CYUT Team Chinese Grammatical Error Diagnosis System Report in NLPTEA-2020 CGED Shared Task

Shih-Hung Wu*, Jun-Wei Wang
Chaoyang University of Technology,
Taichung, Taiwan, R.O.C
shwu@cyut.edu.tw, s10827605@gm.cyut.edu.tw
*Contact author

Abstract

This paper reports our Chinese Grammatical Error Diagnosis system in the NLPTEA-2020 CGED shared task. In 2020, we sent two Runs with two approaches. The first one is a combination of conditional random fields (CRF) and a BERT model deep-learning approach. The second one is CRF approach. The official test results shows that our Run1 achieved the highest precision rate 0.9875 with the lowest false positive rate 0.0163 on detection, while Run2 gives a more balanced performance.

1 Introduction

Learning Chinese is very popular for foreigners, but it is difficult for them to write correct sentence. Grammatical error detection is a big challenge for the Chinese learners as a second language. Learning Chinese sentences will rely too much on the teacher to correct the wrong sentences. It is not easy for learners to get timely feedback. Therefore, how to use existing technology to detect and correct the grammatical errors that learners make has become a hot topic.

Since 2014 (Yu et al.,2014) (Lee et al. 2015) (Lee et al. 2016) (Rao et al., 2017) (Rao et al., 2018), the NLP-TEA workshop provides a series Chinese Grammar Error Detection (CGED) shared tasks to promote the research on grammar error diagnosis. The organizers ask professional teachers to label the errors in learners' sentences. There are four types of label in the sentences: Redundant (R), Selection (S), Disorder (W), and Missing (M). The goal of the task is to build a system that can predict whether a sentence is wrong and correct it. In previous years, we participated in the NLPTEA CGED (Wu et al., 2018) and shows that such a system can be precision oriented or recall oriented for different users.

Since the emerging of deep learning, we find that sequence-to-sequence models have good effect on grammar correction, and the BERT model (Devlin et al., 2019; Xu et al., 2019) is the best sequence-to-sequence pre-training language model using a large number of data sets. The pre-trained model is trained with mask language model (MLM) to enhance the strength of the model.

In Run1 of 2020, we use BERT as the first level of our identification. We fine-tuning the BERT model with the Lang-8[1] corpus and all the data from NLPTEA since 2016 to 2020, so that the model can be used to predict correct and incorrect sentences, and reproduce the wrong sentences. The error types are determined by CRF. In Run2 is not used to determine the wrong and correct sentences. In the following sections, we will introduce related work and our approaches, then discuss the formal test results, and give conclusion and future works.

2 Related Work

Grammar error detection and correction is now a popular research topic in natural language processing (Li et al., 2018; Fu et al., 2018). Previous works show that CRF model can be used to integrate various features to build a good system. Better results can be achieved by using the pre-collected collocation word database.

Recently, researchers use deep learning models to solve this issue. The most common models are sequence-to-sequence (Ge et al., 2018) and convolutional neural network (Li et al., 2019) models. The idea of sequence-to-sequence is to translate the wrong into correct sentences just like translation between two languages. A corrected

[1] https://lang-8.com/

Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 91–96
Suzhou, China, December 4, 2020.

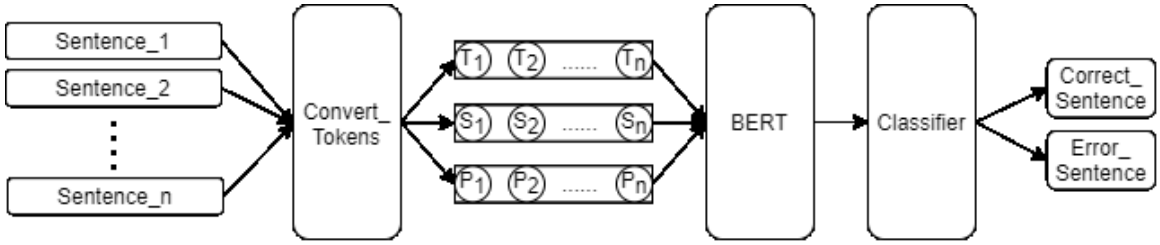

Figure 1 Fine-tune system architecture

sentence is generated from the wrong sentence, and it is believed that multiple revisions will give better results. The convolutional neural network originally is used to process images, now shifted to process text. With the two-dimension processing power, it is easier for the model to read the context of the text.

Since 2019, the BERT gives many state-of-the-art results on several NLP applications. This shows the great influence of BERT on natural language processing. Spelling check is a similar task to the grammar correction (Cheng et al., 2020; Zhang et al., 2020). The authors use the BERT internal model to find typos. Although its effect is not the best, it achieves the purpose the author wants.

3 Method

This year, we mainly focused on minimizing the false alarm on error detection. Since the system is to help foreign learners, we hope that less errors judged by the model will not cause learners to feel frustration.

BERT is a pre-trained language model. Since the original pre-training model was not trained for Chinese grammar correction, we have to train it with our corpus. For different task, better results can be achieved by fine-tuning the pre-trained language models with additional training corpus. Moreover, BERT has achieved excellent results in various projects, such as single classification tasks, sentence-labelling tasks, and question answering tasks. In addition to the BERT model, we also use conditional random fields (CRF) to double check tine wrong sentences detected by BERT, and select the type and location of the errors.

3.1 Fine-tune Language model

We use the BERT pre-training language model provided by huggingface[2]. The pre-train model is "bert-base-chinese". The fine-tuning training data set is the Lang-8 data set provided by NLPCC and all the training data and test data from 2016 to 2020 provided by NLPTEA excepting 2020 test data. Figure 1 shows the BERT fine-tune system architecture. The data set {Sentence _1, Sentence _2, …, Sentence _n} has been preprocessed. Our system compares the original sentence and the modified sentence from the data set. If the sentence is wrong, mark it as "Error_ Sentence ", otherwise mark it as "Correct_ Sentence ". Given a source token = $\{T_1, T_2, ..., T_n\}$ with its segment = $\{S_1, S_2, ..., S_n\}$ and position = $\{P_1, P_2, ..., P_n\}$, we can fine-tune the BERT and obtain the classify results. After classifying the correct and error sentences, the next step the error sentences need input the Conditional Random Fields (CRF). Table 1 shows the number of wrong and correct sentences and their average length in the fine-tuning data set.

	Quantity	Average length
Correct sentence	1,241,126	21
Error sentence	1,117,577	20

Table 1: Fine-tuning data set statistics

3.2 Conditional Random Fields

We use CRF model in both two Runs. Run1 uses a pre-trained language model + CRF and Run2 uses only CRF. We want to see what changes will happen if the pre-trained language model is added. CRF is used to mark the error type and location.

CRF is regarded as a sequence label model. As show in Figure 2, the model will be trained according to the sequence label S we provide, and

[2] https://huggingface.co/transformers/model_doc/bart.html

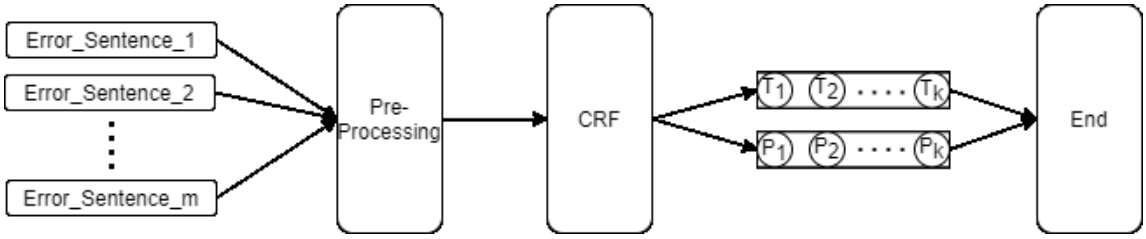

Figure 3 CRF system architecture

the trained model is used to predict the corresponding sequence label Y. The sequence tags we provide to the model contain the words and parts of speech that have been hyphenated by Jieba. The part of speech (POS) can have a very good effect during training. In the first column of the sequence, we place the hyphenated words and the second column. The part of speech of the word, and the label of the wrong type { T_1, T_2, …., T_n} and the position of the word {P_1, P_2, …. , P_n}. Finally, the error type and location are transformed into the format specified by the seminar

3.3 Pre-processing

Figure 3 shows the pre-processing flowchart. The Lang-8 and NLPTEA data are used for fine-tuning the pre-train language model. The sentence before correction must be regarded as an error and the sentence after correction is correct. When preparing the dataset for CRF, our system compares the Lang-8 sentences before and after correction using Jieba segmentation and edit distance. The differences between the two sentences will then be used to determine the three different error types and positions within the edit distance. With the help of Jieba, our system can extract the words in the original sentence and obtain the part-of-speech (POS) tag. The error types include redundant words (R), word selection errors (S), and missing words (M). Next we use the three methods in edit distance. In these methods, insert means missing words, delete means redundant words, and replace means word selection errors. Calculate the position of the wrong word through three ways of editing distance.

We bypass the word ordering errors (W) here because it is very difficult and the training data is too little. The different training materials of NLPTEA and Lang-8 have been marked with error types and positions.

During CRF training, because using too much training data will lead to poor training results, only 57,386 error sentences are used during training.

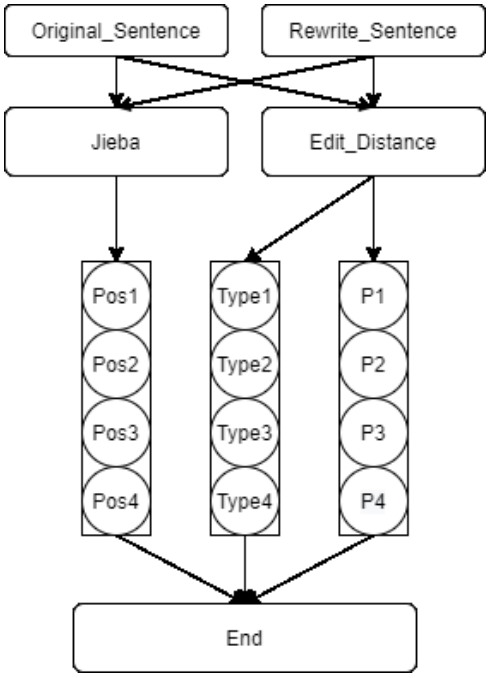

Figure 2 Pre-processing system architecture

Submission	False Positive Rate (the lower the better)
Run1	**0.0163**
Run2	0.5472
Average of 43 runs	0.3920

Table 2: False positive rate in CGED 2020

	Detection Level		
	precision	**Recall**	**F1**
Run1	**0.9875**	0.3443	0.5106
Run2	0.8117	0.6296	0.7091
Average of all 43 runs	0.8922	0.7828	0.8234

Table 3: Detection level in CGED 2020

Each sentence will be processed through Jieba[3] for word segmentation, and finally the corresponding type will be placed in the corresponding position.

4. Experiment Results

4.1 Official test result

Table 2 and Table 3 shows the official results of our system in CGED 2020 shared task. The result shows that our Run1 achieved the highest precision rate 0.9875 with the best false positive rate 0.0163 on detection. In Run2 we improved the recall greatly from 0.3443 to 0.6296 with a drop at the precision rate from 0.9875 to 0.8117. The trade-off of precision and recall is still obvious.

4.2 Error analysis

When encountering sentences that are too long. Our model cannot predict the correct result very well. Here we think that in the fine-tuned training set, it can be seen that the average length is only 20-21. However, as shown in Table 5, the average length of sentences judged by BERT are all above 38 and only a few are below 38. So in the future, we will try to use GPT2 or GPT3 to detect errors in long sentences. Table 4 shows some examples that includes errors but our BERT system fails to detect. As shown in Table 6, we can see the number of errors for the three types of errors. The most numerous are all dependent on one word. Error types R and S almost have similar errors including "的", "是" and "了" and so on. The error type M is mostly punctuation. Because most people usually filter out punctuation because of the convenience of training. Punctuation can make a bad article easier to read. In the future work, we will modify the model towards the above problems.

5. Discussion and Conclusion

In 2020 NLP-TEA CGED shared task, we submitted two Runs, the result shows that our Run1 achieved the highest precision rate (0.9875) with the lowest false positive rate (0.0163) on error detection. The result shows that our system can point out errors with very a high confidence. With very low false alarm, the system can help learners to notice that they really make a grammar error. However, the recall rate of our system is only not very high in Run1. In Run2 we improved the recall greatly from 0.3443 to 0.6296 with a drop at the precision rate from 0.9875 to 0.8117. The trade-off of precision and recall needs more attention.

In the future, we will combine the methods of BERT and GPT2 to improve sentences that our current system cannot detect effectively. About the correction level, we also hope to filter out the best alternative words through GPT2's sentence rewriting method.

# or Sentence	Average length
460	38

Table 5: The sentence statistics of test set

Example	Sentences	Length
1	一个月干下来，大山看上去都没有什么变化。愚公不理会嘲笑，带着全家，继续搬。	37
2	一个兵人暗想：“我要做这个东西了，不然我要被征罚了。”他不暗想：“我不要这样行动因为是不过的”。	48
3	旅行营会把所有的景点、交通和住宿都安排好了。所以自己不能决定哪里去。而且参加旅行营会跟很多不认识的人一起旅行。碰到讨厌的人就没办法，一定要跟他们在一起。所以我喜欢自己一个人旅行。这一样就自由多了。	98
4	星期一上午在大学上课。星期二下牛跟同学一起打排球。	25
5	我有三个哥哥。我最小年轻了。我爸爸妈妈住在法国巴黎。我爸爸是日内瓦大学的老师，所以他很忙。但是，他只教每周只教三天。我妈妈是研究员。她每天上班，所以更忙。	77

Table 4: Examples of long sentences with grammar errors in CGED 2020

[3] https://github.com/fxsjy/jieba

R	R_num	S	S_num	M	M_num
的	433	的	265	。	362
了	308	而	77	，	224
是	198	个	68	不	144
在	98	有	56	的	133
有	82	做	55	一	113
上	63	在	52	我	96
也	51	得	48	是	75
我	43	是	44	有	74
对	42	對	38	很	71
而	42	也	37	人	54
要	39	对	37	这	46
会	37	了	36	在	45

Table 6. NlPTEA CGED 2016 – 2020 Testset, the most frequent errors in each error type

Acknowledgments

This study was supported by the Ministry of Science and Technology under the grant number MOST 109-2221-E-324-024.

References

Xingyi Cheng; Weidi Xu; Kunlong Chen; Shaohua Jiang; Feng Wang; Taifeng Wang; Wei Chu; Yuan Qi, SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p.p 871–881, July 5 - 10, 2020.

Jacob Devlin; Ming-Wei Chang; Kenton Lee; Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019.

Ruiji Fu; Zhengqi Pei; Jiefu Gong; Wei Song; Dechuan Teng; Wanxiang Che; Shijin Wang; Guoping Hu; Ting Liu, Chinese Grammatical Error Diagnosis using Statistical and Prior Knowledge driven Features with Probabilistic Ensemble Enhancement, in Proceedings of The 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia, July 19, 2018.

Tao Ge; Furu Wei; Ming Zhou, Fluency Boost Learning and Inference for Neural Grammatical Error Correction, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, p.p 1-11, Melbourne, Australia, July 15 – 20, 2018.

Lung-Hao Lee, Gaoqi Rao,Liang-Chih Yu, Li-Ping Chang, Xun Endong, Baolin Zhang, Li-Ping Chang. 2016. Overview of the NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'16), Osaka, Japan.

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.

Chen Li; Junpei Zhou; Zuyi Bao; Hengyou Liu; Guangwei Xu; Linlin Li, A Hybrid System for Chinese Grammatical Error Diagnosis and Correction, in Proceedings of The 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia, July 19, 2018.

Si Li; Jianbo Zhao; Guirong Shi; Yuanguap Tan; Huifang Xu; Guang Chen; Haibo Lan; Zhiqing Lin, Chinese Grammatical Error Correction Based on Convolutional Sequence to Sequence Model, IEEE Access., vol. 7, pp. 72905 - 72913, May 17, 2019. doi: 10.1109/ACCESS.2019.2917631

Gaoqi Rao, Baolin Zhang, Endong Xun. 2017. IJCNLP-2017 Task1: Chinese Grammatical Error Diagnosis. 8th International Joint Conference of Nature Language Processing (IJCNLP2017), Taipei, Taiwan.

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, Melbourne, Australia, July 19, 2018.

Shih-Hung Wu; Jun-Wei Wang; Liang-Pu Chen; Ping-Che Yang, CYUT-III Team Chinese Grammatical Error Diagnosis System Report in NLPTEA-2018 CGED Shared Task, in Proceedings of The 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia, July 19, 2018.

Hu Xu; Bing Liu; Lei Shu; Philip S. Yu, BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis, arXiv:1904.02232v2,May 4, 2019.

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.

Shaohua Zhang; Haoran Huang; Jicong Liu; Hang Li, Spelling Error Correction with Soft-Masked BERT, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 882–890, July 5 - 10, 2020.