Chinese Grammatical Error Detection Based on BERT Model

Yong Cheng Mofan Duan

Faculty of Arts. , Ludong University.

Abstract

Automatic grammatical error correction is of great value in assisting second language writing. In 2020, the shared task for Chinese grammatical error diagnosis(CGED) was held in NLP-TEA. As the LDU team, we participated the competition and submitted the final results. Our work mainly focused on grammatical error detection, that is, to judge whether a sentence contains grammatical errors. We used the BERT pre-trained model for binary classification, and we achieve 0.0391 in FPR track, ranking the second in all teams. In error detection track, the accuracy, recall and F-1 of our submitted result are 0.9851, 0.7496 and 0.8514 respectively.

1 Introduction

With the development on economy and international influence of China,, more and more foreigners begin to learn Chinese. However, Chinese language is one of the most complex and difficult languages in the world, which is hard to master for foreigners. As a result, foreigners may produce a lot of grammatical errors in their written compositions which involve various error types such as Word Redundant Error, Word Missing Error, Word Selection Error, and Word Disorder Error (Gaoqi, 2018). So It has become an important task and challenge for TCFL teachers to help foreign students to detect, understand and correct these grammatical errors.

In recent years, natural language processing technology has been developing rapidly, and the latest deep learning technology has promoted the overall development of artificial intelligence greatly（LeCun, 2015; Schmidhuber, 2015）. In this background, computer-aided education has received more and more attention in NLP area, and one of the representative work is automatic chinese grammatical error diagnosis. In this task, computers are trained to detect and correct the grammatical errors in the compositions written by foreigners. And this work can provide a pretty assist on teaching second language writing.

In 2020, the evaluation on Chinese Grammatical Error Diagnosis (CGED) is held as a shared task in NLPTEA workshop. CGED evaluation has been held for six times, which has greatly promoted the development of related technologies. We participated the CGED 2020 as team LDU. We submitted three runs and achieved the second place in the FPR track. In this paper we will introduce our work in detail.

2 Related Work

Early research on grammar error correction mainly focused on English. And the rule-based approach was popular in early research **(**Michael, 2008; Gabor, 2013**)**, For example, the grammar checker in Mircrosoft Word is a broad-coverage rule-based proofreading system. However, this system was designed for native speakers, and cannot detect errors in texts written by second language learners（Claudia, 2014）. With the development of machine learning technology, the data-driven based approach has become the gold standard method in the field of grammatical error correction（William, 1992; Na, 2006; Kevin, 1994）. Since 2011, four shared tasks are held to evaluate grammar error correction technology, that is the 2011 HOO shared task, the 2012 HOO shared task, the 2013 CONLL shared task, and the 2014 shared task **(**Robert, 2011&2012; Hwee, 2013&2014**),** These shared tasks has greatly promoted the

Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 108–113

development of English grammar error correction approaches. And these approches mainly focus on particular errors such as Article errors, prepositional errors and so on. In the recent 2019 Bea shared task (Christopher, 2019), the focus of the research has moved to the correction of the whole sentence. Meanwhile, the Transformer-based machine translation neural network model has been widely adoptesd in error correction task, which greatly improve the state of the art（Yo, 2019).

In contrast, the research on Chinese grammatical error correction started relatively late. One major driving force in this area is the Shared Task of Chinese Grammatical Error Diagnosis (CGED) organized by Beijing Language and Culture University. It has been held for six times from 2014 to 2020 (Liang, 2014; Lung, 2015-2016; Gaoqi, 2017-2018). The goal of the CGED is to identify the types and positions of grammatical errors in sentences and correct them. In previous studies, the mainstream method is to treat the error correction task as a sequence annotation task, and LSTM, CRF and other models to used to diagnose the error (Chen, 2018). In addition, NLPCC has also held an shared task in 2018 **(**Yuanyuan, 2018**)**, which forcus on the whole sentence correction task, and the sequence to sequence neural network model has been widely used in the participating teams (Kai 2018.).

3 Our Approach

3.1 Task Description

The goal of CGED shared task is to develop NLP techniques to automatically diagnose grammatical errors in Chinese sentences written by Chinese as Foreign Language (CFL) learners. Four error types are defined as redundant words (denoted as a capital “R”), missing words (“M”), word selection errors (“S”), and word ordering errors (“W”). one or more such errors may occur in the input sentences. The developed system should indicate whether the sentence contain any errors, If the inputs contain no grammatical errors, the system should return ”correct”. Otherwhise, the error types and the position in sentence should be returned, and the system can provide their word for the missing word in error “M” and the wrong word in error “S”, Figure1 show the input and system output.

In the evaluation stage, five tracks were set up, including false positive, error detection, error type, error location, S & M error correction. The evaluation results and final ranking were given

INPUT: 即使父母好好指导孩子，如果父母每天玩的话，对孩子的效果也没有。
OUTPUT:
Error1: 26, 26 Missing(M)
Error2: 25, 25 Redundant (R)
Error3: 26, 30 Disorder (W)

INPUT: 现在必须得考虑怎样对待这种社会问题了。
OUTPUT: Correct

Figure 1. eamples of system input and output.

respectively according the tracks. Our team LDU focuses on the first two tracks, false positives and error detection. Which is to detect whether a sentence contains grammatical errors. We regard the error detection task as basic problem of grammatical error correction, so it is worthy of in-depth study. In this next, we will introduce our datasets and models.

3.2 Datasets

In our work, we use three different datasets: HSK dynamic composition corpus (HSK Dataset), language8 corpus (Lang8 Dataset), and primary and secondary school error corpus (School Dataset). We will introduce these three types of corpora respectively.

(1) HSK Dataset

"HSK dynamic composition corpus" (Endong, 2018) is a corpus collected from the HSK advanced writing test for foreigners whose mother tongue is not Chinese. This corpus was collected by Beijing Language and Culture University which contain the composition and corresponding answers of some foreign candidates from 1992 to 2005. The total number of the composition is 11569, with a total of 4.24 million words. The corpus was manually annotated with the error types and corrections from different text levels such as word level, sentence level, and paragraph level. Most of the training sets used by CGED are from the HSK Dataset.

(2) Lang8 Dataset

Lang8 Dataset **(**Yuanyuan, 2018**)** is collected from http://lang-8.com/, a language-learning website where native speakers freely choose learners’ essays to correct. This dataSet are used in the Grammatical Error Correction shared task in NLPCC2018. They collect a large-scale Chinese Mandarin learners’ corpus by exploring “language exchange” social networking services (SNS).

There are about 68,500 Chinese Mandarin learners on this SNS website. By collecting their essays written in Chinese and the revised version by Chinese natives, we set up an initial corpus of 1,108,907 sentences from 135,754 essays.

(3) School Dataset

School dataset is collected from the homework books, composition books, weekly notebooks, diaries, examination papers and so on of the students in primary and secondary schools by Ludong University. The collected data was then processed and annotated manually according to the error types. The total number of records of the corpus is 100631, and the total number of words is more than 3 million and it's all from native Chinese speaker. Dataset not open source for now since it's may not completed.

3.3 Our Model

In this paper, we regard the grammatical error detection task as a binary classification problem to judge whether a sentence contains grammatical errors by training a classification model. On data preprocessing, we transform the dataset D into the form of pairs< T, S>, where S denotes the sentece, and T denote the tags which contains "correct" and "wrong". And In training stage, we adapt the BERT pre-trained model and fine tuning method to train the classification model.

The Bert model (Devlin, 2018) was proposed by Google in 2018. Its full name is Bidirectional Encoder Representation from Transformers. This neural network contains three layers: embedding layer, transformer layer and prediction layer, as in Fig 2. In embedding layer, the input consists of three parts: token embedding, segment embedding and position embedding. The function of transformer layer is to encode the input information. Different from convolutional neural network and recurrent neural network, the main characteristic of the transformer architecture is the use of self attention mechanism to mine and search the hidden relations within text sequences, which can effectively improve the performance of various sequence tasks. Finally, in the prediction stage, the model contains two subtasks, one is to predict the relationship between sentences and the other is to predict the cover words. Specifically, the Bert model extracts two sentences A and B from the dataset, in which the probability of sentence B being the next sentence of sentence A is 50%, and 15% of the words in the two sentences are randomly masked. Then the information is input into the embedding layer and the transformer layer for encoding. Finally, the output of the transformer layer is used to predict the hidden words in sentences A and B, and the probability that sentence B is the next sentence of sentence A.

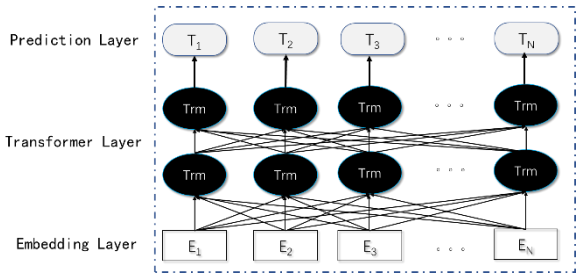

Figure2. Architecture of BERT model

In this paper, we adopt three different pre-training models based on BERT, which named Bert-base (Devlin, 2018), Roberta（Yinhan, 2019) and Roberta wwm（Yiming, 2019). And we train the classification model by fine tuning on the grammatical error corpus. On this basis, the label category of the sentence is predicted. The overall structure of our work is shown in the Figure 3.

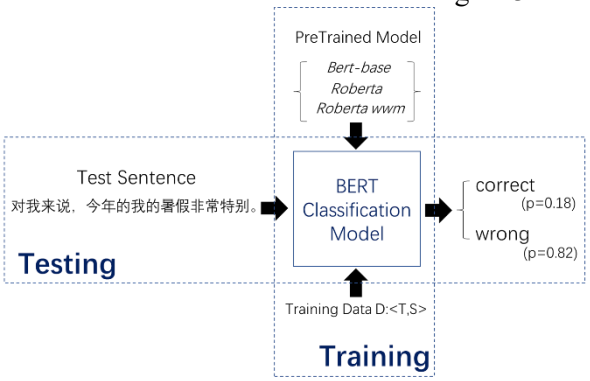

Figure 3. Structure of our work

4 Experiment

4.1 Experiment Setting

As for data, we use four different datasets. The first dataset is provided by CGED committee, in which all the data from 2014 to 2018 are used as training data, and the training set and testing set from 2020 are used as the validation set and the test set in our experiment. The other three datasets are HSK dataset (excluding the duplicate items with CGED dataset), school dataset and lang8 data set. The scale of these four datasets are shown in the Table 1.

Dataset	CGED dataset	HSK dataset	School dataset	Lang8 dataset
Number (sens)	91,967	202,671	25,951	1,786,290

Table 1. Scale of four datasets

Our work focuses on the task of grammar error detection, and we use False Positive Rate (FPR)、Precision(P)、Recall(R) and F-Measure(F-1) as our evaluating indicator.

4.2 Experiment Result

(1) Single Model Comparison

In this paper, we tried three kinds of pre-trained Bert models, named Bert_base, Roberta and Roberta_wwm. in addition, we compare these models with several baseline models based on the TF-IDF features, we use the following classification models: Gauss naive Bayes (gnb), random forest (rf). The results are shown in Table 2.

Result on Validation Set		FPR	P	R	F-1
Base Line	gnb	0.53	0.519	0.571	0.544
	rf	0.568	0.526	0.632	0.574
BERT	BERT-Base	0.134	0.833	0.667	0.741
	RoBERTa	0.131	0.837	0.675	0.747
	RoBERTa -wwm	0.136	0.835	0.688	0.754
Result on Test Set		FPR	P	R	F-1
Base Line	gnb	0.536	0.506	0.550	0.527
	rf	0.604	0.507	0.621	0.558
BERT	BERT-Base	0.052	0.98	0.685	0.807
	RoBERTa	0.065	0.975	0.684	0.804
	RoBERTa -wwm	0.062	0.977	0.691	0.810

Table 2. Experiment result on different models

It can be seen that compared with baseline model, the model based on Bert has a significant improvement in both the validation set and the test set. When comparing different pre-trained models, we can see that the performance of Roberta_wwm is higher than the other two. The FPR and F-1 on verification set are 0.136 and 0.754, and 0.062 and 0.810 on test set. In general, the performance on the test set is better than that on the verification set. It's shown in the first submitted run of LDU.

(2) Experiment on Sample Proportion

In this part, we adjust the proportion of "corret" samples and "wrong" samples in the training set, based on which we can see the influence of the proportion of positive and negative samples on the results. The number of positive and negative samples of the CGED dataset is shown in the table 3. It can be seen that the proportion on training set and validation set is close to 1:1, while the proportion on test set is close to 4:1.

	Training Set	Valid Set	Test Set
Wrong number	44536	1129	1150
Correct number	43716	1129	307

Table 3. Experiment result on sample proportion

On this basis, we adjust the proportion of "correct" and "wrong" samples in the training set by reduce the proportion of correct samples and wrong samples by 10%, 30%, 50%, 70%, 90% respectively, noted as c-10%, c-30%, c-50%, c-70%, c-90%, w-10%, w-30%, w-50%, w-70%, w-90%. On this basis, we use the best Roberta wwm model to carry out experiments on the verification set and test set, the result is shown in Figure 4.

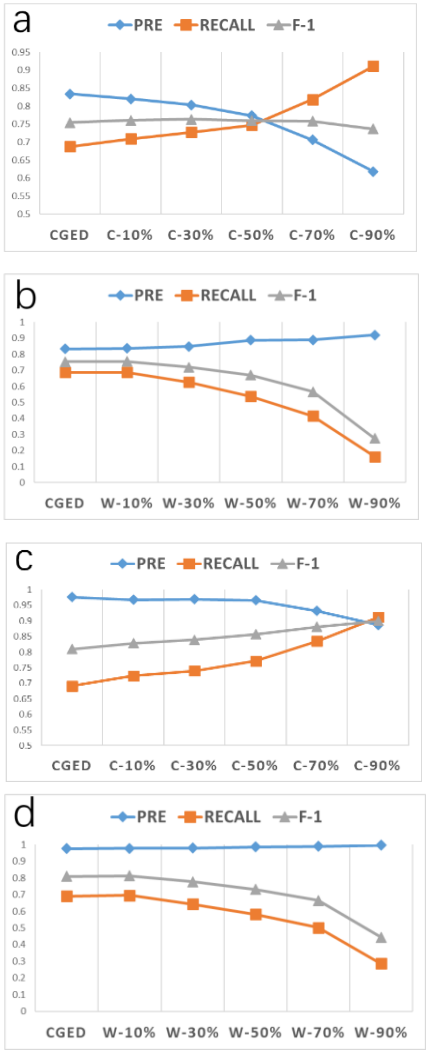

Figure 4. Result on different proportion.(a) Correct samples reduction on validation set. (b) Wrong samples reduction on validation set. (c) Correct samples

reduction on test set. (d) Wrong samples reduction on test set.

From the above results, we can see that when the proportion of "wrong" samples is reduced in both validation set and test set, the precision increases slightly, while the recall decreases rapidly, so the overall F-1 also decreases. However, when the proportion of "correct" samples is reduced, the results are different between validation set and test set. In validation set, The precision decrease rapidly, while the recall increase rapidly, so the overall F-1 value is relatively stable. While in test set, it is obvious that the precision decreases slightly, but both the recall and F-1 value increase rapidly. When the proportion of "correct" lables are reduced to 90%, the F-1 value reaches about 0.9, which is close to the highest level in all the participating teams. For the performance change, We think it is related to the high proportion of false tags in the test set. Therefore, when the proportion of false tags in the training set increases, the overall performance on the test set will increase. It can be seen that the proportion of "correct" and "wrong" samples in the training set has a great impact on the grmmar error detection performance. It's shown in the second submitted run of LDU.

(3) Experiment on Data Augmentation

In this part, we try the data augmentation experiment, three other dataset(hsk, school, lang8) in table are added to the original CGED training set, namely CGED+hsk, CGED+school, CGED+lang8, and we conducted experiments on the verification set and test set respectively, the result is shown in Table 4.

	Valid Set			Test set		
	P	R	F-1	P	R	F-1
CGED	0.835	0.688	0.754	0.977	0.691	0.810
CGED +hsk	0.942	0.702	0.804	0.993	0.624	0.766
CGED +school	0.761	0.791	0.776	0.953	0.815	0.879
CGED +lang8	0.871	0.268	0.410	0.977	0.262	0.413

Table 4. Experiment result on data augmentation

It can be seen that when hsk data is added, the performance on the verification set increases greatly, while the performance on the test set decreases. We believe that the difference between the verification set and the test set is mainly caused by the proportion difference of "correct" and "wrong" samples. When the school data is added, the performance increase slightly both on on the verification set and test set. We think that the CGED test set may contain some homologous errors that primary and secondary school students may also make, so increasing the school data will improve the overall performance. When the lang8 data is added, the performance on the verification set and the test set decrease rapidly. We think lang8 dataset itself has a large number of noise, which is quite different from cged data, so the performance decreases. It's shown in the third submitted run of LDU.

5 Conclusion

This is a technical report of our LDU team participating in CGED2020 shared task. We mainly participated in the task of grammar error detection. We regard the task as a binary classification task, and use different Bert pre-trained models and datasets in our experiments. The experimental results show that the BERT model can greatly improve the accuracy of grammar error detection. And we conclude that the homogeneity and the smaples proportion distribution in training set and test set have a great impact on the final performance.

Acknowledgments

This paper is supported by the research project of Chinese dictionary research center of the state language and writing Commission:"A Study on the Construction of a Hierarchical Retrieval System for Multi-source Sample Sentences for Lexicography Media", whose project number is CSZX-YB-202004.

References

Gaoqi R, Qi G, Baolin Z and Endong X. 2018. Overview of NLPTEA-2018 Share Task Chinese Grammatical Error Diagnosis. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 42–51.

LeCun Y, Bengio Y, Hinton G. 2015. Deep Learning. Nature 521(7553):436-44.

Schmidhuber J. Stockmeyer. 2015. Deep learning in neural networks: An overview. Neural Networks, 6: 85-117.

Michael G, Jianfeng G, Chris B, et al. 2008. Using contextual speller techniques and language modeling for ESL error correction. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), pages 449–456.

Gabor B, Veronika V, Sina Z, et al. 2013. LFG-based features for noun number and article grammatical errors. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pages 62–67.

Claudia L, Martin C, Michael G and Joel. 2014. T. Automated Grammatical Error Detection for Language Learners, Second Edition. Published by Morgan & Claypool.

William G, Ken C and David Y. 1992. Work on statistical methods for word sense disambiguation. In Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, pages 54–60.

Na R, Martin C and Claudia L. 2006. Detecting errors in English article usage by non-native speakers. Natural Language Engineering, 12(2):115–129.

Kevin K and Ishwar C. 1994. Automated postediting of documents. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI), pages 779–784.

Robert D and Adam K. 2011. Helping Our Own: Text massaging for computational linguistics as a new shared task. in Sixth International Natural Language Generation Conference.

Robert D, llya A and George N. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 54–62.

Hwee T, Siew M, Yuanbin W, et al. 2013. The CoNLL2013 Shared Task on Grammatical Error Correction. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pages 1–12.

Hwee T, Siew M, Ted B, et al. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–14.

Christopher B, Mariano F, et al. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 52–75.

Yo J, Jiyeon H, Kyubyong P and Yeoil Yoon. 2019. A Neural Grammatical Error Correction System Built On Better Pre-training and Sequential Transfer Learning. Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 213–227.

Liang C, Lung H, Li C. 2014. Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. Proceedings of the 22nd International Conference on Computers in Education.

Lung H, Liang C, Li C. 2015. Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis. Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 1–6.

Lung H, Liang C, Li C et al. 2016. Overview of NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, pages 40–48.

Gaoqi R, Qi G, Baolin Z and Endong X. 2017. IJCNLP-2017 Task1: Chinese Grammatical Error Diagnosis. Proceedings of the 8th International Joint Conference on Natural Language Processing., Shared Tasks, pages 1-9.

Chen L, Junpei Z, et al. 2018. A Hybrid System for Chinese Grammatical Error Diagnosis and Correction. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 60–69.

Yuanyuan Z, Nan J, WeiWei S, et al. 2018. Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction. In NLPCC 2018, LNAI 11109, pp. 439–445.

Kai F, Jin H, and Yitao D. 2018. Youdao's Winning Solution to the NLPCC-2018 Task 2 Challenge: A Neural Machine Translation Approach to Chinese Grammatical Error Correction. In NLPCC 2018, LNAI 11108, pp. 341–350.

Endong X. 2018. "HSK dynamic composition corpus.", http://hsk.blcu.edu.cn/.

Devlin J, Chang M, Lee K, et al. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In arixv: 1810.04805.

Yinhan L, Myle O, Naman G, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In arxiv: 1907.11692.

Yiming C, Wanxiang C, Ting L, et al. Pre-Training with Whole Word Masking for Chinese BERT. In arxiv: 1906.08101.