MultiKE: A Multi-view Knowledge Graph Embedding Framework for Entity Alignment

Wei Hu, Qingheng Zhang, Zequn Sun and **Jiacheng Huang**
National Key Laboratory for Novel Software Technology, Nanjing University, China
whu@nju.edu.cn, {qhzhang, zqsun, jchuang}.nju@gmail.com

1. Introduction

Background:

- Entity alignment, a.k.a. entity matching or resolution, aims to find entities in different KGs referring to the same real-world identity.
- Entities in KGs have various features, but the current embedding-based entity alignment methods exploit just one or two types of them.
- The existing embedding-based entity alignment methods rely on abundant seed entity alignment as labeled training data[1].

MultiKE:

- We define three representative views based on the name, relation and attribute features to learn entity embeddings.
- Two cross-KG identity inference methods are designed to preserve and enhance the alignment between different KGs.
- We present three different strategies to combine multiple view-specific entity embeddings.

2. Problem Formulation

Preliminaries: We formalize a KG as a 7-tuple $\mathbf{G} = (\mathbf{E}, \mathbf{R}, \mathbf{A}, \mathbf{V}, \mathbf{N}, \mathbf{X}, \mathbf{Y})$, where $\mathbf{E}$, $\mathbf{R}$, $\mathbf{A}$ and $\mathbf{V}$ denote the sets of entities, relations, attributes and literals, respectively. $\mathbf{N} \subseteq \mathbf{E}\times\mathbf{V}$ denotes the name view, $\mathbf{X} \subseteq \mathbf{E}\times\mathbf{R}\times\mathbf{E}$ denotes the relation view, and $\mathbf{Y} \subseteq \mathbf{E}\times\mathbf{A}\times\mathbf{V}$ denotes the attribute view. The name, relation and attribute views are marked by $^{(1)}$, $^{(2)}$ and $^{(3)}$, respectively.

Problem: Given a source KG $\mathbf{G}_\mathrm{a} = (\mathbf{E}_\mathrm{a}, \mathbf{R}_\mathrm{a}, \mathbf{A}_\mathrm{a}, \mathbf{V}_\mathrm{a}, \mathbf{N}_\mathrm{a}, \mathbf{X}_\mathrm{a}, \mathbf{Y}_\mathrm{a})$ and a target KG $\mathbf{G}_\mathrm{b} = (\mathbf{E}_\mathrm{b}, \mathbf{R}_\mathrm{b}, \mathbf{A}_\mathrm{b}, \mathbf{V}_\mathrm{b}, \mathbf{N}_\mathrm{b}, \mathbf{X}_\mathrm{b}, \mathbf{Y}_\mathrm{b})$, entity alignment aims to find a set of identical entities $\mathbf{M} = \{(\mathbf{e}_\mathrm{i}, \mathbf{e}_\mathrm{j}) \in \mathbf{E}_\mathrm{a}\times\mathbf{E}_\mathrm{b} | \mathbf{e}_\mathrm{i} \equiv \mathbf{e}_\mathrm{j}\}$, where "≡" denotes the equivalence relationship.

3. Multi-view KG Embedding

Literal embedding:

Let $l = (o_1, o_2, \dots, o_n)$ denote a literal of n tokens. We employ an autoencoder to encode a list of token embeddings into one literal embedding :

$$\varphi(l) = \mathrm{encode}([\mathrm{LP}(o_1); \mathrm{LP}(o_2); \dots; \mathrm{LP}(o_n)]),$$

where encode$(\cdot)$ returns the compressed representation of the input embeddings, LP$(\cdot)$ is defined as a lookup function that maps the input to an embedding, and [;] denotes the concatenation operation.

Name view embedding:

We embed the name view using the above literal embeddings. Given an entity h, its name embedding is defined as follows:

$$\mathbf{h}^{(1)} = \varphi(\mathrm{name}(h)),$$

where name$(\cdot)$ extracts the name of the input object.

Relation view embedding:

We adopt TransE [2] to preserve relational structures. Given a relation fact (h, r, t), we use the following score function to measure the plausibility:

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

Then, we define the probability of (h, r, t) being a real relation fact as follows:

$$P_{rel}(\zeta_{(h,r,t)} = 1 | \Theta^{(2)}) = sigmoid\left(f_{rel}(\mathbf{h}^{(2)}, \mathbf{r}, \mathbf{t}^{(2)})\right),$$

where $\Theta^{(2)}$ denotes the relation view embeddings and $\zeta_{(h,r,t)}$ denotes the label (1 or -1) of (h, r, t). We parameterize $\Theta^{(2)}$ by minimizing the loss below:

$$L(\Theta^{(2)}) = \sum_{(h,\ r,\ t) \in \mathbf{X}^+ \cup \mathbf{X}^-} \log\left(1 + \exp\left(-\zeta_{(h,r,t)} f_{rel}(\mathbf{h}^{(2)}, \mathbf{r}, \mathbf{t}^{(2)})\right)\right),$$

where $\mathbf{X}^+ = \mathbf{X}_\mathrm{a} \cup \mathbf{X}_\mathrm{b}$, denotes real relation fact set, while $\mathbf{X}^-$ denotes the set of faked ones sampled by replacing the head or tail entities with random ones.

Attribute view embedding:

For the attribute view, we use a convolutional neural network(CNN) to extract features from the attributes and values of entities. Given an attribute fact (h, a, v) in KGs, we define the following score function to measure its plausibility:

$$f_{attr}(\mathbf{h}^{(3)}, \mathbf{a}, \mathbf{v}) = -\|\mathbf{h}^{(3)} - \mathrm{CNN}(\langle \mathbf{a}; \mathbf{v} \rangle)\|,$$

where CNN$(\cdot)$ denotes a convolution operation. This objective can be achieved by minimizing the following logistic loss:

$$L(\Theta^{(3)}) = \sum_{(h,\ a,\ v) \in \mathbf{Y}^+} \log\left(1 + \exp\left(-f_{attr}(\mathbf{h}^{(3)}, \mathbf{a}, \mathbf{v})\right)\right),$$

where $\mathbf{Y}^+ = \mathbf{Y}_\mathrm{a} \cup \mathbf{Y}_\mathrm{b}$ denotes the real attribute fact set, and $\Theta^{(3)}$ denotes the attribute view embeddings.

4. Cross-KG Training for Entity Alignment

Entity Identity Inference:

Given a relation fact (h, r, t), if $(h, \hat{h})$ appears in the seed entity alignment, we add the following auxiliary probability:

$$P_{rel}(\zeta_{(h,r,t)} = 1 | \Theta^{(2)}) = sigmoid\left(f_{rel}(\hat{\mathbf{h}}^{(2)}, \mathbf{r}, \mathbf{t}^{(2)})\right),$$

We maximize these auxiliary probabilities over the relation facts having those entities in the seed entity alignment. The loss is computed as follows:

$$L_{CE}(\Theta^{(2)}) = \sum_{(h,\ r,\ t) \in \mathbf{X}'} \log\left(1 + \exp\left(-f_{rel}(\hat{\mathbf{h}}^{(2)}, \mathbf{r}, \mathbf{t}^{(2)})\right)\right) + \sum_{(h,\ r,\ t) \in \mathbf{X}''} \log\left(1 + \exp\left(-f_{rel}(\mathbf{h}^{(2)}, \mathbf{r}, \hat{\mathbf{t}}^{(2)})\right)\right),$$

where $\mathbf{X}'$ and $\mathbf{X}''$ refer to the sets of relation facts whose head and tail entities are in the seed entity alignment, respectively.

Relation and Attribute Identity Inference:

We add the auxiliary probabilities for the cross-KG relation and attribute identity inference. Let $sim(r, \hat{r})$ denote the similarity of $\langle r, \hat{r} \rangle$, and the loss is:

$$L_{CRA}(\Theta^{(2)}) = \sum_{(h,\ r,\ t) \in \mathbf{X}'''} sim(r, \hat{r}) \log\left(1 + \exp\left(-f_{rel}(\mathbf{h}^{(2)}, \hat{\mathbf{r}}, \mathbf{t}^{(2)})\right)\right),$$

where $\mathbf{X}'''$ denotes the set of relation facts having the soft alignment relations.

5. View Combination

Weighted View Averaging:

Let $\tilde{\mathbf{h}}$ denote the combined embedding for h. Without loss of generality, let D be the number of views, and we have $\tilde{\mathbf{h}} = \sum_{i=1}^{D} w_i \mathbf{h}^{(i)}$, where w_i is the weight of $\mathbf{h}^{(i)}$, and can be calculated by:

$$w_i = \frac{\cos(\mathbf{h}^{(i)}, \bar{\mathbf{h}})}{\sum_{j=1}^{D} \cos(\mathbf{h}^{(j)}, \bar{\mathbf{h}})},$$

Where $\bar{\mathbf{h}}$ is the average of multi-view embeddings of h.

Shared Space Learning:

Let $\tilde{\mathbf{H}}$ be the combined embedding matrix for all entities, and $\mathbf{H}^{(i)}$ be the entity embedding matrix under the i^{th} view. We minimize the mapping loss:

$$L_{SSL}(\tilde{\mathbf{H}}, \mathbf{Z}) = \sum_{i=1}^{D} \left(\|\tilde{\mathbf{H}} - \mathbf{H}^{(i)}\mathbf{Z}^{(i)}\|_F^2 + \|\mathbf{I} - \mathbf{Z}^{(i)\mathrm{T}}\mathbf{Z}^{(i)}\|_F^2\right),$$

where $\mathbf{Z}^{(i)}$ serves as the mapping from the i^{th} view-specific embedding space to the shared space, and $\mathbf{I}$ is the identity matrix.

In-training Combination:

This combination participates in the joint training, and the loss is:

$$L_{ITC}(\tilde{\mathbf{H}}, \mathbf{H}) = \sum_{i=1}^{D} \left(\|\tilde{\mathbf{H}} - \mathbf{H}^{(i)}\|_F^2\right),$$

where $\mathbf{H} = \cup_{i=1}^{D} \mathbf{H}^{(i)}$.

6. Experiments

In our experiments, we reused two datasets, namely DBP-WD and DBP-YG, recently proposed in [3]. The comparison results of MultiKE and other embedding-based entity alignment methods are shown in the following table:

Features		Methods	DBP-WD				DBP-YG			
			Hits@1	Hits@10	MR	MRR	Hits@1	Hits@10	MR	MRR
Relation only		MTransE	28.12	51.95	656	0.363	25.15	49.29	512	0.334
		IPTransE	34.85	63.84	265	0.447	29.74	55.76	158	0.386
		BootEA	74.79	89.84	109	0.801	76.10	89.44	34	0.808
		GCN-Align	47.70	75.96	1,988	0.577	60.05	84.14	299	0.686
Rel. +	Attr.	JAPE	31.84	58.88	266	0.411	23.57	48.41	189	0.320
	Desc.	KDCoE	57.19	69.53	182	0.618	42.71	48.30	137	0.446
	Literal	AttrE	38.96	66.77	142	0.487	23.24	42.70	706	0.300
Multi-view		MultiKE-WVA	90.42	94.59	**22**	0.921	85.92	94.99	**19**	0.891
		MultiKE-SSL	**91.86**	**96.26**	39	**0.935**	82.35	93.30	21	0.862
		MultiKE-ITC	91.45	95.19	114	0.928	**88.03**	**95.32**	35	**0.906**

Experimental results:

- MultiKE **significantly outperformed** the others on all the metrics.
- The three variants all achieved similar results.
- The three views all contributed to entity alignment.

Codes and datasets of MultiKE are available at https://github.com/nju-websoft/MultiKE.

7. Conclusion

We proposed a multi-view KG embedding framework for entity alignment.

- The framework learns entity embeddings from three representative views of KGs, name view, relation view, and attribute view.
- We introduced two cross-KG training methods for alignment inference.
- We designed three kinds of strategies to combine view-specific embeddings.

8. References

[1] Zequn Sun, Wei Hu, and Chengkai Li. Cross-lingual entity alignment via joint attribute preserving embedding. In Proceedings of ISWC, pages 628–644, 2017.

[2] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In Proceedings of NIPS, pages 2787-2795, 2013.

[3] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping entity alignment with knowledge graph embedding. In Proceedings of IJCAI, pages 4396–4402, 2018.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (Nos. 61872172, 61772264) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.