An Industry Evaluation of Embedding-based Entity Alignment

Ziheng Zhang[1*], **Jiaoyan Chen**[2*], **Xi Chen**[1*†],
Hualuo Liu[1], **Yuejia Xiang**[1], **Bo Liu**[1], **Yefeng Zheng**[1]
[1]Tencent Jarvis Lab, Shenzhen, China
[2]Department of Computer Science, University of Oxford, UK
{zihengzhang,jasonxchen}@tencent.com, jiaoyan.chen@cs.ox.ac.uk
lhl18@mails.jlu.edu.cn, {yuejiaxiang,raymanliu,yefengzheng}@tencent.com

Abstract

Embedding-based entity alignment has been widely investigated in recent years, but most proposed methods still rely on an ideal supervised learning setting with a large number of unbiased seed mappings for training and validation, which significantly limits their usage. In this study, we evaluate those state-of-the-art methods in an industrial context, where the impact of seed mappings with different sizes and different biases is explored. Besides the popular benchmarks from DBpedia and Wikidata, we contribute and evaluate a new industrial benchmark that is extracted from two heterogeneous knowledge graphs (KGs) under deployment for medical applications. The experimental results enable the analysis of the advantages and disadvantages of these alignment methods and the further discussion of suitable strategies for their industrial deployment.

1 Introduction

Knowledge graphs (KGs), such as DBpedia (Auer et al., 2007), Wikidata (Vrandečić and Krötzsch, 2014) and YAGO (Suchanek et al., 2007) are playing an increasingly important role in various applications such as question answering and search engines. The construction of KGs usually includes several components, such as Named Entity Recognition (NER) (Li et al., 2018), Relation Extraction (RE) (Zhang et al., 2019a), and Knowledge Correction (Chen et al., 2020). However, the content of an individual KG is often incomplete, leading to a limited knowledge coverage especially in supporting applications of a specific domain (Färber et al., 2018; Demartini, 2019). One widely adopted solution is to merge multiple KGs (e.g., an enterprise KG with fine-grained knowledge of a specific domain and a general-purpose KG with an extensive coverage) with the assistance of an alignment system which discovers cross-KG mappings of entities, relations, and classes (Otero-Cerdeira et al., 2015; Yan et al., 2016).

Embedding-based *entity alignment* has recently attracted more attention due to the popularity of KGs with big data (i.e. a large number of facts) such as Wikidata. Traditional alignment systems such as PARIS (Suchanek et al., 2011) and LogMap (Jiménez-Ruiz and Grau, 2011), which usually reply on lexical matching and semantic reasoning (e.g., for checking the violation of relation domain and range), are believed to be weak in utilizing the contextual semantics especially the graph structure of such large KGs. To address this problem, some novel embedding-based methods have been proposed with the employment of different KG embedding methods such as TransE (Bordes et al., 2013) and Graph Neural Networks (GNNs) (Scarselli et al., 2008) as well as some algorithms from active learning (Berrendorf et al., 2020), multi-view learning (Zhang et al., 2019b) and so forth.

We find all these embedding-based entity alignment methods rely upon *seed mappings* for supervision or semi-supervision in training. They are usually evaluated by benchmarks extracted from DBpedia, Wikidata and YAGO, all of which are constructed from the same source, namely Wikipedia. These methods typically build their models with 30% (or even higher) of all the ground-truth mappings, and the training and validation sets are randomly extracted, sharing the same distribution as the test set.

[*] The first three authors contributed equally.

[†] Xi Chen is the corresponding author.

Proceedings of the 28th International Conference on Computational Linguistics: Industry Track, pages 179–189
Barcelona, Spain (Online), December 12, 2020

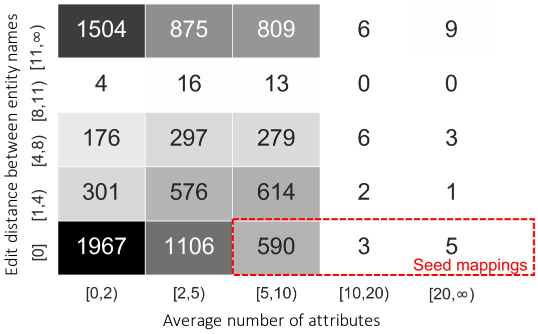


Figure 1: Distribution of mappings of two sampled medical KGs. The horizontal axis denotes the average number of attributes and the vertical axis denotes the edit distance between entity names.

In industrial applications, however, such seed mappings require not only expertise but also much human labour for annotation, especially when the two large KGs come from totally different sources. Even though a small number of seed mappings can be annotated, they are usually biased in comparison with the remaining for prediction with respect to entity name, attribute, graph structure and so on. Figure 1 shows the distribution of all the mappings of two sampled medical KGs from Tencent Technology (cf. Section 3.1 for more details), with two dimensions – the similarity between names of mapping entities and the average attribute number of mapping entities. When we directly invited experts or utilized downstream applications to annotate mappings, the annotated mappings, which could act as the seed mappings for training, usually lie in the bottom right area (seen in the red block in Figure 1) with high name similarity and large attribute number. Thus, we believe that the seed mappings should have the following characteristics to make the evaluation of these supervised methods more practical. Firstly, the seed mappings should take a small proportion of all the mappings, such as 3% that is far smaller than previous experimental settings. Secondly, the seed mappings should be biased towards the remaining mappings with respect to the entity name similarity, the average attribute number, or both. Such biases are ignored in the current evaluation.

In this work, we systematically evaluate four state-of-the-art embedding-based KG alignment methods in an industrial context. The experiment is conducted with one open benchmark from DBpedia and Wikidata, one industry benchmark from two enterprise medical KGs with heterogeneous contents, and a series of seed mappings with different sizes, name biases and attribute biases. The performance analysis considers all the testing mappings as well as different splits of them for fine-grained observations. These methods are also compared with the traditional system PARIS. To the best of our knowledge, this is the first work to evaluate and analyse the embedding-based entity alignment methods from an industry perspective. We find that these methods heavily rely on an ideal supervised learning setting and suffer from a dramatic performance drop when being tested in an industrial context. Based on these results, we can further discuss the possibility to deploy them for real-world applications as well as suitable sampling strategies. The new benchmark and seed mappings can also benefit the research community for future studies, which are publicly available at `https://github.com/ZihengZZH/industry-eval-EA`.

2 Preliminaries and Related Work

2.1 Embedding-based Entity Alignment

Most of the existing embedding based entity alignment methods conform to the following three-step paradigm: *(i)* embedding the entities into a vector space by either a translation based method such as TransE (Bordes et al., 2013) or Graph Neural Networks (GNNs) (Scarselli et al., 2008) which recursively aggregate the embeddings of the neighbouring entities and relations; *(ii)* mapping the entity embeddings in the space of one KG to the space of another KG by learning a transformation matrix, sharing embeddings of the aligned entities, or swapping the aligned entities in the associated triples; *(iii)* searching an entity's counterpart in another KG by calculating the distance in the embedding space using metrics such as the cosine similarity. It is worth noting that the role of the seed mappings mainly lies in the second step, aligning the embeddings of two KGs.

Specifically, we evaluate four methods, namely **BootEA** (Sun et al., 2018), **MultiKE** (Zhang et al., 2019b), **RDGCN** (Wu et al., 2019) and **RSN4EA** (Guo et al., 2018). On the one hand, they have achieved the state-of-the-art performance in the ideal supervised learning setting, according to their own evaluation and the benchmarking study (Sun et al., 2020); on the other hand, they are representative to different techniques that are widely used in the literature. The four methods are introduced as follows.

BootEA is a semi-supervised approach, which adopts translation-based models for embedding and iteratively trains a classifier by bootstrapping. In each iteration, new likely mappings are labelled by the classifier and those causing no conflict are added for training in the following iteration.

MultiKE utilizes multi-view learning to encode different semantics into the prediction model. Specifically, three views are developed for entity names, entity attributes, and the graph structure respectively.

RDGCN applies a GCN variant, Dual-Primal GCN (Monti et al., 2018) to utilize the relation information in KG embedding. It can better utilize the graph structure than those translation-based embedding methods, especially in dealing with the triangular structures.

RSN4EA firstly generates biased random walks (long paths) of both KGs as sequences and then learns the embeddings by a sequential model named Recurrent Skipping Network. The seed mappings here are used to generate cross-KG walks, thus exploring correlations between cross-KG entities.

2.2 Seed Mappings

As far as we know, the current embedding-based entity alignment methods mostly rely on the seed mappings, whose roles are introduces in Section 2.1, for supervised or semi-supervised learning. Specially, we can consider some heuristic rules with, for example, string and attribute matching to generate the seed mappings, as done by the method IMUSE (He et al., 2019), but the impact of the seed mappings is similar and the study of such impact also benefit the distant supervision methods.

In addition, although some semi-supervised approaches such as BootEA (Sun et al., 2018) and SEA (Pei et al., 2019) are less dependent on the seed mappings, their performance, when trained on a small set of seed mappings, may vary from data to data and be impacted by the bias of the seed mappings.

In the own evaluation of these methods and the recent benchmark study (Sun et al., 2020), 20% and 10% of all the ground truth mappings are used for training and validation respectively, and more importantly, they are randomly selected, thus maintaining the same distribution as the testing mappings. This violates the real-world scenarios in the industry, where annotating seed mappings is costly and the annotated ones are usually biased, as discussed in Section 1. Actually, there are relatively few studies that investigate the seed mappings and those investigated only consider the proportion of the seeding mappings. In Sun et al. (2018) and Wu et al. (2019), the proposed methods are evaluated with the proportion of the seed mapping for training varying from 10% to 40%. However, the minimum proportion still leads to a very large number (e.g., 1.5K) of seed mappings in aligning two big KGs.

2.3 Benchmarks

The current benchmarks used to evaluate the embedding-based methods are typically extracted from DBpedia, Wikidata, and YAGO. They can be divided into two categories. The first includes those for cross-lingual entity alignment such as DBP15K (Sun et al., 2017) and WK3l60k (Chen et al., 2018), both of which support the alignment between DBpedia entities in English and DBpedia entities in other languages, such as Chinese or French. These benchmarks usually only support within KG alignment. The second includes those for cross-KG entity alignment such as DWY15K (Guo et al., 2018) and DWY100K (Sun et al., 2018), both of which are for the alignment between DBpedia and Wikidata/YAGO.

As discussed in Sun et al. (2020), entities in these aforementioned benchmarks have a significant bias in comparison with normal entities in the original KGs; for example, those DBpedia entities in WK3l60k have an average connection degrees of 22.77 while that of all DBpedia entities is 6.93. Thus, these benchmarks are not representative to DBpedia, Wikidata, and YAGO. To address this issue, Sun et al. (2020) proposed a new iterative degree-based sampling algorithm to extract new benchmarks for both cross-lingual entity alignment within DBpedia and cross-KG entity alignment between DBpedia and Wikidata/YAGO. Although the new benchmarks are more representative w.r.t. the graph structure, the entity labels defined by *rdfs:label* are removed, which include important name information, which makes them less representative to real-world alignment contexts. More importantly, since DBpedia, Wikidata, and YAGO are constructed from the same source Wikipedia, the entities for alignment often have similar names, attributes, or graph structures. These benchmarks are therefore not applicable in the real-world alignment which in contrast, aims at KGs from different sources to complement each other. To make an industry evaluation, we constructed a new benchmark from two industrial KGs (cf. Section 3.1).

It is worth noting that Ontology Alignment Evaluation Initiatives[1] has been organizing a KG track since 2018 (Hertling and Paulheim, 2020). The benchmarks used are those KGs extracted from several different Wikis from Fandom;[2] for example, starwars-swg is a benchmark with mappings between two KGs from Star Wars Wiki and Star Wars Galaxies Wiki. Multiple benchmarks are adopted, but their scales are limited; for example, 4 out of 5 used in 2019 have less than 2K entity mappings. As the two KGs of a benchmark are about two hubs of one concrete topic (such as the movie and the game of Star Wars), the entity name has little ambiguity and becomes a superior indicator for alignment. Thus they are not suitable industrial benchmarks for evaluating the embedding-based entity alignment methods.

3 Data Generation

3.1 Industrial Benchmark

To evaluate the embedding-based entity alignment methods in an industrial context as discussed above, we first extract a benchmark from two real-world medical KGs for alignment. One KG is built upon multiple authoritative medical resources, covering fine-grained knowledge about illness, symptoms, medicine, etc. It is deployed to support applications such as question answering and medical assistants in our company. However, some of its entities have incomplete information with many important attributes missing, which limits its usability. We extract around 10K such entities according to the feedback from downstream applications. They are then aligned with another KG to improve the information completeness. That KG is extracted from the information boxes of Baidu Baike[3], the largest Chinese encyclopedia, via NLP techniques (such as NER and RE) as well as some handcrafted engineering work. We refer to crowdsourcing for annotating the mappings, where heuristic rules, based on labels and synonyms, and a friendly interface for supporting information check are used for assistance. Finally, we obtain $9,162$ one-to-one entity mappings, based on which one sub-KG is extracted from one original KG. Specifically, the sub-KG includes triples that are composed of entities associated with these mappings. The two sub-KGs are named as MED and BBK, and the new benchmark is named as MED-BBK-9K.

Table 1: Statistics of MED-BBK-9K and D-W-15K.

Benchmark	KGs	#Entities	Relation			Attribute		
			#Relations	#Triples	Degree	#Attributes	#Triples	Degree
MED-BBK-9K	MED	9,162	32	158,357	34.04	19	11,467	1.24
	BBK	9,162	20	50,307	10.96	21	44,987	4.91
D-W-15K	DBpedia	15,000	167	73,983	8.55	175	66,813	4.40
	Wikidata	15,000	121	83,365	10.31	457	175,686	11.59

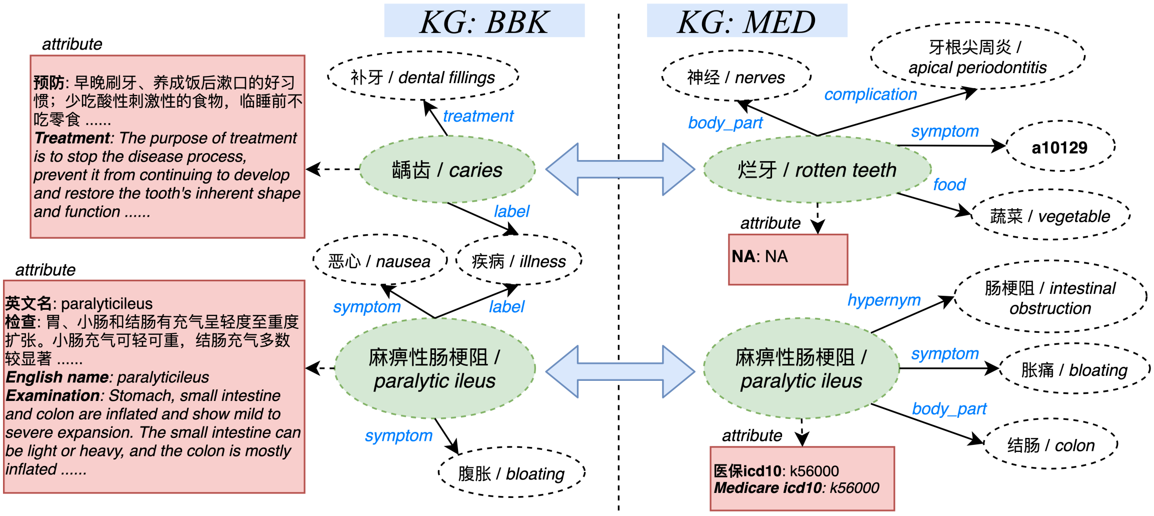
More details of MED-BBK-9K and another benchmark D-W-15K, which is extracted by the iterative degree-based sampling method under the setting of V2 (Sun et al., 2020), are shown in Table 1, where # denotes the number and degree is the rate between the triple number and the entity number. Statistics of relation triples and attribute triples are separately presented in Table 1. Note that a relation is equivalent to an object property connecting two entities, while an attribute is equivalent to a data property associating an entity with a value of some data type. Two entity mapping examples of MED-BBK-9K are depicted in Figure 2, where the green ellipses indicate the aligned entities across KGs, the white ellipses and the solid arrows indicate their relation triples[4], and the red rectangles and the dash arrows indicate the attributes

[1] http://oaei.ontologymatching.org/

[2] http://www.fandom.com/

[3] https://baike.baidu.com/

[4] *label* here indicates a specific relation. Please do not be confused with *rdfs:label* of the W3C standard.



预防: 早晚刷牙、养成饭后漱口的好习惯；少吃酸性刺激性的食物，临睡前不吃零食 ......
Treatment: *The purpose of treatment is to stop the disease process, prevent it from continuing to develop and restore the tooth's inherent shape and function ......*

英文名: paralyticileus
检查: 胃、小肠和结肠有充气呈轻度至重度扩张。小肠充气可轻可重，结肠充气多数较显著 ......
English name: *paralyticileus*
Examination: *Stomach, small intestine and colon are inflated and show mild to severe expansion. The small intestine can be light or heavy, and the colon is mostly inflated ......*

Figure 2: Two mapping examples from MED-BBK-9K with *English translations*.

which include normal values, sentence descriptions, and noisy values. Through the statistics and the examples, we can conclude that KGs in MED-BBK-9K are quite different from KGs in D-W-15K, with a higher relation degree, less attributes, higher heterogeneity, etc.

3.2 Biased Seed Mappings

Besides the industrial benchmark, we also develop a new approach to extract biased seed mappings for the industrial context. We first introduce two variables, s_{name} and n_{attr}, in which s_{name} is the normalized Levenshtein Distance – an edit distance metric (Navarro, 2001) in $[0, 1]$ for the name strings of entities of each mapping, and n_{attr} is the average number of attributes of entities of each mapping. For Wikidata entities in D-W-15K, we use the attribute values of *P373* and *P1476* as the entity names, while for DBpedia entities we use the entity name in the URI. Note when one or both entities in one mapping has multiple names, we adopt the two names leading to the highest similarity i.e., the lowest s_{name}. Meanwhile, all the names are pre-processed before calculating s_{name}: dash, underline and backslash are replaced by the white space, punctuation marks are removed, letters are transformed into lowercase.

With s_{name} and n_{attr} calculated, we divide all the mappings into three different splits according to either the name similarity or the attribute number. For the name similarity, the mappings are divided into "same" (s_{name}=1.0), "close" ($s_{name} < 1.0$) and "different" (s_{name} is NA, i.e., no valid entity name) for both MED-BBK-9K and D-W-15K. From the attribute number, the mappings are divided into "large" ($n_{attr} \geq k_1$), "medium" ($k_2 \leq n_{attr} < k_1$) and "small" ($n_{attr} < k_2$), where (k_1, k_2) are set to $(5, 2)$ for MED-BBK-9K and set to $(10, 4)$ for D-W-15K.

We further develop an iterative algorithm to extract the seed mappings with name bias and attribute bias. Its steps are shown below, with two inputs, namely the set of all the mappings $\mathcal{M}_{all}$ and the size of seed mappings N_{seed}, and one output, namely the set of biased seed mappings $\mathcal{M}_{seed}$.

(1) Initialize the biased seed mapping set $\mathcal{M}_{seed}$.

(2) Assign each mapping in $\mathcal{M}_{all}$ a score: $z = z_{name} + z_{attr}$, where z_{name} is set to 4, 3 and 1 if the mapping belongs to "same", "close" and "different" respectively, and z_{attr} is set to 4, 3 and 1 if the mapping belongs to "large", "medium" and "small" respectively. Note all the mappings in $\mathcal{M}_{all}$ are assigned a score of 8, 7, 6, 5, 4, or 2.

(3) Move the mapping with the highest score in $\mathcal{M}_{all}$ to $\mathcal{M}_{seed}$. Randomly select one if multiple mappings in $\mathcal{M}_{all}$ have the highest score.

(4) Check whether the size of $\mathcal{M}_{seed}$ has been equal to or larger than N_{seed}. If yes, return $\mathcal{M}_{seed}$; otherwise, go to Step (3).

With the above procedure, we can also obtain seed mappings that are name biased alone by setting $z = z_{name}$, and seed mappings that are attribute biased alone by setting $z = z_{attr}$. Note the seed mappings $\mathcal{M}_{seed}$ include both training mappings and validation mappings. In our experiment, the former occupies two thirds of the seed mappings while the latter occupies one third.

4 Evaluation

4.1 Experimental Setting

We first conduct the overall evaluation (cf. Section 4.2). Specifically, the methods BootEA, MultiKE, RDGCN, and RSN4EA are tested under *(i)* an **industrial context** where the seed mappings are both name biased and attribute biased, and the rate of training (resp. validation) mappings is 2% (resp. 1%), and *(ii)* an **ideal context** where the seed mappings are randomly selected without bias, and the rate of training (resp. validating) mappings is 20% (resp. 10%). We then conduct ablation studies where three impacts of seed mappings are independently analysed, including size, name bias, and attribute bias.

In both overall evaluation and ablation studies, we calculate metrics Hits@1, Hits@5, and mean reciprocal rank (MRR) with all the testing mappings. For each testing mapping, the candidate entities (i.e., all the entities in the target KG) are ranked according to their predicted scores; Hits@1 (resp. Hits@5) is the ratio of testing mappings whose ground truths are ranked in the top 1 (resp. 5) entities; MRR is the Mean Reciprocal Rank of the ground truth entity. Meanwhile, to further analyse the impact of the seed mappings on different kinds of testing mappings, we divide the testing mappings into two three-fold splits – "same", "close" and "different" from the name biased aspect, and "small", "medium" and "large" from the attribute biased aspect.

We adopt the implementation of BootEA, MultiKE, RDGCN, and RSN4EA in OpenEA, while their hyperparameters are adjusted with the validation set. Specifically, the batch size is set to 5000, the early stopping criterion is set to when Hits@1 begins to drop on the validation set (checked for every 10 epochs), the maximum epoch number is set to 2000. As MultiKE and RDGCN utilize literals, the word embeddings are produced using a fastText model pre-trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset[5]. To run them on MED-BBK-9K, the Chinese word embeddings are obtained via a medical-specific BERT model pre-trained on big medical corpora from Tencent Technology[6].

We finally compare these embedding-based methods with a state-of-the-art conventional system named PARIS (v0.3)[7], which is based on lexical matching and iterative calculation of relation mappings, class mappings and entity mappings with their correlations (logic consistency) considered (Suchanek et al., 2011). We adopt the default hyperparameters to PARIS. Note that PARIS requires no seed mappings for supervision. As PARIS does not rank all the candidate entities, we use Precision, Recall, and F1-score as the evaluation metrics. For the embedding-based methods, Hits@1 in our one-to-one mapping evaluation is equivalent to Precision, Recall, and F1-score.

4.2 Overall Results

Table 2 presents the results of those embedding-based methods on both D-W-15K and MED-BBK-9K under the ideal context and the industrial context. On one hand, we find that *the performance of all four methods dramatically decreases when the testing context is moved from the ideal to the industrial*, the latter of which is much more challenging with less and biased seed mappings. For instance, considering the average MRR of all four methods on all testing mappings, it drops from 0.661 to 0.262 on D-W-15K, and from 0.327 to 0.118 on MED-BBK-9K.

We also find that the performance decreasement, when moved to the industrial context, varies from one testing mapping split to another. Considering the name-based splitting, the decreasement is the most significant on the "different" split, and the least significant on the "same" split. Take MultiKE on MED-BBK-9K as an example, its Hits@1 decreases by 11.4%, 13.9% and 43.1% on the "same", "close" and "different" splits respectively. As a result, the methods including MultiKE and RDGCN perform better on the "same" split than on the "close" and the "different" splits. It meets our expectations because the seed mappings in the industrial context, which are sampled with a bias toward those with high name

[5] The word embeddings are publicly available at `https://fasttext.cc/docs/en/english-vectors.html`.

[6] Other Chinese word embedding models would suffice to reproduce comparable experimental results.

[7] `http://webdam.inria.fr/paris/`

Table 2: Overall results under the ideal context and the industrial context.

		Models	Name-based Splits (Hits@1)			Attr-based Splits (Hits@1)			All Test Mappings		
			Same	Close	Diff.	Small	Medium	Large	Hits@1	Hits@5	MRR
D-W-15K	Ideal	BootEA	.868	.902	.753	.721	.821	.912	.818	.922	.864
		MultiKE	.977	.254	.216	.306	.488	.661	.484	.622	.554
		RDGCN	.942	.934	.305	.330	.734	.827	.629	.756	.687
		RSN4EA	.718	.718	.579	.536	.663	.753	.650	.797	.717
	Industrial	BootEA	.050	.051	.023	.015	.040	.053	.037	.092	.065
		MultiKE	.968	.211	.036	.086	.392	.605	.368	.426	.402
		RDGCN	.945	.872	.062	.110	.559	.759	.489	.539	.514
		RSN4EA	.055	.060	.029	.016	.046	.065	.043	.092	.068
MED-BBK-9K	Ideal	BootEA	.334	.259	.328	.388	.201	.265	.307	.495	.399
		MultiKE	.342	.173	.072	.269	.149	.195	.213	.367	.289
		RDGCN	.550	.217	.056	.348	.270	.242	.306	.425	.365
		RSN4EA	.238	.121	.226	.277	.114	.095	.195	.311	.253
	Industrial	BootEA	.006	.003	.003	.006	.002	.004	.004	.011	.010
		MultiKE	.303	.149	.041	.218	.137	.155	.179	.322	.252
		RDGCN	.329	.083	.013	.201	.120	.086	.158	.239	.199
		RSN4EA	.008	.002	.007	.009	.001	.000	.005	.013	.011

similarity, are close to the "same" split and far away from the "different" split. However, such a regular is violated when we consider the attribute based seed mapping splits. As to MultiKE tested by the "large" testing split, its performance decreasement when moved to the industrial context is the least significant on D-W-15K, which is as expected, but is the most significant on MED-BBK-9K. Thus MultiKE performs worse on the "large" testing split than on the "small" testing split (with 28.9% lower Hits@1), although the former is more close to the seed mappings. One potential explanation is that mappings with more than 5 attributes (mappings in the "large" testing split) in MED-BBK-9K tend to have duplicate attributes and some attribute values are sentences that cannot be fully utilized by these methods.

On the other hand, we find that *MultiKE and RDGCN are much more robust than BootEA and RSN4EA in the industrial context on both D-W-15K and MED-BBK-9K*. Although MultiKE and RDGCN do not perform as well as in the ideal context, their performance is still promising. Specifically, when measured by all testing mappings, RDGCN performs better than MultiKE on D-W-15K with 27.9% higher MRR and 32.9% higher Hits@1 but performs worse than MultiKE on MED-BBK-9K with 21.3% lower MRR and 11.7% lower Hits@1. The performance of BootEA and RSN4EA is poor in the industrial context; their Hits@1, Hits@5, and MRR on all testing mappings or on different testing splits are all lower than 0.1 for both benchmarks. This means that they are very sensitive to the size or/and the bias of the seed mappings (cf. Section 4.3 for the ablation studies).

4.3 Ablation Studies

4.3.1 Size Impact

According to the results in the "With No Bias" setting in Table 3, we can first find that *MultiKE and RDGCN are relatively robust w.r.t. a small training mapping size*. Considering their Hits@1 measured on all the test mappings, it drops slightly from 0.484 to 0.394 and from 0.629 to 0.513 respectively when the training mapping size is significantly reduced from 20% to 2%. On the "same" testing split and the "large" testing split, both of which are close to the training mappings, the performance of MultiKE and RDGCN keeps relatively good when trained by 2% of the mappings. On the other two splits, which are more biased compared with training mappings, the performance of MultiKE and RDGCN, however, decreases more significantly.

Furthermore, we find that *BootEA and RSN4EA are very sensitive to the training mapping size.* For example, the MRR of BootEA (resp. RSN4EA) measured by all the test mappings decreases from 0.864 to 0.153 to 0.051 (resp. from 0.717 to 0.132 to 0.044) when the training ratio decreases from 20% to 4% to 2%. The performance of BootEA is beyond our expectation as it is a semi-supervised algorithm designed for a limited number of training samples. Besides all the testing mappings, their performance decreasement is also quite significant on different testing splits including the "same" and the "large".

Table 3: Results on D-W-15K under different settings (biases and ratios) of the training mappings.

Settings		Models	Name-based Splits (Hits@1)			Attr-based Splits (Hits@1)			All Test Mappings		
			Same	Close	Diff.	Small	Medium	Large	Hits@1	Hits@5	MRR
With No Bias	20%	BootEA	.868	.902	.753	.721	.821	.912	.818	.922	.864
		MultiKE	.977	.254	.216	.306	.488	.661	.484	.622	.554
		RDGCN	.942	.934	.305	.330	.734	.827	.629	.756	.687
		RSN4EA	.718	.718	.579	.536	.663	.753	.650	.797	.717
	4%	BootEA	.104	.087	.092	.078	.085	.125	.096	.206	.153
		MultiKE	.975	.217	.088	.159	.440	.647	.413	.513	.467
		RDGCN	.898	.901	.123	.163	.650	.754	.521	.605	.562
		RSN4EA	.105	.079	.090	.071	.078	.133	.093	.168	.132
	2%	BootEA	.024	.022	.030	.028	.025	.026	.026	.073	.051
		MultiKE	.969	.224	.048	.121	.428	.639	.394	.463	.433
		RDGCN	.900	.895	.107	.147	.636	.761	.513	.582	.547
		RSN4EA	.026	.015	.031	.025	.021	.034	.027	.056	.044
With Name Bias	20%	BootEA	.871	.903	.535	.433	.737	.931	.645	.766	.702
		MultiKE	.978	.285	.080	.085	.230	.318	.185	.335	.261
		RDGCN	.966	.924	.111	.102	.521	.641	.362	.441	.402
		RSN4EA	.786	.800	.391	.271	.631	.827	.514	.656	.580
	4%	BootEA	.733	.817	.358	.260	.633	.802	.554	.642	.596
		MultiKE	.971	.209	.053	.106	.391	.609	.358	.427	.398
		RDGCN	.956	.905	.076	.128	.616	.766	.491	.544	.518
		RSN4EA	.198	.185	.087	.051	.147	.228	.138	.228	.182
	2%	BootEA	.031	.031	.017	.013	.026	.034	.024	.069	.049
		MultiKE	.968	.195	.027	.093	.389	.617	.360	.404	.388
		RDGCN	.956	.871	.056	.118	.606	.766	.490	.541	.516
		RSN4EA	.054	.040	.027	.018	.036	.062	.038	.084	.062
With Attribute Bias	20%	BootEA	.789	.870	.397	.365	.734	.936	.565	.682	.621
		MultiKE	.975	.358	.078	.145	.488	.767	.334	.459	.398
		RDGCN	.946	.919	.109	.168	.667	.885	.437	.522	.479
		RSN4EA	.725	.816	.309	.277	.670	.834	.489	.611	.546
	4%	BootEA	.704	.819	.337	.245	.622	.800	.538	.611	.574
		MultiKE	.972	.211	.057	.115	.430	.662	.383	.450	.421
		RDGCN	.922	.908	.091	.133	.630	.798	.501	.557	.529
		RSN4EA	.192	.213	.083	.056	.156	.228	.141	.232	.185
	2%	BootEA	.052	.051	.023	.017	.039	.059	.037	.094	.066
		MultiKE	.968	.229	.041	.104	.426	.651	.384	.449	.421
		RDGCN	.915	.895	.078	.122	.615	.785	.497	.552	.524
		RSN4EA	.068	.073	.027	.018	.050	.083	.049	.096	.073

4.3.2 Name Bias Impact

The name bias impact from the seed mappings can be evaluated by comparing the settings of "With Name Bias" and "With No Bias" in Table 3. With 20% of the mappings for training, MultiKE and RDGCN are more negatively impacted by the name bias than BootEA and RSN4EA; for example, the MRR measured by all the test mappings drops by 52.9% and 41.5% respectively, while that of BootEA and RSN4EA drops only by 18.8% and 19.1% respectively.

Specifically, considering different testing mapping splits, the negative impact on MultiKE and RDGCN mainly lies in the "different" split (e.g., Hits@1 of RDGCN drops from 0.305 to 0.111), while the impact on the "same" and the "close" is relatively limited and sometimes even positive. Mappings in the "different" testing split, which have very biased distributions as the training mappings, are sometimes known as long-tail prediction cases, and the above phenomena indicate their universality and difficulty in an industrial context. On the other hand, the negative impact of name bias on MultiKE and RDGCN is still much less than the negative impact of the small size on BootEA and RSN4EA. Thus when impacted by both small size (using 2% of the mappings for training) and name bias, BootEA and RSN4EA perform poorly. It is also worth noting that RDGCN outperforms other methods by a large margin in the "close" split under all the experimental settings; for example, its Hits@1 reaches 0.905 and 0.871 with 4% and 2% training mappings while that for MultiKE is only 0.209 and 0.195 respectively.

4.3.3 Attribute Bias Impact

The attribute bias impact from the seed mappings can be analysed by comparing the settings of "With Attribute Bias" and "With No Bias" in Table 3. When 20% mappings are used for training, its negative impact on all four methods are similar; for example, the MRR of BootEA, MultiKE, RDGCN, and RSN4EA on all testing mappings drops by 28.1%, 28.2%, 30.3%, and 23.8% respectively. The negative impact is especially significant on the "small" testing split as its average attribute number is very different from that of the training mappings. In contrast, the impact on the "large" testing split is even positive for all four methods; for example, when trained by 4% of the mappings, Hits@1 of RSN4EA increases from 0.133 to 0.228. Especially, under the attribute bias, reducing the training mappings size has limited impact on MultiKE and RDGCN, and sometimes the impact is even positive that for example, the MRR of MultiKE and RDGCN on all testing mappings increases by 5.8% and 10.4% respectively when the training mapping ratio drops from 20% to 4%.

4.4 Comparison with Conventional System

This subsection presents the comparison between the embedding-based methods and the conventional system PARIS (Suchanek et al., 2011), using results in both Table 2 and Table 4. Note that Hits@1 in Table 2 is equivalent to Precision, Recall, and F1-Score in our evaluation with all one-to-one mappings. Although PARIS is an automatic system needing no supervision, it still significantly outperforms all four embedding based methods on both D-W-15K and MED-BBK-9K. On MED-BBK-9K whose two KGs for alignment are more heterogeneous, the outperformance of PARIS is even more significant; for example, the F1-score of PARIS is 0.493, while the best of the four embedding based methods is 0.307 (resp. 0.179) when trained in the ideal (resp. industrial) context. One important reason we believe is that these embedding based methods ignore the overall reasoning and the correlation of different mappings, while PARIS utilizes them by an iterative workflow and makes holistic decisions. Luckily, such reasoning capability and inter-mapping correlations can also be considered in the embedding-based methods, and this indicates an important direction for the future industrial application.

Table 4: Results of conventional system PARIS on D-W-15K and MED-BBK-9K.

Benchmark	Metric	Name-based Splits			Attr-based Splits			All Test Mappings
		Same	Close	Diff.	Small	Medium	Large	
D-W-15K	Precision	.998	.998	.900	.868	.980	.999	.956
	Recall	.980	.975	.707	.640	.914	.987	.846
	F1-score	.989	.986	.792	.736	.946	.993	.898
MED-BBK-9K	Precision	.910	.669	.778	.879	.748	.757	.814
	Recall	.505	.248	.258	.417	.293	.314	.354
	F1-score	.649	.362	.388	.565	.422	.444	.493

5 Conclusion and Discussion

In this study, we evaluate four state-of-the-art embedding-based entity alignment methods in an ideal context and an industrial context. To build the industrial context, a new benchmark is constructed with two real-world KGs, and the seed mappings are extracted with different sizes, different name and attribute biases. The performance of all four investigated methods dramatically drops when being evaluated in the industrial context, worse than the traditional system PARIS. Specifically, MultiKE and RDGCN are sensitive to name and attribute bias but robust to seed mapping size; BootEA and RSN4EA are extremely sensitive to seed mappings size, leading to poor performance in the industrial context.

Based on these empirical findings, we recommend to specifically design strategies in crowdsourcing (with tool assistance) to ensure the annotated samples in different name and attribute distributions. In our industrial context where the seed mappings are limited, adopting MultiKE or RDGCN is demonstrated to be a better choice for cross-KG alignments. Meanwhile, as mentioned in the evaluation, an ensemble of such embedding based methods with PARIS or LogMap, which considers the correlation between mappings, is also a promising solution for better performance. Finally, we also plan to develop a robust model that can utilize a complete set of attributes, especially those with values of textual descriptions.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer.

Max Berrendorf, Evgeniy Faerman, and Volker Tresp. 2020. Active learning for entity alignment. *arXiv Preprint*, January. arXiv: 2001.08943.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795.

Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2018. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3998–4004.

Jiaoyan Chen, Xi Chen, Ian Horrocks, Erik B. Myklebust, and Ernesto Jimenez-Ruiz. 2020. Correcting knowledge base assertions. In *Proceedings of The Web Conference 2020*, WWW '20, page 1537–1547, New York, NY, USA. Association for Computing Machinery.

Gianluca Demartini. 2019. Implicit bias in crowdsourced knowledge graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 624–630.

Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. 2018. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1):77–129.

Lingbing Guo, Zequn Sun, Ermei Cao, and Wei Hu. 2018. Recurrent skipping networks for entity alignment. *arXiv Preprint arXiv:1811.02318*.

Fuzhen He, Zhixu Li, Yang Qiang, An Liu, Guanfeng Liu, Pengpeng Zhao, Lei Zhao, Min Zhang, and Zhigang Chen. 2019. Unsupervised entity alignment using attribute triples and relation triples. In *International Conference on Database Systems for Advanced Applications*, pages 367–382. Springer.

Sven Hertling and Heiko Paulheim. 2020. The knowledge graph track at OAEI. In *European Semantic Web Conference*, pages 343–359. Springer.

Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. 2011. LogMap: Logic-based and scalable ontology matching. In *International Semantic Web Conference*, pages 273–288. Springer.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2018. A Survey on Deep Learning for Named Entity Recognition. *CoRR*, abs/1812.09449.

Federico Monti, Oleksandr Shchur, Aleksandar Bojchevski, Or Litany, Stephan Günnemann, and Michael M Bronstein. 2018. Dual-primal graph convolutional networks. *arXiv preprint arXiv:1806.00770*.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

Lorena Otero-Cerdeira, Francisco J. Rodríguez-Martínez, and Alma Gómez-Rodríguez. 2015. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949–971.

Shichao Pei, Lu Yu, Robert Hoehndorf, and Xiangliang Zhang. 2019. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *The World Wide Web Conference*, pages 3130–3136.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. 2011. PARIS: Probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment*, 5(3).

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *International Semantic Web Conference*, pages 628–644. Springer.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4396–4402.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A Benchmarking Study of Embedding-Based Entity Alignment for Knowledge Graphs. *Proc. VLDB Endow.*, 13(12):2326–2340, July.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57(10):78–85.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5278–5284.

Zhuang Yan, Li Guoliang, and Feng Jianhua. 2016. A survey on entity alignment of knowledge base. *Journal of Computer Research and Development*, 1:165–192.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019a. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3016–3025, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019b. Multi-view knowledge graph embedding for entity alignment. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5429–5435.