Chinese Grammatical Correction Using BERT-based Pre-trained Model

Hongfei Wang, Michiki Kurosawa[*], **Satoru Katsumata**[†] and **Mamoru Komachi**
Tokyo Metropolitan University
wang-hongfei@ed.tmu.ac.jp, m-kurosawa@nri.co.jp
satoru.katsumata@retrieva.jp, komachi@tmu.ac.jp

Abstract

In recent years, pre-trained models have been extensively studied, and several downstream tasks have benefited from their utilization. In this study, we verify the effectiveness of two methods that incorporate a BERT-based pre-trained model developed by Cui et al. (2020) into an encoder-decoder model on Chinese grammatical error correction tasks. We also analyze the error type and conclude that sentence-level errors are yet to be addressed.

1 Introduction

Grammatical error correction (GEC) can be regarded as a sequence-to-sequence task. GEC systems receive an erroneous sentence written by a language learner and output the corrected sentence. In previous studies that adopted neural models for Chinese GEC (Ren et al., 2018; Zhou et al., 2018), the performance was improved by initializing the models with a distributed word representation, such as Word2Vec (Mikolov et al., 2013). However, in these methods, only the embedding layer of a pre-trained model was used to initialize the models.

In recent years, pre-trained models based on Bidirectional Encoder Representations from Transformers (BERT) have been studied extensively (Devlin et al., 2019; Liu et al., 2019), and the performance of many downstream Natural Language Processing (NLP) tasks has been dramatically improved by utilizing these pre-trained models. To learn existing knowledge of a language, a BERT-based pre-trained model is trained on a large-scale corpus using the encoder of Transformer (Vaswani et al., 2017). Subsequently, for a downstream task, a neural network model is initialized with the weights learned by a pre-trained model that has the same structure and is fine-tuned on training data of the downstream task. Using this two-stage method, the performance is expected to improve because downstream tasks are informed by the knowledge learned by the pre-trained model.

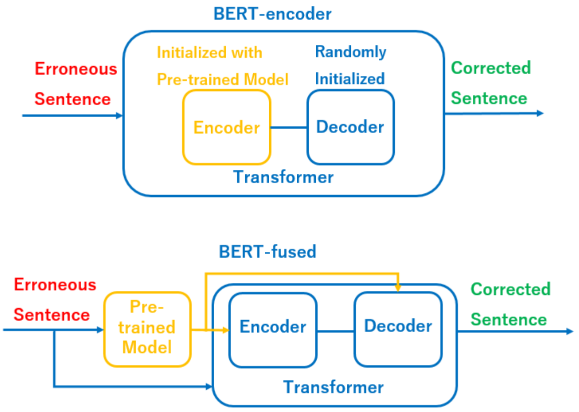

Figure 1: Two methods for incorporating a pre-trained model into the GEC model.

Recent works (Kaneko et al., 2020; Kantor et al., 2019) show that BERT helps improve the performance on the English GEC task. As the Chinese pre-trained models are developed and released continuously (Cui et al., 2020; Zhang et al., 2019), the Chinese GEC task may also benefit from using those pre-trained models.

In this study, as shown in Figure 1, we develop a Chinese GEC model based on Transformer with a pre-trained model using two methods: first, by initializing the encoder with the pre-trained model (BERT-encoder); second, by utilizing the technique proposed by Zhu et al. (2020), which uses the pre-trained model for additional features (BERT-fused); on the Natural Language Processing and Chinese Computing (NLPCC) 2018 Grammatical Error Correction shared task test dataset (Zhao et al., 2018), our single models obtain $F_{0.5}$ scores of 29.76 and 29.94 respectively, which is similar to the performance of ensemble models developed

[*] Currently at Nomura Research Institute, Ltd.
[†] Currently at Retrieva, Inc.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 163–168
December 4 - 7, 2020.

by the top team of the shared task. Moreover, using a 4-ensemble model, we obtain an $F_{0.5}$ score of 35.51, which outperforms the results from the top team by a large margin. We annotate the error types of the development data; the results show that word-level errors dominate all error types and that sentence-level errors remain challenging and require a stronger approach.

2 Related Work

Given the success of the shared tasks on English GEC at the Conference on Natural Language Learning (CoNLL) (Ng et al., 2013, 2014), a Chinese GEC shared task was performed at the NLPCC 2018. In this task, approximately one million sentences from the language learning website Lang-8[1] were used as training data and two thousand sentences from the PKU Chinese Learner Corpus (Zhao et al., 2018) were used as test data. Here, we briefly describe the three methods with the highest performance.

First, Fu et al. (2018) combined a 5-gram language model-based spell checker with subword-level and character-level encoder-decoder models using Transformer to obtain five types of outputs. Then, they re-ranked these outputs using the language model. Although they reported a high performance, several models were required, and the combination method was complex.

Second, Ren et al. (2018) utilized a convolutional neural network (CNN), such as in Chollampatt and Ng (2018). However, because the structure of the CNN is different from that of BERT, it cannot be initialized with the weights learned by the BERT.

Last, Zhao and Wang (2020) proposed a dynamic masking method that replaces the tokens in the source sentences of the training data with other tokens (e.g. [PAD] token). They achieved state-of-the-art results on the NLPCC 2018 Grammar Error Correction shared task without using any extra knowledge. This is a data augmentation method that can be a supplement for our study.

3 Methods

In the proposed method, we construct a correction model using Transformer, and incorporate a Chinese pre-trained model developed by Cui et al. (2020) in two ways as described in the following sections.

3.1 Chinese Pre-trained Model

We use a BERT-based model as our pre-trained model. BERT is mainly trained with a task called Masked Language Model. In the Masked Language Model task, some tokens in a sentence are replaced with masked tokens ([MASK]) and the model has to predict the replaced tokens.

In this study, we use the Chinese-RoBERTa-wwm-ext model developed by Cui et al. (2020). The main difference between Chinese-RoBERTa-wwm-ext and the original BERT is that the latter uses whole word masking (WWM) to train the model. In WWM, when a Chinese character is masked, other Chinese characters that belong to the same word should also be masked.

3.2 Grammatical Error Correction Model

In this study, we use Transformer as the correction model. Transformer has shown excellent performance in sequence-to-sequence tasks, such as machine translation, and has been widely adopted in recent studies on English GEC (Kiyono et al., 2019; Junczys-Dowmunt et al., 2018).

However, a BERT-based pre-trained model only uses the encoder of Transformer; therefore, it cannot be directly applied to sequence-to-sequence tasks that require both an encoder and a decoder, such as GEC. Hence, we incorporate the encoder-decoder model with the pre-trained model in two ways as described in the following subsections.

BERT-encoder We initialize the encoder of Transformer with the parameters learned by Chinese-RoBERTa-wwm-ext; the decoder is initialized randomly. Finally, we fine-tune the initialized model on Chinese GEC data.

BERT-fused Zhu et al. (2020) proposed a method that uses a pre-trained model as the additional features. In this method, input sentences are fed into the pre-trained model and representations from the last layer of the pre-trained model are acquired first. Then, the representations will interact with the encoder and decoder by using attention mechanism. Kaneko et al. (2020) verified the effectiveness of this method on English GEC tasks.

4 Experiments

4.1 Experimental Settings

Data In this study, we use the data provided by the NLPCC 2018 Grammatical Error Correction

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

shared task. We first segment all sentences into characters because the Chinese pre-trained model we used is character-based. In the GEC task, source and target sentences do not tend to change significantly. Considering this, we filter the training data by excluding sentence pairs that meet the following criteria: i) the source sentence is identical to the target sentence; ii) the edit distance between the source sentence and the target sentence is greater than 15; iii) the number of characters of the source sentence or the target sentence exceeds 64. Once the training data were filtered, we obtained 971,318 sentence pairs.

Because the NLPCC 2018 Grammatical Error Correction shared task did not provide development data, we opted to randomly extract 5,000 sentences from the training data as the development data following Ren et al. (2018).

The test data consist of 2,000 sentences extracted from the PKU Chinese Learner Corpus. According to Zhao et al. (2018), the annotation guidelines follow the minimum edit distance principle (Nagata and Sakaguchi, 2016), which selects the edit operation that minimizes the edit distance from the original sentence.

Model We implement the Transformer model using fairseq 0.8.0.[2] and load the pre-trained model using pytorch_transformer 2.2.0.[3]

We then train the following models based on Transformer.

Baseline: a plain Transformer model that is initialized randomly without using a pre-trained model.

BERT-encoder: the correction model introduced in Section 3.2.

BERT-fused: the correction model introduced in Section 3.2. We use the implementation provided by Zhu et al. (2020).[4]

Finally, we train a 4-ensemble BERT-encoder model and a 4-ensemble BERT-fused model.

More details on the training are provided in the appendix A.

Evaluation As the evaluation is performed on word-unit, we strip all delimiters from the system output sentences and segment the sentences using the pkunlp[5] provided in the NLPCC 2018 Grammatical Error Correction shared task.

[Our models]	**P**	**R**	**$\mathrm{F}_{0.5}$**
Baseline	25.14	14.34	21.85
BERT-encoder	32.67	22.19	29.76
BERT-fused	32.11	**23.57**	29.94
BERT-encoder (4-ensemble)	41.94	22.02	35.51
BERT-fused (4-ensemble)	32.20	23.16	29.87
[SOTA Result]			
Zhao and Wang (2020)	44.36	22.18	**36.97**
[NLPCC 2018]			
Fu et al. (2018)	35.24	18.64	29.91
Ren et al. (2018)	41.73	13.08	29.02
Ren et al. (2018) (4-ensemble)	**47.63**	12.56	30.57

Table 1: Experimental results on the NLPCC 2018 Grammatical Error Correction shared task.

Based on the setup of the NLPCC 2018 Grammatical Error Correction shared task, the evaluation is conducted using *MaxMatch* (M2).[6]

4.2 Evaluation Results

Table 1 summarizes the experimental results of our models. We run the single models four times, and report the average score. For comparison, we also cite the result of the state-of-the-art model (Zhao and Wang, 2020) and the results of the models developed by two teams in the NLPCC 2018 Grammatical Error Correction shared task.

The performances of BERT-encoder and BERT-fused are significantly superior to that of the baseline model and are comparable to those achieved by the two teams in the NLPCC 2018 Grammatical Error Correction shared task, indicating the effectiveness of adopting the pre-trained model.

The BERT-encoder (4-ensemble) model yields an $\mathrm{F}_{0.5}$ score nearly 5 points higher than the highest-performance model in the NLPCC 2018 Grammatical Error Correction shared task. However, there is no improvement for the BERT-fused (4-ensemble) model compared with the single BERT-fused model. We find that the performance of the BERT-fused model depends on the warm-up model. Compared with Kaneko et al. (2020) using a state-of-the-art model to warm-up their BERT-fused model, we did not use a warm-up model in this work. The performance noticeably drops when we try to warm-up the BERT-fused model from a weak baseline model, therefore, the BERT-fused model may perform better when warmed-up from a

[2] https://github.com/pytorch/fairseq

[3] https://github.com/huggingface/transformers

[4] https://github.com/bert-nmt/bert-nmt

[5] http://59.108.48.12/lcwm/pkunlp/downloads/libgrass-ui.tar.gz

[6] https://github.com/nusnlp/m2scorer

src	**持 别** 是 北京 ， 没有 “ 自然 ” 的 感觉 。	人们 在 一 辈子 **经验** 很多 事情 。
gold	**特别** 是 北京 ， 没有 “ 自然 ” 的 感觉 。	人们 在 一 辈子 **经历** 很多 事情 。
baseline	**持 别** 是 北京 ， 没有 “ 自然 ” 的 感觉 。	人们 在 一辈子 **经历** 了 很多 事情 。
BERT-encoder	**特别** 是 北京 ， 没有 “ 自然 ” 的 感觉 。	人们 一辈子 **会 经历** 很多 事情 。
Translation	**Especially** in Beijing, there is no *natural* feeling.	People **experience** many things in their lifetime.

Table 2: Source sentence, gold edit, and output of our models.

Error Type	**Number of errors**	**Examples**
B	9	最后 ， 要 关主{**关注**} 一些 关于 天气 预报 的 新闻 。 (Finally, pay attention to some weather forecast news.)
CC	35	有 一 天 晚上 他 下 了 决定{**决心**} 向 富丽 堂皇 的 宫殿 里 走 ， 偷偷 的{**地**} 进入 宫内 。 (One night he decided to walk to the magnificent palace, and sneaked in it secretly.)
CQ	30	在 上海 我 总是 住 NONE{**在**} 一家 特定 NONE{**的**} 酒店 。 (I always stay in the same hotel in Shanghai.)
CD	21	我 很 喜欢 念{**NONE**}读 小说 . (I like to read novels.)
CJ	35	…… 但是 同时 也 对 环境 问题{**NONE**} 日益 严重 造成 了{**造成 了 日益 严重 的**} 空气 污染 问题 。 (But on the meanwhile, it also aggravated the problem of air pollution.)

Table 3: Examples of each error type. The underlined tokens are detected errors that should be replaced with the tokens in braces.

Type	**Detection**			**Correction**		
	P	**R**	**F$_{0.5}$**	**P**	**R**	**F$_{0.5}$**
BERT-encoder						
B	**80.0**	**55.6**	**73.5**	**80.0**	**55.6**	**73.5**
CC	62.5	31.4	52.2	43.8	20.0	35.4
CQ	65.0	43.3	59.1	45.0	30.0	40.9
CD	58.3	28.6	48.3	50.0	28.6	43.5
CJ	56.5	42.9	53.1	4.3	2.9	3.9
BERT-fused						
B	**80.0**	44.4	**69.0**	**80.0**	44.4	**69.0**
CC	61.9	42.9	56.9	38.1	22.9	33.6
CQ	69.0	**63.3**	67.8	44.8	**46.7**	45.2
CD	71.4	42.9	63.0	57.1	38.1	51.9
CJ	63.2	34.3	54.1	15.8	8.6	13.5

Table 4: Detection and correction performance of BERT-encoder and BERT-fused models on each type of error.

stronger model (e.g., the model proposed by Zhao and Wang (2020)).

For the state-of-the-art result achieved by Zhao and Wang (2020), both the precision and the recall are comparatively high, and they therefore obtain the best F$_{0.5}$ score.

Additionally, the precision of the models that used a pre-trained model is lower than that of the models proposed by the two teams; conversely, the recall is significantly higher.

5 Discussion

Case Analysis Table 2 shows the sample outputs.

In the first example, the spelling error 持别 is accurately corrected to 特别 (which means *especially*) by the proposed model, whereas it is not corrected by the baseline model. Hence, it appears that the proposed model captures context more efficiently by using the pre-trained model through the WWM strategy.

In the second example, the output of the proposed model is more fluent, although the correction made by the proposed model is different from the gold edit. The proposed model not only changed the wrong word 经验 (which usually means the noun *experience*) to 经历 (which usually means the verb *experience*), but also added a new word 会 (*would, could*); this addition makes the sentence more fluent. It appears that the proposed model can implement additional changes to the source sentence because the pre-trained model is trained with a large-scale corpus. However, this type of change may affect the precision because the gold edit in this dataset followed the principle of minimum edit distance (Zhao et al., 2018).

Error Type Analysis To understand the error distribution of Chinese GEC, we annotate 100 sentences of development data and obtain 130 errors (one sentence may contain more than one error). We refer to the annotation of the HSK learner corpus[7] and adopt five categories of error: B, CC, CQ, CD, and CJ. B denotes character-level errors, which are mainly spelling and punctuation errors. CC, CQ, and CD are word-level errors, which are word selection, missed word, and redundant word errors, respectively. CJ denotes sentence-level errors which contain several complex errors, such as word order and lack of subject errors. Several

[7] http://hsk.blcu.edu.cn/

examples are presented in Table 3. Based on the number of errors, it is evident that word-level errors (CC, CQ, and CD) are the most frequent.

Table 4 lists the detection and correction results of the BERT-encoder and BERT-fused models for each error type. The two models perform poorly on sentence-level errors (CJ), which often involve sentence reconstructions, demonstrating that this is a difficult task. For character-level errors (B), the models achieve better performance than for other error types. Compared with the correction performance, the systems indicate moderate detection performance, demonstrating that the systems address error positions appropriately. With respect to the difference in performance of the two systems on each error type, we can conclude that BERT-encoder performs better on character-level errors (B), and BERT-fused performs better on other error types.

6 Conclusion

In this study, we incorporated a pre-trained model into an encoder-decoder model using two methods on Chinese GEC tasks. The experimental results demonstrate the usefulness of the BERT-based pre-trained model in the Chinese GEC task. Additionally, our error type analysis showed that sentence-level errors remain to be addressed.

Acknowledgments

This work has been partly supported by the programs of the Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS KAKENHI) Grant Numbers 19K12099 and 19KK0286.

References

Shamil Chollampatt and Hwee Tou Ng. 2018. A multi-layer convolutional encoder-decoder neural network for grammatical error correction. In *AAAI*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Kai Fu, Jun Huang, and Yitao Duan. 2018. Youdao's winning solution to the NLPCC-2018 task 2 challenge: A neural machine translation approach to Chinese grammatical error correction. In *NLPCC*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *NAACL-HLT*.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *ACL*.

Yoav Kantor, Yoav Katz, Leshem Choshen, Edo Cohen-Karlik, Naftali Liberman, Assaf Toledo, Amir Menczel, and Noam Slonim. 2019. Learning to combine grammatical error corrections. In *BEA@ACL*.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *EMNLP-IJCNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Ryo Nagata and Keisuke Sakaguchi. 2016. Phrase structure annotation and parsing for learner English. In *ACL*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *CoNLL*.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *CoNLL*.

Hongkai Ren, Liner Yang, and Endong Xun. 2018. A sequence to sequence learning for Chinese grammatical error correction. In *NLPCC*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *ACL*.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *NLPCC*.

Zewei Zhao and Houfeng Wang. 2020. MaskGEC: Improving neural grammatical error correction via dynamic masking. In *AAAI*.

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. 2018. Chinese grammatical error correction using statistical and neural models. In *NLPCC*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating BERT into neural machine translation. In *ICLR*.

A Appendices

Table 5 shows the training details for each model.

Baseline	
Architecture	Encoder (12-layer), Decoder (12-layer)
Learning rate	1×10^{-5}
Batch size	32
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$)
Max epochs	20
Loss function	cross-entropy
Dropout	0.1
BERT-encoder	
Architecture	Encoder (12-layer), Decoder (12-layer)
Learning rate	3×10^{-5}
Batch size	32
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$)
Max epochs	20
Loss function	cross-entropy
Dropout	0.1
BERT-fused	
Architecture	Transformer (big)
Learning rate	3×10^{-5}
Batch size	32
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Max epochs	20
Loss function	label smoothed cross-entropy ($\epsilon_{ls} = 0.1$)
Dropout	0.3

Table 5: Training details for each model.