Aligning Cross-Lingual Entities with Multi-Aspect Information

Hsiu-Wei Yang[1*], **Yanyan Zou**[2*], **Peng Shi**[1*], **Wei Lu**[2], **Jimmy Lin**[1], and **Xu Sun**[3,4]

[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] StatNLP Research Group, Singapore University of Technology and Design
[3] MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[4] Center for Data Science, Beijing Institute of Big Data Research, Peking University

{h324yang,p8shi,jimmylin}@uwaterloo.ca
yanyan_zou@mymail.sutd.edu.sg
luwei@sutd.edu.sg, xusun@pku.edu.cn

Abstract

Multilingual knowledge graphs (KGs), such as YAGO and DBpedia, represent entities in different languages. The task of cross-lingual entity alignment is to match entities in a source language with their counterparts in target languages. In this work, we investigate embedding-based approaches to encode entities from multilingual KGs into the same vector space, where equivalent entities are close to each other. Specifically, we apply graph convolutional networks (GCNs) to combine multi-aspect information of entities, including topological connections, relations, and attributes of entities, to learn entity embeddings. To exploit the literal descriptions of entities expressed in different languages, we propose two uses of a pretrained multilingual BERT model to bridge cross-lingual gaps. We further propose two strategies to integrate GCN-based and BERT-based modules to boost performance. Extensive experiments on two benchmark datasets demonstrate that our method significantly outperforms existing systems.

1 Introduction

A growing number of multilingual knowledge graphs (KGs) have been built, such as DBpedia (Bizer et al., 2009), YAGO (Suchanek et al., 2008; Rebele et al., 2016), and BabelNet (Navigli and Ponzetto, 2012), which typically represent real-world knowledge as separately-structured monolingual KGs. Such KGs are connected via inter-lingual links (ILLs) that align entities with their counterparts in different languages, exemplified by Figure 1 (top). Highly-integrated multilingual KGs contain useful knowledge that can benefit many knowledge-driven cross-lingual NLP tasks, such as machine translation (Moussallem et al., 2018) and cross-lingual named entity recognition
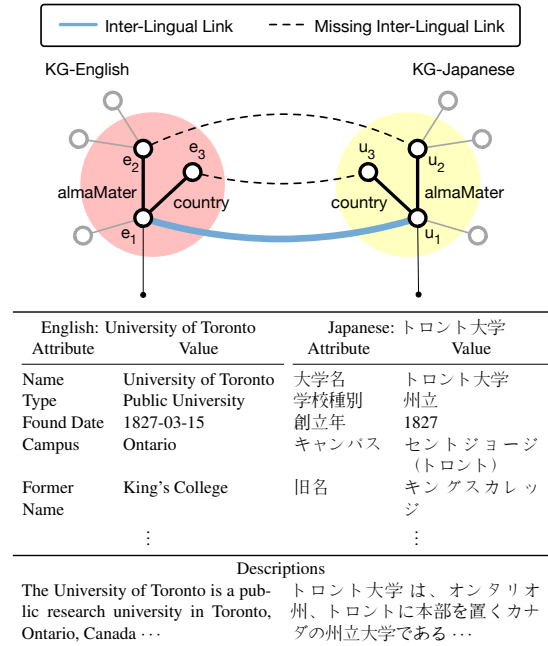

Figure 1: An example fragment of two KGs (in English and Japanese) connected by an inter-lingual link (ILL). In addition to the graph structures (top) consisting of entity nodes and typed relation edges, KGs also provide attributes and literal descriptions of entities (bottom).

(Darwish, 2013). However, the coverage of ILLs among existing KGs is quite low (Chen et al., 2018): for example, less than 20% of the entities in DBpedia are covered by ILLs. The goal of cross-lingual entity alignment is to discover entities from different monolingual KGs that actually refer to the same real-world entities, i.e., discovering the missing ILLs.

Traditional methods for this task apply machine translation techniques to translate entity labels (Spohr et al., 2011). The quality of alignments in the cross-lingual scenario heavily depends on the quality of the adopted translation systems. In addition to entity labels, existing KGs also provide multi-aspect information of entities,

---

including topological connections, relation types, attributes, and literal descriptions expressed in different languages (Bizer et al., 2009; Xie et al., 2016), as shown in Figure 1 (bottom). The key challenge of addressing such a task thus is how to better model and use provided multi-aspect information of entities to bridge cross-lingual gaps and find more equivalent entities (i.e., ILLs).

Recently, embedding-based solutions (Chen et al., 2017b; Sun et al., 2017; Zhu et al., 2017; Wang et al., 2018; Chen et al., 2018) have been proposed to unify multilingual KGs into the same low-dimensional vector space where equivalent entities are close to each other. Such methods only make use of one or two aspects of the aforementioned information. For example, Zhu et al. (2017) relied only on topological features while Sun et al. (2017) and Wang et al. (2018) exploited both topological and attribute features. Chen et al. (2018) proposed a co-training algorithm to combine topological features and literal descriptions of entities. However, combining these multi-aspect information of entities (i.e., topological connections, relations and attributes, as well as literal descriptions) remains under-explored.

In this work, we propose a novel approach to learn cross-lingual entity embeddings by using all aforementioned aspects of information in KGs. To be specific, we propose two variants of GCN-based models, namely MAN and HMAN, that incorporate multi-aspect features, including topological features, relation types, and attributes into cross-lingual entity embeddings. To capture semantic relatedness of literal descriptions, we fine-tune the pretrained multilingual BERT model (Devlin et al., 2019) to bridge cross-lingual gaps. We design two strategies to combine GCN-based and BERT-based modules to make alignment decisions. Experiments show that our method achieves new state-of-the-art results on two benchmark datasets. Source code for our models is publicly available at https://github.com/h324yang/HMAN.

2 Problem Definition

In a multilingual knowledge graph $\mathcal{G}$, we use $\mathcal{L}$ to denote the set of languages that $\mathcal{G}$ contains and $\mathcal{G}_i = \{E_i, R_i, A_i, V_i, D_i\}$ to represent the language-specific knowledge graph in language $L_i \in \mathcal{L}$. E_i, R_i, A_i, V_i and D_i are sets of entities, relations, attributes, values of attributes, and literal

descriptions, each of which portrays one aspect of an entity. The graph $\mathcal{G}_i$ consists of relation triples $\langle h_i, r_i, t_i \rangle$ and attribute triples $\langle h_i, a_i, v_i \rangle$ such that $h_i, t_i \in E_i$, $r_i \in R_i$, $a_i \in A_i$ and $v_i \in V_i$. Each entity is accompanied by a literal description consisting of a sequence of words in language L_i, e.g., $\langle h_i, d_{h,i} \rangle$ and $\langle t_i, d_{t,i} \rangle$, $d_{h,i}, d_{t,i} \in D_i$.

Given two knowledge graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ expressed in source language L_1 and target language L_2, respectively, there exists a set of pre-aligned ILLs $I(\mathcal{G}_1, \mathcal{G}_2) = \{(e, u) | e \in E_1, u \in E_2\}$ which can be considered training data. The task of cross-lingual entity alignment is to align entities in $\mathcal{G}_1$ with their cross-lingual counterparts in $\mathcal{G}_2$, i.e., discover missing ILLs.

3 Proposed Approach

In this section, we first introduce two GCN-based models, namely MAN and HMAN, that learn entity embeddings from the graph structures. Second, we discuss two uses of a multilingual pretrained BERT model to learn cross-lingual embeddings of entity descriptions: POINTWISEBERT and PAIRWISEBERT. Finally, we investigate two strategies to integrate the GCN-based and the BERT-based modules.

3.1 Cross-Lingual Graph Embeddings

Graph convolutional networks (GCNs) (Kipf and Welling, 2017) are variants of convolutional networks that have proven effective in capturing information from graph structures, such as dependency graphs (Guo et al., 2019b), abstract meaning representation graphs (Guo et al., 2019a), and knowledge graphs (Wang et al., 2018). In practice, multi-layer GCNs are stacked to collect evidence from multi-hop neighbors. Formally, the l-th GCN layer takes as input feature representations $H^{(l-1)}$ and outputs $H^{(l)}$:

$$H^{(l)} = \phi\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l-1)} W^{(l)}\right) \quad (1)$$

where $\tilde{A} = A + I$ is the adjacency matrix, I is the identity matrix, $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$, $\phi(\cdot)$ is ReLU function, and $W^{(l)}$ represents learnable parameters in the l-th layer. $H^{(0)}$ is the initial input.

GCNs can iteratively update the representation of each entity node via a propagation mechanism through the graph. Inspired by previous studies (Zhang et al., 2018; Wang et al., 2018), we also

adopt GCNs in this work to collect evidence from multilingual KG structures and to learn cross-lingual embeddings of entities. The primary assumptions are: (1) equivalent entities tend to be neighbored by equivalent entities via the same types of relations; (2) equivalent entities tend to share similar or even the same attributes.

Multi-Aspect Entity Features. Existing KGs (Bizer et al., 2009; Suchanek et al., 2008; Rebele et al., 2016) provide multi-aspect information of entities. In this section, we mainly focus on the following three aspects: topological connections, relations, and attributes. The key challenge is how to utilize the provided features to learn better embeddings of entities. We discuss how we construct raw features for the three aspects, which are then fed as inputs to our model. We use X_t, X_r and X_a to denote the topological connection, relation, and attribute features, individually.

The topological features contain rich neighborhood proximity information of entities, which can be captured by multi-layer GCNs. As in Wang et al. (2018), we set the initial topological features to $X_t = I$, i.e., an identity matrix serving as index vectors (also known as the featureless setting), so that the GCN can learn the representations of corresponding entities.

In addition, we also consider the relation and attribute features. As shown in Figure 1, the connected relations and attributes of two equivalent entities, e.g., "*University of Toronto*" (English) and "トロント大学" (Japanese), have a lot of overlap, which can benefit cross-lingual entity alignment. Specifically, they share the same relation types, e.g., "country" and "almaMater", and some attributes, e.g., "foundDate" and "創立年". To capture relation information, Schlichtkrull et al. (2018) proposed RGCN with relation-wise parameters. However, with respect to this task, existing KGs typically contain thousands of relation types but few pre-aligned ILLs. Directly adopting RGCN may introduce too many parameters for the limited training data and thus cause overfitting. Wang et al. (2018) instead simply used the unlabeled GCNs (Kipf and Welling, 2017) with two proposed measures (i.e., functionality and inverse functionality) to encode the information of relations into the adjacency matrix. They also considered attributes as input features in their architecture. However, this approach may lose information about relation types. Therefore, we re-

gard relations and attributes of entities as bag-of-words features to explicitly model these two aspects. Specifically, we construct count-based *N-hot* vectors X_r and X_a for these two aspects of features, respectively, where the (i, j) entry is the count of the j-th relation (attribute) for the corresponding entity e_i. Note that we only consider the top-F most frequent relations and attributes to avoid data sparsity issues. Thus, for each entity, both of its relation and attribute features are F-dimensional vectors.

MAN. Inspired by Wang et al. (2018), we propose the *Multi-Aspect Alignment Network* (MAN) to capture the three aspects of entity features. Specifically, three l-layer GCNs take as inputs the triple-aspect features (i.e., X_t, X_r, and X_a) and produce the representations $H_t^{(l)}$, $H_r^{(l)}$, and $H_a^{(l)}$ according to Equation 1, respectively. Finally, the multi-aspect entity embedding is:

$$H_m = [H_t^{(l)} \oplus H_a^{(l)} \oplus H_r^{(l)}] \qquad (2)$$

where $\oplus$ denotes vector concatenation. H_m can then feed into alignment decisions.

Such fusion through concatenation is also known as *Scoring Level Fusion*, which has been proven simple but effective for capturing multi-modal semantics (Bruni et al., 2014; Kiela and Bottou, 2014; Collell et al., 2017). It is worth noting that the main differences between MAN and the work of Wang et al. (2018) are two fold: First, we use the same approach as in Kipf and Welling (2017) to construct the adjacency matrix, while Wang et al. (2018) designed a new connectivity matrix as the adjacency matrix for the GCNs. Second, MAN explicitly regards the relation type features as model input, while Wang et al. (2018) incorporated such relation information into the connectivity matrix.

HMAN. Note that MAN propagates relation and attribute information through the graph structure. However, for aligning a pair of entities, we observe that considering the relations and attributes of neighboring entities, besides their own ones, may introduce noise. Merely focusing on relation and attribute features of the current entity could be a better choice. Thus, we propose the *Hybrid Multi-Aspect Alignment Network* (HMAN) to better model such diverse features, shown in Figure 2. Similar to MAN, we still leverage the l-th layer of a GCN to obtain topological embeddings $H_t^{(l)}$, but exploit feedforward neural networks to obtain
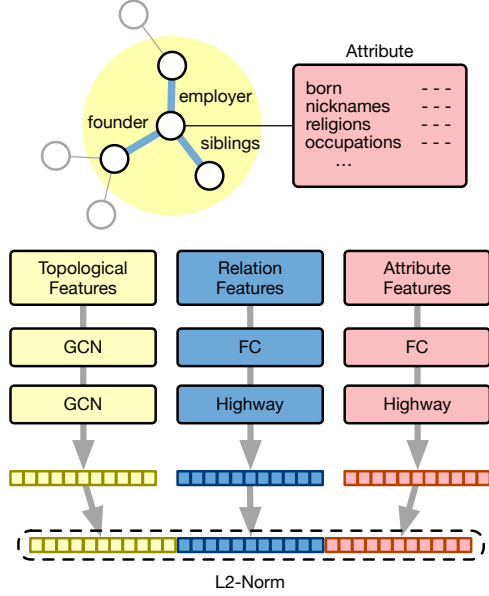
Figure 2: Architecture of HMAN.

the embeddings with respect to relations and attributes. The feedforward neural networks consist of one fully-connected (FC) layer and a highway network layer (Srivastava et al., 2015). The reason we use highway networks is consistent with the conclusions of Mudgal et al. (2018), who conducted a design space exploration of neural models for entity matching and found that highway networks are generally better than FC layers in convergence speed and effectiveness.

Formally, these feedforward neural networks are defined as:

$$S_f = \phi(W_f^{(1)} X_f + b_f^{(1)})$$
$$T_f = \sigma(W_f^t S_f + b_f^t) \qquad (3)$$
$$G_f = \phi(W_f^{(2)} S_f + b_f^{(2)}) \cdot T_f + S_f \cdot (1 - T_f)$$

where $f \in \{r, a\}$ and X_f refer to one specific aspect (i.e., relation or attribute) and the corresponding raw features, respectively, $W_f^{(1,2,t)}$ and $b_f^{(1,2,t)}$ are model parameters, $\phi(\cdot)$ is ReLU function, and $\sigma(\cdot)$ is sigmoid function. Accordingly, we obtain the hybrid multi-aspect entity embedding $H_y = [H_t^{(l)} \oplus G_r \oplus G_a]$, to which ℓ_2 normalization is further applied.

Model Objective. Given two knowledge graphs, $\mathcal{G}_1$ and $\mathcal{G}_2$, and a set of pre-aligned entity pairs $I(\mathcal{G}_1, \mathcal{G}_2)$ as training data, our model is trained in a supervised fashion. During the training phase, the goal is to embed cross-lingual entities into the same low-dimensional vector space where equiv-

alent entities are close to each other. Following Wang et al. (2018), our margin-based ranking loss function is defined as:

$$J = \sum_{(e_1, e_2) \in I} \sum_{(e_1', e_2') \in I'} [\rho(h_{e_1}, h_{e_2}) + \beta$$
$$- \rho(h_{e_1'}, h_{e_2'})]_+ \qquad (4)$$

where $[x]_+ = \max\{0, x\}$, I' denotes the set of negative entity alignment pairs constructed by corrupting the gold pair $(e_1, e_2) \in I$. Specifically, we replace e_1 or e_2 with a randomly-chosen entity in E_1 or E_2. $\rho(x, y)$ is the ℓ_1 distance function, and $\beta > 0$ is the margin hyperparameter separating positive and negative pairs.

3.2 Cross-Lingual Textual Embeddings

Existing multilingual KGs (Bizer et al., 2009; Navigli and Ponzetto, 2012; Rebele et al., 2016) also provide literal descriptions of entities expressed in different languages and contain detailed semantic information about the entities. The key observation is that literal descriptions of equivalent entities are semantically close to each other. However, it is non-trivial to directly measure the semantic relatedness of two entities' descriptions, since they are expressed in different languages.

Recently, Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al., 2019) has advanced the state-of-the-art in various NLP tasks by heavily exploiting pretraining based on language modeling. Of special interest is the multilingual variant, which was trained with Wikipedia dumps of 104 languages. The spirit of BERT in the multilingual scenario is to project words or sentences from different languages into the same semantic space. This aligns well with our objective—bridging gaps between descriptions written in different languages. Therefore, we propose two methods for applying multilingual BERT, POINTWISEBERT and PAIRWISE-BERT, to help make alignment decisions.

POINTWISEBERT. A simple choice is to follow the basic design of BERT and formulate the entity alignment task as a text matching task. For two entities e_1 and e_2 from two KGs in L_1 and L_2, denoting source language and target language, respectively, their textual descriptions are d_1 and d_2, consisting of word sequences in two languages. The model takes as inputs [CLS] d_1 [SEP] d_2 [SEP], where [CLS] is the special classification token,
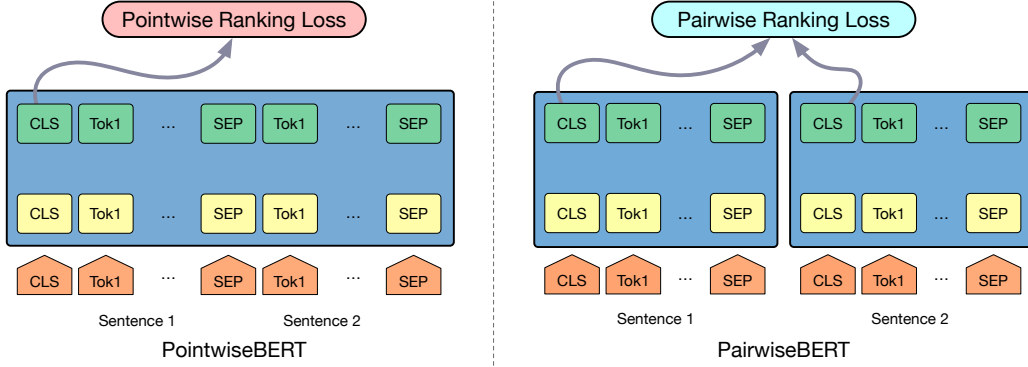
Figure 3: Architecture overview of PᴏɪɴᴛᴡɪsᴇBᴇʀᴛ (left) and PᴀɪʀᴡɪsᴇBᴇʀᴛ (right).

from which the final hidden state is used as the sequence representation, and [SEP] is the special token for separating token sequences, and produces the probability of classifying the pair as equivalent entities. The probability is then used to rank all candidate entity pairs, i.e., ranking score. We denote this model as PᴏɪɴᴛᴡɪsᴇBᴇʀᴛ, shown in Figure 3 (left).

This approach is computationally expensive, since for each entity we need to consider all candidate entities in the target language. One solution, inspired by the work of Shi et al. (2019), is to reduce the search space for each entity with a *reranking strategy* (see Section 3.3).

PᴀɪʀᴡɪsᴇBᴇʀᴛ. Due to the heavy computational cost of PᴏɪɴᴛᴡɪsᴇBᴇʀᴛ, semantic matching between all entity pairs is very expensive. Instead of producing ranking scores for description pairs, we propose PᴀɪʀᴡɪsᴇBᴇʀᴛ to encode the entity literal descriptions as cross-lingual textual embeddings, where distances between entity pairs can be directly measured using these embeddings.

The PᴀɪʀᴡɪsᴇBᴇʀᴛ model consists of two components, each of which takes as input the description of one entity (from the source or target language), as depicted in Figure 3 (right). Specifically, the input is designed as [CLS] $d_1(d_2)$ [SEP], which is then fed into PᴀɪʀᴡɪsᴇBᴇʀᴛ for contextual encoding. We select the hidden state of [CLS] as the textual embedding of the entity description for training and inference. To bring the textual embeddings of cross-lingual entity descriptions into the same vector space, a similar ranking loss function as in Equation 4 is used.

3.3 Integration Strategy

Sections 3.1 and 3.2 introduce two modules that separately collect evidence from knowledge graph structures and the literal descriptions of entities, namely graph and textual embeddings. In this section, we investigate two strategies to integrate these two modules to further boost performance.

Reranking. As mentioned in Section 3.2, the PᴏɪɴᴛᴡɪsᴇBᴇʀᴛ model takes as input the concatenation of two descriptions for each candidate–entity pair, where conceptually we must process every possible pair in the training set. Such a setting would be cost prohibitive computationally.

One way to reduce the cost of PᴏɪɴᴛᴡɪsᴇBᴇʀᴛ would be to ignore candidate pairs that are unlikely to be aligned. Rao et al. (2016) showed that uncertainty-based sampling can provide extra improvements in ranking. Following this idea, the GCN-based models (i.e., Mᴀɴ and Hᴍᴀɴ) are used to generate a candidate pool whose size is much smaller than the entire universe of entities. Specifically, GCN-based models provide top-q candidates of target entities for each source entity (where q is a hyperparameter). Then, the PᴏɪɴᴛᴡɪsᴇBᴇʀᴛ model produces a ranking score for each candidate–entity pair in the pool to further rerank the candidates. However, the weakness of such a reranking strategy is that performance is bounded by the quality of (potentially limited) candidates produced by Mᴀɴ or Hᴍᴀɴ.

Weighted Concatenation. With the textual embeddings learned by PᴀɪʀᴡɪsᴇBᴇʀᴛ denoted as H^B and graph embeddings denoted as H^G, a simple way to combine the two modules is by weighted concatenation:

$$H^C = \tau \cdot H^G \oplus (1 - \tau) \cdot H^B \qquad (5)$$

where H^G is the graph embeddings learned by either Mᴀɴ or Hᴍᴀɴ, and τ is a factor to balance the contribution of each source (where τ is a hyperparameter).

3.4 Entity Alignment

After we obtain the embeddings of entities, we leverage ℓ_1 distance to measure the distance between candidate–entity pairs. A small distance reflects a high probability for an entity pair to be aligned as equivalent entities. To be specific, with respect to the reranking strategy, we select the target entities that have the smallest distances to a source entity in the vector space learned by MAN or HMAN as its candidates. For weighted concatenation, we employ the ℓ_1 distance of the representations of a pair derived from the concatenated embedding, i.e., H^C, as the ranking score.

4 Experiments

4.1 Datasets and Settings

We evaluate our methods over two benchmark datasets: DBP15K and DBP100K (Sun et al., 2017). Table 1 outlines the statistics of both datasets, which contain 15,000 and 100,000 ILLs, respectively. Both are divided into three subsets: Chinese-English (ZH-EN), Japanese-English (JA-EN), and French-English (FR-EN).

Following previous work (Sun et al., 2017; Wang et al., 2018), we adopt the same split settings in our experiments, where 30% of the ILLs are used as training and the remaining 70% for evaluation. *Hits@k* is used as the evaluation metric (Bordes et al., 2013; Sun et al., 2017; Wang et al., 2018), which measures the proportion of correctly aligned entities ranked in the top-k candidates, and results in both directions, e.g., ZH-EN and EN-ZH, are reported.

In all our experiments, we employ two-layer GCNs and the top 1000 (i.e., F=1000) most frequent relation types and attributes are included to build the N-hot feature vectors. For the MAN model, we set the dimensionality of topological, relation, and attribute embeddings to 200, 100, and 100, respectively. When training HMAN, the hyperparameters are dependent on the dataset sizes due to GPU memory limitations. For DBP15K, we set the dimensionality of topological embeddings, relation embeddings, and attribute embeddings to 200, 100, and 100, respectively. For DBP100K, the dimensionalities are set to 100, 50, and 50, respectively. We adopt SGD to update parameters and the numbers of epochs are set to 2,000 and 50,000 for MAN and HMAN, respectively. The margin β in the loss function is set to 3. The balance factor τ is determined by grid search, which

Datasets		DBP15K				
		Entities	Rel.	Attr.	Rel.triples	Attr.triples
ZH-EN	Chinese	66,469	2,830	8,113	153,929	379,684
	English	98,125	2,317	7,173	237,674	567,755
JA-EN	Japanese	65,744	2,043	5,882	164,373	354,619
	English	95,680	2,096	6,066	233,319	497,230
FR-EN	French	66,858	1,379	4,547	192,191	528,665
	English	105,889	2,209	6,422	278,590	576,543
Datasets		DBP100K				
		Entities	Rel.	Attr.	Rel.triples	Attr.triples
ZH-EN	Chinese	106,517	4,431	16,152	329,890	1,404,615
	English	185,022	3,519	14,459	453,248	1,902,725
JA-EN	Japanese	117,836	2,888	12,305	413,558	1,474,721
	English	118,570	2,631	13,238	494,087	1,738,803
FR-EN	French	105,724	1,775	8,029	409,399	1,361,509
	English	107,231	2,504	13,170	513,382	1,957,813

Table 1: Statistics of DBP15K and DBP100K. Rel. and Attr. stand for relations and attributes, respectively.

shows that the best performance lies in the range from 0.8 to 0.7. For simplicity, τ is set to 0.8 in all associated experiments. Multilingual BERT-base models with 768 hidden units are used in POINTWISEBERT and PAIRWISEBERT. We additionally append one more FC layer to the representation of [CLS] and reduce the dimensionality to 300. Both BERT models are fine-tuned using the Adam optimizer.

4.2 Results on Graph Embeddings

We first compare MAN and HMAN against previous systems (Hao et al., 2016; Chen et al., 2017a; Sun et al., 2017; Wang et al., 2018). As shown in Table 2, MAN and HMAN consistently outperform all baselines in all scenarios, especially HMAN. It is worth noting that, in this case, MAN and HMAN use as much information as Wang et al. (2018), while Sun et al. (2017) require extra supervised information (relations and attributes of two KGs need to be aligned in advance). The performance improvements confirm that our model can better utilize topological, relational, and attribute information of entities provided by KGs.

To explain why HMAN achieves better results than MAN, recall that MAN collects relation and attribute information by the propagation mechanism in GCNs where such knowledge is exchanged through neighbors, while HMAN uses feedforward networks to capture expressive features directly from the input feature vectors without propagation. As we discussed before, it is not always the case that neighbors of equivalent entities share similar relations or attributes. Propagating such features through linked entities in GCNs may introduce noise and thus harm performance.

Model	ZH → EN			EN→ ZH			JA → EN			EN→ JA			FR → EN			EN→ FR		
	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50
DBP15K																		
Hao et al. (2016)	21.2	42.7	56.7	19.5	39.3	53.2	18.9	39.9	54.2	17.8	38.4	52.4	15.3	38.8	56.5	14.6	37.2	54.0
Chen et al. (2017a)	30.8	61.4	79.1	24.7	52.4	70.4	27.8	57.4	75.9	23.7	49.9	67.9	24.4	55.5	74.4	21.2	50.6	69.9
Sun et al. (2017)	41.1	74.4	88.9	40.1	71.0	86.1	36.2	68.5	85.3	38.3	67.2	82.6	32.3	66.6	83.1	32.9	65.9	82.3
Wang et al. (2018)	41.2	74.3	86.2	36.4	69.9	82.4	39.9	74.4	86.1	38.4	71.8	83.7	37.2	74.4	86.7	36.7	73.0	86.3
MAN	46.0	79.4	90.0	41.5	75.6	88.3	44.6	78.8	90.0	43.0	77.1	88.7	43.1	79.7	91.7	42.1	79.1	90.9
MAN w/o TE	21.5	55.0	79.4	20.2	53.6	78.8	15.0	44.0	69.9	14.3	44.0	70.6	10.2	34.5	59.5	10.8	35.2	60.3
MAN w/o RE	45.6	79.1	89.5	41.1	75.0	87.3	44.2	78.7	89.8	43.0	76.9	88.1	42.8	79.7	91.4	42.1	78.9	90.6
MAN w/o AE	43.7	77.1	87.8	39.2	72.9	85.5	43.2	77.6	88.4	41.2	74.9	86.6	42.9	79.6	91.0	41.5	78.9	90.5
HMAN	**56.2**	**85.1**	**93.4**	**53.7**	**83.4**	**92.5**	**56.7**	**86.9**	**94.5**	**56.5**	**86.6**	**94.6**	**54.0**	**87.1**	**95.0**	**54.3**	**86.7**	**95.1**
HMAN w/o TE	3.2	16.7	38.3	3.5	17.2	38.5	5.4	22.3	45.5	5.2	22.0	45.5	2.4	13.9	35.3	2.2	13.7	35.5
HMAN w/o RE	50.2	78.4	86.5	49.3	78.6	87.0	52.6	81.6	89.1	52.4	81.1	89.8	52.7	84.2	91.4	52.0	83.9	91.1
HMAN w/o AE	49.2	81.0	89.8	48.8	80.9	90.0	52.2	83.3	91.6	51.5	83.1	91.6	52.3	85.6	93.7	52.3	85.1	93.2
HMAN w/o HW	46.8	76.1	84.1	46.0	76.2	84.6	50.5	79.5	87.5	49.9	79.1	87.5	51.9	82.7	90.9	51.6	82.5	90.6
DBP100K																		
Hao et al. (2016)	-	16.9	-	-	16.6	-	-	21.1	-	-	20.9	-	-	22.9	-	-	22.6	-
Chen et al. (2017a)	-	34.3	-	-	29.1	-	-	33.9	-	-	27.2	-	-	44.8	-	-	39.1	-
Sun et al. (2017)	20.2	41.2	58.3	19.6	39.4	56.0	19.4	42.1	60.5	19.1	39.4	55.9	26.2	54.6	70.5	25.9	51.3	66.9
Wang et al. (2018)	23.1	47.5	63.8	19.2	40.3	55.4	26.4	55.1	70.0	21.9	44.4	56.6	29.2	58.4	68.7	25.7	50.5	59.8
MAN	27.2	54.2	**72.8**	24.7	50.2	**69.0**	30.0	60.4	**77.3**	26.6	54.4	71.2	31.6	64.0	77.3	28.8	59.3	73.4
MAN w/o TE	11.8	28.6	47.7	11.2	28.3	47.9	7.4	21.7	39.4	7.2	21.6	39.8	5.4	19.4	38.2	5.1	18.8	37.1
MAN w/o RE	26.5	53.4	72.1	23.9	49.2	67.9	29.8	60.3	77.1	26.3	53.9	70.6	31.0	63.2	76.4	28.4	58.4	72.2
MAN w/o AE	25.5	51.7	70.4	22.8	47.6	66.3	29.4	59.4	76.1	25.9	52.9	69.7	30.8	62.7	75.8	28.1	57.8	71.5
HMAN	**29.8**	**54.6**	69.5	**28.7**	**53.3**	69.0	**34.3**	**63.3**	76.1	**33.8**	**63.0**	**76.7**	**37.5**	**67.7**	**77.7**	**37.6**	**68.1**	**78.5**
HMAN w/o TE	6.8	20.3	39.2	7.2	21.0	39.4	3.0	11.5	27.3	3.3	11.8	28.0	0.5	3.5	11.1	0.5	3.4	11.4
HMAN w/o RE	28.0	50.3	62.3	28.2	50.6	62.9	30.3	54.9	64.8	30.2	55.9	66.9	32.8	60.3	69.1	33.3	60.9	69.8
HMAN w/o AE	25.7	46.4	57.3	25.5	64.7	57.9	29.6	55.1	66.1	29.9	56.1	67.4	32.5	59.2	67.8	32.9	59.4	68.4
HMAN w/o HW	25.2	46.0	57.9	25.2	45.9	57.9	28.6	52.6	62.2	28.5	53.0	63.0	32.8	60.9	70.0	32.9	60.2	70.3

Table 2: Results of using graph information on DBP15K and DBP100K. @1, @10 and @50 refer to Hits@1, Hits@10 and Hits@50, respectively.

Moreover, we perform ablation studies on the two proposed models to investigate the effectiveness of each component. We alternatively remove each aspect of features (i.e., topological, relation, and attribute features) and the highway layer in HMAN, denoted as w/o TE (RE, AE, and HW). As reported in Table 2, we observe that after removing relation or attribute features, the performance of HMAN and MAN drops across all datasets. These figures prove that these two aspects of features are useful in making alignment decisions. On the other hand, compared to MAN, HMAN shows more significant performance drops, which also demonstrates that employing the feedforward networks can better categorize relation and attribute features than GCNs in this scenario. Interestingly, looking at the two variants MAN w/o TE and HMAN w/o TE, we can see the former achieves better results. Since MAN propagates relation and attribute features via graph structures, it can still implicitly capture topological knowledge of entities even after we remove the topological features.

However, HMAN loses such structure knowledge when topological features are excluded, and thus its results are worse. From these experiments, we can conclude that the topological information is playing an indispensable role in making alignment decisions.

4.3 Results with Textual Embeddings

In this section, we discuss empirical results involving the addition of entity descriptions, shown in Table 3. Applying literal descriptions of entities to conduct cross-lingual entity alignment is relatively under-explored. The recent work of Chen et al. (2018) used entity descriptions in their model; however, we are unable to make comparisons with their work, as we do not have access to their code and data. Since we employ BERT to learn textual embeddings of descriptions, we consider systems that also use external resources, like Google Translate,[1] as our baselines. We directly

---

[1] https://cloud.google.com/translate/

Model	ZH → EN			EN→ ZH			JA → EN			EN→ JA			FR → EN			EN→ FR		
	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50
DBP15K																		
Translation∗	55.7	67.6	74.3	40.3	54.2	62.2	74.6	84.5	89.1	61.9	72.0	77.2	-	-	-	-	-	-
JAPE + Translation∗	73.0	90.4	96.6	62.7	85.2	94.2	82.8	94.6	98.3	75.9	90.7	96.0	-	-	-	-	-	-
PairwiseBert	74.3	94.6	98.8	74.8	94.7	99.0	78.6	95.8	98.5	78.3	95.4	98.4	95.2	99.2	99.6	94.9	99.2	99.7
Man (Rerank)	84.2	93.6	94.8	82.1	91.8	93.1	89.4	94.0	94.8	88.2	93.3	94.0	93.1	95.2	95.4	93.1	95.3	95.4
Hman (Rerank)	86.5	95.9	96.9	85.8	94.1	95.3	89.0	96.0	97.3	89.0	96.0	97.5	95.3	97.7	97.8	95.2	97.9	98.1
Man (Weighted)	85.4	98.2	99.7	83.8	97.7	99.5	90.8	98.8	99.7	89.9	98.5	99.5	96.8	99.6	99.8	96.7	99.7	**99.9**
Hman (Weighted)	**87.1**	**98.7**	**99.8**	**86.4**	**98.5**	**99.8**	**93.5**	**99.4**	**99.9**	**93.3**	**99.3**	**99.9**	**97.3**	**99.8**	**99.9**	**97.3**	**99.8**	99.9
DBP100K																		
PairwiseBert	65.1	85.1	92.6	66.2	85.8	92.9	67.7	86.5	93.1	67.9	86.4	93.2	93.2	97.9	98.9	93.4	98.0	98.9
Man (Rerank)	59.5	62.1	62.2	55.9	58.2	58.2	65.5	68.2	68.4	59.9	62.1	62.3	69.7	70.4	70.5	65.5	66.2	66.2
Hman (Rerank)	58.9	61.2	61.3	57.9	60.2	60.3	66.9	69.4	69.6	67.0	69.6	69.8	72.1	72.9	73.0	72.7	73.5	73.5
Man (Weighted)	**81.4**	**94.9**	**98.2**	**80.5**	94.1	97.7	84.3	95.4	98.3	81.5	94.2	97.6	96.2	99.3	**99.7**	95.7	99.1	99.6
Hman (Weighted)	81.1	94.3	97.8	80.3	**94.5**	**97.9**	**85.2**	**96.1**	**98.4**	**84.6**	**96.1**	**98.5**	**96.5**	**99.4**	99.7	**96.5**	**99.5**	**99.8**

Table 3: Results of using both graph and textual information on DBP15K and DBP100K. @1, @10, and @50 refer to Hits@1, Hits@10, and Hits@50, respectively. ∗ indicates results are taken from (Sun et al., 2017).

take results reported by Sun et al. (2017), denoted as "Translation" and "JAPE+Translation".

The PointwiseBert model is used with GCN-based models, which largely reduces the search space, as indicated by Man (Rerank) and Hman (Rerank), where the difference is that the candidate pools are given by Man and Hman, respectively. For DBP15K, we select top-200 candidate target entities as the candidate pool while for DBP100K, top-20 candidates are selected due to its larger size. The reranking method does lead to performance gains across all datasets, where the improvements are dependent on the quality of the candidate pools. Hman (Rerank) generally performs better than Man (Rerank) since Hman recommends more promising candidate pools.

The PairwiseBert model learns the textual embeddings that map cross-lingual descriptions into the same space, which can be directly used to align entities. The results are listed under PairwiseBert in Table 3. We can see that it achieves good results on its own, which also shows the efficacy of using multilingual descriptions. Moreover, such textual embeddings can be combined with graph embeddings (learned by Man or Hman) by weighted concatenation, as discussed in Section 3.3. The results are reported as Man (Weighted) and Hman (Weighted), respectively. As we can see, this simple operation leads to significant improvements and gives excellent results across all datasets. However, it is not always the case that KGs provide descriptions for every entity. For those entities whose descriptions

	English	Chinese
ILL pair	Casino_Royale_(2006_film) (3)	007大戰皇家賭場 (3)
Features	starring, starring, distributor	starring, starring, language
Neighbors	Daniel_Craig (1), Eva_Green (4), Columbia_Pictures (9)	丹尼爾·克雷格 (1), 伊娃·格蓮 (4), 英語 (832)

Table 4: Case study of the noise introduced by the propagation mechanism.

are not available, the graph embeddings would be the only source for making alignment decisions.

4.4 Case Study

In this section, we describe a case study to understand the performance gap between Hman and Man. The example in Table 4 provides insights potentially explaining this performance gap. We argue that Man introduces unexpected noise from heterogeneous nodes during the GCN propagation process. We use the number in parentheses (*) after entity names to denote the number of relation features they have.

In this particular example, the two entities "Casino_Royale_(2006_film)" in the source language (English) and "007大戰皇家賭場" in the target language (Chinese) both have three relation features. We notice that the propagation mechanism introduces some neighbors which are unable to find cross-lingual counterparts from the other end, marked in red. Considering the entity "英語" (English), a neighbor of "007大戰皇家賭場", no counterparts can be found in the neighbors of "Casino_Royale_(2006_film)". We also observe that "英語" (English) is a pivot node in

the Chinese KG and has 832 relations, such as "語言" (Language), "官方語言" (Official Language), and "頻道語言" (Channel Language). In this case, propagating features from neighbors can harm performance. In fact, the feature sets of the ILL pair already convey information that captures their similarity (e.g., the "starring" marked in blue are shared twice). Therefore, by directly using feedforward networks, HMAN is able to effectively capture such knowledge.

5 Related Work

KG Alignment. Research on KG alignment can be categorized into two groups: monolingual and multilingual entity alignment. As for monolingual entity alignment, main approaches align two entities by computing string similarity of entity labels (Scharffe et al., 2009; Volz et al., 2009; Ngomo and Auer, 2011) or graph similarity (Raimond et al., 2008; Pershina et al., 2015; Azmy et al., 2019). Recently, Trsedya et al. (2019) proposed an embedding-based model that incorporates attribute values to learn the entity embeddings.

To match entities in different languages, Wang et al. (2012) leveraged only language-independent information to find possible links cross multilingual Wiki knowledge graphs. Recent studies learned cross-lingual embeddings of entities based on TransE (Bordes et al., 2013), which are then used to align entities across languages. Chen et al. (2018) designed a co-training algorithm to alternately learn multilingual entity and description embeddings. Wang et al. (2018) applied GCNs with the connectivity matrix defined on relations to embed entities from multilingual KGs into a unified low-dimensional space.

In this work, we also employ GCNs. However, in contrast to Wang et al. (2018), we regard relation features as input to our models. In addition, we investigate two different ways to capture relation and attribute features.

Multilingual Sentence Representations. Another line of research related to this work is aligning sentences in multiple languages. Recent works (Hermann and Blunsom, 2014; Conneau et al., 2018; Eriguchi et al., 2018) studied cross-lingual sentence classification via zero-shot learning. Johnson et al. (2017) proposed a sequence-to-sequence multilingual machine translation system where the encoder can be used to produce cross-lingual sentence embeddings (Artetxe and Schwenk, 2018). Recently, BERT (Devlin et al., 2019) has advanced the state-of-the-art on multiple natural language understanding tasks. Specifically, multilingual BERT enables learning representations of sentences under multilingual settings. We adopt BERT to produce cross-lingual representations of entity literal descriptions to capture their semantic relatedness, which benefits cross-lingual entity alignment.

6 Conclusion and Future Work

In this work, we focus on the task of cross-lingual entity alignment, which aims to discover mappings of equivalent entities in multilingual knowledge graphs. We proposed two GCN-based models and two uses of multilingual BERT to investigate how to better utilize multi-aspect information of entities provided by KGs, including topological connections, relations, attributes, and entity descriptions. Empirical results demonstrate that our best model consistently achieves state-of-the-art performance across all datasets. In the future, we would like to apply our methods to other multilingual datasets such as YAGO and BabelNet. Also, since literal descriptions of entities are not always available, we will investigate alternative ways to design graph-based models that can better capture structured knowledge for this task.

Acknowledgments

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Michael Azmy, Peng Shi, Jimmy Lin, and Ihab F. Ilyas. 2019. Matching entities across different knowledge graphs with graph embeddings. *arXiv preprint arXiv:1903.06607*.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia: A crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, 7(3):154–165.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of NIPS*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

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 IJCAI*.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017a. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of IJCAI*.

Muhao Chen, Tao Zhou, Pei Zhou, and Carlo Zaniolo. 2017b. Multi-graph affinity embeddings for multilingual knowledge graphs. In *Proceedings of NIPS Workshop on Automated Knowledge Base Construction*.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Proceedings of AAAI*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019a. Attention guided graph convolutional networks for relation extraction. In *Proceedings of ACL*.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019b. Densely connected graph convolutional networks for graph-to-sequence learning. *Transactions of the Association of Computational Linguistics*.

Yanchao Hao, Yuanzhe Zhang, Shizhu He, Kang Liu, and Jun Zhao. 2016. A joint embedding method for entity alignment of knowledge bases. In *Proceedings of China Conference on Knowledge Graph and Semantic Computing*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of ACL*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*.

Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1–19.

Sidharth Mudgal, Han Li, Theodoros Rekatsinas, An-Hai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep learning for entity matching: A design space exploration. In *Proceedings of International Conference on Management of Data*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Axel-Cyrille Ngonga Ngomo and Soren Auer. 2011. Limes: a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of IJCAI*.

Maria Pershina, Mohamed Yakout, and Kaushik Chakrabarti. 2015. Holistic entity matching across knowledge graphs. In *Proceedings of IEEE International Conference on Big Data*.

Yves Raimond, Christopher Sutton, and Mark B. Sandler. 2008. Automatic interlinking of music datasets on the semantic web. In *Proceedings of WWW workshop on Linked Data on the Web*.

Jinfeng Rao, Hua He, and Jimmy Lin. 2016. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of CIKM*.

Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. 2016. YAGO: A multilingual knowledge base from

Wikipedia, WordNet, and GeoNames. In *Proceedings of International Semantic Web Conference*.

François Scharffe, Yanbin Liu, and Chuguang Zhou. 2009. RDF-AI: an architecture for RDF datasets matching, fusion and interlink. In *Proceedings of IJCAI workshop on Identity and Reference in Web-based Knowledge Representation*.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *Proceedings of European Semantic Web Conference*.

Peng Shi, Jinfeng Rao, and Jimmy Lin. 2019. Simple attention-based representation learning for ranking short social media posts. In *Proceedings of NAACL*.

Dennis Spohr, Laura Hollink, and Philipp Cimiano. 2011. A machine learning approach to multilingual and cross-lingual ontology matching. In *Proceedings of International Semantic Web Conference*.

Rupesh K. Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Proceedings of NIPS*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A large ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *Proceedings of International Semantic Web Conference*.

Bayu Distiawan Trsedya, Jianzhong Qi, and Rui Zhang. 2019. Entity alignment between knowledge graphs using attribute embeddings. In *Proceedings of AAAI*.

Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. 2009. Discovering and maintaining links on the web of data. In *Proceedings of International Semantic Web Conference*.

Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of WWW*.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of EMNLP*.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of AAAI*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of EMNLP*.

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative entity alignment via joint knowledge embeddings. In *Proceedings of IJCAI*.