A Survey of Entity Alignment of Knowledge Graph Based on Embedded Representation

Jing Huang¹, Jiaqi Wang¹, Yahui Li¹, Wenbin Zhao¹,*

¹School of information science and technology, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, China
*email: zhaowb2013@stdu.edu.cn, zhaowb_email@163.com

Abstract. This paper summarizes the main methods of knowledge representation learning. Representation learning represents the entity information of the knowledge graph as a low dimensional vector. Its vector dimension is low, which helps to improve the computational efficiency and make full use of the semantic information between entities. In order to embed two KGs into a unified semantic space, joint embedding is used to achieve this goal. With the development of research, there are many improved embedding methods, such as iteration, multi view embedding, knowledge graph entity alignment based on graph neural network and so on.

1. Introduction

Entities refer to different objects that exist objectively and can be clearly distinguished in the real world, such as people, things or abstract concepts. Entity alignment refers to judging whether there are two entities in the heterogeneous knowledge graph. Their manifestations are different but they all point to the same object. It is also called entity resolution or entity matching. Entity alignment can help build a large knowledge graph, including multiple sources and languages. However, in the task of entity alignment, there are many challenges in cross language and cross source alignment. Any organization and individual can design and build a knowledge graph according to their own needs, and the data in it contains different relationships and structures. Aligning the entities in the knowledge graph can make it more convenient for people to acquire knowledge. Compared with the traditional representation, the computational complexity of embedded representation is lower in application. In addition, the similarity between entities and relationships can be obtained by measuring the similarity of low dimensional entity and relationship embedding. This paper summarizes the entity alignment of knowledge graph based on embedded representation, introduces the related concepts of entity alignment, the main methods of knowledge representation learning, the application and development of entity alignment based on embedded representation.

2. Methods based on embedding

2.1. Main methods of knowledge representation learning

The low dimensional vector representation obtained by representation learning is a distributed representation. The semantic information of the research object is represented as a dense low dimensional real valued vector through machine learning. The representation learning model mainly includes the following points:

(1) Distance model
SE[1] projects the head entity vector and tail entity vector into the corresponding space of the relationship, and then calculates the distance between the two projection vectors in the space. The distance model is very flexible and can adapt to many knowledge bases. It has strong generalization ability. However, the distance model has an important disadvantage: it uses two different matrices to project the head and tail entities into the space. It has poor synergy and can not well describe the semantic relationship between the two entities and relationships.

(2) Single layer neural network model

Single layer neural network model (SLM)[2] connects entity vectors through the nonlinearity of a single-layer neural network. To a certain extent, the single-layer neural network model alleviates the problem that distance model can not accurately describe the semantic relationship between entities and relationships. Although SLM further improves the SE model, the interaction between two entity vectors in the nonlinear operation of single-layer neural network model is weak and the computational complexity is higher.

(3) Implicit variable model

Implicit variable model (LFM)[3] is based on bilinear structure. The model can capture the order of data interaction and describe the second-order connection between entities and relationships. Compared with other previous models, the implicit variable model well describes the semantic relationship between entities and relationships. The model is probabilistic and takes into account the uncertainty, resulting in low computational complexity.

(4) Tensor neural network model

In the tensor neural network model (NTN)[2], the average value of all word vectors in the entity is taken as the entity vector. This can make the number of words in entities far less than the number of entities. When constructing entity representation, word vectors can be reused to solve the sparsity problem in entity representation learning and strengthen the semantic relationship between entities. Because the tensor operation is introduced into the tensor neural network model, the computational complexity is very high, and a large number of triples are needed in learning.

(5) Translation model

TransE[4] projects entities and relationships into a continuous low dimensional vector space, and interprets the relationship as a conversion operation between head entities and tail entities. The model is simple and effective, but it ignores the important multi-step path information in the knowledge graph, and has defects in dealing with complex relationships. TransH models the relationship as a hyperplane and performs translation operations on the hyperplane, which makes up for the deficiency of TransE in dealing with complex relationships to a certain extent, and the complexity of the model does not increase. TransR can model entities and relationships in different spaces, make the head entities and tail entities with actual relationships close to each other, and keep those entities without relationships away. Although TransR has some improvements compared with previous models, it still has limitations. Then, a TransD model is proposed. TransD not only considers the diversity of relationships, but also considers the diversity of entities.

2.2. Joint Embedding

If learning embedding on two KGs, they will be embedded in two different vector spaces independently. In order to evenly embed the two KGs in the vector space, some entities need to be aligned first, called seed entity alignment. Use KG1 and KG2 to represent two KGs, and the facts in the two KGs are represented by triple \((h,r,t)\), \(h\) represents the head entity, \(t\) represents the tail entity, and \(r\) represents the relationship. Different from previous embedding learning methods, joint embedding learning and the relationship between two KGs, specifically, first use a simple method to obtain some initial alignment entities, these alignments use some additional information or other metrics. As shown in Figure 1, the alignment entities of the same color are the seed entities. KG1 and KG2 are connected through the seed alignment entities, so that the joint embedding of the two KGs can be learned in a unified framework.
3. Development of alignment task model

3.1. Iteration
Although many scholars have done a lot of research on embedding models in the past few years, the KG embedding of alignment needs to be further explored. Secondly, entity alignment based on embedding often requires seed entity alignment, which is used as training. Data, the proportion of available prior comparisons is very small, and the accuracy is usually very low. Literature[5] introduced rule-oriented embedding (RUGE). In the learning process, RUGE obtains soft labels of unlabeled triples through iterative query rules to integrate and update the embedding model. Through this iterative process, the knowledge of logical rules can be better utilized in learning embedding. Due to the high degree of freedom of existing training data, many end-to-end learning methods are limited to rigid registration. Literature[6] solves this limitation by expressing non-rigid transformations as a point-by-point combination of several rigid transformations, using a recursive framework to solve it iteratively, and the difficulty of learning is greatly reduced. Literature[7] proposes to combine the entity name information not affected by the node degree with the structure information. The basic entity alignment framework of the model is shown in Figure 2. The iterative training method based on course learning is used to select high confidence results from easy to difficult to add to the training data to expand the scale of the labeled data.

3.2. Multi-view embedding
Although the existing embedding based entity alignment methods have achieved good results, they still have the following limitations: first, there are various types of entity features in KGs, but most of the recent embedding based entity alignment methods only include one or two types. Different types of features represent different aspects of entities. If they are used together, its robustness and accuracy will be improved; Secondly, the existing entity alignment methods based on embedding often need
seed alignment entities as training data. However, in practical applications, this kind of seed entity alignment is mostly inaccessible, and the cost of aligning seed entities is high. If you can learn embedding from different features, you will automatically get more alignment information and no longer rely too much on seed alignment. In view of the above limitations, literature[8] proposed an entity alignment framework based on multi-view KG embedding-MultiKE, which combines entity name, relationship and attribute views to learn entity alignment embedding, and adopts a variety of combination strategies. Literature[9] combined entity structure information and attribute information to jointly embed learning, and proposed a joint attention method in the model to jointly learn the attention of attribute types and attribute values.

3.3. Semi-supervised entity alignment method
The current reasons that affect the accuracy of entity alignment mainly include two aspects. One is that it is difficult to obtain the labeled data for initial seed alignment and the cost is high, and a large amount of unlabeled data cannot be used; The alignment of low-frequency entities presents challenges. Existing entity alignment methods require additional resources as a training set, such as entity or relationship information, so that such methods require a large amount of label data, which is usually difficult. In response to this problem, scholars have done a lot of research, but they are stuck on how to obtain more aligned seeds without considering a large number of unaligned entities. In this regard, a semi-supervised entity alignment method (SEA)[10] is proposed, and at the same time, it learns from aligned entity seeds and unaligned entities, which solves the problem caused by the degree difference of entities in different KGs. The recent embedding models are roughly divided into two categories: KG-based models and graph-based models. However, the KG-based method requires a sufficient number of seed entity pairs, which requires a high price. In the graph-based model, there is a problem of heterogeneity between different KGs, and the relationship type is not considered, and the entity alignment effect is not ideal. Literature[11] proposed a semi-supervised entity alignment method through knowledge embedding model and cross graph model (KECG). Specifically, for the knowledge embedding model, TransE is used to implicitly complete the consistency of the two KGs. For the cross-graph model, the graph attention network with projection constraints is extended to transfer structural knowledge, and the unimportant neighbors are ignored for alignment through the attention mechanism.

3.4. Unsupervised alignment
The setting of the alignment problem is that there are some pre-matched entities, but this setting is sometimes unsatisfactory, so many researchers are exploring how to unsupervised entity alignment. Many KGs contain a large number of attribute triples, but there is still a lack of research on the entity alignment of these attribute triples. The attribute similarity between two KGs helps attribute embedding to embed the two KGs into a unified semantic space. Reference[12] proposed a new embedding model. The model first obtains attribute embedding by using attribute triples, and then moves the entity embedding of two KGs to the same vector space through attribute embedding, so that the two KGs can obtain the similarity between entities through entity embedding. In natural language processing and computer vision, some studies have shown that bilingual dictionaries can be inferred by aligning word embeddings trained on monolingual data without supervised data. Many of the mappings between these two embeddings are based on antagonistic training. Literature[13] uses an alternative formula based on orthogonal matrix and permutation matrix to initialize the optimization algorithm through convex relaxation. This method is flexible, it only takes a few minutes to converge, and the computational complexity is low. In addition, literature[14] no longer maps each language to a language space, but maps it all to a common space, and proposes a new formula to ensure that the mapping can be combined, so as to achieve more Good alignment.

3.5. Entity alignment based on Neural Network
The research on graph based neural network is the focus of many scholars recently. Some scholars have proposed graph coding methods from different angles. This kind of methods generally use recursive aggregation method to encode the neighbor node information around the node, so that the
graph structure information can be used in the subsequent alignment process. However, such methods often only encode the surrounding first-order neighbors, and still can not overcome the problem of different graph structure caused by different relationship types in different knowledge bases. Recently, people have also proposed to deal with the graph matching task in the way of data dependence, or rely on graph similarity for graph matching. Literature[15] proposed a two-stage neural structure for learning and determining the structural correspondence between graphics. First, the local node embedding calculated by the graph neural network is used to obtain the soft correspondence between nodes. Second, the synchronous message passing network is used to achieve the matching agreement of the local neighborhood between the graph and the graph. The matching architecture is shown in Figure 3.

![Figure 3 Model matching architecture[15]](image)

4. Summary and outlook

4.1. Future research direction
In multivariate fusion, the embedded representation is still limited. The model usually considers either the fusion of text and knowledge base or the embedded representation model of entity description. Both information sources and fusion means are very limited. It is necessary to fuse other information of entities and relationships to improve the presentation ability. When facing different types of relationships, existing work divides relationships into four categories, including 1-1, 1-N, N-1, and N-N. When modeling different types of relationships, it is necessary to design special models and relationship types. The division of is not very accurate, and the type characteristics cannot be expressed intuitively. In the future, the latest research results of artificial intelligence and cognitive science can be combined to carry out research on the alignment of embedded representation entities for different complex relationship types.

4.2. Challenge
Although the embedded representation research has achieved good results, it still faces the following challenges:

1) The current existing models do not consider external knowledge information such as knowledge graphs, and rich external knowledge information has a certain impact on the alignment results. Therefore, how to properly consider external information such as knowledge graphs during alignment is an important challenge in the future.

2) The existing models can only be effectively trained for small-scale graphs, and there are hundreds of millions of nodes in the large-scale knowledge graph. How to solve the problems of storage, training efficiency, and heterogeneity faced in the alignment of large-scale knowledge graphs? The key challenges that need to be solved in actual application scenarios.

3) During the training process of the current model, in order to indicate the true structure of the network as closely as possible, the nodes are ignored to indicate subsequent applications. Therefore, how to explore more application scenarios of embedded representation and consider specific
subsequent application scenarios is an important challenge in the application of embedded representation.

**Acknowledgments**
This work was supported by National Natural Science Foundation of China (61373160), Natural Science Foundation of Hebei province (F2021210003), and Education Department Project of Hebei Province (QN2020197)

**References**

[1] Bordes A, Weston J and Collobert R 2011 *Aaai Conference on Artificial Intelligence Vancouver* British vol 1 pp 301-306

[2] Socher R, Chen D and Manning C D 2013 Reasoning With Neural Tensor Networks For Knowledge Base Completion *Proceedings of NIPS* 1 pp 926-934

[3] Jenatton R, Roux N L and Bordes A 2012 *International Conference on Neural Information Processing Systems* Doha Qatar vol 2 p 22

[4] Bordes A, Usunier N and Garcia-Duran A 2013 *Proceedings of the 26th International Conference on Neural Information Processing Systems* NSW Australia vol 2 pp 2787-2795

[5] Guo S, Wang Q, Wang L, Wang B and Guo L 2018 Knowledge graph embedding with iterative guidance from soft rules *Artificial Intelligence* 30 pp 4816–4823

[6] Feng W, Zhang J and Cai H 2021 Recurrent Multi-view Alignment Network for Unsupervised Surface Registration *Computer Vision and Pattern Recognition* 2 p 104

[7] Zeng W, Zhao X, Tang J Y, Tan Z and Wang W 2020 Iterative Entity Alignment via Re-Ranking *Journal of Computer Research and Development* 57 pp 1460-1471

[8] Xie Z, Zhu R and Zhao K 2020 A Contextual Alignment Enhanced Cross Graph Attention Network for Cross-lingual Entity Alignment *Computational Linguistics* 10 p 520

[9] Yang K, Liu S and Zhao J 2020 COTSAE: CO-Training of Structure and Attribute Embeddings for Entity Alignment *AAAI* 34 pp 3025-3032

[10] Pei S, Yu L and Hoehndorf R 2019 Semi-Supervised Entity Alignment via Knowledge Graph Embedding with Awareness of Degree Difference *AAAI* 13 pp 3130–3136

[11] Pei S C, Yu L and Zhang X L 2020 *The World Wide Web Conference* Ljubljana Slovenia vol 2 pp 3130–3136

[12] Risedya B D, Qi J and Zhang R 2019 Entity Alignment between Knowledge Graphs Using Attribute Embeddings *AAAI* 33 pp 297-304

[13] Grave E, Joulin A and Berthet Q 2018 Unsupervised Alignment of Embeddings with Wasserstein Procrustes *Machine Learning* 16 p 1805

[14] Alaux J, Grave E and Cuturi M 2019 Unsupervised Hyperalignment for Multilingual Word Embeddings *Computation and Language* 2 p 1811

[15] Fey M and Lenssen J E, Morris C 2020 Deep Graph Matching Consensus *Machine Learning* 27 p 2001