Research on Chinese Error Correction based on Bert

Gehao Liang*, Longxing Lin*, Xiping Lin*, Hengsu Liu*, Yixian Liu*
The Institute of Artificail Intelligence , Xiamen University

Abstract

Text errors are common in everyday life. Among them, the most important text errors can be divided into two types:homophone errors and homograph errors. With the rapid development of science and technology, the Chinese text error correction system based on deep learning develops very fast. The AI intelligent learning machine of iFLYTEK is a representative. In Natural Language Processing based on deep learning,many scholars have proposed various models to correct Chinese text. The Soft-Masked BERT(Bidirectional Encoder Representation of Transformer) model proposed by the research team of Fudan University has achieved good results in Chinese text correction. The MAC BERT(Mask as correction Bidirectional Encoder Representation of Transformer) model proposed by the research institute of Harbin Institute of Technology has changed the text masking strategy of BERT during pre-training, which has also improved the BERT model in Chinese text correction tasks.However,MAC BERT doesn't have the ability to detect which word is wrong and the accuracy of the model doesn't reach our expectation.So based on the above research, We propose our model,which add a detection network to MAC BERT.Our model has some improvement in the SIGHAN2015 compared to baseline models.

Introduction

Text error correction is also an important task of Natural Language Processing in artificial intelligence. As early as the 1960s, some researchers in the UK had conducted research on English text correction. In 1960, Karen Kukich[6] implemented the TYPO English spelling test in UNIX on the IBM/360 and IBM/370. With the development of modern technology, the technology of text proofreading will continue to get breakthroughs. Automatic text correction is the core of automatic text proofreading. It detects the input sentence, determines whether there is a wrong word in the sentence, and assists users to correct it. The quality of the proposed modification to the wrong text is the main index to measure the proofreading ability of an algorithm.

With a long history of culture, Chinese is a relatively complex language in the world, and the grammatical results in Chinese sentences are usually rich. Therefore, it is more challenging to auto-correct Chinese text. In the 1990s, Chinese scholars also began to study the automatic proofreading of Chinese texts. Due to its late start and relatively immature development, there are few discussions on automatic error correction.

Rule-based and statistics-based models are pioneers in the field of Chinese text automatic error correction. The core of a rule-based model is rules. However, in real life, text errors are of various types, and it is difficult to solve all problems with simple rules. Therefore, it is necessary to increase rules constantly, which makes the system increasingly large and difficult to maintain. In the 2017 International Chinese Grammatical Error Diagnosis (CGED) sharing task held by IJCNLP, Zheng's team regarded the Chinese text error correction task as the sequence labeling task. The LSTMCRF[11] model was used to solve the task.In the experiment, they found that the CRF model had higher recall rate and the LSTM model had higher accuracy, so the model with better effect could be obtained by combining the two models. The research team of Beijing Language and Culture University (BLCU) has established a Chinese GEC system[7] based on BPE-level ConvS2S model and four models through integrated decoding technology. In terms of input, they use the BPE algorithm to divide uncommon words and unknown words into Subword units. In the case of Embedding, he used word2vec, which was modified based on the Chinese language, and at the same time connected the location vector to Embedding and the generated words to serve as the location information between the input and output layers. In the first CGED sharing task held in NLPCC2018, NetEase Youdao NLP team combined spelling error correction components and Transformer NMT[4] models with multiple configurations. They take the error correction task as a translation task.Firstly, 5-gram language with similar syllabary is used to deal with low-level errors, and then the word-level and word-level Transformer models are used to deal with higher-order errors. Finally, 5-gram language model is used to calculate the difficulty of the modified statements to select the sentence with the lowest difficulty.In 2019, the advent of BERT[3]changed the NLP neighborhood considerably. Many scholars also began to study the application of Bert in the field of Chinese text error correction. In 2020, a team from Fudan University pro-

posed the Soft-Masked BERT[10] model, which had a more complex network architecture based on the Bert model and achieved good results in correcting errors in Chinese texts. In the same year, the Institute of HIT proposed the Mac BERT[2]model, which was also modified on the basis of Bert model,and achieved amazing results on various NLP tasks.

MAC BERT has the same architecture with BERT, but BERT doesn't have the ability to recognize which word is wrong.So we proposed a model with a detection network based on MAC BERT, which named MAC BERT-change.

Related Work

The traditional text correction method is rule-based. First of all, language experts first summarize the common rules,construct an error correction knowledge base, compare the words that need to be corrected with the knowledge base,judge whether there is an error, and then formulate corresponding rules to correct the error.

In the application direction of text error correction, statistical learning can also play a great role. For example, we use a large amount of data to count the probability of two adjacent words appearing at the same time, and then calculate the co-occurrence probability of each two words in the text that needs to be corrected. The higher the co-occurrence probability, the more likely the word is correct, and vice versa. Therefore, the words with low probability of co-occurrence are labeled as error text. After obtaining the error text, the words that may be wrong are corrected according to the homophone or homograph. N-Gram[1] algorithm is a common statistical machine learning algorithm.In the application of realistic scenes, the most commonly used models are Bi-Gram[8][9] and Tri-Gram[5].

The spirit of the model Bidirectional Encoder Representation from Transformers comes from the Encoder of Transformer. Three main elements of Encoder are Multi-head Attention, Feed Forward and residual. All of them can construct a block, and Encoder can pad many blocks like this in order to achieve a stronger learning ability.As is shown in figure 1,the input of BERT contains three embeddings, namely Token Embedding, Segment Embedding and Position Embedding. Token Embedding can code words from sentences, Position Embedding is the position vector of some certain words in sentences, and Segment Embedding, unique to BERT, used by BERT to discriminate a word belongs to which sentences.Two tasks of BERT in training process are Mask Language Model and Next Sentence Prediction. In Mask Language Model, 15% inputting text will be randomly covered, 80%of covered text will be especially token by [MASK], 10%replaced by other token, and no changes to the other 10%.In the Next Sentence Prediction task, it is necessary to train the model to predict the content of the next sentence from the previous sentence, which is very important in Natural Language Processing task.

In the field of text correction, BERT model improved the performance that is difficult for other models. However,Soft-Masked BERT model, a method of correction of Chinese text based on BERT proposed by Fudan University laboratory, shows a better effect in the field of Chinese text correction.

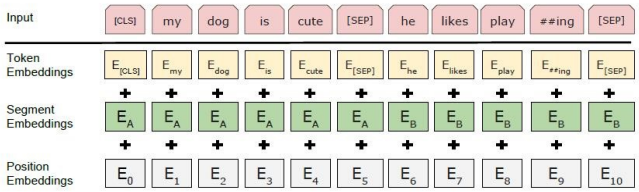

Figure 1: the input of BERT

Soft-Masked BERT is mainly divided into two parts, one of which is the error detection network and the other is the error correction network. The structure of the error detection network is a two-layer GRU network, which is used to locate the error location. The main part of the error correction network is the BERT model, and the input of the model is the output of the processed error detection network. The final output of the whole model is the result of linear addition of the output of the error correction network and the initial input, and then Softmax processing.

Mac BERT, also called MLM as correction BERT, follows the whole architecture of BERT and improves the training objective. In the pre-training task of Mask Language Model, the text masking strategy of the model is improved. There is a big gap between pre-training and finetuning in the Bert model, so the model replaces the characters of mask with words similar to the target word during pre-training. The Mac Bert model also adds an important task to the pre-training: Sentence Order Prediction (SOP),which is the sentence sequence predictor. In the experiment, negative samples were obtained by the sequence of sentences converted into consecutive sentences, and pre-trained the model. Experiment shows that the improvement of Mac Bert on the pre-training task can improve the performance of the model on the downstream task.

Methodology

Overview

We propose a neural network model called MAC BERT-change, as illustrated in Figure 2.MAC BERT-change is composed of a detection network based on Fully Connected Layer and a correction network based on BERT. Next, we describe the details of the model.

Detection Network

The Correction Network is a multi-classification task model based on Mac Bert.The internal structure of the Mac Bert model is shown in Figure 3. The Trm block in figure2 represents the encoder of Transformer.It has three main elements : multi-head attention mechanism, full connection layer and residual connection.The multi-head attention mechanism is an extension of the scaled dot-product attention. Specific calculation is in formula (1).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \qquad (1)$$

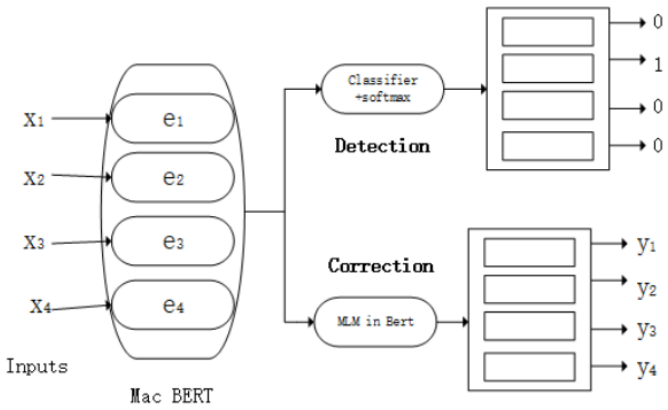

Figure 2: the architecture of MAC BERT-change

K and V in the formula are digital vectors numbered by position, and Q is the output of a certain time node on the module.

The fully connected layer plays the role of classifier in the neural network and feeds the learned feature information back to the data space.It generally uses the Relu activation function. The Relu function is defined as the following formula(2).The model processed by this function has better effect in feature extraction.

$$f(x) = max(0, x) \tag{2}$$

The model uses residual connection, so that Transformer can reach a deeper level and extract deeper features.

Correction Network

The Detection network of the model is a binary classification model, which is composed of a fully connected layer and a SoftMax function to extract sentence features and determine the location of sentence errors.The output of Mac Bert goes through the following formula (3) to get the result.

$$D_n = W_{n1} * O_1 + W_{n2} * O_2 + ... + W_{nn} * O_n + b_n \tag{3}$$

The output of the fully connected layer gets the output of the Detection network via SoftMax.

Learning

When the model is trained, the total loss function includes the Correction part and the Detection part. While fine-tuning, it is assumed that the training data is $\{ (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) \}$ .The Detection network and Correction network loss functions are shown in formulas (4) and formula(5).

$$L_d = -\sum_{i=1}^{n} logP_d(o_i|X) \tag{4}$$

$$L_c = -\sum_{i=1}^{n} logP_c(y_i|X) \tag{5}$$

The final loss function of the whole model is as follows.

$$L = w * L_d + (1 - w) * L_c \tag{6}$$

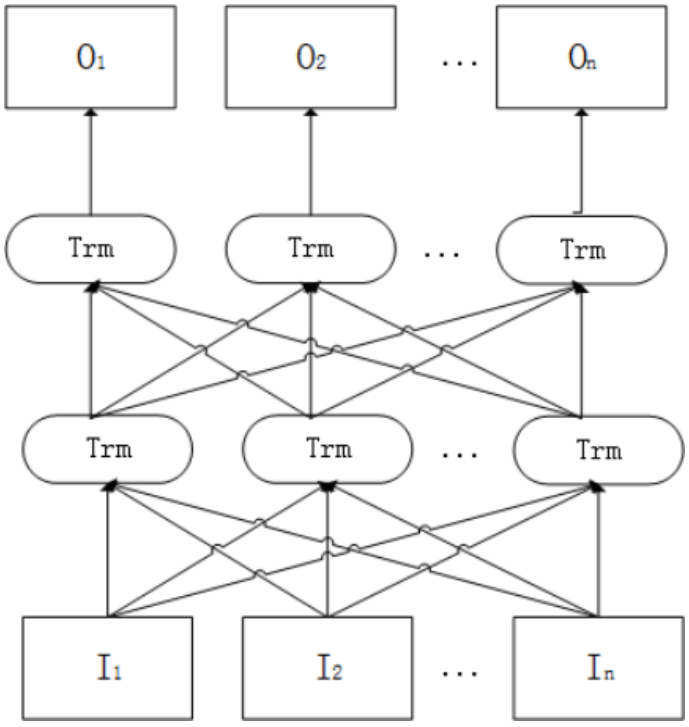

Figure 3: the architecture of MAC BERT

When the model makes predictions, only the weights of the Correction section are used and the weights of the Detection section are not considered.

Experiments

Experimental Setting

Experimental dataset The training dataset is shown in Table 1, which is mainly divided into two parts: error correction dataset of SIGHAN2015 and 270,000 error correction dataset disclosed by 2018ENMLP. In addition, SIGHAN2015 testset and Corpus500 dataset with 500 false sentences are selected as the test dataset

Dataset Name		Dataset Volume(unit:row)
Train Dataset	Wang 271k	271329
	SIGHAN2015	3437
Total		270000+
Test Dataset	SIGHAN2015	1100
	Corpus500	500

Table 1: Experimental Dataset.

Experimental environment The experimental environment of this paper is shown in Table 2 below

Evaluation index The model evaluation index selected in this paper is sentence-level Accuracy, word-level Precision, Recall, and F1 score. The calculation formula of the above four assessment objectives is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

In this equation, TP is the total number of accurately predicted positive samples, FP is the total number of incorrectly predicted positive samples,TN is the total number of accurately predicted negative samples, and FN is the total number of incorrectly predicted negative samples

Experimental environment	Specific information
Operating system	Windows10
CPU	Intel(R) Core(TM) i7-8750H
The graphics card	GeForce RTX 1060
Memory	16.0 GB
Programming language	Python3.9

Table 2: Experimental environment.

Model parameter setting When training the Mac Bert model with modified network structure, the initial learning rate is set to 5E-5, the weight decay parameter is 0.01, the batch data size is 4, and the number of training iterations is 10 rounds. The hyperparameter is 0.3(the weight of the loss function between the detection network and the error correction network)

Experimental Results and Analysis

Comparison of experimental results and analysis The comparative experimental results of different models are shown in Tables 3 and 4. In the comparison experiments, we can see that there is a significant gap between the effect of rule-based model and the effect of pre-trained model in Chinese text error correction. This shows that the pre-trained language model has a strong ability in sentence feature extraction and is able to obtain sentence context information to detect errors in the sentence and correct them. There is also

Method	Acc.	Prec.	Rec.	F1.
BERT-Finetune	0.6060	0.8643	0.4047	0.5513
Soft-Masked Bert	0.6800	0.8714	0.6051	0.0.7142
Mac Bert-change	0.7260	0.9133	0.5987	0.7232

Table 3: Comparative test results of the model on the Corpus dataset.

a large gap in error correction ability between the Mac Bert model after modifying the structure and the Bert model after fine-tuning.Through analysis, we believe that the difference in error correction effect between the Bert model after Finetune and the Mac Bert model with modified structure can be attributed to the improvement of the Mac Bert model on the Bert model's pre-training task. The masking strategy of the Mac Bert model in the pre-training of the text is replaced by [MASK] from the original random selection to replace with homophone or homograph characters, which reduces the model gap between the characters used to cover up in pre-training and fine-tuning, and makes the Mac Bert model have a stronger ability in Chinese text error correction.

Method	Acc.	Prec.	Rec.	F1.
BERT-Finetuned	0.6573	0.8029	0.4052	0.5386
Soft-Masked Bert	0.6964	0.8065	0.5064	0.6222
Mac Bert-change	0.7827	0.8102	0.7311	0.7686

Table 4: Comparative test results on SIGHAN2015 dataset.

When testing the model, by analyzing the situation that the model judged the sentence correctly and incorrectly, we found that the Mac Bert model with the modified structure performed better than the Bert model after fine-tuning on some wrong sentences. For example, the wrong word "man" in the sentence "I can speak a little, but I can't understand a man, so I'm lost" should be changed to "Chinese character". This error cannot be detected by Bert model after Finetune, but Mac Bert with modified structure can detect it. Because this type of error needs to be detected in combination with context information.

The two models with better error correction effects also fail to detect the wrong sentences that involve connections between sentence contexts. For example, the wrong word "heart was very high" in the sentence "he took the girl's hand actively, his heart was very high, and his mouth pretended to be angry" became "cold" after the error correction of the model, and this word should be changed to "happy" in combination with the above sentence. From this result, it can be seen that the current Chinese text error correction model still needs to be improved in the reasoning ability of sentences.

In addition, there is a proprietary vocabulary type of error model that makes some errors when performing error correction. For example, the "Qingge River" of the sentence "woman fell into the Qingge River, all the people rescued" is the name of a river. As long as the word is known to the people who know the river, it is known as the Qingge River, but the model cannot know this aspect of the situation, and the word is mistakenly changed to "Qingyi River".

Therefore, we can see that there is still a lot to be improved in these models in Chinese text error correction.

Experiments on the Impact of hyperparameters At the end of the experiment, the influence of hyperparameters in the model on the performance of the model is studied, and different hyperparameter values are set, such as 0.3, 0.4, 0.5. The other parameters of the model have not changed, and the comparison model performance changes are as follows:

W	Acc.	Prec.	Rec.	F1.
0.3	0.7909	0.8254	0.7311	0.7754
0.4	0.7864	0.8143	0.7348	0.7725
0.5	0.7891	0.8206	0.7330	0.7743

Table 5: Test result on SIGHAN2015 dataset with different hyperparameters.

As can be seen from the above table, when the hyperparameter is set to 0.3, the performance of the model achieves the

best effect in precision and F1 score, but the recall rate is relatively poor.

Conclusion

In this paper, we add the detection network to the MAC BERT model to get better result in Chinese error correction.Our trained model has better performance compared to the baseline models on SIGHAN2015 and Corpus500.However,our model also doesn't correct some sentences because its complex meaning.The further study should be taken to get the understanding of Chinese semantics and sentence cohesion so that better models in Chinese error correction will be found.

References

[1] Peter F. Brown et al. "Class-Based n-Gram Models of Natural Language". In: *Comput. Linguist.* 18.4 (Dec. 1992), pp. 467–479. ISSN: 0891-2017.

[2] Yiming Cui et al. *Revisiting Pre-Trained Models for Chinese Natural Language Processing*. Apr. 2020.

[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[4] Kai Fu, Jin Huang, and Yitao Duan. "Youdao's Winning Solution to the NLPCC-2018 Task 2 Challenge: A Neural Machine Translation Approach to Chinese Grammatical Error Correction". In: *Natural Language Processing and Chinese Computing*. Ed. by Min Zhang et al. Cham: Springer International Publishing, 2018, pp. 341–350. ISBN: 978-3-319-99495-6.

[5] Qiang Huang et al. "Chinese Spelling Check System Based on Tri-gram Model". In: *CIPS-SIGHAN*. 2014.

[6] Karen Kukich. "Techniques for Automatically Correcting Words in Text". In: *ACM Comput. Surv.* 24.4 (Dec. 1992), pp. 377–439. ISSN: 0360-0300. DOI: 10.1145/146370.146380. URL: https://doi.org/10.1145/146370.146380.

[7] Hongkai Ren, Liner Yang, and Endong Xun. "A Sequence to Sequence Learning for Chinese Grammatical Error Correction". In: *Natural Language Processing and Chinese Computing*. Ed. by Min Zhang et al. Cham: Springer International Publishing, 2018, pp. 401–410. ISBN: 978-3-319-99501-4.

[8] Weijian Xie Weijian Xie et al. "Chinese Spelling Check System Based on N-gram Model". In: Jan. 2015.

[9] Weijian Xie et al. "Chinese Spelling Check System Based on N-gram Model". In: *SIGHAN@IJCNLP*. 2015.

[10] Shaohua Zhang et al. "Spelling Error Correction with Soft-Masked BERT". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 882–890. DOI: 10.18653/v1/2020.acl-main.82. URL: https://aclanthology.org/2020.acl-main.82.

[11] Bo Zheng et al. "Chinese Grammatical Error Diagnosis with Long Short-Term Memory Networks". In: *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 49–56. URL: https://aclanthology.org/W16-4907.