



A Chinese Dataset Building Method Based on Data Hierarchy and Balance Analysis in Knowledge Graph Completion

Yunwei Shi, Kuangrong Hao^(✉), Xuesong Tang, and Tong Wang

Donghua University, Shanghai, China
krhao@dhu.edu.cn

Abstract. This paper mainly studies the problem of knowledge graph completion. After analyzing the characteristics of the English dataset of knowledge graph, it is pointed out that many triples are meaningless. At the same time, it is found that there is no good construction method for the Chinese dataset. Then, based on two metrics, the method to create a Chinese dataset is proposed, and two Chinese data subsets are then created, which can well represent the special relations and the overall hierarchy of the dataset. Then, by selecting the existing model, it is concluded that the effect of the created Chinese datasets is better than the effect of the existing English datasets. At the same time, it is found that the uneven amount of data in the dataset may affect the accuracy of the model. Experiments results can conclude that when the amount of data is more uneven, the accuracy of the model will be worse.

Keywords: Knowledge graph completion · Chinese dataset building · ATTH

1 Introduction

Semantic web is made up of a lot of semantic data [1]. It provides semantic understanding of what people look up on the Internet and returns information they want to know, and knowledge graph is a technology to help semantic web understand. The concept of knowledge graph was proposed by Google Company. It is a structured semantic knowledge base that describes concepts and their interrelations through semantic information [2]. The basic unit of knowledge graph is the triple (head entity, relation entity, tail entity), and the entities are connected with each other through relations to form a networked knowledge structure. Through knowledge graph, web links can be transformed from web pages to conceptual links, so that users can accurately locate and acquire knowledge without browsing a lot of web pages.

Since the number of entities and relations in the knowledge graph is limited, there may be some entities and relations not in the knowledge base, which leads to the completion task of the knowledge graph. The main goal of the knowledge graph completion task is to predict the missing part of the triple and make the knowledge graph more complete. According to the specific prediction objects in the triple, the knowledge graph completion

task can be divided into three sub-tasks: head entity prediction, tail entity prediction and relation prediction. For the head entity prediction, the tail entities and their relations of the triples should be given. For tail entity prediction, we need to give the head entity and the relation of the triple, and then predict the entities that can form the correct triple. For relation prediction, it is given the header and tail entities, and then predicts the possible relation between the two entities.

In recent years, through the research of many scholars, the effect of knowledge graph completion has been continuously improved and developed. On the traditional model, [3] proposes a model to deal with multiple relation data. Based on the distributed vector representation of entities and relations, the TransE model ensures that the addition of the header and the relation vector is as equal as possible to the tail entity vector. [4] believes that different relations should focus on different attributes of entities and have different semantic spaces. [5] presents an idea to solve the heterogeneity and imbalance of entities and relations in knowledge graph by using sparse matrix instead of dense matrix in the original model. [6] thinks to use Gaussian distribution to represent entities and relations to deal with the uncertainty inherent in the semantics of relations and entities in the knowledge graph. [7] puts forward a new linear model to learn the knowledge graph representation, using the core tensor scale to measure the interaction level between the elements of the triple by using the mathematical Tucker decomposition. [8] adopts the method of mapping triples to complex vector space. On the neural network model, a new graph convolution network is proposed, which defines a different weight for different relations, and transforms a multi-graph into several single graphs with different strength relations [9, 10]. In [11], a new embedded model is proposed, which is used to model triple relations by capsule network. [12] uses the method of LSTM memory gate to relearn the relation between triples.

However, although the above scholars have made great contributions to the task of knowledge graph completion, they all focus on model improvement and innovation on the basis of the existing knowledge graph dataset, without paying much attention to the possible problems in the dataset itself. Although some articles believe that the existing English datasets have unreasonable problems in the data, such as FB15k [13] and WN18 [14], but none of them analyze Chinese datasets. Moreover, it is worth affirming that both English and Chinese datasets have the problem of meaningless data and unbalanced data quantity, and they all present a tree structure. In terms of semantic richness, Chinese datasets are more polysemy hierarchical than English datasets. Therefore, a more effective method to analyze and build Chinese datasets is urgently needed. In order to solve the above problems, this paper gives a method to optimize the Chinese dataset and builds two Chinese datasets. This method is more hierarchical than other methods to build English datasets, and can better reflect the special structure of different entities and relations, which is more convenient for the subsequent processing of model completion tasks. Then, this paper uses related model to test the effect of the established Chinese datasets and the existing English datasets.

This paper consists of the following parts. Section 2 analyzes some unreasonable problems in the existing English datasets and the methods given in related articles. Section 3 introduces the method of optimizing the Chinese dataset and how to construct a Chinese dataset in detail, and then introduces related model and reason. Section 4

shows the effects of different models on the constructed Chinese dataset, and analyzes the reasons for the good and bad effects and related verification. Section 5 summarizes the main conclusions and some work that can be carried out in the future.

2 Problem Analysis

Nowadays, there are many open-source English datasets for knowledge graph completion tasks, such as FB15k and WN18, they are extracted from the huge knowledge base Freebase and WordNet through related methods, respectively. They can represent a certain knowledge graph structural, and it is widely used. But this kind of dataset is not perfect and has two problems of meaningless triples and unbalanced data volume.

2.1 Existence of Meaningless Triples

Obviously, there are a large number of reverse triples and repeated triples in the open-source English dataset. The meaning of reverse triples is that there are two triples, and the semantics expressed by their relation are opposite. [15] finds that in the FB15k test set and WN18 test set, about 70% of the reverse triples of the triples exist in the training set. However, for these triples, when one of them appears in the test set, the other There is no need for triples to appear in the test set anymore, because their reverse relation can already be determined, so the existence of such a large number of reverse triples will cause the test set and the training set to have a high similarity. It will improve the accuracy of the model in disguise.

At the same time, there is a Cartesian product relation in the dataset. For a Cartesian product relation, the head entity-tail entity pairs in the involved triples constitute the corresponding Cartesian product. For example, if there are city entity a and month entity b widely in the dataset, then for the triples $(a, climate, b)$, each possible city a and month b are valid triples, because for each $climate$, different cities can be established in different months, then $climate$ is a Cartesian product relation.

For such a Cartesian product relation, the problem of complementing the knowledge map becomes predicting whether a city has its climate in a certain month. The existence of these relations will improve the accuracy of the model in disguise, and such prediction tasks are of little significance. Cartesian product relations and reverse relations are both artificially constructed and widely exist in the FB15k dataset and WN18 dataset. Therefore, it is necessary to artificially delete the head and tail entity pairs that exist in these relations. This problem will cause the existing performance indicators of the embedded model to be inconsistent with reality.

2.2 Unbalanced Data Volume

In addition to the widespread existence of meaningless triples, these datasets also have the problem of uneven data volume. [15] also finds through experiments that among the 37 relations in the YAGO3-10 dataset, there are two relations, *isAffiliatedto* and *playsFor*, whose triples account for 35% and 30% of all triples, respectively, which shows that the amount of triple data contained in the other 35 relations is very different from them.

At the same time, these two relations can be regarded as the same semantic relation in semantics, and there are also problems described before.

However, the work of [15] is aimed at open-source English datasets. Currently, there is no open-source Chinese dataset. There are only some general knowledge bases similar to WordNet, such as CN-DBpedia and zhishi.me. Similarly, there is no open-source Chinese dataset after the meaningless and unreasonable relation has been eliminated. Therefore, it is necessary to propose a method that can construct a certain rigorous Chinese dataset.

3 Methods

To build a Chinese dataset, we need to structure the dataset itself, and then refer to the method of constructing the English open-source dataset for analysis and summary.

3.1 Use Indicators to Measure Dataset Structure

We use two mathematical indicators to measure the overall and local structure of the knowledge graph dataset. The indicator that can reflect the overall structure of the dataset is ξ_G , and the formula is:

$$\xi_G = \frac{d(a, m)^2 + \frac{d(b,c)^2}{4} - \frac{d(a,b)^2 + d(a,c)^2}{2}}{2d(a, m)} \tag{1}$$

where a, b, c represent the three vertices of the triangle, m represents the midpoint of the shortest distance between the two vertices b and c of the triangle. $d(\cdot)$ indicates the shortest distance between two points.

The specific principle of this formula is to use the mathematical Toponogov theorem, which can show that the smaller the negative value, the better it reflects the structure of the tree. This theorem is the parallelogram law in Euclidean space, and the value is negative in spaces with negative curvature, such as hyperbolic space.

We define a parameter to represent the weight of different triples under the same relation, this parameter is:

$$k = \frac{e^{n_i}}{\sum_{i=1}^q e^{n_i}} \tag{2}$$

where q represents the number of all connected triples in the knowledge graph dataset, and n_i represents the number of nodes in the triple. We sample 2000 k triangles from these triples that contain this relation in the dataset. By calculating ξ_G of these 2000 k triangles, we use the average value $\overline{\xi_G}$ to represent the overall structure of the dataset under this relation.

$$\overline{\xi_G} = \frac{\sum \xi_G}{2000k} \tag{3}$$

For the entire knowledge graph dataset, we use a weighted average method to obtain the index value of the entire dataset. The weight is:

$$w = \frac{k_r}{\sum_{r=1}^p k_r} \quad (4)$$

where p represents the number of relations in the whole dataset.

Another indicator Khs_G is used to measure whether there is a reverse relation between entities and entities within the dataset. The formula is:

$$Khs_G = \frac{\sum R_{i,j}(1 - R_{j,i})}{\sum R_{i,j}} \quad (5)$$

where i, j represent the interconnected entities. For each relation, if there is only one edge from entity i to entity j , then

$$R_{i,j} = 1 \quad (6)$$

else

$$R_{i,j} = 0 \quad (7)$$

It can be found that if there is a set of reverse relations between every two entities, then $R_{i,j} = 0$, if every two entities are connected in a one-way relation, then $R_{i,j} = 1$.

3.2 Method of Constructing Chinese Dataset

With the above indicators as the criterion, we summarize several steps to build Chinese dataset.

Table 1. Five relations.

Coupling between triple relations	Examples
Symmetric	Marriage relation
Anti-symmetric	Dependency relation
Reversed	Inclusion relation
Combined	Family relation
Independent	Ubiquitous

Firstly, we make it clear that most of the relations in the dataset are coupled. This means that they follow some kind of logical connection. The coupling relations can be divided into four categories: symmetric, anti-symmetric, reversed and combined. These four types of relations need to be focused on and obtained from the dataset for subsequent

model analysis. For ease of representation, we add a class to represent relations without any coupling. These relations and corresponding examples are shown in Table 1. Then, we give the specific implementation steps. The overall step flow chart is shown in Fig. 1.

- 1) Find that the data library basically presents a tree structure by using $\bar{\xi}_G$,
- 2) Screen out entities with higher frequency,
- 3) Sort the results obtained by calculating the relation between the selected entities in ascending order by using $\bar{\xi}_G$ and Khs_G ,
- 4) Extract the first 100 –200 relations and all entities corresponding to these relations,
- 5) Extract most symmetric, anti-symmetric, reversed or combined relations and all related head and tail entities from these relations as the final data subset.

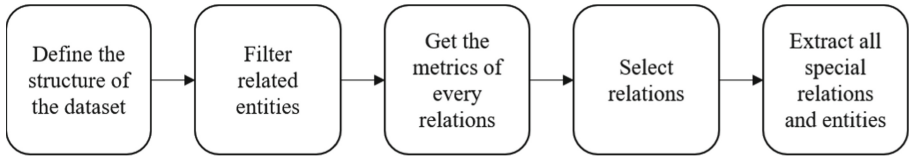


Fig. 1. Flow chart of dataset establishment steps.

3.3 Knowledge Graph Completion Model Selection

After establishing the basic steps of Chinese dataset construction, we need to use the model to verify whether the experimental effect of Chinese dataset constructed by us is better than that of English dataset.

We take the ATTH model proposed by [16] as our selected model into the task of knowledge graph completion. The ATTH model maps entities and relations in triples to hyperbolic-tangent space, and uses $Rot(\cdot)$ and $Ref(\cdot)$ to learn the coupling between different relations in triples. These two parameters were originally used to deal with the hyperbolic space to express the tangent formula, by constructing the diagonal matrix of sine and cosine, and then using the parameter matrix to map the tangent points, complete the mapping of space. $Rot(\cdot)$ can deal with anti-symmetric and combined relations, $Ref(\cdot)$ can deal with symmetric and reversed relations.

The formula related to the ATTH model is as follows.

$$\mathbf{q}_{Rot} = Rot(\cdot)\mathbf{x} \tag{8}$$

$$\mathbf{q}_{Ref} = Ref(\cdot)\mathbf{x} \tag{9}$$

where \mathbf{x} represents the header and tail entities in triples.

$$Att(\mathbf{q}_{Rot}, \mathbf{q}_{Ref}, \mathbf{a}) = \exp^c(\alpha_1\mathbf{q}_{Rot} + \alpha_2\mathbf{q}_{Ref}) \tag{10}$$

where \mathbf{a} represents an attention vector, α_1 is the result of softmax between \mathbf{a} and \mathbf{q}_{Rot} , and α_2 in the same way. $\exp^c(\cdot)$ represents the mathematical formula for the transformation from hyperbolic space to tangent space.

$$Q = Att(\mathbf{q}_{Rot}, \mathbf{q}_{Ref}, \mathbf{a}) \oplus \mathbf{r} \tag{11}$$

where \mathbf{r} represents the relations in triples, \oplus represents the Möbius addition. Finally, Q is put into the distance function and trained.

The reason why we choose ATTH model is that ATTH model is very sensitive to data with hierarchical structure, and it can well separate the coupling of different relations in knowledge graph dataset. Any other model is far less effective than it. The Chinese dataset we build is the dataset with this hierarchical structure, so we can observe the construction effect of Chinese dataset with this model.

4 Experiments

In this section, we need to use the model to verify whether the Chinese dataset we created is more efficient than the English dataset.

First of all, according to the method of creating Chinese dataset proposed previously, we select two general knowledge bases, one is CN-DBpedia [17], and the other is zhishi.me [18]. Then, we screen 30000 entities with high occurrence frequency, and then measure these triples according to the relation with ξ_G and Khs_G . And then we pick 200 of those relations in ascending order, then we find that in CN-DBpedia dataset, most of the 200 relations are coupled with each other, but in zhishi.me dataset, 130 of the 200 relations are uncoupled, we manually remove 54 uncoupled relations and retain 146 relations.

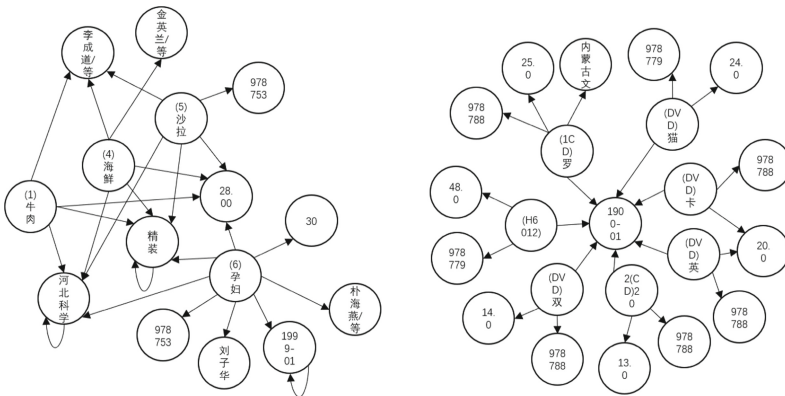


Fig. 2. The schematic diagram of the complexity of the internal relation of the Chinese dataset.

Finally, we constructed Chinese datasets CNDBp-200 and zhishi-146, the related information of the created Chinese datasets is shown in Table 2. The English datasets

Table 2. Details of datasets.

Dataset	#entities	#relations	#triples
CNDBp-200 (Chinese)	13k	200	103k
zhishi-146 (Chinese)	9k	146	85k
FB15k-237 (English)	15k	237	310k
WN18RR (English)	41k	11	93k

we selected are also shown in Table 2. The schematic diagram of CNDBp-200 dataset is shown in Fig. 2.

For the measurement index of model accuracy, we choose H@10 and MRR. H@10 checks whether the correct result of each triple in the test set is in the top ten of the sequence, and then calculate the probability that the correct result of all triples is in the top ten. MRR is an international general mechanism to evaluate search algorithms. When the rank of each result is obtained, the reciprocal of each result is calculated first and the average is calculated. The ATTH model test results are shown in Table 3.

Table 3. Test results of ATTH model on different datasets.

Dataset	MRR	H@10
CNDBp-200	0.291	0.483
zhishi-146	0.325	0.494
FB15k-237	0.307	0.49
WN18RR	0.266	0.447

From the results, we find that the ATTH model is the best on the zhishi-146 Chinese dataset and the worst on the WN18RR English dataset in H@10 index and MRR.

We notice that the effect of WN18RR is much worse than that of other datasets, which may be caused by the large difference in the amount of data between different relations in the dataset.

We also analyzed the proportion of the two relations with the largest difference in the number of relations in each dataset. The results are shown in Table 4.

Then we find that the proportion of the two groups with the largest difference in the number of relations in the WN18RR English dataset is very different from other datasets. Therefore, in order to prove that too much difference in the amount of data leads to poor model effect, we delete some triples with less relation data from our CNDBp-200 dataset to enlarge the difference, and finally maintain it at about 15:1.

Table 4. Comparison of the difference of relation data volume among different datasets.

Dataset	The proportion of the two relations with the biggest difference in quantity
CNDBp-200	3:1
FB15k-237	6:1
WN18RR	12:1

Then we use ATTH model to test again on the modified dataset, and the results are shown in Table 5.

Table 5. Comparison between the modified Chinese dataset and the original one.

Dataset	MRR	H@10
CNDBp-200	0.291	0.483
CNDBp-200(modified)	0.280	0.467

It can be apparently seen from the results that the problem of unbalanced data quantity will affect the accuracy of the model and make the accuracy worse.

5 Conclusions

This paper focuses on the completion of the knowledge graph. It first analyzes the characteristics of the English dataset of the knowledge graph, and then points out that the English dataset is unreasonable in terms of data. Many triples in the English dataset are meaningless. At the same time, it also found that although some of the problems in the English dataset have been improved by others, there is still no good way to improve the Chinese dataset. Then, on the basis of avoiding these existing problems, a method to create a Chinese dataset was given, and two Chinese data subsets were created. Then choose the ATTH model, and conclude that the effect of the Chinese dataset obtained by this method is better than that of the existing English dataset. This means that these Chinese datasets can have a better effect on the completion model. At the same time, it is found that the uneven amount of data in the dataset may affect the accuracy of the model. Through experiments it is concluded that when the amount of data is more uneven, the accuracy of the model will be worse.

The main tasks in the future are as follows. The method of dataset analysis and creation studied in this paper is relatively static, this means that adding some newly emerging entities, such as new network terms, to the existing knowledge graph dataset may have a significant impact on the completion task performance. If this method can be further improved in the future, the position and hierarchy of these newly introduced entities can be well reflected in the whole knowledge graph, and they can perform better in some intelligent scene dialogues and question answering systems. These are the contents that can continue to be studied in the future.

Acknowledgement. This work was supported in part by the Fundamental Research Funds for the Central Universities (2232021A-10, 2232021D-37), National Natural Science Foundation of China (61806051), and Natural Science Foundation of Shanghai (20ZR1400400, 21ZR1401700).

References

1. Hitzler, P.: A review of the semantic web field. *Commun. ACM* **64**(2), 76–83 (2021)
2. Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. *Semant. Web* **8**(3), 489–508 (2017)
3. Bordes, A., Usunier, N., Garcia-Duran, A., et al.: Translating embeddings for modeling multi-relational data. In: *Advances in Neural Information Processing Systems*, pp. 2787–2795 (2013)
4. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, February 2015
5. Ji, G., Liu, K., He, S., Zhao, J.: Knowledge graph completion with adaptive sparse transfer matrix. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*, pp. 985–991. AAAI Press (2016)
6. He, S., Liu, K., Ji, G., et al.: Learning to represent knowledge graphs with Gaussian embedding. In: *Proceedings of the CIKM*, pp. 623–632. ACM, New York (2015)
7. Balazevic, I., Allen, C., Hospedales, T.M.: TuckER: tensor factorization for knowledge graph completion. [arXiv:1901.09590](https://arxiv.org/abs/1901.09590) (2019). n. pag
8. Sun, Z., Deng, Z., Nie, J., Tang, J.: RotatE: knowledge graph embedding by relational rotation in complex space. [arXiv:1902.10197](https://arxiv.org/abs/1902.10197) (2019)
9. Shang, C., Tang, Y., Huang, J., Bi, J., He X., Zhou, B.: End-to-end structure-aware convolutional networks for knowledge base completion. [arXiv:1811.04441](https://arxiv.org/abs/1811.04441) (2019). n. pag
10. Nguyen, D.Q., Nguyen, T.D., Nguyen, D.Q., Phung, D.: A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv preprint* [arXiv:1712.02121](https://arxiv.org/abs/1712.02121) (2017)
11. Nguyen, D.Q., Vu, T., Dinh Nguyen, T., Phung, D.: A capsule network-based embedding model for search personalization. *arXiv e-prints*, arXiv-1804 (2018)
12. Santoro, A., et al.: Relational recurrent neural networks. [arXiv:1806.01822](https://arxiv.org/abs/1806.01822) (2018)
13. Cai, H., Zheng, V.W., Chang, K.C.C.: A comprehensive survey of graph embedding: problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.* **30**(9), 1616–1637 (2018)
14. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading wikipedia to answer open-domain questions. *arXiv preprint* [arXiv:1704.00051](https://arxiv.org/abs/1704.00051) (2017)
15. Akrami, F., Saeef, M.S., Zhang, Q., Hu, W., Li, C.: Realistic re-evaluation of knowledge graph completion methods: an experimental study. In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 1995–2010, June 2020
16. Chami, I., Wolf, A., Juan, D.C., Sala, F., Ravi, S., Ré, C.: Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint* [arXiv:2005.00545](https://arxiv.org/abs/2005.00545) (2020)
17. Xu, B., et al.: CN-DBpedia: a never-ending chinese knowledge extraction system. In: Benferhat, S., Tabia, K., Ali, M. (eds.) *IEA/AIE 2017. LNCS*, vol. 10351, pp. 428–438. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60045-1_44
18. Wu, T., Qi, G., Li, C., Wang, M.: A survey of techniques for constructing Chinese knowledge graphs and their applications. *Sustainability* **10**(9), 3245 (2018)