Chinese text multi-classification based on Sentences Order Prediction improved Bert model

Guanping Fu¹ and Jianwei Sun²

¹North China Institute of Computing Technology, Beijing, 10089, China
²North China Institute of Computing Technology, Beijing, 10089, China
Fu Guanping’s e-mail: 3220191188@bit.edu.cn

Abstract. For the strong noise interference brought by the NSP mechanism (Next Sentences Prediction) in Bert to the model, in order to improve the classification effect of the Bert model when it is used in text classification, an SOP (Sentences Order Prediction) mechanism is used to replace the Bert model of the NSP mechanism-Multi-classification of Chinese news texts. At first, use randomly sorted adjacent sentence pairs for segment embedding. Then use the Transformer structure of the Bert model to encode the Chinese text, and obtain the final CLS vector as the semantic vector of the text. Finally, connect the different semantic vectors to the multi-category Classification. After ablation experiments, the improved SOP-Bert model obtained the highest F1 value of 96.69. The results show that this model is more effective than the original Bert model on text multi-classification problems.

1. Introduction
Text classification is a basic work in the field of natural language processing, which is used in information retrieval, deception detection and recognition, sentiment analysis and spam detection. The technical basis of news classification is text classification in natural language processing[1]. With the rapid development of the Internet, the information on the Internet has exploded. Thus, it is more and more difficult for people to get useful information, and it is urgent to classify the positive and negative information of mass news. Text classification can solve the problem of information confusion in news data, and to a large extent help people locate the source of information more. Traditional text classification methods such as TF-IDF model[2], LDA topic model[3], etc, but the classification effect is not good for most cases; while machine learning for text classification requires complicated manual features design and feature extraction process[4]. With the development of the field of deep learning, the effect of text classification technology based on deep learning is significant[5]. In 2018, Google released the Bert model based on Pre-train&Fine-tune[6], which was applied for the field of natural language processing. Here, we use the SOP mechanism to replace the NSP mechanism in Bert based on the design shortcomings of the Bert model in the field of text classification and have achieved good results. Due to the strong noise interference brought by the NSP (next sentences prediction) mechanism of the Bert to the model[7], its effect on text classification is not good. In our paper, SOP (Sentences Order Prediction) is used as the training target of the model, and Chinese text multi-classification is used as the downstream task for experiments. The results show that the Bert model improved by SOP is more effective in the task of text classification.
2. The core of the Bert model

2.1. Tokenization and Embedding

The text segmentation stage is called Tokenization in Bert, which aims to divide the original input text into token levels. For example: "I love natural language processing", afterword segmentation, it will become: "[CLS][I][Love][Natural][Language][Processing]". The English word segmentation in the Bert model uses a simple space division method. Because of the particularity of the English text, there will be spaces between each token to distinguish it. The Chinese text is different, it does not have obvious signs or symbols to distinguish tokens. After reading its source code, I found that the Chinese version of Bert uses the word level as the token division. Obviously, this is not a good division method, because usually, we use words as the smallest granularity division method of Chinese. We can use the Jieba word segmentation tool to divide the Chinese text into word-level tokens, and then create a vocabulary for the divided data set. The Jieba word segmentation tool will not be introduced here. An important operation afterword segmentation is a random mask. The Bert model uses an unsupervised Mask Language Model (MLM) to pre-train the target. The specific operation is: randomly replace the original token with the [mask] character, the replacement probability is 15%, and the training goal is to predict the masked word based on the context. In addition to MLM, Bert also uses "Next Sentences Processing" (NSP) to jointly train the feature representation of text pairs. Considering that the interference in random different text pairs on the encoder is particularly serious, it is considered in this article Use "Sentences Order Prediction" (SOP) to replace the NSP mechanism, and its detailed description will be introduced into Section 3.

2.2. Transformer

The Bert model uses a multi-layered transformer structure as a feature extractor. Like the traditional seq2seq model, Transformer is also composed of two parts: encoder and decoder. Bert only uses encoder stage for feature extraction. The encoder of Transformer is composed of the same 6 components, each of which has the same structure. This component consists of two parts. The first part is the Multi-Head Attention+LN (Layer Normalization) layer. The multi-head attention mechanism is a parallel operation performed by the Bert model on the self-Attention structure. It is composed of multiple attention head distributions. Each attention head is calculated in parallel and added after the calculation is completed, and then multiplied by a parameter matrix to obtain the final attention result, and then normalize the attention result. The second part of the component are the FFN+LN layer, which consists of a feed-forward neural network and a normalization layer. The feed-forward neural network can extract the deep features of the result of the previous component and can get the required matrix transformation Vector features of specific dimensions. For the LN structure, it is similar to the BN (Batch Normalization) structure. The batch normalization is to use a small batch of samples for normalization and convert it to data onto a mean value of 0 and a variance in 1. On the other hand, LN calculates the mean and variance between each sample. The calculation formula is as follows.

\[ LN \left( x_i \right) = \alpha \times \frac{x_i - \mu_L}{\left( \sigma^2_L + \epsilon \right)^{\frac{1}{2}}} + \beta \]  

3. Bert model with SOP

3.1. Embedding

The Embedding stage of Bert uses 3 layers of Embedding, and each layer is added at the end to get the output of the final Embedding layer. These three layers are token embedding layer, segment embedding layer and position embedding layer. As shown in figure 1 below.
Figure 1. Embedding process of Bert model.

For Embedding at the token level, each token will occupy a position in the vocabulary. The Bert model uses one-hot encoding to encode the token. For example, suppose vocabulary is composed of 7 tokens [CLS,i,like,dog,he,cat,SEP], then the token embedding vector of "CLS" is [1,0,0,0,0,0,0], the token embedding vector of "i" is [0,1,0,0,0,0,0] in the same way. For segment embedding, the same sentence will be assigned the same embedding vector. The original Bert model uses the NSP mechanism for segment embedding. Here we use SOP to transform it. See Section 3 for details. The last embedding vector added is the position embedding vector. As we all know, the traditional RNNs structure such as LSTM, GRU, Bi-LSTM and other temporal structures can carry the word order information of the token through the encoding layer, but the transformer model based on the attention mechanism does not contain this word order information. Therefore, the Bert model introduces position embedding vectors to express word order features. After the three-layer encoding, the encoded three-layer results are finally added to obtain the input encoding feature vector of the final embedding layer.

3.2. SOP instead of NSP

3.2.1. Why SOP

NSP is based on two random sentence pairs, while SOP is a text pair based on a sorted sequence of two consecutive text pairs. Coherence and cohesion in texts have been studied, and many researchers have discovered the phenomenon of connecting adjacent text segments. Several researchers have experimented with discourse coherence as a pre-training goal. But most of the pre-training targets that have valid results from the experiment are very simple. Skip-thought[9] and FastSent[10] sentence embedding uses sentence encoding to predict words in adjacent sentences. There are also embedding learning for other purposes as the pre-training target, such as predicting future sentences instead of predicting adjacent sentences, predicting clear discourse markers and so on. SOP is different from the above method. It calculates the loss most like the sentence ordering goal of Jernite’s[11], in which sentence embedding is learned to determine the order for two consecutive sentences. However, SOP varies according to the text segment, which is different from Jernite’s design method.

3.2.2. Model structure

Modelling between sentences is an important aspect of language understanding. For the SOP structure Bert model, the text pair loss is used, which avoids topic prediction and focuses on modelling the coherence of sentences. SOP loss uses two consecutive segments of the same document as Bert on the
positive example and uses the same consecutive segment of the negative example, but the order is reversed. This forces the model to learn more fine-grained differences in discourse coherence. This paper uses the SOP improved Bert model for text multi-class classification, and its model structure is shown in figure 2.

![Text multi-class classification model structure](image)

In the Embedding layer, the addition of the three encoding vectors of Masked Word Embedding, SOP Embedding and Position Embedding is used as the input vector of the model, and the feature extraction is carried out through the superimposed layer of Transformer. Finally, the vector can be classified by the classifier of the downstream task.

### 4. Experiment

We use the Sohu News text corpus to perform classification experiments to verify the effect of the improved SOP method of the condition that the downstream task is text classification, and try Logistics, SVM, Dense network, etc. in the selection of downstream classifiers.

#### 4.1. Method

The experimental training data set comes from the Sohu news text data set. After sorting and dividing, finally divided into 14 candidate classification categories: finance, lottery, real estate, stocks, home furnishing, education, technology, society, fashion, current affairs, sports, constellations, games, entertainment text data. Among them, 20W samples of training data and 5W samples of the test set. And in the data preprocessing stage, the length of the text is counted, which facilitates the interception of the text length and zero-filling in the later stage.
In this experiment, the Bert base model is used as the Base-line. The activation function adopts the gelu activation function, the drop-out ratio is 0.1, the hidden layer size is 12*768, the attention head is 12, and the vocabulary size is 21128. The training parameters train_batch_size is 32, eval_batch_size is 8, predict_batch_size is 8, train_epochs is 3, and the learning rate lr is 5e-5.

4.2. Analysis
In the experiment, the classification index F1 is used to measure. In the multi-class model, there are two common evaluation indicators, macro and micro. Among them, the macro average calculates the F1 of each category separately and then finds the average value. Since it treats each category equally, it is more susceptible to the influence of the F1 value of the category with a small sample. Instead, micro-averaging calculates the F1 value. It is easily affected by the F1 value of multi-sample categories. In this experiment, the macro average evaluation index is used, that is, macro-F1.

5. Results
Our experiment uses Logistics, SVM, and Dense network as downstream classifiers, and comparative experiments are carried out on the Bert base model and the SOP Bert model respectively. The experimental results are shown in table 1.

|          | Macro_F1 | Logistics | SVM     | Dense   |
|----------|----------|-----------|---------|---------|
| Bert base| 76.15%   | 79.43%    | 83.23%  |
| SOP Bert | 80.01%   | 80.92%    | 86.51%  |

Experimental results show that the improved SOP mechanism is better than the Bert NSP mechanism in classification tasks, and its F1 value is about 3% higher than that of downstream classifiers, and the classification effect of the Dense network is the best, and its F1 value reaches That’s 86.51%.

6. Conclusion
With the development of artificial intelligence, text automatic classification technology, as an effective way of information organization, has received extensive attention and applications. As a basic technology in the fields of information retrieval, information filtering and document management, it has a wide range of applications and development prospects. This paper is centered on Chinese text classification, and the research is focused on the vector encoding model in the text classification problem. We proposed a vector encoding model based on SOP, and conducted experiments to verify its efficiency. However, due to the limitation of time and personal ability, there are still many shortcomings in this work. Future work will focus on the following points:

- Optimize the data structure in the classification algorithm to further improve the classification speed.
- Consider the combined use of multiple classifiers, which will play a key role in text classification of some nested categories.
- Strengthen the proportion of semantic analysis in the model. Semantic analysis is an important direction of future research, and it is of great significance to natural language processing.

References
[1] Wu Xingsu, Chen Jinhui. (2020) Text Classification on Large Scale Chinese News Corpus using Character-level Convolutional Neural Network[J]. Journal of Physics: Conference Series, 1693(1).
[2] GQ W, SZ N. (2020) Research on Text Information Classification of Social E-commerce Based on TF-IDF[J]. Network Space Security, 11(12):32-38.
[3] F Z, DH W, S H. (2020) Text classification method based on LDA and deep learning[J]. Computer Engineering and Design, 41(08):2184-2189.
[4] D W. (2020) Analysis and Research on Text Classification Technology Based on Machine Learning[J]. Science and Technology Innovation Guide, 17(08):90+92.
[5] DY L. (2020) Research on the Application of Chinese Text Multi-classification Based on LSTM[J]. Journal of Shanghai Electric Power University, 36(06):598-602.
[6] Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention Is All You Need[J]. arXiv:1706.03762.
[7] Lan Z, Chen M, Goodman S, et al. (2019) ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[J]. arXiv:1909.11942.
[8] FG S. (2020) Implementation of Chinese text corpus preprocessing module based on jieba Chinese word segmentation[J]. Computer Knowledge and Technology, 16(14):248-251+257.
[9] Kiros R, Zhu Y, Salakhutdinov R R, et al. (2015) Skip-Thought Vectors[J]. Advances in Neural Information Processing Systems, 28.
[10] Hill F, Cho K, Korhonen A. (2016) Learning Distributed Representations of Sentences from Unlabelled Data[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
[11] Jernite Y, Bowman S R, Sontag D. (2017) Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning[J]. arXiv:1705.00557.