IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 20, 2019, accepted August 13, 2019, date of publication September 10, 2019, date of current version October 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940607

A Two-Stage Model for Chinese Grammatical Error Correction

ZHAOQUAN QIU AND YOULI QU
School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
Corresponding author: Youli Qu (ylqu@bjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61976015.

ABSTRACT Chinese grammatical error correction (GEC) is more challenging than English GEC due to its language characteristics. In this paper, a two-stage model was proposed to solve the Chinese GEC problem. The model consists of two components: a spelling check model and a GEC model. The spelling check model based on language model focuses on correcting spelling errors, while the GEC model based on neural sequence-to-sequence (seq2seq) model focuses on correcting grammatical errors. In addition, two generative methods allow the seq2seq model to correct an erroneous sentence incrementally with repeated inference steps. Furthermore, only one seq2seq model is used for grammatical correction rather than ensemble multiple models, which greatly speeds up the generation of final results and saves computing resources. The two-stage model achieves 31.01 $F_{0.5}$ on NLPCC 2018 test set, significantly outperforms all prior approaches on this task.

INDEX TERMS Chinese grammatical error correction, spelling check, seq2seq model.

I. INTRODUCTION

Grammatical error correction (GEC) is an important task in natural language processing (NLP), which aims to detect and correct errors in text. The errors include not only grammatical errors, but also spelling errors and collocation errors.

In recent years, the most common approach for solving GEC problem is treating it as a machine translation problem from "bad" text to "good" text. Due to sequence to sequence (seq2seq) [1], [2] models' impressive performance in machine translation, applying seq2seq models to GEC has attracted widely attention of NLP researchers [3]–[7].

As the most widely used language in the world, English has always been the main focus of GEC task It has many shared tasks such as CoNLL-2013 [8] and CoNLL-2014 [9]. However, the research on Chinese GEC is much less. Previous work has mainly focused on the diagnosis of grammatical error [10], [11] rather than correction. The NLPCC 2018 shared task provides NLP researchers an opportunity to study and develop Chinese GEC.

The performance of seq2seq models is bottlenecked by the need for a large dataset of error-annotated sentence pairs [5] and of good quality. However, there is less error-annota8ted training data for Chinese GEC compared with English GEC. The dataset that we can only get currently is provided by NLPCC 2018 shared task. Furthermore, Chinese has unique characteristics in many aspects compared with English or other alphabetical languages, so we can't rely entirely on the work that the researchers have done in English GEC. To alleviate the data sparsity problem and language difference problem, we decompose the Chinese GEC task into two subtasks: spelling check based language model and grammatical error correction based seq2seq model. For a noisy, ungrammatical sentence, first the typos were corrected by spelling check model. Second, the intermediate results are translated into a clean, grammatically correct sentence by seq2seq model.

Unlike previous neural approaches for GEC, which pretrain word vectors to initialize the embeddings in both encoder and decoder sides or initialize the embeddings randomly [12], [13], we only initialize the embeddings in the decoder with pre-trained word vectors. Furthermore, we adopt two methods to correct a sentence incrementally with repeated inference in the grammatical correction stage. Seq2seq model may not be able to correct all errors in a sentence with multiple grammatical errors by just a single round inference. For a sentence, the first corrected part will benefit the model to correct the remaining errors. Also, we only use one seq2seq model for grammatical correct rather than ensemble multiple models or re-rank results produced by multiple NMT models, which greatly accelerates the generation of final results and saves computing resources.

In summary, this paper is organized as follows. Section II gives an overview of related work on GEC in recent years. In

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

Section III, a Two-stage model for Chinese grammatical error correction including two submodels is introduced. Section IV describes recycles generation methods for correcting the sentence better. Data preprocessing cycle generation method, experimental setup and results are described in Section V and VI. In Section VII a conclusion of the experiment is made.

II. RELATED WORK

For English GEC, the previous methods mainly focus on classifier-based techniques [14], [15]. A classifier is trained to correct a specific type of error and the classifier-based methods combine multiple classifiers for specific errors to build a hybrid system for English GEC. Statistical machine translation (SMT)-based systems attract widely attention due to its similarity with GEC task and its superior performance in correcting various types of errors [16], [7]. However, SMT framework suffers from its weak generalization capabilities and limited ability to capture global dependencies. Recently, various neural machine translation (NMT)-based methods have been proposed and achieved amazing effect.

Yuan and Briscoe [17] first applied NMT model to grammatical error correction task. They used an encoder-decoder recurrent neural network with attention mechanism [18] and achieved a better result than all prior systems. Chollampatt and Ng [3] improved the performance of GEC using a multilayer convolutional encoder-decoder neural network, which completely eliminated the huge performance gap between the neural and statistical methods of this task. Junczys-Dowmunt *et al.* [5] treated GEC problem as low-resource MT problem and proposed some model-independent methods that can be easily applied in GEC problem. More recently, Ge *et al.* [19] proposed a system based on 7-layer convolutional seq2seq models which combine fluency boost learning and fluency boost inference. The system achieves the state-of-the-art performance, becoming the first English GEC system that reaches human-level performance.

For Chinese, previous work mainly focuses on the diagnosis of grammatical error rather than correction. Both Yang *et al.* [10] and Zheng *et al.* [11] treated Chinese Grammatical error diagnosis as a sequence labeling task and built a system that mainly based on conditional random field (CRF) model and Long Short-Term Memory (LSTM) model. In 2018, NLPCC shared Chinese GEC task and it boosted the development of Chinese GEC.

Similar to Chollampatt and Ng [3], Ji *et al.* [4] built a Chinese GEC system that is based on the convolutional seq2seq model. Zhou *et al.* [20] combines rule-based models, SMT-based GEC models and NMT-based GEC model. Fu *et al.* [12] tackles Chinese GEC problem using stage approach. By combining a spelling error correction model and transformer model, they achieved the highest performance.

III. SYSTEM DESCRIPTION

This paper presents a two-stage model for Chinese GEC. As figure 1 shows, this model consists of two separate submodels that deal with different grammatical errors. The spelling check model mainly focuses on correcting the spelling errors in a sentence, especially nonword errors. In Chinese, nonword errors mean a word segmented by a word segmentation that can't be found in dictionary. The seq2seq model trained with original error-corrected sentence pairs mainly focuses on correcting the grammatical errors in the sentence, and it will correct the remaining spelling errors.

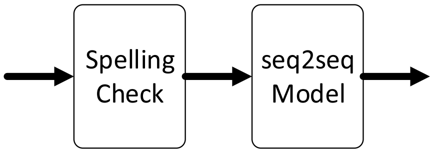


FIGURE 1. The two-stage model.

Figure 1 illustrates the two-stage model for Chinese GEC. In the following sections, the two submodels of the model are introduced in detail.

A. SPELLING CHECK BASED LANGUAGE MODEL

As we discuss in Section 1, there is no sufficient error-annotated sentence pairs to train seq2seq model perfectly in Chinese GEC. In addition, for a Chinese character, there will be different spelling mistakes in different contexts. Therefore, the seq2seq model cannot maximize its performance in this task. By taking the factors into consideration, spelling check is conducted to alleviate spelling errors problems before using seq2seq model to correct grammatical errors.

Since Chinese is an ideogram, the reasons for spelling errors are quite different from English or other alphabetic languages. In Chinese, a character is very likely to be mistakenly written in a form with similar pronunciation or similar shape. For example, the character '传' may be misspelled as '穿' or '转' in different context. So, for spelling check, the similar shape set and similar pronunciation set are essential to generate candidates to correct spelling mistakes. The above two similar character sets (SCS) are provided by SIGHAN 2013 CSC [21], [22]. Here are some examples of the similar character set:

-Similar Shape set: 也, 他地她弛池迆牠拖迤施

-Similar Pronunciation set: 工, 共攻宫红蚣恭肱躬功弓供龚公矿巩共汞贡供拱

The other important component of spelling check is language model. In our experiments, we use a n-gram language model to select the most probable word from candidates, which makes the original sentence get highest probability. Here, we set n=5 and the equation for calculating the probability of the sentence X=$w_1w_2w_3w_4w_5$ is refer to (1)

$$\begin{aligned} \mathrm{p}\,(\mathrm{X}) &= \mathrm{p}\,(w_1)\,p\,(w_2|w_1)\,p\,(w_3|w_1w_2)\cdots p\,(w_5|w_1w_2w_3w_4) \\ &= \prod_{i=1}^{5} p(w_i|w_1\cdots w_{i-1}) \qquad (1) \end{aligned}$$

In this model, Jieba is used for word segmentation. In a Chinese sequence, if the misspelled word is a nonword, it is very likely that jieba word segmentation still regards it as one word rather than segments it into two or more words. For example, if “不管” is mistakenly written as “不官 ”, Jieba will do nothing on this word, but other word segmentation tools will segment it into “不/官”

I. The spelling check algorithm is summarized as follows. First, a character sequence T is segmented by Jieba word segmentation into a segmented sequence T_1. For each word w in T_1, if it is not in dictionary D, each character of w is replaced with SCS to generate candidate substitution word set S_w. Then a language model LM is used to select the most likely word from S_w that makes the sequence T_1 get highest probability. The spelling check algorithm is shown in algorithm 1.

Algorithm 1 Spelling Check

```
Input: character sequence $T$
Input: dictionary $D$, similar character set $S$, language model $LM$
Output: $T^*$, character sequence without spelling error
$T_1 \leftarrow$ Segment sentence T using jieba word segmentation;
$T^* \leftarrow []$
//$|T_1|$ is the number of words in $T_1$
// $| w |$ is the number of character in $w$.
for $i \leftarrow 0$ to $|T_1|$ do
    $w \leftarrow T_1[i]$
    if $|w| == 1$ or $w \in D$ then
        continue
    $S_w \leftarrow \{w\}$
    for $j \leftarrow 0$ to $|w|$ do
        $x \leftarrow w[j]$
        if $x \in S$ then
            $L \leftarrow S[x]$
            for $c \in L$ do
                $w' \leftarrow w[:j] + c + w[j+1:]$
                if $w' \in D$ then
                    $S_w \leftarrow S_w \cup \{w'\}$
            end
    end
    $w_{best} \leftarrow \arg\max_{w' \in s_w} LM.score(T_1[:i]) + w' + T_1[i+1]:])$[1]
  $T^* \leftarrow T^* + [w'_{best}]$
end
return $T^*$
```

B. SEQ2SEQ-BASED GEC MODEL

After some spelling errors are corrected by spelling check model, the GEC task is treated as a translation task that translates a sequence of “bad” to a sequence of “good”. Seq2seq models for translation are widely used in GEC problem due to its superior performance. For an input sequence $X = (x_1, x_2, x_3, x_4, x_5)$, the model “translates” it into the corresponding target sequence $Y = (y_1, y_2, y_3, y_4, y_5)$, as show in Eq (2):

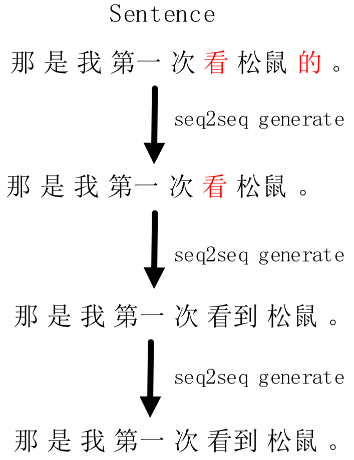

Sentence
那 是 我 第一 次 看 松鼠 的 。
seq2seq generate
那 是 我 第一 次 看 松鼠 。
seq2seq generate
那 是 我 第一 次 看到 松鼠 。
seq2seq generate
那 是 我 第一 次 看到 松鼠 。

FIGURE 2. Recycle generating.

$$p(Y|X) = \prod_{t=1}^{n} p(y_t|X, y_1 : \mathrm{y}_{t-1}; \theta) \tag{2}$$

There are many variants in seq2seq model. Transformer as a seq2seq model to GEC task in our experiments, which is proposed by Vaswani *et al.* [24], is completely based on attention mechanisms, dispensing with recurrence and convolutions entirely. Compared with other seq2seq models, transformer captures global dependencies between input and output sequence easily.

IV. GENERATE STYLE

A. RECYCLE GENERATING

All of multiple grammatical errors in a sentence may not be corrected with only single-round inference by seq2seq model.

So, a sentence is corrected through multi-round inference with the method, called Recycle Generating followed Ge *et al.* [19]. This procedure summarized as follows. First, an original sentence X is inputted into seq2seq model and a hypothesis X_1 is outputted after calculation. Then X_1 is regarded as the input of the model again and waits for next output X_2 instead of regarding X_1 as the final result. Until the probability of sentence X_t is bigger than X_{t-1}, this process will be terminated. Figure 2 gives an example of the method of recycle generation

B. RECYCLE GENERATING AND RE-RANKING

Due to small quantity of dataset used to train seq2seq model and the complex characteristics of Chinese, seq2seq model may correct a sentence incorrectly even if through recycle generation. This will leads the generated result to contain more errors than the original sentence. Therefore, we should try to avoid this situation.

Based on the idea of recycle generating, we make a re-ranking between the final result X_t and the original

[1] https://kheafield.com/code/kenlm/

sentence X. We use the language model to select a sentence with higher probability as our final result Y, as Eq (3) shows.

$$y = \max(p(X_t), p(X)) \tag{3}$$

where $p(X)$ is Eq (1), the probability of given sentence X.

V. PRE-TRAINING OF WORD EMBEDDINGS

Similar to Chollampatt and Ng [3], we use fastText [25] to train word embeddings with a large Chinese monolingual dataset. Each sentence of the dataset is first segmented by pkuseg word segmentation and then each word is splited into sub-word units by byte pair encoding (BPE) algorithm [18]. Compared with word2vec [26], [27], fastText takes into account both the morphological structure of words and the correlation between words. Word embeddings trained by fasettext empirically outperforms word2vec embeddings whether the embeddings are initialized randomly or not.

Unlike Chollampatt and Ng [3] or Junczys-Dowmunt *et al.* [5], who initialize the embeddings for the source and target words with pre-trained word embeddings, the word embeddings of target side only is initialized by pre-trained word embeddings and the embeddings of source words is initialized randomly. Experiments show this way to initialize the target word with pre-trained word embeddings can greatly improve the performance of the model and get better results.

VI. EXPERIMENTS

A. DATASETS

The error-corrected dataset provided by NLPCC 2018 Shared Task 2 was used to train our seq2seq model. The dataset is collected from Lang-8[2] website and contains approximately 1.22M sentence pairs. Similar to Hassan *et al.* [28], the dataset was filtered according to the following criteria:

- The length of source sentence and target sentence are between 6 and 100.
- Pairs where (source length >1.5* target length or target length >1.5* source length) are removed.
- Traditional Chinese in the sentence are converted to simplified Chinese using OpenCC[3] converter.
- The full-width characters in the sentence are converted to half-width characters.

The sentence pairs that are identical and have no grammatical error were retained, they are in favour of performance of seq2seq model. After filtration, the quantity of dataset is reduced to 1.09M. Following Ji *et al.* [4], we randomly select 5k sentence pairs from the dataset as our validation set. Thus, the remaining 1.08M error-corrected sentence pairs were used as our training set. The test set is also provided by NLPCC 2018.

Sogou[4] dataset and wikipedia dataset were used to pre-train word embeddings. All the data (parallel and monolingual) have been segmented with pkuseg word segmentation and the BPE algorithm is applied to split rare words into multiple frequent sub-words. We learn a BPE model with 35K merge operations.

TABLE 1. Statistics for datasets.

Data	Sents
Training set	1086933
Validation set	5000
Test set	2000
Sogou dataset	55.9M
Wikipedia dataset	9.1M
News dataset	52.7M

We use a large Chinese dataset from Sogou, Wikipedia and news dataset crawled from the internet to train 5-gram kenLM language model. The total dataset we use is about 16GB and it has approximately 117.6 million sentences. We use language model as an assistant model to provide features for selecting the most likely result. Table 1 illustrates the details of the datasets which are used in our experiments.

In spelling check model, the dictionary D used for checking non-word errors is SogouW[5] from Sogou Inc. The similar character set S mentioned in Section 3.1 is provided by SIGHAN 2013 CSC datasets, which is used to substitute characters to construct the correct word. As the original character in CSC is traditional Chinese, we use OpenCC converter to convert it to simplified Chinese. Jieba were chosen to segment Chinese sentence. The former is employed in the stage of spelling check, and latter is used for segmenting the error-corrected dataset and other dataset.

B. MODEL AND TRAINING SETTING

In our experiments, the Transformer sequence-to-sequence model implemented[6] by FAIR with PyTorch was used. The dimensionality of source and target word embeddings is set to 512. And both encoder and decoder have 6 identical layers and 8 attention heads. We set the dimensionality of the inner-layer in position-wise feed-forward network to 2048. In total, the model has 101M parameters. During training process, the initial learning rate was set to 0.0005, and the model was optimized with Nesterov Accelerated Gradient [29]. The momentum value was set to 0.99 and dropout rate was set to 0.2. The training stopped if the learning rate dropped below 10^{-9} or the number of parameter updates exceeded 250000 times. In the stage of training Transformer, checkpoints are produced approximately every 30 minutes.

Similar to Heafield *et al.* [23] and Junczys-Downmunt *et al.* [5], the average of the 7 checkpoints near the best checkpoint was taken as our final seq2seq model. In the decoding time, the beam size was set to 12 as Ge *et al.* [19].

C. EXPERIMENTAL RESULTS

We compare our system to the following three Chinese GEC systems proposed in NLPCC 2018:

[2] http://lang8.com/

[3] http://code.google.com/p/opencc/.

[4] https://www.sogou.com/labs/resource/list_news.php

[5] https://www.sogou.com/labs/resource/w.php

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

TABLE 2. Comparison of GEC systems on NLPCC 2018 test set.

System	Precision	Recall	$F_{0.5}$
Ren et al. 2018	41.73	13.08	29.02
Zhou et al. 2018	41.00	13.75	29.36
Fu et al. 2018	35.24	18.64	29.91
Our Model	36.88	18.94	31.01

TABLE 3. Results on the NLPCC 2018 test set.

System	Precision	Recall	$F_{0.5}$
Spelling Check	48.01	4.84	17.26
Base Transformer	29.49	14.47	24.42
Base +embed	31.19	15.17	25.75
Base +embed + RG1	31.75	15.98	26.52
Base+ embed + RG2	34.41	15.25	27.50
Base+ embed+RG2+ Spelling Check	36.88	18.94	31.01

+embed denotes to initialize the word embeddings of decoder by pre-trained word embeddings. +RG1 refers to recycle generating mentioned in section 4.1 and +RG2 refers to recycle generating and re-ranking mentioned in section 4.2. +Spelling Check denotes spelling check model proposed in section 3.1

TABLE 4. Results of different embedding initializations on the NLPCC 2018 test set.

Initialization	Precision	Recall	$F_{0.5}$
Random	29.49	14.47	24.42
Word2vec	29.82	14.44	24.58
FastText	31.19	15.17	25.75

- Ren et al. [13]: A seq2seq model based entirely on convolutional neural network.
- Zhou et al. [20]: The system combines rule-based model, SMT-based model and NMT-based model.
- Fu et al. [12]: the state-of-the-art Chinese GEC system on NLPCC 2018 dataset, which is based on spelling error correction model and NMT model.

Table 1 shows the results of Chinese GEC system on NLPCC 2018 test set. Without ensem bling multiple models, our model outperforms all previous Chinese GEC systems. It achieves 31.01 $F_{0.5}$ score, which improves $+1.1$ $F_{0.5}$ points than state-of-the-art systems.

We evaluate each components of our system on the test set, and the results are shown in table 3. Our base Transformer model with initializing the word embeddings of decoder using pre-trained word embeddings achieves 25.75 $F_{0.5}$. With the two generation methods we mentioned in section 4, the performance reaches 26.52 $F_{0.5}$ and 27.50 $F_{0.5}$ respectively. Recycle generation method improves both precision (from 31.19 to 31.75) and recall (from 15.17 to 15.98). Compared to recycle generation, recycle generation and re-ranking method improve precision greatly (from 31.19 to 34.41) while recall value is basically unchanged (from 15.17 to 15.25). After combining the spelling check and the method of recycle generation and re-ranking, the performance of our system has increased to 31.01 $F_{0.5}$, significantly outperforming the previous published best result of $F_{0.5} = 29.91$(Fu *et al.* [12]).

We also make a comparison between different methods of initializing the target word embeddings, the results are shown in table 4. We use default parameters to train word2vec and fastText embeddings. As result shows, initializing with faseText works better than word2vec and random initializing. Therefore, we choose fasetText as our tools to pre-train word embeddings.

VII. CONCLUSION

In this work, a Two-stage model was presented for Chinese grammatical error correction. The model includes two independent models, a spelling check model based language model and a seq2seq-based GEC model. First, spelling check model was used to solve the non-word spelling error problem. Then the seq2seq model was used to correct the grammatical errors. We also utilize pre-trained word embeddings to initialize the decoder of seq2seq model and adopt two generate methods to improve a sentence's fluency. By combining these two models, our system achieves tremendous improvement compared with state-of-art results on NLPCC 2018 benchmark datasets.

REFERENCES

[1] 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," in *Proc. EMNLP*, Doha, Qatar, Oct. 2014, pp. 1724–1734.

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

[3] S. Chollampatt and H. T. Ng, "A multilayer convolutional encoder-decoder neural network for grammatical error correction," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 5755–5762.

[4] 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. 55th ACL*, Vancouver, BC, Canada, Jul. 2017, pp. 753–762.

[5] M. Junczys-Dowmunt, R. Grundkiewicz, S. Guha, and K. Heafield, "Approaching neural grammatical error correction as a low-resource machine translation task," in *Proc. NAACL-HLT*, New Orleans, LA, USA, Jun. 2018, pp. 595–606.

[6] Z. Xie, A. Avati, N. Arivazhagan, D. Jurafsky, and A. Y. Ng, "Neural language correction with character-based attention," *CoRR*, vol. abs/1603.09727, 2016. [Online]. Available: https://arxiv.org/abs/1603.09727

[7] Z. Yuan and M. Felice, "Constrained grammatical error correction using statistical machine translation," in *Proc. CoNLL*, Sofia, Bulgaria, Aug. 2013, pp. 52–61.

[8] H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, "The CoNLL-2013 shared task on grammatical error correction," in *Proc. 17th Comput. Natural Lang. Learn. Conf.*, Sofia, Bulgaria, Aug. 2013, pp. 1–12.

[9] 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. 18th Conf. Comput. Natural Lang. Learn.*, Baltimore, MD, USA, Jun. 2014, pp. 1–14.

[10] 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. IJCLP*, Taipei, Taiwan, Nov. 2017, pp. 41–46.

[11] B. Zheng, W. Che, J. Guo, and T. Liu, "Chinese grammatical error diagnosis with long short-term memory networks," in *Proc. Natural Lang. Process. Techn. Educ. Appl.*, Osaka, Japan, Dec. 2016, pp. 49–56.

[12] 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*, Hohhot, China, Aug. 2018, pp. 341–350.

[13] H. Ren, L. Yang, and E. Xun, "A sequence to sequence learning for Chinese grammatical error correction," in *Proc. NLPCC*, Hohhot, China, Aug. 2018, pp. 401–410.

[14] R. Dale and A. Kilgarriff, "Helping our own: The HOO 2011 pilot shared task," in *Proc. 13th ENLG*, Nancy, France, Sep. 2011, pp. 242–249.

[15] D. Dahlmeier, H. T. Ng, and E. J. F. Ng, "NUS at the HOO 2012 shared task," in *Proc. Building Educ. Appl.*, Montreal, QC, Canada, Jun. 2012, pp. 216–224.

[16] M. Junczys-Dowmunt and R. Grundkiewicz, "The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation," in *Proc. CoNLL*, Baltimore, MD, USA, Jun. 2014, pp. 25–33.

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

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," presented at the 3rd ICLR, San Diego, CA, USA, May 2015.

[19] T. Ge, F. Wei, and M. Zhou, "Reaching human-level performance in automatic grammatical error correction: An empirical study," *CoRR*, vol. abs/1807.01270, 2018. [Online]. Available: https://arxiv.org/abs/1807.01270

[20] J. Zhou, C. Li, H. Liu, Z. Bao, G. Xu, and L. Li, "Chinese grammatical error correction using statistical and neural models," in *Proc. NLPCC*, Hohhot, China, Aug. 2018, pp. 117–128.

[21] C.-L. Liu, M.-H. Lai, K.-W. Tien, Y.-H. Chuang, S.-H. Wu, and C.-Y. Lee, "Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications," *ACM Trans. Asian Lang. Inf. Process.*, vol. 10, no. 2, p. 10, Jun. 2011.

[22] 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.*, Nagoya, Japan, Oct. 2013, pp. 35–42.

[23] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, Sofia, Bulgaria: ACL, Aug. 2013, pp. 690–696.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[25] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

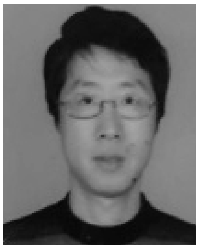
[26] T. Mikolov, K. Chen, G. Corradb, and J. Dean, "Efficient estimation of word representations in vector space," presented at the 1st ICLR, Scottsdale, AZ, USA, May 2013.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.

[28] H. Hassan *et al.*, "Achieving human parity on automatic Chinese to English news translation," *CoRR*, vol. abs/1803.05567, 2018. [Online]. Available: https://arxiv.org/abs/1803.05567

[29] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML*, Atlanta, GA, USA, Jun. 2013, pp. 1139–1147.

ZHAOQUAN QIU received the B.E. degree in computer science from the China University of Petroleum, in 2017. He is currently pursuing the master's degree with the School of Information and Technology, Beijing Jiaotong Univesity. His main focus is NLP.



YOULI QU received the Ph.D. degree in computer science from the Department of Computer Science, Beijing Institute of Technology, China, in 2000.

He is currently a Senior Engineer with Beijing Jiaotong University. His research interests include web and text mining, semantic web, and natural language processing.