IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 31, 2019, accepted April 22, 2019, date of publication May 17, 2019, date of current version June 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917631

Chinese Grammatical Error Correction Based on Convolutional Sequence to Sequence Model

SI LI[1], JIANBO ZHAO[1], GUIRONG SHI[2], YUANPENG TAN[3], HUIFANG XU[3], GUANG CHEN[1], HAIBO LAN[2], AND ZHIQING LIN[1]

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]State Grid Jibei Electric Power Company Limited, Beijing 100053, China
[3]China Electric Power Research Institute, Beijing 100192, China

Corresponding author: Jianbo Zhao (zhaojianbo@ bupt.edu.cn)

This work was supported by the State Grid Corporation of China through the Project Research on Key Technologies of Knowledge Graph in Power System Fault Management under Grant 52010119000F.

ABSTRACT Chinese grammatical error correction (CGEC) is practically useful for learners of Chinese as a second language, but it is a rather challenging task due to the complex and flexible nature of Chinese language so that existing methods for English cannot be directly applied. In this paper, we introduce a convolutional sequence to sequence model into the CGEC task for the first time, since many Chinese grammatical errors are concentrated between three and four words and convolutional neural network can better capture the local context. A convolution-based model can obtain the representations of the context by fixed size kernel. By stacking convolution layers, long-term dependences can be obtained. We also propose two optimization methods, shared embedding and policy gradient, to optimize the convolutional sequence to sequence model through sharing parameters and reconstructing loss function. Besides, we collate the existing Chinese grammatical correction corpus in detail. The results show that the models we proposed two different optimization methods both achieve large improvement compared with the natural machine translation model based on a recurrent neural network.

INDEX TERMS Chinese grammatical error correction, sequence to sequence, convolutional.

I. INTRODUCTION

Nowadays, more and more people have learned Chinese as their second foreign language. Meanwhile, Chinese writing has received more and more attention. Grammatical error, also called usage error, is a big challenge for the Chinese learners as a second language. Traditional learning methods for learning Chinese writing rely on teachers or others to make corrections to wrong sentences. This is time consuming and labor intensive for a large number of writing text corrections. It is also impossible for learners to get timely feedback. Therefore, people are beginning to pay attention to automatic grammatical error corrections. A lot of research [1]–[3] has been done for English automatic grammatical error correction. The works of Chinese grammatical error correction (CGEC) [4] are limited by the lack of corpus and are still in exploring. Some shared tasks have been conducted for English [5], [6] and Chinese [4], [7] to promote the research on grammatical error correction. In this paper, we focus on Chinese grammatical error correction.

Chinese grammatical errors are different from English grammatical errors in terms of commonly observed mistakes. For example, in English, the improper use of article like ''a'' and ''the'', the error use of singular and plural form, and spelling mistakes often appear. However, misnomers and improper use of auxiliary words are the two most common cases of Chinese writing errors, which is different from English. Similar pronunciation, character form and meaning in Chinese words lead to misnomers. Auxiliary words are sometimes as the essential part of Chinese grammar, however, these words may have no clear semantics, such as ''了'' and ''的''. Correction of English grammatical error needs to pay attention to not only the errors between words, but also the errors inside of the words. Correction of Chinese grammatical error focuses more on the fixed match error and the overall sentence structure. Grammatical errors in Chinese sentences are defined as four types according to NLPTEA 2016 shared task [8] in accordance with the method of correction, redundant words (denoted as a capital ''R''), missing

The associate editor coordinating the review of this manuscript and approving it for publication was Zhanyu Ma.

TABLE 1. Examples of each error type in Chinese grammatical errors.

Error Type	Incorrect Sentence	Correct Sentence
R	这样可以带来了好处。(This can brought benefits.)	这样可以带来好处。(This can bring benefits.)
M	我来谈谈我老师教我的知识。(Let me talk about what I teacher taught me.)	我来谈谈我的老师教我的知识。(Let me talk about what my teacher taught me.)
S	实际上这种措施还没实现。(In fact, this measure has not been realized.)	实际上这种情况还没实现。(In fact, this situation has not been realized.)
W	我们拿自己的钱应该救他们的命。(We save their lives should with our own money.)	我们应该拿自己的钱救他们的命。(We should save their lives with our own money.)

words (''M''), word selection errors (''S''), and word ordering errors (''W''). Examples of grammatical error are shown in Table 1.

Grammatical error correction is usually considered as a translation task [2]. Sentences with grammatical errors are translated into correct sentences. The most commonly used grammatical error correction model is the sequence to sequence model based on recurrent neural network (RNN) [2]. Unlike previous models, we use convolutional sequence to sequence model [9]. Many Chinese grammatical errors are concentrated between three and four words and convolutional neural network (CNN) can better capture the local context. Therefore, the CNN-based model is more suitable for Chinese grammatical error correction. Convolution obtains representations of the context through fixed size windows. The maximum length of the dependencies to be modeled can be more precisely controlled by stacking convolution layers. The convolutional network no longer relies on the output of the previous time step, so parallel calculations can be performed to reduce computation time. Based on the basic convolutional model, we propose two optimization methods to make the model perform better.

Our main contribution consists of four parts. First, we introduce the convolutional sequence to sequence model to CGEC task for the first time. Second, we add shared embedding and policy gradient respectively on the basis of the convolutional sequence to sequence model. Third, we organize and analyze existing corpus for Chinese grammatical error correction. Finally, the results of experiments show that the models we proposed of two different optimization methods are effective in Chinese grammatical error correction and achieve better results in all evaluation methods compared with basic model.

II. RELATED WORK

Grammatical error correction, a sequence-to-sequence problem, now is often seen as a translation task [2]. Translate sentences containing grammatical errors into correct sentences. Represent words as word vectors [10]. The maximum probability [11], [12] of the current word is determined by the original sentence and the previous word.

English grammatical error correction has developed for a long time. CoNLL-2013 shared task pays attention to error correction in English and classifies various grammatical errors in detail. Xiang *et al.* [1] combined machine learning and rule-based methods together to correct five types of errors, determiner, preposition, noun number, verb form and subject-verb agreement. Yuan and Briscoe [2] first introduced translation model to grammatical error correction task. Sakaguchi *et al.* [13] trained grammatical error correction model with reinforce learning. Ji *et al.* [14] proposed a new hybrid neural model with nested attention layers for grammatical error correction (GEC) which can correct errors by incorporating word and character-level information. Due to using a wrong verb is the most common grammatical errors, Wu *et al.* [15] described a system for detecting and correcting potential verb errors in a given sentence. Lo *et al.* [16] presented a GEC system which trained on EF-Cambridge Open Language Database (EFCAMDAT), a large learner corpus annotated with grammatical errors and corrections. Ge *et al.* [3] used fluency boosting learning which generated fluency-boost sentence pairs during training, improving the performance of the error correction model. New evaluation methods have also been proposed. Chollampatt and Ng [17] proposed a sentence-level analysis indicates that comparing GLEU and M2, one metric may be more useful than the other depending on the scenario. Chollampatt and Ng [18] proposed the first neural approach to automatic quality estimation of GEC output sentences that did not employ any hand-crafted features. The approach trained in a supervised manner on learner sentences and corresponding GEC system outputs with quality score labels computed using human-annotated references.

Compared with English, Chinese grammatical error correction, due to the lack of corpus of corresponding sentence pairs, often used statistical method in the early stage. In recent years, machine learning has been applied to Chinese grammatical error correction task, such as CNN [19], [20] and RNN [21], [22]. Hu *et al.* [23] presented statistical data and performed analysis in detail on basic information of short message corpus and built an automatic error correction system to correct the misapplication of Chinese characters in short messages. Yu *et al.* [24] corrected sentences by providing candidate corrections for all or partially identified characters in a sentence, and scoring all altered sentences and identifying which was the best corrected sentence. Chang *et al.* [25] used phonological similarity and orthographic similarity co-occurrence to train linear regression model to detect and correct misspelled words in documents. Cheng *et al.* [26] focused on word ordering errors with SVM.

Xiong *et al.* [27] proposed a unified framework for Chinese essays spelling correction based on extended HMM and ranker-based models, together with a rule-based model. Chen *et al.* [28] measured the likelihood of correction candidates generated by deleting or inserting characters or words, moving substrings to different positions, substituting prepositions with other prepositions, or substituting words with their synonyms or similar strings. Shiue *et al.* [29] treated the error correction task as a translation task from erroneous Chinese to well-formed Chinese. Li *et al.* [30] proposed a hybrid system with two stages: the detection stage with BiLSTM-CRF and GEC models and the correction stage. GEC models contained rule-based model, NMT model and SMT model. Fu *et al.* [31] regarded the CGEC task as a translation problem which translated the wrong sentence into the correct one. Fu *et al.* [32] built the detection model through bidirectional Long Short-Term Memory with a conditional random field layer (BiLSTM-CRF) and the correction model based on the ePMI values and seq2seq model.

III. CONVOLUTIONAL SEQUENCE TO SEQUENCE MODEL

A. MODEL

Current grammatical error correction task is often considered as a translation problem. For grammatical error correction, the source sentences of translation are sentences containing grammatical errors and the target sentences are the corrected sentences. The aim of translation task is to convert the two languages, so the words in source language and the target language basically do not overlap. Different from the translation problem, for grammatical error correction problem, the source language is consistent with the target language.

Different from English, Chinese sentences have no spaces, which means that word segmentation has not been made for the original sentences. Although the word is the smallest semantic unit of Chinese, the number of commonly used words is huge and sparse problem may exist. Compared to words, the number of commonly used characters in Chinese is relatively small, and characters also contain certain semantic information. So in our model, we use characters as the smallest unit.

The embedding consists of two parts, one is the character embedding and the other is the position embedding, for both the input of encoder and the output elements generated by decoder. The embedding is represented by $v_i = c_i + p_i$, where c_i and p_i represent the character embedding and position embedding respectively. The calculation of convolution does not contain sequence absolute position information like RNN. The absolute position of the word or character in the sequence facilitates the translation task of the sequence [9], so the position embedding is combined with character embedding. Both embeddings are trained with other parameters in the network.

Both encoder and decoder share same block structure which contains a one dimensional convolution followed by a non-linearity. The convolution kernel can only focus on the context with a window size of k. Stacking convolution layer can increase the number of input elements represented in a state. Consider the input context $S \in R^{k \times d}$, where d is the dimension of the word embedding. $2d$ convolution kernels of size $k \times d$ are used to convolve the input context, and the output is represented as $Y \in R^{2 \times d}$. Gated linear units (GLU) [33] is used as the following non-linearity. Y can also be represented as $[Y_1, Y_2]$, each has a dimension d. The output of the convolutional layer can be calculated as (1),

$$f(Y) = Y_1 \cdot \sigma(Y_2), \tag{1}$$

where $f(Y)$ represents the output representation, $\cdot$ is an element-wise multiplication and σ refers to the non-linearity. The role of non-linearity is similar to the gate function of long short-term memory (LSTM) module which can control the correlation between the current input and the current context simply.

Residual connection is added to the convolutional blocks to enable the multi-layer convolutional network contain the underlying information to have better performance as (2),

$$h_i^l = f(Y) + h_i^{l-1}, \tag{2}$$

where h_i means the output of encoder, l refers to the convolutional layer.

For decoder, two padding symbols are added at the beginning when generate the target sentence. In each decoder layer, convolution is also followed by a non-linearity. Residual connection is also added like (3).

$$w_i^l = f(Z) + w_i^{l-1}, \tag{3}$$

where w_i^l represents the decoder state at time step i at layer l. Z represents the output of decoder after convolution.

Multi-step attention is used in each decoder layer. The calculation of attention weights is shown as (5) and (5).

$$s_i^l = W_d^l w_i^l + b_d^l + t_i, \tag{4}$$

$$a_{ij}^l = \frac{exp(s_i^l \cdot h_j^u)}{\sum_{t=1}^{m} exp(s_i^l \cdot h_t^u)}. \tag{5}$$

To compute attention weights, decoder state summary s_i^l is needed to compute by the current decoder state w_i^l and previous target element t_i. The attention weight is computed as a dot product of the decoder state s_i^l and each output h_j^u of the last encoder block u.

The calculation of the source context vector q_i^l is the weighted summation of the encoder outputs and input embeddings as (6).

$$q_i^l = \sum_{j=1}^{m} a_{ij}^l (h_j^u + v_j). \tag{6}$$

The final decoder output vector is map to e_i whose dimension is the target vocabulary size. Dropout is used before every encoder and decoder layer, decoder output, and embedding layer. Softmax is used to calculate the probability for each character.

B. POLICY GRADIENT

We introduce Monte-Carlo policy gradient[1] method from reinforce algorithm to optimize our model to address the problem of non-differentiable in text generation. After using the policy gradient, the objective function aims to maximize the reward when generating the sentence, as illustrate in (7),

$$L_M = \Sigma_{Y_{1:T}} M_\theta(Y_{1:T}|X) \cdot R(Y_{1:T}, X) \tag{7}$$

where L_M is the objective function with policy gradient, $Y_{1:T}$ refers to the sentence of length T generated by our model M, R is a Monte-Carlo estimation of the reward. The reward we defined as the score of GLEU.[2] Since the loss of the generation model is calculated after the target sentence is predicted. So we multiply loss for one sentence and reward for one sentence directly.

C. SHARED EMBEDDING

The biggest difference between the grammatical error detection model and the translation model is the difference between the source language and the target language. The source language of the translation model is different from the target language and the process is translating the source language into the target language. In the grammatical error correction task, the source language is the same as the target language. The task is to translate a sentence containing grammatical errors into a correct sentence. So in the sequence to sequence model, the encoder and decoder can share same embeddings which contain meanings of words or characters. The method of sharing embedding optimizes embedding parameters at both the encoder layer and the decoder layer.

IV. EXPERIMENT

A. DATASET

The dataset we used is from NLPCC2018 shared task [4] and NLPTEA shared task [7].

The corpus published by NLPTEA[3] is derived from the HSK corpus[4] which is a Chinese proficiency test. This corpus provides not only the corrected sentences, but also the location and type of the errors. The advantage of the corpus is that the data has high quality since the correct sentence is modified by experts and there are almost no errors in correction information. However, this dataset is too small to support the grammatical error correction task.

The corpus provided by NLPCC2018[5] is derived from lang-8.[6] Since the corrections provided by lang-8 are mostly from netizens who use Chinese as their native language, the dataset is relatively confusing. There is a case in the corpus that the error sentence does not match the correct sentence. Some corrections in the corpus are just corrective methods and there are no complete corrected sentences.

[1]https://github.com/ZhenYangIACAS/NMT_GAN
[2]http://www.nltk.org/api/nltk.translate.html
[3]http://www.cged.science
[4]http://bcc.blcu.edu.cn/hsk
[5]http://github.com/zhaoyyoo/NLPCC2018_GEC
[6]http://lang-8.com/

TABLE 2. Information of corpus: Correct means the number of sentences which do not contain any errors. Error means the number of error sentences in the corpus. Train, valid and test are the datasets we use in our experiment.

Corpus	NumOfPairs	Correct	Error
NLPTEA2016	14254	102	14152
NLPTEA2017	16170	90	16080
NLPTEA2018	634	8	626
NLPCC2018	1172014	123405	1048609
Train	1193115	122587	1070528
Valid	9957	1018	8939
Test	1967	16	1951

At the same time, there are sentences written in traditional Chinese. Therefore, cleaning is required for NLPCC2018 corpus. In order to ensure that the data is sufficient, traditional Chinese is changed into simplified Chinese by wiki,[7] since our task is for simplified Chinese. At the same time, in order to avoid the mismatch problem of data pairs, the sentence pair will be discarded if the length of the corrected sentence exceeds 1.5 times the length of the error sentence.

Chinese sentences written by foreigners generally do not exceed 75 characters in length, so we keep sentences with less than 75 characters. The cleaned corpus is divided into two parts, training set and validation set. The test set in NLPCC2018 is used directly as our test set. The test set is annotated correction information by two experts.

Detailed information is shown in Table 2.

B. HYPER PARAMETERS

Our code is based on the public pytorch-Fairseq code.[8] The parameters used in experiments are the default settings in Fairseq. The initial learning rate is set to 0.25. The input dimensions of encoder and decoder are both 512. The character embedding is initialized randomly. The optimizer we used is Nesterov's Accelerated Gradient Descent (NAG) [34] with a simplified formulation for Nesterov's momentum. When generate the target sentence, the beam is set to 5.

C. EVALUATION CRITERIA

Precision, recall and $F_{0.5}$ are used to evaluate the model which are computed with MaxMatch (M_2) scorer. M_2 algorithm is widely used method for grammatical error correction task which prefers to choose the hypothesis that holds highest overlap with the gold edits from annotators.

Here, $\{a_1, a_2, \ldots, a_n\}$ represents the gold edit set from annotator and $\{s_1, s_2, \ldots, s_n\}$ represents the system edit set. The calculation is shown as follows.

$$P = \frac{\sum_{i=1}^{n} |a_i \cap s_i|}{\sum_{i=1}^{n} |s_i|}, \tag{8}$$

$$R = \frac{\sum_{i=1}^{n} |a_i \cap s_i|}{\sum_{i=1}^{n} |a_i|}, \tag{9}$$

[7]http://zh.wikipedia.org/wiki/Wikipedia:繁简处理
[8]https://github.com/pytorch/fairseq

TABLE 3. Performance of models: Annotator 1&2 refers to the gold sentences corrected by two experts. NMT_{RNN} is the natural machine translation model based on RNN. $CGEC_{CSS}$ is the convolutional sequence to sequence model. +share represents the model with shared embedding. +policy represents the model with policy gradient. $CGEC_{CSS_4}$, $CGEC_{CSS_6}$ and $CGEC_{CSS_8}$ represent models are trained based on the corpus whose ratio of error sentences to correct sentences is 4:1, 6:1, and 8:1 respectively.

Model	Annotator 1			Annotator 2			Annotator 1&2		
	Precision	Recall	$F_{0.5}$	Precision	Recall	$F_{0.5}$	Precision	Recall	$F_{0.5}$
NMT_{RNN}	34.25%	6.82%	18.99%	35.12%	6.89%	19.31%	35.76%	7.19%	19.93%
$CGEC_{CSS}$	39.96%	6.06%	18.85%	41.30%	6.16%	19.30%	41.77%	6.39%	19.82%
$CGEC_{CSS}$ +share	41.10%	6.14%	19.21%	42.20%	6.22%	19.56%	43.33%	6.55%	20.41%
$CGEC_{CSS}$ +policy	38.09%	7.10%	20.33%	38.56%	7.11%	20.46%	39.53%	7.46%	21.25%
$CGEC_{CSS_4}$	41.55%	4.99%	16.85%	42.60%	5.06%	17.14%	43.41%	5.28%	17.76%
$CGEC_{CSS_6}$	44.14%	6.19%	19.84%	44.94%	6.24%	20.07%	45.83%	6.52%	20.79%
$CGEC_{CSS_8}$	40.68%	6.25%	19.38%	42.14%	6.38%	19.87%	42.60%	6.61%	20.39%

$$F_{0.5} = 5 \times \frac{P \times R}{P + 4 \times R}, \tag{10}$$

$$a_i \cap s_i = \{s \in s_i | \exists a \in a_i (match(s, a))\}. \tag{11}$$

V. RESULTS

The results are shown in Table 3 and demonstrate that the models we proposed have good performance in all evaluation method.

A. CONVOLUTIONAL SEQUENCE TO SEQUENCE MODEL

Compared with the sequence to sequence model with RNN, convolutional based model has achieved better precision. The precision in annotator 1&2 increases over 6%. Since convolution can focus directly on the information around words, it may have better performance.

B. SHARED EMBEDDING

$CGEC_{CSS}$ and $CGEC_{CSS}$+share represent the model without shared embedding and with shared embedding respectively. Compared with results from $CGEC_{CSS}$, $CGEC_{CSS}$+share improves precision significantly. Especially for annotator 1&2, $CGEC_{CSS}$+share can improve 1.56%. For recall and $F_{0.5}$, the model also has a slightly improvement. Since the encoder and the decoder share embeddings, the meaning of the same word or character in encoder and decoder layer is same. During the training process, there will be no different meanings due to share embedding. Although there are many words or characters in Chinese that may have different meanings in different contexts, in the grammatical error correction, the source language and the target language are consistent and the meaning of the sentence expression is essentially the same. Therefore, share embedding model can find errors and correct more accurately.

C. POLICY GRADIENT

Compared with the results of $CGEC_{CSS}$ and $CGEC_{CSS}$+share, $CGEC_{CSS}$+policy achieves the highest recall and $F_{0.5}$ in all annotators. Compared with the basic convolutional sequence to sequence model, the values of recall and $F_{0.5}$ increase by 1.07% and 1.43% respectively in annotator 1&2. However model with policy gradient decreases the accuracy in all annotators. The change of the loss function makes the model tend to save the parameters which reach highest reward. We simply use GLEU as our reward and the reward is not applied to the characters but the sentence which may cause the decline of accuracy.

TABLE 4. Information of added corpus: NumOfSens represent sentences in added corpus after cleaning.

Corpus	CTB9	News' Commentary	News
NumOfSens	36591	99239	208598

D. THE RATIO OF ERROR SENTENCES TO CORRECT SENTENCES

After experiments, we found that some simple errors, such as character writing errors, were not corrected in the test set. So we want to explore the effect of the number of correct sentences in the corpus on the model. In the original training corpus, the ratio of error sentences to correct sentences is about 10:1. We get the added corpus from CTB9 and WMT[9] respectively. WMT consists of two parts, news' commentary and news. All corpus is cleaned with only Chinese. The detailed information is shown in Table 4. All trained models are based on shared embedding.

We can see from the results that the model trained on the corpus whose ratio of error to correct reaches 6:1 achieves best results. The model trained on the corpus with ratio 6:1 improves by 3.04%, 2.74%, and 2.5% for precision compared with the model only with shared embedding with annotator 1, 2 and 1&2 respectively. Also, $F_{0.5}$ in all annotator has a slightly improvement. Adding correct sentences can help the model find and correct errors more accurately. When fewer correct sentences are added, it will be slightly helpful for the model to correct sentences but not enough. When the correct sentences is added to a certain number, the model can have a significant improvement in the correction. When more correct sentences are added, fewer errors may be observed for the model, which is not conducive to training. In our model with shared embedding, the ratio of error sentences to correct sentences is 6:1, which would have a better effect.

[9]http://www.statmt.org/wmt18/

TABLE 5. Examples for case study: the words marked in blue refer to the word error, the word marked in red refer to the structure error.

Example	Source	Sentences
1	*Error*	那些空气污染也没有助于人生的身体建康。
		Those air pollution also do not help people's helth.
	NMT_{RNN}	那些空气污染也没有助于人生的身体健康。
	$CGEC_{CSS}$	那些空气污染也没有助于人生的身体建康。
	$CGEC_{CSS}$+share	那些空气污染也没有助于人生的身体健康。
	$CGEC_{CSS}$+policy	那些空气污染也没有助于人生的身体健康。
	$CGEC_{CSS_6}$	那些空气污染也没有助于人生的身体健康。
	Gold	那些空气污染也无助于人的身体健康。
		Those air pollution also do not help people's health.
2	*Error*	但是眼前的场景是不可容纳的，也不可允许的。
		But the scene at present is unacceptable and unacceptable.
	NMT_{RNN}	但是眼前的场景是不可容纳的，也不可允许。
	$CGEC_{CSS}$	但是眼前的场景是不可容纳的，也不可允许。
	$CGEC_{CSS}$+share	但是眼前的场景是不可容纳的，也不可允许的。
	$CGEC_{CSS}$+policy	但是眼前的场景是不可容纳的，不可允许的。
	$CGEC_{CSS_6}$	但是眼前的场景是不可容纳的，也不允许的。
	Gold	但是眼前的场景是不可容忍的、也不可允许的。
		But the scene at present is intolerable and unacceptable.
3	*Error*	它们的美逐渐使我所迷住。
		Their beauty gradually make me been fascinated.
	NMT_{RNN}	它们的美逐渐使我所迷住。
	$CGEC_{CSS}$	它们的美逐渐使我所迷住。
	$CGEC_{CSS}$+share	它们的美逐渐使我所迷住。
	$CGEC_{CSS}$+policy	它们的美逐渐使我所迷住。
	$CGEC_{CSS_6}$	它们的美逐渐使我迷住。
		Their beauty gradually make me fascinated by.
	Gold	它们的美逐渐把我迷住。
		Their beauty gradually fascinated me.
4	*Error*	中国是楼的规模比韩国的不一样。
		The size of China is buildings is different compare that of South Korea.
	NMT_{RNN}	中国的规模比韩国的不一样。
		The size of China is different compare that of South Korea.
	$CGEC_{CSS}$	中国是楼的规模和韩国的不一样。
		The size of China is buildings is different from that of South Korea.
	$CGEC_{CSS}$+share	中国的规模和韩国的不一样。
		The size of China is different from that of South Korea.
	$CGEC_{CSS}$+policy	中国是楼的规模和韩国的不一样。
	$CGEC_{CSS_6}$	中国的规模和韩国的不一样。
	Gold	中国的楼的规模和韩国的不一样。
		The size of China's buildings is different from that of South Korea.
5	*Error*	结果那一天不吃了午饭。
		As a result, we not have lunch on that day did.
	NMT_{RNN}	结果那一天不吃午饭。
		As a result, we do not have lunch on that day.
	$CGEC_{CSS}$	结果那一天不吃午饭。
	$CGEC_{CSS}$+share	结果那一天不吃午饭。
	$CGEC_{CSS}$+policy	结果那一天不吃午饭。
	$CGEC_{CSS_6}$	结果那一天不吃午饭了。
		As a result, we did not have lunch on that day.
	Gold	结果那一天不吃午饭了。
6	*Error*	在星期六，他早上去了散步中在路上遇到了一位邮递人。
		On Saturday, he met a postman when he went for a walk on the way in the morning.
	$CGEC_{CSS}$+share	在星期六，他早上去散步时，在路上遇到了一位邮递人。
		On Saturday, when he went for a walk in the morning, he met a postman on his way.
	$CGEC_{CSS}$+policy	在星期六，他早上去散步时在路上遇到了一位邮递人。
		On Saturday, when he went for a walk in the morning he met a postman on his way.
	Gold	在星期六早上，他在散步的途中遇到了一位邮递人。
		On Saturday morning, he met a postman on his way for a walk.
7	*Error*	无论你在哪里，我们可以联系。
	$CGEC_{CSS_6}$	无论你在哪里，我们都可以联系。
	Gold	无论你在哪里，我们可以联系。
		No matter where you are, we can contact you.

VI. CASE STUDY

In this section, we will analyze the corrected sentences by different models. Sentences are shown in Table 5. If the corrected sentences are the same, we just provide one translation.

A. THE EFFECT OF DIFFERENT ERRORS ON CORRECTIONS

In Chinese, words can be considered as the smallest semantic unit. The semantics of words are basically clear. Words consist of many characters and the most common case is a

word consists of two characters. Since a word are composed of fixed characters, it is easy to find errors in a word. Example 1 illustrates this situation. The error in the word is marked in blue. The word "健康" (health) is a common word in Chinese. This kind of word often appears in Chinese text and the probability of co-occurrence of the characters in the word is high. Therefore, it is easy for model to correct errors in common words.

It is a bit difficult to correct this situation if one character in the word is written incorrectly and the word becomes another commonly used word. Because this is not an internal error of a word, but related to the semantics of the whole sentence. In Example 2, the word marked blue "容纳" (unacceptable) and "容忍" (intolerable) are both common words in Chinese, but represent different meaning and they cannot be substituted for each other. In this case, the model needs to consider the expression of the whole sentence in order to make corrections.

A large part of the above two situations occur in nouns and adjectives, because many of these words are composed of more than one characters. Therefore, when a word differs from the original one, it is likely to be found and corrected. In Chinese, the difficulties of correcting verbs include two aspects, one is that most verbs contain only one character, and the other is how to choose the appropriate verbs according to the context. As in English, there are many verbs can be chosen for a noun regardless of context. In Example 3, both "使" and "把" are verbs. In Chinese, we can correct the sentence like "使我入迷" (fascinate me) or "把我迷住" (fascinate me). The two correct sentences express same meaning. Therefore, the verb correction of this kind of words which contain only one character is very difficult. The meaning of the sentence and some structures need to be considered.

The above-mentioned errors are limited to relatively small range. It is quite difficult to correct the overall structure of the sentence. In Example 5, errors in the order of multiple words lead to structural errors in the sentence. These errors sometimes cannot be corrected by the model. Corrections to structural errors in the sentences are not unique. Therefore, the sentences corrected by the model are inconsistent with those gold sentences. But for Chinese, these corrected sentences are feasible. Hence, it is inadequate to simply evaluate the corrected sentences given by the annotator for the sentence structure error. In Example 6, sentences corrected by model and given by gold express exactly the same meaning, and the grammar is correct. However, the ways of modification are different. It would be inappropriate to use gold statements as the simple criterion.

Although gold sentences are corrected by experts, sometimes there are cases where sentences corrected by model are superior to gold ones. As shown in Example 7, the modification of the model is more in line with Chinese language habits which represents the same meaning and translates the same as gold in English.

B. THE EFFECT OF DIFFERENT MODELS ON CORRECTIONS

Compared with $CGEC_{CSS}$ model, in the case of using the same corpus, model with shared embedding and with policy gradient can make word correction better. For instance in Example 1, $CGEC_{CSS}$ cannot correct the word "建康" (helth). However, models with shared embedding and with policy gradient correct it with same training corpus. For some word errors, although the models have not provided the correction, the methods of modification are given. It can be seen that the model detects errors, which is also an improvement of the performance, as Example 4. The performance of the model also improves in the correction of sentence structure, when the model is provided with a larger corpus, as Example 5. It can be seen that providing the right number of correct sentences can promote model learning of sentence structure.

VII. CONCLUSION

This paper introduces the convolutional sequence to sequence model into the Chinese grammatical error correction task for the first time and the models we proposed of two different optimization methods both achieve good performance. In Chinese grammatical error correction tasks, using shared embedding can improve the precision and using policy gradient to achieve greater reward can improve recall and $F_{0.5}$ significantly. When adding additional correct sentences to the training data, the number of correct sentences need to be paid attention to. In our model, when the ratio of error sentences to correct sentences is 6:1, the model have the best performance. Compared with the basic convolutional sequence to sequence model, the model $CGEC_{CSS_6}$ improves by 4.06% on precision and 0.97% on $F_{0.5}$ in annotator 1&2 and $CGEC_{CSS}$+policy improves by 1.07% and 1.43% on recall and $F_{0.5}$ in annotator 1&2 respectively. Besides, we collate the corpus of Chinese grammatical error correction and provide detailed information.

In the future, we will consider correcting the structural errors in the sentences and try to introduce structural information to make the model perform better.

REFERENCES

[1] Y. Xiang, B. Yuan, Y. Zhang, X. Wang, W. Zheng, and C. Wei, "A hybrid model for grammatical error correction," in *Proc. CoNLL*, Sofia, Bulgaria, 2013, pp. 115–122.

[2] Z. Yuan and T. Briscoe, "Grammatical error correction using neural machine translation," in *Proc. NAACL*, San Diego, CA, USA, 2016, pp. 380–386.

[3] T. Ge, F. Wei, and M. Zhou, "Fluency boost learning and inference for neural grammatical error correction," in *Proc. ACL*, Melbourne, VIC, Australia, 2018, pp. 1055–1065.

[4] Y. Zhao, N. Jiang, W. Sun, and X. Wan, "Overview of the NLPCC 2018 shared task: Grammatical error correction," in *Proc. NLPCC*, 2018, pp. 439–445.

[5] H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, "The CoNLL-2013 shared task on grammatical error correction," in *Proc. CoNLL*, Sofia, Bulgaria, 2013, pp. 1–12.

[6] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant, "The CoNLL-2014 shared task on grammatical error correction," in *Proc. CoNLL*, Baltimore, MD, USA, 2014, pp. 1–14.

[7] G. Rao, Q. Gong, B. Zhang, and E. Xun, "Overview of NLPTEA-2018 share task chinese grammatical error diagnosis," in *Proc. NLPTEA*, Melbourne, VIC, Australia, 2018, pp. 42–51.
[8] H.-H. Chen, Y.-H. Tseng, V. Ng, and X. Lu, "Proceedings of the 3rd workshop on natural language processing techniques for educational applications (NLPTEA2016)," in *Proc. NLPTEA*, Osaka, Japan, 2016.
[9] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. ICML*, 2017, pp. 1243–1252.
[10] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018. doi: 10.1109/TNNLS.2016.2616445.
[11] Z. Ma, J. Xie, Y. Lai, J. Taghia, J.-H. Xue, and J. Guo, "Insights into multiple/single lower bound approximation for extended variational inference in non-Gaussian structured data modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published. doi: 10.1109/TNNLS.2019.2899613.
[12] Z. Ma, Y. Lai, W. B. Kleijn, Y.-Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2019. doi: 10.1109/TNNLS.2018.2844399.
[13] K. Sakaguchi, M. Post, and B. Van Durme, "Grammatical error correction with neural reinforcement learning," in *Proc. IJCNLP*, Taipei, Taiwan, 2017, pp. 366–372.
[14] J. Ji, Q. Wang, K. Toutanova, Y. Gong, S. Truong, and J. Gao, "A nested attention neural hybrid model for grammatical error correction," in *Proc. ACL*, Vancouver, BC, Canada, 2017, pp. 753–762.
[15] Y.-H. Wu, J.-J. Chen, and J. Chang, "Verb replacer: An English verb error correction system," in *Proc. IJCNLP*, Tapei, Taiwan, 2017, pp. 49–52.
[16] Y.-C. Lo, J.-J. Chen, C. Yang, and J. Chang, "Cool English: A grammatical error correction system based on large learner corpora," in *Proc. COLING*, Santa Fe, NM, USA, 2018, pp. 82–85.
[17] S. Chollampatt and H. T. Ng, "A reassessment of reference-based grammatical error correction metrics," in *Proc. COLING*, Santa Fe, NM, USA, 2018, pp. 2730–2741.
[18] S. Chollampatt and H. T. Ng, "Neural quality estimation of grammatical error correction," in *Proc. EMNLP*, Brussels, Belgium, 2018, pp. 2528–2539.
[19] M. D. Zeiler and R. Fergus. (2013). "Stochastic pooling for regularization of deep convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1301.3557
[20] Z. Ma *et al.*, "Fine-grained vehicle classification with channel max pooling modified CNNs," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3224–3233, Apr. 2019, 10.1109/TVT.2019.2899972.
[21] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 194–197.
[22] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 121–128, Jan. 2019. doi: 10.1109/TVT.2018.2879361..
[23] R. Hu, Y. Tang, C. Li, and X. Wang, "Statistical analysis on large scale chinese short message corpus and automatic short message error correction," in *Proc. PACLIC*, 2008, pp. 397–403.
[24] L.-C. Yu, C.-H. Liu, and C.-H. Wu, "Candidate scoring using Web-based measure for Chinese spelling error correction," in *Proc. SIGHAN*, Nagoya, Japan, 2013, pp. 108–112.
[25] T.-H. Chang, H.-C. Chen, Y.-H. Tseng, and J.-L. Zheng, "Automatic detection and correction for chinese misspelled words using phonological and orthographic similarities," in *Proc. SIGHAN*, Nagoya, Japan, 2013, pp. 97–101.
[26] S.-M. Cheng, C.-H. Yu, and H.-H. Chen, "Chinese word ordering errors detection and correction for non-native Chinese language learners," in *Proc. COLING*, Dublin, Ireland, 2014, pp. 279–289.
[27] J. Xiong, Q. Zhao, J. Hou, Q. Wang, Y. Wang, and X. Cheng, "Extended HMM and ranking models for chinese spelling correction," in *Proc. SIGHAN*, Wuhan, China, 2014, pp. 133–138.
[28] S.-H. Chen, Y.-L. Tsai, and C.-J. Lin, "Generating and scoring correction candidates in Chinese grammatical error diagnosis," in *Proc. NLPTEA*, Osaka, Japan, 2016, pp. 131–139.
[29] Y.-T. Shiue, H.-H. Huang, and H.-H. Chen, "A Chinese writing correction system for learning Chinese as a foreign language," in *Proc. COLING*, Santa Fe, NM, USA, 2018, pp. 137–141.
[30] C. Li, J. Zhou, Z. Bao, H. Liu, G. Xu, and L. Li, "A hybrid system for Chinese grammatical error diagnosis and correction," in *Proc. NLPTEA*, Melbourne, VIC, Australia, 2018, pp. 60–69.
[31] K. Fu, J. Huang, and Y. Duan, "Youdao's winning solution to the NLPCC-2018 task 2 challenge: A neural machine translation approach to Chinese grammatical error correction," in *Proc. NLPCC*, 2018, pp. 341–350.
[32] R. Fu *et al.*, "Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement," in *Proc. NLPTEA*, Melbourne, VIC, Australia, 2018, pp. 52–59.
[33] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 933–941.
[34] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in optimizing recurrent networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8624–8628.

SI LI received the Ph.D. degree from Beijing University of Posts and Telecommunications, in 2012, where she is currently a Lecturer in School of Information and Communication Engineering. Her current research interests include natural language processing and machine learning.

JIANBO ZHAO received the bachelor's degree from Beijing University of Posts and Telecommunications, in 2017, where she is currently pursuing the master's degree in School of Information and Communication Engineering. Her current research interests include natural language processing and machine learning.

GUIRONG SHI is currently a Senior Engineer and the Director of the Power Dispatching Control Center, State Grid Jibei Electric Power Company Limited. His research interests include power system optimization and dispatching operation, source network load layered coordinated control, reliable grid connection, and efficient consumption.

YUANPENG TAN received the D.E. degree from North China Electric Power University, in 2017. He is currently an AI R&D Engineer with the Artificial Intelligence Application Department, China Electric Power Research Institute. His research interests include pattern discovery, object detection, natural language processing, and smart grid.

HUIFANG XU received the master's degree from Dalian University of Technology, in 2013. She is currently an AI R&D Engineer with the Artificial Intelligence Application Department, China Electric Power Research Institute. Her research interests include natural language processing, machine learning, and recommender systems.

GUANG CHEN received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, in 2006, where he is currently an Associate Professor in School of Information and Communication Engineering. His current research interests include information retrieval, text mining, and visualization.

HAIBO LAN is currently a Senior Engineer and the Director of the Power Dispatching Control Center, Dispatching Control Department, State Grid Jibei Electric Power Company Limited. His research interests include power grid operation and mode management.

ZHIQING LIN received the degree from Beijing University of Posts and Telecommunications, in 1982, where she is currently a Professor in School of Information and Communication Engineering. Her current research interest includes intelligent information processing and its application in networks.