LETTER *Special Section on Data Engineering and Information Management*

Iterative Cross-Lingual Entity Alignment Based on TransC*

Shize KANG[†a)], **Lixin JI**[†], **Zhenglian LI**[††], **Xindi HAO**[††], ***Nonmembers*, *and* Yuehang DING**[†], ***Member***

SUMMARY The goal of cross-lingual entity alignment is to match entities from knowledge graph of different languages that represent the same object in the real world. Knowledge graphs of different languages can share the same ontology which we guess may be useful for entity alignment. To verify this idea, we propose a novel embedding model based on TransC. This model first adopts TransC and parameter sharing model to map all the entities and relations in knowledge graphs to a shared low-dimensional semantic space based on a set of aligned entities. Then, the model iteratively uses reinitialization and soft alignment strategy to perform entity alignment. The experimental results show that, compared with the benchmark algorithms, the proposed model can effectively fuse ontology information and achieve relatively better results.

key words: *cross-lingual entity alignment, ontology, knowledge embeddings, iterative alignment*

1. Introduction

Knowledge graph has been proposed to organize knowledge and applied to various intelligent applications. Knowledge graphs are often constructed from different languages and multilingual knowledge graphs are of great significance for the global sharing of knowledge. Thus, it is necessary to align entities with their counterparts from knowledge graphs of different language.

Recently, some cross-lingual entity alignment models based on embedding models have been proposed to connect knowledge graphs of different language versions. Most of these models only use the structural information of the knowledge graph, but there is some other information that can be utilized. For example, JAPE [1] fuses the attribute information of entities when performing cross-lingual entity alignment.

However, as far as we know, there is no related work about entity alignment based on ontology information. In knowledge graph, ontology defines the classes of entities and the relationships among these classes. Multilingual knowledge graphs such as Dbpedia first defines a unified ontology, and then knowledge graphs of different language versions are constructed according to the ontology. Figure 1 shows an example that constructs knowledge graphs of different language versions through a shared ontology. Therefore, ontology, as a kind of information that can be shared among different knowledge graphs, may be able to assist entity alignment tasks.

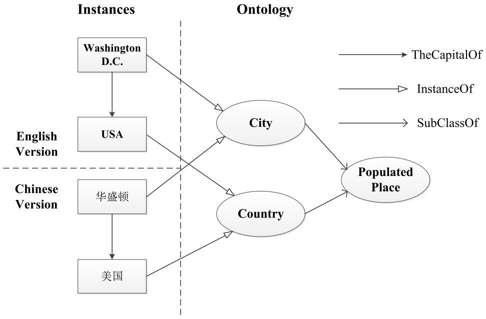

Fig. 1 An example of instances of different language versions and their corresponding ontology

To integrate ontology information, we propose a novel embedding model. We first introduce TransC [2] to encode all the entities and relations into knowledge embeddings. Meanwhile, we adopt parameter sharing model [3] to map knowledge embeddings of different languages into a shared space. In the end, we perform iterative alignment using reinitialization and soft alignment strategy. As far as we know, this is the first work that proposes a cross-lingual entity alignment model based on ontology. We construct a dataset and conduct extensive experiments on this dataset. The experimental results show the effectiveness of our model compared with the benchmark algorithms.

2. Problem Definition

A multilingual knowledge graph has two layers: the ontology layer and the instance layer. The ontology layer regulates the relationships among the classes of the entities in the instance layer. In this paper, we only choose one relationship in the ontology layer: "SubclassOf". C represents the set of classes in the ontology layer. The entities in the instance layer are connected to the classes in the ontology layer by relationship "InstanceOf". In the instance layer, each piece of knowledge can be represented as a triple (h, r, t), where h and t represent the head entity and tail entity respectively; r represents relationships among entities.

I represents a collection of languages. For each $i \in I$, the corresponding language-specific knowledge graph in in-

---

Manuscript received June 6, 2019.
Manuscript revised September 9, 2019.
Manuscript publicized January 9, 2020.

[†] The authors are with Information Engineering University, Zhengzhou, 450001, China.

[††] The authors are with Beijing Information and Communication Research Center, Beijing, 100000, China.

[*] This work is supported by the National Natural Science Foundation of China (61521003).

a) E-mail: xiaozebixia@163.com (Corresponding author)

DOI: 10.1587/transinf.2019DAL0001

Copyright © 2020 The Institute of Electronics, Information and Communication Engineers

stance layer can be represented as KG_i. I^2 represents an unordered combination between any two languages. For a pair of language combinations $(i_1, i_2) \in I^2$, $\mathrm{ILL}(i_1, i_2)$ denotes the aligned entity pairs between i_1 and i_2.

3. The Model

The proposed model can be divided into three parts: the knowledge embeddings part that trains knowledge embeddings using TransC; the joint embedding part that utilizes parameter sharing (PS) model to join knowledge embeddings of different languages into a shared semantic space; the alignment part that adopts reinitialization strategy (RE) and soft alignment (SA) strategy to perform entity alignment iteratively.

3.1 Knowledge Embeddings

We introduce TransC [2] to learn embeddings for the entities and relationships in a multilingual knowledge graph. In this work, all the entities, classes and relationships are mapped into the same embedding space. The learning can be divided into the following three part:

The triples in the instance layer. We use TransE as the basic knowledge embedding model for the triples in the instance layer, which transforms both entities and relationships into low-dimensional embedding space and assumes that the relationship is the translation from head entity to tail entity in embedding space, i.e. given any triple (h, r, t), it is expected to satisfy $\mathbf{t} \approx \mathbf{h} + \mathbf{r}$, and the corresponding energy function is:

$$f(h, r, t) = \|h + r - t\| \tag{1}$$

For any language $i \in I$, the corresponding score function is:

$$L_{G_i} = \sum_{(h,r,t)\in KG_i,(h',r',t')\in KG_i^-} [\gamma + f(h, r, t) - f(h', r', t')]_+ \tag{2}$$

where $[x]_+ = \max\{0, x\}$, KG_i^- is the negative sampling set of KG_i [1]. γ is the margin which is used to separate the positive triples and the negative triples. The total score function is:

$$L_G = L_{G_{i_1}} + L_{G_{i_2}} \tag{3}$$

InstanceOf triple embeddings. For a given triple (e, r_i, c), where $e \in E$ is an entity in the instance layer, $c \in C$ is a class in the ontology layer and r_i represents the relationship "InstanceOf". All triples of this type form a set S_i. TransC models the vector of each class c as a sphere $s(\mathbf{p}, m)$, where $\mathbf{p} \in \mathrm{R}^k$ is the sphere center and m is the radius. If an entity e in the instance layer is the instance of a class in the ontology, the vector $\boldsymbol{e}$ of e should be inside the sphere and the corresponding energy function is:

$$f_i(e, r_i, c) = \|\mathbf{e} - \mathbf{p}\| - m \tag{4}$$

For all the triples in the S_i, the corresponding score function is:

$$L_i = \sum_{\varepsilon\in S_i} \sum_{\varepsilon'\in S_i'} [\gamma_i + f_i(\varepsilon) - f_i(\varepsilon')]_+ \tag{5}$$

where S_i' is the negative sampling set of S_i.

SubClassOf triple embeddings. For a given triple (c_i, r_s, c_j) where c_i and c_j are two different classes in the ontology layer, and r_s represents the relationship "SubClassOf". All triples of this type form a set S_s. As c_i and c_j are two spheres in the embedding space, the sphere of c_i should be inside the sphere of c_j if c_i is the subclass of c_j. Thus the energy function for the triple (c_i, r_s, c_j) is:

$$\begin{cases} f_s(c_i, r_s, c_j) = d - m_j \\ d = \left\|\mathbf{p}_i - \mathbf{p}_j\right\|_2 \end{cases}, \tag{6}$$

where d is the distance between the centers of c_i and c_j. $\mathbf{p}_i$ and $\mathbf{p}_j$ are the sphere centers of c_i and c_j. m_j is the radius of c_i.

For all the triples in the S_s, the corresponding score function is:

$$L_s = \sum_{\varepsilon\in S_s} \sum_{\varepsilon'\in S_s'} [\gamma_i + f_s(\varepsilon) - f_s(\varepsilon')]_+ \tag{7}$$

where S_s' is the negative sampling set of S_s.

3.2 Joint Embeddings

To implement entity alignment, knowledge embeddings of different languages in the instance layer should be adjusted together in the training process. The method adopted in this paper is based on the entity pairs that have already been aligned (ILLs) and the parameter sharing model proposed by [5]. Parameter sharing model enables aligned entities to share the same knowledge embeddings, i.e. for any entity pair (e_1, e_2) of the aligned entity pairs $\mathrm{ILL}(i_1, i_2)$, we let the corresponding embeddings $\mathbf{e}_1$ and $\mathbf{e_2}$ of them be the same during the training process:

$$\mathbf{e}_1 \equiv \mathbf{e_2} \tag{8}$$

3.3 Alignment

In the training process of multilingual knowledge embeddings, the similarity of the knowledge embeddings of the entity pairs that can be aligned will gradually approach due to the adjustment of the parameter sharing model. In order to determine whether two entities will be aligned, a threshold σ_1 is set. When the similarity of the knowledge embeddings of the two entities exceeds the threshold, the two entities can be aligned. In this paper, the similarity of knowledge embeddings is measured using the cosine measure.

We use the newly aligned entities to find more aligned entities iteratively. However, this process can lead to error propagation problem as mentioned in [5]. In order to alleviate this problem, we propose two strategies. One is to reinitialize the knowledge embeddings when performing iterative alignment, and the other is the method of soft alignment.

Reinitialization Strategy In each iteration, the multilingual knowledge embeddings are first trained until the performance on the validate set is getting worse. At this time, in the unaligned entity pairs whose knowledge embedding similarity is greater than the threshold σ_1, the ones with the highest similarities is chosen as the newly aligned entity pair set ILL_{new}. After ILL_{new} is chosen, the knowledge embeddings are reinitialized to start a new iteration of training. In the new iteration of training, when the training of the knowledge embedding is finished, it needs to clear the ILL_{new} of the last iteration and regenerate the new ILL_{new}. By reinitializing the knowledge embeddings and ILL_{new} in each iteration, the propagation of error to the next iteration will be reduced.

Soft Alignment For any entity pair (e_1, e_2) in ILL_{new}, we define a score function according to the soft alignment method proposed by [3]:

$$\begin{cases} L_K = \sum\limits_{(e_1,e_2)\in\partial_{new}} \left(S(e_1,e_2) + S(e_2,e_1)\right) \\ S(e_1,e_2) = \sum\limits_{(e_1,r,t)} f(e_2,r,t) + \sum\limits_{(h,r,e_1)} f(h,r,e_2) \end{cases}, \tag{9}$$

For KG_{i_1} and KG_{i_2}, the entity pairs in ILL_{new} will not join the already aligned $\mathrm{ILL}(i_1, i_2)$ in the parameter sharing model but only use the soft alignment method to participate in the training, so the total objective function of the model in each iteration is:

$$L = L_G + L_K + L_s + L_i \tag{10}$$

4. Experiment

Dataset We construct a multilingual knowledge graph date set (MKG) based on Dbpedia, including En-Fr (English to French part) and En-De (English to German part). We first randomly sample 1000 entity pairs from the ILLs of Dbpedia (2016-10) for En-Fr and En-De respectively. We restrict that each entity in the entity pairs has appeared in the mapping based_objects files of Dbpedia at least 5 times. Based on the 1000 sampled entity pairs, the triples that match the entities in the 1000 entity pairs are selected from the mapping based_objects of each language. Finally, in these selected triples, we use Dbpedia's ILLs to match more aligned entity pairs (For example, the sampled 1000 pairs include A-B, and A-B matched (A, r1, C) and (B, r2, D). Although C-D is not in the sampled 1000 entity pairs, C-D is in the ILLs of Dbpedia), these entity pairs and the previously sampled 1000 entity pairs form the ILLs of this data set (ILLs (MKG)). Besides, we restrict that there is only one matched entity for each entity in the ILLs (MKG).

The selected triples mentioned above form the instance layer of MKG. For each InstanceOf triple (e, r_i, c) of Dbpedia, if e exists in the instance layer of MKG, this triple will be included in the InstanceOf set of MKG. For each triple (e, r_i, c) in the InstanceOf set of MKG, the class c will form the class set of MKG. For each (c_i, r_s, c_j) of Dbpedia, it is included in the ontology of MKG when c_i or c_j is in the class set of MKG.

Table 1 The statistics of the constructed dataset

DataSets	En-Fr		En-De	
	En	Fr	En	De
# Instance Entity	48735	19262	32914	27134
# Relation in Instance Layer	253	196	239	104
# Instance Triple	53851	21269	35268	30316
# Classes	462		457	
#InstanceOf Triple	67997		60048	
#SubclassOf Triple	405		402	
# ILL	6943		5124	

In the process of training, the ILLs of the data set are divided into training set, validation set and test set according to the proportion of 30%, 10% and 60%. The statistics about the data set are as follows:

4.1 Cross-Lingual Entity Alignment

We use two metrics to measure the performance of the proposed model. (1) Mean Rank: the average rank of the correct entities; (2) Hits@k: the proportion of the correct entities in the top k entities. Hits@1 and Hits@10 are used in this paper. Higher Hits@k and lower Mean indicate better experimental results.

We represent the proposed model as Ps-TransC (RE+SA). Besides, we also propose three variants of the proposed model:

Ps (RE+SA) In order to test the influences of the ontology information on the task, we proposed Ps (RE+SA) that only keeps the triples in the instance layer when training knowledge embeddings. The joint embedding and alignment part for Ps (RE+SA) is the same as the proposed model.

Ps-TransC (SA) In order to evaluate the effectiveness of adopting reinitialization strategy in the alignment part, we proposed Ps-TransC (SA) that abandons reinitialization strategy and only preserve soft alignment strategy in the training process. The joint embedding and knowledge embedding part for Ps-TransC (SA) is the same as the proposed model.

Ps-TransC (RE+HA) To verify the effectiveness of the proposed model's adoption of the soft alignment strategy, we proposed Ps-TransC (RE+HA) that replaces the soft alignment strategy in the proposed model with hard alignment strategy (HA) proposed by [3]. The other part of Ps-TransC (RE+HA) is the same as the proposed model.

For comparison, we also introduce LM (Linear Mapping) and MTransE [5] as baselines. For two aligned entities $e_1 \in \mathrm{KG}_{i_1}$ and $e_2 \in \mathrm{KG}_{i_2}$, both LM and MtransE learn a mapping from i_1 to i_2:

$$\|M\mathbf{e}_1 - \mathbf{e_2}\| \tag{11}$$

Table 2 The performance of the cross-lingual entity alignment models

Language	En-Fr			En-De		
Metric	Hits@1	Hits@10	Mean Rank	Hits@1	Hits@10	Mean Rank
LM	0.0356	0.1475	2331.3	0.0547	0.1726	2026.9
MTransE	0.1340	0.3274	885.7	0.1467	0.3358	881.2
Ps (RE+SA)	0.2634	0.4876	632.7	0.2952	0.4938	629.2
Ps-TransC (SA)	0.2875	0.5064	589.2	0.3172	0.5187	572.5
Ps-TransC (RE+HA)	0.2912	0.5160	562.9	0.3198	0.5271	544.8
The proposed	**0.3128**	**0.5263**	**549.4**	**0.3306**	**0.5386**	**527.1**

LM first trains knowledge embeddings, and then learns the mapping between these trained embeddings. However, MtransE trains the knowledge embeddings and mappings simultaneously. From Formula (11), we can see that the knowledge embeddings of different languages generated by LM and MtransE belong to different spaces. Since all embeddings generated by TransC are in a shared space, LM and MtransE can only be applied to instance layer in our experiment. Since our dataset doesn't include attribute information, we haven't choose JAPE as a baseline.

Setting In each iteration of training, the knowledge embeddings are initialized based on a truncated normal distribution, and the optimal dimension of the knowledge embedding is experimentally selected to be 75. In this paper, the AdaGrad algorithm [4] is used to optimize the objective function, and the normalization of the knowledge embedding is always maintained during the training process. We terminate the experiment using the early stop method.

Results Table 2 shows the performance of each models. Because LM and MTransE can be calculated in two directions, we show the average of the two results of different directions in the table.

As shown in the above Table, LM performs worst of all methods. The possible reason is that the knowledge embeddings of different languages are independently learned in the LM method, while other methods consider the association between different knowledge graphs in the process of training, so the knowledge embeddings of other methods have better correlation.

Although LM, MTransE and Ps (RE+SA) are all trained in the instance layer, Var1 outperforms LM and MTransE. The reason may be that Ps (RE+SA) has adopted iterative alignment which contributes to the performance of entity aligment.

Besides, the proposed model performs better than all the variants, indicating that all of the three factors (ontology information; the adoption of the reinitialization strategy and the adoption of the soft alignment strategy) can promote the performance of entity alignment especially the ontology information. For the iterative models in this paper, it will take at least 5 iterations for them to achieve the best results. As the TransE model is effective with less parameters compared with deep neural models, this cost is affordable on our dataset.

5. Conclusion

In this paper, we propose a novel embedding model to perform cross-lingual entity alignment integrating ontology information. We first introduce TransC and parameter sharing model to connect knowledge embeddings of different languages, and then propose two strategies to perform entity alignment. The experimental results show that, the proposed model can achieve a relatively better results compared with the baselines.

References

[1] Z. Sun, W. Hu, and C. Li, "Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding," The 16th International Semantic Web Conference, Vienna, Austria, vol.10587, pp.628–644, 2017.

[2] X. Lv, L. Hou, J. Li, and Z. Liu, "Differentiating Concepts and Instances for Knowledge Graph Embedding," Empirical Methods in Natural Language Processing (2018), Brussels, Belgium, pp.1971–1979, 2018.

[3] H. Zhu, R. Xie, Z. Liu, and M. Sun, "Iterative Entity Alignment via Joint Knowledge Embeddings," International Joint Conference on Artificial Intelligence (2017), Melbourne, Australia, pp.4258–4264, 2017.

[4] M.C. Mukkamala and M. Hein, "Variants of RMSProp and Adagrad with logarithmic regret bounds," International Conference on Machine Learning (2017), Cancun, Mexico, pp.2545–2553, 2017.

[5] M. Chen, Y. Tian, M. Yang, and C. Zaniolo, "Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment," International Joint Conference on Artificial Intelligence (2017), Melbourne, Australia, pp.1511–1517, 2017.