The V International AMMCS
Interdisciplinary Conference
Waterloo, Ontario, Canada
August 18-23, 2019

Utilizing Bidirectional Encoder Representations from Transformers for Answer Selection Task

Md Tahmid Rahman Laskar[1,3], Enamul Hoque[2], Jimmy Huang[2,3]

[1] *Department of Electrical Engineering and Computer Science, York University, Toronto, Canada, tahmedge@cse.yorku.ca*
[2] *School of Information Technology, York University, Toronto, Canada, enamulh@yorku.ca, jhuang@yorku.ca*
[3] *Information Retrieval & Knowledge Management Research Lab, York University, Toronto, Canada, jhuang@yorku.ca*

The Bidirectional Encoder Representations from Transformers (BERT) [2] model has been found very effective in different Natural Language Processing (NLP) tasks recently. The BERT model is pre-trained on language modeling task and it can provide contextualized representations of each token in a sentence. Though the fine-tuned BERT model has provided state-of-the-art results in different NLP problems, to our knowledge it has not been evaluated for the answer selection task yet. In this work, we fine-tune the pre-trained BERT model (see Figure 1) for the answer selection task. In the task, a question is given along with some candidate answers. The goal then is to rank the candidate answers based on their similarity with the question. More specifically, we take the question $X = x_1, x_2, ..., x_m$ and the candidate answer $Y = y_1, y_2, ..., y_n$ as input to the BERT model. Then the sentences are combined together into a single sequence, separated by a special token $[SEP]$. The final hidden state C of the first token ($[CLS]$), which is the aggregate representation of the sequence, is used for classification. During fine-tuning, parameters are added for an additional classification layer W. All the parameters of the pre-trained BERT model along with W are fine-tuned jointly to maximize the probability of the correct label. The label probabilities $P \in \mathbb{R}^K$ (where K is the total number of classifier labels) are calculated as follows:

$$P = softmax(CW^T) \tag{1}$$

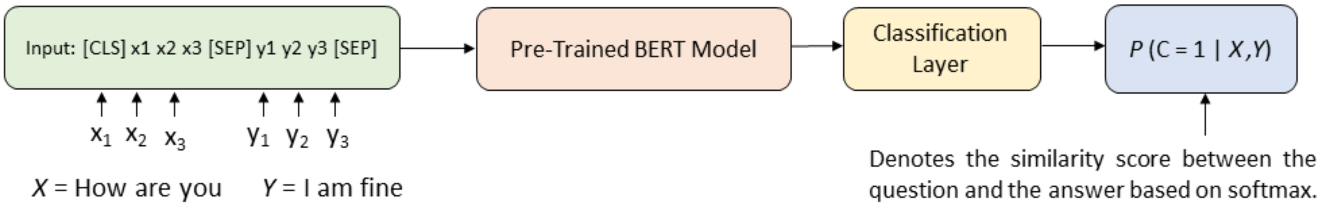
Input: [CLS] x1 x2 x3 [SEP] y1 y2 y3 [SEP]
Pre-Trained BERT Model
Classification Layer
P (C = 1 | X,Y)
x1 x2 x3
y1 y2 y3
X = How are you
Y = I am fine
Denotes the similarity score between the question and the answer based on softmax.

Figure 1: Fine-tuning the BERT model: The tokens of question X and a candidate answer Y are combined together as input to the pre-trained BERT model. The parameters are then fine-tuned based on the classification output.

We evaluate the effectiveness of the fine-tuned BERT model on two different datasets, namely, the TREC-QA and the WikiQA. We observe new state-of-the-art results for the answer selection task in both datasets. In terms of Mean Average Precision (MAP) metric, the fine-tuned BERT model has an improvement of 6.18% in the TREC-QA and an improvement of 13.72% in the WikiQA datasets respectively over the state-of-the-art approaches [1].

References

[1] Q. Chen, Q. Hu, J. X. Huang, and L. He. *CA-RNN: Using Context-Aligned Recurrent Neural Networks for Modeling Sentence Similarity*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, (2018).

[2] 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).