Hybrid Chinese text classification model based on pretraining model

Xing Zhaoye¹, Liu Xiaojun*, Sun Peijie³

¹School of Information Engineering, College of architectural engineering, Zhangjiakou, Hebei Province, 075000, Chinese
²Corresponding author’s e-mail: lxq@hebiace.edu.cn
³School of Information Engineering, College of architectural engineering, Zhangjiakou, Hebei Province, 075000, Chinese

Abstract: In order to solve the problems of fuzzy boundary of Chinese words, polysemy of one word and the inability of traditional models to reflect the importance of each word in the text, a hybrid text classification model based on Ernie CNN and bilstm attention (mecba) is proposed. Firstly, Ernie model is used to generate the corresponding word vector, which can retain rich semantic information and enhance the semantic representation of words. Then, the word vector is input into CNN to extract local information, and the context feature is extracted by bilstm attention, and the output of the two is spliced; Finally, softmax classifier is used for classification. The experimental results show that the model is better than CNN, bilstm and other classification models, and can effectively improve the performance of Chinese text classification.

1. Introduction

Text classification is a classic problem in natural language processing (NLP). Its purpose is to assign tags to text units, such as sentences, paragraphs and documents. It has a wide range of applications, including question answering system, spam detection, sentiment analysis, news classification, user intention classification, content audit and so on. Text data can come from different sources, such as network data, e-mail, chat, social media, air tickets, insurance claims, user comments, customer service questions and answers. Text information contains a lot of important information, so how to get the information efficiently is of great significance.

In recent years, because deep learning can express deeper information in the text without prior knowledge, and can accommodate the advantages of massive data and feature extraction, most text classification tasks now use this method. At present, convolution neural network (CNN) and recurrent neural network (RNN) are two commonly used deep learning methods in the field of text classification. Kalchbrenner proposed a CNN based classification model, which greatly improves the accuracy of text classification. Although CNN model has good classification effect, it ignores the semantic information of text context. Miolovt and others use RNN to capture word correlation and text structure, and capture the semantic information of text context for text classification. However, due to the gradient ladder elimination or gradient explosion problem in the practical application of RNN model, long-term and short-term memory network (LSTM) is widely used, which aims to better capture long-term dependence and effectively alleviate the gradient disappearance or explosion problem faced by RNNs. Li et al. [1] proposed CNN bilstm model, which used CNN to extract local features, bilstm to extract context semantics, and combined them to get feature vectors, which significantly improved the
accuracy of classification, but did not pay attention to more important words. To solve this problem, Hou et al. [2] introduced the attention mechanism and achieved good results in text classification.

Chinese is different from English in that every word in English is separated by a space, while every word in Chinese is connected with each other, which leads to fuzzy word boundary and polysemy in different contexts, which greatly affects the classification results. In order to solve this problem, Google puts forward the pre-processing model of Bert language. The best model can capture the whole sentence sequence information, context or semantic information of context, and solve the problem of polysemy. It can also perform parallel operations to further enhance the representation ability of text information. In this paper, enhanced language representation with information entities (Ernie) model is adopted. Based on the best model, knowledge is explicitly added to the best model, so that the model can learn more knowledge information. At the same time, most of the pre training data is thucnews Chinese data set, so it is more suitable for the task of Chinese text classification. In addition, CNN and LSTM attention hybrid network is added to Ernie model to further extract feature information, which significantly improves the classification performance.

2. Hybrid Text Classification Model of CNN and bilstm attention (mecha) based on Ernie

Upstream task language preprocessing has always been a hot topic in text classification task research. Ernie language preprocessing model uses multiple information entities in knowledge map as external knowledge to improve language representation, so that the model can learn more knowledge information, obtain high-quality word vector, and achieve the purpose of improving classification effect.

The mecha model structure constructed in this paper consists of five parts: Ernie model, Bi LSTM layer, CNN layer, attention layer and output layer. As shown in Figure 1. Firstly, the Ernie pre training language model is used to annotate the corpus to obtain the corresponding word vector. Then, the word vector is input to CNN to extract the local information. At the same time, the context information is extracted by bilstm layer, and the attention mechanism is added behind the layer to obtain the score of the information, so as to reflect the importance of each word. Finally, the two outputs are spliced and classified by softmax. So as to complete the whole process of text classification.

Figure 1 mecha model
2.1. ERNIE pretraining language model
Ernie (enhanced language representation with informational entities) model adds k-encoder based on Bert to realize the fusion of knowledge information and original semantic information of token. It does not need to train word vector and word vector in advance. It only needs to input sequence in Ernie to automatically extract word level features of sequence. It has the characteristics of grammatical structure and semantics.

![Figure 2 Ernie model](image)

2.2. BiLSTM layer
Recurrent neural network (RNN) is a kind of neural network which can capture time series information and contain cycles. However, gradient explosion is easy to occur when obtaining long-distance dependent information, which makes it difficult to deal with long-distance sequences. Long short term memory network (LSTM) realizes the selective utilization of historical information through gating mechanism, so as to effectively solve the problem of long-term dependence.

The bilstm layer uses two layers of LSTM to compute and learn the sequence data from the front and back respectively, and fuses the obtained information to capture better context information.

2.3. CNN layer
The CNN layer uses the basic textcnn structure to extract features. The first layer is the input layer, which takes the semantic vector extracted from LSTM layer as the input; The second layer is convolution layer, which uses multiple convolutions of different sizes to check the input semantic vector to perform convolution; The third layer is pooling layer, which performs the maximum pooling operation and reconstructs a new feature vector y.

2.4. Attention layer
Attention mechanism is a selective resource allocation model, which can focus on more important content and ignore unimportant information just like human beings. In text classification, the more important words are in the text, the more important they are to the classification effect. Focusing on more important information can significantly improve the classification effect.
In order to prove the effectiveness of mecba model algorithm, this paper tests the classification results on data sets. The experimental environment and configuration are shown in Table 1.

### Table 1 experimental environment and configuration

|                |                 |
|----------------|-----------------|
| operating system | Ubuntu          |
| CPU            | i7-7700HQ@2.8GHz|
| GPU            | GTX1660ti       |
| Python         | 3.6.0           |
| Deep learning framework | Pytorch     |

#### 3.1. Experimental data set

The experimental data set is from the thucnews corpus. It contains about 740000 news documents, which are generated by filtering historical data of Sina News RRAs from 2005 to 2011, all in UTF-8 format. This paper selects ten categories, including stock, swimming, sports, finance, music, trend, military and so on. There are 10000 corpora in each category, with a total of 100000 corpora. The 100000 pieces of data are divided into training set, verification set and test set, including 80000 pieces of training set, 10000 pieces of verification set and 10000 pieces of test set.

#### 3.2. Experimental process

In this paper, some parameters of Ernie model are fixed, and only the parameters of LSTM, Attention and CNN are updated. That is to say, the optimal parameters of mecba model are obtained by continuously adjusting the model parameters.

In order to explore the performance of the model, the current mainstream deep learning text classification model is used as the comparison model, and the accuracy and F1 value are mainly used...
for evaluation. The model comparison is shown in Table 2.

| Model              | Accuracy | Recall rate | F1 value |
|--------------------|----------|-------------|----------|
| CNN                | 90.78    | 90.80       | 0.908    |
| BiGRU              | 90.29    | 90.47       | 0.904    |
| transform          | 89.77    | 89.60       | 0.893    |
| LSTM-CNN           | 90.88    | 90.83       | 0.912    |
| CNN-BiGRU-Attention| 91.02    | 90.93       | 0.923    |
| bert               | 92.59    | 92.57       | 0.926    |
| MECBA              | 94.71    | 94.69       | 0.946    |

It can be seen from Table 2 that the classification effect of CNN model and bigru model is not very different, which indicates that the text features extracted from the embedding part affect the change of downstream model, that is, the text features extracted from the embedding part need to be accurate in order to have an effective impact on the change of downstream model. The poor classification effect of transform model indicates that the downstream model after embedding has an important impact on the effect of text classification.

CNN-BiGRU-Attention model is better than LSTM-CNN model because the introduction of Attention mechanism can focus on more important words and better capture the information hidden in the text, so as to improve the classification performance. Compared with the progressive structure of CNN-BiGRU-Attention, mecba model adopts parallel connection, which solves the problem of gradient disappearance and gradient explosion to a certain extent, and enhances the feature extraction advantages of CNN and LSTM more effectively.

The performance of mecb model in three indicators is better than other models. The reason is that the model can not only effectively obtain the semantic information and local key features of the text, but also focus on the words that have a greater impact on the classification results through the attention mechanism, so that they can play a greater role in the classification process and improve the classification accuracy. At the same time, the mecb model uses ERNIE model based on Knowledge Enhancement for pretraining. Through massive modeling of entity concepts and other prior knowledge in the data, it can enhance the semantic representation of words, better represent the semantic information of words, and significantly improve the effect of text classification. Experimental results show that mecb model has some advantages over the contrast model.

4. Conclusion

In this paper, a Chinese text classification model based on mecb is proposed based on thucnews data set. The model uses Ernie pre training model as the embedding layer, and combines bilstm attention and CNN parallel structure. It can not only pay more attention to the words that have a greater impact on the classification effect, but also effectively alleviate the problems of gradient disappearance and gradient explosion, and further improve the text effect. Compared with the mainstream CNN, bigru, transform and other models, the superiority of mecb model is further proved. However, this model also has some limitations, such as the length limitation of Ernie model, which is not ideal in the task of long text
classification. And the training time is long. Therefore, how to apply the model to long text tasks and reduce the parameters is the direction of future research. Considering that Ernie pre training model itself needs a large number of parameters, this paper mainly studies the optimization method of Ernie pre training model in the future.

Acknowledgments
Thank you teachers and classmates for guidance and help.

Project fund: Research on high precision positioning technology of ice and snow accidents in VR scene based on 5g (Project No.: 20470302d)

Author's contact details:
Xing Zhaoye (the first author) (1994529), male, Han nationality, Zhangjiakou City, Hebei Province, student, master, majoring in natural language processing, Zhangjiakou College of architectural engineering, Hebei Province

Project fund: Research on high precision positioning technology of ice and snow accidents in VR scene based on 5g (Project No.: 20470302d) Post code: 075000

Tel: 15531367113
Micro signal: 15531367113
QQ: 1211422940

Reference
[1] Li Wenhui, Zhang Yingjun, pan Lihu. Short text classification based on improved bilstm network [J]. Computer engineering and design, 2020,41 (03): 880-886
[2] Yu Bing. Short text classification based on attention mechanism [J]. Computer knowledge and technology, 2020,16 (28): 185-186 + 201
[3] Teng Jinbao, Kong Weiwei, Tian Qiaoxin, Wang Zhaqian. Text classification method based on LSTM attention and CNN hybrid model [J]. Computer engineering and application: 1-9 [2021-05-10]http://kns.cnki.net/kcms/detail/11.2127.TP.20210331.1507.016.html.
[4] Niu Yuting, Chen Boqi, Chen Bin. Chinese text classification based on improved ernie-dpcnn model [J]. Journal of Jiangsu Normal University (NATURAL SCIENCE EDITION), 2021,39 (01): 47-52
[5] Niu Yuting, Chen Boqi, Chen Bin. Chinese text classification based on improved ernie-dpcnn model [J]. Journal of Jiangsu Normal University (NATURAL SCIENCE EDITION), 2021,39 (01): 47-52
[6] Xu xukan, Zhou Zeyu. Emotional classification model and application of wechat tweets based on multiscale bilstm CNN [J]. Information science, 2021,39 (05): 130-137
[7] Cao Yukun, Zhao Tian. Based on at. Research on automatic classification of power grid fault information based on CNN and attention bigru fusion network [J]. Computer applications and software, 2021,38 (05): 93-98 + 116
[8] Wei Shangfei, Qiao Baojun, Yu junyang, Yao Xiangyu. Research on sentiment analysis based on word vector fusion of pre training language model [J]. Computer applications and software, 2021,38 (05): 152-157
[9] Yu you, Fu Yu, Wu Xiaoping. A method of news text classification based on revolutionary network of words and event topics [J]. Computer applications and software, 2021,38 (05): 170-174 + 240
[10] Li Dan, Xu Tong, Zheng Yi, Wang Zhefeng, Chen Enhong. Chinese medical named entity recognition based on radical perception [J]. Chinese Journal of information, 2020,34 (12): 54-64