2022 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2022, Virtual, December 12-15, 2022

Common Space Learning with Gaussian Embedding for Multi-Modal Entity Alignment

Kenta Hama† and Takashi Matsubara†

†Graduate School of Engineering Science, Osaka University
1-3, Machikaneyama, Toyonaka, Osaka, Japan
Email: hama@hopf.sys.es.osaka-u.ac.jp, matsubara@sys.es.osaka-u.ac.jp

Abstract—Structuring data is important in information retrieval, and knowledge graphs are used as structured knowledge representations. As knowledge graphs become larger, it becomes more important to complement missing or erroneous information. Multi-modal information such as images and attribute values are useful as supplemental information. For this reason, there has been a lot of research on entity alignment, which finds entities of the same concept in different multi-modal knowledge graphs. However, if the supplemental information itself is missing or incorrect, the addition of that information will negatively affect information retrieval. If we can quantify the usefulness of the information for retrieval as a degree of importance, the influence of unimportant supplementary information can be reduced. In this study, we proposed a method that expresses the importance of each piece of information by using a probability distribution. The proposed method outperformed multi-modal entity alignment in the entity alignment task of two multi-modal knowledge graphs.

1. Introduction

A knowledge graph is a data structure that represents human knowledge as a directed graph with edges representing relationships and nodes representing entities. It is used in tasks such as question answering, recommendation systems, and information retrieval. As datasets in machine learning become larger and larger, it is becoming increasingly important to integrate information from multiple knowledge graphs to fill in missing information in a single knowledge graph [1]. However, since the purposes of creating knowledge graphs are diverse, and the domains and languages used are different, there are gaps in the description and graph structure of concepts that refer to the same object in different knowledge graphs. The task of dealing with these gaps and linking entities that refer to the same object across different knowledge graphs is called entity alignment.

To improve the accuracy of entity alignment, multi-modal information could be used. A knowledge graph with images, numerical information, descriptions, and other information representing entities is called a multi-modal knowledge graph (MMKG). However, if multi-modal information is automatically collected for general knowledge graphs, there is a risk that noisy information will be added, degrading the performance of entity alignment. If unimportant information can be automatically removed during model training, the effect of noise can be reduced. MMEA [4] (multi-modal entity alignment) is a typical embedding-based method for multi-modal entity alignment. It is generic and scalable because it learns and integrates each embedding point independently using information from each modal.

However, it does not take into account the importance of information for each modal when integrating information In this study, we propose a new common-space learning method for multi-modal entity alignment that transforms the information of each modal into a multivariate normal distribution instead of a point in space, so that the importance of information can be expressed as the size of variance of the distribution. The proposed method achieves accuracy significantly better than MMEA on two datasets of MMKG [5], which is a common evaluation dataset for multi-modal entity alignment.

2. Related Work

A knowledge graph (KG) is structured data consisting of an entity representing a concept and a relation between two entities. The MMKG is an extension that each entity in the KG has multi-modal information such as images and attributes. Entities are used in machine learning methods such as deep learning, so they are often converted to a distributed representation that can compute the similarity between entities [2]. This is called knowledge graph embedding.

MMKG has been studied to improve the quality of representation by using information from other modalities for embedding. IKRL [6] (image-embodied knowledge representation learning) integrates knowledge graph embedding with features and attention mechanisms obtained from images of entities. MKBE [7] (multi-modal knowledge base embeddings) also simultaneously uses numerical and categorical information as well as relations and images. These methods use information from other modalities as supplementary information for embedded representation acquisi-

ORCID iDs Kenta Hama: 0000-0002-2338-9229, Takashi Matsubara: 0000-0003-0642-4800

tion. For entity alignment, MMEA is a generic and scalable model: it learns the relational information between entities, the image information of entities, and the numerical information as different embedding representations, and integrates the embeddings of each modality by bringing them closer to one point in the common space. Since the representation of each modality is obtained independently and then integrated, it is easy to add modal information and transfer learned models. However, because MMEA treats the information of each modality equivalently, the information of a modality that is not important for an entity can negatively affect the embedding. Therefore, it is necessary to learn the knowledge representation while weighting the information of each modality according to its importance.

3. Proposed Method

In this study, entities are transformed into a multivariate normal distribution instead of a point in space in order to be able to express the uncertainty of each modal embedding. In this section, we explain the terminology used in this paper, and then describe in detail how to learn the representation by probability distribution from each modal information. Finally, we describe common-space learning, which integrates each embedding using uncertainty.

3.1. Definitions of Terms

Denote the MMKG as $KG = (\hat{E}, R, I, N, X, Y, Z)$. where $\hat{E}$,R,I,N are the sets of entities, relations, images, and numbers, and X,Y,Z are the sets of triples of relations, entity-image pairs, and numbers, respectively.

Entity alignment is the task of matching entities that describe the same thing in the real world from different knowledge graphs. Let $KG_1 = (\hat{E}_1, R_1, I_1, N_1, X_1, Y_1, Z_1)$ and $KG_2 = (\hat{E}_2, R_2, I_2, N_2, X_2, Y_2, Z_2)$ be two different KGs, then $H = \{(e_1, e_2) \mid e_1 \in \hat{E}_1, e_2 \in \hat{E}_2\}$ denotes the set of pairs of entities that describe the same thing in the whole knowledge graph.

3.2. Probability Distribution Embedding

3.2.1. Relational Distribution

A triple consisting of two entities and a relation between them is called a fact and is represented as $(h, r, t) \in X$. In the proposed method, $(h, r, t) \in X$ are represented by $\mathcal{N}(\mu_h, \Sigma_h)$, $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_t, \Sigma_t)$, where are multivariate normal distributions.

Similarly to [8], we define the similarity score between $\mathcal{N}(\mu_t, \Sigma_t) - \mathcal{N}(\mu_h, \Sigma_h)$ and $\mathcal{N}(\mu_r, \Sigma_r)$ using Kullback-Leibler (KL)-divergence as follows:

$$f_{rel}(h, r, t) = -D_{KL}(\mathcal{N}(\mu_h - \mu_t, \Sigma_h - \Sigma_t), \mathcal{N}(\mu_r, \Sigma_r)). \quad (1)$$

The KL-divergence between two multivariate normal distributions can be calculated as follows:

$$\begin{aligned} &D_{KL}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) \\ &= \int_{x \in R^{k_1}} \mathcal{N}(x \mid \mu_2, \Sigma_2) \log \frac{\mathcal{N}(x \mid \mu_1, \Sigma_1)}{\mathcal{N}(x \mid \mu_2, \Sigma_2)} dx \\ &= \frac{1}{2}\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) \\ &\qquad - \log \frac{\det(\Sigma_1)}{\det(\Sigma_2)} - k_1\}, \end{aligned}$$

where $\mathrm{tr}(\Sigma)$ are the trace of the covariance matrix Σ, Σ^{-1} are the inverse of Σ, and k_1 is the dimension of entity in the embedding space. In this study, the covariance matrix Σ is assumed to be a diagonal matrix to simplify these calculations.

The loss function for learning the relation embedding is defined by using the margin γ as follows:

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

where D^+ and D^- are the sets of positive and negative examples of the fact, respectively. The positive examples are given as $\tau = (h, r, t) \in X$ at training time, but in this study, as in [4], X is extended to D^+ using an exchange strategy. The exchange strategy is that given $(h, r, t) \in X$, if $(h, \bar{h}) \in H$, then $(\bar{h}, r, t)$ is also added to D^+. This is done for t as well. Once D^+ is obtained, the set of negative examples, D^-, is generated with the following definition.

$$\begin{aligned} &D^- = \\ &\{(h', r, t) \mid h' \in \hat{E} \wedge h' \neq h \wedge (h, r, t) \in D^+ \wedge (h', r, t) \notin D^+\} \\ &\cup \{(h, r, t') \mid t' \in \hat{E} \wedge t' \neq t \wedge (h, r, t) \in D^+ \wedge (h, r, t') \notin D^+\} \end{aligned}$$

3.2.2. Visual Distribution

Each entity is given image information represented as $(e^{(i)}, \boldsymbol{i}) \in Y$. From the image given for an entity, we extract a 4096 dimensional vector before the classification layer as a feature using VGG16, which has already been trained on ImageNet. The proposed method transforms $e^{(i)}$ into a normal distribution $\mathcal{N}(\mu_{e^{(i)}}, \Sigma_{e^{(i)}})$. We make a distribution $\mathcal{N}(\mu_i, \Sigma_i)$ from feature $\boldsymbol{i}$, the output of VGG16. The mean vector is $\mu_i = \tanh(\boldsymbol{M_1 i})$, where $\boldsymbol{M_1}$ is a $4096 \times d$ matrix and d is the dimension of the mean vector of the entity. The Σ_i is a diagonal matrix with all elements fixed at 0.5. For $(e^{(i)}, i) \in Y$ given as pairs, we define the following score functions:

$$f_{vis}(e^{(i)}, i) = -D_{KL}(\mathcal{N}(\mu_{e^{(i)}}, \Sigma_{e^{(i)}}), \mathcal{N}(\mu_i, \Sigma_i)). \quad (3)$$

The loss function for embedding image information is defined as follows:

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

3.2.3. Numerical Distribution

A triple of numeric information is represented as $(e^{(n)}, a, n_{(e^{(n)},\, a)}) \in Z$, where a is the attribute name and $n_{(e^{(n)},\, a)}$ is its numerical value. Here, the numerical values are real numbers, so they are converted into a distributed representation using the radial basis function (RBF) as follows:

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

where $\boldsymbol{c}_i$ is the radial kernel center vector and $\boldsymbol{\sigma}_i$ is the variance vector. The numerical values are normalized by attribute name.

The proposed method transforms $e^{(n)}$ into the normal distribution $\mathcal{N}(\mu_{e^{(n)}}, \Sigma_{e^{(n)}})$. We make a distribution $\mathcal{N}(\mu_n, \Sigma_n)$ from $(a, n_{(e^{(n)},\, a)})$. The mean vector is $\mu_n = \tanh(\mathrm{vec}(\mathrm{CNN}(\tanh(\boldsymbol{M_2}))\boldsymbol{W})$ where $\mathrm{vec}(\cdot)$ denotes the projection, $\mathrm{CNN}(\cdot)$ is the l-layer convolutional layer, and $\boldsymbol{W}$ means a fully-connected layer. $\boldsymbol{M_2}$ is a matrix of $2 \times d$, which is a concatenation of $\boldsymbol{a}$, an embedding of attribute names, and $\phi(n_{(e^{(n)},a_i)})$, a distributed representation of numbers. Σ_i is a diagonal matrix with all elements fixed at 0.5. For $(e^{(n)}, a, n) \in Z$, we define the following score function:

$$f_{num}(e^{(n)}, a, n) = -D_{KL}(\mathcal{N}(\mu_{e^{(n)}}, \Sigma_{e^{(n)}}), \mathcal{N}(\mu_n, \Sigma_n)). \tag{6}$$

The loss function of the entire numerical information embedding is as follows:

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

3.2.4. Common Space Learning

In the case of the proposed method, each modal information is represented by a multivariate normal distribution, so the loss function to be integrated in the common space is defined as follows:

$$\begin{aligned} L_{csl}(e, e^{(r)}, e^{(i)}, e^{(n)}) = \\ D_{KL}(\mathcal{N}(\mu_e, \Sigma_e), \mathcal{N}(\mu_{e^{(r)}}, \Sigma_{e^{(r)}})) \\ + D_{KL}(\mathcal{N}(\mu_e, \Sigma_e), \mathcal{N}(\mu_{e^{(i)}}, \Sigma_{e^{(i)}})) \\ + D_{KL}(\mathcal{N}(\mu_e, \Sigma_e), \mathcal{N}(\mu_{e^{(n)}}, \Sigma_{e^{(n)}})), \end{aligned} \tag{8}$$

where e is an entity in the common space and $\mathcal{N}(\mu_e, \Sigma_e)$ is its normal distribution embedding representation. Finally, to bring corresponding entities closer together across different knowledge graphs, we minimize the following loss function minimize the following loss function

$$L_{ac}(e_1, e_2) = D_{KL}(\mathcal{N}(\mu_{e_1}, \Sigma_{e_1}), \mathcal{N}(\mu_{e_2}, \Sigma_{e_2})) \tag{9}$$

The defined $L_{rel}, L_{vis}, L_{num}, L_{csl}, L_{ac}$ are trained by iteratively updating the parameters one epoch at a time as in MMEA.

4. Experiment

4.1. Dataset

In this study, we used two MMKG datasets, FB15K-DB15K and FB15K-YAGO15K, created by [5] et al. FB15K is a dataset commonly used in knowledge graph completion, and the entities in FB 15K entities and related entities were selected from DBpedia and YAGO, and DB15K and YAGO15K were created. The datasets used in this study are the same as those used in MMEA, and the training dataset was randomly split 5 times into 20%, 50%, and 80% splits, and the comparison between the models was based on the mean value of the training evaluation results in each dataset.

4.2. Evaluation Metrics

Hits@n, MRR (mean reciprocal rank), and MR (mean rank), which are commonly used in ranking evaluation, were used as evaluation indices for entity alignment. Hits@n is the percentage of correct entities within the top n of the ranking obtained by the similarity calculation, MR is the average of the ranks of correct entities, and MRR is the average of the inverse of the ranks of the correct entities. Therefore, the higher Hits@n and MRR and the lower MR, the better the performance.

4.3. Experimental Settings

In this study, OpenEA [13] was used for implementation. All models used in the comparison experiments used the default training settings of OpenEA, except MMEA, which was trained similarly to the training parameter settings in [4]. Each model used cross-domain similarity local scaling [14] during the evaluation.

In order to stabilize the learning process, a two-stage learning process was used: the mean of the normal distribution was learned first, and then the parameter for variance was unfixed and relearned. To avoid divergence of the variance values during learning, the range of $[C_{min}, C_{max}] = [0.5, 50]$ was used. The initial value of all variances was set to 0.5.

4.4. Experimental Results

The results of the evaluation of each model when the training data is divided by 20 percent of the total are shown in Table 1. Compared to the models using only relational information, the models using relational + numerical and relational + numerical + visual information are more accurate. The proposed method achieves higher accuracy than MMEA.

The evaluation results of MMEA and the proposed method when the training data is split at 20, 50, and 80% are shown in Table 2. It can be seen that the proposed method outperforms MMEA for all dataset splits.

Table 1: Evaluation results for each model in FB15K-DB15K and FB15K-YAGO15K when the training dataset is divided into 20% of the total dataset. R, N, and V stand for relational, numerical, and visual, respectively.

Modal	Model	FB15K-DB15K					FB15K-YAGO15K				
		H@1	H@5	H@10	MR	MRR	H@1	H@5	H@10	MR	MRR
R	MtransE [3]	0.68	2.72	5.02	580.3	0.025	0.40	1.80	3.30	651.1	0.017
	IPtransE [9]	13.69	32.47	42.32	143.4	0.231	11.24	26.31	34.62	181.4	0.191
	TransE	23.21	41.26	49.85	122.5	0.320	16.51	30.17	37.20	178.2	0.237
	SEA [10]	29.13	50.43	60.02	79.3	0.394	25.67	44.40	53.95	84.8	0.351
R+N	GCN [11]	6.73	16.80	23.14	357.0	0.123	4.48	11.61	16.53	420.9	0.087
	IMUSE [12]	35.14	57.03	66.13	63.4	0.455	29.65	48.47	56.74	73.8	0.388
R+N+V	MMEA	41.28	62.52	70.58	53.7	0.513	37.05	56.22	64.59	52.9	0.464
	proposed	**44.74**	**66.30**	**74.14**	**41.2**	**0.548**	**39.97**	**60.50**	**68.88**	**42.5**	**0.497**

Table 2: Evaluation results of MMEA and the proposed method when the training dataset is divided into 20%, 50%, and 80% of the total dataset.

Split	Model	FB-DB		FB-YAGO	
		H@10	MRR	H@10	MRR
20%	MMEA	70.58	0.513	64.59	0.464
	proposed	**74.14**	**0.548**	**68.88**	**0.497**
50%	MMEA	81.28	0.646	76.87	0.606
	proposed	**84.93**	**0.689**	**81.94**	**0.657**
80%	MMEA	89.11	0.749	86.80	0.711
	proposed	**91.10**	**0.776**	**89.22**	**0.750**

5. Conclusion

In this paper, we proposed a novel method of multi-modal KG embedding that uses uncertainty to evaluate the importance of supplementary information. The proposed method weights the information of each modality according to its importance and reduces the effects of missing or erroneous KGs. Experimental results show that the proposed method significantly outperformed the baseline model on two benchmark datasets for multi-modal entity alignment. This indicates that weighting important information is effective in information retrieval.

Acknowledgments

This study was partially supported by JST PRESTO (JPMJPR21C7), and JSPS KAKENHI (19H04172, 19K20344), Japan.

References

[1] K. Zeng et al., "A comprehensive survey of entity alignment for knowledge graphs," *AI Open*, vol. 2, pp. 1–13, 2021.

[2] A. Bordes et al., "Translating embeddings for modeling multi-relational data," in *Proc. NIPS*, 2013.

[3] M. Chen et al., "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment," in *Proc. IJCAI*, pp. 1511–1517, 2017.

[4] L. Chen et al., "MMEA: entity alignment for multi-modal knowledge graph," in *Proc. KSEM*, 2020.

[5] Y. Liu et al., D. Oñoro-Rubio, and D. S. Rosenblum, "MMKG: multi-modal knowledge graphs," in *Proc. ESWC*, 2019.

[6] R. Xie et al.,"Image-embodied knowledge representation learning," in *Proc. IJCAI*, 2017.

[7] P. Pezeshkpour et al., "Embedding multimodal relational data for knowledge base completion," in *Proc. EMNLP*, 2018.

[8] S. He et al., "Learning to represent knowledge graphs with gaussian embedding," in *Proc. CIKM*, pp. 623–632, 2015.

[9] H. Zhu et al., "Iterative entity alignment via joint knowledge embeddings," in *Proc. IJCAI*, pp. 4258–4264, 2017.

[10] S. Pei et al., "Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference," in *Proc. WWW*, pp. 3130–3136, 2019.

[11] Z. Wang et al., "Cross-lingual knowledge graph alignment via graph convolutional networks," in *Proc. EMNLP*, pp. 349–357, 2018.

[12] F. He et al., "Unsupervised entity alignment using attribute triples and relation triples," in *Proc. DASFAA*, pp. 367–382, 2019.

[13] Z. Sun et al., "A benchmarking study of embedding-based entity alignment for knowledge graphs," *Proc. VLDB Endow.*, pp. 2326–2340, 2020.

[14] G. Lample et al., "Word translation without parallel data," in *Proc. ICLR*, 2018.