Preliminary Evaluation of BERT Embedding for Semantic Search in Nuclear Engineering Domain

Byoungchan Han*, Byeong-hyeok Ha, Byeongmun Ahn, Tongkyu Park, Sung-Kyun Zee
FNC Technology Co., Ltd., 32F, 13 Heungdeok 1-ro, Giheung-gu, Yongin-si, Gyeonggi-do, Korea
**Corresponding author: bchan007@fnctech.com*

1. Introduction

Nuclear power plant (NPP) design documents have been produced since the first NPP was built in Korea, and a vast amount of information has gradually been accumulated. For minimizing the cost and the time for NPP design and maximizing the design quality, design engineers' experiences and know-how are necessary. As a result, the necessity for an intelligent semantic search and analysis system that can provide information quickly among the huge plant design documents and materials has arisen.

As the first step of developing intelligent semantic search and analysis system for nuclear engineering, we introduce the methodology and its potential for semantic search system using machine learning model, named BERT.

2. Preliminaries

In this section, we briefly review the state-of-the-art language representation model called BERT.

2.1. BERT

BERT [1], which stands for Bidirectional Encoder Representations from Transformers, is a language representation model widely used in natural language processing (NLP) tasks. It is designed to train deep bidirectional representations from unlabeled text through a contextual layer, which is composed of stacked attention layers [2] and feed forward networks. This model is commonly used to generate word or sentence embedding vectors. For solving automated NLP tasks, these embedding vectors are connected with other machine learning models and transfer learning is performed. BERT model achieved state-of-the-art performances on eleven NLP benchmarks since it was published.

There are two steps for applying BERT model; pre-training and fine-tuning. In pre-training step, the BERT model is trained on unlabeled data. For fine-tuning, parameters of pre-trained model are fine-tuned based on labeled data from the various downstream tasks.

2.2. Pre-training step

Devlin et al. proposed two methods for unsupervised pre-training on the BERT model [1]. Masked Language Model (MLM) randomly replaces some of the tokens with [MASK] placeholders, usually 10% to 15%, in the input sequence and trains the model to predict the original words of the placeholders based on other tokens. Unlike standard transformer based models which learn the context of the input sequence within one direction (left-to-right or right-to-left), MLM put all tokens including [MASK] placeholders to transformer encoder layer and then predicts the targeted placeholder. Therefore, the BERT model can learn the context of the input sequence in both directions.

Second pre-training method is Next Sentence Prediction (NSP), which makes the BERT model to learn longer-term dependencies across sentences. 50% of NSP dataset consists of subsequent sentences pairs, and the rest of unconnected pairs. Such method is necessary for Question Answering (QA) and Natural Language Inference (NLI) tasks as they are based on learning the relationship between two sentences.

2.3. Fine-tuning step

The fine-tuning procedure depends on the NLP task to be performed. Pairs of task-specific input and output data are plugged into the pre-trained BERT model, and parameters are fine-tuned end-to-end for certain NLP task.

3. Methodology and Results

In this section, we present the performance of semantic search system using the BERT model in nuclear engineering domain.

3.1. Methodology

In order to implement a semantic search engine based on machine learning algorithms, derivation of sentences or documents embedding vectors is required. In the original BERT framework, sentence embedding is calculated using the mean, maximum pooling of token vectors, or treating the [CLS] token's output vector as a sentence embedding. However, such approaches have performed worse in NLP tasks than using the Glove [3] embedding method.

For this reason, we used the Sentence-BERT model [4] to derive sentence embeddings from the input sequence. This model has additional pooling layer to obtain a vector on the BERT model pre-trained in the form of Siamese network. It is noted that the term "Siamese network" comes from the characteristic of

sharing same weights of neural network while learning two input data. [5] Fig. 1 depicts the structure of the Sentence-BERT model.

In this study, the Sentence-BERT model was firstly initialized with multi-qa-mpnet-base-dot-v1 pre-trained model [6]. The model was then MLM pre-trained using phrases from articles in the Nuclear Engineering and Technology journal (NET) as training data. Finally, cosine similarity scores between the search query and each article were used to obtain semantic search results in the Proceedings of the KNS Meeting based on the MLM pre-trained model. The input materials were 17,459 KNS articles published between 2005 and 2020.

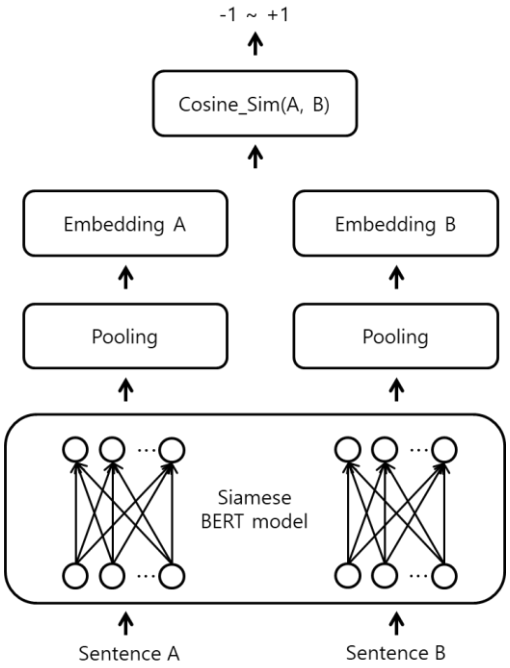

Fig. 1. Architecture of the Sentence-BERT model to compute cosine similarity scores.

3.2. Semantic Search Results

Each division name from the Proceedings of the KNS Meeting was used as the search query in Table I to evaluate the search engine's performance. Following that, we compared the actual divisions of the publications that were searched with the search query. Table II displays the outcome.

According to the findings, the BERT model performed well while searching with the search queries of case 3, 5, 8, 11 and 12. On the other hand, search queries of case 1, 2, 6, 9 and 10 lead to the poor performances of semantic searching. Such differences arose from the range of the topics. For instance, documents about "reactor system technology" and "nuclear safety" categories can be appeared across the overall nuclear engineering domain. For this reason, we infer that it is natural which the performance of the search engine on a wider variety of searches provide poor accuracy as academic publications can cover a wide range of topics.

Table I: List of Search Queries

No.	Search Query
1	reactor system technology
2	reactor physics and computational science
3	nuclear facility, decommissioning and radioactive waste management
4	nuclear fuel and materials
5	nuclear thermal hydraulics
6	nuclear safety
7	radiation protection
8	radiation utilization and instrumentation
9	quantum engineering and nuclear fusion
10	nuclear power plant construction and operation technology
11	nuclear policy, human resources and cooperation
12	nuclear I&C, human factors and automatic remote systems

Table II: The Number of Papers that Match the Divisions with the Search Query in Top 100 Most Similar Results

No.	Matched Results	Total Results
1	34	100
2	27	
3	90	
4	42	
5	64	
6	16	
7	50	
8	78	
9	27	
10	33	
11	85	
12	79	

4. Conclusion

The applicability of the BERT word embedding method to the domain of nuclear engineering was demonstrated in this work. As the results showed that the model is capable of obtaining the documents that the user requires, evaluated word embedding vectors can be employed in semantic search tasks, particularly in the field of NPP Architectural Engineering & Design (A/E).

The computed model, however, still has a possibility to be changed. On this model, we had only done unsupervised learning. With the specified pairings of task-specific input and output data, the model can be fine-tuned even further. Therefore, we will be concentrating our efforts in the future on creating labeled dataset for supervised learning in nuclear engineering domain, as well as fine-tuning of the BERT model.

ACKNOWLEDGMENTS

This work was supported by the collaborative research program with KEPCO Engineering & Construction Company.

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is All you Need, Advances in Neural Information Processing Systems, Vol. 30, 2017.
[3] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing, pp. 1532-1543, 2014.
[4] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084, 2019.
[5] G. Koch, R. Zemel, and R. Salakhutdinov, Siamese neural networks for one-shot image recognition, International Conference on Machine Learning deep learning workshop, Vol. 2, 2015.
[6] Hugging Face, https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1.