CVTE at IJCNLP-2017 Task 1: Character Checking System for Chinese Grammatical Error Diagnosis Task

Xian Li, Peng Wang, Suixue Wang, Guanyu Jiang, Tianyuan You
CVTE Central R&D Institute
lixian@cvte.com

Abstract

Grammatical error diagnosis is an important task in natural language processing. This paper introduces CVTE Character Checking System in the NLP-TEA-4 shared task for CGED 2017, we use Bi-LSTM to generate the probability of every character, then take two kinds of strategies to decide whether a character is correct or not. This system is probably more suitable to deal with the error type of bad word selection, which is one of four types of errors, and the rest are words redundancy, words missing and words disorder. Finally the second strategy achieves better F1 score than the first one at all of detection level, identification level, position level.

1 Introduction

Nowadays, Chinese language gains more popularity in the world, many foreigners begin to learn Chinese. Unlike English, Chinese has no verb tenses and pluralities, and a sentence can be expressed in many ways, a native Chinese speaker can handle well all of these different grammatical phenomena, but for the foreigners, these are difficult parts to learn the Chinese well. In the HSK (Hanyu Shuiping Kaoshi), which is an international standard test for Chinese language proficiency of non-native speakers, after analyzing a considerable number of examination papers, it shows that foreigners who study Chinese often make grammatical mistakes by having redundant words, missing words, bad word selection and disorder words due to their language false analogy, over-generalization, teaching methods, learning strategies and other reasons. For all grammatical errors, it is proposed that a task named CGED (Chinese Grammatical Error Diagnosis) as one of share task of NLPTEA in three consecutive years 2014-2016, CGED 2014 (Yu etal., 2014) defined four kinds of grammatical errors: words redundancy, words missing, bad word selection and words disorder. At most one error occurred in one sentence. The evaluation was based on error detection and error classification in sentence level. CGED 2015 (Lee et al., 2015) further required the positions of the errors. CGED 2016 tested on the ability to detect multiple errors in one sentence.

2 Task Definition

The shared task of CGED is defined as below: There are four types of grammatical errors in a sentence, which are redundancy (R), words missing (M), bad selection (S) and disorder (D). The systems participating this shared task should detect whether the sentence contains errors (Detection-level), find out which type the error belongs to (Identification-level), and where the errors are (Position-level).

Table1 and Table2 show two examples in test data:

我$_1$真$_2$不$_3$明$_4$白$_5$。$_6$她$_7$们$_8$可$_9$能$_{10}$是$_{11}$追$_{12}$求$_{13}$一$_{14}$些$_{15}$前$_{16}$代$_{17}$的$_{18}$浪$_{19}$漫$_{20}$		
Correction	我$_1$真$_2$不$_3$明$_4$白$_5$。$_6$她$_7$们$_8$可$_9$能$_{10}$是$_{11}$追$_{12}$求$_{13}$一$_{14}$些$_{15}$前$_{16}$代$_{17}$的$_{18}$浪$_{19}$漫$_{20}$	
Detection-level	correct	
Identification-level	-	-
Position-level	-	-

Table 1: The sentence is correct.

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 78–83,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

我$_1$根$_2$本$_3$不$_4$能$_5$了$_6$解$_7$这$_8$妇$_9$女$_{10}$辞$_{11}$职$_{12}$回$_{13}$家$_{14}$的$_{15}$现$_{16}$象$_{17}$		
Correction	我$_1$根$_2$本$_3$不$_4$能$_5$理$_6$解$_7$妇$_8$女$_9$辞$_{10}$职$_{11}$回$_{12}$家$_{13}$的$_{14}$现$_{15}$象$_{16}$	
Detection-level	-	
Identification-level	S	R
Position-level	6,7	8,8

Table 2: Two errors are found in the sentence above, one is bad word selection (S) error from position 6 to 7, the other one is redundant words (R) error at position 8.

3 The Unified Framework of CVTE Character Checking System

For this shared task, we propose a unified framework called CVTE Character Checking System in Figure 1, just as its name implies, the system can handle with Chinese character errors which are almost S errors in CGED-2017, hence our system mainly focuses on S errors for the moment.

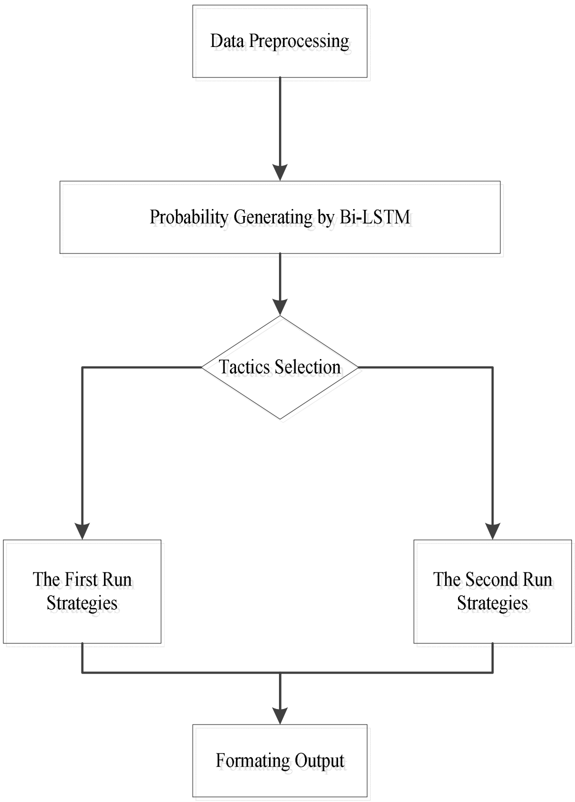

Figure 1: A unified framework of CVTE Character Checking System

Data preprocessing step is to split a sentence into sub-sentences by punctuation. Probability Generating by Bi-LSTM step generates the probability of each character of the input sub-sentence. In Tactics Selection step, the sole purpose is to choose which error deciding strategy we are going to use. The First Run Strategies and The Second Run Strategies are error deciding steps. Format output step is to format the output into CGED-2017 style.

3.1 Data preparation

Data provided by organizer is in the form of long sentences and contains some non-Chinese characters. In our system, only short sentences are supported, and they can't contain non-Chinese characters and punctuations. In order to meet the input requirement of our system, long sentences were splitted into some short sub-sentences by punctuation, and non-Chinese characters were removed determined by its unicode.

This task is an open test and some training data is provided, but considering that a larger training set can improve the performance of our framework, we crawl corpus in addition from composition website and novel website as our training data.

3.2 Probability Generating by Bi-LSTM Model

In order to achieve good performance of neural network language model (Bengio Y.,2003), we implement RNN neural network (Mikolov T.,2012) to train our language model. A Bi-LSTM multi-layer network is applied into the structure of training model, so that both previous and posterior sub-sentences could be taken into consideration.

In the Chinese information processing, the performance of word segmentation determines the upper bound of tasks. In addition, the number of words is much larger than the number of characters. Words based features will bring sparseness to the training data, and will also reduce the training speed of the neural network. In order to get rid of these problems, the Chinese characters are taken as the input of the neural network language model. $S = C_1C_2...C_n$ stands for the sentence to be detected, as shown in figure 1, and

C_i is a single Chinese character. After sentence S is put into the neural network, the probability of every character at its position will be exported, which we name "position probability distribution". The forward input sequence is $<s> C_1 ... C_{n-1}$, and the backward input sequence is $C_2 ... C_{n-1} </s>$, and the label sequence is $C_1 C_2 ... C_n$. In all of forward and backward input sequences, $<s>$ and $</s>$ represent the start and the end of a sequence respectively. Taking the probability distribution of position of C_2 as an example, Bi-LSTM utilizes the context information of both $<s> C_1$ and $C_3 ... </s>$.

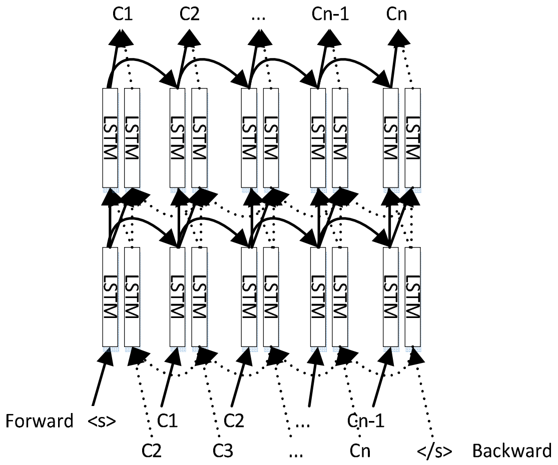

Figure 2: The Bi-LSTM language model based on characters

3.3 Error Detection Strategies

Based on the neural network language model mentioned in the previous section, we propose two Strategies for the final detection of the error position.

3.3.1 The First Run Strategy

After a sequence is input, we can get the position probability distribution and the probability of each character. The main process of this strategy has two parts. Firstly, whether a character is correct may be judged directly by the character probability. Secondly, if it is hard to make such a decision, generating the scores of all candidate sentences which are built by replacing certain characters with confusion characters which own the top 3 probability values in confusion set, and choosing the one with top score in candidate sentences and original sentence. Confusion characters share the same Pinyin with each other. The score of a sentence is computed by multiplying every character probability which is the output of Bi-LSTM model in the sentence. The same position where the character is different between top score sentence and original sentence is the error position the system finds out. The details are shown as pseudo code in Table 3.

3.3.2 The Second Run Strategy

We also put forward another strategy which can improve the recall rate. The specific process is shown in the pseudo code of Table 4 and Table 5.

The pseudo code of the first strategy

```
1: FOR each character $C_i^{origin}$ with probability $P_i^{origin}$ in sentence $S_{origin}$:
        IF $P_i^{origin} \geq 0.1$ THEN:
            $C_i^{origin}$ correct and continue
        IF $0.001 \leq P_i^{origin} < 0.1$ THEN:
            Get the maximum probability value $Q_{C_i}$ of character in confusion set $F_{C_i}$
            IF $P_i^{origin} == Q_{C_i}$ THEN:
                $C_i^{origin}$ correct and continue

        Select the characters $T_1^{C_i^{origin}}$, $T_2^{C_i^{origin}}$, $T_3^{C_i^{origin}}$ with top 3 probability values in $F_{C_i}$
3: Generate the candidate sentences by replace $C_i^{origin}$ with $T_1^{C_i^{origin}}$, $T_2^{C_i^{origin}}$, $T_3^{C_i^{origin}}$ in $S_{origin}$
4: Input these candidate sentences $S_{T_j^{C_i^{origin}}}$ back into model，and get $Score_{T_j^{C_i^{origin}}}$,
   where $Score_{T_j^{C_i^{origin}}} = \sum_{1}^{N} \log P_k^{T_j^{C_i^{origin}}}$
5: FOR position $i$ in $S_{origin}$:
        IF $Score_{C_i^{origin}} > Score_{T_j^{C_i^{origin}}}$ THEN:
                $C_i$ correct
        IF $Score_{C_i^{origin}} <= Score_{T_j^{C_i^{origin}}}$ THEN:
                $C_i$ error
```

Table 3: The pseudo code of the firs strategy

```
Rule_Top_3
FOR position $i$, Bi-LSTM output set of softmax probabilities is $D_i$, $p_1$, $p_2$, $p_3$ are top three probabilities in $D_i$ THEN:
    $sum\_top\_3 = p_1 + p_2 + p_3$
    IF $sum\_top\_3 \geq 0.99$ THEN:
        $N_1 = 10$ and $M_1 = 30$
        select top $N_2$ characters that make the sum of probabilities is greater than 0.998
        select top $M_2$ characters that make the sum of probabilities is just greater than 0.97
    IF $sum\_top\_3 < 0.99$ AND $sum\_top\_3 \geq 0.95$ THEN:
        $N_1 = 20$ and $M_1 = 40$
        select top $N_2$ characters that make the sum of probabilities is greater than 0.995
        select top $M_2$ characters that make the sum of probabilities is greater than 0.965
    IF $sum\_top\_3 < 0.95$ AND $sum\_top\_3 \geq 0.65$ THEN:
        $N_1 = 30$ and $M_1 = 70$
        select top $N_2$ characters that make the sum of probabilities is greater than 0.992
        select top $M_2$ characters that make the sum of probabilities is greater than 0.96
    IF $sum\_top\_3 < 0.65$ THEN:
        $N_1 = 50$ and $M_1 = 100$
        select top $N_2$ characters that make the sum of probabilities is greater than 0.99
        select top $M_2$ characters that make the sum of probabilities is greater than 0.95
    $N = min\ (N_1, N_2)$
    $M = min\ (M_1, M_2)$
The N number of characters make up set one $U_1$
The M number of characters make up set two $U_2$
```

Table 4: The pseudo code of Rule_Top_3

```
The process of strategy two
1: FOR each character $C_i^{origin}$ with probability $P_i^{origin}$ in sentence $S_{origin}$:
        IF $P_i^{origin} \geq 0.1$ THEN:
            $C_i^{origin}$ correct and continue
        IF $P_i^{origin} \leq 0.0001$ THEN:
            $C_i^{origin}$ error and continue
        Select two sets based on Rule_Top_3 shown in Table 4
        IF $C_i^{origin}$ in $U_1$ THEN:
            $C_i^{origin}$ correct and continue
        IF $C_i^{origin}$ not in $U_2$ THEN:
            $C_i^{origin}$ error and continue

2: Generate the candidate sentences $S_{T_j^{C_i^{origin}}}$ by replace $C_i^{origin}$ with all character $T_j$ in confusion set $F_{C_i}$
3: Input these candidate sentences $S_{T_j^{C_i^{origin}}}$ back into model， and get $Score_{T_j^{C_i^{origin}}}$ ,
where $Score_{T_j^{C_i^{origin}}} = \sum_1^N \log P_k^{T_j^{C_i^{origin}}}$
        IF $Score_{C_i^{origin}}$ is in top 20 percent of set $\left(Score_{C_i^{origin}}, Score_{T_j^{C_i^{origin}}}\right)$ THEN:
            $C_i^{origin}$ correct and continue
        ELSE:
            $C_i^{origin}$ error and continue
```

Table 5: The pseudo code of the second strategy

4 Experiments

In the formal run of CGED2017 shared task, there are 5 participants in HSK, 13 runs in total. Two runs (CVTE-Run1, Run2) of HSK were submitted to CGED 2017 shared task for official evaluation. The submission of Run1 is generated by The First Run Strategies system and the Run2 is generated by The Second Run Strategies system. Table 6 shows the false positive rate, our system has relatively low false positive rate comparing with other participants.

Table 7, Table 8, and Table 9 show the formal run result of our system in Detection level, Identification level and Position level, respectively. Our system mainly focuses on the Detection level, as for this task, Run2 plays better than Run1, and it has relatively better performance on Accuracy, Precision and F1-score indicators. As for Identification level task, Run1 achieves the highest precision rate comparing with other teams, but the recall rate of our system is fare.

Submission	False Positive Rate
CVTE-Run1	0.1441 (169/1173)
CVTE-Run2	0.3154 (370/1173)

Table 6: False Positive Rate

Submission	Acc	Pre	Rec	F1
CVTE-Run1	0.475	0.745	0.250	0.374
CVTE-Run2	0.539	0.708	0.452	0.552

Table 7: Detection Level

Submission	Acc	Pre	Rec	F1
CVTE-Run1	0.446	**0.606**	0.121	0.202
CVTE-Run2	0.471	0.539	0.205	0.297

Table 8: Identification Level

Submission	Acc	Pre	Rec	F1
CVTE-Run1	0.331	0.118	0.020	0.034
CVTE-Run2	0.260	0.109	0.046	0.065

Table 9: Position Level

5 Conclusion

This paper proposes a unified framework called CVTE Character Checking System which only aims to handle with bad word selection error. Bi-LSTM and two kinds of strategies are applied into our system. However, the other types of errors such as words redundancy, words missing, and words disorder are not considered in the system, which may not give fine results. Chinese character error and Chinese grammatical error are different levels of error in a sentence, so the solutions are quite different.

In future studies, works on both Chinese character check and Chinese grammatical error diagnosis could be done to improve our system, which include: (1) Taking the word level Bi-LSTM model for Chinese character check. (2) Containing a sequence to sequence model for Chinese grammatical error diagnosis. (3) Implementing an online toolkit and service for Chinese character check and Chinese grammatical error diagnosis as a stimulator for this empirical research topic.

Acknowledgments

This research was funded by the Guangzhou Huangpu Postdoctoral Science Foundation.

References

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang (2014). Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan, 30 November, 2014, pp. 42-47.*

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang. 2015. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. *In Proceedings of the 2nd*

Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'15), pages 1-6, Beijing, China.

Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. *Journal of machine learning research*, 2003, 3(Feb): 1137-1155.

Mikolov T. Statistical language models based on neural networks[J]. *Presentation at Google, Mountain View*, 2nd April, 2012.