MaskGEC: Improving Neural Grammatical Error Correction via Dynamic Masking

Zewei Zhao, Houfeng Wang
MOE Key Lab of Computational Linguistics, Peking University, Beijing, 100871, China
{zhaozewei, wanghf}@pku.edu.cn

Abstract

Grammatical error correction (GEC) is a promising natural language processing (NLP) application, whose goal is to change the sentences with grammatical errors into the correct ones. Neural machine translation (NMT) approaches have been widely applied to this translation-like task. However, such methods need a fairly large parallel corpus of error-annotated sentence pairs, which is not easy to get especially in the field of Chinese grammatical error correction. In this paper, we propose a simple yet effective method to improve the NMT-based GEC models by dynamic masking. By adding random masks to the original source sentences dynamically in the training procedure, more diverse instances of error-corrected sentence pairs are generated to enhance the generalization ability of the grammatical error correction model without additional data. The experiments on NLPCC 2018 Task 2 show that our MaskGEC model improves the performance of the neural GEC models. Besides, our single model for Chinese GEC outperforms the current state-of-the-art ensemble system in NLPCC 2018 Task 2 without any extra knowledge.

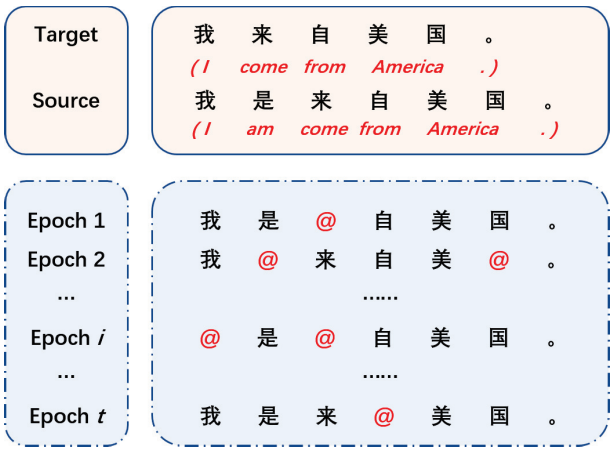

Figure 1: An example of error-corrected sentence pairs and the generated noisy sentence pairs during the whole training time. The placeholder '@' denotes any possible word that is chosen as a replacement.

Introduction

Grammatical error correction (GEC) has attracted much interest as a natural language processing (NLP) application in recent years. The definition of the grammatical error correction task is: given a sentence which may contain grammatical errors, one is required to detect and correct the errors presented in the sentence, and return its error-free natural language representation. Regarding the incorrect sentences as source language and the corrected sentences as target language, the GEC task can be treated as a machine translation (MT) task. For example, English GEC can be converted to the translation from "bad" English to "good" English.

With the rapid development of deep learning, neural machine translation (NMT) approaches based on sequence-to-sequence (seq2seq) models have become mainstream in the field of machine translation. Recently, quite a few works (Yuan and Briscoe 2016; Ji et al. 2017; Chollampatt and Ng 2018) have applied the neural seq2seq models to the grammatical error correction tasks and have made some progress. However, these NMT-based models for GEC face a problem. Due to the limited size of the parallel corpus of error-corrected sentence pairs, the seq2seq models for GEC, which usually contain millions of parameters, are difficult to be trained sufficiently. Therefore, even if a test case sentence is just slightly different from a training instance, the models may fail to correct it.

In order to overcome the drawback of neural grammatical error correction models which is mentioned above, we propose a simple yet effective dynamic masking method to enhance the performance of neural GEC models.

In the training procedure, We add various kinds of random noises to the inputs via masking to generate noisy source sentences dynamically, but keep the target sentences unchanged. By pairing the new source sentences with the corresponding target sentences, we can obtain more abundant error-corrected sentence pairs, as Figure 1 shows. For the sake of convenience, we call the newly constructed error-corrected sentence pairs through random noising *noisy sentence pairs*. Rather than use the aforementioned noisy sen-

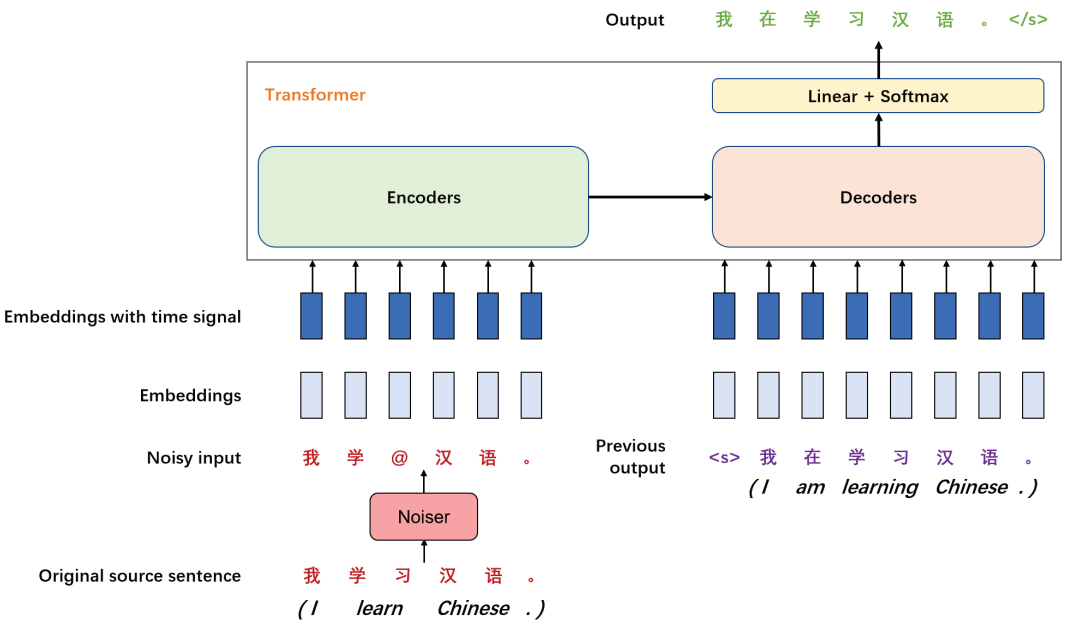

Figure 2: An illustration of the training process of our dynamic masking approach for Chinese grammatical error correction. The seq2seq architecture is Transformer. '@' is a placeholder which denotes a possible substitution word. '<s>' and '</s>' refer to 'BOS' (begin of sentence) and 'EOS' (end of sentence), respectively.

tence pairs as additional training instances, we substitute the original sentences with the noisy ones on the source side directly. In this way, our grammatical error correction model can obtain more samples of error-corrected sentence pairs during the whole training process, without increasing the size of the training set. By the introduction of noise, the generalization ability of the grammatical error correction model is enhanced in our approach.

Experiments demonstrate that grammatical error correction model with the proposed dynamic masking method outperforms the baseline seq2seq model, and achieves state-of-the-art results in the Chinese GEC task.

In short, this paper makes the following contributions:

- We propose a simple yet effective dynamic masking method to address the limitation of the Chinese neural GEC model. To the best of our knowledge, it is the first work to introduce dynamic masking technique to Chinese GEC tasks.
- Our model achieves state-of-the-art results in NLPCC 2018 Task 2 without additional resources, which proves the effectiveness of our approach in the Chinese GEC task.

Model

Neural GEC Model

A seq2seq model basically consists of an encoder-decoder architecture. Seq2seq models have been proven to be effective in many NLP tasks, such as machine translation (Sutskever, Vinyals, and Le 2014), text summarization (Rush, Chopra, and Weston 2015), dialogue systems (Serban et al. 2016), and so on. To correct the potential errors, GEC systems have to understand the meaning of the sentences. It could be hard as long distance dependencies may exist between the words in a natural language sentence. Recurrent neural networks (RNNs) are good at modeling word sequences and capturing the context of sentences. Therefore, RNN is mostly adopted by former neural models for GEC (Yuan and Briscoe 2016), especially for its variety gated recurrent unit (GRU) network (Xie et al. 2016; Ge, Wei, and Zhou 2018). Because most of the grammatical errors are localized and dependent on the nearby words, it is essential for GEC systems to capture local contexts. Convolutional neural networks (CNNs) are able to capture local information effectively by window operations. Through hierarchical multi-layer convolutional network (Chollampatt and Ng 2018), wider contexts between distant words can also be captured by higher layers. Attention mechanism (Bahdanau, Cho, and Bengio 2014; Luong, Pham, and Manning 2015) has led to great achievements on sequence learning tasks since proposed. Recent neural grammatical error correction models (Xie et al. 2016; Ji et al. 2017; Chollampatt and Ng 2018; Ge, Wei, and Zhou 2018) have introduced attention mechanism to let the models concentrate on the relevant parts with the grammatical errors in the sentences.

Most former neural GEC models use RNN or CNN as encoder and decoder, while Transformer (Vaswani et al. 2017) is a new kind of encoder-decoder framework. Transformer, proposed by Google recently, is based solely on attention mechanisms. Transformer has demonstrated its strong ability to model word sequences, and has achieved the best per-

Algorithm 1 Dynamic masking method

```
Initialize the neural grammatical error correction model Θ with random weights.
for each training epoch t do
    S^(t) ← ∅.
    for each (X, Y) ∈ S do
        Establish a noisy sentence pair (X̃^(t), Y) by applying a noising scheme f to X.
        S^(t) ← S^(t) ∪ {(X̃^(t), Y)}.
    end for
    Update model Θ with S^(t).
end for
```

formance in the machine translation task.

Our grammatical error correction model adopts the Transformer as the NMT framework. It is worth mentioning that the choice of the NMT framework is not a focus of this paper. We expect that other seq2seq models would benefit from our approach.

Given a source sequence

$$X = (x_1, x_2, \ldots, x_m) \tag{1}$$

and its corresponding corrected sequence

$$Y = (y_1, y_2, \ldots, y_n) \tag{2}$$

where m and n are the lengths of sequence X and Y respectively, the grammatical error correction model needs to estimate the following conditional probability:

$$P(Y|X) = \prod_{i=1}^{n} P\left(y_i|y_1, \ldots, y_{i-1}, X; \Theta\right) \tag{3}$$

where Θ is model parameters. The model is trained by maximum likelihood estimate (MLE), i.e. minimizing the negative log likelihood (NLL) loss:

$$l(\Theta) = -\sum_{i=1}^{n} \log\left(P\left(y_i|y_1, \ldots, y_{i-1}, X; \Theta\right)\right) \tag{4}$$

Dynamic Masking

For neural network models, the size of a training corpus is usually one of the key factors of the model performance. In order to obtain more training samples conveniently and efficiently, we add noises to source sentence X with a certain probability in the j-th epoch of the training process dynamically (Figure 2), and get the noisy text

$$\tilde{X}^{(j)} = (\tilde{x}_1^{(j)}, \ldots, \tilde{x}_i^{(j)}, \ldots, \tilde{x}_m^{(j)}) \tag{5}$$

where the i-th word in $\tilde{X}^{(j)}$ is given by:

$$\tilde{x}_i^{(j)} = \begin{cases} f(x_i), & p \leq \delta \\ x_i, & p > \delta \end{cases} \tag{6}$$

where f is a word substitution function, p is a random number generated by a uniform distribution over the interval [0.0, 1.0], and δ is the threshold of substitution probability[1].

[1] we set $\delta = 0.3$

Figure 3: Some examples of grammatical errors in Chinese texts caused by confusions between homophones. Their parts of speech are also listed to show the universality of this error type.

During the t rounds of iteration, a group of noisy source texts $\{\tilde{X}^{(1)}, \tilde{X}^{(2)}, \ldots, \tilde{X}^{(t)}\}$ are generated. Our GEC model is required to map this group of noisy texts to the target sentence Y. We describe our noise training method in Algorithm 1, where S is the set of original error-corrected sentence pairs in the training corpus, and $S^{(t)}$ is the set of noisy sentence pairs in epoch t.

Different noising schemes may have different impacts on the model performance. We consider the following noising schemes and conduct a set of experiments to compare them:

Padding Substitution: Every word in the source sentence has a certain probability δ to be chosen and substituted with a padding symbol '<pad>' . Through padding substitution, we can increase the training samples exponentially in the training process of the GEC model, and decrease the repetition of training instances. Besides, we can reduce the GEC model's dependency on specific words by replacing some words with padding symbols. In this way, the GEC model is forced to learn the meaning of the substituted words from the context in hidden layers, which helps to boost the performance.

Random Substitution: Similarly to padding substitution, the GEC model randomly picks out some words from the source sentence with a probability δ. However, the model uses the random words from the vocabulary V to replace them, instead of the padding symbols. The words for substitution are uniformly sampled from the vocabulary V with a probability of $1/|V|$, where $|V|$ is the size of V. Random substitution is more suitable for GEC task than padding substitution, as it can make the model generate noisy samples which are closer to the real wrong texts.

Word Frequency Substitution: Generally, the grammatical errors in natural language tend to mistake high frequency words for low frequency ones. Therefore, we believe that words with higher frequencies should appear in the incorrect source sentences as substitution errors more often. Accordingly, we propose a substitution method which is based on word frequency. Our GEC model counts the occurrences of each word in the target sentences within the training corpus to get the word frequency. Then it calculates the probability

Split	# Sent.	# Token src.	# Token tgt.
Train	1.2 M	23.7 M	25.0 M
Dev	5,000	99.3 K	104.1 K
Test	2,000	58.9 K	-

Table 1: Overview of NLPCC-2018 dataset

distribution $Prob(V)$ of the vocabulary V. In the training procedure, the GEC model samples the words for substitution according to word frequency rather than uniformly.

Homophone Substitution: Errors caused by homophony phenomenon account for a large proportion in grammatical errors in Chinese text. There exist lots of homophones in Chinese characters. They have the same pronunciation, but differ in shapes and meanings. Figure 3 demonstrates some grammatical errors caused by this phenomenon. We use *pypinyin*[2] to obtain the characters' corresponding pinyin, the official romanization system for standard Chinese. Then we categorize the words in the target sentences according to their pinyin, and count word frequencies based on the pinyin categories. Thus we can get the probability distribution of words of each pinyin type. While training the model, we choose the words to be replaced by means of the method described before. Then we look up the pinyin of these words and select homophones for substitution based on the corresponding word frequency distribution.

Mixed Substitution: In addition to the single noising schemes above, we also propose a mixed substitution method. For each training instance, our Chinese GEC model randomly select a single noising scheme or the empty scheme (keep it unchanged), and apply it to the training procedure. In this way, we integrate all the single noising schemes and obtain more diverse noisy sentence pairs.

Experiments

Setup

To validate the effectiveness of our approach in the Chinese grammatical error correction task, we conduct a set of experiments on the dataset of NLPCC 2018 Task 2 (Zhao et al. 2018)[3]. The statistics of the dataset are given by Table 1. This shared task provides the first and latest benchmark dataset for Chinese GEC. The corpus of this task is collected from the Lang-8[4] website, a language learning platform where native speakers correct what you write. The essays in the corpus are written by CSL (Chinese as a Second Language) students and corrected by native speakers in China. It should be noted that our model does not use any additional natural language resources.

As an incorrect sentence may have several corrected versions in this dataset, we combine the source sentence with each corrected sentence one by one to build the parallel corpus. In this way, we get 1.2 million sentence pairs in all. Since there is no official development data, we randomly sample 5,000 sentence pairs as our development set from the training set following prior work (Ren, Yang, and Xun 2018). The official test set contains 2,000 sentences extracted from PKU Chinese Learner Corpus, which is composed of essays written by foreign college students. Two groups of annotations are provided to give out the gold edits of grammatical errors in these sentences.

We use the official MaxMatch (M^2) (Dahlmeier and Ng 2012) scorer to evaluate our GEC models and compare them to previous systems in this shared task. Given a source sentence and a system hypothesis, M^2 scorer computes all the possible sequences of phrase-level edits between them, and finds the edit sequence that achieves the highest overlap with the gold standard annotation. Then the optimal sequence is used to compute the value of precision, recall and $F_{0.5}$. The official evaluation metric is the $F_{0.5}$ value given by M^2 scorer with all groups of annotations taken into consideration.

Model and Training Details

We implement our Chinese grammatical error correction model using OpenNMT-py[5], a neural machine translation toolkit developed in PyTorch[6]. The details for the training process and our GEC model are as follows: the architecture of the seq2seq model is the base model of Transformer. The encoder is a stack of 6 identical layers with two sub-layers, which are a multi-head self-attention layer and a position-wise fully connected feed-forward network. The decoder is also a stack of 6 identical layers. However, in the middle of each layer there is a third sub-layer which performs multi-head attention over the output of the encoder stack. The number of heads for Transformer self-attention is set to 8. The size of hidden Transformer feed-forward is 2,048. Both the dimension of word vectors on the source side and the target side are 512. The parameters of our model are initialized with Xavier's method (Glorot and Bengio 2010). Position encoding is applied as suggested. In the Chinese grammatical error correction task, we use the tokenization script from the BERT project[7] to tokenize the Chinese texts and keep the non-Chinese words unchanged. We apply dropout (Srivastava et al. 2014) operations on the encoders and decoders, with a probability of 0.1. Our model adopts the Adam optimizer with an initial learning rate of 2, and a beta value of (0.9, 0.998). We use Noam's learning rate decay scheme (Vaswani et al. 2017), warmup_steps = 8,000. We add label smoothing with a epsilon value of 0.1. The batch size is set to 4,096 tokens. The beam size is 12 during the time of model inference. We adopt the early-stopping technique and choose the best models according to the validation perplexity on the development set.

[2]https://github.com/mozillazg/python-pinyin

[3]http://tcci.ccf.org.cn/conference/2018/taskdata.php

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

[5]https://github.com/OpenNMT/OpenNMT-py

[6]https://pytorch.org

[7]https://github.com/google-research/bert

System	Model type	Resources	P	R	$F_{0.5}$
YouDao (Fu, Huang, and Duan 2018)	Ensemble	*LM*, *SCS*	35.24	18.64	29.91
AliGM (Zhou et al. 2018)	Ensemble	*LM*, *Emb.*	41.00	13.75	29.36
BLCU (Ren, Yang, and Xun 2018)	Single	*Emb.*	41.73	13.08	29.02
BLCU (ensemble) (Ren, Yang, and Xun 2018)	Ensemble	*Emb.*	**47.63**	12.56	30.57
Char-Transformer	Single	-	36.57	14.27	27.86
Dropout-Src (Junczys-Dowmunt et al. 2018)	Single	-	39.08	18.80	32.15
Ours	Single	-	44.36	**22.18**	**36.97**

Table 2: Performance of systems on the NLPCC-2018 dataset. *Ours* refers to applying dynamic masking method with the mixed substitution noising scheme based on Char-Transformer. In the Resources column, *LM* denotes language model, *SCS* denotes similar character set, *Emb.* denotes pre-trained word embeddings.

Baselines

We compare our models to all previous systems for Chinese grammatical error correction in the NLPCC 2018 Task 2.

The best performing systems evaluated on NLPCC-2018 dataset are listed as follows:

- YouDao (Fu, Huang, and Duan 2018): The Chinese GEC system developed by NetEase Youdao Co., LTD. Five different hybrid models correct sentences stage by stage independently. A language model is used to re-rank the outputs for ensembling.
- AliGM (Zhou et al. 2018): The Chinese GEC system developed by Alibaba, which combines NMT-based approaches, SMT-based approaches and a rule-based approach together with various modules.
- BLCU and BLCU (ensemble) (Ren, Yang, and Xun 2018): The Chinese GEC systems developed by BLCU. The former system is based on a multi-layer convolutional seq2seq model, and the latter is an ensemble of four single models with different initialization.

In order to verify the effectiveness of our dynamic masking method on Chinese neural grammatical errors correction models, we implement a character-based Transformer model (Char-Transformer) for Chinese GEC as our baseline.

Based on the Char-Transformer, we also re-implement the source-word dropout method proposed by Junczys-Dowmunt et al. (2018). Following their work, we set the full embedding vector for a source word to 0 with a probability p_{src}, all other embedding values are scaled with $1/(1-p_{\text{src}})$. They presented that dropout over source words can bring gains for neural grammatical error correction. By the introduction of corruption on the source side, the model is taught to reduce trust into the input and to apply corrections more aggressively.

Results

We compare our best model Char-Transformer with dynamic masking using the mixed substitution scheme to the state-of-the-art systems evaluated on NLPCC-2018 dataset. Table 2 demonstrates the evaluation results of our approach and prior systems on this Chinese GEC benchmark dataset using the official scorer.

Masking strategy	P	R	$F_{0.5}$
Static	43.73	21.71	36.35
Dynamic	44.36	22.18	36.97

Table 3: Comparison between static and dynamic MaskGEC models on the NLPCC-2018 dataset.

The Char-Transformer model gets an $F_{0.5}$ score of 27.86. There is a fairly big gap between this baseline model and the leading Chinese GEC systems of the shared task. However, after applying dynamic masking with the mixed substitution scheme, our model reaches 36.97 $F_{0.5}$ score. The result significantly outperforms the YouDao system ($F_{0.5} = 29.91$), which ranks first in the contest, by a large margin of 7.06 $F_{0.5}$. On the basis of BLCU system, Ren, Yang, and Xun build an ensemble model and achieve an $F_{0.5}$ score of 30.57, which is the best published result in NLPCC 2018 Task 2 so far. Despite this, the proposed model for Chinese GEC yields a higher $F_{0.5}$ than this current best one, which establishes a new state-of-the-art result on the NLPCC-2018 dataset. It is noteworthy that all the top three models in NLPCC 2018 Task 2 are ensemble models, but our single model still surpasses them. Our proposed approach is totally orthogonal to these ensembling methods, which means that our GEC model may achieve better results through these methods.

The approach of Junczys-Dowmunt et al. (2018) reaches 32.15 $F_{0.5}$, an improvement of 4.29 $F_{0.5}$ over the Char-Transformer model. Despite this, our dynamic masking model still beats it by a significant margin of 4.82 $F_{0.5}$. The reason is that our proposed dynamic masking method yields more diverse noisy sentence pairs, which may benefit the generalization ability of our GEC model.

Effect of Dynamic Masking

Different timings to perform masking operations lead to two kinds of noising strategies.

The *static* masking strategy performs masking once during data preprocessing, resulting in a single static mask. Training data can be duplicated k times to avoid using the same mask for each training instance in every epoch. Therefore, each source sentence is masked in k different ways over the dozens of epochs of training. Each training instance

Model	Precision	Recall	$F_{0.5}$ (official)	F_1 (unofficial)
Char-Transformer	36.57	14.27	27.86	20.53
+ padding	41.59	18.21	33.09	25.33
+ random	37.87	20.47	32.37	26.58
+ word frequency	32.25	**23.80**	30.11	27.39
+ homophone	34.69	18.95	29.75	24.51
+ mixed	**44.36**	22.18	**36.97**	**29.57**

Table 4: Performance of our NMT-based models for Chinese grammatical error correction with different noise training schemes on the NLPCC-2018 dataset.

is seen with the same mask several times during the whole training procedure.

By contrast, the *dynamic* masking strategy generates a new random masking pattern every time we feed a source sentence into the encoder. As a result, each training instance may be seen with a different mask in different epochs. It's vital when we need to pretrain models for more steps or with larger datasets.

Table 3 shows that our MaskGEC model with dynamic masking performs better than the static one in the Chinese grammatical error correction task.

Effect of Noising Schemes

We also investigate the influence of various noising schemes. The results are shown in Table 4.

It can be observed that all the proposed noising schemes improve the performance over the normal NMT-based Chinese GEC model via comparing by rows in Table 4. Take the simplest padding substitution noising scheme as an example. Our model based on the padding scheme gets a precision rate of 41.59 and a recall rate of 18.21, which means an increment of 5.02 precision and 3.94 recall over the baseline model. The result shows that our dynamic masking method allows the neural GEC model to detect more errors as well as correct errors better than the original one.

One reason for the improvement on the precision metric is that the substitution of words reduces the GEC model's dependency on wrong patterns like specific words or collocations. As a result, the neural GEC model is forced to capture the context information of the replaced or missing words through the noise training process, which may contribute to the correction of grammatical errors.

On the other hand, the introduction of random noises raises the number of training samples exponentially in the training procedure, and avoids the over-fitting problem caused by meaningless repetition of training instances. Consequently, our dynamic masking approach also gets higher recall rates in Chinese GEC task, since extremely diverse incorrect sentences are generated to allow the model to detect errors more aggressively.

The system performance among our neural GEC models applying the four single noising methods is quite different. The padding substitution scheme achieves the highest precision rate of 41.59, while its recall rate (18.21) is lowest. By contrast, the word frequency substitution scheme gets the lowest precision rate of 32.25, but the highest recall rate of 23.80. It seems that sampling words as replacements from the vocabulary according to the word frequency does help the GEC model pick out more sentence errors. However, if too many high frequency words appear in a sentence, the GEC model may be confused, resulting in a decrease of precision. The homophone substitution noising scheme which is specially designed for the Chinese GEC task obtains a relatively unsatisfactory score. Homophone noising generates more realistic wrong sentences, as a large proportion of grammatical errors in Chinese are related to the homophony phenomenon. It's harder for the model to decide whether a word should be corrected. Besides, due to a limitation of word choices, homophone noising can not generate as many noisy sentence pairs as the random noising and word frequency noising. As a consequence, the recall rate of homophone noising model falls behind. Lastly, the random substitution noising scheme keeps a balance between the precision (37.87) and the recall (20.47), and it obtains the second highest $F_{0.5}$ score as a result.

After mixing the five noising schemes (including no substitution) above, our model achieves the best result of 36.97 $F_{0.5}$ score. This mixed noising method increases the precision further and maintains a relatively high recall.

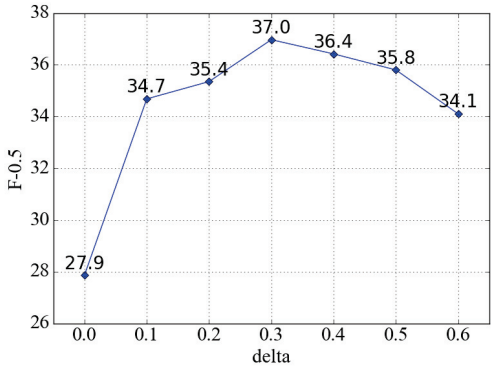

Figure 4: The effect of δ on dynamic masking.

Effect of Threshold δ

We explore the effect of hyper-parameter δ in Equation 6 on dynamic masking. The result is illustrated in Figure 4.

Model	P	R	$F_{0.5}$
Subword-Transformer	34.06	12.05	24.94
Char-Transformer	36.57	14.27	27.86

Table 5: Performance of subword-based and character-based Chinese GEC models on the NLPCC-2018 dataset. The seq2seq architecture is Transformer.

δ ranges from 0 to 1. From Figure 4, we can see that the model achieves the best performance when δ is 0.3. When δ decreases, the diversity of generated noisy source sentences is drastically reduced, making the dynamic masking gradually degrade to normal seq2seq learning. When δ increases, the exponential growth of noises will harm the training procedure. Therefore, the threshold of substitution probability should be carefully selected to strike a balance between generalization ability and robustness.

Effect of Word Segmentation

Because of the characteristics of Chinese, our grammatical error correction model adopts the character-based NMT approach. We also implement a subword-based Transformer model (Subword-Transformer) for comparison.

In the Subword-Transformer model, BPE (byte pair encoding) (Sennrich, Haddow, and Birch 2016) algorithm is applied to segment the rare words into subword units. According to the performance on the development set, the number of BPE operations is set to 8,000. The result is demonstrated in Table 5. We can find that the Char-Transformer model gets an $F_{0.5}$ score of 27.86, outperforming the Subword-Transformer model (24.94 $F_{0.5}$) by 2.92 $F_{0.5}$ score. The reason may be that, by regarding Chinese characters as segmentation units, the vocabulary can be reduced to a suitable size for GEC task. Besides, segmentation errors may lead to a decline of model performance.

Related Work

Early grammatical error correction systems use type-specific classifiers (De Felice and Pulman 2008; Rozovskaya et al. 2014). The emergence of statistical machine translation (SMT) approaches promotes the development of grammatical error correction systems (Behera and Bhattacharyya 2013; Junczys-Dowmunt and Grundkiewicz 2016). Besides, SMT-based systems can be easily joined with manually designed rules (Felice et al. 2014), classifiers (Rozovskaya and Roth 2016) and neural models (Chollampatt, Hoang, and Ng 2016; Chollampatt, Taghipour, and Ng 2016), which helps to improve their performance in GEC tasks.

Recently, many neural machine translation (NMT) approaches have been proposed for GEC tasks. Typical neural GEC approaches use RNN-based seq2seq models (Xie et al. 2016; Yuan and Briscoe 2016; Ji et al. 2017). However, CNN-based NMT models also demonstrate impressive results in grammatical error correction tasks (Schmaltz et al. 2017; Chollampatt and Ng 2018). As a powerful encoder-decoder architecture, Transformer is also introduced into the NMT approaches of GEC task in recent times (Junczys-Dowmunt et al. 2018).

In order to address the problem of data sparsity, several methods are proposed for synthesizing parallel corpus in the tasks of grammatical error correction (Felice and Yuan 2014; Xie et al. 2016). This process is also known as error generation, which creates artificial data as a method of data augmentation. Ge, Wei, and Zhou (2018) propose a novel fluency boost learning and inference mechanism, allowing the model to generate fluency-boost sentence pairs. Xie et al. (2018) propose a noising and denoising scheme to synthesize "realistic" static parallel data for grammatical error correction by back-translation. Experiments show that the artificial data synthesized by their model have the same effect as the additional non-synthesized data. The main difference between our dynamic masking method and the approaches above is that our method serves as a kind of regularization to some extend while training the seq2seq model. We do not synthesize new training data explicitly. Instead, our dynamic masking approach is more like a token-level dropout.

Until the release of NLPCC-2018 dataset, Chinese Grammatical Error Diagnosis (CGED) (Lee, Yu, and Chang 2015; Lee et al. 2016; Fung et al. 2017; Rao et al. 2018) has been a focus in the field of automatic CSL error correction. CGED can be treated as a sequence labeling problem (Zheng et al. 2016; Yang et al. 2017) and solved by using the LSTM-CRF architecture, which combines the traditional method of conditional random fields (CRF) and long short-term memory (LSTM) network to predict the sequence of output labels.

Conclusion

In this paper, we present that dynamic masking methods can promote the normal neural machine translation (NMT) approaches for Chinese grammatical correction (GEC). To address the drawbacks of neural approaches for GEC, we propose five noising schemes to be applied in the training procedure on the source side of the seq2seq model. Our proposed noising schemes are able to generate extremely diverse error-corrected sentence pairs, which improves the performance over normal seq2seq GEC models significantly. Our simple yet effective dynamic masking method of NMT-based models enables our Chinese GEC systems to exceed all published results on the NLPCC-2018 benchmark datasets and establish a new state-of-the-art for the challenging task.

Acknowledgments

We thank all the anonymous reviewers for their constructive feedback. Our work is supported by National Natural Science Foundation of China under Grant No.61433015 and the National Key Research and Development Program of China under Grant No.2017YFB1002101. The corresponding author of this paper is Houfeng Wang.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Behera, B., and Bhattacharyya, P. 2013. Automated grammar correction using hierarchical phrase-based statistical machine translation. In *Proc. IJCNLP*.

Chollampatt, S., and Ng, H. T. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proc. AAAI*.

Chollampatt, S.; Hoang, D. T.; and Ng, H. T. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proc. EMNLP*.

Chollampatt, S.; Taghipour, K.; and Ng, H. T. 2016. Neural network translation models for grammatical error correction. In *Proc. IJCAI*.

Dahlmeier, D., and Ng, H. T. 2012. Better evaluation for grammatical error correction. In *Proc. NAACL-HLT*.

De Felice, R., and Pulman, S. G. 2008. A classifier-based approach to preposition and determiner error correction in l2 english. In *Proc. COLING*.

Felice, M., and Yuan, Z. 2014. Generating artificial errors for grammatical error correction. In *Proc. of the Student Research Workshop at EACL*.

Felice, M.; Yuan, Z.; Andersen, Ø. E.; Yannakoudakis, H.; and Kochmar, E. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proc. CoNLL: Shared Task*.

Fu, K.; Huang, J.; and Duan, Y. 2018. Youdao's winning solution to the nlpcc-2018 task 2 challenge: a neural machine translation approach to chinese grammatical error correction. In *Proc. NLPCC*.

Fung, G.; Debosschere, M.; Wang, D.; Li, B.; Zhu, J.; and Wong, K.-F. 2017. NLPTEA 2017 shared task – Chinese spelling check. In *Proc. NLPTEA*.

Ge, T.; Wei, F.; and Zhou, M. 2018. Fluency boost learning and inference for neural grammatical error correction. In *Proc. ACL*.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proc. AISTATS*.

Ji, J.; Wang, Q.; Toutanova, K.; Gong, Y.; Truong, S.; and Gao, J. 2017. A nested attention neural hybrid model for grammatical error correction. In *Proc. ACL*.

Junczys-Dowmunt, M., and Grundkiewicz, R. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proc. EMNLP*.

Junczys-Dowmunt, M.; Grundkiewicz, R.; Guha, S.; and Heafield, K. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proc. NAACL-HLT*.

Lee, L.-H.; Rao, G.; Yu, L.-C.; XUN, E.; Zhang, B.; and Chang, L.-P. 2016. Overview of NLP-TEA 2016 shared task for Chinese grammatical error diagnosis. In *Proc. NLPTEA*.

Lee, L.-H.; Yu, L.-C.; and Chang, L.-P. 2015. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. In *Proc. NLPTEA*.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*.

Rao, G.; Gong, Q.; Zhang, B.; and Xun, E. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proc. NLPTEA*.

Ren, H.; Yang, L.; and Xun, E. 2018. A sequence to sequence learning for chinese grammatical error correction. In *Proc. NLPCC*.

Rozovskaya, A., and Roth, D. 2016. Grammatical error correction: Machine translation and classifiers. In *Proc. ACL*.

Rozovskaya, A.; Chang, K.-W.; Sammons, M.; Roth, D.; and Habash, N. 2014. The illinois-columbia system in the conll-2014 shared task. In *Proc. CoNLL: Shared Task*.

Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. In *Proc. EMNLP*.

Schmaltz, A.; Kim, Y.; Rush, A.; and Shieber, S. 2017. Adapting sequence models for sentence correction. In *Proc. EMNLP*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. In *Proc. ACL*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. AAAI*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Proc. NeurIPS*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. NeurIPS*.

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

Xie, Z.; Genthial, G.; Xie, S.; Ng, A.; and Jurafsky, D. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proc. NAACL-HLT*.

Yang, Y.; Xie, P.; Tao, J.; Xu, G.; Li, L.; and Si, L. 2017. Alibaba at IJCNLP-2017 task 1: Embedding grammatical features into LSTMs for Chinese grammatical error diagnosis task. In *Proc. IJCNLP 2017, Shared Tasks*.

Yuan, Z., and Briscoe, T. 2016. Grammatical error correction using neural machine translation. In *Proc. NAACL-HLT*.

Zhao, Y.; Jiang, N.; Sun, W.; and Wan, X. 2018. Overview of the nlpcc 2018 shared task: Grammatical error correction. In *Proc. NLPCC*.

Zheng, B.; Che, W.; Guo, J.; and Liu, T. 2016. Chinese grammatical error diagnosis with long short-term memory networks. In *Proc. NLPTEA*.

Zhou, J.; Li, C.; Liu, H.; Bao, Z.; Xu, G.; and Li, L. 2018. Chinese grammatical error correction using statistical and neural models. In *Proc. NLPCC*.