Letter

Self-Supervised Entity Alignment Based on Multi-Modal Contrastive Learning

Bo Liu, Ruoyi Song, Yuejia Xiang, Junbo Du,
Weijian Ruan, and Jinhui Hu

Dear Editor,

This letter proposes an unsupervised entity alignment method, which realizes integration of multiple multi-modal knowledge graphs adaptively.

In recent years, Large-scale multi-modal knowledge graphs (LMKGs), containing text and image, have been widely applied in numerous knowledge-driven topics, such as question answering, entity linking, information extraction, reasoning and recommendation.

Since single-modal information contains unilateral knowledge, which makes LMKGs become more and more important. Considering that we can extract new facts from scratch, it is reasonable and practicable to align existing incomplete knowledge graphs (KGs) to complement each other. Entity alignment (EA) aims at aligning entities having the same real-world identities from different knowledge graphs. Among the studies of EA, there exist two main problems as follows: 1) Most existing EA methods [1], [2] only focus on utilizing textual information, in which the visual modality is yet to be explored for EA. 2) The previous works [1] rely heavily on the supervised signals provided by human labeling, which would cost a lot and may introduce inferior data in constructing LMKGs. As a result, the EA problem remains far from being solved. To demonstrate the benefit from injecting images and help our readers to understand the task of entity alignment, we present an example of "Times" and "泰晤士报" in Fig. 1. Without images, it is possible that "Times" and "clock" will have the similar embeddings. But with image embeddings, it will be more easily to align them.

To solve the above problems, we propose a novel self-supervised entity alignment method via multi-modal contrastive learning, namely SelfMEA, which embeds text and images [3] in a unified network. The purpose is to increase the accuracy of unsupervised learning and improve the accuracy of automatic entity alignment through contrastive learning and multi-modal method. Concretely, the framework of our method can be divided into two components as shown in Fig. 2, the first is the vectorization and representation of multi-modal knowledge graphs, and the second is the alignment of multi-lingual entities. The former one can be achieved by encoding the embeddings of graph structure, image in knowledge graphs and auxiliary information, and then integrate them to serve as a final embedding. For the latter one, we utilize the corresponding relationship of entities with the same meanings between two graphs through neighborhood component analysis, [4] and iterative learning, so as to realize the multi-language unsupervised entity alignment.

The main contributions of our work are as follows:

1) We propose a novel self-supervised entity alignment method via multi-modal contrastive learning, which is the first work to introduce

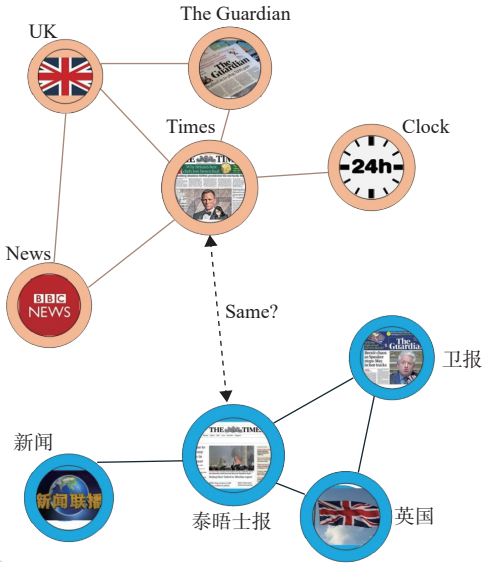

Fig. 1. An example of "Times" and "泰晤士报" of our EA task.

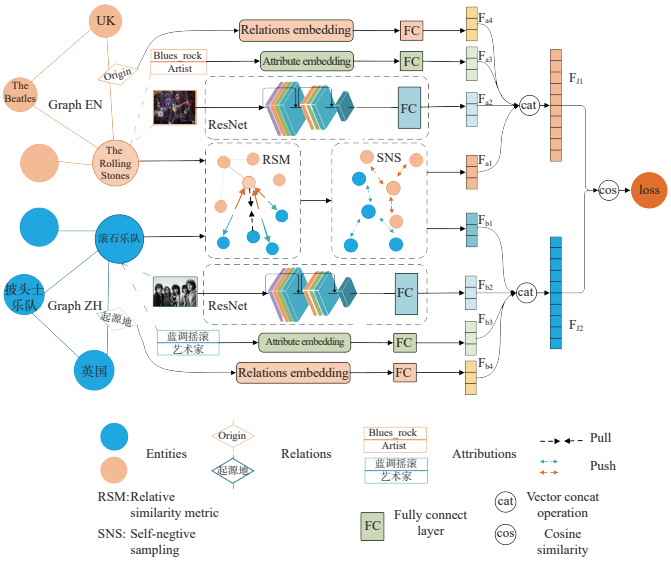

Fig. 2. The flow chart of our SelfMEA.

Corresponding author: Weijian Ruan.

Citation: B. Liu, R. Y. Song, Y. J. Xiang, J. B. Du, W. J. Ruan, and J. H. Hu, "Self-supervised entity alignment based on multi-modal contrastive learning," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 11, pp. 2031–2033, Nov. 2022.

B. Liu is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710000, and also with the China Electronics Technology Group Corporation (CETC) Key Laboratory of Smart City Model Simulation and Intelligent Technology, the Smart City Research Institute of CETC and National Center for Applied Mathematics Shenzhen (NCAMS), Shenzhen 518000, China (e-mail: liubo69@cetc.com.cn).

R. Y. Song is with the Chinese University of Hong Kong, Shenzhen 518000, China (e-mail: ruoyisong@link.cuhk.edu.cn).

Y. J. Xiang is with Tencent, Shenzhen 518000, China (e-mail: yuejiaxiang@tencent.com).

J. B. Du and J. H. Hu are with the CETC Key Laboratory of Smart City Model Simulation and Intelligent Technology, the Smart City Research Institute of CETC and National Center for Applied Mathematics Shenzhen (NCAMS), Shenzhen 518000, China (e-mail: dujunbo@cetc.com.cn; hujinhui@cetc.com.cn).

W. J. Ruan is with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, and also with the CETC Key Laboratory of Smart City Model Simulation and Intelligent Technology, the Smart City Research Institute of CETC and National Center for Applied Mathematics Shenzhen (NCAMS), Shenzhen 518000, China (e-mail: ruanweijian@cetc.com.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105962

contrastive learning into multi-modal knowledge for EA and achieve great performance.

2) The proposed method provides an effective insight for LMKGs construction, which can avoid expensive labelling cost and break the information unicity of single-modal knowledge graph.

3) Experiments on benchmark data sets DBP15k show that our SelfMEA achieves state-of-the-art performance on multi-modal knowledge graphs entity alignment.

Problem formulation: The task of entity alignment is to discover the unique equivalent entities in knowledge graphs, especially cross-language entities. Inspired by [2], we can generate a formula to encode entities into vectors and calculate the cosine similarity of all entities between two graphs, and the aligned entity is the one with the greatest similarity. The knowledge graph is defined as $G = \{E, R, T\}$, where E represents an entity, R represents a relationship, and T represents a triple. $T = \{E_1, R, E_2\}$, where E_1 is a head entity and E_2 is a tail entity. $S = \{(ei, ej) \mid ei \in E_1,\ ej \in E_2,\ ei \Leftrightarrow ej\}$ where $\Leftrightarrow$ represents equivalence. Graph structure is one of the most intuitive way to measure the cross language knowledge graphs. If the structures of two knowledge graphs are similar, that is, the distribution of the entities in $G_1 = \{E_1, R_1, T_1\}$ is similar to $G_2 = \{E_2, R_2, T_2\}$, then they have aligned entities to a large extend. In order to further improve the accuracy of the model, SelfMEA use not only graph structure embedding, but also image, attribute and relationship embeddings. Then, Neighbourhood component analysis loss is used for capturing the correspondence between counterpart entities from cross-lingual KGs.

As shown in Fig. 2 we firstly calculate entity embedding, relationship embedding, attribute embedding, and graph structure embedding, and fully connect them as the final embedding. Then we compute the the cosine similarity between the entities from two multilingual knowledge graphs for finding aligned entity. To explain the EA task, we give an example selected from DBP15K(ZH_EN): "The rolling stones" and "滚石乐队" are entities, which are the entities we need to align. Their relationships, attributes and graph structures are the raw materials used in the model to generate embeddings.

Graph structure embedding: Entity expressions of different knowledge graphs are distributed in different vector spaces. In this work, we use LaBSE, another most advanced multilingual pre-trained language model to map entity expressions to the same space. In the alignment task, [2] demonstrates that it is more effective to stay away from negative samples than to draw close to positive samples So, according to this discovery, we use contrastive learning in constructing graph structure embedding. When constructing this part, we use the absolute similarity metric (ASM) theorem [2]. Based on ASM, the relative similarity metric (RSM) is proposed. For fixed $\tau > 0$ and encoder f qualifications $\|f(\cdot)\| = 1$, we have

$$\begin{aligned}\mathcal{L}_{\mathrm{RSM}} &= -\frac{1}{\tau} + \mathop{\mathbb{E}}_{\{y_i^-\}_{i=1}^{M} \overset{\mathrm{iid}}{\sim} p_{\mathrm{y}}} \left[\log\left(e^{1/\tau} + \sum_i e^{f(x)^T f(y_i^-)/\tau}\right)\right] \\ &\le \mathcal{L}_{\mathrm{ASM}} \\ &\le \mathcal{L}_{\mathrm{RSM}} + \frac{1}{\tau}\left[1 - \min_{(kx,y)\sim p_{\mathrm{pos}}} \left(f(x)^T f(y)\right)\right]. \end{aligned} \tag{1}$$

In the process of contrastive learning, SelfMEA also uses the method of negative sampling, we sample negatives x_i^- from KG_x for a positive target $y_i^- \in \mathrm{KG}_{\mathrm{y}}$. This method can effectively avoid the error of accidentally deleting positive samples, and even if there were a small amount of repetition in KG_x or KG_{y}, it would not cause large error to the results [2].

Image embedding: As for image embedding, it is not advisable to directly use text-picture conversion to obtain images because of the polysemy problem, especially in Chinese. So we use images crawled from website. Then, We use ResNet-152 [5] to vectorize it. The dimension reduction model is like a fully connected layer without activation function.

$$F_I = W_I \times RES(I) + b_I \tag{2}$$

where RES stands for RESNET-152, W_I and b_I are parameters.

Attribute and relation embedding: Entity attributes and relationship also contain abundant information. Reference [6] found that spatially adjacent entities may interfere with each other, thus polluting entity representations in the GCN modeling process, which we don't want to see. Therefore, SelfMEA adopts a simple full connection layer to map relationship and attribute features to low-dimensional space, thus reducing the dimensions of entity attributes and relationship.

$$F_R = W_R \times R + b_R \tag{3}$$

$$F_A = W_A \times A + b_A \tag{4}$$

where R represents the relationship matrix, A represents the attribute matrix.

Final embedding: Finally, we need to integrate graph structure embedding, image embedding, attribute embedding and relationship embedding. Because the four embedding dimensions are different and their contributions to the model are not the same, we first unify the four embedding matrices at the row level, and then splice them together

$$\mathbf{F}_J = \mathop{\Big\oplus}_{i=1}^{n} \left[\frac{e^{w_i}}{\sum_{j=1}^{n} e^{w_j}} \cdot \mathbf{F}_i\right] \tag{5}$$

where F_J is the final embedding and F_i is the four embedding; n is the number of modalities; w_i is an attention weight for the i-th modality. In the process of training model, we also introduce a mechanism to automatically adjust the proportion of four embeddings.

Cosine similarity: SelfMEA uses cosine distance to measure the similarity between two final embeddings in inner product space. Let F_h and F_t denote the embeddings of entity E_h and target entity E_t. By computing their cosine distance matrix $S = \langle \mathbf{F}_J^s, \mathbf{F}_J^t \rangle$, we get each attribute's S_{ij}, the cosine-similar between the i-th entity in E_s and the j-th entity in E_t.

Loss function: After knowing how to measure the similarity of final embedding, we consider what kind of loss function to use. Referring to the neighborhood component analysis (NCA) based text image matching method proposed by [4], we adopt NCA loss. It used local and global statistics to measure the importance of samples, and uses soft weighting scheme to punish hard negative numbers. This is to alleviate the hubness problem in embedding space. The loss formula is as follows:

$$\begin{aligned}\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\Bigg(&\frac{1}{\alpha}\log\left(1 + \sum_{m\neq i} e^{\alpha \mathbf{S}_{mi}}\right) \\ &+ \frac{1}{\alpha}\log\left(1 + \sum_{n\neq i} e^{\alpha \mathbf{S}_{in}}\right) - \log\left(1 + \beta \mathbf{S}_{ii}\right)\Bigg)\end{aligned} \tag{6}$$

where α, β are temperature scales; N is the number of pivots within the mini-batch. We apply NCA loss to each mode respectively, select the parameters with reference to [3], and finally combine them to obtain global loss, which is recorded as

$$\mathcal{L}_{\mathrm{Joint}} = \sum_{i}^{n} \mathcal{L}_i \tag{7}$$

where $\mathcal{L}_i$ represents the loss term for aligning the i-th modality.

Iterative learning: Iterative learning control has been an active research area for more than a decade. Since SelfMEA is an unsupervised learning model, in order to improve the effect of the model without labels, we refer to [3] and adopt iterative learning strategy to propose more seeds from unaligned entities. Once epoch loops to a specific node, we will put forward a new round of suggestions and add each pair of intersection graph entities in the nearest neighbor to the candidate list. Therefore, the list of candidates will be updated in certain epochs, which yields a stable iteration greatly by enlarging seeds pool. The algorithm of obtaining the candidate list is described in Algorithm 1.

Experimental results: We conduct extensive experiments on DBP15k [7] dataset, as well as images obtained from Wikipedia, to evaluate the proposed SelfMEA. The DBP15k dataset contains three pairs of graph correspondences in different languages, namely, Chinese and English (zh_en), French and English (fr_en), and Japanese

Table 1. Comparison With the State-of-the-Art Methods and SelfMEA Results on DBP15k. "–" Means not Reported by the Original Paper. Underlined Bold Numbers are the Best Models

Model	FR_EN			JA_EN			ZH_EN		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
KECG (Li *et al*. 2019) [8]	0.486	0.851	0.61	0.49	0.844	0.61	0.478	0.835	0.598
HMAN (Yang *et al*. 2019) [6]	0.543	0.867	−	0.565	0.866	−	0.537	0.834	−
GCN-JE (Wu *et al*. 2019b) [9]	0.483	0.778	−	0.466	0.746	−	0.459	0.729	−
GMN (Xu *et al*. 2019) [10]	0.596	0.876	0.679	0.465	0.728	0.58	0.433	0.681	0.479
ALINET (Sun *et al*. 2020a) [11]	0.552	0.852	0.657	0.549	0.831	0.645	0.539	0.826	0.628
RKDEA (Li *et al*. 2022) [12]	0.622	0.912	0.721	0.597	0.881	0.698	0.603	0.872	0.703
EVA (Liu *et al*. 2021) [3]	0.715	0.936	0.795	0.716	0.926	0.792	0.72	0.925	0.793
SelfKG (Liu *et al*. 2021) [2]	0.957	0.992	−	0.816	0.913	−	0.745	0.866	−
SelfMEA	**0.963**	**0.992**	**0.975**	**0.861**	**0.939**	**0.889**	**0.790**	**0.916**	**0.833**

and English (ja_en). Each pair has 15K pairs of aligned entities, of which 60% are selected as the test set and the rest as the verification set. There are about 165k−222k relationships in each subdataset, and the three relational datasets have 2k−3k classes. The proportion of images in each data set is about 66%−78%. We conduct comparisons with 8 state-of-the-art methods on DBP15k dataset, as shown in Table 1. The evaluation results indicate that the proposed SelfMEA outperforms the other unsupervised classical alignment models, Specifically, our SelfMEA model leads to 4−5% absolute improvement in H@1 over the best baseline. This shows that the multi-modal method with contrastive learning can effectively improve the representation of cross language entities and infer their correspondence without additional supervision tags. There are obvious gaps between different subdatasets. For example, in ZH_EN, all the methods achieve the smallest Hit@1, Hit@10, and mean reciprocal rank (MRR), while achieve the middle in JA_EN, and the best in FR_EN. This phenomenon can be attributed to that there are more relationships between entities in the FR_EN dataset, and fewer types of relationships, thus the graph structure embeddings can better capture the information in FR_EN than ZH_EN and JA_EN. In addition, when obtaining graph structure embedding, we also tried to add pictures to the comparative learning model. As a result, the subtask was slightly improved, but the application was not improved on SelfMEA.

Algorithm 1 Iterative Learning

Input: Image embeddings of entities from two graphs F_1, F_2; new size n
Output: The new candidate list S
1: Similarity matrix $M = < F1, F2 >$;
2: Sort elements of M;
3: **While** $S\,! = n$ **do**
4 **if** $m.ri \notin R_u \& m.ci \notin C_u$ **then**
5 $S \leftarrow S \cup (m.ri, m.ci)$;
6 $R_u \leftarrow R_u \cup m.ci$;
7 $C_u \leftarrow C_u \cup m.ri$;
8 **return** new train list
9 **end**
10 **end**

Conclusion: We propose a multi-model self-supervised method with contrastive learning for entity alignment in this letter, which can break the information unicity of single-modal knowledge graph and avoid expensive labelling cost. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods. In the future, we plan to supplement the image dataset of DBP15K, and further improve the model accuracy through enhancing the interaction between information in different modality.

Acknowledgments: This work was supported by the National Key Research and Development Project (2019YFB2102500) and the National Nature Science Foundations of China (U20B2052).

References

[1] W. Zeng, X. Zhao, J. Tang, and X. Lin, "Collective embedding-based entity alignment via adaptive features," arXiv preprint arXiv: 1912.08404, 1912.

[2] X. Liu, H. Hong, X. Wang, Z. Chen, E. Kharlamov, Y. Dong, and J. Tang, "A self-supervised method for entity alignment," arXiv preprint arXiv: 2106.09395, 2021.

[3] F. Liu, M. Chen, D. Roth, and N. Collier, "Visual pivoting for (unsupervised) entity alignment," in *Proc. AAAI Conf. Artificial Intelligence*, 2021, pp. 4257–4266.

[4] F. Liu, R. Ye, X. Wang, and S. Li, "HAL: Improved text-image matching by mitigating visual semantic HUBs," in *Proc. AAAI Conf. Artificial Intelligence*, 2020, vol. 34, no. 7, pp. 11563−11571.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf Computer Vision and Pattern Recognition*, 2016, pp. 770−778.

[6] H.-W. Yang, Y. Zou, P. Shi, W. Lu, J. Lin, and X. Sun, "Aligning cross-lingual entities with multi-aspect information," arXiv preprint arXiv: 1910.06575, 2019.

[7] Z. Sun, W. Hu, and C. Li, "Cross-lingual entity alignment via joint attribute-preserving embedding," in *Proc Int. Semantic Web Conf*. Springer, 2017, pp. 628−644.

[8] C. Li, Y. Cao, L. Hou, J. Shi, J. Li, and T.-S. Chua, "Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model," in *Proc. Conf. Empirical Methods Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing,* 2019, pp. 2723– 2732.

[9] Y. Wu, X. Liu, Y. Feng, Z. Wang, and D. Zhao, "Jointly learning entity and relation representations for entity alignment," arXiv preprint arXiv: 1909.09317, 2019.

[10] K. Xu, L. Wang, M. Yu, Y. Feng, Y. Song, Z. Wang, and D. Yu, "Cross-lingual knowledge graph alignment via graph matching neural network," arXiv preprint arXiv: 1905.11605, 2019.

[11] Z. Sun, C. Wang, W. Hu, M. Chen, J. Dai, W. Zhang, and Y. Qu, "Knowledge graph alignment network with gated multi-hop neighborhood aggregation," in *Proc. AAAI Conf. Artificial Intelligence*, 2020, vol. 34, no. 1, pp. 222−229.

[12] X. H. Li, Y Zhang, and C. X. Xing. "Jointly learning knowledge embedding and neighborhood consensus with relational knowledge distillation for entity alignment." arXiv preprint arXiv: 2201.11249, 2022.