Knowledge Graph Augmentation with Entity Identification for Improving Knowledge Graph Completion Performance

Shuichi Chikatsuji(✉), Kenta Yamamoto, Ryu Takeda, and Kazunori Komatani

The Institute of Scientific and Industrial Research (SANKEN), Osaka University, Osaka, Japan
s-chikatsuji@ei.sanken.osaka-u.ac.jp

Abstract. A knowledge graph often lacks some existent triples. Knowledge graph completion is a technique for complementing such triples and its performance can be improved by augmenting triples from other external databases. However, entity names often differ between the original knowledge graph and an external database, which reduce the augmentation's efficiency. In this study, we identify the same entities that have different names (orthographic variants) that come from different sources, merge them into one entity, and augment the knowledge graphs. Our proposed method exploits in the original knowledge graph and the external database the similarity of triples, which were embedded using BERT. Experimental evaluation on our knowledge graph completion performance showed that our proposed method with graph information effectively outperformed two baselines.

Keywords: Knowledge graph · Knowledge graph completion · Orthographic variants

1 Introduction

Many studies have investigated knowledge graphs (KGs) as databases for dialogue systems [14,19–21,24]. A KG is represented as a set of triples (e_s, r, e_o) where e_s is a subject entity, r is a relation, and e_o is an object entity. The relations between two entities can be flexibly represented in KGs. On the other hand, it is basically impossible to represent every triple in the real world.

We can estimate the missing triples in KGs using knowledge graph completion (KGC) [2,8], which can be utilized to generate the response sentences of a dialogue system [7]. However, the more missing triples that exist, the lower is the KGC performance. To improve the KGC performance, KGs can be augmented using a different external database, as exemplified in Fig. 1. Increasing the number of relations per entity by augmenting KG will improve the KGC performance.

F. Liu et al. (Eds.): PRICAI 2023, LNAI 14325, pp. 480–486, 2024.
https://doi.org/10.1007/978-981-99-7019-3_43

A crucial problem in this augmentation is that entity names often differ between an existing KG and an external database. We call such different names having identical meanings *orthographic variants*. For example, “chocolate cake” is often abbreviated to “choco cake” in Japanese.

We identify entities whose meanings are identical and merge them, as shown in the “chocolate cake” example in Fig. 1. In our study, *entity identification* refers to associating two entities with the same meanings. If such entities are successfully merged, we can augment more relations between existing entities, which will improve the KGC performance.

Our proposed entity identification uses the similarity of feature vectors generated by BERT [5] by considering the graph information. We evaluated its effectiveness by the KGC performance obtained after augmenting a KG with entity identification.

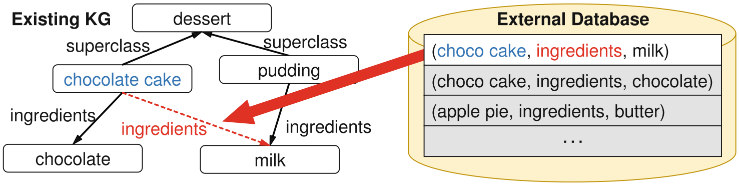

Fig. 1. Augmentation of KG using different databases

2 Related Work

Although some studies have addressed KG augmentation or construction, most did not take into account orthographic variants [1,4,9,22]. Meng et al. [10] constructed a KG from Chinese literature by merging orthographic variants using the Word2vec [11] model trained from the original literature. However, the KG and the external database considered in our study have no original literature to train a model.

Ikeda et al. [6] and Saito et al. [13] used language models to remove Japanese orthographic variants without KGs. Turson et al. [17] also studied a similar method for Uighur. Unlike our study, these works assume the availability of sufficient documents for training models.

Zhang et al. [23] and Sun et al. [15] input entity or triple information to language models to perform NLP tasks. However, both works assume that the original KG has enough relations between its entities.

3 Entity Identification Based on Graph Information

3.1 Augmentation with Entity Identification

Figure 2 shows the augmentation of a KG using entity identification, which is done on entities e_s and e_o in triples (e_s, r, e_o) of an external database used for

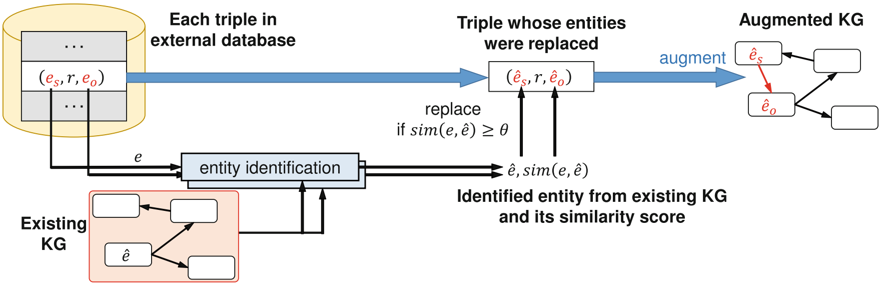
Each triple in external database
…
(e_s, r, e_o)
…
Triple whose entities were replaced
$(\hat{e}_s, r, \hat{e}_o)$
augment
Augmented KG
$\hat{e}_s$
$\hat{e}_o$
replace
if $sim(e,\hat{e}) \geq \theta$
e
entity identification
$\hat{e}, sim(e,\hat{e})$
Identified entity from existing KG and its similarity score
Existing KG
$\hat{e}$

Fig. 2. Augmentation details with entity identification

augmentation. The entity identification module outputs the most similar entities, $\hat{e}_s$ and $\hat{e}_o$, in the existing KG and their similarity scores. If the similarity scores are larger than or equal to threshold θ, e_s and e_o in the original triple are replaced with $\hat{e}_s$ and $\hat{e}_o$. Then the triple is augmented into the KG. We did not use triples that have unreplaced entities for augmentation because they may degrade the KGC performance.

3.2 Feature Vectors of Entities with BERT Considering Graph Information

Entity identification calculates the cosine similarity between the feature vectors of the entities in the KG and the external database. An entity with the largest cosine similarity in the KG is identified as the most similar. The feature vectors are computed with graph information using BERT.

We use the name of each entity and the triples containing it as input to BERT. Figure 3 shows an example. The triples are grouped by relations, and the sentences about them are connected by [SEP] tokens. When computing the feature vector for entity "chocolate cake," the input is "[CLS] chocolate cake [SEP] ingredients are egg and chocolate [SEP] superclass is dessert [SEP]" based on the graph structure. A [CLS] token is always used at the beginning of the BERT input.

Mean-pooling was applied to the sequence of output vectors from BERT. The pooled vector is a feature vector of each entity. In addition, we normalized each feature vector by subtracting the mean of all the feature vectors from it to improve the KGC performance after augmentation.

4 Experiments and Evaluations

4.1 Settings

We used a food subgraph from Wikidata [18] as the original KG. We extracted a portion of it and used it as test and validation data. The remaining graph after the extraction was used as the augmentation target. The target data had 14454

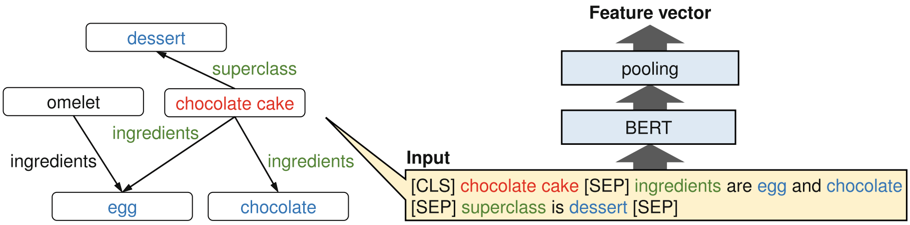

Fig. 3. Input format for BERT based on graph information

triples, the validation data had 242, and the test data had 243. They contained 8423 entities and 110 kinds of relations.

We used Rakuten Recipe from the Rakuten public data[1] for the external database. It has about 800,000 recipes. The entities to be augmented came from the names of dishes and the ingredients in the recipes.

We used TransE [3] and RotatE [16] as KGC models. Using the validation data, we set the embedding dimension to 300 for both models. For each triple (e_s, r, e_o) in the test data, we evaluated the performance of randomly predicting either e_s or e_o. Hits@$N (N = 1, 10)$ and mean reciprocal rank (MRR) were used as evaluation metrics. The BERT model was fine-tuned from a pre-trained model for Japanese[2]. Its hyperparameters are based on a previous paper [12]. Threshold θ (Fig. 2) was experimentally set to 0.4 using the validation data.

We set two baselines. One was "EditDist-based," in which similarity scores were computed by subtracting the normalized edit distance between the entity names from 1. The edit distance was normalized by the length of the longer entity name. Each entity name was treated as letters representing its Japanese pronunciation in this baseline. Entity pairs with similarity scores over 0.9 were regarded as identical. The other baseline was "BERT" without graphs, i.e., only each entity name was used to compute its feature vector by BERT.

4.2 Results and Discussion

Table 1 shows the KGC performance and the number of triples of the augmented KG for each method. Our proposed method is "BERT+graph."

Our BERT+graph method outperformed the other methods in every metric, especially the BERT baseline, and its number of triples decreased from the BERT baseline. This result indicates that the graph information reduced the triples that do not contribute to the KGC performance and positively impacted it.

Comparing the performance of each method, the increase from the BERT baseline to our BERT+graph method exceeded that from the EditDist-based baseline to the BERT baseline in all the metrics. This also confirms the effectiveness of graph information.

[1] https://rit.rakuten.com/data_release_ja/.

[2] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking.

Table 1. KGC performance and number of triples of augmented KG for each method

Method	TransE			RotatE			Number of triples
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	
No augmentation	0.014	0.072	0.035	0.010	0.072	0.032	14454
EditDist-based	0.072	0.327	0.157	0.121	0.377	0.209	55169
BERT	0.128	0.422	0.228	0.222	0.504	0.315	290475
BERT+graph	**0.191**	**0.531**	**0.302**	**0.383**	**0.724**	**0.497**	211168

Table 2 shows some examples of similarity scores? "Target entity" is an entity of Rakuten Recipe, and "Existing entity" is an entity of the food subgraph. We include the Japanese entity names and the pronunciations in parentheses. Our BERT+graph method successfully computed more appropriate similarity scores. For example, its similarity scores were high for similar pairs, such as 酢イカ (vinegared squid) and イカ (squid), and low for dissimilar pairs, such as 酢イカ (vinegared squid) and スイカ (watermelon) or かき揚げ (vegetable tempura) and カキ (oyster). On the other hand, even for a pair indicating exactly the same thing, such as そば(soba) and 蕎麦 (soba), the scores of another similar pair, such as そば(soba) and かけそば (kakesoba), were higher. Kakesoba is a kind of soba. While the KG augmentation improved KGC performance as demonstrated in Table 1, its negative effects were mitigated by preventing the erroneous merging of dissimilar pairs, such as 酢イカ (vinegared squid) and スイカ (watermelon).

Table 2. Examples of entity identification results

		BERT+graph	BERT	EditDist-based
酢イカ[suika]	スイカ[suika]	0.30	0.50	**1.00**
酢イカ[suika]	イカ[ika]	**0.66**	**0.70**	0.60
かき揚げ[kakiage]	カキ[kaki]	0.35	**0.71**	0.57
かき揚げ[kakiage]	から揚げ[karaage]	**0.59**	0.52	**0.71**
そば[soba]	かけそば[kakesoba]	**0.73**	0.62	0.50
そば[soba]	蕎麦[soba]	0.64	**0.79**	**1.00**

5 Conclusion

We augmented a KG with entity identification based on graph information and evaluated its KGC performance effectiveness after augmentation. Our experiment's results indicate that our proposed method outperformed the two baselines. In the future, we will verify whether our proposed method remains effective with another KG and external databases.

Acknowledgement. This work was partly supported by JST Moonshot R&D Grant Number JPMJPS2011 and JSPS KAKENHI Grant Numbers JP23H03457 and JP22H00536.

References

1. Al-Khatib, K., et al.: End-to-end argumentation knowledge graph construction. In: Proceedings of AAAI, vol. 34, pp. 7367–7374 (2020)
2. Bordes, A., et al.: Learning structured embeddings of knowledge bases. In: Proceedings of AAAI, vol. 25, pp. 301–306 (2011)
3. Bordes, A., et al.: Translating embeddings for modeling multi-relational data. In: Proceedings of NIPS, pp. 2787–2795 (2013)
4. Cannaviccio, M., et al.: Leveraging Wikipedia table schemas for knowledge graph augmentation. In: Proceedings of WebDB, pp. 1–6 (2018)
5. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
6. Ikeda, T., et al.: Japanese text normalization with encoder-decoder model. In: Proceedings of WNUT, pp. 129–137 (2016)
7. Komatani, K., et al.: Knowledge graph completion-based question selection for acquiring domain knowledge through dialogues. In: Proceedings of IUI, pp. 531–541 (2021)
8. Lao, N., et al.: Random walk inference and learning in a large scale knowledge base. In: Proceedings of EMNLP, pp. 529–539 (2011)
9. Luan, Y., et al.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Proceedings of EMNLP, pp. 3219–3232 (2018)
10. Meng, F., et al.: Creating knowledge graph of electric power equipment faults based on BERT-BiLSTM-CRF model. J. Electr. Eng. Technol. **17**(4), 2507–2516 (2022)
11. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS, pp. 3111–3119 (2013)
12. Reimers, N., et al.: Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP, pp. 3982–3992 (2019)
13. Saito, I., et al.: Improving neural text normalization with data augmentation at character-and morphological levels. In: Proceedings of IJCNLP, pp. 257–262 (2017)
14. Sarkar, R., et al.: Suggest me a movie for tonight: leveraging knowledge graphs for conversational recommendation. In: Proceedings of COLING, pp. 4179–4189 (2020)
15. Sun, T., et al.: CoLAKE: contextualized language and knowledge embedding. In: Proceedings of COLING, pp. 3660–3670 (2020)
16. Sun, Z., et al.: RotatE: knowledge graph embedding by relational rotation in complex space. In: Proceedings of ICLR, pp. 1–18 (2018)
17. Tursun, O., Cakıcı, R.: Noisy Uyghur text normalization. In: Proceedings of WNUT, pp. 85–93 (2017)
18. Vrandečić, D., et al.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
19. Xu, L., et al.: End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In: Proceedings of AAAI, vol. 33, pp. 7346–7353 (2019)
20. Yao, X., Van Durme, B.: Information extraction over structured data: question answering with freebase. In: Proceedings of ACL, vol. 1, pp. 956–966 (2014)

21. Yasunaga, M., et al.: QA-GNN: reasoning with language models and knowledge graphs for question answering. In: Proceedings of NAACL-HLT, pp. 535–546 (2021)
22. Yoo, S., Jeong, O.: Auto-growing knowledge graph-based intelligent chatbot using BERT. ICIC Express Lett. **14**(1), 67–73 (2020)
23. Zhang, Z., et al.: ERNIE: enhanced language representation with informative entities. In: Proceedings of ACL, pp. 1441–1451 (2019)
24. Zhou, H., et al.: Commonsense knowledge aware conversation generation with graph attention. In: Proceedings of IJCAI, pp. 4623–4629 (2018)