amounts of labelled data, are unable to encode multi-modal data simultaneously, and enforce only a few consistency constraints. In this paper, we propose an end-to-end, unsupervised entity alignment framework for cross-lingual KGs using multi-order graph convolutional networks. An evaluation of our method using real-world datasets reveals that it consistently outperforms the state-of-the-art in terms of accuracy, efficiency, and label saving.**

***Index Terms*—knowledge graph, entity alignment, network embedding, graph convolutional neural network**

I. Introduction

Knowledge graphs (KGs) are used to represent real-world entities, the relationships between them, and the relationships between their attributes [1], [2]. Entity alignment (the task of identifying corresponding entities between monolingual KGs) is the foundation for the integration of multiple KGs [3]. The problem of KG entity alignment has emerged recently with graph embedding techniques [4], [5]. The first generation methods of this paradigm [6] learn the embeddings on the assumption that if two entities have a relation, the distance between their respective embeddings is equal to the embedding of their relation. Avoiding this strict assumption, the second generation of embedding techniques employ graph neural networks, which encode the structural relationship based on neighbourhood information [6].

However, we argue that the above approaches overload the embedding model with unrelated objectives. On the one hand, the entity embeddings must encode the syntactic information (e.g. neighbourhood, topology, degree) for each KG, while on the other, they also need to reflect the semantic alignment of entities across KGs. Furthermore, existing models have not fully utilised the attribute information of entities (e.g. the age attribute of a person, the population of a country) due to the high levels of inconsistency and linguistic differences.

In this paper, we meet the above requirements via a unified, unsupervised and adaptive entity alignment model for cross-lingual KGs. In essence, our idea is to fully leverage the richness of a KG by simultaneously comparing the relational and attributional information of the entities to be aligned. The fusion of these types of information helps them to complete each other and to mitigate the high levels of consistency violation for each type. To efficiently extract the relational data, we propose to use the multi-layer characteristics of graph convolutional networks (GCNs) to model the relational correlation at different orders without the need for supervision data (e.g. pre-aligned entities). We published the source code and the datasets [1] for use by the community.

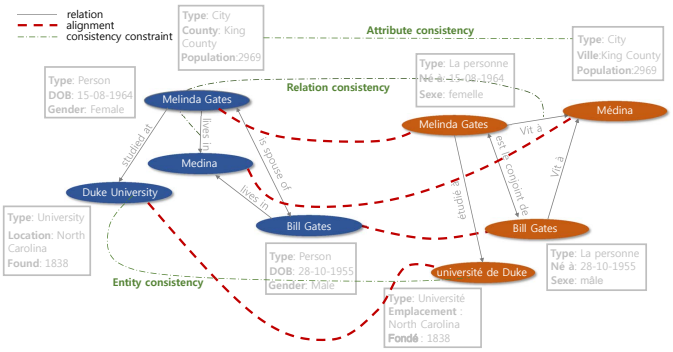

Fig. 1: Example of knowledge graph entity alignment

II. Problem and Approach

Problem. KG entity alignment aims to find all of the corresponding entities of two given KGs. Fig. 1 shows an example of knowledge graph alignment with several entities from YAGO, in which the entities in the English version are aligned with their counterparts in the French version.

Consistency guarantees. In general, entity alignment cross-lingual KGs needs to respect three types of constraints: (i) *entity consistency* – the names of corresponding entities should be equivalent, (ii) *relation consistency* – the relation is preserved from source KG to target KG, and (iii) *attribute consistency* – the corresponding entities should have equivalent attributes and equivalent attribute values. Most baseline models struggle to address all three types of consistency requirements simultaneously, since the attribute imbalance between the two

[1]https://github.com/thanhtrunghuynh93/EMGCN

KGs and modality inconsistency are frequently observed in real-world [6]. Going beyond the state-of-the-art, our multi-order embedding model can naturally satisfy the above constraints at the same time, proven by theoretical analyses and empirical experiments on the real-world datasets [6].

Relation-aware Multi-order Embedding. Our GCN-based model consists of k layers, and each hidden feature layer simultaneously encodes the topological and attributional information using a message passing scheme [7]. Most of GCN models use the embeddings in the final layer as the node representation [8], since the deepest layer aggregates the information from all previous layers. However, while the deeper layers contain richer topological information, they are also prone to noise from inconsistent nodes in previous layers, which is fairly common in real-world KG datasets. The topological information may also be diluted in the deeper layers, especially for expander-like networks, since the collective information of a large neighbourhood would overshadow individual nodes. To address these challenges, we instead use the learned embeddings at all layers to identify the nodes. This strategy allows our framework to exhaustively exploit the topological information in both a local and global manner.

Unsupervised loss function. We design an *unsupervised* loss function to minimise the distance between embeddings of neighbouring nodes while maximising those of unrelated nodes [6]. Our loss function goes beyond the state-of-the-art to incorporate all low-order and high-order embeddings. The former are embeddings at shallower layers, which capture locally topological patterns of nodes, although irrelevant nodes may share a similar patterns. The latter are embeddings at deeper layers, which capture larger neighbourhood information, but risk pulling different communities of nodes too close.

Weight-sharing training. To make sure the source and target networks are embedded into the same vector space, we use weight-sharing training. That is, the forward pass of the GCN model for both source and target networks uses the same weight matrices for each layer. This mechanism guarantees that the learned embeddings of the source and target relational network stay within a common embedding space, thus allowing their direct usage without a reconciliation step for the two different spaces. The mechanism also assures that the consistency constraints are satisfied [6].

III. Excerpt of Experimental Results

We conducted experiments with real-world DBP15K [6] datasets, which were generated from DBpedia, a large-scale multilingual knowledge base containing rich inter-language links between different languages [6]. In total, there are 500K entities, 13K relations, 38K attributes, and 4.2M triples.

Table I reports an end-to-end comparison of our model against baseline methods. Our *EMGCN* model outperformed the others in all scenarios, without requiring any supervision data. Though using similar multi-order GCN mechanism as *GAlign*, the gain of 10-20% of Success@1 demonstrates the efficiency of the fusion of rich properties in KGs proposed in *EMGCN*, especially for noisy datasets such as *ZH-EN*.

TABLE I: End-to-end effectiveness

Dataset	Metric	EMGCN	GAlign	RDGCN	GCNA	MuGNN
ZH-EN	Success@1	**0.8625**	0.6943	0.7029	0.4057	0.4779
	MAP	**0.8931**	0.7513	0.7250	0.5270	0.6000
JA-EN	Success@1	**0.8663**	0.7481	0.7630	0.4072	0.4866
	MAP	**0.8987**	0.8025	0.8110	0.5270	0.6103
FR-EN	Success@1	**0.9395**	0.8695	0.8775	0.3910	0.4896
	MAP	**0.9582**	0.9052	0.9060	0.5270	0.6171

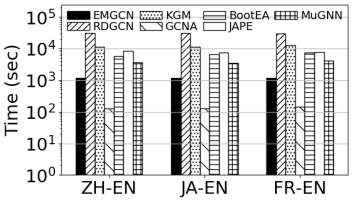

Fig. 2: Computation time

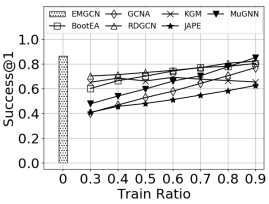

Fig. 3: Label savings

Fig. 2 reports the running time of our model against the baselines. The reported times for the supervised baselines included training and testing times. It can be seen that *GCNA* is the fastest baseline, with a running time of approximately $100s$, but accuracy is sacrificed to achieve this (the value of *Success@1* for this method is less than 0.41, as shown in Table I). Our model *EMGCN* was the next best, since it does not require any training time.

Fig. 3 shows the power of our unsupervised approach against supervised baselines. We varied the training ratio from 0.3 to 0.9 and compared the values for *Success@1* on the ZH-EN dataset. It can be seen that our *EMGCN* model, with no training data, outperformed or was on a par with all the other baselines, even when they used 90% of the data for training.

IV. Conclusion

We proposed an unsupervised entity alignment framework for cross-lingual KGs with no prior information, reducing the labeling effort. The framework is built on top of a multi-order GCN model that combines both relation and attribute information as well as satisfy the consistency constraints.

References

[1] L. Bellomarini, D. Fakhoury, G. Gottlob, and E. Sallinger, "Knowledge graphs and enterprise ai: the promise of an enabling technology," in *ICDE*, 2019, pp. 26–37.

[2] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *TKDE*, vol. 29, no. 12, pp. 2724–2743, 2017.

[3] M. C. Phan, A. Sun, Y. Tay, J. Han, and C. Li, "Pair-linking for collective entity disambiguation: Two could be better than all," *TKDE*, vol. 31, no. 7, pp. 1383–1396, 2018.

[4] H. Chen, H. Yin, T. Chen, Q. V. H. Nguyen, W.-C. Peng, and X. Li, "Exploiting centrality information with graph convolutions for network representation learning," in *ICDE*, 2019, pp. 590–601.

[5] M. Chen, I. W. Tsang, M. Tan, and T. J. Cham, "A unified feature selection framework for graph embedding on high dimensional data," *TKDE*, vol. 27, no. 6, pp. 1465–1477, 2014.

[6] T. Nguyen, T. Huynh, H. Yin, V. Tong, D. Sakong, B. Zheng, and Q. Nguyen, "Entity alignment for knowledge graphs with multi-order convolutional networks," *TKDE*, vol. 32, no. 13, pp. 1–14, 2021.

[7] H. T. Trung, T. Van Vinh, N. T. Tam, H. Yin, M. Weidlich, and N. Q. V. Hung, "Adaptive network alignment with unsupervised and multi-order convolutional networks," in *ICDE*, 2020, pp. 85–96.

[8] C. T. Duong, H. Yin, D. Hoang, M. H. Nguyen, M. Weidlich, Q. V. H. Nguyen, and K. Aberer, "Graph embeddings for one-pass processing of heterogeneous queries," in *ICDE*, 2020, pp. 1994–1997.