Chinese Grammatical Error Diagnosis Based on Policy Gradient LSTM Model

Changliang Li
Kingsoft
lichangliang@kingsoft.com

Ji Qi
Communication University of China
ji.qi@cuc.edu.cn

Abstract

Chinese Grammatical Error Diagnosis (CGED) is a natural language processing task for the NLPTEA2018 workshop held during ACL2018. The goal of this task is to diagnose Chinese sentences containing four kinds of grammatical errors through the model and find out the sentence errors. Chinese grammatical error diagnosis system is a very important tool, which can help Chinese learners automatically diagnose grammatical errors in many scenarios. However, due to the limitations of the Chinese language's own characteristics and datasets, the traditional model faces the problem of extreme imbalances in the positive and negative samples and the disappearance of gradients. In this paper, we propose a sequence labeling method based on the Policy Gradient LSTM model and apply it to this task to solve the above problems. The results show that our model can achieve higher precision scores in the case of lower False positive rate (FPR) and it is convenient to optimize the model on-line.

1 Introduction

In English and many other languages , the space is a good approximation of a word divider (word delimiter), a sentence separated by spaces into multiple words. Unlike the English, Chinese does not have a separator on the written scripts, a sentence consists of Chinese characters that are next to each other, where sentences but not words are delimited. This is very difficult for the machine or learner without a Chinese foundation to analyze Chinese grammar, because it first has to face the problem of Chinese word segmentation (Xue, 2003). Compared to English, Chinese has neither singular/plural change, nor the tense changes of the verb, and it uses more short sentences but less clauses. In addition, the same word may express different meanings in different contexts, namely ambiguity. All these problems make learning Chinese very difficult. Most non-native Chinese language learners usually need professional Chinese teachers to guide them and correct grammatical errors. However, online teaching has recently become the main channel for language learning, which requires the system to automatically diagnose and give advice to a large number of learners' grammatical errors. Therefore, the study of Chinese grammatical error automatic diagnosis system is very important. The goal of Chinese Grammatical Error Diagnosis (CGED) is to build a system that can automatically diagnose errors in Chinese sentences. Such errors are defined as redundant words (denoted as a capital "R"), missing words ("M"), word selection errors ("S"), and word ordering errors ("W"). Evaluation includes three levels, which are detection level, identification level and position level.

At present, most methods regard the Chinese grammatical error diagnosis task as a sequence labeling task (Settles and Craven, 2008), such as using a conditional random field construction sequence labeling model (Lafferty et al., 2001) and a sequence labeling model constructed using LSTM (Hochreiter and Schmidhuber, 1997). However, the characteristics of Chinese language leads to a obvious problem in constructing Chinese grammatical error diagnosis model, which is the imbalance between positive and negative samples. For example, a sentence to be labeling is: "人战胜了饥饿，才努力为了下一代作更好的、更健康的东西。" , The correct labeling result should be: "NNNNNNNNPNNNNNNPNNNNNNNNNNN",

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 77–82
Melbourne, Australia, July 19, 2018.

where N denotes a negative label, ie there is no wrong label, P denotes a positive label, ie there is a wrong label. We can see that the proportion of positive and negative sample labels in a not very long sentence is seriously unbalanced, in the above example, the ratio is 2:27, which is a serious problem faced by the Chinese grammatical error diagnosis model. In order to solve the above problems, we propose a Policy Gradient-based model to tag Chinese sentences. Similar to the recent work, we also use the LSTM model to handle this task as a sequence labeling problem (Zheng et al., 2016). Moreover, we use the Policy Gradient method to deal with the imbalance of positive and negative samples. The results show that our method can achieve better results.

This paper is organized as follows. Section 2 introduces some related work . Section 3 briefly describes the CGED Shared Task. Section 4 illustrates our methodology, including data preparation, model description and the details of policy gradient method. Section 5 shows the experiment settings and results. And finally, section 6 concludes the paper and presents future work.

2 Related works

The English Grammatical Error Correction task has been held for two consecutive years as one of the natural language processing tasks of the Conference on Computational Natural Language Learning (CoNLL). The researchers used many different methods to study the task and achieved good results (Tou et al., 2017). where (Junczys-Dowmunt and Grundkiewicz, 2014) usesd 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. As a universal language model, the Long Short-Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997) 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. (Yuan and Briscoe, 2016) used the Encoder-Decoder model similar to neural machine translation to process the English Grammatical error correction Task and achieved good results. Compared with English, the research time of Chinese grammatical error diagnosis system is short, the data sets and effective methods are lacking. (Yu and Chen, 2012) uses the CRF-based model to construct a Chinese word ordering error detection model and obtains a higher accuracy on the experimental data set. In recent years, Chinese grammatical error diagnosis has been cited as a shared task of NLPTEA CGED. Many researchers in the field of natural language processing have researched and proposed several effective methods (Yu et al., 2014; Lee et al., 2015, 2016). HIT 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).

3 CGED Task Description

The goal of The NLPTEA CGED task is to use a model to perform a grammar diagnosis on a data set containing Chinese sentences, these datasets are written by Chinese Foreign Language (CFL) leaner. These datasets contain the following four errors, such errors are defined as redundant words (denoted as a capital ”R”), missing words (”M”), word selection errors (”S”), and word ordering errors (”W”). The input sentence may contain one or more such errors, and there may also be no errors. The developed system should indicate which error types are embedded in the given sentence and the position at which they occur. Some typical examples are shown in Table 1:

Sentence	人战胜了饥饿，才努力为了下一代作更好的、更健康的东西。
Correction	人战胜了饥饿，才能努力为了下一代做更好的、更健康的东西。
Errors	9, 9, M, 能 16, 16, S, 做

Table 1: Typical Error Examples

Table1 shows the CGED shared task input data and output data samples. Each sentence contains a single id, each output error contains the sentence id, and the number in Errors indicates the index of the error location. The criteria for judging correctness are determined at three levels as detection level, identification level and position level.

4 Methodology

In this section, we will introduce our entire process of the CGED task, including data preprocess-

ing, model construction, and the construction of objective functions based on the Policy Gradient. Same as previous work, we treat the CGED task as a sequence labeling problem. Such as given a sentence x, our model generates a corresponding label sequence y. Each label in y is a token from a specific tag set. We use "O" to indicate the correct character's tag, 'B-X' indicating the beginning positions for errors of type 'X' and 'I-X' as middle and ending positions for errors of type 'X'.

First, we will introduce our CGED task data preprocessing process, including Bigram feature construction, POS data annotation, and data label settings. Second, we will introduce the construction of the ensemble model that combines Biggram feature, POS feature, and character embedding. Finally, we will introduce the idea and mathematical formula of the objective function based on the Policy Gradient.

4.1 Data Preparation

First, we use the Word2vec tool to train the Bigrams of all Chinese sentences in the data set into word vectors. These word vectors will be used to generate input sentence features during model building. we first convert the original character sequence to a bigram sequence. Then we can train bigram embeddings readily using word2vec (Mikolov et al., 2013) on the resulting bigram sequences.

We use the Part-of-speech (POS) feature to improve the performance of the system. Therefore, we use the part-of-speech (POS) feature to generate a corresponding POS tag sequence for each Chinese sentence sequence of the data set, B-pos indicating the beginning character's POS tag while I-pos indicating the middle and end characters'.

We define each character in the sentence as a separate tag that contains the character's position in the word. We use "O" to indicate the correct character's tag, 'B-X' indicating the beginning positions for errors of type 'X' and 'I-X' as middle and ending positions for errors of type 'X'. In the CGED task, we will get 8 labels: B-W, I-R, B-R, B-M, I-S, I-W, B-S, O. After the data is preprocessed, each sample can be represented as the structure shown by Table 2. The input of each sample during training is composed of three parts as shown in the inputs features of Table 2, and the label sequence of each sample is composed of 8 pre-defined labels.

4.2 Model Description

We regard the Chinese grammatical error diagnosis task as a sequence labeling task, and first use LSTM to construct a sequence labeling model. LSTM network is a variant of recurrent neural network (RNN) and have better ability to capture long term dependencies. Given a sequence of input vectors $X = x_1, x_2, \ldots, x_T = \{x_t\}_1^T$, a recurrent unit $\mathcal{H}$ computes a sequence of hidden vectors h=$h_1, h_2, \ldots, h_T = \{h_t\}_1^T$ and a sequence of output symbols $\hat{Y} = \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T = \{\hat{y}_t\}_1^T$ by iterating the following equations,

$$h_t = \mathcal{H}(x_t, h_{t-1}) \quad (1)$$

$$\hat{y}_t = argmax(softmax(W_{hy} h_t)) \quad (2)$$

where $softmax(z_m) = e^{z_m} / \sum_i e^{z_i}$, The LSTM recurrent unit $\mathcal{H}$ represents the calculation process of the LSTM network. A typical LSTM network consists of input gates, oblivion gates, output gates, and memory cells. Which input gate controls the current time step which information will be input into the memory cell, the forgotten gate controls the current time step which history information will be forgotten by the memory cell, and the output gate controls which information will be output as h_t according to the current memory cell state. Each gate consists of a sigmoid neural net layer and a point-wise multiplication operation.

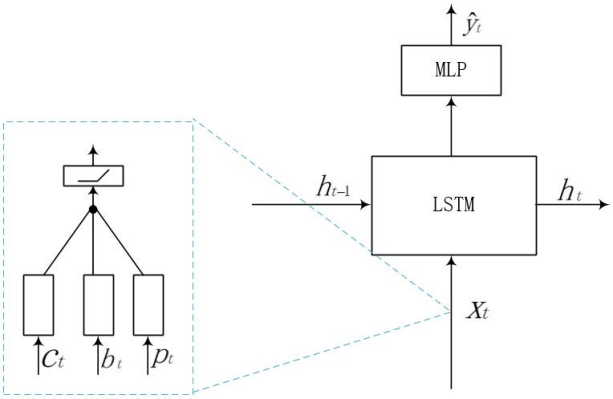

Figure 1: An illustration of LSTM model

In this work, we denote the input of the time step i as:

$$x_t = \boldsymbol{\sigma}(W_x(c_t, b_t, p_t)) \quad (3)$$

Where $\boldsymbol{\sigma}$ represents the nonlinear activation function, c_t is the character embeddings that are initialized immediately, b_t represents the bigram

Sentence	我根本不能了解这妇女辞职回家的现象。
Char	我 根 本 不 能 了 解 这 妇 女 辞 职 回 家 的 现 象 。
Bigram	<s>我 我根 根本 本不 不能 能了 了解 解这 这妇 妇女 女辞 辞职 职回 回家 家的 的现 现象 象。 。</s>
POS	B-r B-d I-d B-d B-v B-v I-v B-r B-n I-n B-v I-v B-v I-v B-u B-n I-n B-wp
Label	O O O O O B-S I-S B-R O O O O O O O O O O

Table 2: A snapshot of our training data after the pre-processing

vector of the current time step, and p_t represents the POS discrete feature. These three simple features are combined as the input vector for the time step t. The ensemble model is shown in Figure 1.

4.3 Policy Gradient

Deep Reinforcement Learning (DRL) is divided into Value-Based Deep RL (Mnih et al., 2015) and Policy-Based Deep RL (Lillicrap et al., 2015) in terms of implementation[16]. Value-Based Deep RL is a Neural Network usually used as a Q function to estimate the return of an action which can be obtained in the current environment, namely Deep Q-network (DQN). Such as (Mnih et al., 2013) present the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using reinforcement learning, model is a convolutional neural network, trained with a variant of Q-learning. The Policy-Based Deep RL is Represent policy by deep network with weights u, as shown below:

$$\mathrm{a} = \pi(\mathrm{a}|\mathrm{s},\mathrm{u}) \quad \text{or} \quad \mathrm{a} = \pi(\mathrm{s},\mathrm{u}) \tag{4}$$

Where π is the policy expressed by the neural network and u is the network learning parameter. Define objective function as total discounted reward:

$$\mathrm{L}(\mathrm{u}) = \mathbb{E}[\mathrm{r}_1 + \gamma \mathrm{r}_2 + \gamma^2 \mathrm{r}_3 + ...|\pi(\cdot,\mathrm{u})] \tag{5}$$

$\mathrm{L}(\mathbf{u})$ denotes the objective function, $\mathrm{r}_1, \mathrm{r}_2, ...$ denotes the returns obtained in each step. In this paper, the value of the return of the tagged result of each token is indicated. $\gamma \in [0, 1]$ is the discount factor, which indicates the importance of future returns. In this article we set $\gamma = 0.9$. To make high-value actions more likely, the gradient of a stochastic policy $\pi(\mathrm{a}|\mathrm{s},\mathrm{u})$ is given by:

$$\frac{\partial \mathrm{L}(\mathrm{u})}{\partial u} = \mathbb{E}[\frac{\partial \log \pi(\mathrm{a}|\mathrm{s},\mathrm{u})}{\partial \mathrm{u}} \mathrm{Q}^{\pi}(\mathrm{s},\mathrm{a})] \tag{6}$$

Where Q^{π} is a function value that measures the return of each action. In this article, we define that the return value of the tag "O" is successfully marked as 1, and the return value of the failed tag is -1. Defining all other error labels "B-W, I-W, B-M, I-W ..." is marked with a score of 10 for a successful return, and a return of -10 for a failed tag. Finally, update parameters u by stochastic gradient ascent. Our ensemble model is shown in Figure 2.

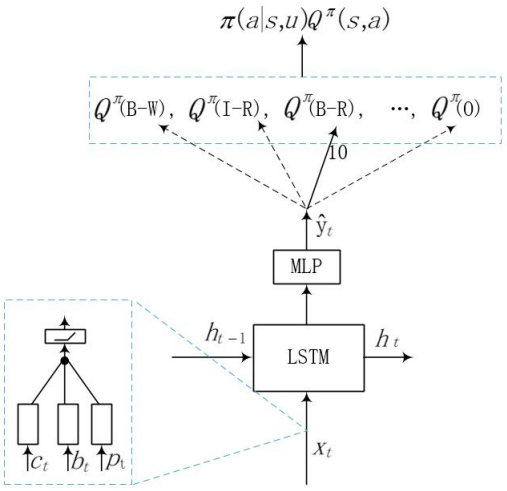

Figure 2: An illustration of Policy Gradient-based LSTM model

Where $Q^{\pi}(\mathrm{X})$ represents the reward after label "X" was tagged, for example, the "X" is "B-R", $\hat{y}_t$ represents the policy obtained by the network. Finally, the final output $\pi(\mathrm{a}|\mathrm{s},\mathrm{u})\mathrm{Q}^{\pi}(\mathrm{s},\mathrm{a})$ of the network is obtained with the policy π and reward Q known. This output is used to calculate the policy gradient $\frac{\partial \mathrm{L}(\mathrm{u})}{\partial u}$, and then the gradient is used to update the network parameters.

5 Experiments

In this section, we introduce the entire process of the experiment. First of all, we introduce the use of data sets and division, and then briefly introduce the CGED experimental results evaluation

method. Finally, we introduce the results on the validation dataset and the results from the evaluation dataset based on our proposed model.

5.1 Dataset and criteria

During the training of the model, we use the collection of training set of CGED2017 and training set of CGED2018 as the training dataset. In CGED2017 training set, provide 10,449 training units with a total of 26,448 grammatical errors, categorized as redundant (5,852 instances), missing (7,010), word selection (11,591) and word ordering(1,995). In the CGED2018 training set, contain total of 1,067 grammatical errors, categorized as redundant (208 instances), missing (298), word selection (87) and word ordering(474). In addition, use CGED2017's test set as the validation set during training, it's contain total of 4,871 grammatical errors, categorized as redundant (1,060 instances), missing (1,269), word selection (2,156) and word ordering(386). Table 3 shows the data distribution in the training data.

	R error	M error	S error	W error
Train	6060	7308	11678	2469
Validation	1060	1269	2156	386

Table 3: Data statistics

The criteria for judging correctness are determined at three levels, (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. The False Positive Rate(FPR), Accuracy (Acc), Precision (Pre), Recall (Rec) and F1 score(F1) are measured at all levels with the help of the confusion matrix.

5.2 Experiment results

We use the above data partitioning to train and converge the training set based on our proposed Policy Gradient-based model, the trained model was tested on the validation set and evaluation set.

5.2.1 Results on Validation Dataset

We refer to the model's results on the validation dataset and select the best hyper-parameters model. Table 4 shows the results.

5.2.2 Results on evaluation Dataset

We testing on the final evaluation dataset for CGED2018 test set, the result showing with table 5. As we can see, our model can obtain better identification score and position score while obtaining a better detection level score.

Our model obtains good results at three levels, and the Policy Gradient-based model can be easily applied to online tasks to optimize the network structure through continuous interaction and attempting to obtain maximum rewards.

5.3 Conclusion and Future Work

This paper proposes a method based on policy gradient applied to NLPTEA 2018 CGED shared task. We use the value function method of deep reinforcement learning to map the labeling results to rewards to solve the problem of imbalanced positive and negative samples in Chinese grammatical error diagnosis. Moreover, our system can be applied to online optimization as easily as a depth-enhanced model. In this paper, we verify the effectiveness of the Policy Gradient through experiments on the validation dataset and the evaluation dataset.

In the future, we hope to betterly solve the problem of serial labeling with imbalanced positive and negative samples in Chinese grammatical error diagnosis through deep reinforcement learning strategies. In terms of Policy Gradients, we hope to be able to define reward functions that are more in line with the mission requirements and optimize the entire network. In addition, we hope to optimize the network through multiple rounds of online annotation results and further conduct relevant online experiments. Ultimately, the network can achieve good labeling results while also being able to cope with the challenges posed by online data changes.

References

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Model runs	Detection Level			Identification Level			Position Level		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
1	0.64	0.26	0.37	0.45	0.28	0.35	0.17	0.02	0.03
2	0.71	0.47	0.52	0.48	0.17	0.25	0.21	0.01	0.02

Table 4: Results on Validation Dataset

Model runs	Detection Level			Identification Level			Position Level		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
1	0.6349	0.4232	0.5079	0.4792	0.1995	0.2817	0.1185	0.0442	0.0644
2	0.6698	0.2494	0.3634	0.5139	0.1323	0.2105	0.1585	0.0331	0.0547
3	0.6346	0.5426	0.5850	0.4735	0.2646	0.3395	0.1129	0.0609	0.0792

Table 5: Results on Evaluation Dataset

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The amu system in the conll-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33.

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

Lung-Hao Lee, RAO Gaoqi, Liang-Chih Yu, XUN Endong, Baolin Zhang, and Li-Ping Chang. 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)*, pages 40–48.

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang. 2015. Guest editoral: Special issue on chinese as a foreign language. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 1, June 2015-Special Issue on Chinese as a Foreign Language*, 20(1).

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1070–1079. Association for Computational Linguistics.

Ng Hwee Tou, Wu Siew Mei, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2017. Conll-2014 shared task: Grammatical error correction.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, 8(1):29–48.

Chi-Hsin Yu and Hsin-Hsi Chen. 2012. Detecting word ordering errors in chinese sentences for learning chinese as a foreign language. *Proceedings of COLING 2012*, pages 3003–3018.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. 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*, pages 42–47.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386.

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.