TK-BERT: Effective Model of Language Representation using Topic-based Knowledge Graphs

2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM) | 978-1-6654-5348-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IMCOM56909.2023.10035573

Chanwook Min
Department of Artificial Intelligence Convergence
Kwangwoon University
Seoul, South Korea
a4073631@kw.ac.kr

Jinhyun Ahn
Department of Management Information Systems
Jeju National University
Jeju, South Korea
jha@jejunu.ac.kr

Taewhi Lee
Smart Data Research Section
Electronics and Telecommunications Research Institute
Daejeon, South Korea
taewhi@etri.re.kr

Dong-Hyuk Im
School of Information Convergence
Kwangwoon University
Seoul, South Korea
dhim@kw.ac.kr

***Abstract*— Recently, the K-BERT model was proposed to add knowledge for language representation in specialized fields. The K-BERT model uses a knowledge graph to perform transfer learning on the pre-trained BERT model. However, the K-BERT model adds the knowledge that exists in the knowledge graph rather than the data relevant to the topic of the input data when using the knowledge graph of the corresponding field. Hence, the K-BERT model can cause confusion in the training. To solve this problem, this study proposes a topic-based knowledge graph BERT (TK-BERT) model, which uses the topic modeling technique. The TK-BERT model divides the knowledge graph by topic using the knowledge graph's topic model and infers the topic for the input sentence, adding only knowledge relevant to the topic. Therefore, the TK-BERT model does not add unnecessary knowledge to the knowledge graph. Moreover, the proposed TK-BERT model outperforms the K-BERT model.**

Keywords*— *Knowledge Graphs, Language Representation, BERT, K-BERT, Topic Modeling

I. INTRODUCTION

Pre-trained models, such as the BERT model, have shown good performances in various fields of natural language processing [1]. Such models are pre-trained to acquire general language knowledge and then customized for the user's task using transfer learning. The pre-training of the BERT model acquires general language knowledge through two approaches: One is the masked language model (MLM), which masks tokens with large scale documents and prediction, which predicts and learns the next sentence of a predicts the masked tokens, and the other is the next sentence prediction (NSP), which predicts and learns the next sentence of a sentence. Transfer learning is used to tune the pre-trained model according to the user's task. However, even if general language knowledge is learned through pre-learning, pre-trained models have the disadvantage of not performing well in specialized fields that require deep knowledge. There are two methods to solve this problem: 1) train the model with a lot of data from the corresponding field over a long time during pre-training and 2) add knowledge using a knowledge graph [2]. The K-BERT model that uses a knowledge graph for training has been proposed [3]. This K-BERT model adds knowledge about the input sentence using the knowledge graph when performing transfer learning on the pre-trained BERT model. In this process, Knowledge Noise(KN) is generated. KN is a phenomenon in which the meaning of a sentence changes differently due to the addition of too much knowledge. This is caused by adding a large amount of knowledge. The K-BERT model uses a visible matrix to prevent KN. A visible matrix is a matrix that expresses the relationship between the added knowledge in each token, indicating the relationship between the token in the input sentence and the added knowledge. However, even with the visible matrix, it is possible to induce learning that is inconsistent with the original intention because the entire large-scale knowledge graph is referred to. Therefore, the knowledge that deviates from the input sentence's topic may be added.

In this study, we propose a topic-based knowledge graphs BERT (TK-BERT) model, which divides the knowledge graph using topic modeling and adds knowledge relevant to the input sentence's topic. Some previous works incorporate the topic modeling technique using a knowledge graph [4]. We propose a method for applying the knowledge graph that has been divided into topics using the topic modeling technique to the BERT model for the first time. The main contributions of this study are as follows:

- We divide the knowledge graph by topic using the topic modeling technique so that only knowledge relevant to the input sentence's topic was added. Thus, the knowledge noise problem can be suppressed.
- Our TK-BERT model can use a large-scale knowledge graph effectively because it divides the large-scale knowledge graph by topic when using the knowledge graph. The TK-BERT model can also use the knowledge graph efficiently because the hash table size and memory allocation are reduced by dividing the large-scale knowledge graph.

978-1-6654-5348-6/23/$31.00 ©2023 IEEE

- Through the experimental results on various datasets and two knowledge graphs, we demonstrate that the TK-BERT model outperforms the K-BERT model.

II. KNOWLEDGE GRAPHS

Fig.1 shows an example of structured knowledge, which can be expressed as a predicate indicating the relationship between entities corresponding to the subject and objects. In other words, the relationship between entities can be expressed as a triple structure of the <subject, predicate, object> format. The triple structured information can clearly express the relationship between entities and also describes the object of the image because the structure is simple[5]. This structure can compensate for insufficient information in input sentences.

The K-BERT model using such knowledge graph can outperform the existing BERT models in cases that the knowledge about the input sentence is insufficient. However, because the K-BERT model uses the entire knowledge graph, the knowledge that deviates from the input sentence's topic is also added, which may confuse the model's learning. To relieve this problem, we divided the knowledge graph using topic modeling so that only the knowledge relevant to the input sentence's topic is added and learned in this study.

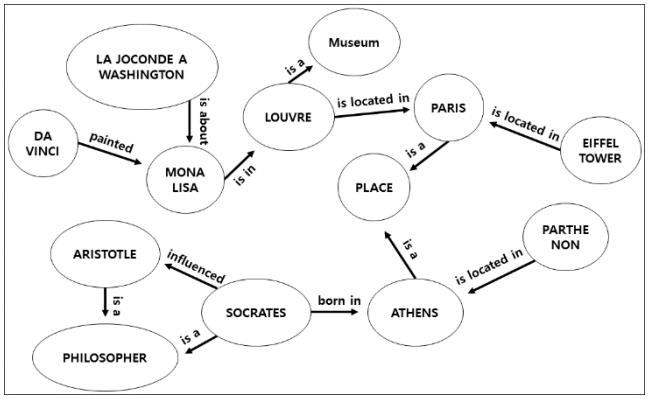

Fig. 1. Example of a knowledge graph

III. TOPIC MODELING

Topic modeling is a statistical model for discovering abstract topics in a set of documents. It identifies the topics based on the number of documents and topics by assigning the corresponding words to the topics. We employ the Latent Dirichlet Allocation (LDA) [6, 7] as our topic model. The LDA assumes that documents are composed of various topics and the topics generate words based on the Dirichlet probability distribution. The distribution of the topics in each document and words corresponding to the topics are estimated using the assumption. Fig. 2 shows the creation of a topic-based knowledge graph using LDA. G denotes the total number of triple sentences in the knowledge graph, K denotes the total number of topics, N denotes the number of entities in the g-th triple sentences, $z_{g,n}$ denotes the topic of the n-th word in the g-th triple sentence and $e_{g,n}$ denotes the n-th entity in the g-th triple sentences. The Gibbs sampling technique is used to estimate the LDA's parameters. The entity's topic is reassigned based on two criteria: $p(topic|triple\ sentences)$ and $p(entity|topic)$. The parameters are estimated by repeating this process until all entity topic assignments converge equally.

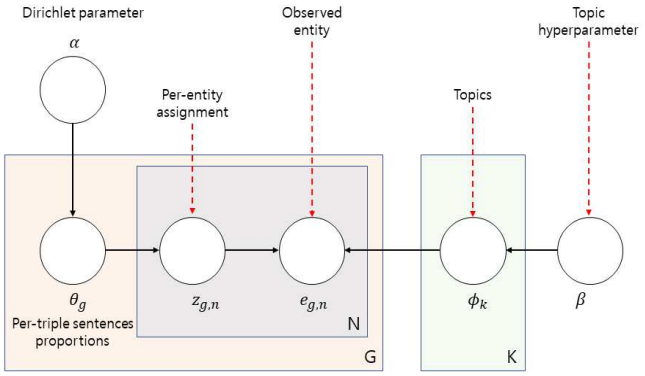

Fig. 2. Knowledge Graph creation process of the LDA

$$p(z_{g,i}=j|\boldsymbol{z}_{-i},e)=\left\{\frac{n_{g,k}+\alpha_k}{\sum_{i=1}^{K}(n_{g,i}+\alpha_i)}\right\}\times\left\{\frac{E_{k,e_{g,n}}+\beta_{e_{g,n}}}{\sum_{j=1}^{E}(E_{k,j}+\beta_j)}\right\} \quad \text{Eq. (1)}$$

(1) represents the probability that the topic $z_{g,i}$ of the i-th entity in the g-th triple sentence is allocated to the j-th topic. E denotes all the entities that appear on the knowledge graph, $n_{g,k}$ denotes the entity frequency of the g-th triple sentence assigned to the k-th topic, $n_{g,i}$ denotes the entity frequency of the g-th triple sentence assigned to the i-th topic, $E_{k,j}$ denotes the frequency of entity j assigned to the k-th topic in the entire knowledge graph, $E_{k,e_{g,n}}$ denotes the frequency of entity $e_{g,n}$ assigned to the k-th topic in the entire knowledge graph, α denotes the Dirichlet distribution parameter for generating the topic distribution of the knowledge graph and β denotes the Dirichlet distribution parameter for generating the word distribution of the topic.

IV. TOPIC-BASED KNOWLEDGE GRAPHS BERT

Fig. 3 shows the overall structure of the topic-based knowledge graphs BERT (TK-BERT) model. The TK-BERT model consists of three stages:

1) Creating a topic model and dividing the knowledge graph by topic
2) Inferring the topic of the input sentence
3) Adding knowledge relevant to the topic

In the first stage, a topic model is created through the knowledge graph and the LDA technique, and the knowledge graph is divided through the created topic model. In the second stage, the topic of the input sentence is inferred through the topic model created in the first stage. Finally, in the third stage, knowledge relevant to the topic is added from the knowledge graph corresponding to the topic of the input sentence inferred in the previous stages. This knowledge is used as an input to the BERT model for learning and classification. Since K-BERT learns the entire knowledge graph, sometimes it expresses inaccurate information. For example, according to the input sentence tree in Fig. 3, K-BERT also adds knowledge of "Athens is a place", which is not related to "Eiffel Tower", so the input sentence can be classified incorrectly. However, the TK-BERT model does not add knowledge of "Athens is a place" that does not fit the topic because it adds knowledge that fits the topic of the input sentence. In such cases, TK-BERT shows better performance than the K-BERT model because it does not add unnecessary knowledge.

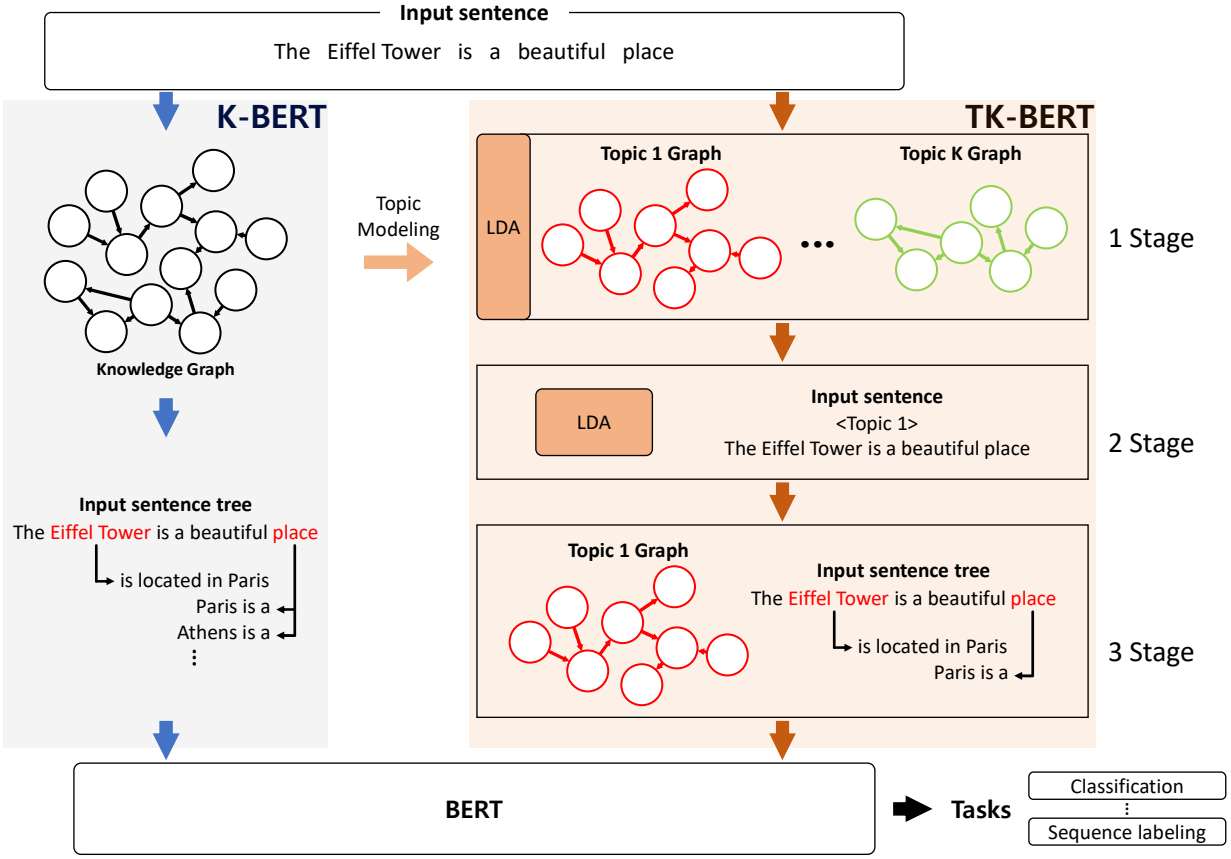

Fig. 3. Differences between TK-BERT and K-BERT Knowledge Reference

V. Experiment

In this study, the experiment was conducted to compare F1-Score between the K-BERT and the TK-BERT models. The K-BERT model was trained using two knowledge graphs, Cn-DBpedia and HowNet. The TK-BERT model was trained using the topic-based Cn-DBpedia and HowNet knowledge graphs that have been divided using the LDA topic model.

A. Experimental Environment

Pre-trained Model The Google BERT model, which is a pre-trained model published by Google, was used as the pre-trained BERT model in our experiment. This model has been pre-trained with the WikiZh data. WikiZh is a Chinese Wikipedia, with 12 million Chinese sentences in its database.

Topic Model The topic model was created using the LDA technique. Two knowledge graphs, Cn-DBpedia and HowNet, were used as the data to create the topic model. Two topic models were created--one for each knowledge graph. The number of topics was selected considering perplexity and topic coherence. The generated topic model was used to divide each knowledge graph by topic and estimate the topic of the input sentence.

Knowledge graphs Two knowledge graphs were used: Cn-DBpedia and HowNet. Cn-DBpedia is a knowledge graph created in a large open domain encyclopedia that defines tens of millions of entities and hundreds of millions of relationships, but in this study, objects containing special characters and stopwords were removed and used. The refined Cn-DBpedia consists of approximately 5.16 million triple data. In this study, the Cn-DBpedia data were divided into ten topics, with each topic containing approximately 510,000 triple data on average. HowNet is a Chinese vocabulary knowledge graph, with approximately 52,000 triple data. In this study, the HowNet data were divided into 15 topics, with each topic containing approximately 5,000 triple data on average.

Domain task Six datasets were used in the experiment. The datasets consist of five open domain task datasets, which are composed of Book_review, Chnsenticorp, Shopping, Weibo and LCQMC datasets and one specific domain task dataset consisting of Law_Q&A dataset. The Book_review dataset contains data about positive and negative reviews, with 20,000 data about online shopping reviews, with approximately 21,000 positive reviews and 19,000 negative reviews. The Weibo dataset contains data about Sina Weibo's emotion analysis, with 60,000 positive reviews and 60,000 negative reviews. The LCQMC dataset contains Chinese question matching data. The Law_Q&A dataset contains legal question-and-answer data, which selects the most appropriate

TABLE I.
COMPARISON BETWEEN THE KNOWLEDGE GRAPH AND THE DIVIDED KNOWLEDGE GRAPH

	Memory Allocation (KB)		Hash Table Size (# of keys)	
	KG	Topic-KG	KG	Topic-KG
Cn-DBpedia	172032	18432	3379703	410745
HowNet	2662	143	45124	3454

TABLE II. COMPARATIVE EXPERIMENTAL RESULTS OF TK-BERT MODEL AND K-BERT

Model\Dataset	Book_review		Chnsenticorp		Shopping		Weibo		LCQMC		Law_Q&A	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
K-BERT (HowNet)	86.8	85.85	94.5	94.6	96.5	96.05	98.25	98.35	84.8	86.15	85.65	86
TK-BERT (Topic-based HowNet)	**87.35**	**86.45**	94.5	**95.6**	**96.75**	**96.45**	98.25	98.35	**85.3**	**86.6**	**86.25**	**86.9**
K-BERT (Cn-DBpedia)	87.4	86.25	93.95	94.5	96.45	96.3	98.23	98.3	88.35	86.45	85.6	86.6
TK-BERT (Topic-based Cn-DBpedia)	**88.2**	**86.9**	**94.5**	**95.5**	**96.5**	**96.35**	98.25	**98.35**	88.35	**86.5**	**86.15**	**87**

answer to a question.

B. Experimental Result

First, Table I shows the average memory allocation and hash table size in TK-BERT are reduced by assigning and using existing knowledge graphs by subject. The TK-BERT model can effectively use a large-scale knowledge graph because it uses a partial knowledge graph for each topic rather than the entire knowledge graph. Second, Table II shows the F1-Score of K-BERT and TK-BERT. when the K-BERT and TK-BERT models using the HowNet knowledge graph were compared, the F1-Score of TK-BERT increased by 0.6, 1.0, 0.45, 0.45 and 0.9 for the Book_review, Chnsenticorp, Shopping, LCQMC and Law_Q&A data sets. Moreover, the K-BERT and TK-BERT models performed similarly for the Weibo data set. Overall, performance was improved for most of the data sets when the TK-BERT model was used. When the K-BERT and TK-BERT models using the Cn-DBpedia knowledge graph were compared, the F1-Score of TK-BERT increased by 0.65, 1.0, 0.05, 0.05, 0.05 and 0.04 for the Book_review, Chnsenticorp, Shopping, Weibo, LCQMC and Law_Q&A data sets. It is shown that the TK-BERT model showed better performance for most of the data set. In this experiment, the knowledge graph was divided into topics using topic modeling techniques to reduce knowledge noise that adds knowledge that is unhelpful to learning. Using a topic-based knowledge graph, it is possible to identify the topic of an input sentence and to avoid knowledge noise because knowledge that does not fit the topic is not added. Through experiments, it can be seen that using knowledge graph divided by topic showed better performance. In other words, if the knowledge graph is divided and used for each topic, the knowledge graph can be used more effectively for learning.

VI. CONCLUSIONS

In this study, we proposed the TK-BERT model to improve the problem with the existing K-BERT model. The K-BERT model has the problem of adding knowledge that is irrelevant to the input sentence's topic, which the TK-BERT solves using topic modeling. The TK-BERT model creates a topic model using the LDA technique and knowledge graph during topic modeling. It then divides the knowledge graph by topic using the created topic model and infers the input sentence's topic. After that, only the knowledge relevant to the topic is added. We demonstrated that the TK-BERT model outperforms the K-BERT model on various datasets. These results showed that large-scale knowledge graphs can be used more effectively. In the future, we plan to conduct comparison experiments to find a better model for knowledge graphs using various topic modeling techniques.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00231, Development of Approximate DBMS Query Technology to Facilitate Fast Query Processing for Exploratory Data Analysis). This work was also supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2022-2018-08-01417) supervised by the Institute for Information & Communications Technology Promotion (IITP). This research was also funded by industry-academic Cooperation R&D program funded by LX Spatial Information Research Institute (LXSIRI, Republic of Korea).

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, 2018.

[2] W. Liu, "K-BERT: Enabling Language Representation with Knowledge Graph." Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 3, pp. 2901-2908, 2020.

[3] C. Gutierrez and J. F. Sequeda, "Knowledge graphs," Communications of the ACM, vol. 64, no. 3. Association for Computing Machinery (ACM), pp. 96–104, Mar. 2021.

[4] L. Yao et al., "Incorporating Knowledge Graph Embeddings into Topic Modeling," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1. Association for the Advancement of Artificial Intelligence (AAAI), Feb. 12, 2017.

[5] S. Kim, T. H. Jeon, I. Rhiu, J. Ahn, and D.-H. Im, "Semantic Scene Graph Generation Using RDF Model and Deep Learning," Applied Sciences, vol. 11, no. 2, p. 826, Jan. 2021.

[6] Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, pp. 993-1022. Jan. 3, 2003

[7] J. Sleeman, T. Finin, and A. Joshi. "Topic modeling for rdf graphs." 3rd International Workshop on Linked Data for Information Extraction, 14th International Semantic Web Conference. Vol. 1267. 2015.