CMMC-BDRC Solution to the NLP-TEA-2018 Chinese Grammatical Error Diagnosis Task

Yongwei Zhang[1,2], **Qinan Hu**[1,2], **Fang Liu**[1,3], **and Yueguo Gu**[1]
[1]China Multilingual & Multimodal Corpora and Big Data Research Centre, Beijing, China
[2]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
[3]School of Software and Microelectronics, Peking University, Beijing, China
zhangyw@cass.org.cn, qinan.hu@qq.com, liu_fang@pku.edu.cn, gyg@beiwaionline.com

Abstract

Chinese grammatical error diagnosis is an important natural language processing (NLP) task, which is also an important application using artificial intelligence technology in language education. This paper introduces a system developed by the Chinese Multilingual & Multimodal Corpus and Big Data Research Center for the NLP-TEA shared task, named Chinese Grammar Error Diagnosis (CGED). This system regards diagnosing errors task as a sequence tagging problem, while takes correction task as a text classification problem. Finally, in the 12 teams, this system gets the highest F1 score in the detection task and the second highest F1 score in mean in the identification task, position task and the correction task.

1 Introduction

With the development of Chinese economy and the growing popularity of Chinese culture, more and more foreigners begin to learn Chinese. However, Chinese and English are different. For instance, Chinese grammar is more flexible and more complex than English grammar and there are few morphological changes in Chinese. Consequently, it is quite difficult for the second language (L2) learners to master. In addition, the huge number of Chinese characters and no space between word and word cause the difficulty in Chinese natural language processing. In short, regarding how to use artificial intelligence to correct L2 learners, Chinese writing meets both opportunities and challenges.

In order to promote the development of automatic detection of syntactic errors in Chinese writing, the Natural Language Processing Techniques for Educational Applications (NLP-TEA) have taken CGED as one of the shared tasks since 2014. Thanks to the CGED task, some research achievements have been made in Chinese grammar error detection. Based on those previous research results, this paper puts forward a new thinking direction of enriching training dataset for the CGED task.

The structure of this article is as follows: Section 2 briefly introduces the CGED shared task. Section 3 introduces some related work. Section 4 talks about the methodology. Section 5 presents the data augmentation method used in the system, and section 6 shows the experiment result. Finally, conclusion and future work are drawn in Section 7.

2 Task Definition

CGED has been held in five consecutive years since 2014. It aims to develop a NLP system to automatically diagnose grammatical errors in Chinese sentences written by L2 learners. Such errors are divided into four types: redundant words (‘R’), missing words (‘M’), word selection errors (‘S’), and word ordering errors (‘W’). The input sentence may contain one or more such errors. For each sentence, the developed system would detect the following four levels (or tasks):

(1) Detection-level: whether the sentence is correct or not?

(2) Identification-level: which error types are embedded?

(3) Position-level: where the error positions occur?

(4) Correction-level: what is the correct

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 180–187
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

word?

M and S type errors are required to offer 1 to 3 corrections. The other type errors only need to be identified.

The training dataset provided by CGED includes original error text, correct text, error types as well as error intervals. But the correct words of errors are not given explicitly. Table 1 shows two examples of the training dataset.

In table 1, there are two errors in example 1. One is S type from position 23 to 24, and the other is M type at position 28. There are also two errors in example 2. One is R type at position 8, and the other is W type from position 9 to 14. It has been found that，in example 1, ‘原故’ is an error word and ‘缘故’ is the correct form. Beside this, ‘了’ is omitted in example 1.

3 Related Work

Yu and Chen (2012) proposed a CRF-based model to detect Chinese word ordering errors. In 2014, Cheng et al. (2014) proposed an SVM model to further study the Chinese word ordering problems. Lee et al. (2013) used a series of manual linguistic rules to detect grammatical errors in Chinese learners’writings. Lee et al. (2014) then further proposed a system which integrated both handcrafted linguistic rules and N-gram models to detect Chinese grammatical errors in sentences. Those two aforementioned models are based on linguistic rules, which need to be summarized manually. And because of the flexibility of Chinese syntax, the performance of existing models is not ideal. In recent years, artificial neural networks have been extensively used to do NLP tasks. However, due to the lack of large writing data of interlanguage, the performance of deep learning algorithms is limited a lot. In order to integrate more linguistic information into neural networks, HIT team (Zheng et al., 2016) used Part-of-Speech (POS) tag as a feature, and Alibaba team (Yang et al., 2017) further integrated Part-of-Speech-Tagging Score (POS Score), Point-wise Mutual Information (PMI), and dependency word collocation etc. into deep learning networks. These efforts made two teams achieved pretty good results in 2016 and 2017 CGED tasks respectively.

4 Methodology

We treat the first three tasks which are detection task, identification task and position task (DIP tasks) as a sequence tagging problem, and correction task as a classification problem.

4.1 Methodology of DIP Tasks

4.1.1 Model Description

Same with the methods used by HIT team (Zheng et al., 2016) and Alibaba team (Yang et al., 2017), we treat DIP tasks as a sequence tagging problem. Specifically, we tag each character of the sentences and then use the LSTM-CRF model (Huang et al., 2015) for training and prediction. Each character is tagged with BIO encoding (Collier and Kim, 2004), also the same as the method adopted by HIT team (Zheng et al., 2016) and Alibaba team (Yang et al., 2017). We use the bidirectional LSTM unit as the RNN model. The structure of the model we adopted in our research is shown in Figure 1.

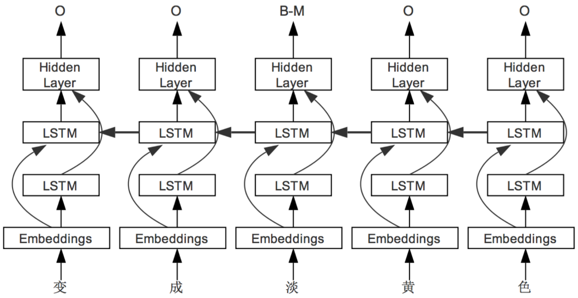


Figure 1: The structure of LSTM-CRF model we used.

4.1.2 Word Embedding Feature

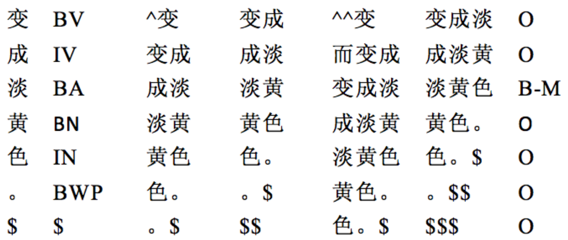
We use char feature, POS feature, two char bigram features, and two char trigram features as the input features of the neural network. Language Technology Platform[1] (LTP) is used to segment words and do the POS tagging. If a word’s POS tag is ‘X’, the POS tag of the first character of the word is ‘B-X’, and the POS tags of the rest characters of the word are all ‘I-X’. When training bigram embeddings or trigram embeddings，we need to add a ‘^’ character at the start of a sentence and a ‘$’ character at the end of the sentence. In addition, in order to mark the missing words

[1] https://github.com/HIT-SCIR/ltp/

Example 1	Original Text	1 此 2 外 3 ， 4 吸 5 烟 6 也 7 影 8 响 9 了 10 美 11 观 12 ， 13 洁 14 白 15 的 16 牙 17 齿 18 因 19 为 20 吸 21 烟 22 的 23 原 24 故 25 而 26 变 27 成 28 淡 29 黄 30 色 31 。 32	
	Correct Text	此外，吸烟也影响了美观，洁白的牙齿因为吸烟的缘故而变成了淡黄色。	
	Error Type	S (word selection)	M (missing word)
	Error Interval	23, 24	28, 28
	Error-Correct Word	原故-缘故	-了
Example 2	Original Text	1 一 2 般 3 的 4 吸 5 烟 6 的 7 人 8 把 9 时 10 间 11 管 12 理 13 不 14 好 15 。 16	
	Correct Text	一般的吸烟的人管理不好时间。	
	Error Type	R (redundant word)	W (word ordering error)
	Error Interval	8, 8	9, 14
	Error-Correct Word	把-	时间管理不好-管理不好时间

Table 1: Two examples of training sentence of the CGED training dataset.

error occurred at the end of the sentence, a '$' character is also need to be added at the end of the sentence. Figure 2 shows an example of the embedding features we used as the input for the neural networks.



变	BV	^变	变成	^^变	变成淡	O
成	IV	变成	成淡	而变成	成淡黄	O
淡	BA	成淡	淡黄	变成淡	淡黄色	B-M
黄	BN	淡黄	黄色	成淡黄	黄色。	O
色	IN	黄色	色。	淡黄色	色。$	O
。	BWP	色。	。$	黄色。	。$$	O
$	$	。$	$$	色。$	$$$	O

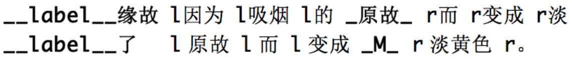
Figure 2: Embedding features of each character of '变成淡黄色。'. Each line represents one character's embedding features. These embedding features can be categorized as char feature, POS feature, two char bigram features, two char trigram features, and error tag. Different features are separated by using a tab character.

4.2 Methodology of Correction Task

4.2.1 Model Description

The goal of text classification is to assign documents to one or multiple categories. Such categories can be spam v.s. non-spam, review scores or animal names. For correction task, the correct word can be seen as another type of category, and its context including its error form can be seen as a short document belonging to the category. In example 1, '缘故' is mistakenly written as '原故'. So we take '缘故' as a category, and '原故' as well as its context（N-gram) as the document.

In order to distinguish an error word from its left and right contexts, for correction task, we add a '_' character before and after the error word, a 'l' letter before each left word, and a 'r' letter before each right word. In addition to this, a prefix '___label___' is also required before the category name. And for M type error, we use '_M_' to denote the missing word, as shown in Figure 3.



```
__label__缘故 l因为 l吸烟 l的 _原故_ r而 r变成 r淡
__label__了   l 原故 l 而 l 变成 _M_ r 淡黄色 r。
```

Figure 3: The categories and their corresponding documents (texts) generated from example 1. Each line contains a category, followed by a corresponding document (text) which takes the error words, leftward three words, and rightward three words as its content.

However, using a text classifier to provide correct words also has a disadvantage—all proposed words must be correct forms of error words or missing words in the training dataset. The classifier can not provide correct words which do not contain. But the number of words that L2 learners used is limited. For this reason, text classifier can be used to provide correct words for the most common error words and missing words.

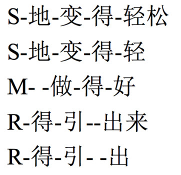
5 Data Augmentation

5.1 Rule format

The training dataset of CGED is relatively small for training neural network models. Increasing the scale of the training dataset may improve the performance of the models. We can study error rules from the training dataset of CGED. In addition, we find that L2 learners often make mistakes that native speakers are frequently to make. So, to identify linguistic mistakes often made by native speakers frequently also helps to identify linguistic errors of L2 learners. Therefore, there are two sources of data augmentation rules in this paper: (i) the training dataset of CGED; (ii)native speakers' error-prone language knowledge.

Error rules can be extracted from the training dataset of CGED, and be studied from the native speakers' error-prone language knowledge. And then, we can use those rules to generate more error sentences to enrich the training dataset. Therefore, error rule is an important medium for data augmentation.

The error rule consists of error type, error word, prefix of the error word, correct form of the error word (correct word), and suffix of the error word. The error rule types include S, M, and R types.



S-地-变-得-轻松
S-地-变-得-轻
M- -做-得-好
R-得-引--出来
R-得-引- -出

Figure 4: An example of the rule format.

If figure 4, each line represents one error rule. The items of an error rule involved can be categorized as error type, error word, prefix, correct word, and suffix. Different items are separated by using a '-' character from left to right. The rule 'S-地-变-得-轻松' expressing the meaning of '变得轻松' is wrongly written as '变地轻松'.

5.2 Rules from CGED Training Dataset

The steps of extracting rules from the training dataset of CGED are indicated as follows:

(1) Count the number of sentences in each training document that contains the original error text and correct text, and discard documents that are not equal in number and cannot be corrected manually.

(2) Split the original error text and correct text of each document into sentences by LPT toolkit.

(3) Each error of the sentence can generate an error rule. The components of an error rule can be calculated based on the sentence original error text, correct text, and error interval. The prefix and suffix can be a word or a character. If it is a word, the left and right strings of the error word in the sentence need to do word segmentation respectively. After the word segmentation, the prefix becomes the rightmost word of the left string, and the suffix is the leftmost word of the right string.

For example, example 3 in Table 2 contains a S type error. Through the original text and error interval, we can know that '教养' is a bad word selection. The content before '抚养' in the correct text is the same as the content before '教养' in original text, and the content behind '抚养' in correct text is also the same as the content behind '教养' in original text. This can be inferred that the correct writing of '教养' should be '抚养' in this context. Therefore, the rules 'S -教养-孩子-抚养-成人' and 'S -教养-子-抚养-成' can be derived from the example 3.

Not all the correct form of an error word can be inferred. It is difficult to infer the correct word if the following conditions occur:

(1) Two errors have crossed position, or one error is contained in another.

(2) Two errors next to each other in position, but they are a S type error and a M type error.

(3) Two errors next to each other in position, but one of them is a W type error.

5.3 Rules from Native Speakers

There are many resources in Baidu WenKu[2], such as similar Chinese characters, commonly confused words, homonyms, and easily-misused characters, which are collected and uploaded by many teachers or students' parents. In addition, many Chinese researchers have written different kinds of books and dic-

[2] http://wenku.baidu.com/

Example 3	Original Text	1 怎 2 样 3 把 4 孩 5 子 6 教 7 养 8 成 9 人 10 呢 11? 12
	Segment Text	怎样, 把, 孩子, 教养, 成人, 呢,?
	Correct Text	怎样把孩子抚养成人呢?
	Error Type	S (word selection)
	Error Interval	6,7
	Error-Correct Word	教养-抚养

Table 2: An example of training sentence that contains only one S type error.

tionaries to review these resources (Li, 2005; Pang, 2006; Ran, 2010; Tian, 2012; Ye, 1978).

Although all of the aforementioned resources can be converted to error rules. Although these resources provide only a correct word or an error word of an error rule, the prefix and suffix can be obtained from text corpus. We count the cluster (trigram) of the words in a textbook corpus, and the words located before or after the central words are regarded as prefixes or suffixes respectively. For example, the highest frequency clusters which take ‘录’ as the central error word are ‘报录的’, ‘记录下’ and ‘听录音’. ‘录’ and ‘陆’ are easily-misused Chinese characters. Taking the misuse of ‘录’ as ‘陆’ for an example, we can generate the error rules of ‘S-陆-报-录-的’, ‘S-陆-记-录-下’ and ‘S-陆-听-录-音’ with the help of the high frequency clusters extracted from the textbook corpus.

In addition to the S type error, the M type error and the R type error can also be generated similarly. In order to reduce the number of rules and make the rules more accurate in predicting, the Chinese characters of the error word and correct word are all from the *Essential Chinese Dictionary* (Xu and Yao, 2009) and the top 1500 frequency characters high frequency in the list of the training dataset of CGED. These two wordlists contain 1,535 different Chinese characters. Based on the wordlists, 97.48% (49706/50471) of the correct words of the CGED dataset are formed.

5.4 Data Generation

5.4.1 Raw Data

In order to make the generated sentences more similar to the sentences written by L2 learners, we select candidate sentences from a textbook corpus, which covers 12 sets of textbooks compiled for foreign students and 7 sets of textbooks compiled for Chinese students, provided by the Research Center for Lexicology & Lexicography, the Chinese Academy of Social Sciences. Although large-scaled, it is still failed to provide enough candidate sentences. Therefore, we also select the People’s Daily (1946-2017) provided by the Library of the Chinese Academy of Social Sciences as a supplementary corpus.

5.4.2 Preprocessing

The processing of text corpus includes the following steps:

(1) Use OpenCC[3] toolkit to convert all traditional CGED dataset to simplify dataset.

(2) Use LTP toolkit to do Chinese sentence segmentation.

(3) Filter the sentences by following methods: discard sentences whose characters are less than 5 or more than 40; discard sentences, in which the proportion of Chinese characters is less than 50%; if a sentence contains any character or word out of *National Syllabus of Graded Words and Characters for Chinese Proficiency* (Hanban, 2001) and *Chinese Proficiency Test Syllabus Level 1-6* (Hanban, 2010), the sentence should also be discarded.

The rest sentences are candidate sentences for generating error sentences.

5.4.3 Error Sentences Generation

Error sentences are generated based on error rules. We can replace the ‘prefix+correct word+suffix’in a filtered candidate correct sentence with ‘prefix+error word+suffix’ to get an error sentence. For example, there is a correct sentence ‘他又当爹又当妈，把儿子抚养成人。’ and an error rule ‘S-教养-子-抚养-成’. When ‘子抚养成’ in the sentence is replaced with ‘子教养成’, a new error sentence ‘他又当爹又当妈，把儿子教养成人。’ is generated. The newly generated training sentence is shown in table 3.

[3] https://github.com/BYVoid/OpenCC

Example 4	Original Text	1 他 2 又 3 当 4 爹 5 又 6 当 7 妈 8， 9 把 10 儿 11 子 12 教 13 养 14 成 15 人 16。 17
	Correct Text	他又当爹又当妈，把儿子抚养成人。
	Error Type	S (word selection)
	Error Interval	12,13
	Error-Correct Word	教养-抚养

Table 3: An example of training sentence generated from an error rule 'S-教养-子-抚养-成'.

6 Experiment Results

6.1 Implementation Details

We merge all the historical CGED training dataset and test dataset, and obtain 76,117 error sentences after sentence segmentation, of which 58,521 sentences have corresponding correct sentences. We use 80% of the error sentences and their corresponding correct sentences for training (119,414 sentences) and the rest for validation. In DIP tasks, we generated 79,131 rules from CGED dataset and 61,149 **different** rules from other corpus mentioned in section 5.4.1. With the help of these error rules, we generated 19,1331 error sentences. We use TensorFlow[4] to implement the LSTM-CRF model, and use FastText[5] directly for the correction task. We only use pre-trained embeddings for LSTM-CRF model which are pre-trained with the textbooks corpus and People's Daily (1946-2017) text corpus.

6.2 Results on Validation Dataset

We used the validation dataset to select the best hyper-parameters for both the LSTM-CRF model of DIP tasks and the classification model for correction task. From the results of table 4, it has been found that the model with added trigram embeddings performs better than that with only character embedding and bigram embeddings when using the same dataset. The model trained with increased new data is superior to the model that only trained with CGED dataset.

Table 5 shows the results of the correction task. MN refers to model N. For example, M2 refers to model 2. N stands for the number of aforementioned prefixes and suffixes in section 5.1. The smaller the N is, the more effective the model is.

Detection Task			
Model	**Precision**	**Recall**	**F1**
CGED (U+B)	**0.6137**	0.6586	0.6354
CGED (U+B+T)	0.5686	**0.8102**	0.6682
CGED+G (U+B+T)	0.5969	0.7615	**0.6692**
Identification Task			
Model	**Precision**	**Recall**	**F1**
CGED (U+B)	0.4204	0.4236	0.422
CGED (U+B+T)	0.3973	**0.4974**	0.4418
CGED+G (U+B+T)	**0.4213**	0.4905	**0.4533**
Position Task			
Model	**Precision**	**Recall**	**F1**
CGED (U+B)	0.2995	0.2634	0.2803
CGED (U+B+T)	0.2499	0.2831	0.2655
CGED+G (U+B+T)	**0.3161**	**0.3057**	**0.3108**

Table 4: Results on Validation Dataset of DIP tasks. CGED indicates that only CGED training dataset is used. G stands for using generated dataset, U stands for character embedding, B stands for bigram embeddings, and T stands for trigram embeddings.

In table 5, model 1 has the best predictive effect, while the other models can predict the correct suggestions rather than model 1. Therefore, we take the results of model 1 as basis. If three results of the other four models are inconsistent with those of model 1, they will be taken as the priority result.

[4] https://github.com/tensorflow/tensorflow

[5] https://github.com/facebookresearch/fastText

Correction Task (Top1)			
Model	**Precision**	**Recall**	**F1**
M1	**0.323**	**0.323**	**0.323**
M2	0.310	0.310	0.310
M3	0.297	0.297	0.297
M4	0.287	0.287	0.287
M5	0.278	0.278	0.278
Correction Task (Top3)			
Model	**Precision**	**Recall**	**F1**
M1	**0.136**	**0.408**	**0.204**
M2	0.130	0.389	0.195
M3	0.122	0.367	0.183
M4	0.121	0.362	0.181
M5	0.118	0.354	0.177

Table 5: Results on Validation Dataset of Correction task.

6.3 Results on Evaluation Dataset

While testing on the final evaluation dataset, we merged all the training dataset and validation dataset, and added generated sentences to retrain our models. Table 6 and Table 7 show the final results of DIP tasks and correction task.

We used the same parameters for training 9 different models, but obtained 9 different test results. Hence, we selected the best performing model in detection task in evaluating dataset of 2017 as run 1, and the best performing model in position task in evaluating dataset of 2017 as run 2. During this process, we didn't apply any model stacking.

Finally, 12 teams submitted 32 DIP task results. The first run of our system (run1) achieved the highest F1 scores in the detection task. In the identification task, the F1 of run1 and run2 ranked the second and the third respectively. And in the position task, the F1 of run2 gained third place among 32 results.

As for the correction task, the new task of this year, 9 teams submitted a total of 23 results. Run2 got better result than run1 in both top1 and top3 tasks. In top1 correction task, the F1 of run2 ranked 2/9 according to teams and 2/23 according to results, which is lower than the highest result by only 0.0001. In top3 correction task, the F1 of run2 ranked 2/9 according to teams and 3/23 according to results.

Detection Task			
Runs	**Precision**	**Recall**	**F1**
Run1	0.6736	**0.8621**	**0.7563**
Run2	**0.7266**	0.7408	0.7336
Identification Task			
Model	**Precision**	**Recall**	**F1**
Run1	0.4834	0.5952	0.5335
Run2	**0.5831**	**0.4955**	**0.5357**
Position Task			
Model	**Precision**	**Recall**	**F1**
Run1	0.2741	**0.3177**	0.2943
Run2	**0.3839**	0.2966	**0.3346**

Table 6: Results on Evaluation Dataset of DIP Tasks.

Correction Task (Top1)			
Runs	**Precision**	**Recall**	**F1**
Run1	0.1364	**0.1651**	0.1494
Run2	**0.1852**	0.1609	**0.1722**
Correction Task (Top3)			
Runs	**Precision**	**F1**	
Run1	0.1432	0.1569	
Run2	**0.1934**	**0.1798**	

Table 7: Results on Evaluation Dataset of Correction Task.

7 Conclusion and Future Work

In this shared task paper, we mainly describe how to generate more error sentences based on the CGED training dataset and large filtered corpus. Based on the original training data and augmented data, we trained LSTM-CRF models ranking 1/12, 2/12 and 2/12 separately in DIP tasks. In the correction task, we regarded it as a classification problem and ranked 2/9. Our final submitted results achieved 2nd place in mean ranking. All of this proves the effectiveness of the data augmentation algorithm proposed in this paper.

In the future work, we will blend more grammatical features in error detection and correction, and integrate more second language teaching experience in the model.

References

Shuk-Man Cheng, Chi-Hsin Yu, and Hsin-Hsi Chen. 2014. Chinese word ordering errors detection and correction for non-native chinese language learners. In *Proceedings of COL-*

ING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 279–289.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, NLPBA/BioNLP 2004, Geneva, Switzerland, August 28-29, 2004*.

Hanban. 2001. *National Syllabus of Graded Words and Characters for Chinese Proficiency*. Economic Science Press.

Confucius Institute Headquarters Hanban. 2010. *Chinese Proficiency Test Syllabus Level 1-6*. The Commercial Press.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Lung-Hao Lee, Li-Ping Chang, Kuei-Ching Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2013. 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)*, pages 27–29.

Lung-Hao Lee, Liang-Chih Yu, Kuei-Ching Lee, Yuen-Hsien Tseng, Li-Ping Chang, and Hsin-Hsi Chen. 2014. A sentence judgment system for grammatical error detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 67–70.

Xingjian Li. 2005. *Similar Chinese Character Discrimination Dictionary*. Jiangsu Education Publishing House.

Chenguang Pang. 2006. *Common Similar Chinese Character Discrimination Dictionary*. World Publishing Corporation.

Hong Ran. 2010. *Similar Chinese Character Dictionary*. Foreign Language Teaching and Research Press.

Juanhua Tian. 2012. *Similar Chinese Characters Discrimination*. Shanghai Brilliant Publishing House.

Lin Xu and Xishuang Yao. 2009. *Essential Chinese Dictionary*. Foreign Language Teaching and Research Press.

Yi Yang, Pengjun Xie, Jun Tao, Guangwei Xu, Linlin Li, and Si Luo. 2017. Alibaba at IJCNLP-2017 task 1: Embedding grammatical features into lstms for chinese grammatical error diagnosis task. In *Proceedings of the IJCNLP 2017, Shared Tasks, Taipei, Taiwan, November 27 - December 1, 2017, Shared Tasks*, pages 41–46.

Yu Ye. 1978. *Easily-Misused Chinese Characters*. Shanghai Educational Publishing House.

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.

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, Osaka, Japan. The COLING 2016 Organizing Committee.