Detecting Simultaneously Chinese Grammar Errors Based on a BiLSTM-CRF Model

Yajun Liu, Hongying Zan, Mengjie Zhong, Hongchao Ma
College of Information and Engineering, Zhengzhou University
liuyajun_gz@163.com, iehyzan@zzu.edu.cn
1837361628@qq.com, ma-hc@foxmail.com

Abstract

In the process of learning and using Chinese, many learners of Chinese as foreign language(CFL) may have grammar errors due to negative migration of their native languages. This paper introduces our system that can simultaneously diagnose four types of grammatical errors including redundant (R), missing (M), selection (S), disorder (W) in NLPTEA-5 shared task. We proposed a Bidirectional LSTM CRF neural network (BiLSTM-CRF) that combines BiLSTM and CRF without hand-craft features for Chinese Grammatical Error Diagnosis (CGED). Evaluation includes three levels, which are detection level, identification level and position level. At the detection level and identification level, our system got the third recall scores, and achieved good F1 values.

1 Introduction

With the rapid development of China's economy, "Chinese Fever" has been set off in the world and more foreigners begin to learn Chinese. Writing is an important part of Chinese learning, and the grammar is the basis of writing. In the process of writing and communicating with each other using Chinese, learners of Chinese as foreign language(CFL) may have grammar errors due to negative migration of their native languages.

Traditional learning methods for CFL rely on heavily manual work to point out grammar errors, which costs a lot of time and labor. In order to reduce the workload of manual identification, it is necessary to explore effective methods for Chinese Grammatical Error Diagnosis (CGED). In the field of natural language processing, CGED is a great challenge because of the flexibility and irregularity in Chinese, so a series of CGED evaluation tasks are arranged.

The CGED evaluation tasks provided a platform for many researchers to study the automatic detection of Chinese grammatical errors. The CGED 2018 evaluation task defines Chinese grammatical errors as four categories: redundant(R), selection (S), missing(M), disorder(W). As shown in Table 1, the example sentences corresponding to each error are given.

In this paper, we regarded the CGED 2018 shared task as a character-based sequence labeling task. We proposed a Bidirectional LSTM CRF(BiLSTM-CRF) neural network that combines LSTM and CRF for sequence labeling without any hand-craft features. Firstly, we use BiLSTM network to learn the information in the sentence and extract features, then we utilize CRF for sequence labeling to complete automatically Chinese grammatical errors detection.

Error Type	Error Sentence	Correct Sentence
R(Redundant)	时间是无价之宝**的**。	时间是无价之宝。Time is priceless.
W(Word Order)	你**采取几种方法应该**帮助他们。	你应该采取几种方法帮助他们。 You should take several steps to help them.
M(Missing)	任何婴儿心都是白纸似的清白。	任何婴儿的心都是白纸似的清白。 Any baby's heart is white innocence.
S(Selection)	大家都知道吸烟是**害**健康的。	大家都知道吸烟是损害健康的。 Everyone knows that smoking is harmful to health.

Table 1: The examples given.

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 188–193
Melbourne, Australia, July 19, 2018.

The rest of this paper is organized as follows: Section 2 briefly introduces related work in this field. Section 3 introduces the model that we proposed. Section 4 discusses experiments and results analysis, including data preprocessing, hyperparameters and experiment results. Finally, conclusion and prospects are arranged.

2 Related Work

Automatic detection of grammatical errors is one of the most important tasks in the field of natural language processing. Researchers have already done a lot of work in the field of English grammatical errors diagnosis. For example, Helping Our Own (HOO) is a series of shared tasks in correcting textual errors (Dale and Kilgarriff,2011; Dale et al., 2012). The CoNLL2013 and CoNLL2014 shared tasks (Ng et al., 2013; Ng et al., 2014) focused on grammatical error correction, and many approaches were proposed, such as based N-gram language model methods (Hdez et al., 2014), statistical machine translation methods (Felice et al., 2014), machine learning methods (Wang et al., 2014), etc.

Compared with English, the study for Chinese grammatical errors diagnosis started later. The researchers also proposed many methods, such as statistical learning methods (Chang et al., 2012), ruled-based methods (Lee et al., 2013), and hybrid-based model methods (Lee et al., 2014). However, due to the lack of corpora and the limitations of technology, the research progress is limited greatly. The CGED shared tasks (Yu et al., 2014; Lee et al., 2015, 2016; RAO et al., 2017) provided researchers with a good platform to present their work. In CGED2016 shared task, a CRF-based model achieved good precision (Liu et al., 2016) and a model based on CRF+LSTM get good results (Zheng et al., 2016). In CGED 2017, researchers used some features such as part of speech, collocation words, N-gram etc., and put forward the BiLSTM+CRF model to train models for each error type respectively, then analyzed the errors by model fusion, finally made great progresses for CGED (Xie et al., 2017; Liao et al., 2017).

In this paper, we propose a bidirectional LSTM CRF Neural Network (BiLSTM-CRF) for CGED. The model is described as follows:

(1) Different from the previous methods that train models for each error type, in our system, only one model is trained for all error types, and multiple error types are predicted at the same time.

(2) Our model captures sentence-level features based on the powerful long-term memory ability of BiLSTM and uses CRF for sequence labeling.

(3) The model only learns from word information without any handcraft features.

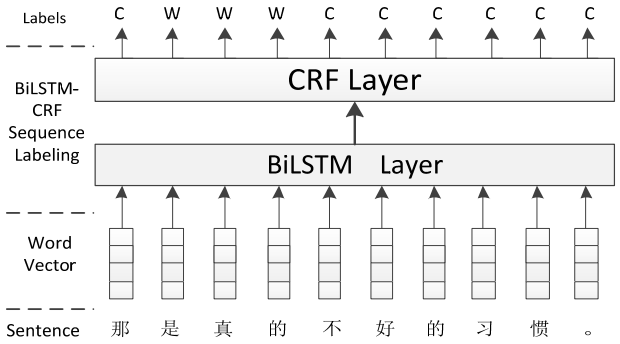

Fig 1 The proposed BiLSTM-CRF model.

3 Model

In this paper, we regard Chinese Grammatical errors diagnosis as the sequence labeling task based on character level, and the tag sets are R (Redundant), S (Selection), M (Missing), W (Word Order), C (Correct). The BiLSTM-CRF model presented in this paper is shown in Figure 1, which includes Embedding Layer, BiLSTM Layer and CRF layer.

(1) Embedding Layer: transforms the index of word into word vector.

(2) BiLSTM Layer: learns the information of each word and extracts features from sentence.

(3) CRF Layer: decodes and produces labels for words.

3.1 Embedding Layer

Embedding Layer aims to transform words into distributed representations which capture syntactic

and semantic meanings of words. Therefore, we use word embeddings to represent words in the sentence.

Given a sentence S, then we can describe it as $S = \{w_1, w_2, w_3, \dots, w_{n-1}, w_n\}$, which contains a sequence of words, and each word is derived from a vocabulary V. Words are represented by distributional vectors $w \in R^d$ which are drawn from a word embedding matrix $W \in R^{|V| \times d}$. After Embedding Layer, then we can get X:

$X = \{x_1, x_2, x_3, \dots, x_{n-1}, x_n\}$.

3.2 BiLSTM Layer

Due to the powerful long-term memory ability of LSTM, LSTM based neural networks, which have access to both past and future contexts, are proven to be effective in sequence labeling task. The hidden states in bidirectional LSTM can capture both past and future context information and accomplish sequence labeling for each token.

Basically, a LSTM unit is composed of three multiplicative gates which control the proportions of information to forget and to pass on to the next time step. Three components composite the LSTM-based recurrent neural networks: one input gate i_t with corresponding weight matrix $W^{(xi)}, W^{(hi)}, W^{(ci)}, b^{(i)}$; one forget gate f_t with weight matrix $W^{(xf)}, W^{(hf)}, W^{(cf)}, b^{(f)}$; one output gate o_t with corresponding weight matrix $W^{(xo)}, W^{(ho)}, W^{(co)}, b^{(o)}$. Formally, the formulas (1) to update an LSTM unit at time t are:

$$\begin{aligned} i_t &= \sigma(W^{(xi)}x_t + W^{(hi)}h_{t-1} + W^{(ci)}c_{t-1} + b^{(i)}) \\ f_t &= \sigma(W^{(xf)}x_t + W^{(hf)}h_{t-1} + W^{(cf)}c_{t-1} + b^{(f)}) \\ u_t &= tanh(W^{(xc)}x_t + W^{(hc)}h_{t-1} + W^{(cc)}c_{t-1} + b^{(c)}) \\ c_t &= i_t \odot u_t + f_t \odot c_{t-1} \\ o_t &= \sigma(W^{(xo)}x_t + W^{(ho)}h_{t-1} + W^{(co)}c_t + b^{(o)}) \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \quad (1)$$

where σ is the element-wise sigmoid function and $\odot$ is the element-wise product. x_t is the input vector at time t, and h_t is the hidden state vector storing all the useful information at (and before) time t.

Mathematically, the input of the BiLSTM layer is a sequence X of word vectors from Embedding Layer, where $X = \{x_1, x_2, x_3, \dots, x_{n-1}, x_n\}$. The output of the BiLSTM Layer is a sequence of the hidden states for each input word vectors, denoted as $h = \{h_1, h_2, h_3, \dots, h_{n-1}, h_n\}$. Each final hidden state is the concatenation of the forward $\overrightarrow{h_t}$ and backward $\overleftarrow{h_t}$ hidden states, then we can get h_t:

$$\overrightarrow{h_t} = lstm(x_t, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = lstm(x_t, \overleftarrow{h_{t+1}})$$
$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$$

3.3 CRF Layer

Since there are many syntactic constraints in natural language sentences, the relationship among adjacent tags is very important for CGED shared task. If we simply transfer directly the hidden states of BiLSTM Layer to a Softmax layer for tag prediction, it is possible to break the syntactic constraints and it is difficult to consider the correlation among adjacent tags. Conditional random field (CRF) is the most commonly used method in structural prediction, and its basic idea is to use a series of potential functions to approximate the conditional probability of the output label sequence for the input word sequence.

The sequence of hidden states in the BiLSTM Layer can be described as $h = \{h_1, h_2, h_3, \dots, h_{n-1}, h_n\}$, then we treat it as the input to the CRF Layer. The output of CRF Layer is our final prediction label sequence, we can see that $y = \{y_1, y_2, y_3, \dots, y_{n-1}, y_n\}$, where $y_i \in Y$ and Y represents the set of all possible label sequences. So we can use the hidden state sequence to get the conditional probability of the output sequence, and the conditional probability is:

$$p(y|h; W, b) = \frac{\prod_{i=1}^{n} exp(W_{y_{i-1},y_i}^T h + b_{y_{i-1},y_i})}{\sum_{y' \in Y} \prod_{i=1}^{n} exp(W_{y'_{i-1},y'_i}^T h + b_{y'_{i-1},y'_i})} \quad (2)$$

Where W, b is the two weight matrices, and the subscription indicates that we extract the weight vector for the given label pair (y_i, y_j). At the same time, in order to train the CRF Layer, we use the classical maximum conditional likelihood estimation to train our model. The final log-likelihood of the weight matrix is as follows:

$$L(W, b) = \sum_{(h_i, y_i)} log\, p(y_i|h_i; W, b) \quad (3)$$

Finally, the Viterbi algorithm is used to train the CRF Layer and decode the optimal output sequence.

4 Experiments and Results Analysis

In this paper, based on the CGED series evaluations, we adopted the dataset of CGED 2016 and CGED 2018 shared tasks as out training dataset, then we manually deleted some incorrect sentenc-

es in the training set and rebuilt the dataset. The CGED 2017 test set was selected as the validation set and the CGED 2018 test set was used as the test set. We selected BiLSTM-CRF model for CGED 2018 shared task. This part mainly includes data preprocessing, parameter settings, results analysis on the validation set and the test set.

4.1 Data Preprocessing

Since the CGED evaluation task involves identification of incorrect boundary positions, word segmentation may cause the misalignment between the end points of words and corresponding error intervals. At the same time, it may also result in overlapping problems among multiple types of errors. Therefore, in this paper we employed characters for Chinese grammatical error diagnosis. Different from previous methods that trained models for each error type, only one model which can identity simultaneously four types of errors is trained in our system.

Using previous data preprocessing method (Liu et al., 2016), we extracted correct sentences and wrong sentences from the corpus according to the manual annotation, and then respectively marked characters with the corresponding labels that include redundant(R), missing(M), selection(S), disorder(W), correct (C). we give some preprocessing examples that are shown in Table 2.

Error sentence: 他们是不但我父母，而且是人生的先辈。 **Correction sentence**: 他们不但是我父母，而且是人生的导师。 (They are not only my parents but also mentors in life.)
Manual annotation: (3,5) W (16,17) S
Preprocessing results: 他/C 们/C 是/W 不/W 但/W 我/C 父/C 母/C，/C 而/C 且/C 是/C 人/C 生/C 的/C 先/S 辈/S。/C 他/C 们/C 不/C 但/C 是/C 我/C 父/C 母/C，/C 而/C 且/C 是/C 人/C 生/C 的/C 导/C 师/C。/C

Table 2: The examples of data preprocessing.

Methods		**CRF**	**BiLSTM-CRF**
False Positive Rate		**0.1881**	0.9643
Detection Level	Precision	**0.7514**	0.6016
	Recall	0.3093	**0.9481**
	F1-Score	0.4382	**0.7361**
Identification Level	Precision	**0.6328**	0.3375
	Recall	0.1763	**0.32**
	F1-Score	0.2758	**0.3285**
Position Level	Precision	**0.3913**	0.0015
	Recall	**0.0658**	0.0009
	F1-Score	**0.1126**	0.0011

Table 3: The results on the validation set.

4.2 Parameter Settings

In this paper, word vector is randomly initialized, and word vector dimension is 50. Here is the overview of optimized parameters:

- Word vector dimension 50
- Hidden size 50
- Adam learning rate 0.001
- Epoch 300

4.3 Experiments Results

In this paper, we use two different models to conduct experiments respectively, which are CRF model (M1) and BiLSTM-CRF model (M2).

CRF model: The CRF model adds a variety of grammatical features such as bigram and trigram features. The selection of features directly affects the performance of the model. Therefore, this experiment adopts the feature length of 7 and uses bigram and trigram to extract features.

BiLSTM-CRF model: The BiLSTM-CRF model combines LSTM and CRF for sequence labeling. Firstly, we use BiLSTM network to learn information in the sentence and extract features, then we utilize CRF for sequence labeling to complete automatically CGED shared work.

The results on the validation set: The valuation set used in this paper is the test set in the CGED2017 shared task. Two different models are

used to conduct experiments on the valuation set, results are shown in Table 3.

From Table 3, we can see that CRF model has lower False Positive Rate (FPR) than BiLSTM-CRF model, and CRF model achieves better precision performance at the detection level and the identification level, because that CRF model has more features information such as bi-gram, tri-gram. However, CRF model and BiLSTM-CRF model are not good at position level. We think that our models are short of identification of position boundary. Next, we will focus on the position level by adding character position features.

The results on the test set: The test set is the test set in the CGED 2018 shared task. We submitted only one result in this task. The Table 4 lists the result Run1 we submitted and the test result based on CRF model.

At the error detection level and error identification level, our system achieves a third recall rate and gets a good F1 value. However, our system has a poor performance at the error position level and FPR. Since our system recognizes four types of errors at the same time, increasing the difficulty of recognition, it is easier to identify a correct sentence as an error sentence, it results in lower FPR performance on the test set. In addition, our system is based on character level, although the BiLSTM network has a powerful long-term memory function, the lack of word collocation information also results in lower position level efficiency. Another reason for low position level efficiency is that tag does not distinguish among locations. For example,

Error: *我/C 朋/C 友/C 的/C 努/C 力/C 真/C 是/C 可/S 看/S 的/C。/C*

Correction: *我朋友的努力真是有效的。*

(My friend's efforts are really effective)

In this sentence, "*可看*" should be corrected as "*有效*". There was no distinction in two "*/S*", so we think it leads to lower position level efficiency.

Methods		**CRF**	**Run1**
False Positive Rate		**0.0851**	0.9309
Detection Level	Precision	**0.8506**	0.5441
	Recall	0.3449	**0.9179**
	F1-Score	0.4908	**0.6926**
Identification Level	Precision	**0.7373**	0.3144
	Recall	0.17	**0.6266**
	F1-Score	0.2763	**0.4187**
Position Level	Precision	**0.5037**	0.0078
	Recall	**0.0615**	0.0189
	F1-Score	**0.1096**	0.0110

Table 4: The results on the test set.

5 Conclusion

On the basis of CGED series evaluation tasks, this paper proposes a neural network model based on BiLSTM-CRF, which is used for Chinese grammatical error detection. It has good effect at the detection level and identification level, especially the high recall rate. But it has low performance at the position level. Next, we will add some external features, such as parts of speech, character position features and collocation features to improve the performance of our system.

References

Chang, Ru-Yng, Chung-Hsien Wu, and Philips Kokoh Prasetyo. 2012. Error diagnosis of Chinese sentences using inductive learning algorithm and decomposition-based testing mechanism. *ACM Transactions on Asian Language Information Processing (TALIP)*, *11*(1), 3.

Dale, Robert, and Adam Kilgarriff. 2011, September. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 242-249). Association for Computational Linguistics.

Dale, Robert, Ilya Anisimoff. 2012, June. 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* (pp. 54-62). Association for Computational Linguistics.

Felice, Mariano, et al. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 15-24).

Gaoqi, R. A. O., et al. 2017. IJCNLP-2017 Task 1: Chinese Grammatical Error Diagnosis. *Proceedings of the IJCNLP 2017, Shared Tasks*, 1-8.

Hdez, S. David, and Hiram Calvo. 2014. CoNLL 2014 shared task: Grammatical error correction with a syntactic n-gram language model from a big corpora. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 53-59).

Lee, Lung-Hao, et al. 2013, November. Linguistic rules based Chinese error detection for second language learning. In *Work-in-Progress Poster Proceedings of the 21st International Conference on Computers in Education (ICCE-13)* (pp. 27-29).

Lee, Lung-Hao, et al. 2014. A sentence judgment system for grammatical error detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (pp. 67-70).

Lee, Lung-Hao, et al. 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* (*NLPTEA'15*), Beijing, China, 31 July, 2015, pp. 1-6.

Lee, Lung-Hao, et al. 2016. Overview of nlp-tea 2016 shared task for chinese grammatical error diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)* (pp. 40-48).

Liao, Quanlei, et al. 2017. YNU-HPCC at IJCNLP-2017 Task 1: Chinese Grammatical Error Diagnosis Using a Bi-directional LSTM-CRF Model. *Proceedings of the IJCNLP 2017, Shared Tasks*, 73-77.

Liu, Yajun, et al. 2016. Automatic Grammatical Error Detection for Chinese based on Conditional Random Field. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)* (pp. 57-62).

Ng, Hwee Tou, et al. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. *Seventeenth Conference on Computational Natural Language Learning: Shared Task*(pp.1-12).

Ng, Hwee Tou, 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* (pp. 1-14).

Wang, Peilu, Zhongye Jia, and Hai Zhao. 2014. Grammatical error detection and correction using a single maximum entropy model. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 74-82).

Xie, Pengjun. 2017. Alibaba at IJCNLP-2017 Task 1: Embedding Grammatical Features into LSTMs for Chinese Grammatical Error Diagnosis Task. *Proceedings of the IJCNLP 2017, Shared Tasks*, 41-46.

Yu, Liang-Chih, Lung-Hao Lee, and Li-Ping Chang. 2014, November. Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 42-47).

Zheng, Bo, et al. 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)* (pp. 49-56).