Chinese Grammatical Error Diagnosis Based on RoBERTa-BiLSTM-CRF Model

Yingjie Han [1,2], **Yingjie Yan**[1,2], **Yangchao Han**[1], **Rui Chao**[1], **Hongying Zan** [1,2]

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

Abstract

Chinese Grammatical Error Diagnosis (CGED) is a natural language processing task for the NLPTEA6 workshop. The goal of this task is to automatically diagnose grammatical errors in Chinese sentences written by L2 learners. This paper proposes a RoBERTa-BiLSTM-CRF model to detect grammatical errors in sentences. Firstly, RoBERTa model is used to obtain word vectors. Secondly, word vectors are input into BiLSTM layer to learn context features. Last, CRF layer without hand-craft features work for processing the output by BiLSTM. The optimal global sequences are obtained according to state transition matrix of CRF and adjacent labels of training data. In experiments, the result of RoBERTa-CRF model and ERNIE-BiLSTM-CRF model are compared, and the impacts of parameters of the models and the testing datasets are analyzed. In terms of evaluation results, our recall score of RoBERTa-BiLSTM-CRF ranks fourth at the detection level.

1 Introduction

The number of foreigners learning Chinese is constantly increasing. Some foreign countries even regard Chinese as their second language. Learners of Chinese as a foreign language (CFL) may make grammatical errors in writing Chinese. And the goal of Chinese grammatical error diagnosis (CGED) shared task is to develop NLP techniques to automatically diagnose grammatical errors in Chinese sentences written by L2 learners. Such errors fall into four categories: redundant words (denoted as a capital "R"), missing words ("M"), word selection errors ("S"), and word ordering errors ("W").

The criteria for judging correctness are determined at three levels as follows. (1) Detection-level: to distinguish whether a sentence contains grammatical errors; (2) Identification-level: to identify the types of those errors type; (3) Position-level: to detect positions where errors occur. The quality of diagnosis is measured by FPR (False Positive Rate), Pre (Precision), Rec (Recall), and F1.

CGED shared task has been held since 2014 (YUa et al.,2014). In CGED of NLP-TEA 2018 (Rao et al.,2018), deep learning models are widely used, LSTM-CRF has been a standard implementation (Fu et al.,2018; Zhou et al., 2018). While, in recent years, pre-training models, such as BERT, XLNET, ERNIE(Sun et al.,2019) and RoBERTa (Liu et al., 2019) achieve good performance in various NLP tasks (Qiu et al.,2020) because of their fast convergence speed and less cost.

This paper proposes a RoBERTa-BiLSTM-CRF model to detect grammatical errors. The model is described as follows:

(1) The RoBERTa model contains general domain data features and fine-tunes the CGED training data to obtain the corresponding word vectors.
(2) The BiLSTM layer captures sentence-level features based on the powerful long-term memory ability, and CRF works for adjusting labels. The CRF layer only learns from word information without any handcraft features.
(3) In this CGED shared task, our model is only used to detect grammatical errors but not correct them.

Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 97–101
Suzhou, China, December 4, 2020. ©2020 Association for Computational Linguistics

2 Models

We regard the CGED task as a sequence labeling task. The illustrative graph of RoBERTa-BiLSTM-CRF is shown in Figure 1. Chinese characters are input into RoBERTa，and RoBERTa converts each character into a one-dimensional vector. Vector T_1, T_2, …T_n fused with semantic features are output.

The BiLSTM layer makes full use of the context information of the input sequence in the sequence labeling task so that it can predict label more accurately.

The CRF layer fully considers the context correlation when predicting the label. More importantly, the Viterbi algorithm of CRF uses the dynamic programming method to find the path with the highest probability. Therefore, it fits better with the task of CGED and avoids illegal sequences, such as ‘B-R’ tag followed by ‘I-R’ tag.

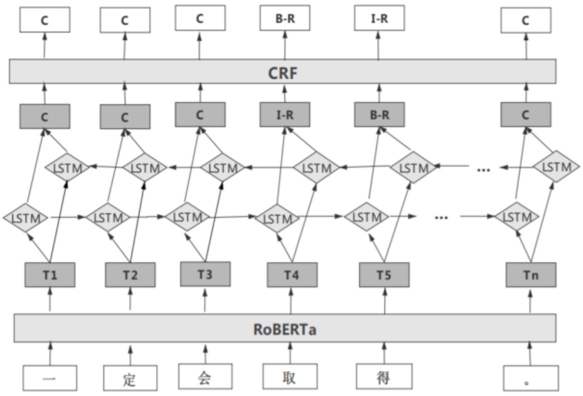

Figure 1: A RoBERTa-BiLSTM-CRF model

2.1 RoBERTa model

RoBERTa model can represent relationships between various words and extract important features in the text. The transformer structure of RoBERTa can get vector representations of sentences from inputting tokens. The RoBERTa model uses a dynamic mask strategy, the model will gradually adapt to different mask strategies processing continuous input data. Compared with training ERNIE model，training RoBERTa model needs larger data sizes and batches. Besides, RoBERTa-large has more network layers and a more complex structure.

2.2 BiLSTM layer

BiLSTM (Bidirectional Long-Short Term Memory) model is composed of a forward LSTM (Long-Short Term Memory) model and a backward LSTM model (Hochreiter et al.,1997). Each word contains information from forward and backward at any time. LSTM model remembers or forgets previous information through the internal gate structure: forgetting gate, memory cell, input gate, and output gate. Figure 2 shows a basic unit of LSTM.

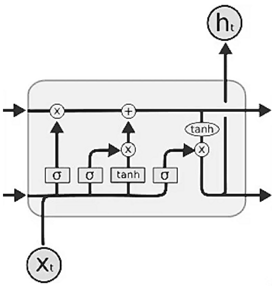

Figure 2: Basic unit in LSTM, it contains forgetting gate, memory cell, input gate and output gate.

1) Forgetting gate as shown in formula (1) selects information to forget, in which h_{t-1} indicates the previous moment; X_t indicates input words, and f_t indicates the output of forgetting data.

$$f_t = (W_t \cdot [h_{t-1}, X_t] + b_f) \quad (1)$$

2) Memory gate selects information to remember, as shown in formula (2), in which i_t indicates the output of the memory gate, and $C_t$indicates the temporary cell’s state shown in formula (3).

$$i_t = (W_i \cdot [h_{t-1}, X_t] + b_i) \quad (2)$$

$$C_t = tanh(W_c \cdot [h_{t-1}, X_t] + b_c) \quad (3)$$

3) Memory cell records cell state C_t in the current moment, as shown in formula (4). The last cell state is C_{t-1}.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \quad (4)$$

4) Formula (5) and (6) show output gate result O_t and state of this moment h_t.

$$O_t = W_o \cdot [h_{t-1}, X_t] + b_o \quad (5)$$

$$h_t = O_t \cdot \tanh(C_t) \quad (6)$$

2.3 CRF layer

The last layer CRF (conditional random field) (Lafferty et al., 2001) are used to learn an optimal path (Liu et al., 2018). The output dimension of the Bi-LSTM layer is tag size, and the score of input sequence X corresponds to the output tag sequence y is defined as formula (7).

$$s(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \quad (7)$$

P represents an output matrix of Bi-LSTM, where P_{ij} represents the non-normalized probability of word w_i mapped to tag_j, and A_{ij} represents the transition probability of tag_i to tag_j.

Softmax function work for defining a probability value $P(y|X)$ as shown in formular (8) for each correct tag sequence y.

$$P(y|X)=\frac{e^{s(X,y)}}{\sum_{\tilde{y}\in Y_X}e^{s(X,\tilde{y})}} \quad (8)$$

In training, maximizing the likelihood probability $P(y|x)$. Therefore, we define the loss function as $-log(P(y|X))$, and then use the gradient descent method to learn the network. It is shown in formula (9).

$$\log(p(y|X))=\log\left(\frac{e^{s(X,y)}}{\sum_{\tilde{y}\in Y_X}e^{s(X,\tilde{y})}}\right)$$
$$=\mathrm{S(X,y)}-\log\left(\sum_{\tilde{y}\in Y_X}e^{s(X,\tilde{y})}\right) \quad (9)$$

3 Dataset

We collect training datasets of CGED2016 (HSK) (Lee et.al, 2016), CGED2017 (Rao et.al, 2017), CGED2018, and CGED2020 as training dataset and validation dataset, with a total of 21938 data units. The ratio of training dataset size to validation dataset size is about 8:2. We adopt the CGED2018 testing dataset as our experimental testing dataset, with a total of 3549 data units. CGED2020 testing dataset has a total of 1457 data units. Table1 shows the number of data units, the number of errors grouped by error types in the training dataset, validation dataset, test2018, and test2020.

	Training dataset	Validation dataset	Test 2018	Test 2020
Units	17461	4476	3541	1457
Errors	42335	10583	5040	3595
R	9507	2377	1119	768
M	10963	2741	1381	816
S	19072	4768	2167	1688
W	3157	789	373	323

Table 1: The number of data units, number of errors and distributions of error types in training dataset, validation dataset, test2018, and test2020.

We segment sentences into separate characters, and tag label for every character. Label 'C' indicates correct character; 'B-X' indicates the beginning position for an error of type 'X' and 'I-X' shows the middle or ending position for an error of type 'X'. Eight kinds of labels in our data: 'B-R', 'I-R', 'B-M', 'B-S', 'I-S', 'B-W', 'I-W', and 'C'. The sample of processed data is shown in Table 2.

Original data format: <DOC> <TEXT id="200505109525100098_2_9x1"> 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。 </TEXT> <CORRECTION> 即使父母好好指导孩子，如果父母每天玩的话，对孩子的教育效果也没有。 </CORRECTION> <ERROR start_off="26" end_off="26" type="M"></ERROR> <ERROR start_off="25" end_off="25" type="R"></ERROR> <ERROR start_off="26" end_off="30" type="W"></ERROR> </DOC>
Processed data format: 即 C\n 使 C\n 父 C\n 母 C\n 好 C\n 好 C\n 指 C\n 导 C\n 孩 C\n 子 C\n， C\n 如 C\n 果 C\n 父 C\n 母 C\n 每 C\n 天 C\n 玩 C\n 的 C\n 话 C\n，C\n 对 C\n 孩 C\n 子 C\n 的 B-R\n 效 I-W\n 果 I-W\n 也 I-W\n 没 I-W\n 有 I-W \n。C\n

Table 2: A data unit sample of original data and processed data, every character has a label.

4 Experiments

4.1 Experimental results and discussions

In the shared task, RoBERTa-BiLSTM-CRF model (Model1) and RoBERTa-CRF model (Model2) are used. Different epochs are set on Model1 and the general parameters of two models are shown below:

- Learning rate 1e-5
- Batch size 16
- Embedding size 1024
- Hidden size 128
- Max length 100

Methods		Model1(epoch=50)	Model2(epoch=50)	Model1(epoch=60)
False Positive Rate		**0.5265**	0.722	0.6933
Detection-level	**Pre.**	0.6817	0.6247	0.6355
	Rec.	0.8896	0.9481	0.9536
	F1	**0.7719**	0.7532	0.7627
Identification-level	**Pre.**	0.5553	0.5274	0.5564
	Rec.	0.5802	0.6412	0.6513
	F1	0.5675	0.5689	**0.6001**
Position-level	**Pre.**	0.3108	0.3078	0.4389
	Rec.	0.2946	0.3129	0.4287
	F1	0.3025	0.3103	**0.4337**

Table 3：Results of three experiments (two models) at three levels on test2018. Model1 represents for RoBERTa-BiLSTM-CRF model, and Model2 for RoBERTa-CRF model

Methods		Run1	Run2	Run3
False Positive Rate		0.8708	0.7557	**0.6938**
Detection-level	**Pre.**	0.8118	0.8182	0.8254
	Rec.	0.9304	0.9078	0.8757
	F1	**0.8671**	0.8607	0.8498
Identification-level	**Pre.**	0.5899	0.6150	0.64
	Rec.	0.5126	0.5076	0.5214
	F1	0.5485	0.5562	**0.5746**
Position-level	**Pre.**	0.29	0.2874	0.2783
	Rec.	0.1941	0.1892	0.2042
	F1	0.2326	0.2282	**0.2356**

Table 4： Results of three runs submitted in shared CGED task. Model2(epoch=50), Model1 and Model2(epoch=60) on test2020.

The following metrics at detection-level, identification-level, and position-level are Pre, Rec, F1, besides an integrated FPR. The results of Model1 (epoch=50; epoch=60) and Model2 on test2018 are shown in Table 3.

F1 scores of Model1 are higher than Model2 at detection-level but lower than Model2 at identification-level and position-level. Since the BiLSTM model learns the dependency relationship between sentences, Model1 may capture error information accurately from the global sequences.

F1 scores of models with larger epoch at identification-level and position-level are higher. This is because larger epoch may lead to overfitting of Model1 at detection-level but not at identification- level and position-level.

Table 4 shows the three runs submitted to the CGED2020 shared task. Run1 is based on the Model1 with 50 epochs; Run2 is based on Model2 with 50 epochs, and Run3 is based on Model1 with 60 epochs.

In this shared task, we get a good recall score of Model1 at the detection-level with bad FPR score. The reason may be as follows. The training corpus of the pre-training model, which comes from news, community discussions, and encyclopedias, is different from the training dataset of CGED, and may easily recognize correct sentences as sentences with grammatical errors.

The performances of three runs on test2020 are consistent with that on test2018 in sum. But F1 scores of three runs on test2020 at detection-level are all higher than that of test2018. According to statistics of errors in Table 1, a data unit contains an average of 1.4233 errors on test2018, while a data unit contains an average of 2.467 errors on test2020. This may lead to diagnosis models more easily to predict whether a sentence contains grammatical errors or not.

4.2 Follow-up experiments and discussions

After the CGED2020-TEA, we use ERNIE-BiLSTM-CRF model (Model3) to do this task. F1 score of Model1 and Model3 on test2018 and test2020 can be seen in Table 5 and Table 6. Model1 gets a worse performance than Model3 at three levels on test2018 but better performance on test2020. The reason is that RoBERTa includes 24 transformers, 16 attention head, and 1024 hidden layer units, which make the generalization ability of RoBERTa-BiLSTM-CRF strong.

	Model1	**Model3**
Detection-level	0.7719	**0.7755**
Identification-level	0.5675	**0.6138**
Position-level	0.3025	**0.4451**

Table 5： F1 scores of Model1 and Model3 on test2018

	Model1	**Model3**
Detection-level	**0.8671**	0.8311
Identification-level	**0.5485**	0.527
Position-level	**0.2326**	0.2153

Table 6： F1 scores of Model1 and Model3 on test2020

5 Conclusion and Future work

This paper proposes a RoBERTa-BiLSTM-CRF model to detect grammatical errors in the CGED shared task. The results of experiments show RoBERTa-BiLSTM-CRF is a good model for detecting grammatical errors in general since RoBERTa model obtains word vector according to data feature, and BiLSTM-CRF captures sentence-level features to predict and adjust labels. In the three runs submitted, our recall ranks fourth at detection- level in the CGED shared task.

In addition, we find that the performance of ERNIE-BiLSTM-CRF is unreasonable on test2020 in our experiments, we will try to pursue reasons from model structure and characters of datasets in the future work.

Acknowledgments

Special thanks to the organizers of CGED 2020 for their great job. We also thank the anonymous reviewers for insightful comments and suggestions.

References

Fu, Ruiji, et al. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. 2018.

Hochreiter, Sepp, and Jürgen Schmidhuber. Long short-term memory. *Neural computation* 9.8 (1997): 1735-1780.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

Lee, Lung-Hao, Liang-Chih Yu, and Li-Ping Chang. 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*. 2015.

Liu, Yinhan, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

Liu, Yajun, et al. Detecting simultaneously Chinese grammar errors based on a BiLSTM-CRF model. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. 2018.

Qiu, Xipeng, et al. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271* (2020).

Rao, Gaoqi, et al. IJCNLP-2017 task 1: Chinese grammatical error diagnosis. *Proceedings of the IJCNLP 2017, Shared Tasks*. 2017.

Rao, Gaoqi, et al. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. 2018.

Sun, Yu, et al. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223* (2019).

YUa, Liang-Chih, Lung-Hao LEE, and Li-Ping CHANG. Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language.

Zhou, Yujie, Yinan Shao, and Yong Zhou. Chinese Grammatical Error Diagnosis Based on CRF and LSTM-CRF model. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. 2018.