Chinese Grammatical Errors Diagnosis System Based on BERT at NLPTEA-2020 CGED Shared Task

Hongying Zan[1,2], **Yangchao Han**[1], **Haotian Huang**[1], **Yingjie Yan**[1,2], **Yuke Wang**[1], **Yingjie Han**[1,2]

[1]School of Information Engineering, Zhengzhou University, Zhengzhou Henan, China
[2]Zhengzhou Zoneyet Technology Co., Ltd.
{iehyzan,ieyjhan}@zzu.edu.cn, hanyangchao@foxmail.com
grenouillehuang@gmail.com, yjyan@gs.zzu.edu.cn, ykwangyoko@163.com

Abstract

In the process of learning Chinese, second language learners may have various grammatical errors due to the negative transfer of native language. This paper describes our submission to the NLPTEA 2020 shared task on CGED. We present a hybrid system that utilizes both detection and correction stages. The detection stage is a sequential labelling model based on BiLSTM-CRF and BERT contextual word representation. The correction stage is a hybrid model based on the n-gram and Seq2Seq. Without adding additional features and external data, the BERT contextual word representation can effectively improve the performance metrics of Chinese grammatical error detection and correction.

1 Introduction

With the improvement of China's international status, more and more foreigners begin to learn Chinese. Unlike English, Chinese grammar lacks morphology and singular and plural changes, and its sentence patterns are flexible and changeable. In learning Chinese, foreigners are prone to introduce grammatical errors due to the complexity of Chinese itself, the negative transfer of mother tongue and target language, and the cultural differences of different countries.

In order to promote the development of automatic detection of syntactic errors in Chinese writing, the Natural Language Processing Techniques for Educational Applications(NLP-TEA) have taken CGED as one of the shared tasks since 2014. Thanks to the CGED task, some research achievements have been made in Chinese grammatical error detection. Based on those previous research results, this paper puts forward a new thinking direction for the CGED task. Some typical examples are shown in Table 1:

TEXT:他们知不道吸烟对未成年年会造成的各种害处。
GED:<3,4,W>,<12,12,S>,<22,23,S>
GEC :他们不知道吸烟对未成年人会造成的各种伤害。

Table1: Typical error example of CGED dataset

CGED has four subtasks:

(1) Detection-level: Binary classification of a given sentence, that is, correct or incorrect, should be completely identical with the gold standard. All error types will be regarded as incorrect.

(2) Identification-level: This level could be considered as a multi-class categorization problem. All error types should be clearly identified. A correct case should be completely identical with the gold standard of the given error type.

(3) Position-level: In addition to identifying the error types, this level also judges the occurrence range of the grammatical error. That is to say, the system results should be perfectly identical with the quadruples of the gold standard.

(4) Correction-level: For the error types of Selection and Missing, recommended corrections are required. At most 3 recommended corrections are allowed for each S and M type error. In this level, the amount of the corrections recommended would need influent the precision and F1 in this level. The trust of the recommendation would be tested.

This paper is organized as follows: Section 2 describes some related works in English and Chinese grammar error diagnosis. Section 3 introduces the hybrid system that we proposed.

Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 102–107
Suzhou, China, December 4, 2020.

Section 4 shows the evaluation and discussion of our system. Section 5 concludes the paper and discusses future work.

2 Related Work

The automatic diagnosis of grammatical errors is a topic of natural language processing. More research on the task of automatic grammatical error recognition focuses on English. In the 1960s, the study of automatic proofreading of English texts was carried out abroad. The HOO (Helping Our Own) (2011) task related to grammatical errors in the task are all about English, which attracts many English grammatical errors researchers. Researchers have proposed a variety of technologies suitable for automatic detection and correction of English grammatical errors, such as rule-based methods (Foster et al., 2004), phrase-based statistical methods (Gamon., 2010), machine learning-based methods (Rei et al., 2016).

However, there are few studies on grammatical errors in modern Chinese. Starting in 2014, the Natural Language Processing Techniques for Educational Applications (NLPTEA) has added modern Chinese grammatical error recognition tasks. These evaluations The task provides a good platform for researchers to showcase their work, and it also speeds up the progress of modern Chinese grammatical errors in automatic recognition methods. At different stages of the development of science and technology, the research methods of modern Chinese grammatical error recognition are different, from rule-based to statistics-based, and then to deep learning-based methods. Zheng (2016) proposed a model based on stacked LSTM and CRF in 2016, which improved the accuracy and recall rate of automatic grammatical error recognition. In the 2017 IJCNLP-2017 CGED evaluation, Yang (2017) proposed a sequence labelling model based on BiLSTM−CRF, which combines the establishment of parts of speech, n-gram grammar, and dependency features, and uses multiple model results to merge and delete After the last 20% of the results are merged, and the results are voted three different integration mechanisms, the effect of automatic grammatical error recognition has been dramatically improved in the F1 value of the three levels, Fu(2018) in the 2018 NLPTEA-2018 CGED evaluation task. Based on the BiLSTM-CRF model, it combines new features such as Gaussian point-by-point mutual information, and adopts multiple model results for probabilistic integration and mixed multiple results ranking. Two different integration mechanisms are introduced in the post-processing.

GEC is typically formulated as a sentence correction task. A GEC system takes a potentially erroneous sentence as input and is expected to transform it into its corrected version. The CoNLL-2014 shared task test set is the most widely used dataset to benchmark GEC systems. The test set contains 1,312 English sentences with error annotations by two expert annotators. Models are evaluated with the MaxMatch scorer, which computes a span-based Fβ-score.

In the NLPCC2018-task2-CGEC (Zhao et al., 2018), the You Dao team (Fu et al., 2018) regards the error correction task as a translation task. Errors are divided into surface errors and grammatical errors. The similar phonetic table and 5-gram language model are used to solve low-level errors, and the Transformer model based on character granularity and word granularity are used to solve high-level errors. Combine the low-level model and the high-level model and finally use the 5-gram language model to analyze the corrected sentence's perplexity and select the sentence with the lowest perplexity. The Ali team(Zhou et al., 2018) adopts a multi-model parallel structure, using three types of models: rule-based, statistics-based, and neural network. First, the low-level combination, which includes one rule based model, two SMT based models, and four NMT based models, obtains the category candidates, and then the high-level combination merges the candidates generated by the low-level combination.

3 System Description

The system proposed in this paper contains two parts: the error detection stage and the error correction stage. The hybrid model presented in this paper is shown in Figure 1.

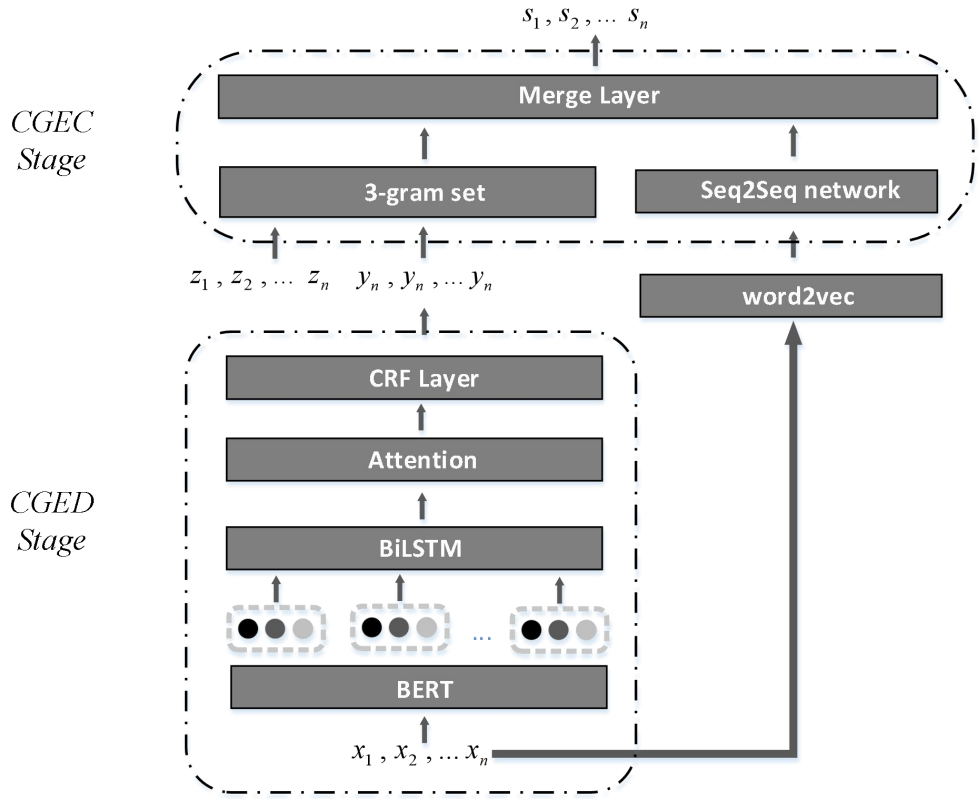

Figure 1: The structure of our system.

First, in the error detection stage, we integrate the BERT method and attention mechanism on the traditional BiLSTM-CRF model. The input is a sequence of characters $\{x_1,x_2,...x_n\}$. The output is the dynamic word vector sequence $\{w_1,w_2,...w_n\}$, after encoding layer and decoding layer, we can get the label sequence $\{y_1,y_2,...y_n\}$ .Then, in the error correction stage, we perform 3-gram extraction based on the corrected sentence sequence $\{z_1,z_2,...z_n\}$ of CGED2016-2018, and construct a quadruple with frequency information. According to the results obtained by the detection stage, we will extract the label sequences containing M or S, merge the error-checking results with the rewrite results of seq2seq, and obtain the final result information $\{s_1,s_2,...s_n\}$.

3.1 Detection Stage

The experimental training data set in this article is a CGED training set that integrates 2016-2018, and the test set is the CGED 2018 test set. First, we need to preprocess the data set. Set the label set to {C, R, M, S, W} to indicate no grammatical error, R type error, M type error, S type error, W type error. According to the grammatical error information marked in the data set, each word is marked with the corresponding label. The processed form is: char, word / POS / dependency / label. The processed data is input into the model for experimentation. After deleting 114 units without control, 21937 units are left for training. Our training set statistics are shown in Table 2.

Training Set	units	errors
CGED2016	10071	24797
CGED2017	10449	26448
CGED2018	402	1067
CGED 2020	1129	2909
Sum	22051	55221
Invalid Data	114	203
Using Data	21937	55018

Table 2: Training set statistics

Example of sentences before processing is shown as follows:

<DOC>
<TEXT id="20030710952310053 8_2_4x1">
农作物也是不例外。
</TEXT>
<CORRECTION>
农作物也不例外。
</CORRECTION>
<ERROR start_off="5" end_off="5" type="R">
</ERROR>

The example of preprocessed data is shown in Table 3.

Char	Word	POS	DEP	Label
农	农作物	B-n	B-SBV	C
作	农作物	I-n	I-SBV	C
物	农作物	I-n	I-SBV	C
也	也	B-d	B-ADV	C
是	是	B-v	B-HED	B-R
不	不	B-d	B-ADV	C
例	例外	B-v	B-VOB	C
外	例外	I-v	I-VOB	C
。	。	B-wp	B-WP	C

Table 3: The example of preprocessed data

BERT embedding layer: The semantic information between sentence sequences in the traditional model is extracted by BiLSTM. The vectors of words in the embedding layer are the same in different semantic environments. This may confuse the semantic information of the sentence. BERT uses a two-way Transformer structure. Transformer uses a multi-head attention mechanism, each layer has the same structure but different weights, each layer focuses on different features, and the overall feature is obtained. It can learn the contextual relationship between texts by paying attention to important information between sequences. Since the model does not pay attention to the sequence order, the position is introduced Information features to strengthen the extraction of location information, making it a deeper understanding of the context. The input is a sequence of characters $\{x_1,x_2,...x_n\}$,through the BERT neural network, the beginning of each sentence is marked by [CLS], and the mark [SEP] is added to the end of each sentence, which means that a sentence embedding is added to each In terms of characters, token embedding, sentence embedding, and transformer position embedding respectively represent character vectors, sentence vectors, and position vectors. In Chinese grammatical error recognition. In the model, the model input is a single sentence. Add a position embedding to each character to indicate its position in the sequence. The output is the dynamic word vector sequence $\{w_1,w_2,...w_n\}$ after BRRT encoding, which is input into the encoding module as a word vector feature.

Encoding layer: The encoding module uses BiLSTM. The input of this module is a sequence of dynamic word vectors encoded by BERT, which can be expressed as $\{w_1,w_2,...w_n\}$. BiLSTM generates the hidden state sequence corresponding to each character by encoding the character vector. The bidirectional LSTM obtains the forward and backward hidden states by reading the sequence from left to right and reading the sequence information from right to left, respectively. The layer output is the splicing of the front and backs hidden states, and the output hidden layer sequence is $\{h_1,h_2,...h_n\}$.

Decoding layer: The decoding module uses BiLSTM-CRF, and at the same time adds an attention mechanism. Although BiLSTM extracts contextual information, there is no correlation between the output sequences. It only predicts the optimal at each moment. In order to capture useful information for the error recognition task, an attention mechanism is added to give different information obtained by decoding. Attention weight. CRF uses transition features to constrain the output sequence and output the final predicted label sequence, where, represents the set of all predicted labels. The output of the CRF layer is the final predicted label, the label set is {C, R, M, S, W}, and each word in the input sequence is labeled with a corresponding label.

3.2 Correction Stage

The experimental data set of the error correction part uses the data set of NLPCC2018-TASK2, which has a total of 717241 sentence pairs. After deleting 123501 data sets without grammatical error and 513 data sets without control, 593227 data sets are left. This paper divides the data set and conducts experiments according to the ratio 10000:10000:573227 of test set, validation set, and training set.

Seq2seq model: This article uses the encoding method for each Chinese character, that is, the sentence is converted into a sequence of Chinese characters, and the Chinese characters are encoded by character vectors, and the word2vec character vector is adopted, and the character vector dimension is 200 dimensions. Input the word vector into seq2seq for training. The optimizer uses rmsprop, and the loss function uses cross entropy. After 40 rounds of training, the model can output the corrected sentence well. The output of the model is shown in Table 4:

Input sentence: 请 把 我 修 改 一 下 ！
Decoded sentence: 请 帮 我 修 改 一 下 ！
Target sentence: 请 帮 我 修 改 一 下

Table 4: The example of seq2seq model's prediction

n-gram model: We extracted 400,490 3-gram combinations from 20,000 correct sentences through the NLTK tool. If the previous word and the next word in the error position are the same as the beginning and end of the triple, the middle word will be the recommended word. Then the model will use the frequency of 3-gram appearance as the answer score, sort according to the score and get the best answer.

4 Experiment Results

CGED evaluation indicators include false positive rate, accuracy, precision, recall rate, F1 value, in order to evaluate the performance of the system at the four levels of grammatical errors.

$$False\ Positive\ Rate = FP/(FP + TN) \quad (1)$$

$$Accuracy = TP + TN/(TP + FP + FN + TN) \quad (2)$$

$$Precision = TP/(TP + FP) \quad (3)$$

$$Recall = TP/(TP + FN) \quad (4)$$

$$F1 = 2 * P * R/(P + R) \quad (5)$$

Table 5 shows the experiments results that the system BERT-BiLSTM-CRF+Correction model performs best among many models. This is because BERT encodes sentences, effectively extracts the dynamic word vector features of sentences, and adds an attention mechanism to the decoding. It further extracts meaningful information from the decoded tags and improves the correction effect.

We also find out that our result didn't perform well in FPR. Because the CGED task belongs to the cost unequal experiment, we should try to increase the cost of marking the sentences with non-error type as error in the experiment instead of treating the cost as the same.

Methods		**BERT-BiLSTM-Attention-CRF+Correction (epoch=100)**	**Char/Word/POS/DEP+BiLSTM-CRF+Correction (epoch=100)**	**Char/Word+BiLSTM-CRF (epoch=100)**
False Positive Rate		**0.6645**	0.6775	0.7394
Detection-level	**Pre.**	**0.8262**	0.8145	0.8136
	Rec.	0.8435	0.7939	**0.8617**
	F1	0.8348	0.8041	**0.8370**
Identification-level	**Pre.**	**0.5856**	0.5053	0.5018
	Rec.	0.4416	0.4127	**0.5060**
	F1	0.5035	0.4543	**0.5039**
Position-level	**Pre.**	**0.2502**	0.0996	0.067
	Rec.	**0.1472**	0.0665	0.0613
	F1	**0.1854**	0.0798	0.0640
Correction-level	**Pre.**	**0.0027**	0.0009	
	Rec.	**0.0012**	0.0004	
	F1	**0.0017**	0.0006	

Table 5: Results on the test data

	Detection Level		
	Precision	**Recall**	**F1**
Run1	0.8262	**0.8435**	**0.8348**
Average of 43runs	0.89	0.78	0.82

Table 6: Performance evaluation in Detection Level

The performance of our hybrid system is shown in the following tables comparing to the average of all 43 formal runs in 2020. Table 6 shows our metrics on detection level. As we expected, BERT-BiLSTM-CRF+Correction model gives the perform well in both recall and F1.

5 Conclusion

Aiming at the problems of the traditional models for automatic recognition of grammatical errors in Chinese, such as the complex features and the large number of model integrations that are difficult to train, this paper proposes a BERT-BiLSTM-Attention-CRF+Correction model that combines the BERT word vector and attention mechanism. Compare it with the multi-feature BiLSTM-CRF and CRF models. The experimental results on the 2020 NLPTEA evaluation data set show that the BERT-BiLSTM-Attention-CRF model performs better than other models we submitted, proving the superiority of BERT word vectors in feature representation.

On the basis of the model proposed in this article, comparing the effects of embedding different pre-trained word vectors on the recognition effect, and how to add a large amount of external knowledge to the recognition model to improve performance are issues worth exploring in our future work.

References

Bo Zheng, Wanxiang Che, Jiang Guo, and Ting Liu. 2016. 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)*, pages 49–56.

Fu, K.,Huang, J.,& Duan, Y. 2018. Youdao's winning solution to the nlpcc-2018 task 2 challenge: a neural machine translation approach to Chinese grammatical error correction. *Lecture Notes in Computer Science*, *vol 11108*. Springer, Cham. Pages 341-350. https://doi.org/10.1007/978-3-319-99495-6_29

Jennifer Foster and Carl Vogel. 2004. Parsing ill-formed text using an error grammar. *Artificial Intelligence Review*, 21(3-4):269–291.

Marek Rei and Helen Yannakoudakis. 2016. *Compositional sequence labelling models for error detection in learner writing. arXiv preprint* arXiv:1607.06153.

Michael Gamon. 2010. Using mostly native data to correct errors in learners′ writing: a meta-classifier approach. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistic*s, pages 163 − 171. Association for Computational Linguistics.

Dale, Robert, and Adam Kilgarriff. 2011, September. elping our own: The HOO 2011 pilot shared task. *In Proceedings of the 13th European Workshop on Natural Language Generation*. pp242-249.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu,and Ting Liu. 2018. Chinese grammatical error diagnosis using statistical and prior knowledge riven features with probabilistic ensemble enhancement. *In Proceedings of the 5th Workshop on atural Language Processing Techniques for Educational Applications*, pages 52–59.

Yi Yang, Pengjun Xie, Jun Tao, Guangwei Xu, Linlin Li, and Luo Si. 2017. Alibaba at ijcnlp-2017 task 1: Embedding grammatical features into lstms for Chinese grammatical error diagnosis task. *In Proceedings of the IJCNLP 2017, Shared Tasks*, pages 41–46.

Zhao Y, Jiang N, Sun W, & Wan X. 2018. Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction: 7th CCF International Conference, NLPCC 2018, Proceedings, Part II. *Natural Language Processing and Chinese Computing. Springer*,pages439-445. https://doi.org/10.1007/978-3-319-99501-4_41

Zhou J., Li C., Liu H., Bao Z., Xu G., Li L. 2018.Chinese Grammatical Error Correction Using Statistical and Neural Models. In: Zhang M., Ng V., Zhao D., Li S., Zan H. *Natural Language Processing and Chinese Computing*. NLPCC 2018. Lecture Notes in Computer Science, vol 11109. Springer, Cham. https://doi.org/10.1007/978-3-319-99501-4_10