Incorporating BERT into Document-Level Neural Machine Translation

1810409 GUO Zhiyu

In recent years, we have witnessed the rapid development of deep learning technology, and the application of deep learning in the field of machine translation has continued to be deepening. Among them, the attention-based encoder-decoder Neural Machine Translation (NMT) framework surpassed the traditional statistical machine translation framework in performance significantly. Further, the Transformer framework has improved the performance of neural machine translation to a new level.

Due to the limitation of training methods, these advanced frameworks consider one sentence as a whole in the process of translation. In the actual translation process, the text we use is often composed of multiple sentences. As the document has special characteristics, the translation of these sentence-level models is often lacking coherence and cohesiveness when translating documents.

Since late 2018, large-scale pre-trained representations such as BERT have been widely used in many natural language understanding tasks, including machine reading comprehension, text classification. The methods of incorporating BERT into document-level machine translation are still being explored. BERT is able to understand sentence relationship since one of the BERT pre-training task is the next sentence prediction task, the sentence relationship information is very important for document-level machine translation. Therefore, in our work, we leverage pre-trained BERT to improve the performance of document-level machine translation.

In this research, we propose a novel method to incorporate pre-trained BERT into document-level NMT. The BERT model performs as a context encoder to model the document-level contextual information. We concatenate the document-level context and the current sentence as the input for the BERT context encoder. The contextual-representation encoded by BERT is then integrated into both the encoder and the decoder of the Transformer NMT model using the multi-head attention mechanism. The attention mechanism can also deal with the case that BERT module and Transformer NMT module might use different word segmentation rules. Given the fact that translating different sentences may require a different amount of contextual information, we propose to use context gates to integrate the output of the multi-head attention mechanism.

The parameter size of our model is very huge, to save training time, we propose a two-step training strategy for our model. Firstly, we split the

document-level training data into separate sentences, we train a sentence-level Transformer NMT model. After that, we use the sentence-level Transformer NMT model to initialize the parameter of the Transformer NMT module in our model, and we train the document-level NMT model with the parameter of the BERT module fixed.

We tested our model on English-German and Chinese-English datasets. The results showed huge improvements over the sentence-level Transformer model, and our proposed model outperformed several strong document-level NMT baselines. Especially, our model achieved new state-of-the-art performance on the English-German News Commentary dataset. The effectiveness of our model has been proved.

We tried to integrate the contextual representation encoded by BERT into a different part of the Transformer NMT model. The results showed integrating contextual representation into the encoder can achieve more improvements than integrating into the decoder. Integrate the contextual representation into both the encoder and the decoder of the NMT model can achieve the best results.

Regrading previous research argues that the context encoder in document-level NMT can not capture contextual information, we follow their experimental setting presenting three inputs for BERT context encoder. The results showed that the BERT context encoder in our model can capture contextual information to improve translation performance.

In future work, we would like to compress our model into a light version. Also, we would like to use more than one context sentences. Furthermore, we would like to test the performance of our model in some low-resource languages.

Keywords: *Deep Learning, Neural Machine Translation, Pre-trained Model, Document-level, Attention Mechanism, Context Gate*