A knowledge graph completion model integrating entity description and network structure

Chuanming Yu†
School of Information and Safety Engineering
Zhongnan University of Economics and Law
Wuhan, China
yuchuanming2003@126.com

Zhengang Zhang
School of Information and Safety Engineering
Zhongnan University of Economics and Law
Wuhan, China
zhangzhengang2020@126.com

Lu An
Wuhan University
School of Information Management
Wuhan,China
anlu97@163.com

ABSTRACT

In recent years, knowledge graph completion has gained increasing research focus and shown great improvements. However, most existing models only uses the structures of knowledge graph triples when obtaining the entity and relationship representations, while the integration of the entity description and the knowledge graph network structure has been ignored. To address this issue, we propose a novel Entity Description Augmented Knowledge Graph Completion (EDA-KGC) model, which incorporates the entity description and network structure. The model is divided into three components, i.e., the representation component, the deep interaction component, and the inference component. The first component utilizes entity description to obtain the pre-training representation of entities. The deep interaction component acquires the features of the deep interaction between entities and relationships. The inference component performs matrix manipulations with the deep interaction feature vector and entity representation matrix, and thus obtains the probability distribution of tail entities. We conduct intensive experiments on the FB15K, WN18, FB15K237 and WN18RR data sets to validate the effect of the proposed model. The experiments demonstrate that the proposed model outperforms the traditional structure-based knowledge graph completion model and the entity-description-enhanced knowledge graph completion model. The research shows that the integration of entity description and network structure can significantly increase the effect of the knowledge graph completion task. The research has significant reference for completing the missing information in knowledge graph and improving the application effect of knowledge graph in the tasks of information retrieval, question answering and other fields.

CCS CONCEPTS

Computing methodologies • Artificial intelligence • Natural language processing • Knowledge base completion

KEYWORDS

Knowledge graph completion; Knowledge Entity Graph; Entity description; Deep learning; Pre-training model

ACM Reference format:

Chuanming Yu, Zhengang Zhang and Lu An. 2021. A knowledge graph completion model integrating entity description and network structure. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (workshop: EEKE'2021). ACM, Online, 5 pages.*

1 Introduction

This study aims to predict the missing information in the knowledge graph, that is, the prediction of tail entity under the premise of given head entity and relationship, or the prediction of head entity under the premise of given tail entity and relationship.

In recent years, knowledge graph completion has gained increasing research focus and shown great improvements. For instance, researchers have applied various knowledge graph representation learning models such as TransE[1] and TransH[2] to the task of knowledge graph completion. By mapping entities and relationships to a low-dimensional vector space, these models can retain the potential semantic information of entities and relationships, and achieve good results. It is worth noting that most knowledge graph representation learning models, which include the

EEKE'2021, September, 2021, Online

https://doi.org/**.****/*********

translation-based models [3-6], and the semantic-matching models [7-16], only use the triple structure in the knowledge graph when training entity and relationship representation, while ignoring many rich information from other sources such as entity description, attributes, etc. Although the triple contains the structural information and the semantic associations between entities and relationships, the potential semantic representation information learned by the knowledge graph representation model is limited due to the simple form of the triple.

Several models such as DKRL[17], EDGE [18], KG-BERT[19], and SDT[20] have attempted to introduce the entity description, attributes and other information into knowledge graph representation learning. The above models have achieved some improvements, but they are still facing the following challenges in practical applications. First, the models such as DKRL and EDGE extract the entity description information by convolutional neural networks (CNN) and long short-term memory (LSTM). The potential semantic information of entity description obtained is limited, which fails to significantly improve the performance of the knowledge base completion task. Second, for KG-BERT and other models that use the pre-training method, it may lead to very expensive model overhead as they score all possible triples during reasoning, which restricts the generalizability of the models.

To address the above issues, we propose a novel Entity Description Augmented Knowledge Graph Completion (EDA-KGC) model, and investigate the following research questions on this basis.

1) Compared with the traditional knowledge graph completion models and the ones that incorporate external information, does the proposed model have better effect in the knowledge graph completion task?

2) For the knowledge graph completion model integrating entity description and network structure, what are the key influencing factors?

2 Methodology

To investigate the above research question, we propose an entity description enhanced Knowledge graph completion model (EDA-KGC). The EDA-KGC model consists of three modules, i.e., a representation module, a deep interaction module, and an inference module, as shown in Figure 1. The division is based on the high cohesion and low coupling principle. That is, we keep subparts that are relevant to each other in a single component, and separate subparts that are unrelated in different components. Among them, the representation module takes the entity description as the input, and encodes it by employing the pre-training language model BERT. The deep interaction module first uses the relationship representation to obtain the convolution kernel corresponding to the relationship. Then, the convolutional kernel is employed to extract the features of the entity representation vector, as well as the features of the interaction between the entity and the relationship. Upon this, the deep interaction feature vector of the entity and the relationship is acquired through feature aggregation. The reasoning module performs matrix operation between the deep interaction feature vector and entity representation matrix to obtain the probability distribution of tail entities. Consider that the number of relationships in the knowledge graph is relatively small, the representation of relationships is initialized in a random manner.

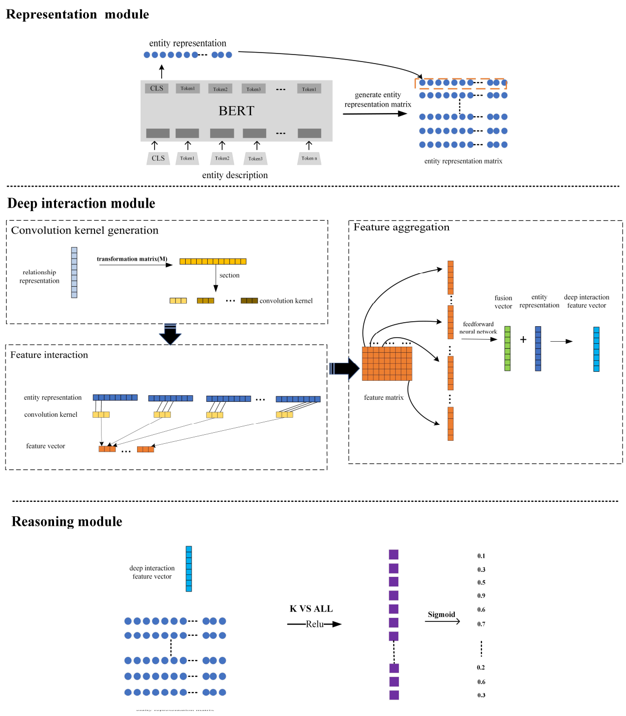

Figure 1: The architecture of the proposed EDA-KGC model

3 Experimental results

3.1 The comparative analysis

To investigate the research question, we conduct intensive experiments on four data sets, i.e., FB15K, WN18, FB15K-237, and WN18RR, which have been frequently used in the knowledge graph completion task. There are four evaluation indicators, namely Hits@1, Hits@3, Hits@10, and MRR (Mean Reciprocal Rank). The experimental results are shown in table 1-2.

In general, compared with the baseline models, the EDA-KGC model achieves the best results in respect to all the four indicators on the FB15K and FB15K237 datasets. On the WN18 dataset, EDA-KGC achieves the best results in Hits@10; On the WN18RR dataset, EDA-KGC achieves the best results in terms of MRR, Hits@1 and Hits@3.

Specifically, it can be seen from Table 1, on the FB15K data set, the proposed model performs significantly better than the 17 structure-based knowledge graph completion models. Additionally, the Hits@10 value of the proposed model is at least 7.9 points higher than that of the four text-enhanced models, i.e., DKRL, EDGE, SDT, and TKRL. On the WN18 data set, EDA-KGC achieves the highest Hits@10 value compared with all of the baseline methods.

It is worth mentioning that sometimes the performance of the proposed method is a little worse than state-of-the-art method. For instance, on the WN18 dataset, the MRR value of Tucker is 0.002 point higher than the proposed model. This may be due to the fact both two factors, i.e., network structure and entity description, could improve the total the performance, and TuckER leverages better the network structure information than the proposed model.

Table 1 The experimental results on the FB15K and WN18 datasets

	FB15K				WN18			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
AcrE	0.815	76.4	85.2	89.8	0.948	94.3	95.2	95.7
TransE	0.463	29.7	57.8	74.9	0.495	11.3	88.8	94.3
TransH	-	-	-	64.4	-	-	-	86.7
TransR	-	-	-	70.2	-	-	-	92.3
TransD	-	-	-	77.3				92.5
RotatE	0.797	74.6	83.0	88.4	0.949	94.4	95.2	95.9
DistMult	0.350	-	-	57.7	0.830	-	-	94.2
ComplEx	0.692	59.9	75.9	84.0	0.941	93.6	94.5	94.7
ConvE	0.657	55.8	72.3	83.1	0.942	93.5	94.7	95.5
HypER	0.790	73.4	82.9	88.5	0.951	94.7	**95.5**	95.8
TuckER	0.795	74.1	83.3	89.2	**0.953**	**94.9**	95.5	95.8
DKRL	-	-	-	67.4	-	-	-	89.1
EDGE	-	-	-	82.1	-	-	-	95.5
SDT	-	-	-	75.3	-	-	-	95.4
TKRL	-	-	-	73.4				93.4
EDA-KGC	**0.824**	**77.8**	**85.7**	**90.0**	0.951	94.4	95.3	**96.1**

3.2 Ablation Experiment

To investigate the influence of different model architectures on the effect of the EDA-KGC model, we conduct the following ablation experiment on the FB15K237 dataset. That is, we remove the residual connection (-Resid) and the entity bias score (-Bias) from the EDA-KGC model to compare the effect differences. As can be seen from Table 3, after removing the residual join and the entity bias score, the model effect decreases to different degrees, indicating that the residual connection and the entity bias score are effective in improving the performance of the knowledge graph completion task.

Table 2 The experimental results on the FB15K-237 and WN18RR data sets

	FB15K-237				WN18RR			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
AcrE	0.358	26.6	39.3	54.5	0.459	42.2	47.3	53.2
RotatE	0.338	24.1	37.5	53.3	0.476	42.8	49.2	57.1
ConvE	0.312	22.5	34.1	49.7	0.430	40.0	44.0	52.0
ConvR	0.350	26.1	38.5	52.8	0.475	44.3	48.9	53.7
HypER	0.341	25.2	37.6	52.0	0.465	43.6	47.7	52.2
TuckER	0.358	26.6	39.4	54.4	0.470	44.3	48.2	52.6
D4-STE	0.320	23.0	35.3	50.2	0.480	45.2	49.1	53.6
R-GCN	0.164	10.0	18.1	30.0	0.123	8.0	13.7	20.7
RNNLogic	0.344	25.2	38.0	53.0	0.483	44.6	49.7	55.8
MuIDE	0.328	23.7	-	51.5	0.481	43.3	-	**57.4**
KG-BERT	-	-	-	42.0	-	-	-	52.4
EDA-KGC	**0.362**	**27.0**	**39.6**	**55.1**	**0.493**	**45.2**	**50.4**	57.3

Table 3 The influence of different model architectures on the performance

	FB15K237			
	MRR	Hits@1	Hits@3	Hits@10
EDA-KGC	**0.362**	**27.0**	**39.6**	**55.1**
-Resid	0.358	26.7	39.1	54.6
-Bias	0.359	26.6	39.3	54.9

3.3 The influence of different text embeddings

To investigate the influence of different text representations on the performance, we replace the text representation generation component with random initialization, the traditional word embedding (GlovE), and the simplified version of BERT (ALBERT), respectively. Specifically, GlovE uses the average value of the GlovE vector of each word to obtain the sentence representation in respect to a given entity. Based on this, we compare the differences in respect to the performance on the knowledge graph completion task. It can be seen from Table 4 that, compared with the random initialization, both GlovE and ALBERT models have improved the performances to different degrees, indicating that the text representations can effectively improve the effect. In addition, among the three different text representations, BERT achieved the best result, followed by ALBERT and GlovE, indicating that different text representations have a certain influence on the performance of the EDA-KGC model. Specifically, the contextual pretrained models, such as BERT and ALBERT, are better than the traditional word embedding approach, i.e., GlovE. The experimental results are consistent with the performance of these two types of text representations in the natural language processing tasks, indicating that the effect of text representation learning is positively correlated with that of the knowledge graph completion tasks.

Table 4 The influence of different embeddings on the performance

	FB15K-237			
	MRR	Hits@1	Hits@3	Hits@10
Random initialization	0.324	23.6	35.5	50.4
GlovE	0.338	25.1	37.0	51.4
BERT	**0.362**	**27.0**	**39.6**	**55.1**
ALBERT	0.342	25.4	37.5	51.9

4 Discussions

Upon the above empirical and comparative studies, we discuss the research questions proposed in this study.

For RQ1, it is found that the proposed model significantly outperforms the structure-based knowledge graph completion models, i.e., TransE, TransH, TransR and TransD, etc., and the entity description enhancement methods, i.e., DKRL, EDGE, SDT, and TKRL, on most data sets. It is worth to be mentioned that, compared with the SOTA text-enhanced model KG-BERT, the experimental results suggest that the EDA-KGC model performs better in reasoning over the complex data sets.

For RQ2, it is suggested that different model architectures have certain influence on the performance of the EDA-KGC model. The residual connection and entity bias score employed in the model are effective in the knowledge base completion task. Additionally, the contextualized pre-training language models such as BERT and ALBERT are better than the traditional models, e.g. GlovE. The performance of vocabulary representation learning is positively correlated with that of the knowledge base completion task.

In respect to the efficiency, we use the K-v-ALL method to accelerate the reasoning process. That is, given a pair of head entity and relation, we score the pair against all entities simultaneously by conducting the matrix multiplication of the transpose of the interactive feature vector and the entity representation matrix. Compared with other models, which take a head entity, a tail entity, and a relation as a triple, and score each triple one by one, the K-v-ALL method could achieve a significant improvement in respect to the evaluation time. In addition, the method has been proved to be scalable to large knowledge base, as increasing the number of entities by 10 times only causes the computation time increase by 25% [9].

5 Conclusion

In this study, we propose a novel knowledge graph completion model (EDA-KGC) integrating entity description and network structure. By employing the contextualized pre-training language model BERT to encode the entity description information, the potential semantic knowledge of entity is obtained. Through the deep interactions between the entity and the relationship representation, the model solves the problem of limited entity relationship interaction in respect of traditional convolutional neural network. The intensive experiments on FB15K, WN18, FB15K237 and WN18RR shows that, compared with the existing structure-based methods and the entity description enhanced ones, the proposed model has achieved a significantly better performance. The research has important reference significance for completing the missing information in the knowledge graph and improving the application effect of the knowledge graph in the fields of information retrieval, question answering and so on.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (71974202, 71921002, 71790612, and 72174153), the project of the Ministry of Education of China (Grant No. 19YJC870029 and 17JZD034), and the Fundamental Research Funds for the Central Universities, Zhongnan University of Economics and Law (Grant No. 2722021AJ011).

REFERENCES

[1] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O, 2013. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, MIT Press, Cambridge, MA, 26:2787-2795.

[2] Wang, Z., Zhang, J., Feng, J., & Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 28, No. 1), AAAI Press, Menlo Park, CA, 1112-1119.

[3] Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In Twenty-ninth AAAI conference on artificial intelligence, AAAI Press, Menlo Park, CA, 2181-2187.

[4] Ji, G., He, S., Xu, L., Liu, K., & Zhao, J. 2015. Knowledge graph embedding via dynamic mapping matrix. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL Press, Stroudsburg, PA, 687-696.

[5] Sun, Z., Deng, Z. H., Nie, J. Y., & Tang, J. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197.

[6] Wang, K., Liu, Y., Ma, Q., & Sheng, Q. Z. 2021. MulDE: Multi-teacher Knowledge Distillation for Low-dimensional Knowledge Graph Embeddings. In Proceedings of the Web Conference 2021, Springer, Ljubljana, Slovenia, 1716-1726.

[7] Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.

[8] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. 2016. Complex embeddings for simple link prediction. In International conference on machine learning, ICML, New York, USA, 2071-2080.

[9] Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. 2018. Convolutional 2d knowledge graph embeddings. In Thirty-second AAAI conference on artificial intelligence, AAAI Press, Menlo Park, CA, 1811-1818.

[10] Balažević, I., Allen, C., & Hospedales, T. M. 2019. Hypernetwork knowledge graph embeddings. In International Conference on Artificial Neural Networks, Springer, Munich, Germany, 553-565.

[11] Balažević, I., Allen, C., & Hospedales, T. M. 2019. Tucker: Tensor factorization for knowledge graph completion. arXiv preprint arXiv:1901.09590.

[12] Ren, F., Li, J., Zhang, H., Liu, S., Li, B., Ming, R., & Bai, Y. 2020. Knowledge Graph Embedding with Atrous Convolution and Residual Learning. arXiv preprint arXiv:2010.12121.

[13] Jiang, X., Wang, Q., & Wang, B. 2019. Adaptive convolution for multi-relational learning. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 978-987.

[14] Xu, C., & Li, R. 2019. Relation embedding with dihedral group in knowledge graph. arXiv preprint arXiv:1906.00687.

[15] Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018, June). Modeling relational data with graph convolutional networks. In European semantic web conference. Springer, Cham, 593-607.

[16] Qu, M., Chen, J., Xhonneux, L. P., Bengio, Y., & Tang, J. 2020. Rnnlogic: Learning logic rules for reasoning on knowledge graphs. In Proceedings of the tenth International Conference on Learning Representations, Elsevier, Vienna, Austria, 1-11.

[17] Xie, R., Liu, Z., Jia, J., Luan, H., & Sun, M. 2016. Representation learning of knowledge graphs with entity descriptions. In Proceedings of the AAAI Conference on Artificial Intelligence , 2659-2665.

[18] Wang, S., & Jiang, C. 2019. Knowledge graph embedding with interactive guidance from entity descriptions. IEEE Access, 7, 156686-156693.

[19] Yao, L., Mao, C., & Luo, Y. 2019. KG-BERT: BERT for knowledge graph completion. arXiv preprint arXiv:1909.03193.

[20] Chen, X., Jia, S., Ding, L., Shen, H., & Xiang, Y. 2020. SDT: An integrated model for open-world knowledge graph reasoning. Expert Systems with Applications, 162:113889.