Effective Use of Knowledge Graphs in a Language Representation Model

Chanwook Min, Hyungsun Na, Yeji Song, Jinhyun Ahn, and Dong-Hyuk Im

Abstract Methods such as knowledge-enabled language representation model (K-BERT) that help train models using external information, such as knowledge graphs, have recently been proposed in the field of natural language processing. However, adding knowledge that does not match the topic when external information is used may hinder training. Hence, a method is required that adds only knowledge matching the topic of the input data by partitioning the external information by topic. Topic-based knowledge graph BERT (TK-BERT), which uses the existing latent Dirichlet allocation (LDA) model to partition the knowledge, does not consider the contextual information. To compensate for this drawback and effectively use external information, this study uses the BERTopic technique to partition the knowledge graph.

1 Introduction

With recent advances in deep learning technology, various studies are being conducted on natural language processing [1–3]. Areas of natural language processing include recommendation systems, question and answer tasks, language translation, and text generation. Among deep learning-based natural language processing models, transformer-based pretrained models, such as GPT and BERT, have achieved high performance in various natural language processing fields [4, 5]. The BERT model uses a large amount of text to generate a pretrained model, and the BERT model consists of two pretraining methods: (1) the masked language model (MLM) and (2) next sentence prediction (NSP). The MLM is a training method

C. Min · H. Na · Y. Song · D.-H. Im (✉)
Kwangwoon University, Seoul, South Korea
e-mail: dhim@kw.ac.kr

J. Ahn
Jeju National University, Jeju, South Korea

that masks a certain number of tokens and predicts the masked tokens, and NSP is a training method that takes two sentences as input and predicts the order of the sentences. The BERT model generates a pretrained model that has acquired general language knowledge based on these two pretraining methods. The pretrained model generated by the BERT model is used in various tasks through transfer learning. However, even if general language knowledge is acquired through pretraining, the pretrained model does not perform well in tasks that require specialized knowledge. To overcome this drawback, pretraining should be performed with data containing specialized knowledge about the corresponding field. However, this method requires considerable time and a large amount of specialized knowledge. Hence, applying this method in practice is difficult. Therefore, research has been conducted on using external data such as a knowledge graph to supplement insufficient knowledge about the input data. A knowledge graph is structured knowledge. It expresses the predicate representing the relationship between entities of the subject and object in a triple structure <subject, predicate, object> format. The K-BERT model in [6] employed a method that adds knowledge graph information, which is external data, to compensate for the BERT model's drawback of having poor performance in specialized tasks. This method improved the performance of the BERT model even in specialized tasks that were not general natural language tasks. However, because the knowledge graph has a lot of information, the K-BERT model uses information outside the topic of the input data, which may confuse the training of the model. To compensate for this drawback of the K-BERT model, [7] used the LDA technique, a topic modeling technique, in the knowledge graph. The LDA technique is a statistical technique for determining a topic in a document. This technique was used in the TK-BERT model to divide a vast knowledge graph into topics and infer the topic of the input data. Based on this, only knowledge matching the topic of the input data is added. Consequently, the TK-BERT model achieved a better performance than that of the K-BERT model. However, the LDA technique takes a document-term matrix (DTM) and term frequency-inverse document frequency (TF-IDF) as input. Hence, the LDA technique does not consider the order of the words. That is, the LDA technique does not consider the contextual information. Because the structure of a knowledge graph is constructed in the <subject, predicate, object> triple structure, the order of each word is important. To compensate for this drawback, this study proposes a method that uses a knowledge graph after dividing the knowledge graph into topics considering the contextual information of the knowledge graph using the BERTopic technique.

2 Knowledge Graphs

Knowledge graphs are data structurally constructed to represent the concept of knowledge. Each node of a knowledge graph consists of entities corresponding to the subjects and objects. In addition, the edge connecting each node is expressed as a predicate representing the relationship between subjects and objects. The constructed

graph can be expressed as a triple structure in the form of <subject, predicate, object>. The triple structure can express the relationship between entities clearly and concisely. This simple structure is suitable for adding insufficient knowledge about the input data in natural language processing. In the K-BERT model, which was investigated in a previous study, the lack of knowledge about input data in various fields of natural language processing was overcome using knowledge graphs. However, because the K-BERT model refers to a vast knowledge graph, it refers to knowledge outside the topic of the input data, as well as knowledge about the input data. This is called the knowledge noise problem. The knowledge noise problem is the phenomenon of confusing the training of models, and it occurs because too much knowledge is added from a knowledge graph. To prevent the occurrence of knowledge noise, Min et al. [7] used a knowledge graph by dividing it into topics using the LDA technique, which is a topic modeling technique. In this study, the BERTopic model is used to solve the LDA technique's problem, which is not considering the order of words. The BERTopic model shows a more effective use of a knowledge graph in natural language processing because it partitions the knowledge graph more appropriately according to the topics.

3 Topic Modeling

Topic modeling is a statistical model for estimating abstract topics that are inherent in a set of documents, and the LDA technique is a common topic modeling technique. The LDA technique begins by assuming that a document is composed of a mixture of topics, and the topics generate words based on the probability distribution. Based on this assumption, the LDA technique backtracks the process through which the document was created and infers the topic of the document and words. However, the LDA technique takes a DTM and TF-IDF as input. Hence, it uses the frequency of the words in the document but does not consider the order of the words. If the order of the words is not considered, the contextual information of each document cannot be known, making the identification of the exact topic difficult. To compensate for this problem, research was recently conducted on the BERTopic technique, which is a topic modeling technique that considers the contextual information [8]. The BERTopic technique uses BERT-based embedding and class-based TF-IDF. Figure 1 shows the structure of the BERTopic technique, and it largely consists of three stages. The first stage is the embedding stage, and the BERT model is used to perform embedding for each document. Here, the BERT model used in the embedding process is a pretrained model. The second stage uses UMAP to reduce the dimension of each document vector and performs clustering using HDBSCAN. Here, similar documents are clustered for each document vector. The third stage determines words that well represent each group using c-TF-IDF and makes adjustments using the maximize candidate relevance algorithm such that words representing each group are selected as diversely as possible. The BERTopic model assigns topics considering the contextual information through this process. To assign topics considering the context of the triple
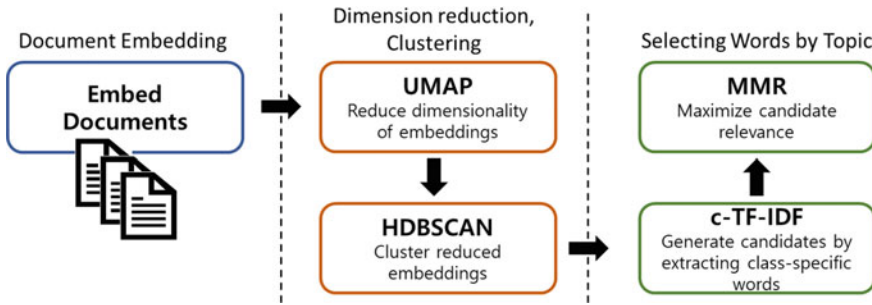
Fig. 1 Structure of the BERTopic model

structure of the knowledge graph, the BERTopic model was used to assign the topics of the knowledge graph, and the knowledge graph was divided and used based on the topics.

4 Method

Figure 2 shows the structure of the overall method of this study. Figure 2 consists of three stages: (1) generating the topic model and partitioning the knowledge graph, (2) inferring the topic of the input sentence, and (3) adding knowledge that matches the topic. In the first stage, a topic model is generated using the knowledge graph and BERTopic technique, and the generated topic model is used to partition the knowledge graph according to the topics. Here, the knowledge that matches the topic of the input data can be determined using the partitioned knowledge graph. In the second stage, the topic of the input data is inferred using the topic model generated in the first stage. Finally, in the third stage, the knowledge of the knowledge graph matching the inferred topic is added. The knowledge added in this stage refers only to the knowledge graph that matches the topic of the input data inferred in the second stage. That is, the BERTopic technique can be used to alleviate the knowledge noise phenomenon, which is a problem of the K-BERT model, and solve the existing LDA technique's drawback of not considering the context.

5 Experiment

In this study, an $F1$-score comparison experiment was conducted to compare the performances of the K-BERT model, which provides insufficient knowledge through the knowledge graph, the TK-BERT model, which uses the LDA technique to partition the knowledge graph by topic, and our model, which uses the BERTopic technique. Two knowledge graphs, Cn-DBpedia and HowNet, were used to train the
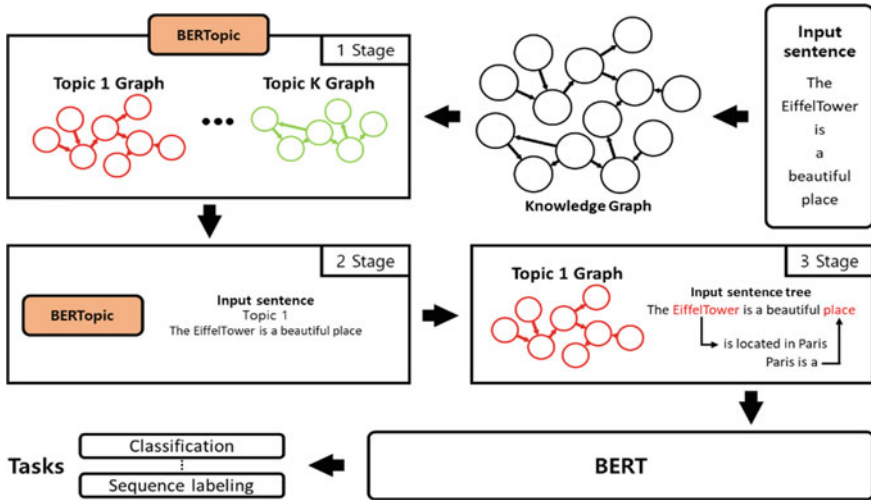
Fig. 2 Structure of the method

K-BERT model. The TK-BERT model was trained using the two knowledge graphs after partitioning them using the LDA technique. Finally, the knowledge graphs were partitioned by topic using the BERTopic technique, and the partitioned knowledge graphs were used to train our model.

5.1 Experiment Environment

Google BERT was used as the pretrained model of BERT in the experiment. This model was pretrained with WikiZh data, which is Chinese Wikipedia data composed of 12 million sentences. In addition, two knowledge graphs, Cn-DBpedia and HowNet, were used in the experiment. The Cn-DBpedia knowledge graph consists of approximately 5.16 million triple data points. The HowNet knowledge graph concerns the Chinese lexicon and consists of approximately 52,000 triple data points. The TopicModel was constructed using the two knowledge graphs and BERTopic technique. The pretrained model of the BERTopic model used models trained with more than 50 languages. Because the knowledge graphs were partitioned by topic using the topic model, the two knowledge graphs were combined and then partitioned by topic. Here, the number of topics was set to 50 because 50 achieved the best performance in an experiment conducted using 50, 100, and 150 as the numbers of topics. In addition, the Book_review, Chnsenticorp, and Shopping datasets were used. The datasets consist of positive and negative reviews. The Book_review dataset consists of 20,000 positive b and 20,000 negative book reviews. The Chnsenticorp dataset is hotel review data, and it consists of 6000 positive and 6000 negative hotel

Table 1 Results of the K-BERT model, TK-BERT model, and our model

Model	Dataset					
	Book_review		Chnsenticorp		Shopping	
	Dev	Test	Dev	Test	Dev	Test
K-BERT (HowNet)	86.8	85.85	94.5	94.6	96.5	96.05
K-BERT (Cn-DBpedia)	87.4	86.25	93.95	94.5	96.45	96.3
TK-BERT (Topic-based HowNet)	87.35	86.45	94.5	95.6	96.75	96.45
TK-BERT (Topic-based Cn-DBpedia)	88.2	86.9	94.5	95.5	96.5	96.35
Our model	**88.55**	**87.1**	**94.5**	**95.65**	96.75	**96.65**

Bold represents high performance

reviews. Finally, the Shopping dataset contains online shopping review data, and it consists of approximately 21,000 positive and 19,000 negative shopping reviews.

5.2 Results

Table 1 compares the $F1$-scores of the K-BERT model, TK-BERT model, and our model; our model performed the best on each dataset. The K-BERT model adds knowledge using the knowledge graph, and the TK-BERT model uses the knowledge graph by partitioning it using the LDA technique. OurModel uses the knowledge graph by partitioning it using the BERTopic model to compensate for the LDA technique's drawback of not considering the contextual information. The results show that when the K-BERT model, which adds knowledge using the knowledge graph, is used; information that benefits learning can be added if the knowledge graph is used after partitioning it by topic. In addition, when the BERTopic technique is used to partition the knowledge graph, the knowledge graph can be divided into topics more effectively than when using the LDA technique to partition the knowledge graph because the topic model is generated considering the contextual information.

6 Conclusion

In this study, we proposed a method for a more effectively implementation of the TK-BERT model, addressing the problem of the existing K-BERT model. The topic model of the TK-BERT model uses the LDA technique. The LDA technique estimates a topic using only the frequency of the words in a document. Hence, it cannot reflect the contextual information. To address this problem, the BERTopic technique was used to implement the topic model because this technique can consider the contextual information. Because the BERTopic model considers the contextual information through document embedding, it can use the knowledge graph by partitioning it

more effectively. The experiment verifies that our model outperformed the TK-BERT model, which uses the existing LDA model. The results indicate that our model can effectively partition even a larger knowledge graph for training.

References

1. Maulud DH et al (2021) State of art for semantic analysis of natural language processing. Qubahan Acad J 1(2):21–28
2. Zhang Z et al (2020) Semantics-aware BERT for language understanding. Proc AAAI Conf Artif Intell 34(05)
3. Xu W et al (2021) Enabling language representation with knowledge graph and structured semantic information. In: 2021 international conference on computer communication and artificial intelligence (CCAI). IEEE
4. Radford A et al (2018) Improving language understanding by generative pre-training
5. Devlin J et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
6. Liu W et al (2020) K-BERT: enabling language representation with knowledge graph. Proc AAAI Conf Artif Intell 34(03)
7. Min C, Ahn J et al (2023) TK-BERT: effective model of language representation using topic-based knowledge graphs. In: International conference on ubiquitous information management and communication (IMCOM) (accepted)
8. Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794