A Hybrid System for Chinese Grammatical Error Diagnosis and Correction

Chen Li* **Junpei Zhou*†** **Zuyi Bao** **Hengyou Liu** **Guangwei Xu** **Linlin Li**
Alibaba Group
969 West Wenyi Road, Hangzhou, China
{puji.lc, zuyi.bzy, hengyou.lhy, linyan.lll}@alibaba-inc.com
jpzhou1996@gmail.com kunka.xgw@taobao.com

Abstract

This paper introduces the DM_NLP team's system for NLPTEA 2018 shared task of Chinese Grammatical Error Diagnosis (CGED), which can be used to detect and correct grammatical errors in texts written by Chinese as a Foreign Language (CFL) learners. This task aims at not only detecting four types of grammatical errors including redundant words (R), missing words (M), bad word selection (S) and disordered words (W), but also recommending corrections for errors of M and S types. We proposed a hybrid system including four models for this task with two stages: the detection stage and the correction stage. In the detection stage, we first used a BiLSTM-CRF model to tag potential errors by sequence labeling, along with some handcraft features. Then we designed three Grammatical Error Correction (GEC) models to generate corrections, which could help to tune the detection result. In the correction stage, candidates were generated by the three GEC models and then merged to output the final corrections for M and S types. Our system reached the highest precision in the correction subtask, which was the most challenging part of this shared task, and got top 3 on F1 scores for position detection of errors.

1 Introduction

More and more people are learning a second or third language as an interest, a career plus, or even a challenge to oneself. Chinese is one of the oldest and most versatile languages in the world. Many people choose to learn Chinese, and the number of CFL leaner grows rapidly.

However, it would be difficult to learn Chinese, because Chinese has a lot of differences from other languages. For example, Chinese has neither the change of singular and plural, nor the tense change of the verb. It has quite flexible expressions and loose structural grammar. These traits bring a lot of trouble to CFL learners, so the demands for Chinese Grammatical Error Diagnosis (CGED) as well as Correction (CGEC) is growing rapidly. GEC for English has been studied for many years, with many shared tasks such as CoNLL-2013 (Ng et al., 2013) and CoNLL-2014 (Ng et al., 2014), while those kinds of studies on Chinese is less yet.

This CGED shared task (Gaoqi et al., 2017; Lee et al., 2016, 2015; Yu et al., 2014) gives researchers an opportunity to build the system and exchange opinions in this field. It could make the community more flourish which benefits all CFL learners. Compared with previous years, this year's NLPTEA CGED shared task requests participants to generate candidate corrections for errors of M and S types. This correction subtask is more challenging and valuable, so we focused on this subtask and got the highest precision in this subtask.

This paper is organized as follows: Section 2 describes some related works in English as well as Chinese. Dataset will be described in Section 3. Section 4 illustrates our hybrid system with two stages, including four models. Section 5 shows the evaluation and discussion of the hybrid model. Section 6 concludes the paper and discusses the future work.

2 Related Work

Earlier attempts to GEC involve rule-based models (Heidorn et al., 1982; Bustamante and León, 1996) and classifier-based approaches (Han et al., 2004; Rozovskaya and Roth, 2011), which can cope with

[*] Equal Contribution

[†] This work was done while the author at Alibaba Group

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 60–69
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

Table 1: Typical examples for four types of errors

Error	Original Sentence	Correct Sentence
M	中国已成了世界拥有最多“烟民”的国家。	中国已成了世界**上**拥有最多“烟民”的国家。
R	孩子的教育不能只靠一个**学校**老师。	孩子的教育不能只靠一个老师。
S	父母对孩子的**爱情**是最重要的。	父母对孩子的**关爱**是最重要的。
W	生产率较低，那**肯定价格**要上升。	生产率较低，那**价格肯定**要上升。

only specific type of errors.

As a sentence may contain multiple errors of different types, a practical GEC system should be able to cope with most of those errors, which is difficult to be achieved by rule-based or classifier models alone. The combination of rule-based and classifier models (Rozovskaya et al., 2013) can correct multiple errors, but it is useful only when the errors are independent of each other, which means that it is unable to solve the problem of dependent errors.

To address more complex errors, MT models are proposed and developed by many researchers. Statistical Machine Translation (SMT) has been dominant for the past two decades. In the work of Brockett et al. (2006), they propose an SMT model used for GEC, and later the round-trip translation is also used in GEC (Madnani et al., 2012). A POS-factored SMT system is proposed (Yuan and Felice, 2013) to correct five types of errors in the text. In the work of Felice et al. (2014), they propose a pipeline of the rule-based system and a phrase-based SMT system augmented by a sizeable web-based language model. The word-level Levenshtein distance between source and target can be used as a translation model feature (Junczys-Dowmunt and Grundkiewicz, 2014) to enhance the model. Rule-based method and n-gram statistical method are combined (Wu et al., 2015) to get a hybrid system for CGED shared task. Recently Napoles and Callison-Bursh (2017) propose a lightweight approach to GEC called Specialized Machine translation for Error Correction.

Nevertheless, Neural Machine Translation (NMT) systems have achieved substantial improvements in this field (Sutskever et al., 2014; Bahdanau et al., 2014). Inspired by this phenomenon, Sun et al. (2015) utilize the Convolutional Neural Network (CNN) for the article error correction. The Recurrent Neural Network (RNN) is also used (Yuan and Briscoe, 2016) to map the sentence from learner space to expert space. Recently Ji et al. (2017) propose a hybrid neural model with nested attention layers for GEC.

3 Dataset Description

The dataset is provided by the 5th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA) 2018 with a Shared Task for CGED. The NLPTEA CGED has been held since 2014, and it provides several sets of training data for this field.

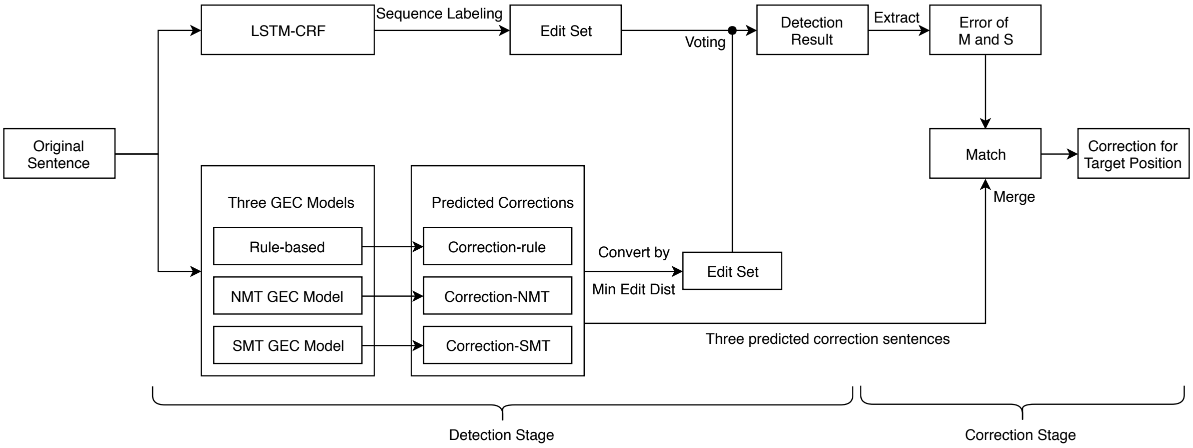
Each instance in the CGED training dataset is composed of an original sentence with a unique sentence number ‘sid’, some ‘target edits’, and a correction sentence. The original sentence contains grammatical errors in Chinese sentences written by CFL learners. All errors are divided into four types, including redundant words (denoted as R), missing words (M), word selection errors (S), and word ordering errors (W). Some typical examples are shown in Table 1.

Each edit in the ‘target edits’ indicates the error type and the position at which it occurs in the original sentence. If an input sentence contains one or more grammatical errors, the ‘target edits’ will include many items, each of which is in the form of [start-off, end-off, error-type], where start-off and end-off respectively denote the starting and ending position of the grammatical error, and the error-type is in the set of R, M, S, and W. For each original sentence given in the test dataset, the developed system should predict the ‘target edits’ in the format which is same as the training set, and for the error type of S and M, the system should predict the candidate corrections.

We also used an external dataset Lang-8[1] to train our GEC models, which contains more than 700,000 items, and each item consists of an original sentence and corresponding corrected sentences. Each original sentence has k correction

[1] provided by NLPCC 2018 GEC shared task

Figure 1: The pipeline of our hybrid system



sentences, where $k \geq 0$.

4 System Description

We proposed a hybrid system for the CGED shared task this year, which contained two stages: the detection stage and the correction stage. In the detection stage, given a sentence s_i, which is composed of characters as $[c_1, c_2, ..., c_n]$, our system generates an edit set E_i which contains one or more errors of this sentence in the form of $[sid, start, end, err]$, where $start$ and end denote a specific part of this sentence $[c_{start}, c_{end}]$ has the error of type err. Then, in the correction stage, for the $err \in \{\text{M}, \text{S}\}$, our system can generate candidate corrections for $[c_{start}, c_{end}]$. If err is M, c_{start} must be equal to c_{end}, and the correction will be inserted at this position. The whole pipeline of our hybrid system is shown in Figure 1.

Our model consists of four models, including the BiLSTM-CRF model for tagging possible errors by sequence labeling at the detection stage, and three GEC models to convert the Chinese sentence from the 'learner space' to the 'expert space'. Those GEC models not only generate candidate corrections for M and S errors at the correction stage, but also help the BiLSTM-CRF model to tag the possible error position at the detection stage. The three GEC models are Rule-based model, NMT model, and SMT model, which are able to cope with different types of grammatical errors.

4.1 BiLSTM-CRF

In the detection stage, we treated the error detection problem as a sequence labeling problem and utilized the BiLSTM-CRF model (Huang et al., 2015) to get the corresponding label sequence in the form of BIO encoding (Kim et al., 2004). More specifically, given an input sentence which is composed of characters as $[c_1, c_2, ..., c_n]$, we utilized this model to predict the label L_i of c_i, for $i \in 1, 2, ..., n$. Since the prior knowledge can be used in this task, we incorporated many additional features for this sequence labeling problem, including Char Bigram, Part-of-speech (POS) tagging, POS score, Adjacent Word Collocation (AWC), Dependent Word Collocation (DWC), as used in (Xie et al., 2017).

4.2 Rule-based Model

The rule-based model starts by segmenting Chinese characters into chunks, which incorporates useful prior grammatical information to identify possible out-of-vocabulary errors. The segments are looked up in the dictionary built by Gigawords (Graff and Chen, 2005), and if a segment is out of vocabulary, it will go through the following steps:

1. If the segment consists of two or more characters, and turn out to be in the dictionary by permuting the characters, it will be added to the candidate list.

2. If the concatenation with a previous or next segment is in the dictionary, it will be added to the candidate list.

3. All possible keys in the dictionary with

the same or similar Pinyin (the Romanization system for Standard Chinese) or similar strokes to the segment are generated. The generated keys for the segment itself, concatenated with those of previous or next segments, will be added to the candidate list of possible corrections.

After the steps, a candidate list of all possible corrections will be processed to identify whether there might be out-of-vocabulary error and it's probability using a language model. The negative log likelihood of a size-5 sliding window suggests whether the top-scored candidate should be a correction of the original segment.

4.3 NMT GEC Model

The NMT model can capture complex relationships between the original sentence and the corrected sentence in GEC. We used the encoder-decoder structure (Bahdanau et al., 2014) with the general attention mechanism (Luong et al., 2015). We used two-layer LSTM model for both encoder and decoder. To enhance the ability of NMT models, we trained four NMT models with different parallel data pairs and configurations as described in Section 5.1. Those four NMT models were denoted as N_j, where $j \in \{1, 2, 3, 4\}$ was the model index. The correction result of sentence s_i generated by N_j was denoted as C_{iN_j}.

We used the character-based NMT because most characters in Chinese has its meaning, which is quite different from English characters, and the Chinese word's meaning often depends on the meaning of its characters. For example, we have two characters 昨天 (yesterday), and we can split it as [yester] + [day]. As in English, the second character 天 means day, and the first one is not a word if taken alone. But it is sufficiently unique to give the whole word its meaning. On the other hand, the errors in original sentences can make the word-based tokenization worse, which will introduce larger and lower quality vocabulary list. So, we chose to use char-based NMT for the CGEC problem.

4.4 SMT GEC Model

The SMT model consists of two components. One is a language model and the other one is a translation model. The language model is learned from a monolingual corpus of the target language, while the parameters of the translation model are calculated from the parallel corpus. We used the noisy channel model (Brown et al., 1993) to combine the language model and the translation model, and incorporated beam search to decode the result.

To explore the ability of SMT models with different configurations, we trained six SMT models with different data granularity and monolingual dataset as described in Section 5.1. Those six SMT models were denoted as S_j, where $j \in \{1, 2, 3, 4, 5, 6\}$ was the model index. The correction result of sentence s_i generated by S_j was denoted as C_{iS_j}.

4.5 Grammatical Error Detection and Correction

For the detection stage, we used the BiLSTM-CRF model as described in Section 4.1 to tag possible errors, by generating labels for each character in sentence s_i. Then each sequence labeling was converted to the editing format $[s_{id}, start, end, err]$. Next, we used the correction results generated by our three different GEC models to help to tune the detection result. For an original sentence s_i, we predicted the corrected sentence C_{iM} with our GEC model M, where M could be NMT N_j or SMT S_j. After getting the predicted correction sentence, we converted it to the editing format $[s_{id}, start, end, err]$, which was consistent with the detection result of the BiLSTM-CRF model.

The conversion from C_{iM} to editing format is based on the minimum editing distance, and we only focused on the error whose type is R, M, or S. On one side, these three types of errors are simple and clear, which can be generated by comparing the s_i and C_{iM} with high confidence. On the other side, the error of type W is more complicated, and the diversity of our GEC model would introduce a great number of noises into the original result on this type of error. Considered that there may exist many kinds of edit trace between a specific pair of s_i and C_{iM}, we kept tracing the edit list which minimized the editing distance between s_i and C_{iM}.

With the edits e_{ij} of sentence s_i, which are generated by BiLSTM-CRF and GEC models, the next step of our system is to ensemble all those edits. When it comes to the ensemble, we tried two methods. One is merging, which combines all detections generated by BiLSTM-CRF model as well as those GEC models, and take the union of their editing sets. The other is voting, in which we

Table 2: Configurations of four NMT models

Model	Network	Embed	Dataset
N_1	LSTM	no	$data_{ed}$
N_2	BiLSTM	enc-dec	$data_{ed}$
N_3	BiLSTM	enc-dec	$data_{all}$
N_4	BiLSTM	dec	$data_{all}$

set a voting threshold $thre$ and accept the edit with $T_{ij} \geq thre$, where T_{ij} is the times of appearance of edit e_{ij} for sentence s_i.

In the correction stage, we used the editing set E_i generated in the detection subtask. For the edit e_{ij} in E_i whose error type is M or S, we selected the candidate characters in the corresponding correction sentence predicted by our GEC models. Finally, all candidates of corrections generated by different GEC models will be collected and merged to create the submission file with detections as well as corrections.

5 Evaluation and Discussion

5.1 Data Split and Experiment Setting

To train the BiLSTM-CRF model, we collected several datasets of CGED, which are 2015, 2016, 2017, and 2018. We split 20% of the 2017 training data as the validation dataset, which is denoted as '17-dev', and all the rest as training. We used the character embeddings and word embeddings pre-trained on the Gigawords and fixed them. For other parameters, we initialized them randomly.

To train our GEC models, we used the external Lang-8 dataset as explained in Section 3. Because each original sentence could have more than one corrected sentences, we used two approaches to generate parallel data pairs to train our GEC models. The first choice is to use only the correct sentence whose edit distance is smallest from the original sentence. The training data generated by the first choice is denoted as $data_{ed}$. The second choice is to use all the correct sentences of the corresponding original sentence. The training data generated by the first choice is denoted as $data_{all}$.

For the NMT model, we used the pre-trained embedding in different parts of the model. The first choice was to use it for the whole model, which forced the model to learn a proper embedding by itself. Considering the dataset is not large enough for the model to learn the embedding from scratch, we also tested the pre-trained embedding used for both encoder and decoder parts. But the embedding was trained on the Gigaword (Graff and Chen, 2005), which was quite different from the sentences written by CFL learners, so we also used the pre-trained embedding only in the decoder part. The configurations of our four different NMT GEC models $N_j, j \in \{1, 2, 3, 4\}$ are shown in Table 2. For the 'Network' column, the 'BiLSTM' means bi-directional LSTM (Schuster and Paliwal, 1997), and for the 'Embed' column, the 'enc-dec' means using pre-trained embedding for both encoder and decoder part in our model.

Table 3: Configurations of six SMT models

Model	Granularity	Corpus	Dataset
S_1	char	Gigawords	$data_{all}$
S_2	char	ChineseWiki	$data_{all}$
S_3	char	CGED+NLPCC	$data_{all}$
S_4	phrase	Gigawords	$data_{all}$
S_5	phrase	ChineseWiki	$data_{all}$
S_6	phrase	CGED+NLPCC	$data_{all}$

For the SMT model, we trained the language model part on different corpora, including the Gigaword, the Chinese Wikipedia corpus (Denoyer and Gallinari, 2006), and the corpus consists of CGED as well as Lang-8 correct sentences which are constructed by ourselves. Besides, we also tested different granularities of the model, which means, used char-level or phrase-level translation model. It is worth to mention that we found that using $data_{all}$ outperformed $data_{ed}$ significantly, so we only did detailed experiments on $data_{all}$ because of the time limitation of the contest. The configurations of our six different SMT models $S_j, j \in \{1, 2, 3, 4, 5, 6\}$ are shown in Table 3

Many excellent tools can emancipate us from the heavy burden of implementing models from scratch. For those NMT GEC models, we implemented it with the *OpenNMT* (Klein et al., 2017) toolkit, and for those SMT GEC models, we implemented the language model with *KenLM* (Heafield, 2011) toolkit and translation model with *Moses* (Koehn et al., 2007).

For the Lang-8 dataset, we found that in those 717,241 lines data, 474,638 lines contained traditional Chinese. The traditional Chinese cannot convey more information than its corresponding simplified Chinese, but will make the size of vocabulary much larger. So, we used the *opencc*

Table 4: Experiments of Grammatical Error Detection on 17-dev dataset by merging eleven models. The corresponding configuration of the models in 'NMT-type' and 'SMT-type' can be found in Table 2 and Table 3. The values for 'Detection', 'Identification', and 'Position' columns are all F_1 values.

NMT-type	SMT-type	FP-rate	Detection	Identification	Position
N_2	S_2	**0.7868**	0.6721	0.3511	0.1846
N_3	S_3	0.8032	**0.6747**	0.3512	0.1853
N_2	S_6	0.8160	0.6719	**0.3566**	0.1834
N_3	S_2	0.8028	0.6746	0.3513	**0.1856**

Table 5: Experiments of Grammatical Error Detection on 17-dev dataset by voting eleven models. The corresponding configuration of the models in 'NMT-type' and 'SMT-type' can be found in Table 2 and Table 3. The values for 'Detection', 'Identification', and 'Position' columns are all F_1 values.

Threshold	NMT-type	SMT-type	FP-rate	Detection	Identification	Position
2	N_4	S_2	0.3336	0.6414	0.4597	**0.2648**
2	N_1	S_6	0.3452	0.6472	**0.4669**	0.2643
2	N_3	S_6	0.3560	**0.6494**	0.4656	0.2643
4	N_4	S_2	**0.1036**	0.4799	0.3435	0.2297

toolkit to convert all the traditional Chinese to simplified Chinese.

5.2 Experiment Result

The evaluation metrics for NLPTEA CGED shared task consists of four subtasks: 'Detection' (determine if the sentence contains errors), 'Identification' (determine the error types), 'Position' (determine the position of errors), and 'Correction' (determine the candidate corrected words for M and S error types). Those four subtasks are from easy to hard, and the last metric is the most valuable, which will be paid more attention by us. The former three metrics are related to the detection stage, and the last metric is related to the correction stage.

Grammatical Error Detection

We used different parameters and initial states of BiLSTM-CRF model to get eight different results on detection stage. Each of three GEC models can generate the result in the editing format as described in Section 4.5. We utilized different methods to ensemble those eleven models, including merging and voting as explained in Section 4.5. Because both NMT and SMT models have different configurations, we tried all combinations of $N_j, j \in \{1, ..., 4\}$ and $S_j, j \in \{1, ..., 6\}$, with the fixed rule-based model, and part of the experiment result with merging is shown in Table 4, while voting method is shown in Table 5.

It's shown in Table 4 and 5 that voting method is more powerful than the merging method on all metrics except for the 'Detection', which is the easiest subtask. We also found out that different combinations of models can cope with different types of errors, and can generate results good at different subtasks. To better utilize the correction generated by our translation model, we preferred the model which performs best on the 'Position' metric, so we chose to use the voting method with threshold 2 to operate on the test dataset with N_2 and S_4.

Grammatical Error Correction

We found that our GEC models can focus on different type of errors, as shown in the Table 6 on the official testing data of CGED 2018, which is denoted as '18-test'. The Table 7 shows some cases in which our different models generated various types of corrections for the original sentence.

As shown in Table 6, the rule-based model can correct those word selection errors which share similar morphology or pronunciation with the ground truth characters. The rule-based model focuses on the correction of word selection errors, so it is able to yield high precision for the error correction problem. The SMT model can handle some errors whose type is R, even that part seems reasonable in the local context. The NMT model is good at correcting many types of errors, including simple errors of word missing or word redun-

Table 6: The cases which can be corrected by our GEC model

Model	Original Sentence	Translation Sentence
Rule	我就会完全知道他的性格，他的爱好，和不好的**密秘**。	我就会完全知道他的性格，他的爱好，和不好的**秘密**。
Rule	学生早恋这问题是很难**结决**的。	学生早恋这问题是很难**解决**的。
Rule	不过我觉得没有个性的文化**是也**没有意义的。	不过我觉得没有个性的文化**也是**没有意义的。
Rule	没有人可以**帮住**我，我是多么的辛苦，多么的劳累啊！	没有人可以**帮助**我，我是多么的辛苦，多么的劳累啊！
NMT	我们**能会做到得**！	我们**能做到**！
NMT	这种措施对个人健康和公众利益有所**好的影响**。	这种措施对个人健康和公众利益有所**好处**。
NMT	这个问题真是个难以解决的。	这个问题真是个难以解决的**问题**。
NMT	这表示你的肺部不是**正常**。	这表示你的肺部不是**正常的**。
NMT	我们从父母学会很多事情	我们从父母**那里**学会很多事情
NMT	我想**也抽烟**不好，但是不能这样对烟民。	我想**抽烟也**不好，但是不能这样对烟民。
NMT	随着社会的变化两代人之间的差异越来越大了。	随着社会的变化**，**两代人之间的差异越来越大了。
NMT	我觉得父母给孩子**的最主要**东西应该是极强的思维方式和美好的内心。	我觉得父母给孩子**最主要的**东西应该是极强的思维方式和美好的内心。
SMT	从小我**也学会**有好的爱清洁的习惯。	从小我**学会**有好的爱清洁的习惯。
SMT	因为化肥和农药，空气污染**了**很严重。	因为化肥和农药，空气污染很严重。
SMT	有些流行歌曲，或是些个体，出的歌曲**的中**带有不文明的话与语言。	有些流行歌曲，或是些个体，出的歌曲**中**带有不文明的话与语言。

Table 7: The same original sentence corrected by different GEC models

Model	Original Sentence	Translation Sentence
Rule	青少年看他们抽烟，引起自己的好奇，后来试抽一次，再抽一次，已经**瘾上**了。	青少年看他们抽烟，引起自己的好奇，后来试抽一次，再抽一次，已经**上瘾**了。
SMT	青少年看他们抽烟，引起自己的好奇，后来试抽一次，再抽一次，已经**瘾**上了。	青少年看他们抽烟，引起自己的好奇，后来试抽一次，再抽一次，已经**迷**上了。
NMT	下面我来**具体的**写一下我的理由。	下面我来**具体地**写一下我的理由。
SMT	下面我来**具体的**写一下我的理由。	下面我来**具体**写一下我的理由。
NMT	我**想**这样的态度是对自己和国家都不好。	我**认为**这样的态度对自己和国家都不好。
SMT	我想这样的态度**是对自己**和国家都不好。	我想这样的态度**对自己**和国家都不好。

Table 8: Ablation Tests of Correction Subtask

Method	Precision	Recall	F_1
$Rule$	0.215	0.00395	0.00775
N_4	0.299	0.0124	0.0238
S_2	**0.348**	0.0178	0.0338
$N_4 + S_2$	0.303	0.0248	0.0459
$Rule + N_4$	0.281	0.0161	0.0304
$Rule + S_2$	0.313	0.0217	0.0406
$Rule + N_4 + S_2$	0.292	**0.0285**	**0.0519**

dancy. It is worth mentioning that the NMT model can correct some more complicated problems including phrase editing and word reordering. For example, it can correct 能会做到得 to 能做到, and also can correct 也抽烟不好 to 抽烟也不好. It can also add punctuations in the middle of the original sentence.

In Table 7, it shows that in some cases, given an original sentence, different GEC models can give different corrections. For the first two rows, the rule-based model and the SMT model give different corrections for the same position of the original sentence, and both of those corrections are reasonable. For the last two rows, the NMT model and the SMT model give corrections at different positions of the original sentence. The ensemble of those models could be helpful because they can generate corrections for many parts of the original sentences, and if they produce different candidates for the same position, we use the voting method to determine the final output.

We explored the ablation test after the release of CGED 2018 ground truth labels. Given error detection results generated by BiLSTM-CRF in the detection stage, we used different combination of three GEC models to generate the candidate corrections for errors of S and M. As we mentioned before, we picked the model combination that performed best on the 'Position' metric in Table 5 to better utilize the candidates generated by our GEC models. It's worth to mention that our rule-based GEC model is not customized for this dataset and the errors made by CFL learners are quite different from native speakers, which leads to relatively low precision. The result of the combination of all three models is slightly better than the version we submitted to CGED shared task because we fixed a small bug in the GEC model. From the ablation study, it showed that the combination of three GEC models improved the F_1 score of Correction Subtask significantly.

6 Conclusion and Future Work

This paper describes our system approach in NLPTEA 2018 shared task of CGED. We proposed a two-stage hybrid system which combined the BiLSTM-CRF model and three GEC models. In the detection stage, we utilized the correction results generated by GEC models to tune the error tags generated by the BiLSTM-CRF model. While in the correction stage, outputs of our GEC models were merged to generate candidate corrections for errors whose type were S or M. Our system achieved the highest precision in the 'Correction' subtask, which is the most challenging part of this shared task and got top 3 on F1 scores for position detection of errors.

In the future, we will further explore the strengths as well as limitations of three GEC models in our system and find a better method to combine them.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chris Brockett, William B Dolan, and Michael Gamon. 2006. Correcting esl errors using phrasal smt techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Flora Ramírez Bustamante and Fernando Sánchez León. 1996. Gramcheck: A grammar and style checker. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 175–181. Association for Computational Linguistics.

Ludovic Denoyer and Patrick Gallinari. 2006. The wikipedia xml corpus. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 12–19. Springer.

Mariano Felice, Zheng Yuan, Øistein E Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24.

RAO Gaoqi, Baolin Zhang, XUN Endong, and Lung-Hao Lee. 2017. Ijcnlp-2017 task 1: Chinese grammatical error diagnosis. *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 1–8.

David Graff and Ke Chen. 2005. Chinese gigaword. *LDC Catalog No.: LDC2003T09, ISBN*, 1:58563–58230.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2004. Detecting errors in english article usage with a maximum entropy classifier trained on a large, diverse corpus. In *LREC*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

George E. Heidorn, Karen Jensen, Lance A. Miller, Roy J. Byrd, and Martin S Chodorow. 1982. The epistle text-critiquing system. *IBM Systems Journal*, 21(3):305–326.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. 2017. A nested attention neural hybrid model for grammatical error correction. *arXiv preprint arXiv:1707.02026*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The amu system in the conll-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Lung-Hao Lee, RAO Gaoqi, Liang-Chih Yu, XUN Endong, Baolin Zhang, and Li-Ping Chang. 2016. Overview of nlp-tea 2016 shared task for Chinese grammatical error diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 40–48.

Lung-Hao Lee, Liang-Chih Yu, and Liping Chang. 2015. Overview of the nlp-tea 2015 shared task for Chinese grammatical error diagnosis.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Exploring grammatical error correction with not-so-crummy machine translation. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 44–53. Association for Computational Linguistics.

Courtney Napoles and Chris Callison-Burch. 2017. Systematically adapting machine translation for grammatical error correction. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 345–356.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

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

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, and Dan Roth. 2013. The university of illinois system in the conll-2013 shared task. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 13–19.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for esl correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 924–933. Association for Computational Linguistics.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Chengjie Sun, Xiaoqiang Jin, Lei Lin, Yuming Zhao, and Xiaolong Wang. 2015. Convolutional neural networks for correcting english article errors. In *Natural Language Processing and Chinese Computing*, pages 102–110. Springer.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Xiupeng Wu, Peijie Huang, Jundong Wang, Qingwen Guo, Yuhong Xu, and Chuping Chen. 2015. Chinese grammatical error diagnosis system based on hybrid model. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 117–125.

Pengjun Xie et al. 2017. Alibaba at ijcnlp-2017 task 1: Embedding grammatical features into lstms for Chinese grammatical error diagnosis task. *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 41–46.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–47.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386.

Zheng Yuan and Mariano Felice. 2013. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 52–61.