IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 19, 2022, accepted March 11, 2022, date of publication March 14, 2022, date of current version April 5, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3159676

Pre-Training-Based Grammatical Error Correction Model for the Written Language of Chinese Hearing Impaired Students

BINBIN CHEN[1] **AND JINGYU ZHANG**[2]
[1]School of International Studies, Qiannan Normal University for Nationalities, Duyun, Guizhou 558000, China
[2]English Literature Institute, Xi'an International Studies University, Xi'an, Shaanxi 710128, China

Corresponding author: Jingyu Zhang (jingyuzhang888@hotmail.com)

This work was supported in part by the Humanities and Social Sciences Research of Chinese Ministry of Education under Grant 17XJA740001.

ABSTRACT Grammatical error correction has been considered as an application closely related to daily life and an important shared task in many prestigious competitions and workshops. The neural machine translation with an encoder-decoder architecture containing language models has been the fundamental solution for the grammatical error correction. Whereas Grammatical error correction task on texts of hearing impaired people or its solution has not been seen yet, and common Grammatical error correction tasks are suffering several challenges, such as insufficient training data, insufficient accuracy due to the unsatisfactory capacity of extracting semantic and grammatical patterns. Under these circumstances, we proposed a novel encoder-decoder architecture based on multi-head self-attention along with multiple strategies, which excels at extracting deep representations from the corrupted sentences of hearing impaired students and further reconstructing the sentences into grammatical ones. Via the re-ranking strategy, our model can correct various kinds of errors including spelling and complex syntax errors. The ablation experiments prove that the semantic extracting of self-attention mechanism excluding the position encoding with the word order shuffle operation can significantly learn the hearing impaired students' sentence patterns whose word order is quite different from the ones of hearing people and improve the correction scores. The pre-training can enhance the restoring efficiency of sentence structure in the decoding process. The comparison experiments with baseline models show that our model obtains superior performance either in the hearing impaired students' grammatical error correction or in a common grammatical error correction shared task.

INDEX TERMS Hearing impaired student, grammatical error correction, self-attention, encoder-decoder, pre-training.

I. INTRODUCTION

Grammatical Error Correction (GEC) is a task designed to automatically correct grammatical errors in text, wildly applied in many scenarios that directly interact with people, such as error correction for search query spelling correction, speech recognition, optical character recognition, etc. There have been considerable remarkable achievements in renowned GEC shared tasks in English, such as the CoNLL-2014 Shared Task, JFLEG, BEA-2019 Shared Task, and tasks in Chinese: shared tasks of NLPCC-2018, series workshops or shared tasks on NLPTEA and SIGHAN.

Despite sharing similar solutions, Chinese Grammatical Error Correction (CGEC) contains its unique characteristics due to the distinct language structure of Chinese. As an analytic language, according to linguistic typology, modern Chinese has few grammatical inflections, such as tense, voice, or number. In order to mark the tense and voice, there are heavy uses of grammatical particles in modern Standard Chinese, involving "le" (了, perfective aspect), "zhe" (着, durative stative), "guo" (过, experiential aspect), and so on. In addition, comprehension of Chinese sentences is highly dependent on syntax, namely, sentence order and structure, rather than morphology, which brings great burden for learners of Chinese as a foreign or second language (CFL/CSL), especially for western learners. It can be seen in Tab.1. that the tiny mistakes of particle usage would cause considerable semantic misunderstandings.

As for congenital Chinese hearing impaired students, the written language is the second language [1], while the first

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

TABLE 1. Erroneous samples.

	Error type	Source	Corrected
CFL/CSL	Perfective aspect	如果家庭破裂，将会让孩子在心灵上蒙上了阴影。 Ruguo jiating polie, jiang hui rang haizi zai xinling shang meng shang le yinxing. ("jiang hui" is the particle in the future tense, while "le" indicates the past tense. The tenses expressed by the two are contradictory)	如果家庭破裂，将会让孩子在心灵上蒙上阴影。 Ruguo jiating polie, jiang hui rang haizi zai xinling shang meng shang yinxing. (If the family breaks up, it will cast shadow in the heart of the child)
CFL/CSL	Durative stative	我们听音乐聊天。 Women ting yinyue liaotian. (There are two omissions of the particle indicating the durative stative "zhe".)	我们听着音乐，聊着天。 Women ting zhe yinyue, liao zhe tian. (We were chatting and listening to the music.)
Deaf learners	Order	离开了她一个人。 Likai le ta yigeren. (The subject "ta" and adverbial "yigeren" should be placed after the verb in declarative sentence order.)	她一个人离开了。 Ta yirenren likai le. (She left alone.)
Deaf learners	Omission	我说："不要哭，我给加油。" Wo shuo: "buyao ku, wo gei jiayou." (There are omission of object "ni" and the omission of future tense particle "hui".)	我说："你不要哭，我会给你加油的。" Wo shuo: "buyao ku, wo hui gei jiayou de." (I said, "Don't cry. I will cheer you on.")

language is a sign language possessing a completely different language system. That is to say, the text written by hearing impaired students suffers similar but more severe spelling and grammatical problems compared to CFL learners. It contains more frequent and serious problems, such as word order, syntax errors, sentence constituent omissions, related subject dropping, etc. (as shown in Tab.1)

GEC tasks frequently come along with Grammatical Error Diagnosis (GED) procedure, which is supposed to locate different types of grammatical errors. However, in this case, the error and its correction should lay in mostly the same position in the source sentence and the corrected sentence, otherwise the diagnosis result would not be satisfied. The error type of Chinese GED (CGED) mainly involves spelling errors, redundant words, missing words, lousy word selection and incorrect word order. Prior GEC and GED studies have obtained the outstanding achievements in this area. Most of them employed n-gram [2], [3], confusion set [4], [5], language model [6] including the BERT [7]–[9] etc. to diagnose the errors.

However, the solutions above-mentioned possess few capabilities of tackling problems such as inappropriate word order and sentence constituent omission.

Encoder-decoder [10], [11], a novel approach, has reached the new achievements in GEC [11], [12] as in machine translation. Notably, most of the recent advanced GEC architectures utilize the advantages of multiple models to solve different sort of problems [3], [5], [13], involving Transformer-based methods [14], [15].

The encoder-decoder architecture can solve many problems, including restoring sentences with incorrect word order to a certain extent, and only it carries two shortcomings. The prominent one is that the considerable demand for ungrammatical and grammatical paired data cannot be fed at the moment, which is crucial for encoder-decoder training. Especially for CGEC, there is substantial paucity of Chinese labeled corpus. Moreover, the encoder-decoder model cannot correct all multiple errors in one sentence in one time, even may revise the correct part into the incorrect one.

In this study, a novel solution involving a novel architecture and other strategies is introduced to deal with grammatical problems in the written language of hearing impaired students. We aim to extract the patterns from the sentences with severely corrupted structure and restore them into the grammatical ones. To the best of our knowledge, this is the first model that concerns about the grammatical error correction of the written language of hearing impaired people. Our contributions are summarized as follows:

1) We propose a backbone model involving a novel architecture with the significant ability to extract deep patterns from sentences with inappropriate word order and corrupted structure leveraging multi-head self-attention mechanism, and then correct the erroneous sentences.
2) The global attention is incorporated to soft-align the target characters to the deep representations of source characters processed by the multi-head self-attention.
3) The pre-training method based on large-scale corpus of hearing people is incorporated to assist the decoder to learn the patterns of grammatical sentences more effective given the situation lack of training data.
4) With the re-ranking strategy, our model can utilize the advantages of different models and their combinations to tackle various errors including spelling and complex syntax errors in sentences.

II. RELATED WORKS

Earlier studies performed well in spelling and grammatical error correction with relatively basic approaches involving n-gram [2], [3] confusion set [4], [5], statistical machine translation architecture [6], [16]–[18]. Generally, these previously mentioned methods are employed as compounds more frequently than individually.

As GEC is an integral part of Natural Language Processing, Grammatical Error Diagnosis (GED) has drawn a lot of attention in the meantime. It would be considered an individual task or the first procedure of GEC. GED is expected to diagnose types of grammatical errors which may be redundant words, missing words, lousy word selection, misplaced words, etc. A number of individual GED tasks are conducted as a classification problem [18]–[23], some of them are treated as a sequence tagging problem and the mission of feature engineering [24], [25], and some of them diagnose and correct grammatical errors via RNN architectures like LSTM [26] or GRU [27]. There is a trend of incorporation of Bert-based architectures [28]–[34] in the Chinese grammatical error diagnosis in the NLPTEA share task.

Various GEC studies modified and utilized the mask mechanism of BERT [7] to detect the errors and further correct them [8], [9]. Asano [35] incorporated the BERT to detect sentences with grammatical errors.

The performance of GEC has been pumped due to the introduction of Neural Machine Translation based on encoder-decoder structure [10], [11]. Recent GEC works based on the seq2seq framework have yielded impressive results [13], [36]–[39].

Despite the achievements prior works yielded, the simplex approach cannot tackle all kinds of grammatical or spelling errors in one go. Strategies with a compound of multiple solutions have been proven to be effective in GEC tasks.

Rozovskaya *et al.* combined machine learning classification and machine translation to improve the performance of grammatical error correction. The advantages of classification and translation frameworks differ in handling different types of error. The classifier is proved good at detecting and correcting the basic errors and therefore is taken as an initial part of the model, which applies the machine translation to fix the other complex errors afterwards [40].

Ge *et al.* proposed boost learning strategies based on the seq2seq framework to construct the error generation and error correction model, whose purposes are both generating error samples to build diverse error-corrected sentence pairs during training. A multiple round bidirectional error correction approach is conducted during the boost inference in order to improve sentence fluency in stages [12].

Zhao *et al.* introduced the copy mechanism in Grammatical Error Correction task based on Transformer framework, allowing the model to copy grammatically correct and out-of-vocabulary tokens directly from the source sequence. Furthermore, the denoising auto-encoder strategy is leveraged to pre-train the decoder model, improving the correction performance [15].

Another highly notable challenge lying in the GEC is the lack of training data. Despite sharing the encoder-decoder architecture with machine translation tasks, Grammatical Error Correction suffers from insufficient parallel training data, thus several solutions were incorporated in previous studies. Lichtarge *et al.* took the versions of texts before and after editing from Wikipedia edit histories with minimal filtration heuristics as training sentence pairs. The other extending data method is back-translating the Wikipedia sentence through another language in order to generate the error combining sentences. All data are transported to pre-train the Transformer model and further conduct the correction task [14]. Zhao *et al.*, inspired by the BERT, enlarge the training corpus by randomly generating errors [15]. Ge *et al.* obtained plenty of extended parallel data from the boost process [12]. Kiyono *et al.* demonstrated that the back-translation approach is effective to enlarge the training data to improve the correction performance, compared with directly introducing the noise to build the training data [41]. Zhao *et al.* proposed a method to improve the grammar error correction model based on neural machine translation through dynamic masking, which solves the model's need for a corpus of ''error-correct'' sentence pairs [42].

III. METHODOLOGY

Prior studies on GEC focus on slightly grammatically erroneous sentences where most other parts of sentence are correct [15]. In contrast, the sentences of hearing impaired students severely suffer from syntax erros. Most sentences lack grammatical and semantic continuity and are more like slices of meaning with order enormously different from the text of hearing people (see Tab.1), which was described as ''telegraphic utterances'' [43]. The trickiest problem is to catch patterns from the sentences of hearing impaired students with chaotic orders.

As proved in many studies, treating GEC as translation tasks with encoder-decoder frameworks like seq2seq and Transformer achieves outstanding performance. However, considering the complex situation of the text of hearing impaired students, either the vanilla seq2seq architecture or Transformer may not be the ideal option. The RNN structure in the encoder is prone to learn the chaotic sequential patterns and transport them to the decoder, thereby causing erroneous outputs.

In this study, we build an architecture (as shown in Fig.1) consisting of an encoder (sec.3.2.1) based on multi-head self-attention mechanism and a decoder (sec.3.4) with GRU RNNs so as to extract the chaotic patterns from the sentences of hearing impaired students and reconstruct them into grammatical ones. The soft attention (sec.3.3) incorporates the representations of encoder into decoder. In order to let the model further learn the out-of-order information and augment the training data, we introduce the shuffle strategy (sec.3.2.2) in the encoding stage. A pre-training strategy (sec.3.5) is incorporated to assist the decoder to learn the patterns of grammatical sentences of hearing people, given the situation lack of training data. All above constitutes our backbone model. Meanwhile, a N-gram language model (sec.3.1) is involved with the backbone model in our solution to remove the spelling errors. A re-ranking strategy (sec.3.6) with the above models and their different combinations is implemented for adapting grammatical errors of varying degrees. In the whole procedure, characters rather than words are

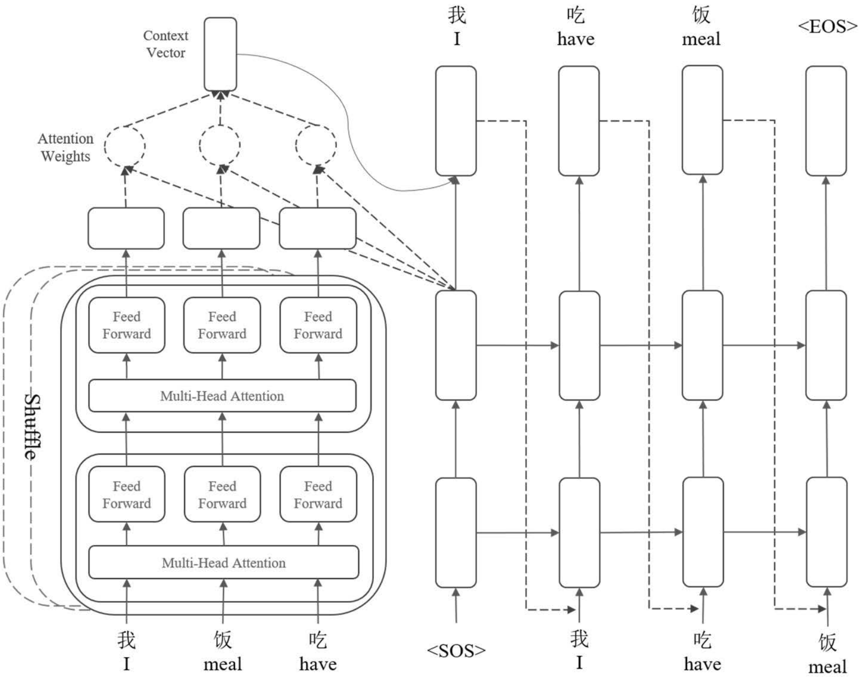

FIGURE 1. The architecture of backbone model.

considered as the input tokens in order to reduce the OOV problem.

The task is described as follows. Given an erroneous sentence $X = (x_1, x_2, \ldots, x_n)$ and its label sentence $L = (l_1, l_2, \ldots, l_m)$, our solution aims to learn the probabilistic mapping from error-corrected sentence pairs to generate output sentences $Y_1, Y_2, \ldots, Y_n$, $[Y = (y_1, y_2, \ldots, y_m)]$. A re-rank strategy is then implemented so that the output one with the lowest score of perplexity (as shown in Eq.1) will be selected.

$$PP(Y) = P(y_1 y_2 \ldots y_N)^{-\frac{1}{N}} \tag{1}$$

A. SPELLING ERROR CORRECTION

In the corpus of hearing impaired students, not every sentence suffers from the sequential order or structure problem, especially that of senior students. Similar to the text of the CFL students, there are a considerable amount of spelling errors of characters, such as erroneous characters with similar shape or pronunciation. For addressing this issue, a tri-gram language model is implemented.

We define the original sentence with errors $X = (x_1, x_2, \ldots, x_n)$, and try to replace every character existing in the confusion sets of SIGHAN-7 CSC Datasets [44]. The character string with the maximum score evaluated by the maximum likelihood estimation is selected as the correct one. The conditional probability of the character string s is calculated as in Eq.2 and then estimated as in Eq.3.

$$P(X) = \prod_{n=3}^{N} P(x_n \mid x_{n-2}, x_{n-1}) \tag{2}$$

$$P(x_n \mid x_{n-2}, x_{n-1}) = \frac{C(x_{n-2}, x_{n-1}, x_1)}{C(x_{n-2}, x_{n-1})} \tag{3}$$

B. FEATURE EXTRACTION FROM WRITTEN LANGUAGE OF HEARING IMPAIRED STUDENTS

1) ENCODER BASED ON MULTI-HEAD SELF-ATTENTION MECHANISM

In the contrary to RNN and CNN networks, Transformer has been proven to comprise the more outstanding semantic feature extraction ability by virtue of self-attention mechanism in natural language processing tasks [45]. The vanilla Transformer constructure [46] consists of encoder and decoder with stacks of identical layers whose cores consist of multi-head self-attention mechanism. Both the encoder and decoder embed the sequential inputs combined with the corresponding positional encoding so as to learn the relative order of the sequence.

In this section, we aim to encode the representations of the sets of individual Chinese character meanings, excluding the sequential order in the model. Self-attention allows the model to attend to the interaction between the target and its entire sequence, thereby extracting the feature of the target. Multiple attention heads are capable of squeezing deep representations of words. We hence build a character-level encoder consisting of a stack of six duplicate layers with two sub-layers: multi-head self-attention mechanism and a fully connected feed-forward network.

The self-attention is illustrated as the relation of the mapping between a query and key-value pair from each corresponding token. The scaled dot products by $1/\sqrt{d_k}$ between a query and all keys is the aligned weights assigned to the corresponding values. The self-attention is the sum of weighted values. The queries, keys, and values are packed into sets of matrices Q, K, V, and computed as in Eq.4 as follows:

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{4}$$

In order to capture more representation, the multi-head attention MA generated via weighting self-attention is introduced as representation of the written language of hearing impaired students, which is defined in Eq.5.

$$MA=Concat\left(head^{1},\ldots,head^{h}\right)w^{O} \tag{5}$$

where $head^{i}=Attention\left(QW_h^{Q},KW_h^{K},VW_h^{V}\right)$, $W_h^{Q}\in\mathbb{R}^{d_{model}\times d_k}$, $W_h^{K}\in\mathbb{R}^{d_{model}\times d_k}$, $W_h^{V}\in\mathbb{R}^{d_{model}\times d_v}$ and $W^{O}\in\mathbb{R}^{\mathrm{h}d_v\times d_{model}}$. In this work, we employ h = 8 parallel attention layers or heads. For each of these, we use $d_k=d_v=d_{model}/h$.

A fully connected feed-forward network is applied identically to each *MA*, which contains two linear transformations with a ReLU activation, according to the formula below:

$$FFN\left(x\right)=\max\left(0,xW_1+b_1\right)W_2+b_2 \tag{6}$$

Upon each sub-layer, a residual connection followed by layer normalization is implemented. The output of each sublayer is described as in Eq.7.

$$output=LayerNorm\left(x+Sublayer\left(x\right)\right) \tag{7}$$

By contrary with the vanilla transformer, given extracted character representations without sequential order, the position encoding is not included in our model.

2) SHUFFLE - DEEPER EXTRACTION VIA ORDER REVISION

Intending to extract deeper patterns of the sequences with chaotic order of hearing impaired students, inspired by the essence of denoising auto-decoders [47], we generate an enlarged training data containing order-corrupted sentences. Given the good result of denoising method in GEC performance [15] we randomly revised the sequential order to compose an erroneous sentence with its label into training pairs so as that the model can learn more character representations under vast sequential order combinations.

A feature observed is that the sentential order of hearing impaired students is not completely chaotic. The chunking pattern based on the word segment can still be analyzed. In this case, character-level shuffling may not benefit from feature extraction. We implement the shuffle before the word segment in this study.

C. THE SOFT ATTENTION

The final output, that is, the output of the encoder, of six encoder layers is MA'_n. Inspired by the attentive seq2seq architecture [48], we calculate the align score between the current hidden state h^t in the decoder with each deep representation of input character MA'_n in the encoder (see eq.8). The context vector c_t is the weighted average over all the MA'_n, and the weight are the corresponding align score a_t. The context vector participates in the next calculation in the decoder.

$$\begin{aligned} a_t &= align\left(h_t, MA'_n\right)\\ &= \frac{exp\left(score\left(h_t, MA'_n\right)\right)}{\Sigma_{n'}exp\left(score\left(h_t, MA'_{n'}\right)\right)} \end{aligned} \tag{8}$$

$$c_t = \sum\nolimits_i^t a_i h_i \Big/ t \tag{9}$$

D. SEQUENTIAL ORDER DECONSTRUCTION

The decoder of our model is expected to capture the long-term dependencies of the texts of hearing people and further to restore the sequential order under the help of the context vector from the encoder. Considering the significant advantage of RNN architecture in learning long-term dependency [49], we introduce the RNNs with the GRU cells in our model as the decoder.

The decoder of our model is two layers of RNN with GRU cells which are trained to generate the output sequence by predicting the conditional distribution of the next symbol y_t given the hidden state h^t, y_{t-1} and the sequence representation c from encoder while the h^t is computed leveraging the similar procedure as fellows,

$$h^t = f\left(h^{t-1}, y_{t-1}, c_t\right) \tag{10}$$

$$P\left(y_t \mid y_{t-1}, y_{t-2}, \ldots, y_1, c_t\right) = g\left(h_{\langle t\rangle}, y_{t-1}, c_t\right) \tag{11}$$

where f and g are activation functions.

In a GRU cell, in order to predict the *i-th* character in time step t, the update gate z_i decides how much of the past information h_{t-1} needs to maintain and be passed along to the future. z_i is the sum of weighted $y_{t-1}, h^{t-1}, and\ c_t$ with the sigmoid activation function.

$$z_i = \sigma\left(\left[O_z y_{t-1}\right]_i + \left[S_z h^{t-1}\right]_i + \left[U_z c_t\right]_i\right) \tag{12}$$

The reset gate r_i helps the model to determine how much the past information to forget.

$$r_i = \sigma\left(\left[O_r y_{t-1}\right]_i + \left[S_r h^{t-1}\right]_i + \left[U_r c_t\right]_i\right) \tag{13}$$

The element-wise product between the reset gate r_i and Sh^{t-1} determines what to remove from the previous and the representation c_t.

$$\tilde{h}_i^t = \tanh\left(\left[O y_{t-1}\right]_i + r_i \odot \left[Sh^{t-1} + Uc_t\right]\right) \tag{14}$$

The memory h^t of t time step is combined from the current memory $\tilde{h}_i^t$ and previous memory h^{t-1}. The update gate is implemented to choose information from $\tilde{h}_i^t$ and h^{t-1}.

$$h_i^t = z_i \odot h_i^{t-1} + \left(1 - z_i\right) \odot \tilde{h}_i^t \tag{15}$$

At the inference time, the beam-search strategy with beam width 3 is incorporated to re-rank the candidates.

E. PRE-TRAINING

Due to the insufficient training data, the unsupervised learning strategy is introduced to help the model to learn more sequence features. We pre-train the GRU RNN in a decoder on the Douban book comments corpus[1] without labels, considering that the comment corpus rather than news corpus contains oral characteristics closer to the language of starter learners like hearing impaired students. After pre-training, the correction model starts to train with the decoder initialized with the pre-trained parameters, and the parameters of other parts of the model initialized randomly.

F. RE-RANKING

Our corpus consists of texts of hearing impaired students from elementary school to high school. There are complex types of grammatical errors in the erroneous sentences. However, students in different grades have different degrees and types of errors in their sentences. Using a complex neural network model to deal with sentences that contain only a small number of simple spelling errors may yield poor results, for example, by changing the correct sentence to the wrong one, or vice versa. Therefore, a flexible approach using appropriate models to tackle varying degrees of error should be proposed. Inspired by Fu *et al.* [50], we configure models into following pipeline containing four solutions, forming our ensemble model:

S1: Spelling correction
S2: Backbone model
S3: Spelling correction + Backbone model
S4: Backbone model + Spelling correction

S1 denotes the N-gram language model for correcting the spelling errors in section 3.1. S2 includes the whole proposed neural network architecture along with the strategies. S3 and S4 incorporate the same components, however, with different order. Because in some out-of-order strings, first applying the N-gram spelling correction may mislead the model and cause the inappropriate correction.

The pipeline processes each input sentence and produces its corrected four candidates. The one with the lowest perplexity score (Eq.1) will be the output sequence (see Fig.2). According to such a configuration, sentences of higher-grade hearing impaired students that are more in line with grammatical rules and contain a small number of spelling errors will be allocated to S1 for processing, instead of being changed into worse sentences by other solutions. The sentences of lower-grade students with serious grammatical errors will be assigned to s2-s4, and the best result will be selected as final output.

IV. EXPERIMENTS AND DISCUSSION

A. DATASETS

We collected the raw texts of hearing impaired students containing compositions and diaries from schools for the hearing impaired and mute, ranging from elementary school to high school. All raw texts are corrected by students and teachers of linguistics. After the segment, the raw texts are shuffled three times, then compose sentence pairs with corresponding labels. 30% of paired sentences are taken as validation data. At the pre-train step, the Douban comment corpus is utilized. All corpora are listed in Tab.2.

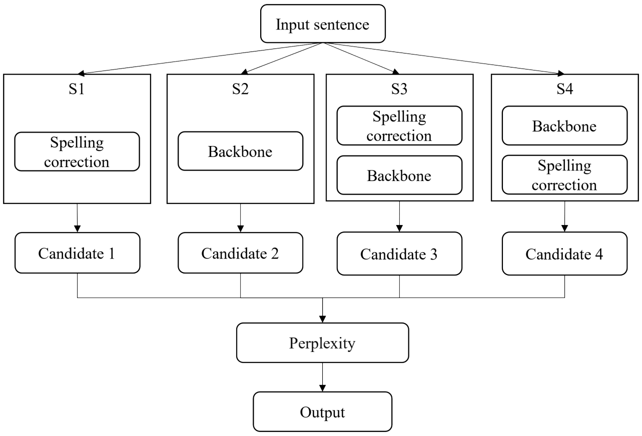

FIGURE 2. Re-ranking processing flow.

B. EVALUATION CRITERIA

In the following experiments, we use the precision, recall, and F0.5 measure to evaluate the performances of the models in NLPCC 2018 shared task, and the GLEU for the other experiments.

As for comparison on the shared GEC task of NLPCC 2018, the MaxMatch(M^2) Scorer [42], [51] is introduced to evaluate the performance of competitors. It allows the phrases from the prediction sentence with the maximal overlap with the gold standard to be selected to form the set of prediction edits $\{e_1, \ldots e_n\}$. The precision, recall, and F0.5 measure between the set of prediction edits and the set of gold edits $\{g_1, \ldots, g_n\}$ for all sentences are computed as eq.16-18.

$$P = \frac{\sum_{i=1}^{n} |e_i \cap g_i|}{\sum_{i=1}^{n} |e_i|} \tag{16}$$

$$R = \frac{\sum_{i=1}^{n} |e_i \cap g_i|}{\sum_{i=1}^{n} |g_i|} \tag{17}$$

$$F_{0.5} = \frac{\left(1 + 0.5^2\right) P \times R}{0.5^2 P + R} \tag{18}$$

Except MaxMatch Scorer, the commonly utilized evaluation indicators for the performance evaluation of text error correction related models are as follows: I-measure, and GLEU. I-measure [52] calculates the model performance score I by comparing the modified text of the model with the source text (the uncorrected original text). GLEU is a simple variant of BLEU [53], which shows better correlation with human judgment on the CoNLL-2014 shared task test set. It is more suitable for evaluating the fluency of corrected text [12], [54].

As for hearing impaired texts, the priority to the pragmatics and accessibility of meaning is expected. Relative to the accuracy of the position of words, the evaluation of hearing

[1] The comment data comes from the site: https://book.douban.com/

TABLE 2. Corpora.

Corpus	Sent.	Token	Type	Application
Corpus of the deaf	313,742	14,305,770	Labeled	Training and validation
Shuffled	941,226	42,917,310	Labeled	Training and validation
NLPCC-2018	2,440,140	54,075,716	Labeled	Training and validation
Douban comments	58,025,759	1,090,115,266	Unlabeled	Pre-train

TABLE 3. Comparison on the corpus of hearing impaired students.

Model	GLEU
Attentive Seq2seq	0.4527
Transformer	0.4312
Our ensemble model	**0.5897**

impaired text pays more attention to the fluency of sentences and the accessibility of meaning. Therefore, GLEU metric is leveraged to evaluate the effect of error correction on hearing impaired texts. To this end, 2 or more manually revised sentences of 1 original sentence are given as its labels.

The calculation equations of GLEU are as follows:

$$Count_B\left(n-gram\right)=\sum\nolimits_{n-gram'\in B} d\left(n-gram, n-gram'\right) \tag{19}$$

$$d\left(n-gram, n-gram'\right)=\begin{cases} 1 & if\ n-gram=n-gram' \\ 0 & otherwise \end{cases} \tag{20}$$

$$BP=\begin{cases} 1 & if\ c>r \\ e^{\left(1-{}^{c}\!/_{r}\right)} & if\ c\leq r \end{cases} \tag{21}$$

$$GLEU\left(C,R,S\right)=BP\cdot exp\left(\sum\nolimits_{n=1}^{N} w_n \log p'_n\right) \tag{22}$$

where, $N=4$, $w_n=\frac{1}{n}$, these two parameters are the standard parameters. S is the source text, R is the standard text, and C is the output text.

C. COMPARISON WITH THE BASELINE MODELS

1) COMPARISON ON THE CORPUS OF HEARING IMPAIRED STUDENTS

In this section of comparison experiments, for the backbone model, the encoder consists of token embedding and a hidden size of 768 dimensions, six layers and eight attention heads, and the position-wise feed-forward network of 2048 dimensions; the decoder incorporates two layers of RNNs with 128 GRU cells with hidden state size of 768 dimensions; The following parameters are set to achieve the best accuracy: epoch = 10, batch size = 512, learning rate = 0.002. We implement the dropout with the rate 0.05 to prevent overfitting, Sparse SoftMax cross entropy with logits as loss function, Adaptive moment estimation algorithm to update the parameters adaptively. A 7,000 character-level vocabulary for the input and output tokens is built from the datasets.

Given that the hearing impaired written language substantially differs from any text from CGEC shared tasks, we take the highly acclaimed vanilla Transformer [46] and attentive seq2seq [48] (seq2seq with global attention) as baseline models instead of modified ones for specific tasks in CGEC contest. The parameters are set the same as our model. The comparison results are shown as in Tab.3.

The results show that the seq2seq obtains the inherent structural advantage of RNN for dealing with sequence problems. Therefore, compared to the transformer, the attentive seq2seq is better at acquiring the sequence features of the source sentence and using it to recover the error sentence on the decoder side and has an advantage of 0.0215 points in the result. It is worth noting that our model combines both the ability of the transformer to extract the deep character representation and the ability of seq2seq to obtain long-term dependence in the sequence. Its performance has been proven by the comparison experiments. Our model has obtained the most superior result.

In the machine translation task, when the word order of the source sentence and the target sentence are the same, the strong attention will be presented along the diagonal of the matrix. This kind of tasks can be handled relatively easily by the translation models. When the word order of the source sentence and the target sentence are inconsistent, the strong attentions will not necessarily show a diagonal distribution. However, the results of basic neural translation models tend to be distributed diagonally, and it is difficult to obtain the associated patterns of source sentences and target sentences with inconsistent word order. Especially when the source sentence and the target sentence are in the same language, and the units of two sentences differ extremely little, normal neural translation models are not as smooth as they used to be in the process of this kind of conversion. This is why a separate seq2seq with attention is not accounted for advantage in GEC task of hearing impaired students.

In A1 and A2 of Fig.3, in Chinese, a location noun should be grounded after “qu去 (go to)” in the source sentence, the mission of GEC model is expected to attach the location noun “da ting大厅 (hall)” after “qu”. Our model (A2) skips “le了 (particle)” and “ting听 (listen to)” to correctly associate “da ting” in the source sentence and in the target sentence, and places it after “qu”. By the contrary, seq2seq failed to convert the wrong collocation “qu (go to) le (particle) ting (listen to)” to “qu da ting (go to the hall)”. Although it can be seen from A1, the model correctly mapped almost every character of the source sentence and target sentence, it failed to capture the syntactic relationship, that is, failed to correct the wrong word order. In the sentence of B1 and B2, there are more complex

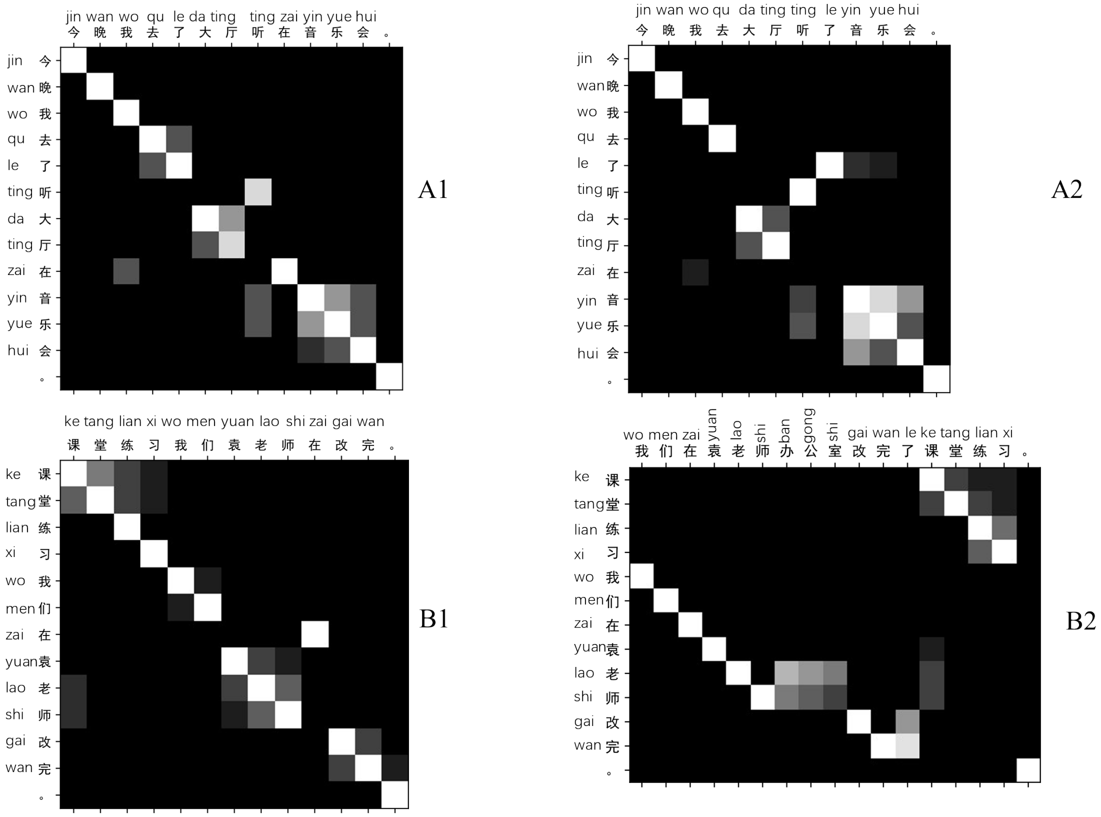

FIGURE 3. Two sample alignments based on global attention (comparison between the attentive seq2seq and our backbone model). The x-axis and y-axis separately correspond to the characters of the corrected sentence and the raw one. Two sentences are not necessarily of equal length. Each grid denotes the attentive weight between the raw character and the corrected character. The weight is presented in the gray scale, the closer to white, the greater the weight, and vice versa. We randomly selected two sentences without "unknown" to be presented in the plots. A1 and B1 are the sentences corrected by the attentive seq2seq, and A2 and B2 are corrected by our backbone model. Besides the important role of the encoder-decoder structure, soft (global) attention also provides a visual illustration of interpretability. We can intuitively see the relevance between the source and target characters after processing by a model. Fig. 3 respectively depicts the comparison results between the seq2seq model and our backbone model. Each intersecting grid in the matrix represents the degree of association (attention weight) between a source character and a character in the corrected sentence. Through this, it is easy to observe which source character is considered more important by the model while generating the corrected character.

TABLE 4. Comparison results in NLPCC2018 shared task.

Model	P	R	F0.5
YouDao	35.24	18.64	29.91
AliGM	41.00	13.75	29.36
BLCU	41.73	12.08	29.02
Our ensemble model	**42.85**	**16.35**	**32.36**

syntax and word order errors, and the length is longer, so it can be seen from the plot that the correction result of seq2seq is less ideal.

2) COMPARISON IN CGEC TASK

Considering the outstanding performance of the baseline models, we try to further test the generalization performance of the model in this section. The most renown and recent Chinese grammatical error correction contest is the shared task in NLPCC 2018 [42], [51]. Thus, we take the data and the test criteria of this task to verify the performance of our model and compare it with the results of the winning models. The results are listed as follows:

The corpus used in this task is from learners of Chinese as a foreign or second language (CFL/CSL). As mentioned in the introduction, the syntactic structure errors of CFL/CSL written language are much more minor than those of hearing impaired people. Our model can solve the grammatical problem of the written language of the hearing impaired to a large extent, thus it can solve the problems of the written language of the second foreign learner better than other competitors relatively easily.

All top three competitors incorporate the encoder-decoder architecture in their backbone models and treat the CGEC as a neural machine translation task. YouDao [50] incorporated the vanilla transformer as the main neural translation model. AliGM [55] took the statistical language and machine translation models and the seq2seq with attention architecture as the neural machine translation to solve the grammatical problems. As for BLCU [56], the convolutional sequence-to-sequence model was introduced as the main model. In conclusion, the competitors adopted combination solutions of a relatively simple neural network model combined with other data processing methods to correct grammatical errors.

TABLE 5. Results of ablation of re-ranking.

Model	GLEU
S1: N-gram	0.2267
S2: Backbone	0.5396
S3: N-gram + Backbone	0.5082
S4: Backbone + N-gram	0.5613

TABLE 6. Results of ablation of pre-training.

Model	GLEU
Backbone − pre-training	0.4887
Backbone	0.5396

In contrary, we proposed a completely novel architecture which excels at correcting the grammatical errors in written language either of hearing impaired students or of hearing people, which the listed results in Tab.4 have proved.

Note that the scores are different from the comparison on the corpus of hearing impaired students, because the Max-Match is adopted in this section.

D. ABLATION EXPERIMENTS

1) ABLATION RESULTS OF RE-RANKING

Aiming to evaluate the effects of each procedure and component of the re-ranking strategy in the proposed model, we conduct the ablation experiments on the corpus of hearing impaired students, and the results are shown in Tab.5.

It can be observed that the separate N-gram solution gains the lowest score (0.3346 points lower than S4), which proves that the grammatical problems of hearing impaired students are way too complex for N-gram to solve. It is worth noting that the score of S3 is lower than S4, even lower than S2 - the separate backbone model. Instead of improvement, the operation of N-gram reduces the performance of the model. That is to say, first applying the N-gram may incorrectly revise sentences or mislead the backbone model to incorrectly revise sentences.

2) ABLATION RESULTS OF PRE-TRAINING

In this section, we try to verify whether the pre-training solves the problems including slow convergence and difficulty in improving accuracy due to the lack of training data. The experiments are conducted on the corpus of hearing impaired students.

Due to not sufficient training data and the bad syntactic structure of the written language of hearing impaired students, the grammatical error correction performance of our model is not completely satisfactory. Whereas, with the help of extracted sequential patterns from the pre-training procedure, the result of grammatical error correction proposed model reaches an outstanding level. The performance gap with or without pre-training reaches 0.0509 points (see Tab.6).

TABLE 7. Results of ablation of shuffle.

Model	GLEU
Attentive Seq2seq	0.4527
Attentive Seq2seq + shuffle	0.3096
Transformer	0.4312
Transformer + shuffle	0.3692
Backbone - shuffle	0.4821
Backbone	0.5396

TABLE 8. Results of ablation of position embedding.

Model	GLEU
Transformer	0.4312
Transformer - position embedding	0.4278
Transformer + shuffle	0.3692
Transformer + shuffle - position embedding	0.4397
Backbone + position embedding	0.5002
Backbone	0.5396

TABLE 9. Results of ablation of decoder.

Model	GLEU
Transformer	0.4312
Backbone - pre-training - shuffle + position embedding	0.4416
Backbone - shuffle + position embedding	0.4624
Backbone	0.5396

There is another achievement that the pre-training procedure has made, that is, the convergence speed of the proposed model has significantly reduced with the pre-training participating. The convergence process can be observed in Fig.4.

3) ABLATION RESULTS OF SHUFFLE

In order to testify the effect of shuffle, we set the shuffled enlarged hearing impaired corpus as training data of vanilla Transformer (Transformer + shuffle) and seq2seq (Attentive Seq2seq + shuffle) with global attention, meanwhile the raw corpus as the training data of the proposed backbone model (Backbone − shuffle).

By comparing the results in Tab.7, it can be seen that the shuffle has reduced the performance of Attentive Seq2seq by 0.1431 points. Although the shuffle also has weakened the performance of Transformer (0.0620 points lower), the degree of weakening is less obvious than seq2seq, which means Transformer is less sensitive to information of sequential position. The poorer result (0.1076 lower) of our backbone model training without shuffle (Backbone−shuffle) compared with performance of the backbone model denotes that the backbone model benefits from the shuffle operation.

4) ABLATION RESULTS OF POSITION EMBEDDING

From the ablation experiments of effectiveness of shuffle operation, it can be observed that our backbone model gains superior performance against Transformer and Attentive

TABLE 10. Case study.

	Samples	Treatment
1a.	我们怎么客气的和人打招乎老师说了。 We how polite greet people teacher say (perfective aspect) (ungrammatical subordination).	Raw
1b.	我们怎么客气地和人打招呼，老师说。 How shall we greet people politely, said the teacher.	S1
1c.	老师告诉我们怎么和客气的人打招呼。 Our teacher said us how to greet polite people.	S2
1d	老师告诉我们应该怎么和客气的人打招呼。 Our teacher told us how to greet polite people.	S3
1e	老师说我们该怎么客气地和人打招呼。 Our teacher told us how to greet people politely.	S4
2a.	我真心希望他过幸福的生话。 I truly hope he will have happy live speech. (with spelling error).	Raw
2b.	我真心希望他过得幸福的生活。 I truly hope he had (inappropriate tense) happily live.	S4
2c.	我真心希望他过幸福的生活。 I truly hope he will have happy live.	S3

seq2seq from shuffling the input tokens. In this section, experiments are conducted to further prove the better results above our model obtained are due to the absence of position embedding.

The results of Tab.8 shows that Transformer performance decreased slightly (0.0034) after position embedding removal. It proved that position embedding should be helpful for the extraction of sequence features, but its effect was limited. However, for the text after shuffle, i.e., the out-of-order, augmented text, compared to the Transformer with position embedding, Transformer without position embedding has a significant gain (0.0705), and even has a certain improvement (0.0085) compared with the vanilla Transformer without shuffle. On the other hand, even with the help of pre-training and shuffle strategy, the correction performance of backbone model still drops (0.0394) after the addition of position embedding.

5) ABLATION RESULTS OF DECODER

In this section, we design ablation experiments to verify the effectiveness of modification of decoder of Transformer. Compared with vanilla Transformer, the main differences in our backbone model structure, except encoder, lie in the addition of shuffle and pre-training strategies, and the absence of position embedding. Therefore, in order to verify the effectiveness of encoder of Transformer and our backbone model, we removed the above components and carried out comparative experiments.

From the comparison results of the first and second rows in Tab.9, it can be seen that the correction result has been improved by 0.0104 after modification of decoder of transformer into our GRU-based structure. However, the main purpose of our decoder modification is to extract the sequence features of the grammatical sentences from the hearing people. The GRU neural networks are the base of the pre-training strategy. After adding the pre-training results based on the texts of hearing people, the proposed backbone model has significantly improved by 0.0208 compared with backbone model excludes the pre-training (see the data in the third line of Tab.9). It demonstrated that the GRU-based decoder modification is effective in our task compared with the vanilla Transformer's encoder.

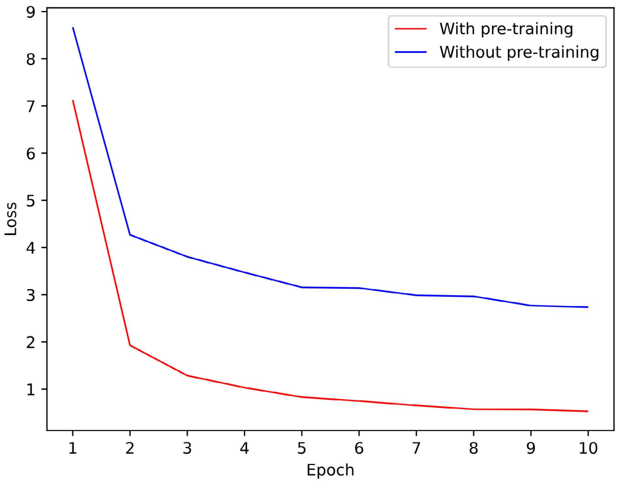

FIGURE 4. The convergence rate with and without pre-training.

E. DISCUSSION

From the comparison, results note that in this task the score of seq2seq is higher than Transformer, which proves that the

RNN based model takes advantage in capturing long-term dependency. Meanwhile, it is worth noting that seq2seq has taken way more time to converge than the Transformer. It is important to consider balancing the performance and execution time in the future works, which means it is necessary to take advantage of different structures of models to design a novel architecture.

The performance of Attentive seq2seq with shuffle is poorer, which reveal that the RNN architecture learns corrupted order patterns and then transports to generate the ungrammatical sentence. The similar result can be observed in the comparison between Transformer with and without shuffle. It is possible that the corrupted sequential patterns are learned due to the RNN structure in seq2seq and the position embedding in Transformers, which is not conducive to the generation of grammatical sentences. In contrast, the proposed backbone model obtains remarkable results in experiments, which verifies the benefit of the position-encoding-excluded procedure in the encoder.

The sentence with complex grammatical errors is processed by the re-ranking procedure, then to output sentences with errors (1b, 1c, and 1d of Tab.10) at different extent except one obtaining the grammatically satisfactory result (1e of Tab.10). The output samples in the case study (Tab.10) basically accord with the ablation experiment results in Tab.5, that is, the S4 procedure superior to the others. However, the best solution is not always S4, it depends on the degree of grammatical error in the sentence.

Despite the better score of S4 in most sentences of hearing impaired students, we have found many cases corrected by S3 are more in line with grammatical rules as in 1c of Tab.10. After analysis, it is found that the sentences that get higher scores through S3 processing come from higher grade students, mostly senior students. These texts suffer slightly from sentence structure but generally spelling errors like those by hearing people. Thus, under these circumstances S3 benefits from the N-gram first conducts before the neural machine translation architecture, our backbone model. Despite the fact that the results of the ablation experiment of re-ranking demonstrated that the neural network architecture obtains more advantages and there seems to be no reason to put the N-gram model in the first place (the S3 solution), the case study shows that due to the complexity of the grammatical error correction task of hearing impaired students, the S3 solution has played an irreplaceable role.

V. CONCLUSION

In this study, we take advantage of a self-attention mechanism to capture the semantic and grammatical patterns of the written language of hearing impaired students. On this basis, we compose an encoder-decoder architecture along with re-ranking strategy and shuffle learning to restore the ungrammatical sentences into grammatical ones. Conducted experiments demonstrate that: 1) the multiple layers consisting of neat self-attention excluding the position embedding are capable of extracting the separate token representation, 2) via the re-ranking, our model is able to correct errors of varying degrees in a sentence at one go, 3) the pre-training operation, utilizing the sequential patterns of hearing people texts, helps the model to restore the wrong sentences to the correct ones, meanwhile accelerates the convergence of the model.

Similar to the regular GEC missions, the grammatical error correction task of hearing impaired student written language suffers the substantial paucity of ground-truth data with reliable correspondingly corrected labels, no matter in which language. In further research, we will figure out more methods so as that the GEC model can work better on sparse data and explore the chance implementing our architecture in other NLP tasks, such as machine translation tasks.

REFERENCES

[1] V. R. Charrow and J. D. Fletcher, "English as the second language of deaf children," *Develop. Psychol.*, vol. 10, no. 4, p. 463, 1974.

[2] W. Xie, P. Huang, X. Zhang, K. Hong, Q. Huang, B. Chen, and L. Huang, "Chinese spelling check system based on N-gram model," in *Proc. 8th SIGHAN Workshop Chin. Lang. Process.*, 2015, pp. 128–136.

[3] L. Zhang and H. Wang, "A unified framework for grammar error correction," in *Proc. 18th Conf. Comput. Natural Lang. Learn., Shared Task*, 2014, pp. 96–102.

[4] C.-J. Lin and W.-C. Chu, "A study on Chinese spelling check using confusion sets and N-gram statistics," *Int. J. Comput. Linguistics Chin. Lang. Process.*, vol. 20, no. 1, pp. 23–48, Jun. 2015.

[5] A. Rozovskaya and D. Roth, "Generating confusion sets for context-sensitive error correction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 961–970.

[6] C. Brockett, W. B. Dolan, and M. Gamon, "Correcting ESL errors using phrasal SMT techniques," in *Proc. ACL*, 2006, p. 249.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[8] Y. Hong, X. Yu, N. He, N. Liu, and J. Liu, "FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm," in *Proc. 5th Workshop Noisy User-Generated Text*, Hong Kong, 2015, pp. 160–169. [Online]. Available: https://www.aclweb.org/anthology/D19-5522

[9] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling error correction with soft-masked BERT," 2020, *arXiv:2005.07421*.

[10] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[12] T. Ge, F. Wei, and M. Zhou, "Reaching human-level performance in automatic grammatical error correction: An empirical study," 2018, *arXiv:1807.01270*.

[13] S. Chollampatt and H. T. Ng, "A multilayer convolutional encoder-decoder neural network for grammatical error correction," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 5755–5762.

[14] J. Lichtarge, C. Alberti, S. Kumar, N. Shazeer, N. Parmar, and S. Tong, "Corpora generation for grammatical error correction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 3291–3301.

[15] W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu, "Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 156–165.

[16] D. Dahlmeier and H. T. Ng, "Correcting semantic collocation errors with L1-induced paraphrases," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 107–117.

[17] D. Dahlmeier and H. T. Ng, "Better evaluation for grammatical error correction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2012, pp. 568–572.

[18] R. Dale and A. Kilgarriff, "Helping our own: The HOO 2011 pilot shared task," in *Proc. 13th Eur. Workshop Natural Lang. Gener.*, 2011, pp. 242–249.

[19] M. Chodorow, J. Tetreault, and N.-R. Han, "Detection of grammatical errors involving prepositions," in *Proc. 4th ACL-SIGSEM Workshop Prepositions*, 2007, pp. 25–30.

[20] R. De Felice and S. G. Pulman, "A classifier-based approach to preposition and determiner error correction in L2 English," in *Proc. 22nd Int. Conf. Comput. Linguistics*, 2008, pp. 169–176.

[21] N.-R. Han, J. R. Tetreault, S.-H. Lee, and J.-Y. Ha, "Using an error-annotated learner corpus to develop an ESL/EFL error correction system," in *Proc. LREC*, 2010, pp. 1–8.

[22] C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault, "Automated grammatical error detection for language learners," *Synth. Lectures Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–134, Jan. 2010.

[23] J. Tetreault, J. Foster, and M. Chodorow, "Using parse features for preposition selection and error detection," in *Proc. Assoc. Conf. Comput. Linguistics*, 2010, pp. 353–358.

[24] M. Kaneko, Y. Sakaizawa, and M. Komachi, "Grammatical error detection using error-and grammaticality-specific word embeddings," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2017, pp. 40–48.

[25] M. Rei and H. Yannakoudakis, "Compositional sequence labeling models for error detection in learner writing," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1–9.

[26] Y. Yang, P. Xie, J. Tao, G. Xu, L. Li, and L. Si, "Alibaba at IJCNLP-2017 task 1: Embedding grammatical features into LSTMs for Chinese grammatical error diagnosis task," in *Proc. IJCNLP*, 2017, pp. 41–46.

[27] C. Wang, R. Li, and H. Lin, "Deep context model for grammatical error correction," in *Proc. 7th ISCA Workshop Speech Lang. Technol. Educ.*, Aug. 2017, pp. 167–171. [Online]. Available: http://www.isca-speech.org/archive/SLaTE_2017/abstracts/SLaTE_2017_paper_5.html

[28] Y. Cao, L. He, R. Ridley, and X. Dai, "Integrating BERT and score-based feature gates for Chinese grammatical error diagnosis," in *Proc. 6th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2020, pp. 49–56.

[29] D. Liang, C. Zheng, L. Guo, X. Cui, and X. Xiong, "BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis," in *Proc. 6th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2020, pp. 57–66.

[30] H. Wang and M. Komachi, "TMU-NLP system using BERT-based pre-trained model to the NLP-TEA CGED shared task 2020," in *Proc. 6th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2020, pp. 87–90.

[31] Y. Zhou, Y. Shao, and Y. Zhou, "Chinese grammatical error diagnosis based on CRF and LSTM-CRF model," in *Proc. 5th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2018, pp. 97–101.

[32] S.-H. Wu, J.-W. Wang, L.-P. Chen, and P.-C. Yang, "CYUT-III team Chinese grammatical error diagnosis system report in NLPTEA-2018 CGED shared task," in *Proc. 5th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2018, pp. 102–107.

[33] Y. Cheng and M. Duan, "Chinese grammatical error detection based on BERT model," in *Proc. 6th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2020, pp. 108–113.

[34] B. B. Shrestha and B. K. Bal, "Named-entity based sentiment analysis of nepali news media texts," in *Proc. 6th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2020, pp. 114–120.

[35] H. Asano, M. Mita, T. Mizumoto, and J. Suzuki, "The AIP-Tohoku system at the BEA-2019 shared task," in *Proc. 14th Workshop Innov. NLP Building Educ. Appl.*, vol. 2019, pp. 176–182.

[36] K. Sakaguchi, M. Post, and B. Van Durme, "Grammatical error correction with neural reinforcement learning," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2017, pp. 366–372.

[37] A. Schmaltz, Y. Kim, A. Rush, and S. Shieber, "Adapting sequence models for sentence correction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2807–2813.

[38] Z. Xie, A. Avati, N. Arivazhagan, D. Jurafsky, and A. Y. Ng, "Neural language correction with character-based attention," 2016, *arXiv:1603.09727*.

[39] Z. Yuan and T. Briscoe, "Grammatical error correction using neural machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 380–386.

[40] A. Rozovskaya and D. Roth, "Grammatical error correction: Machine translation and classifiers," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 2205–2215, doi: 10.18653/v1/P16-1208.

[41] S. Kiyono, J. Suzuki, M. Mita, T. Mizumoto, and K. Inui, "An empirical study of incorporating pseudo data into grammatical error correction," in *Proc. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1236–1242.

[42] Z. Zhao and H. Wang, "MASKGEC: Improving neural grammatical error correction via dynamic masking," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 1226–1233.

[43] L. Namir and I. M. Schlesinger, "The grammar of sign language," *Sign Lang. Deaf*, vol. 4, pp. 97–140, Oct. 1978.

[44] S.-H. Wu, C.-L. Liu, and L.-H. Lee, "Chinese spelling check evaluation at SIGHAN Bake-off 2013," in *Proc. 7th SIGHAN Workshop Chin. Lang. Process.*, 2013, pp. 35–42.

[45] G. Tang, M. Müller, A. R. Gonzales, and R. Sennrich, "Why self-attention? A targeted evaluation of neural machine translation architectures," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4263–4272.

[46] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[47] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[48] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[49] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2978–2988.

[50] 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. Int. Conf. Natural Lang. Process. Chin. Comput.*, 2018, pp. 341–350.

[51] Y. Zhao, N. Jiang, W. Sun, and X. Wan, "Overview of the NLPCC 2018 shared task: Grammatical error correction," in *Proc. Int. Conf. Natural Lang. Process. Chin. Comput.*, 2015, pp. 439–445.

[52] M. Felice and T. Briscoe, "Towards a standard evaluation method for grammatical error detection and correction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2015, pp. 578–587.

[53] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[54] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground truth for grammatical error correction metrics," in *Proc. 7th Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2015, pp. 588–593.

[55] J. Zhou, C. Li, H. Liu, Z. Bao, G. Xu, and L. Li, "Chinese grammatical error correction using statistical and neural models," in *Proc. Int. Conf. Natural Lang. Process. Chin. Comput.*, 2018, pp. 117–128.

[56] H. Ren, L. Yang, and E. Xun, "A sequence to sequence learning for Chinese grammatical error correction," in *Proc. Int. Conf. Natural Lang. Process. Chin. Comput.*, 2018, pp. 401–410.

BINBIN CHEN received the B.E. degree from Xi'an International Studies University, in 2007, the M.S. degree from Southwest University, in 2016, and the Ph.D. degree from Xi'an International Studies University, in 2021. He is currently serving as an Associate Professor with Qiannan Normal University for Nationalities. His current research interests include machine learning, deep learning, cognitive linguistics, and natural language processing.

JINGYU ZHANG studied at Xi'an International Studies University, The University of Utah, and the Guangdong University of Foreign Studies. He is a Doctoral Supervisor for foreign linguistics and applied linguistics. He is a Sino–U.S. Fulbright Research Scholar, a candidate for the New Century Talent Support Program of the Ministry of Education, an outstanding Instructor for the National Master of Education degree, and an outstanding Teacher for the Sasakawa Medical Scholarship Program of the Ministry of Health.