ATTR-INT: A SIMPLE AND EFFECTIVE ENTITY ALIGNMENT FRAMEWORK FOR HETEROGENEOUS KNOWLEDGE GRAPHS

Linyan Yang  Jingwei Cheng*  Chuanhao Xu  Xihao Wang  Jiayi Li  Fu Zhang

School of Computer Science and Engineering, Northeastern University

ABSTRACT

Entity alignment (EA) refers to the task of linking entities in different knowledge graphs (KGs). Existing EA methods rely heavily on structural isomorphism. However, in real-world KGs, aligned entities usually have non-isomorphic neighborhood structures, which paralyses the application of these structure-dependent methods. In this paper, we investigate and tackle the problem of entity alignment between heterogeneous KGs. First, we propose two new benchmarks to closely simulate real-world EA scenarios of heterogeneity. Then we conduct extensive experiments to evaluate the performance of representative EA methods on the new benchmarks. Finally, we propose a simple and effective entity alignment framework called Attr-Int, in which innovative attribute information interaction methods can be seamlessly integrated with any embedding encoder for entity alignment, improving the performance of existing entity alignment techniques. Experiments demonstrate that our framework outperforms the state-of-the-art approaches on two new benchmarks.

Index Terms— Entity Alignment, Knowledge Graph, Representation Learning

1. INTRODUCTION

Recently, knowledge graphs (KGs) have been built and applied in various domains. However, most KGs are constructed by different organizations and individuals, which inevitably brings about heterogeneity problems. Knowledge fusion [1] [2] aims to align and merge heterogeneous and redundant information in KGs to form globally unified knowledge identification and association. Entity Alignment (EA) has been the key technique in knowledge graph fusion. The main purpose of EA is to find equivalent entities among different KGs.

The methods based on knowledge representation learning have gained their popularity in solving the entity alignment problem. These methods assume that equivalent entities from various KGs posses similar neighborhood structures (i.e., isomorphism). However, it is not always the case as different KGs may be highly heterogeneous [3]. Therefore, attempts have been made to alleviate heterogeneity by using the attention mechanism to assign different weights to the relations between entities [4] [5] or by ignoring different neighbors that have a negative impact on the alignment task [6] [7] [8]. The underlying mechanism of these methods to deal with heterogeneity is to transform heterogeneous problems into isomorphic problems, and lack of consideration for the cases that cannot be transformed into isomorphic problems.

Furthermore, recent studies have observed that attribute triples can also provide crucial alignment signals [9]. Most of these methods leverage attribute information by learning attribute name, type and value embeddings, and combine them with structural embeddings to obtain the alignment results [10] [11] [12] [13] [14] [15] [3] [16] [17]. Due to the complexity and diversity of attribute information, there is a great memory and computing power overhead.

To deal with these challenges, we investigate and tackle the problem of entity alignment between heterogeneous KGs. First, we propose two new benchmarks to closely simulate real-world EA scenarios of heterogeneity. Then, we propose a simple and effective entity alignment framework (Attr-Int), in which innovative attribute information interaction methods can be seamlessly integrated with any embedding encoder in entity alignment, improving the performance of original entity alignment techniques. It is worth noting that in our framework, the utilization of attribute information does not require embedding learning, nor does it require any prior alignment of attributes.

In summary, the main contributions are as follows:

- We propose two new benchmarks in which entity pairs share less identical neighbors to reproduce the scenario of EA between heterogeneous KGs, based on DBP15K [10] and OpenEA [18] benchmarks.
- We propose an entity alignment framework called Attr-Int that integrates attribute information in a simple and effective way. In this framework, we provide attribute utilization method, uniqueness of attribute values, and two interaction methods, result correction based on attribute information discrimination and matrix combination based on parameter search.
• We conduct extensive experiments to evaluate representative EA methods on the new benchmarks. Comprehensive experiments demonstrate our framework outperforms the state-of-the-art approaches.

2. PROBLEM DEFINITION

**Definition 1** (Knowledge Graph). A knowledge graph (KG) is denoted as $G = (E, R, A, V, T_r, T_a)$, where $E = \{e_1, e_2, \ldots, e_m\}$, $R = \{r_1, r_2, \ldots, r_n\}$, $A = \{a_1, a_2, \ldots, a_p\}$, $V = \{v_1, v_2, \ldots, v_q\}$, $T_r \subseteq E \times R \times E$ and $T_a \subseteq E \times A \times V$ represent entity set, relation set, attribute set, attribute value set, relation triple set and attribute triple set respectively, and $m, n, p, q$ are the numbers of entities, relations, attributes, and attribute values respectively.

**Definition 2** (Entity Alignment in KGs). Given a source KG $G^1 = \{E^1, R^1, A^1, V^1, T_r^1, T_a^1\}$, and a target KG $G^2 = \{E^2, R^2, A^2, V^2, T_r^2, T_a^2\}$, the aligned entity pairs (training set) is denoted as $P = \{(e^1_i, e^2_j) | e^1_i \in E^1, e^2_j \in E^2, e^1_i \equiv e^2_j\}$, where $\equiv$ stands for equivalence, i.e., the entity $e^1_i$ and the entity $e^2_j$ refer to the same thing in the real world. The goal of the entity alignment task is to find remaining equivalent entity pairs of these two KGs.

**Definition 3** (Coverage Rate of Entity Pairs). Let $(e^1_i, e^2_j)$ be an entity pair, $N(e^1_i)$ and $N(e^2_j)$ be the sets of neighboring entities of $e^1_i$ and $e^2_j$ respectively, then the coverage rate $C(e^1_i, e^2_j)$ of the entity pair $(e^1_i, e^2_j)$ is calculated by $C(e^1_i, e^2_j) = |N(e^1_i) \cap N(e^2_j)| / \min(|N(e^1_i)|, |N(e^2_j)|)$, where $|\cdot|$ represents the size of a set.

3. METHODOLOGY

Our proposed Attr-Int framework is shown in Figure 1, which consists of three major components: encoder, attribute similarity module and interaction module. First, entity embeddings are obtained from the encoder, from which the similarities between entities are calculated, and the preliminary entity alignment result is obtained. Then, the attribute similarity matrix is calculated by the attribute similarity module, and the entity alignment result based on attribute information is obtained. Finally, the interaction module carries out sufficient cross-learning [19] on the above two results to obtain the final result. It is worth noting that, in the attribute similarity module, unlike previous methods leveraging attribute information, there is no need to learn attribute embedding or pre-align any attributes.

3.1. Encoder

Our framework can incorporate any existing entity alignment technique. For the sake of generality, we consider the representative technique (RDGCN [5]) as the encoder in our framework. In order to better incorporate relation information into entity embeddings, RDGCN [5] first constructs a dual relation graph for the input KG, in which vertices denote the relations in the original graph. Then, RDGCN utilizes a graph attention mechanism to encourage interactions between the dual relation graph and the original graph.

3.2. Attribute Similarity Module

We assume that the same entities have similar attribute value information, and the similarity of entity pairs is obtained by calculating the similarity between the attribute values they possess.

We make the following assumption: if an attribute value is observed to occur only once in a KG, then it can be regarded as a crucial attribute value for distinguishing entities. Under this assumption, if two entities from different KGs share a same attribute value, it is determined that these two entities should be aligned.

Formally, given a KG $G = (E, R, A, V, T_r, T_a)$ and $v_i \in V$, if attribute value $v_i$ only occurs once in $G$, such that $(e, a, v_i) \in T_a$, we can consider $v_i$ is a crucial distinguishing attribute value. That is, if $(e^1_i, a^1_j, v_i) \in T_a^1, (e^2_j, a^2_j, v_i) \in T_a^2$, we can get $e^1_i \equiv e^2_j$. Similarly, if the number of occurrences of attribute value $v$ is $k$, and the frequency $v_{freq} = \frac{1}{k}$, then the probability of alignment of two entities with attribute value $v$ is $\frac{1}{k}$. For a given entity pair $(e^1_i, e^2_j)$, we need to obtain the overlapping attribute value sets $(v^1_{freq}, v^2_{freq})$ of the two entities. The similarity of entity pair $S(e^1_i, e^2_j)$ is calculated by $S(e^1_i, e^2_j) = \sum(v^1_{freq}, v^2_{freq})$.

3.3. Interaction module

We propose two interaction approaches in the interaction module: result correction based on attribute information discrimination (RC) and matrix combination based on parameter search (PS) in order to fully combine the information obtained from encoder and attribute similarity module.

3.3.1. Result Correction Based on Attribute Information Discrimination

The encoder gets the entity embedding by training the labeled data, and then gets the preliminary entity alignment result by embedding distance. Attribute similarity module does not require any labeled data training, and attribute similarity of enti-
Fig. 2. Percentage of coverage rate of entity pairs in each stage of the original benchmark datasets and new benchmark datasets. The x-axis represents the coverage rate of entity pairs, while the y-axis represents the proportion of all benchmark datasets.

ties is obtained by comparing the overlapping attribute value sets. Because of the independence of the two modules, the two alignment results are also independent of each other. In order to effectively combine the two results, we retain the result with high confidence in the attribute similarity result, replace the result obtained by the encoder at the same position, and obtain the final result of the entity alignment framework.

3.3.2. Matrix Combination Based on Parameter Search

Inspired by a inference rule in the formal logic NAL [20], we propose another interaction method, which is to combine the entity embedding similarity matrix $S_{EA}$ obtained by encoder and attribute similarity matrix $S_{AT}$, and form the combined similarity matrix that used to evaluate alignment performance. For each pair of test entity $e^1_i \in E^1$ and $e^2_j \in E^2$, $S_{ij}^E$ and $S_{ij}^A$ can be seen as similarity belief statements, as shown in equation (3). In the truth-value, $f_{EA}$ and $f_{AT}$ represent frequency in $S_{EA}$ and $S_{AT}$ respectively, $e_{EA}$ and $e_{AT}$ represent confidence in $S_{EA}$ and $S_{AT}$ respectively. $\leftrightarrow$ represents similarity.

\[
\begin{align*}
  e^1_i \leftrightarrow e^2_j (f_{EA}, e_{EA}) \\
  e^1_i \leftrightarrow e^2_j (f_{AT}, e_{AT})
\end{align*}
\]

Then a revision rule inference can be performed to combine the two beliefs due to approximately disjoint evidence bases. The truth-value function is shown in equation (4), where $f_1/f_2$ and $c_1/c_2$ represents the frequency and confidence of the first/second premise belief, $f$ and $c$ represents those in the conclusion belief.

\[
\begin{align*}
  f &= \frac{f_1 c_1 (1 - c_2) + f_2 c_2 (1 - c_1)}{c_1 (1 - c_2) + c_2 (1 - c_1)} \\
  c &= \frac{c_1 (1 - c_2) + c_2 (1 - c_1)}{c_1 (1 - c_2) + c_2 (1 - c_1) + (1 - c_1)(1 - c_2)}
\end{align*}
\]

Table 1. Details of the new benchmark datasets. Original Rel. Triples represent the triplets in the original benchmark datasets.

| Datasets   | Entities | Rel. | Attr. | Original Rel. Triples | Rel. Triples | Attr. Triples |
|------------|----------|------|-------|------------------------|--------------|--------------|
| DBP15K     | 19370    | 1450 | 1391  | 26638                  | 22368        | 132114       |
| ZH-EN      | 19572    | 1323 | 6933  | 95142                  | 85240        | 567755       |
| FR-EN      | 19814    | 1299 | 5681  | 77241                  | 62737        | 354619       |
| JA-EN      | 19780    | 1153 | 5500  | 93484                  | 83526        | 497230       |
| DBP15K(V1)| 19993    | 1208 | 6161  | 115722                 | 102274       | 576543       |
| EN         | 19587    | 1323 | 6933  | 95142                  | 85240        | 567755       |
| D-W-15K(V1)| 19780    | 1153 | 5500  | 93484                  | 83526        | 497230       |
| EN         | 19587    | 1323 | 6933  | 95142                  | 95240        | 567755       |
| D-Y-15K(V1)| 19993    | 1208 | 6161  | 115722                 | 102274       | 576543       |
| EN         | 19587    | 1323 | 6933  | 95142                  | 85240        | 567755       |

The expectation of the conclusion belief $e^1_i \leftrightarrow e^2_j (f, c)$ is computed by $c \times (f - 1/2) + 1/2$. Finally, the conclusion expectations constitute the elements of the combined similarity matrix. The element of $S_{ij}^E$ and $S_{ij}^A$ corresponds to $f_{ij}^E$ and $f_{ij}^A$, respectively. $c_{ij}^E$ as a parameter represents the overall reliability (or we can call it as the expected amount of evidence to support the belief) of the entity embedding similarity matrix, and analogously $c_{ij}^A$ as another parameter. A parameter grid search for $c_{ij}^E$ and $c_{ij}^A$ is performed on each experiment scenario.

4. EXPERIMENT

4.1. Experimental Settings

4.1.1. Datasets

Table 1 outlines the statistics of new benchmark datasets obtained by lowering the coverage rate of the original benchmarks, DBP15K [10] and two monolingual datasets with sparse version in OpenEA [18]. As shown in Figure 2 (a), we count the percentage of the coverage rate of the entity pairs in each stage of the original benchmark datasets DBP15K, D-W-15K(V1) and D-Y-15K(V1). As we can see, the vast majority of entity pairs have high rates of coverage, which makes it challenging to confirm KG heterogeneity. So we process...
In this paper, we construct new benchmark datasets to closely mimic real-world EA scenarios. Then, we propose a simple and effective entity alignment framework (Attr-Int), in which innovative attribute information interaction methods can be seamlessly integrated with any embedding encoder for entity alignment, improving the performance of original entity alignment techniques. Experimental results show that our framework outperforms the state-of-the-art approaches.

Table 2 compares the overall performance of our method and the baseline methods. BERT-based methods gain the best results which owe to treat entity alignment as the downstream objective to fine-tune a pre-trained BERT model. For BERT-INT, it has a strong dependency on entity name. Since FR-EN (in DBP15K\(_{hs}\)) and D-Y-15K(V1)\(_{hs}\) include well-aligned entity names (which are extracted literally similar), it works well on these datasets as expected. BERT-INT performs poorly on D-W-15K(V1)\(_{hs}\), this is because the two KGs on the datasets do not contain literally matched entity names. SDEA achieves the best results on ZH-EN (in DBP15K\(_{hs}\)) and D-W-15K(V1)\(_{hs}\), which is mainly attributed to the design of identifying various contributions of neighbors and handling alignment of long-tail entities. As a comparison, although our method dose not use BERT to obtain information, we still achieve comparable performance with BERT-INT and SDEA. Moreover, Attr-Int(PC) outperforms BERT-INT (80.1) by 1.7 of hits@1 on JA-EN (in DBP15K\(_{hs}\)).
6. REFERENCES

[1] Tong Shen, Fu Zhang, and Jingwei Cheng, “A comprehensive overview of knowledge graph completion,” *Knowledge-Based Systems*, p. 109597, 2022.

[2] Fu Zhang, Pengpeng Qiu, Tong Shen, and Jingwei Cheng, “Multi-aspect enhanced convolutional neural networks for knowledge graph completion,” in *ECAI 2023*, pp. 2994–3001. IOS Press, 2023.

[3] Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li, “Bert-int: A bert-based interaction model for knowledge graph alignment,” in *IJCAI*, 2020, pp. 420–428.

[4] Y Wu, X Liu, Y Feng, Z Wang, R Yan, and D Zhao, “Relation-aware entity alignment for heterogeneous knowledge graphs,” in *IJCAI*, 2019.

[5] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao, “Mraea: an efficient and robust entity alignment approach for cross-lingual knowledge graph,” in *WSDM*, 2020, pp. 420–428.

[6] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu, “Multi-view knowledge graph embedding for entity alignment,” in *IJCAI*, 2018, vol. 18, pp. 4396–4402.

[7] Renbo Zhu, Meng Ma, and Ping Wang, “Raga: Relation-aware graph attention networks for global entity alignment,” in *PAKDD (1)*. Springer, 2021, pp. 501–513.