MMEA: Entity Alignment for Multi-modal Knowledge Graph

Liyi Chen[1], Zhi Li[1], Yijun Wang[2], Tong Xu[1(✉)], Zhefeng Wang[2], and Enhong Chen[1]

[1] School of Data Science, University of Science and Technology of China, Hefei, China
liyichencly@gmail.com, zhili03@mail.ustc.edu.cn,
{tongxu,cheneh}@ustc.edu.cn
[2] Huawei Technologies, Shenzhen, China
{wangyijun13,wangzhefeng}@huawei.com

Abstract. Entity alignment plays an essential role in the knowledge graph (KG) integration. Though large efforts have been made on exploring the association of relational embeddings between different knowledge graphs, they may fail to effectively describe and integrate the multi-modal knowledge in the real application scenario. To that end, in this paper, we propose a novel solution called Multi-Modal Entity Alignment (MMEA) to address the problem of entity alignment in a multi-modal view. Specifically, we first design a novel multi-modal knowledge embedding method to generate the entity representations of relational, visual and numerical knowledge, respectively. Along this line, multiple representations of different types of knowledge will be integrated via a multi-modal knowledge fusion module. Extensive experiments on two public datasets clearly demonstrate the effectiveness of the MMEA model with a significant margin compared with the state-of-the-art methods.

Keywords: Multi-modal knowledge · Entity alignment · Knowledge graph

1 Introduction

Knowledge graph (KG), which is composed of relational facts with entities connected by various relations, benefits lots of AI-related systems, such as recommender systems, question answering, and information retrieval. However, most KGs are constructed for specific purposes and monolingual settings, which results in the separate KGs with gaps of different descriptions for even the same concepts. Therefore, entity alignment techniques are urgently required to integrate the distinct KGs by linking entities referring to the same real-world identity.

Along this line, many efforts have been made in exploring the associations of distinct KGs and querying knowledge completely by entity alignment. In general, prior arts could be roughly grouped into two categories, i.e., similarity-based methods [9,12] and embedding-based methods [3,20]. Early studies mostly focus

G. Li et al. (Eds.): KSEM 2020, LNAI 12274, pp. 134–147, 2020.
https://doi.org/10.1007/978-3-030-55130-8_12

on the attribute similarity, such as string similarity [12] and numeric similarity [16]. However, these methods often suffer from the attribute heterogeneity, which makes the entity alignment error-prone [15]. Recently, in view of the rapid development of knowledge graph embedding, many researchers have attempted to utilize embedding-based models for the entity alignment problem [10,15]. In spite of the importance of prior arts, existing researches mainly focus on the semantics or concept knowledge graphs alignment but largely ignore the multi-modal knowledge from the real scenarios.

Indeed, in real-world application scenarios, knowledge is usually summarized in various forms, such as relational triples, numerical attributes and images. These distinct knowledge forms not only can play an important role as extra pieces of evidence for the KG completion, but also highly support the entity alignment task. For instance, Fig. 1 illustrates a toy example of entity alignment for multi-modal knowledge graphs, in which the image of "Fuji" can clearly demonstrate that the entity type is the mountain. Moreover, the similar images and numerical attributes (such as "Height" and "Latitude") can be helpful for aligning the same entity between two KGs. Unfortunately, it is not trivial to leverage multi-modal knowledge to the entity alignment problem. On the one hand, the alignment task is challenging in terms of computational complexity, data quality, and acquisition of prior alignment data in large-scale knowledge graphs. On the other hand, the inevitable heterogeneity among different modalities makes it difficult to learn and fuse the knowledge representations from distinct modalities. Therefore, traditional techniques may fail to deal with this task.

To conquer these challenges, in this paper, we propose a novel solution called Multi-Modal Entity Alignment (MMEA) for modeling the entity associations of multi-modal KGs and finding entities referring to the same real-world identity. To be specific, we first propose a multi-modal knowledge embedding method to discriminatively generate knowledge representations of three different types of knowledge, i.e., relational triples, visual contents (images) and numerical attributes. Then, to leverage multi-modal knowledge for the entity alignment task, a multi-modal fusion module is designed to integrate knowledge representations from multiple modalities. Extensive experiments on two large-scale real-world datasets demonstrate that MMEA not only provides insights to take advantages of multi-modal knowledge in the entity alignment task, but also outperforms the state-of-the-art baseline methods.

2 Related Work

Generally, the related work can be classified into two perspectives, i.e., entity alignment and multi-modal knowledge graph.

2.1 Entity Alignment

Actually, the entity alignment problem has been one of the major studies in the knowledge graph area for a long time. Early researchers mainly focus on

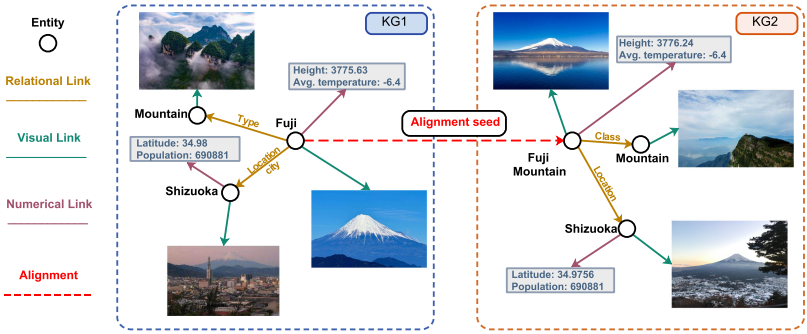

Fig. 1. A toy example of entity alignment between multi-modal knowledge graphs

exploring the content similarity to align the entities between different KGs. LD-Mapper [12] utilizes entity nearest neighbor similarity and string similarity. RuleMiner [9] refines a set of matching-rules with an Expectation-Maximization algorithm. SILK [16] measures entity similarity with string equality and similarity, numeric similarity and so on.

Recently, it is notable that entity alignment based on knowledge graph embedding representation becomes popular in the area. The current methods often embed entity to a low-dimensional space and measure the similarity between entity embeddings. Embedding-based methods concentrate on the semantics or concept so that they have a better analysis of knowledge. IPTransE [20] is an iterative method through joint knowledge embedding. BootEA [14] iteratively labels possible entity alignments as the training data, and employs an alignment editing method to reduce the error accumulation during the iterations. SEA [10] utilizes an awareness of the degree difference in adversarial training and incorporates the unaligned entities to enhance the performance. KDCoE [2] adds entity descriptions for entity alignment with a semi-supervised learning method for joint training. Furthermore, there are several methods utilizing attributes to strengthen the performance of entity alignment model. AttrE [15] uses a large number of attribute triples to generate character embeddings, and employs the relationship transitivity rule. IMUSE [6] achieves entity alignment and attribute alignment with an unsupervised method, and employs bivariate regression to merge alignment results. Additionally, GCN [17] uses relations to build the structures of graph convolutional networks and combines relations and attributes. However, these methods ignore the multi-modal knowledge from the real scenarios.

2.2 Multi-modal Knowledge Graph

In diverse domains, researchers study multi-modal learning in order to extract semantic information from various modalities. Multi-modal information such as structural and visual features is significant for entity alignment. PoE [7] is proposed to find entity alignment in multi-modal knowledge graphs through

extracting relational, latent, numerical and visual features. In addition, the most relevant task to our multi-modal entity alignment is multi-modal knowledge representation. Considering visual features from entity images for knowledge representation learning, IKRL [19] integrates image representations into an aggregated image-based representation via an attention-based method. MKBE [11] models knowledge bases that contain a variety of multi-modal features such as links, images, numerical and categorical values. It applies neural encoders and decoders which embed multi-modal evidence types and generate multi-modal attributes, respectively. [8] proposes a multi-modal translation-based method, which defines the energy of a knowledge graph triple as the sum of sub-energy functions that leverages structural, visual and linguistic knowledge representations. On the whole, multi-modal knowledge graph is still a novel problem, and the entity alignment has not been fully discussed.

3 Methodology

In this section, we formally introduce the entity alignment task for multi-modal knowledge graphs (KGs) and give an overview of our proposed model, i.e., Multi-Modal Entity Alignment (MMEA). Then, we describe the details of MMEA.

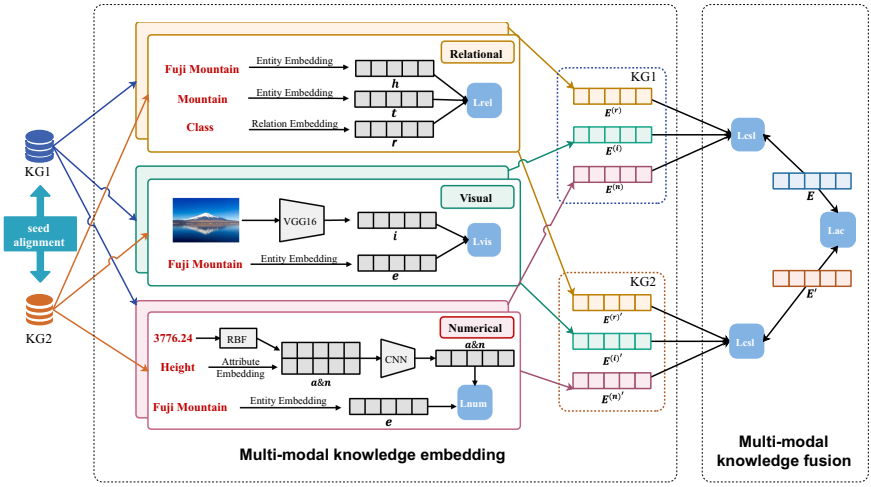

Fig. 2. The framework overview of MMEA.

3.1 Preliminaries and Technical Framework

Notation and Problem Definition. A multi-modal knowledge graph can be noted as $G = (\widehat{E}, R, I, N, X, Y, Z)$, where $\widehat{E}, R, I, N$ denote the sets of entities,

relations, images and numerics, and X, Y, Z denote the sets of relational triples, entity-image pairs and numerical triples, respectively. With multi-modal knowledge embedding, we denote $\mathbf{E}^{(r)}$, $\mathbf{E}^{(i)}$, $\mathbf{E}^{(n)}$ as entity embeddings for relational, visual and numerical information, respectively.

The task of entity alignment refers to matching entities describing the same thing in the real world from different knowledge graphs, which is beneficial for people to acquire knowledge completely, and it is not necessary to find related information of the same entity from multiple knowledge graphs anymore. Let $G_1 = (\widehat{E}_1, R_1, I_1, N_1, X_1, Y_1, Z_1)$ and $G_2 = (\widehat{E}_2, R_2, I_2, N_2, X_2, Y_2, Z_2)$ be two different KGs. $H = \left\{(e_1, e_2)|e_1 \in \widehat{E}_1, e_2 \in \widehat{E}_2\right\}$ denotes the set of aligned entities across knowledge graphs.

Framework Overview. In this paper, we propose a multi-modal model for entity alignment, namely Multi-Modal Entity Alignment (MMEA) model, which can automatically and accurately align the entities in two distinct multi-modal knowledge graphs. As illustrated in Fig. 2, our proposed MMEA consists of two major components, i.e., *Multi-Modal Knowledge Embedding* (MMKE) and *Multi-Modal Knowledge Fusion* (MMKF). In the MMKE module, we extract the relational, visual and numerical information to complement the absence of useful entity features. Then, with the MMKF module, we propose a novel multi-modal knowledge fusion method to minimize the distance of aligned entities from two distinct KGs across the multi-modal knowledge in the common space and design an interactive training stage to optimize the MMEA end-to-end.

3.2 Multi-modal Knowledge Embedding

Multi-modal knowledge plays a significant part in knowledge representations. In our multi-modal knowledge graph, there are three types of data modality, i.e., relational, visual and numerical data. Relational data refer to relational triple with entity associations, visual data mean the image of entities, and numerical data represent the attribute value. We will detail three types of knowledge embedding in the following section.

Relational Knowledge Representations. Relational triples are the main part of KGs, which are essential to judge the association of entities from different KGs. Under the relational data, we adopt the most representative translational distance model: TransE [1]. Given a fact $(h, r, t) \in X$, h and t can be associated by r in a low-dimensional continuous vector space. The process named translation adjusts the distance between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$ in the space constantly, in order that $\mathbf{h}+\mathbf{r}$ is equal to $\mathbf{t}$ as much as possible when (h, r, t) holds. In multi-relational data, there are certain structural similarities. Such as ("`Fuji`", "`Location city`", "`Shizuoka`") and ("`Eiffel`", "`Location city`", "`Paris`") in the embedding space, we have "`Shizuoka`" – "`Fuji`" $\approx$ "`Paris`" – "`Eiffel`". Through the relationship "`Location city`", we can acquire "`Eiffel`" + "`Location city`"

≈ "Paris" from "Fuji" + "Location city" ≈ "Shizuoka" automatically. The scoring function which we take to be L_2-norm is defined as follows:

$$f_{rel}(h, r, t) = -||\mathbf{h} + \mathbf{r} - \mathbf{t}||_2^2. \tag{1}$$

To learn the entity embeddings from relational data, we apply the margin-based [18] loss function with $\gamma > 0$ over the training set:

$$L_{rel} = \sum_{\tau^+ \in D^+} \sum_{\tau^- \in D^-} max(0, \gamma - f_{rel}(\tau^+) + f_{rel}(\tau^-)). \tag{2}$$

Here, D^+ and D^- are positive and negative examples sets, respectively. Given a positive example $\tau^+ = (h,\ r,\ t)$, we supplement the set of positive examples through the exchange strategy. The exchange strategy means that if h has been aligned by $\bar{h}$ in the other knowledge graph, $(\bar{h},\ r,\ t)$ will be expanded into the set D^+. For t, the exchange strategy generates $(h,\ r,\ \bar{t})$ in D^+ identically. The supplementary relational triples benefit linking two diverse knowledge graphs in the unified low-dimensional continuous vector space. The definition of D^- is described as follows:

$$\begin{aligned} D^- = &\left\{ (h', r, t) \,|h' \in \widehat{E} \wedge h' \neq h \wedge (h, r, t) \in D^+ \wedge (h', r, t) \notin D^+ \right\} \\ &\cup \left\{ (h, r, t') \,|t' \in \widehat{E} \wedge t' \neq t \wedge (h, r, t) \in D^+ \wedge (h, r, t') \notin D^+ \right\}. \end{aligned}$$

Negative examples sampled by replacing the head or tail entities of real relational triples at random are arranged to approximate the partition function.

Visual Knowledge Representations. Sometimes the relational structure information of knowledge graphs can cause ambiguity. When finding the entity aligned with "Fuji" in the other knowledge graph, "Fuji Mountain" and "Fujifilm" exist. The visual features characterize the appearance of the entity more intuitively and vividly than relational knowledge, and we can distinguish "Fuji Mountain" from "Fujifilm" because the one is a mountain, and the other one is a company logo. Therefore, visual data serve as a vital part of multi-modal knowledge graphs and visual features disambiguate the relational information to some extent.

In order to extract visual features, we achieve the vectorization of images and each of entity images is embedded into a vector. However, image vectors can not be directly applied in this scene, hence we project them to associate with entity embedding vectors. We learn embeddings for images according to the VGG16 [13] model. The model pre-trained on the ILSVRC 2012 dataset derived from ImageNet [5] is applied in our model. The filters in a stack of convolutional layers have the receptive fields of 3×3. We develop 13 convolutional layers which have different depths in various architectures. They are followed by 3 fully-connected layers, but we remove the last fully-connected layer and the softmax layer, then obtain the 4096-dimensional embeddings for all entity images. Given

a pair $\left(e^{(i)}, i\right) \in Y$ in the visual knowledge, we use the following score function to utilize visual features:

$$f_{vis}(e^{(i)}, i) = -||\mathbf{e}^{(i)} - tanh(vec(\mathbf{i}))||_2^2, \tag{3}$$

where $vec(\cdot)$ denotes the projection, and $tanh(\cdot)$ is a kind of activation function. Based on the above score function, we minimize the following loss function to optimize the visual knowledge representations:

$$L_{vis} = \sum_{(e^{(i)},i)\in Y} \log\left(1 + \exp\left(-f_{vis}(e^{(i)}, i)\right)\right). \tag{4}$$

Numerical Knowledge Representations. The numerical triple is denoted as $\left(e^{(n)}, a, n\right) \in Z$ in the numerical data, where a denotes the attribute key, and n denotes the numerical value. Attribute keys and corresponding numerical values form the *key-value pairs* to describe entities. Formally, relational structures only model the translation between head entities and tail entities while numerical features supplement the information between some entities which can not be constituted of a relational fact in the knowledge graphs. For instance, the "`height`" of "`Fuji`" is 3775.63 and the "`height` of "`Fuji Mountain`" is 3776.24, hence we deduce that they are likely to refer the same thing for entity alignment.

First of all, we deal with numeric since continuous value needs special treatment. Sparse numerical data demands to be fitted to a simple parameter distribution, and the radial basis function (RBF) [4] meets our requirement exactly. The RBF network is able to approximate any non-linear function and handle the issues of analyzing data regularity. It has good generalization ability and has a fast speed of convergence.

We convert numerical information to embeddings in high-dimensional spaces with applying a radial basis function as follows:

$$\phi\left(n_{(e^{(n)},a_i)}\right) = \exp\left(\frac{-\left(n_{(e^{(n)},a_i)} - c_i\right)^2}{\sigma_i^2}\right), \tag{5}$$

where c_i denotes the radial kernel center, σ_i denotes the variance and they are both vectors. Firstly, all corresponding numerical values for each attribute key will be normalized. After normalization, c_i and σ_i can be computed in the RBF neural network through the supervised method.

In addition, we intend to extract features from attribute keys and corresponding numerical values of entities, which indeed form the *key-value pairs*. We concatenate the embedding of an attribute key and its numerical vector got from the RBF layer. This process generates a new $2\times d$ matrix denoted by $\mathbf{M} = \left\langle \mathbf{a}, \phi(n_{(e^{(n)},a)}) \right\rangle$. Then we define the score function to measure the plausibility of the embeddings:

$$f_{num}(e^{(n)}, a, v) = -||\mathbf{e}^{(n)} - tanh(vec(\mathrm{CNN}(tanh(\mathbf{M})))\mathbf{W})||_2^2, \tag{6}$$

where CNN denotes l convolutional layers, and $\mathbf{W}$ means a fully-connected layer. We reshape the feature map to a vector, then project it to the embedding space. The loss function is given as follows:

$$L_{num} = \sum_{(e^{(n)},a,n)\in Z} \log\left(1 + \exp\left(-f_{num}(e^{(n)}, a, v)\right)\right), \tag{7}$$

where Z denotes the set of numerical triples in the numerical data. Exchanging aligned entities in the involved numerical triples, because they refer to the same real-world object across different knowledge graphs and they own the same numerical features. If a numerical triple (e, a, n) exists and $(e, \bar{e})$ appears in the seed entity alignment, $(\bar{e}, a, n)$ is added to Z.

3.3 Multi-modal Knowledge Fusion

Information from different independent sources under different modalities complements each other. Commonly, multi-modal features tend to correlate, which provide additional redundancy for better robustness. The features in the three types of modality could not be directly extracted to one space, therefore we propose a *Multi-Modal Knowledge Fusion* (MMKF) module to integrate knowledge representations from multiple modalities. MMKF migrates multi-modal knowledge embeddings from separate spaces to a common space. Common space learning enables multi-modal features to benefit from each other. It enhances the complementarity of multiple modalities which improves the accuracy of the task of entity alignment. The loss function is designed as follows:

$$L_{csl}(\mathbf{E}, \mathbf{E}^{(r)}, \mathbf{E}^{(i)}, \mathbf{E}^{(n)}) = \alpha_1||\mathbf{E} - \mathbf{E}^{(r)}||_2^2 + \alpha_2||\mathbf{E} - \mathbf{E}^{(i)}||_2^2 + \alpha_3||\mathbf{E} - \mathbf{E}^{(n)}||_2^2, \tag{8}$$

where $\mathbf{E}$ denotes the entity embeddings in the common space, and $\mathbf{E}^{(r)}$, $\mathbf{E}^{(i)}$ and $\mathbf{E}^{(n)}$ are the entity embeddings in the spaces of relational, visual and numerical knowledge, respectively. Besides, α_1, α_2 and α_3 are ratio hyper-parameters for each type of knowledge.

Since aligned entities have identical meaning in different knowledge graphs, it is intuitive for us to make those aligned entities closer in the common space. The distance between aligned entities is calculated as $\|\mathbf{e}_1 - \mathbf{e}_2\|$, where $\mathbf{e}_1, \mathbf{e}_2 \in \mathbf{E}$. Taking the distance into account, we adapt the alignment constraint approach in the common space to minimize the mapping loss:

$$L_{ac}(\mathbf{E}_1, \mathbf{E}_2) = ||\mathbf{E}_1 - \mathbf{E}_2||_2^2, \tag{9}$$

where $\mathbf{E}_1$ and $\mathbf{E}_2$ denote embeddings of entities in the sets of $\widehat{E}_1$ and $\widehat{E}_2$, which are defined as follows:

$$\widehat{E}_1 = \left\{e_1 | e_1 \in KG_1 \wedge e_1 \in \widehat{E} \wedge (e_1, e_2) \in H\right\}$$
$$\widehat{E}_2 = \left\{e_2 | e_2 \in KG_2 \wedge e_2 \in \widehat{E} \wedge (e_1, e_2) \in H\right\},$$

where H denotes the set of aligned entities across different knowledge graphs.

For the purpose of making up for imbalance among different types of knowledge, we design an interactive training stage which learns embeddings of three multi-modal (relational, visual and numerical) knowledge and optimizes the common space learning during an epoch, repeatedly. We constrain all entity embeddings with L_2 normalization to regularize embedding vectors. Firstly, we train image embeddings from VGG16 and obtain the 4096-dimensional embeddings for all the entities. Then, at each step, the parameters are updated by L_{rel}, L_{vis}, L_{num}, L_{csl} and L_{ac}.

4 Experiments

In this section, we evaluate MMEA on two real-world datasets, and demonstrate that MMEA provides insights to take advantages of multi-modal knowledge in the entity alignment task and outperforms the baselines which were shown to achieve state-of-the-art performance for entity alignment.

4.1 Experimental Settings

Datasets. In our experiments, we use two multi-modal datasets which were built in [7], namely FB15K-DB15K and FB15K-YAGO15K. FB15K is a representative subset extracted from the Freebase knowledge base. Aiming to maintain an approximate entity number of FB15K, DB15K from DBpedia and YAGO15K from YAGO are mainly selected based on the entities aligned with FB15K. Table 1 depicts the statistics of multi-modal datasets. Each dataset provides 20%, 50%, and 80% reference entity alignment as training sets, respectively.

Evaluation Metrics. We utilize cosine similarity to calculate the similarity between two entities and employ Hits@n, MRR, and MR as metrics to evaluate all the models. Hits@n means the rate correct entities rank in the top n according to similarity computing. MR denotes the mean rank of correct entities and MRR denotes the mean reciprocal rank of correct entities. The higher values of Hits@n and MRR explain the better performance of the method, while the lower value of MR proves it.

Implementation Details. All the experiments are tuned for both datasets. For MMEA we initialize the embeddings of KGs in each type of knowledge with Xavier initializer and restrain their lengths to 1. The dimensions of all the embeddings are set as 100. We adopt the mini-batch method with the batch size of 5000. We start to valid every 10 epochs after 300 epochs and stop the training when the metric MRR is declining continually in the valid set. We set all the learning rates to 0.01 except that the learning rate of common space learning is 0.004. In addition, the max epochs are set as 600. More specifically, γ in the

relational knowledge representation is 1.5. In the numerical knowledge representation, l and the number of filters are both set as 2. The kernel size is 2×4. α_1, α_2, and α_3 in the common space learning are selected as $\{1, 0.01, 1\}$ on FB15K-DB15K dataset and $\{1, 1, 0.01\}$ on FB15K-YAGO15K dataset, respectively. We optimize all the above loss functions using stochastic gradient descent (SGD).

Table 1. Statistics of multi-modal datasets.

Datasets	Entities	Relations	Attributes	Relational triples	Numerical triples	Images	Links
FB15K	14951	1345	116	592213	29395	13444	–
DB15K	12842	279	225	89197	48080	12837	12846
YAGO15K	15404	32	7	122886	23532	11194	11199

4.2 Compared Methods

To demonstrate that MMEA framework outperforms the state-of-the-art entity alignment models, we compare it with the following methods:

- **TransE** [1] is a typical translational method for knowledge graph embedding. We perform this method in the entity alignment task by sharing the parameters between aligned entities.
- **MTransE** [3] learns the translation matrix to map the aligned entities from different knowledge graphs in the unified space. It acquires a great deal of seed alignment, otherwise the translation matrix will be inaccurate.
- **IPTransE** [20] obtains entity embeddings through employing an iterative and parameter sharing method. Additionally, soft alignment and multi-step relation paths are utilized to align entities from different KGs.
- **SEA** [10] served as a semi-supervised method realizes the adversarial training with an awareness of the degree difference and leverages both labeled entities and the abundant unlabeled entity information for the alignment.
- **GCN** [17] adopts GCNs to encode the structural information of entities, and combine relation and attribute embeddings for the entity alignment task.
- **IMUSE** [6] generates lots of high-quality aligned entities with an unsupervised method. Besides, a bivariate regression is utilized to merge the alignment results of relations and attributes better.
- **PoE** [7] combines the multi-modal features and measures the credibility of facts by matching the underlying semantics of the entities and mining the relations contained in the embedding space. Regarding computing the scores of facts under each modality, it learns the entity embeddings for entity alignment.

4.3 Results and Analyses

We partition the datasets to compare the results of all models. For each dataset, we use the 20%, 50%, 80% data as training sets, and the remains are treated as testing sets, respectively.

Table 2. 20% alignment results on two datasets. (R.: Relational knowledge, N.: Numerical knowledge, V.: Visual knowledge)

Models		FB15K-DB15K					FB15K-YAGO15K				
		Hits@1	Hits@5	Hits@10	MR	MRR	Hits@1	Hits@5	Hits@10	MR	MRR
R.	MTransE	0.359	1.414	2.492	1239.465	0.0136	0.308	0.988	1.783	1183.251	0.011
	IPTransE	3.985	11.226	17.277	387.512	0.0863	3.079	9.505	14.443	522.235	0.07
	TransE	7.813	17.95	24.012	442.466	0.134	6.362	15.11	20.254	522.545	0.112
	PoE-l	7.9	–	20.3	–	0.122	6.4	–	16.9	–	0.101
	SEA	16.974	33.464	42.512	191.903	0.255	14.084	28.694	37.147	207.236	0.218
R. + N.	GCN	4.311	10.956	15.548	810.648	0.0818	2.27	7.209	10.736	1109.845	0.053
	IMUSE	17.602	34.677	43.523	182.843	0.264	8.094	19.241	25.654	397.571	0.142
R. + N. + V.	PoE-lni	12.0	–	25.6	–	0.167	10.9	–	24.1	–	0.154
	MMEA	**26.482**	**45.133**	**54.107**	**124.807**	**0.357**	**23.391**	**39.764**	**47.999**	**147.441**	**0.317**

Performance Comparison. Table 2 lists the results of all the models with 20% alignment data on FB15K-DB15K and FB15K-YAGO15K datasets. The results for PoE are taken from [7]. From the overview, our proposed MMEA achieves the state-of-the-art performance for entity alignment. Specifically, there are several observations. First, MMEA performs better than all the other methods. Compared with these methods, Hits@1, Hits@5, Hits@10, MRR are at least improved by 8.88%, 10.456%, 10.584%, 0.093 and 9.307%, 11.07%, 10.852%, 0.099, and MR is at least decreased by 58.036 and 59.795 on two datasets. The results indicate that MMEA is more suitable for multi-modal knowledge graphs from the real scenarios. Second, solutions with multi-modal knowledge generate better results than solutions with a single modality in most cases. Both MMEA and PoE-lni outperform MTransE, IPTransE, TransE and PoE-l, which indicates that as increasing numerical and visual knowledge leads to improvements, the effects of multi-modal knowledge have been proven. Third, MMEA outperforms PoE-lni absolutely, suggesting our modeling for multi-modal knowledge is more effective, and multi-modal fusion method with common space is better.

Figure 3 shows the experimental results with different test splits on FB15K-DB15K and FB15K-YAGO15K datasets. In most cases, especially when only 20% alignment data is split to the training set, MMEA with a significant margin compared with the state-of-the-art methods could make full use of limited data. Moreover, it demonstrates the robustness and effectiveness of MMEA once again.

Ablation Study. To further validate the effectiveness of multi-modal knowledge in the task of entity alignment, we design two variants for ablation study, namely

MMEA-R and MMEA-RN. MMEA-R is a variant of MMEA with only relational knowledge, and MMEA-RN is a variant of MMEA with relational and numerical knowledge. According to the experimental results, MMEA outperforms both MMEA-R and MMEA-RN, which reveals that multi-modal knowledge complements the absence of useful entity features, and MMEA provides insights to take advantages of multi-modal knowledge. Moreover, it is obvious that our multi-modal knowledge fusion method could leverage multi-modal knowledge for entity alignment.

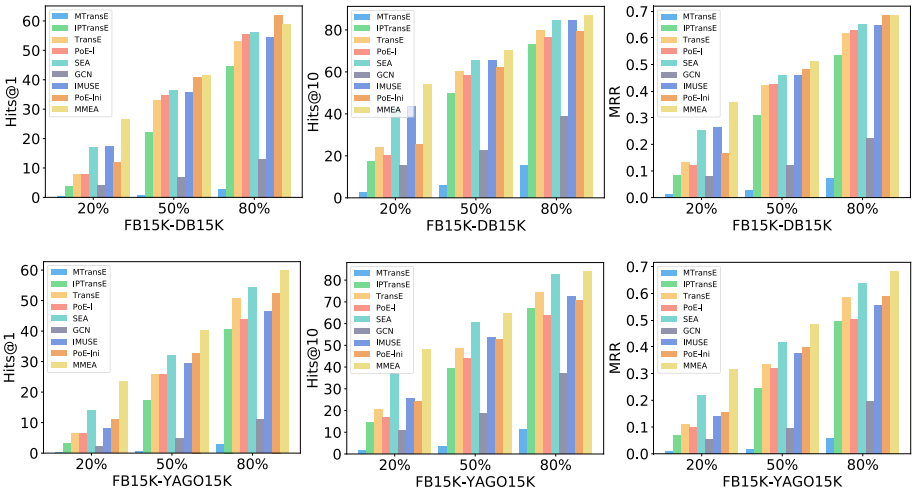

Fig. 3. Experimental results with different test splits on two datasets.

In summary, all above evidences demonstrate that MMEA framework has a good ability to find entities referring to the same real-world identity from different KGs by taking full advantages of multi-modal knowledge and achieves state-of-the-art performance for entity alignment (Table 3).

Table 3. Ablation study.

Models		FB15K-DB15K					FB15K-YAGO15K				
		Hits@1	Hits@5	Hits@10	MR	MRR	Hits@1	Hits@5	Hits@10	MR	MRR
20%	MMEA-R	24.957	43.084	51.581	143.171	0.340	22.199	38.563	46.493	160.576	0.305
	MMEA-RN	26.209	44.982	53.759	125.874	0.355	23.091	39.589	47.689	154.908	0.314
	MMEA	**26.482**	**45.133**	**54.107**	**124.807**	**0.357**	**23.391**	**39.764**	**47.999**	**147.441**	**0.317**
50%	MMEA-R	40.95	61.362	69.721	58.093	0.505	39.161	56.696	63.956	65.848	0.477
	MMEA-RN	41.436	61.691	70.089	54.53	0.51	39.509	56.959	63.979	65.255	0.48
	MMEA	**41.653**	**62.1**	**70.345**	**54.257**	**0.512**	**40.263**	**57.231**	**64.51**	**62.969**	**0.486**
80%	MMEA-R	58.256	80.192	86.466	14.557	0.679	58.803	77.078	83.132	15.308	0.672
	MMEA-RN	58.411	80.355	86.76	14.493	0.681	59.377	78.22	83.34	14.745	0.68
	MMEA	**59.034**	**80.405**	**86.869**	**14.129**	**0.685**	**59.763**	**78.485**	**83.892**	**14.512**	**0.682**

5 Conclusion

In this paper, we proposed a novel solution for the entity alignment task in multi-modal knowledge graphs, which integrated multiple representations of different types of knowledge based on knowledge embedding. Moreover, a multi-modal fusion method was designed through common space learning to migrate features under different knowledge spaces. Extensive experiments on two real-world datasets demonstrated the robustness and effectiveness of our solution for multi-modal entity alignment, which outperformed several state-of-the-art baseline methods with a significant margin.

Acknowledgments. This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2018YFB1402600), and the National Natural Science Foundation of China (Grant No. 61703386, U1605251).

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, pp. 2787–2795 (2013)
2. Chen, M., Tian, Y., Chang, K.W., Skiena, S., Zaniolo, C.: Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. arXiv preprint arXiv:1806.06478 (2018)
3. Chen, M., Tian, Y., Yang, M., Zaniolo, C.: Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. arXiv preprint arXiv:1611.03954 (2016)
4. Chen, S., Cowan, C.F., Grant, P.M.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Trans. Neural Netw. **2**(2), 302–309 (1991)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. He, F., et al.: Unsupervised entity alignment using attribute triples and relation triples. In: Li, G., Yang, J., Gama, J., Natwichai, J., Tong, Y. (eds.) DASFAA 2019. LNCS, vol. 11446, pp. 367–382. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-18576-3_22
7. Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., Rosenblum, D.S.: MMKG: multi-modal knowledge graphs. In: Hitzler, P., et al. (eds.) ESWC 2019. LNCS, vol. 11503, pp. 459–474. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21348-0_30
8. Mousselly-Sergieh, H., Botschen, T., Gurevych, I., Roth, S.: A multimodal translation-based approach for knowledge graph representation learning. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pp. 225–234 (2018)
9. Niu, X., Rong, S., Wang, H., Yu, Y.: An effective rule miner for instance matching in a web of data. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1085–1094 (2012)

10. Pei, S., Yu, L., Hoehndorf, R., Zhang, X.: Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In: The World Wide Web Conference, pp. 3130–3136 (2019)
11. Pezeshkpour, P., Chen, L., Singh, S.: Embedding multimodal relational data for knowledge base completion. arXiv preprint arXiv:1809.01341 (2018)
12. Raimond, Y., Sutton, C., Sandler, M.B.: Automatic interlinking of music datasets on the semantic web. LDOW **369** (2008)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Sun, Z., Hu, W., Zhang, Q., Qu, Y.: Bootstrapping entity alignment with knowledge graph embedding. In: IJCAI, pp. 4396–4402 (2018)
15. Trisedya, B.D., Qi, J., Zhang, R.: Entity alignment between knowledge graphs using attribute embeddings. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 297–304 (2019)
16. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Bernstein, A., et al. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04930-9_41
17. Wang, Z., Lv, Q., Lan, X., Zhang, Y.: Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 349–357 (2018)
18. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2840–2848 (2017)
19. Xie, R., Liu, Z., Luan, H., Sun, M.: Image-embodied knowledge representation learning. arXiv preprint arXiv:1609.07028 (2016)
20. Zhu, H., Xie, R., Liu, Z., Sun, M.: Iterative entity alignment via joint knowledge embeddings. In: IJCAI, pp. 4258–4264 (2017)