Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Multitask Learning for Chinese Grammatical Error Detection

Yu-jie ZHOU[a], Yong ZHOU[b*]
[a,b] *Department of Education Information Technology, East China Normal University, Shanghai*
*yzhou@ied.ecnu.edu.cn

Abstract: Chinese as a Foreign Language (CFL) learners often make grammatical errors such as missing words, selecting wrong words and wrong word order due to language negative migration. In this paper, we propose a neural sequence labeling model with a supplementary objective for Chinese grammatical error detection. We use the manually labeled dataset written by CFL learners to train the models. This multitask learning model has better performance than other sequence labeling model because it can learn the bias in the label distribution and learn richer features for semantic composition.

Keywords: Grammatical error detection, Chinese as a Foreign Language, Multitask learning

1. Introduction

The number of Chinese as a Foreign Language (CFL) learners has continuously increased these years. Unlike English or some other languages, Chinese sentences are composed with the string of characters without spaces to mark word boundaries. Also, Chinese has quite flexible expressions and loose structural grammatical, so it has been regarded as one of the most difficult languages in the world (Bo Zheng et al., 2016). CFL learners often make grammatical errors such as missing words, selecting wrong words and wrong word order due to language negative migration, over-generalization, teaching methods, learning strategies and other reasons.

Automated Grammatical error detection and correction system are very essential and invaluable to language learners because manual correction is time-consuming and laborious (Leacock et al., 2010). GEC for English has been studied for many years, with many shared tasks such as CoNLL-2013 (Ng et al., 2013) and CoNLL-2014 (Ng et al., 2014), however, few grammatical correction applications have been developed to support CFL learners because of the limited labeled data and the complexity of Chinese. The exist Chinese grammatical errors detection applications are based on a range of techniques, such as statistical learning (Chang et al, 2012; Wu et al, 2010; Yu and Chen, 2012), rule-based analysis (Lee et al., 2013) and Deep Learning-based models (Gaoqi et al., 2017; Lee et al., 2016, 2015; Yu et al., 2014).

In this paper, Chinese grammatical errors detection is considered as a sequence labeling problem which assigns each Chinese word in a target sentence with a tag indicating the error types. With limited labeled data, we use multitask learning to solve this problem. To be specific, we propose a neural sequence labeling architecture which consists a supplementary objective of predicting surrounding words in addition to labeling each token to encourage the framework to learn richer features for semantic composition without requiring additional training data.

2. Related Work

Grammatical error detection is a sub-task of sequence labeling in natural language processing which assigns semantic label to each word of the input sentence. Our work builds on previous research exploring sequence labeling model on grammatical error detection.

The researchers used many different methods to study the grammatical error detection task and achieved good results (Tou et al., 2017). As for English, Marcin et al., used phrase based translation optimized for F-score using a combination of kb-MIRA and MERT with augmented language models and task-specific features, and got a good result(Junczys-Dowmunt and Grundkiewicz, 2014). As a universal language model, the Long Short-Term Memory network (LSTM) has achieved good results in many tasks in natural language processing in recent years, including text classification tasks, machine translation tasks, and sequence annotation tasks(Hochreiter and Schmidhuber, 1997). Rei et al., used the Encoder-Decoder model similar to neural machine translation to process the English Grammatical(Rei, Yuan and Briscoe, 2017).

Compared with English, the research time of Chinese grammatical error diagnosis system is short. In recent years, the Natural Language Processing Techniques for Educational Applications (NLPTEA) workshops have hosted a series of shared tasks for Chinese grammatical error diagnosis. Some researchers have done some works based on the given dataset. Bo et al., propose a CRF+BiLSTM model based on character embedding on bigram embedding, on the CGED-HSK dataset of NLP-TEA-3 shared task, their system presents the best F1-scores in all the three levels (Zheng et al., 2016).Ruiji et al., improved the model of bidirectional Long Short-Term Memory with a conditional random field layer (BiLSTM-CRF) with several new features and adopted a probabilistic ensemble approach. This model achieved the best F1 scores on NLPTEA-2018 CGED(Ruiji et al., 2018). Chen et al., proposed a two-stage hybrid system which combined the BiLSTM-CRF model along with some handcraft features and three GEC models which achieved the highest precision(Chen et al., 2018).

3. Task Definition

Table 1

Two errors are found in the sentence below, one is word selection error (S) at positon 8, the other is word ordering error (W) from position 9 to 12.

国^{1} 家^{2} 不^{3} 应^{4} 该^{5} 盲^{6} 目^{7} 的^{8} 经^{9} 济^{10} 发^{11} 展^{12} 。^{13}		
Error Type	S	W
Error Position-start	8	9
Error Position-end	8	12
Correction	国家不应该盲目地发展经济。	

The grammatical errors made by CFL learners are varied. The NLPTEA shared task for Chinese Grammatical Error Diagnosis (CGED) which has been hosted for years has divided those errors into four types, including redundant words (denoted as R), missing words (M), word selection errors (S), and word ordering errors (W) (Gaoqi et al., 2017,2018). The goal of this task is to detect these four types of grammatical errors. The input sentence may contain one or more grammatical errors. Example sentence and corresponding notes are shown in Table 1.

4. Methodology

In this section, we introduce the proposed multitask neural CRF sequence labeling model (MTN-CRF). First, we will introduce the sequence labeling model used in our main task. Then we will introduce the neural language model used in our auxiliary task.

4.1 Bi-LSTM grammatical error detection model

Our Bi-LSTM grammatical error detection model can efficiently use past input features via an LSTM layer. The model is shown in Figure 1, we build Bi-LSTM blocks (h1) for each input word, and concatenate these block in two directions to form the forward Bi-LSTM and backward Bi-LSTM. These Bi-LSTM blocks are parameter sharing.

Compared with traditional grammatical error detection models, Bi-LSTM based model mainly have two advantages. The first is that Bi-LSTM model can capture the context features for each word. For example, consider the input sentence “我 中文 学” in Figure 1, when predicting the tag for the Chinese word “中文”, the Bi-LSTM model can using the features from the whole sentence (i.e., “我” and “学”). While traditional models usually cannot capture such features since the data sparse issue and the computation is costly. For a given input length n, the Bi-LSTM model can learn such features in time complexity O(n), while traditional models will have a time complexity O(n2). The second advantage is that Bi-LSTM model is an end to end model, which can extract features automatically. In traditional models, the model performance largely depends on the feature engineering. Doing such feature engineering is exhaust for the system developer.

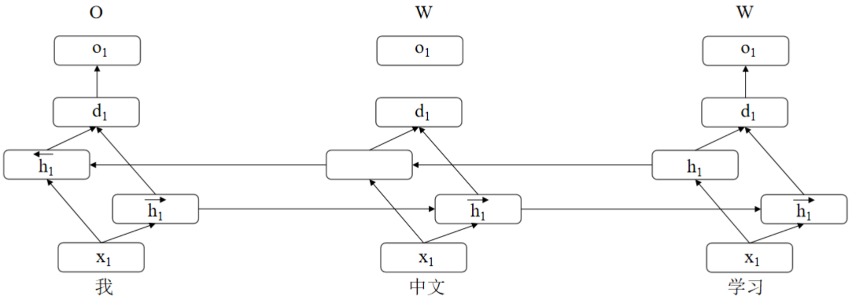

Figure 1. Bi-LSTM grammatical error detection model

4.2 Bi-LSTM language models

In this work, we also combine our grammatical error detection model with a Bi-LSTM language model to enhance the model performance. The Bi-LSTM language model is shown in Figure 2. The training of the Bi-LSTM is a unsupervised learning. It takes an natural language text as input, and models the probability of the next word at each time step. Consider the example in Figure 2, the input of the Bi-LSTM language model is a sentence without any labels (i.e., “我 学 中文”). For the first time step, the input of the LSTM unit is “我”, and the output of the Bi-LSTM is the probability of the next word. Suppose the word vocabulary is 3, includes “我”, “学”, “中文” , then the output of the first time step will be key-value pairs like (“我”: 0.2), (“学” : 0.5), (“中文” : 0.3). The key is the vocabulary, the value is the probability to occur at next time step. If the model is correctly trained, the word “学” will have a higher probability then others. This model is quite similar to the traditional n-gram language model in natural language processing. They both models the probability distribution of the next time step given each input word. In practice, the performance Bi-LSTM language model is better than the traditional n-gram language model.

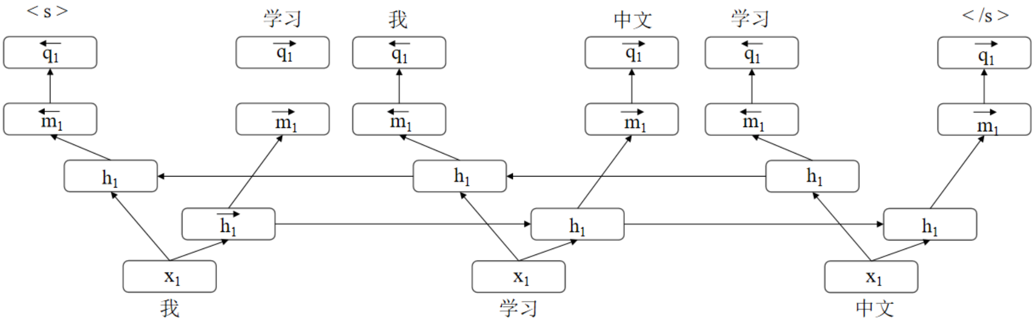

Figure 2.Bi-LSTM language models

4.3 Multitask learning framework

In this section, we will introduce how we combine the grammatical error detection model with the language model to form the multitask learning framework. The multitask learning framework is shown in Figure 3.

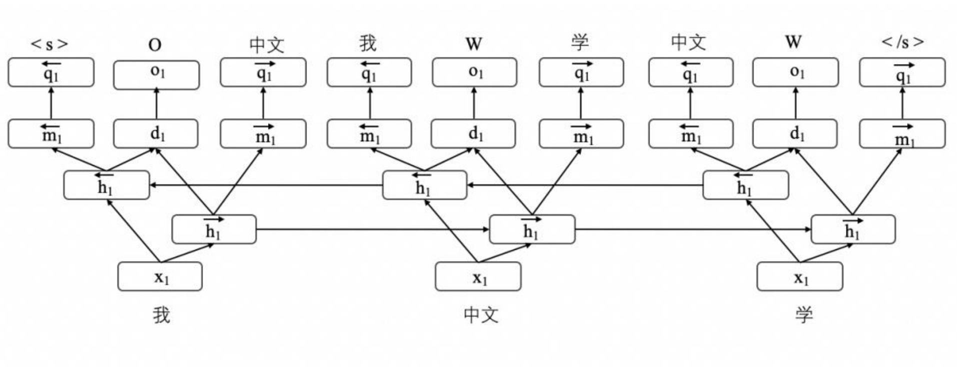

Figure 3. Multitask learning framework

The LSTM units are sharing on this two tasks. The output of the LSTM units contains both language model outputs and the grammatical error detection outputs. That is to say the model needs to predict both surround words and the grammatical error type at the same type. The loss of the mode is as bellow:

$$\text{Loss}_{\text{all}} = \text{Loss}_{\text{lm}} + \alpha\, \text{Loss}_{\text{ged}}$$

Where Lossall is the joint loss of the multitask learning framework, Losslm is the loss of the language model, and the Lossged is the loss of the grammatical error detection model, α is the weight parameter of the Lossged which we set 1, and automatically tuned through training of the multitask learning framework.

5. Experiments and Evaluation

5.1 Dataset

The dataset we use was from the NLPTEA (Natural Language Processing Techniques for Educational Applications) shared task for Chinese Grammatical Error Diagnosis. The corpora used in CGED task were taken from the writing section of HSK (Pinyin of Hanyu Shuiping Kaoshi, Test of Chinese Level) which is for Chinese as a Foreign Language(CFL) learners. The data set which contains the origin sentence, the error type, the position of the error and the correction of the sentence is manually labeled by the experienced teachers.

Table 2 shows the distributions of error types in the training set, validation set and testing set. The ratio of training set size to validation set size is about 10:1. Besides the sentences with grammatical errors, there are over 40% of the sentences contain no error which was simulated the sampling in the writing sessions in HSK to test the performance of the systems in false positive identification.

For the supplementary objective, we use an external dataset Lang-81 to train the model, which contains more than 700,000 items, and each item consists of an original sentence and corresponding corrected sentences.

Table 2

The distributions of error types in datasets

	#R	#M	#W	#S
Training Set	10671(22.55%)	11955(25.26%)	3516(7.43%)	21178(44.75%)
Validation Set	574(21.95%)	682(26.08%)	171(6.54%)	1188(45.43%)
Testing Set	795(21.97)	928(25.64%)	281(7.76%)	1615(44.63%)

5.2 Results and discussion

The evaluation of the test results is determined three levels including detection-level, identification-level and Positon level. Detection-level is binary classification of the given sentence, that is, correct or incorrect, should be completely identical. All error types will be regarded as incorrect. Identification-level could be considered as a multi-class categorization problem to identify the error type. Position-level judges the occurrence range of the grammatical error. Table 11 Table 12 and Table 13 shows the evaluation result for detection level, identification level and position level of multitask learning architecture on grammatical error detection datasets. The two baseline results are from our previous work in 2018 CGED shared task (Yujie et al., 2018).

Table 3

Evaluation results of sequence labeling architectures on Detection Level

Detection Level			
	Precision	Recall	F1
Baseline			
CRF	0.5923	0.5445	0.5993
BiLSTM-CRF	0.8202	0.5652	0.6692
Our Work			
Multitask learning	**0.8314**	**0.5932**	**0.6924**

Table 4

Evaluation results of sequence labeling architectures on Identification Level

Identification Level			
	Precision	Recall	F1
Baseline			
CRF	0.4452	0.2740	0.3392
BiLSTM-CRF	0.6068	0.4183	0.4952
Our Work			
Multitask learning	**0.6342**	**0.4723**	**0.5414**

Table 5

Evaluation results of sequence labeling architectures on Positon Level

Positon Level			
	Precision	Recall	F1
Baseline			
CRF	0.3532	0.1346	0.1949
BiLSTM-CRF	0.4631	0.2568	0.3303
Our Work			
Multitask learning	**0.4735**	**0.2984**	**0.3661**

According to experiment results, we found that Multitask learning model has better performance. The CRF baseline is low, because CRF model largely depends on feature engineering. It is hard to do feature engineering in grammaticaltical error detection, because the training data is sparse. And it is also difficult to find certain feature to capture a specific error type. The BiLSTM-CRF model performs slightly better than CRF model, since it can automatically extract features for CRF models rather than handcraft feature engineering. But it still suffers from the data sparse issue. The multitask learning model performs better than Bi-LSM-CRF on all three level. This is because the sequence labeling model is only optimized based on the labels contains information. While in our test set, over 40% of the sentences in the test set contain no error and 84% of all tokens have the label O (correct). So many of the tokens in the dataset contribute very little to the training process. Multitask learning architecture which contain a supplementary objective is able to learn this bias in the label distribution without obtaining much additional information from the majority labels. It allows the model to make full use of the training data and get better results than other sequence labeling task in Grammatical error detection task.

6. Conclusion and Further Work

In this paper, we propose a neural sequence labeling model with a supplementary objective for Chinese grammatical error detection. We use the manually labeled dataset written by CFL learners to train the models. This multitask learning model has better performance than other sequence labeling model because it can learn the bias in the label distribution and learn richer features for semantic composition. For further work, we plan to address more complex errors in addition to the four-main error type in this paper and focus on Chinese grammatical error correction which may involve Machine Translation models.

References

Lafferty J, Andrew M, and Fernando P, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the eighteenth international conference on machine learning, In proc, ICML, 2001.

Lample G , Ballesteros M , Subramanian S , et al. Neural Architectures for Named Entity Recognition, In proc. NAACL, 2016.

Ma X, and Hovy E, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, In proc. ACL, 2016.

Marek Rei, Semi-supervised Multitask Learning for Sequence Labeling, In proc. ACL, 2017.

Helen Yannakoudakis, Marek Rei, Øistein E. Andersen and Zheng Yuan. 2017. Neural Sequence-Labelling Models for Grammatical Error Correction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2795–2806

Gaoqi Rao Qi Gong Baolin Zhang Endong Xun. 2018. Overview of NLPTEA-2018 Share Task Chinese Grammatical Error Diagnosis. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 42–51

Grundkiewicz, R., & Junczys-Dowmunt, M. (2014). The WikEd Error Corpus: A Corpus of Corrective Wikipedia Edits and Its Application to Grammatical Error Correction. Advances in Natural Language Processing.

Hochreiter, S. , & Jürgen Schmidhuber. (1997). Flat minima. Neural Computation, 9(1), 1.

Rei, M., Felice, M., Yuan, Z., & Briscoe, T. (2017). Artificial error generation with machine translation and syntactic patterns.

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, pages 49–56.

Yujie Zhou, Yinan Shao, Yong Zhou. 2018. Chinese Grammatical Error Diagnosis Based on CRF and LSTM-CRF model. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 165–17.

Po-Lin Chen, Wu Shih-Hung, Liang-Pu Chen, and Ping-Che Yang. 2016. Improving the selection error recognition in a Chinese grammatical error detection system. International Conference on Information Reuse and Integration, pages 525-530.

Yajun Liu, Yingjie Han, Liyan Zhuo, and Hongying Zan. 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, pages 57–62.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese language technology platform. In Proceedings of the Coling 2010: Demonstrations, pages 13-16.

Zhiheng Huang, Wei Xu and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. Computer Science. Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of ACL, pages 1064-1074.