Chinese grammatical error diagnosis based on sequence tagging methods

H Wang and Y J Zhang*, X M Sun

School of Computer Science and Software Engineering, University of Science and Technology, Liaoning, China

Corresponding author. E-mail: zhangyujun@ustl.edu.cn

Abstract: In view of the problems such as exploding gradient or vanishing gradient or inefficiency caused by parallel problems when traditional neural networks deal with long text grammar error correction. In this paper, Chinese grammatical error detection is proposed as a sequence tagging problem for named entity recognition, and a Chinese grammatical error detection model based on BERT BILSTM CRF is designed and implemented by using Bert model to construct word vectors. Meanwhile, three groups of comparative tests were designed: CRF, BILSTM CRF and BERT CRF. In the end, BERT BILSTM CRF model achieves the error detection accuracy of 66.67% and 52.94% in M (missing error) and W (ordering error), respectively. BERT CRF and CRF models have also achieved good results in S (selection error) and R (redundant words). The research shows that it is feasible to solve grammatical error detection as a sequence labeling problem. On the other hand, a Chinese grammatical error detection model based on BERT BILSTM CRF is proposed for the first time.

1. Introduction

In summary, the research on text error correction is mainly divided into spelling error correction and grammatical error correction. The focus of spelling error correction is to judge whether the words used here are correct. Grammatical error correction was first based on the judgment of word order until CGED (Chinese grammatical error diagnosis) was proposed in 2014. Most of the subsequent studies are divided into four types according to the task set provided by it, namely word Redundant words(R), Missing words(M), selection errors(S) and ordering errors(W) \[1\]. The proposal of CGED has aroused widespread attention in the field of natural language processing, and has since been leading the research direction of Chinese grammatical error correction. The research in this article is also based on CGED, training and evaluation are carried out according to the task set provided by CGED.

Before the proposal of CGED, the main methods used in Chinese text error correction were based on statistical language models [2] and rules. With the strength of deep learning in various fields, text error correction is gradually combined with deep learning. In 2015, the Baidu research team first proposed to use the BILSTM CRF model to solve the sequence labeling problem [3]. In 2017, the Ali team also used LSTM CRF based model design improvements to detect grammatical errors and won the first place in CGED [4]. In 2018, Youdao's natural language processing team proposed a neural machine translation method to solve the problem of Chinese grammatical error correction [5]. In the 2018 Chinese grammatical error automatic diagnosis contest, 7 teams of 13 submission systems adopted the LSTM CRF model structure, and the HFL team used the model based on the BILSTM CRF structure to achieve the best results [6]. Although there are many research methods for text error correction, text error correction is still one of the task sets with extremely low baseline in natural...
language processing. Taking the final result of CGED2018 as a reference, the highest accuracy rate of Chinese grammatical error correction is only 72.78%. The highest error correction accuracy rate is only 30.77%, and the optimal F1 value is only 0.2527. It can be seen that there is still high room for improvement in Chinese grammatical error correction.

This paper treats text error correction as a sequence tagging task to solve. The primary reason is that the sequence tagging task includes word segmentation, part-of-speech tagging, named entity recognition, keyword extraction, word meaning role tagging and other tasks that have achieved excellent results. Among them, in terms of word segmentation and part-of-speech tagging, at the ACL 2020 conference, the WMSeg model proposed by the Innovation Workshop Greater Bay Area Artificial Intelligence Research Institute has made the F scores for segmentation and joint tagging reach all five types of test sets Above 96.50 \(^7\). The accuracy of the latest research on named entity recognition has basically remained above 93% \(^8\).

In addition, BERT, as one of the commonly used models for sequence labeling, has greatly refreshed its accuracy in 11 directions in the NLP field since its release in 2018. BERT has been applied to word segmentation \(^9\), and BERT BILSTM CRF has been applied to the research of named entity recognition. Both have achieved good results \(^10\). The current research on grammatical error correction still stays on the BILSTM-CRF model \(^11\). Therefore, by optimizing and improving the above models, this paper treats grammatical error correction as a sequence labeling problem to solve, and finally through experiments it shows that it can effectively improve the error detection effect.

In this paper, based on the original CGED task set, it is converted into a sequence labeling format for training. The traditional machine learning method CRF uses the default word embedding method; BILSTM-CRF embeds each word according to unigram, and finally the confidence level of each label is output by the BILSTM layer. BILSTM can better capture the context between semantics, and add constraints to the CRF to further improve the accuracy; BERT-CRF uses BERT to replace the CRF default word embedding method for sequence, not only the training time is shortened, and the fit of the model has achieved a higher effect; BERT-BILSTM-CRF uses BERT to replace the word embedding method of BILSTM, and uses the model design of BILSTM-CRF for further optimization.

2. Task definition
The training set provided CGED the shared task is two separate text files, representing text data and label data respectively. Text data includes error text id and error text content. Label data includes location information and type of text error. In order to adapt to the problem of sequence annotation, the original data set is treated with certain conversion types.

In the converted data set, there are a total of 868,763 rows in the training set, of which 696,202 rows of data are removed without spaces. Each row has a text or symbol and its corresponding label (see the table above). There are a total of 51,303 data containing error type labels, of which R (There are 11,290 entries for word redundancy, 8941 entries for M (missing words), 27,473 entries for S (selection errors), and 14,889 entries for W (word ordering). The remaining tags are all O, which is the correct type. The test set and validation set data are shown in Table 1 below:

| Error type | Quantity(ratio) | Error type | Quantity(ratio) |
|------------|----------------|------------|----------------|}
| R          | 1172(0.50%)    | R          | 1372(1.53%)    |
| M          | 928(0.40%)     | M          | 1109(1.24%)    |
| S          | 2902(1.24%)    | S          | 1419(1.58%)    |
| W          | 1554(0.66%)    | W          | 3487(3.89%)    |
| O          | 227726(97.20%) | O          | 73341(81.74%)  |
3. Methods
Four sets of experiments were designed in this research, namely CRF, BILSTM-CRF, BERT-CRF and BERT-BILSTM-CRF. The first three groups are comparative experiments, and the model is introduced as follows:

3.1. CRF
The full name of CRF is a Conditional Random Field, which is a discriminative probability model. It was proposed by Lafferty et al (2001) [12]. It combines the characteristics of the maximum entropy model and the hidden Markov model, and is an undirected graph model. As a classic model of sequence labeling, each point on the column is treated as a whole during CRF training. The labeling result of each point has a certain dependency, and the training is carried out in the unit of path. Generally, the model can be defined as:

$$P(y|x) = \frac{1}{Z(x)} \exp(\sum_k \lambda_k f_k)$$ (1)

Where $x$ represents the input sentence, $y$ represents the corresponding error type label, $Z(x)$ represents the normalization factor, $f_k$ is set of features, $\lambda_k$ is the corresponding weight.

3.2. BILSTM-CRF
Adding BILSTM on the basis of CRF can better learn the associated information between sentences, but because BILSTM cannot be calculated in parallel, this model will be extremely time-consuming work during runtime.

Among them, the structure of BILSTM not only better learns bidirectional semantic features, but also effectively solves the problem of long-distance dependence of RNN.

BILSTM stands for Bidirectional LSTM, which can process the input sequence from two directions, so that it can effectively avoid problems such as inverse correlation and strong modification in Chinese text, thereby improving the accuracy of the model. After the input sequence passes through the BILSTM layer, the error tag type weight of each word is obtained, and the final result is further determined by adding restriction conditions in the CRF.

3.3. BERT-CRF
BERT is called the Bidirectional Encoder Representation from Transformers [14], in which Transformers is an extension of attention mechanism (Attention) [15]. Attention can be understood as a method. Although the implementation methods used by different researchers are different, the basic theory of the above all follow a paradigm:

$$Attention(K, Q, V) = \text{soft max}(\text{Similarity}(K, Q)V)$$ (2)

Key is the value corresponding to and used to calculate similarity with the query as the basis for Attention selection; Query is a query when Attention is executed once; Value is the data that has been noticed and selected [16].

BERT was released by Google in 2018. Since BERT pre-training takes a long time to complete, it takes about three days on 16 TPUv3 chips. So Google provides developers with pre-trained models that developers can use directly. The development time can also be greatly shortened by fine-tuning [14].

Applying BERT to text error correction is to convert the n*k classification problems processed by BERT into 1*K classification problem on the basis of the sequence annotation model CRF.

As shown in the figure, input sequence $X = (x_1, x_2, ..., x_n)$, predict label $Y = (y_1, y_2, ..., y_n)$, $y_i \in \{1, 2, ..., K\}$ where $K$ represents the number of labels, the score of the sequence can be defined as:
\[ s(X, y) = \sum_{i=0}^{n} A_{yi, y_{i+1}} + \sum_{i=0}^{n} P_{yi, y_{i+1}} \]

\[ y_0 \text{ is the start-label and } y_{n+1} \text{ is the end-label. The model is trained by maximizing the log probability of the correct label sequence:} \]

\[ \log(p(y \mid X)) = s(X, y) - \log(\sum_{y \in Y} e^{s(X, y)}) \]

\[ s(X, y) = \sum_{i=0}^{n} A_{yi, y_{i+1}} + \sum_{i=0}^{n} P_{yi, y_{i+1}} \]

\[ y_0 \text{ is the start-label and } y_{n+1} \text{ is the end-label. The model is trained by maximizing the log probability of the correct label sequence:} \]

\[ \log(p(y \mid X)) = s(X, y) - \log(\sum_{y \in Y} e^{s(X, y)}) \]

3.4. BERT-BILSTM-CRF

BERT-BILSTM-CRF adds a layer of BERT word embedding on the basis of the BILSTM-CRF model. BERT can better capture the left and right semantic features of the sentence, which not only shortens the training time, but also further improves the performance of the model.

Compared with the BERT-CRF model, the bottom layer of BERT is based on the self-attention mechanism, and the self-attention mechanism only inputs position embedding to determine the position information of the token, which will weaken the position information during the calculation process, and add LSTM to learn the observation sequence Dependency, which can effectively improve the model effect. The framework of the BERT-BILSTM-CRF model is to change the BILSTM-CRF word embedding layer to BERT, and the rest remains unchanged.

4. Experiment

4.1. Evaluation Criteria

The test benchmark for this experiment is based on official documents and uses a confusion matrix to measure the following indicators: Precision = TP / (TP+FP); Recall = TP / (TP+FN); F1 = 2*Precision*Recall / (Precision + Recall)

4.2. Experimental results

The experimental results include the accuracy Pre, recall Rec and F1 values of four models on four error types. The detailed data are shown in table 3 below:

| Models     | Redundant words(R) | Missing words(M) | selection errors(S) | ordering errors(W) |
|------------|--------------------|------------------|---------------------|--------------------|
|            | Pre    | Rec    | F1    | Pre    | Rec    | F1    | Pre    | Rec    | F1    | Pre    | Rec    | F1    |
| CRF        | 34.10% | 8.49%  | 13.59%| 25.00% | 4.31%  | 7.35% | 38.64% | 21.24% | 27.38% | 40.19% | 18.46% | 13.59% |
| BILSTM-CRF | 15.20% | 2.10%  | 3.23% | 8.71%  | 2.91%  | 4.36% | 11.86% | 11.78% | 11.76% | 17.92% | 10.01% | 11.75% |
| BERT-CRF   | 25.43% | 15.73% | 19.44%| 0.00%  | 0.00%  | 0.00% | 49.53% | 35.28% | 41.20% | 30.35% | 21.97% | 25.49% |
| BERT-BILSTM-CRF | 27.46% | 6.79%  | 10.88%| 66.67% | 0.44%  | 0.86% | 36.80% | 16.19% | 22.48% | 52.94% | 2.97%  | 5.62% |

4.3. Experimental parameters

The following table lists several parameters that need to be tuned for the BILSTM-CRF model, and the basic parameters of the called BERT fine-tuning model, and the individual parameters of the BERT-BILSTM-CRF that are different from the above two models. In addition, the model is not
optimal, and the following parameters are only the parameters currently adjusted to achieve the effect:

4.4 Experiment Summary
Through the comparison of the final data of the above four model experiments, it can be seen that the detection effect of BERT-BILSTM-CRF on the error types M and W is significantly better than the other three models. In addition, the error detection effect of R and S has also been achieved. Good results. Although the CRF and BERT-CRF models achieve better results on R and S respectively. However, BERT-CRF has the problem of under-fitting. The result of the label prediction of M (missing error) as 0. This type of error is also a common error in grammatical error correction, and it is because the self-attention mechanism only inputs position embedding to determine the location information of the token is caused by weakening the location information in the calculation process.

Compared with the 90% or more effect that named entity recognition has achieved, the text error correction that is also solved as a sequence labeling task has a high room for improvement, which largely depends on the frequency of the label distribution in the training set and test set. When the text error correction is converted to a sequence labeling task, a sentence ranges from about 10-100 words, but there are only 1-5 errors. Too few training tags will cause the model to underfit and affect the final prediction effect. However, compared with the existing research, the error detection effect of the BERT-BILSTM-CRF model on grammatical errors has refreshed the benchmark in this respect.

5. Related work
This paper converts Chinese grammatical error correction into a sequence labeling problem to solve it, and uses multiple models for training and tuning, and the BERT-CRF and BERT-BILSTM-CRF used have never been applied to Chinese grammatical error correction. Gaps in related research.

The training set used in this article is derived from the data set provided by the CGED shared task, which has certain authority. On the CGED2016 test set, the BERT-BILSTM-CRF model achieved 27.46%, 66.67%, 36.80%, and 36.80% on the four types of errors. The accuracy rate of 52.94%, although the R (word redundancy) and S (selection error) are slightly lacking compared to the CRF and BERT-CRF models, the average effect achieved has refreshed the benchmark of related research. In addition, the use of BERT's attention mechanism to replace RNN to train word vectors effectively avoids the problems of gradient explosion and gradient disappearance, and also achieves the effect of parallel computing, which not only shortens the training model time, but also excellently handles long texts. The problem of insufficient interdependence.

6. Summary
Based on CGED's competition system design over the years, this paper uses sequence annotation to solve Chinese grammatical error correction, and uses a model that has achieved excellent results in named entity recognition to improve and optimize it. It is applied to text error correction. Among them, Bert BILSTM CRF has achieved good results in grammar error correction.

This article aims to uniformly apply the optimal sequence labeling model to grammatical error correction, so as to verify whether the grammatical error correction is applicable to the method of sequence labeling. The results show that although the effects of each model are compared with the current baseline of Chinese grammatical error correction, there is a certain improvement and optimization. However, the accuracy of error correction is maintained at about 50%, which cannot reflect the superiority of sequence labeling for solving such problems.

The writer's ability is limited. Although the effect of the improved model has been improved to some extent, there is still a lot of room for improvement. If more corpora is added, the generalization of the model can be further improved, and the addition of an error correction model can further improve the practicability of the system. In addition, the Transformers XL based on the Transformers framework and the XLNet (2019) model similar to BERT, the former allows to learn more than a fixed length of dependencies without breaking time consistency; the latter is a general autoregressive pre-training method, by maximizing the expected possibility of all permutations of the factorization
order to achieve bidirectional context learning. Whether these two models can be applied to text error correction to achieve better results remains to be determined.

In addition, hope to continue to optimize and improve the model in the following study, and further improve the performance of the grammatical error correction model on the premise of a solid foundation.

References

[1] Lee L H, L C Yu, Chang L P. 2015. Overview of the NLP-TEA2015 Shared Task for Chinese Grammatical Error Diagnosis. *Int. Conf. ACL-IJCNLP 2015*. pp 1-6. (Beijing on July 26th to 31st)

[2] Chang T H, Sung Y T, Hong J F, et al. 2014. Knged: A tool for grammatical error diagnosis of Chinese sentences. *Int. ICCE2014*. pp 48-55. (Nara: Nara Prefectural New Public Hall on November 30th to December 4th)

[3] Huang Z H, Xu W, Yu K. 2015. Bidirectional LSTM-CRF models for sequence tagging. (Preprint arXiv:1508. 01991)

[4] Yang Y, Xie P J, Tao J, et al. 2017. Ali-baba at IJCNLP-2017 Task1: Embedding Grammatical Features into LSTMs for Chinese Grammatical Error Diagnosis Task. *Int. Conf. Proceedings of the IJCNLP 2017, Shared Tasks*. pp 41–46. (Taipei: National Taiwan Normal Universityon November 27 to December 1)

[5] Fu K, Huang J, Duan Y. 2018. Youdao’s Winning solution to the NLPCC-2018 Task2 challenge: a neural machine translation approach to Chinese grammatical error correction. *Int. Conf. Proceedings of NLPCC-2018*. pp 341-350. (Hohhot on August 26-30)

[6] Zhou Y, Shao Y, Zhou Y. 2018. Chinese Grammatical Error Diagnosis Based on CRF and LSTM CRF model. *Int. Conf. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. pp 165–171. (Melbourne)

[7] Tian Y H, Song Y, Xia F, et al. 2020. Improving Chinese Word Segmentation with Word hood Memory Networks. *Int. Conf. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp 8274–85. (online from July 5th through July 10th)

[8] Johnson S, Shen S, Liu Y. 2020. CWPC_BiAtt: Character–Word–Position Combined Bi LSTM Attention for Chinese Named Entity Recognition. *Information (Switzerland)*. Vol 11. pp 11-45.

[9] Tian Y H, Song Y, Xiang A, et al. 2020. Joint Chinese Word Segmentation and Part of speech Tagging via Two-way Atentions of Auto analyzed Knowledge. *Int. Conf. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp 8286–96. (online from July 5th through July 10th)

[10] Dai Z, Wang X, Ni P, et al. 2019. Named Entity Recognition Using BERT Bi LSTM CRF for Chinese Electronic Health Records. *Int. Conf. 12th International Congress on Image and Signal Processing, Bio Medical Engineering and Informatics*. pp 1-5. (Shanghai on October 14-16)

[11] RAO G Q, Gong Q, Zhang B L, et al. 2018. Overview of nlptea-2018 share task Chinese grammatical error diagnosis. *Int. Conf. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. pp 42–51. (Melbourne)

[12] John L, Andrew M C, Fernando C N. 2001. "Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data". *Int. Conf. Proceedings of the 18th International Conference on Machine Learning 2001*. pp 282-9. (MIT Press. 55 Hayward St. Cambridge. MA. United States)

[13] Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Computation*. vol 9. pp 1735–80.

[14] Devlin J, Chang M W, Lee K, et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. (preprint arXiv:1810. 04805)
[15] Ashish V, Noam S, Niki A, et al. 2017. Attention is all you need. Int. Conf. Advances in Neural Information Processing Systems. pp 6000–10. (Montreal on December 8-13)

[16] Dzmitry B, Kyunghyun C, Yo S B. 2014. Neural machine translation by jointly learning to align and translate. arXiv. vol 1409.

[17] Guillaume L, Miguel B, Sandeep S, et al. 2016. Neural architectures for named entity recognition. Computing Research Repository. pp 260-270.