Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction

Yuanyuan Zhao[1,2], Nan Jiang[1,2], Weiwei Sun[1,2,3(✉)], and Xiaojun Wan[1,2]

[1] Institute of Computer Science and Technology, Peking University, Beijing, China
zhaoyy1461@gmail.com,jnhsyxxy@126.com,{ws,wanxiaojun}@pku.edu.cn
[2] The MOE Key Laboratory of Computational Linguistics, Peking University, Beijing, China
[3] Center for Chinese Linguistics, Peking University, Beijing, China

Abstract. In this paper, we present an overview of the Grammatical Error Correction task in the NLPCC 2018 shared tasks. We give detailed descriptions of the task definition and the data for training as well as evaluation. We also summarize the approaches investigated by the participants of this task. Such approaches demonstrate the state-of-the-art of Grammatical Error Correction for Mandarin Chinese. The data set and evaluation tool used by this task is available at https://github.com/zhaoyyoo/NLPCC2018_GEC.

1 Introduction

Grammatical Error Correction (GEC) is a challenging task in natural language processing and it has attracted more and more concerns recently. This year, we organize the first shared task of GEC for Mandarin Chinese, with a focus on speech errors produced by Chinese learners. In particular, our task is defined as to detect the grammatical errors in the essays from non-native speakers and return the corrected texts [1]. The previous research on grammatical errors in Chinese is mainly devoted to error detection [2], while our shared task also include automatic correction of such grammatical errors. To the best of our knowledge, this task provides the first benchmark data set for GEC for Chinese.

The goal of the task is to develop techniques to automatically detect and correct errors made by writers of CSL (Chinese as a Second Language). We provide large-scale Chinese texts written by non-native speakers in which grammatical errors have been annotated and corrected by native speakers. Blind test data is used to evaluate the outputs of the participating teams using a common scoring software and evaluation metric.

A total of 23 teams signed up for the shared task and six of them submitted final results. This overview paper provides detailed descriptions of the shared task and it is organized as follows. Section 2 gives the task definition. Section 3 presents a detailed introduction of the data sets and annotation guidelines. Section 4 provides the evaluation metric and Sect. 5 introduces different approaches from participants. Section 6 shows the final results and Sect. 7 gives the conclusion of the paper.

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 439–445, 2018.
https://doi.org/10.1007/978-3-319-99501-4_41

2 Task Definition

Automatically correcting grammatical errors is a challenging task which has attracted an increasing attention recently. The goal of this shared task is to detect and correct grammatical errors present in Chinese essays written by non-native speakers of Mandarin Chinese. Given annotated training data with corrections of grammatical errors and blind test data, the participating teams are expected to submit automatically corrected version of texts in test data. An example of mistaken quantifiers under the task definition is shown in Table 1.

Table 1. An example of the input and the output under the task definition.

Source Input	那是一个牛。
Segmented Input	那 是 一 个 牛 。
Outputs	那是一头牛。
	那 是 一 头 牛 。

3 Data

This section presents the released training and test data in the shared task.

3.1 Training Data

The training data provided in the shared task is collected from http://lang-8.com/, a language-learning website where native speakers freely choose learners' essays to correct. Following [3], we collect a large-scale Chinese Mandarin learners' corpus by exploring "language exchange" social networking services (SNS). There are about 68,500 Chinese Mandarin learners on this SNS website. By collecting their essays written in Chinese and the revised version by Chinese natives, we set up an initial corpus of 1,108,907 sentences from 135,754 essays.

As correcting specifications are not unified and there is lots of noise in raw sentences, we take a series of measures to clean up the data. First, we drop words surrounded by <spanclass = "sline"> since this indicates redundant contents. As for other kinds of tags, correctors use them in different ways. We just remove the tag and remain inner words for consistency and clarity. Learners often ask questions in their native languages, bringing about extra noise into the corpus. We need to get rid of sentences with too many foreign words by checking their Unicode values. There is one more situation where writers use Chinese phonetic alphabet to represent the word that they want to express but do not know how to write it in Chinese characters. Such nonstandard sentences are excluded from final dataset. To improve compactness, we also drop rather simple sentences such as 大家好 *(Hello everyone)*, 晚上好 *(Good night)*. According to our observation, writers sometimes provide optional corrections using "/", "or", " "或 (or)" (or)" or " "或者 (or)" (or)". In such situations, the first correction is

reserved. Moreover, to explain the reason why the original sentence is ungrammatical, correctors may write comments in the position of revised sentences. We utilize a rule-based classifier to determine whether to include the sentence into the corpus.

Through above cleaning operations, we finally sort out a Chinese Mandarin learners' corpus of 717,241 sentences from writers of 61 different native languages. Among these sentences, there are 123,501 sentences considered to be correct, 300,004 sentences with one correction, 170,407 sentences with two corrections and the maximum number of corrections about one sentence is twenty-one. Sample sentences are shown in Table 2. Besides, we use PKUNLP tool (http://www.icst.pku.edu.cn/lcwm/pkunlp/downloads/libgrass-ui.tar.gz) for word segmentation.

Table 2. Sample sentences from the training data.

Source Sentence	Corrected Sentences
长成大人，我盒饭做的很开心。	长大成人后，我做盒饭做得很开心。
城市里的人能度过多方面的生活。	城市里的人能过丰富多彩的生活。
	城市里的人能过多种多样的生活。
	城市里的人能过多方面的生活。

3.2 Test Data

The test data is extracted from *PKU Chinese Learner Corpus*. *PKU Chinese Learner Corpus* is constructed by Department of Chinese Language and Literature, Peking University. The goal is to promote research on international education and Chinese interlanguage. And it is composed of essays written by foreign college students. We collected 2,000 sentences from the corpus and release the source sentences and the segmented version.

To obtain gold edits of grammatical errors, two annotators annotated these sentences. The annotation guidelines follow the general principle of *Minimum Edit Distance*. This principle regulates how to reconstruct a correct form of a given sentence containing errors and it selects the one that minimizes the edit distance from the original sentence [4]. This means that we choose to follow the original intention of the writer as much as possible. Following [2], errors are divided into four types: redundant words (denoted as a capital "R"), missing words ("M"), word selection errors ("S"), and word ordering errors ("W"). The first annotator marked the edit alone, and the second annotator was asked to check the annotation and make a revision if he thought the current edit was not appropriate. We release evaluation results on both the two kinds of gold annotations and their integration.

4 Evaluation Metric

We use the *MaxMatch* (M_2) Scorer for evaluation [5]. M_2 Algorithm is a widely used method for evaluating grammatical error correction. The general idea is

computing the phrase-level edits between the source sentence and the system output. Specifically, it will choose the system hypothesis that holds the highest overlap with the gold edits from annotators. And [1] extends the M_2 Scorer to deal with multiple alternative sets of gold-standard annotations, in which case there are more than one corrections that are reasonable for the current sentence.

Suppose the gold edit set is $\{g_1, g_2, ..., g_n\}$, and the system edit set is $\{e_1, e_2, ..., e_n\}$. The precision, recall and $F_{0.5}$ are defined as follows:

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

$$R = \frac{\sum_{i=1}^{n} |e_i \cap g_i|}{\sum_{i=1}^{n} |g_i|} \tag{2}$$

$$F_{0.5} = 5 \times \frac{P \times R}{P + 4 \times R} \tag{3}$$

where the intersection between e_i and g_i is defined as

$$e_i \cap g_i = \{e \in e_i | \exists g \in g_i (match(e, g))\}. \tag{4}$$

Take the sentence in Fig. 1 as an example, suppose the source sentence is "随着通迅技术的发达我们的生活也是越来越放便。(With the development of communication technology, our life is becoming more and more convenient.)", the set of gold edits **g** and the set of system edits **e** are shown in this figure. Then there will be $P = 1$, $R = 2/3$, $F_{0.5} = 10/11$.

Source Sentence: 随着通迅技术的发达我们的生活也是越来越放便。
Gold Edits (**g**): {通迅 → 通讯, 也 → ϵ, 放便 → 方便}
System Edits (**e**): {通迅 → 通讯, 放便 → 方便}

Fig. 1. An example of the evaluation metric.

5 Approaches

There are altogether 18 submissions from six teams, at most three submissions per team. The detailed information of participants is shown in Table 3.

Most of the systems treat the GEC problem as a machine translation (MT) task. Rule-based models and language models are also explored. *AliGM* [6] proposes two modules for this problem: the correction module and the combination module. In the former module, correction candidates are generated for each input sentence with two statistical models and one neural model. The statistical models include a rule-based model and a statistical machine translation (SMT) -based

Table 3. The detailed information of participants.

System	Organization
AliGM	Alibaba Group
CU-Boulder	Department of Linguistics, University of Colorado Boulder
YouDao	Department of ML & NLP, Youdao
BUPT	Beijing University of Posts and Telecommunications
PKU	Institute of Computational Linguistics, Peking University
BLCU	School of Information Science, Beijing Language and Culture University

model. The neural model refers to a neural machine translation (NMT) -based model. In the latter module, they combine these models in a hierarchical manner. *CU-Boulder* uses a Bi-LSTM model with attention to make corrections. And they use the character-level minimum edit distance (MED) to select the correction version among multiple candidates. Joint voting of five models is implemented to advance the performance. *YouDao* [7] also casts the problem as a machine translation task. It is worth noting that they use a staged approach and design specific modules targeting at particular errors, including spelling, grammatical, etc. *BUPT* uses a two-stage procedure method. In the first stage, they adopt neural models for error detection. In the second stage, they use a statistical method following [8]. *PKU* uses a character-based MT model to deal with this problem. Besides, they propose a preprocessing module for the correction of spelling errors. First, the error detection is based on the binary features including co-occurrence probability, mutual information and chi-square test. Then confusion sets are introduced to generate candidates at the detected point. The final correction is the candidate with the highest language model probability. To improve the precision score, they set a high threshold. In addition, they check each correction with confidence levels in a post-processing stage. *BLCU* [9] proposes a system mainly based on the convolutional sequence-to-sequence model.

6 Results

We perform evaluations on all the eighteen submissions regarding to both of the two kinds of gold annotations and their integration. The best performance of each system referring to the integrated gold standard edits is shown in Table 4.

From Table 4, we can see that grammatical error correction for Chinese language is a challenging task. There still remains large gaps between automatic GEC systems and native speakers. In detail, *YouDao* gets the highest recall and $F_{0.5}$ score while *BLCU* wins the highest precision score. Both of the two systems treat the GEC problem as a MT task. By contrast, the rule-based models and language models perform unsatisfactorily.

Table 4. Evaluation Results

System name	Precision	Recall	$F_{0.5}$
YouDao	35.24	18.64	29.91
AliGM	41.00	13.75	29.36
BLCU	41.73	13.08	29.02
PKU	41.22	7.18	21.16
CU-Boulder	30.07	6.54	17.49
BUPT	4.22	1.49	3.09

7 Conclusion

This paper provides the overview of the Grammatical Error Correction (GEC) shared task in NLPCC 2018. We release a large Chinese learner corpus and briefly introduce participants' methods. The final results show that it is still a challenging task which deserves more concern.

Acknowledgement. This work was supported by National Natural Science Foundation of China (61772036, 61331011) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the Department of Chinese Language and Literature, Peking University for providing the original inputs of the test data. Weiwei Sun is the corresponding author.

References

1. Ng, H.T., Wu, S.M., Wu, Y., Hadiwinoto, C., Tetreault, J.: The CoNLL-2013 shared task on grammatical error correction. In: Proceedings of the 17th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Sofia, pp. 1–12 (2013)
2. Rao, G., Zhang, B., Xun, E., Lee, L.: IJCNLP-2017 Task 1: Chinese grammatical error diagnosis. In: Proceedings of the IJCNLP 2017, Shared Tasks, pp. 1–8. Asian Federation of Natural Language Processing, Taipei (2017)
3. Mizumoto, T., Komachi, M., Nagata, M., Matsumoto, Y.: Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In: Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 147–155. Asian Federation of Natural Language Processing, Chiang Mai (2011)
4. Nagata, R., Sakaguchi, K.: Phrase structure annotation and parsing for learner English. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1837–1847. Association for Computational Linguistics, Berlin (2016)
5. Dahlmeier D, Ng H T.: Better evaluation for grammatical error correction. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 568–572. Association for Computational Linguistics (2012)

6. Zhou, J., Li, C., Liu, H., Bao, Z., Xu, G., Li, L.: Chinese grammatical error correction using statistical and neural models. In: Proceedings of NLPCC-2018 (2018)
7. Fu, K., Huang, J., Duan Y.: Youdao's Winning solution to the NLPCC-2018 Task 2 challenge: a neural machine translation approach to Chinese grammatical error correction. In: Proceedings of NLPCC-2018 (2018)
8. Chen, S., Tsai, Y., Lin, C.: Generating and scoring correction candidates in Chinese grammatical error diagnosis. In: Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, Osaka, pp. 131–139 (2016)
9. Ren, H., Yang, L. Xun, E.: A sequence to sequence learning for Chinese grammatical error correction. In: Proceedings of NLPCC-2018 (2018)