FOCUS IS ALL YOU NEED FOR CHINESE GRAMMATICAL ERROR CORRECTION

Jingheng Ye¹, Yinghui Li¹, Shirong Ma¹, Rui Xie², Wei Wu², Hai-Tao Zheng¹,³*  

¹Shenzhen International Graduate School, Tsinghua University  
²Meituan, ³Peng Cheng Laboratory

ABSTRACT
Chinese Grammatical Error Correction (CGEC) aims to automatically detect and correct grammatical errors contained in Chinese text. In the long term, researchers regard CGEC as a task with a certain degree of uncertainty, that is, an ungrammatical sentence may often have multiple references. However, we argue that even though this is a very reasonable hypothesis, it is too harsh for the intelligence of the mainstream models in this era. In this paper, we first discover that multiple references do not actually bring positive gains to model training. On the contrary, it is beneficial to the CGEC model if the model can pay attention to small but essential data during the training process. Furthermore, we propose a simple yet effective training strategy called ONETARGET to improve the focus ability of the CGEC models and thus improve the CGEC performance. Extensive experiments and detailed analyses demonstrate the correctness of our discovery and the effectiveness of our proposed method.

Index Terms— Natural Language Processing, Chinese Grammatical Error Correction, Multi-Reference, ONETARGET

1. INTRODUCTION

Grammatical Error Correction (GEC) aims to correct all grammatical errors in a given ungrammatical text, with the constraint of keeping the original semantic as possible [1]. The GEC task has attracted more and more attention due to its practical value in daily life, and has been widely used in machine translation [2], spoken language [3], and user-centric application [4]. In the Chinese natural language processing community, Chinese Grammatical Error Correction (CGEC) also plays an important role [5][6].

In the long-term development of CGEC, multi-reference is always an important setting that cannot be ignored because of the uncertainty and subjectivity of the CGEC task [7]. Take Table 1 as an example, for the ungrammatical source sentence, there often exist two or more grammatically correct and semantically unchanged reference sentences. We show that Lang8 [8], a widely used CGEC training dataset, consists of lots of multi-reference samples, i.e., multiple different target sentences for the same ungrammatical source sentence. In addition, the multi-reference setting is also an important aspect of the recently proposed CGEC benchmark called MuCGEC [8]. But is this setting always perfect and necessary?

Dialectically, we argue that the multi-reference setting is reasonable for better evaluation quality, but introduces more uncertainty into the training process of the model. Particularly, in the process of CGEC dataset annotation, it will be difficult to decide which reference is the best among multiple equally acceptable corrections. Therefore, the compatibility brought by the multi-reference setting can guarantee a more realistic evaluation of the model performance [9]. But every coin has two sides, we find that more correction uncertainty exists in multi-reference training samples, which could confuse models during training. We think that the main reason why the model would be confused by multiple references is that the intelligence of the mainstream models in this era is not enough to support them effectively distinguishing different references, while human language learners can easily establish complex connections between these references and promote their writing skills.

Based on the above observations and intuitions, we investigate how multi-reference samples influence the performance of CGEC models. Specifically, training models with multi-reference samples could be treated as a multi-label classification problem (for Seq2Edit-based models) or multi-target generation problem (for Seq2Seq-based models), which dilutes the prediction probability mass of models. Consequently, models will be confused and get into a dilemma without additional mechanisms introduced. In order to remedy this problem, we propose a simple yet effective training strategy called ONETARGET, which is model-agnostic. ONETARGET picks out only one reference based on different strategies for a source sentence with multiple candidate references and keeps the one-reference sample as it is. Though fine-tuning CGEC models with fewer references, ONETARGET gives the models better focus ability and improves their performance, this is the meaning of our title, “focus is all you need for CGEC”.

We identify distinct advantages to training with our proposed ONETARGET: (1) Training models using only one-reference samples is more efficient since cleaned datasets account for 50-60% paired samples of original datasets. (2) Filtering unnecessary references can achieve better performance, and it holds for both the Seq2Edit-based model and the Seq2Seq-based model.

To be summarized, the contributions of our work are three-fold: (1) We first observe and focus on the negative impact of the multi-reference setting for the training of CGEC models. (2) We propose ONETARGET to construct a smaller but more refined dataset, which not only speeds up training but also improves the performance of CGEC models. (3) We conduct extensive experiments and detailed analyses on MuCGEC and achieve state-of-the-art performance.

| Source | I am competent for this position. |
|--------|---------------------------------|
| Ref. 1  | I am competent for this position. |
| Ref. 2  | I am competent for the position. |

Table 1. The example of CGEC. The correction part is marked.

* Corresponding author. (E-mail: zheng.haitao@sz.tsinghua.edu.cn)

1http://tcci.ccf.org.cn/conference/2018/taskdata.php
We refer to “reference” as “target” in the context of training from now on. Recent works \[8, 10\] in the CGEC community usually train models using publicly available datasets, i.e., Lang8 and HSK. Lang8 is a language learning platform, where native speakers voluntarily correct texts written by second-language learners without strict annotation regulation. HSK (Hanyu Shuiping Kaoshi) is an official Chinese proficiency exam, of which samples come from essays written by Chinese second-language learners in the exam. As for the recently proposed evaluation dataset, MuCGEC, a multi-reference multi-source evaluation dataset, consisting of 7,063 sentences collected from three Chinese-as-a-Second-Language learner sources.

In order to figure out whether and how many multi-reference samples exist in public datasets, we report the statistics of Lang8, HSK and MuCGEC-dev datasets in Table 2. One interesting observation is that in Lang8 dataset, unique source sentences only account for 36.95% of the number of samples while HSK consists of almost unique source sentences. Since we are interested in the impact of the number of target sentences on model training, we focus on Lang8 dataset. Further, we report in detail the statistics of Lang8 dataset in Table 3.

High correction uncertainty in multi-target samples. As seen in Table 3, as the number of target sentences increases, the number of unique source sentences decreases, the length of source sentences and the number of edits per target sentence grow while the Levenshtein ratio decreases with the trend. The Levenshtein ratio is defined as follow:

\[
L(S, T) = \frac{\text{sum} - \text{ldist}(S, T)}{\text{sum}},
\]

where sum is the sum of the length of source sentence S and target sentence T, ldist(S, T) is Levenshtein distance between S and T.

Since the Levenshtein ratio is the ratio of the Levenshtein distance and the sum of length, it could be indicative of the correction difficulty of a source sentence. The correction difficulty is a concept that can be understood more exactly as correction uncertainty. Bryant et al. \[4\] have shown that human annotations of GEC task are very subjective and diversified, and we further find that the correction uncertainty may be an important factor for why multi-target samples exist. Annotators are more likely to give different corrections when facing an ungrammatical source sentence with high correction uncertainty. As for an ungrammatical source sentence of lower correction uncertainty, on the opposite, different annotators are more likely to agree with others, and then only one unique target sentence is generated.

Less correction uncertainty in MuCGEC dataset. In order to encourage more diverse and high-quality annotations, the annotation process of the MuCGEC evaluation dataset involved 21 undergraduate native students, among which 5 outstanding annotators were selected as senior annotators to review and determine final golden references \[8\]. Therefore, MuCGEC is constructed by experts and includes less noise while Lang8 is by voluntaries and includes more noise. The Levenshtein ratio of the MuCGEC dataset (0.90 reported in Table 2) is much higher than that of erroneous samples in the Lang8 dataset (0.82 reported in Table 3), showing that a proper correction includes fewer edits. That is, the annotations of the MuCGEC dataset are of less correction uncertainty since annotations are made by experts, inflecting the goal of GEC: to correct all grammatical errors in a given ungrammatical text, with the constraint of keeping the original semantic of the sentence as possible.

### 3. METHODOLOGY

#### 3.1. Multi-target Training

In this section, we analyze what if we train models using multi-target samples. Without the loss of generalization, we take Seq2Edit-based
models as examples, Seq2Edit-based models treat GEC as a sequence tagging problem and decode tag sequences in parallel. Owing to the existence of multi-target samples, GEC is essentially a multi-label classification problem. As shown in Table 1, when coming to the existence of multi-target samples, GEC is essentially a sequence tagging problem and decode tag sequences in parallel. Owing to the existence of multi-target samples, GEC is essentially a sequence tagging problem and decode tag sequences in parallel.

In order to demonstrate explicitly the effectiveness of our proposed ONETARGET strategies on MuCGEC-dev and then evaluate our best models on MuCGEC-test.

### 4.1. Experimental Setup

#### Datasets and Metrics.
We first limit our training data to Lang8 dataset if not mentioned otherwise, since HSK dataset does not consist of many multi-target samples. Following Zhang et al. [8], we filter duplicate sentences in MuCGEC dataset for real performance and also discard correct sentences. We adopt ERRANT-zh proposed by Zhang et al. [8] as our evaluation metric and evaluate models on MuCGEC dataset.

#### Models.
To show that our proposed method is unrelated to the model structure, we train respectively GECToR [16] as Seq2Edit-based models, which achieves the SOTA performance on EGEC datasets, and BART [17] as Seq2Seq-based models. Following Zhang et al., we initialize our Seq2Edit-based models with StructBERT [18] and Seq2Seq-based models with Chinese-BART [19]. Since some common characters are not contained in the vocabulary of Chinese-BART, we extend its vocabulary. We also train GEC models using random one-target dataset and full original dataset for comparison with our data cleaning method introduced in Section 3.2.

#### Hyper-parameters.
We use AdamW optimizer with learning rate $1e-5$, batch size 32, and gradient accumulation step 1. We set the number of training epochs to 10 for all experiments and evaluate models every epoch. The other hyper-parameters are the same with publicly available code [8]. All experiments are conducted for 4 runs and the averaged metrics are reported.

### 4.2. Results on MuCGEC-dev

Table 4 reports the performance on MuCGEC-dev of models trained with different ONETARGET strategies on MuCGEC-dev and then evaluate our best models on MuCGEC-test.

| Strategy     | P  | R  | F₀.₅ | P  | R  | F₀.₅ |
|--------------|----|----|------|----|----|------|
| lev_sim      | 46.14 | 27.30 | 40.55 | 47.20 | 16.60 | 34.48 |
| lev_dis      | 43.72 | 27.96 | 39.29 | 38.38 | 18.34 | 31.50 |
| jac_sim      | 46.25 | 26.71 | 40.35 | 45.75 | 15.63 | 33.02 |
| jac_dis      | 43.68 | 26.33 | 38.59 | 38.77 | 17.73 | 31.34 |
| edิ_least    | 46.33 | 24.69 | 39.42 | 43.90 | 16.30 | 32.80 |
| edิ_most     | 44.04 | 27.18 | 39.18 | 37.25 | 21.21 | 32.25 |
| random       | 45.45 | 25.44 | 39.27 | 41.42 | 17.27 | 32.37 |

**Table 4.** Performance on MuCGEC-dev of Seq2Seq-based and Seq2Edit-based models with different strategies.
| Dataset | #Sample | P   | R   | F₀.₅ | P   | R   | F₀.₅ |
|---------|---------|-----|-----|------|-----|-----|------|
| 1-target| 127,986 | 42.03| 15.67| 31.45|     |     |      |
| 2-target| 255,972 | 37.50| 18.51| 30.92|     |     |      |
| 3-target| 383,958 | 39.22| 16.74| 30.92|     |     |      |

Table 5. Comparison of our single models with related work on MuCGEC-test. We only report part of the results due to the limitation of space. Models in the first block are trained using the full dataset (with 1,557,438 paired samples) and models in the second block are trained using one-target dataset (with 1,056,600 paired samples). The symbol of “*” means the results of our re-implemented baseline method.

Table 6. Performance on MuCGEC-dev of GECToR models trained with 1~3-target datasets respectively. For the sake of simplicity, we do not introduce pre-training on synthetic parallel dataset and multistage fine-tuning as Omelianchuk et al. [16].

4.4. Ablation Study: Influence of the Target Number

In order to further study the impact of the target number on the model performance, we train models using three artificial datasets with a different number of target sentences but the same number of source sentences. Specifically, we pick out source sentences with three or more target sentences randomly in the Lang8 dataset, and a total of 127,986 unique source sentences are obtained. For each source sentence, we construct 1~3-target datasets, such as 2-target dataset consisting of 127,986×2=255,972 paired samples. Note that each of these datasets includes the same number of unique source sentences regardless of n. Due to the limitation of space, we fine-tune three Seq2Edit-based models with 1~3-target datasets respectively and report the results on MuCGEC-dev in Table 5.

It is clear that 1-target training is the best among 1~3-target training settings, which is counter-intuitive since more training samples should have achieved better performance. The main reason is that 1-target training results in higher Precision without a great decrease in the Recall, which caters to the need for grammar error correction that pays more attention to Precision.

These phenomena are interesting and can be understood after considering the features of 1~3-target datasets. On the one hand, though with a different number of samples, these datasets consist of the same number of unique ungrammatical sentences, as a result of which all these models observe the same grammatical errors while multi-target datasets provide multiple equally reasonable corrections for the same ungrammatical sentence. Not like humans, however, models cannot fully understand these multiple corrections, as a result of which these different corrections confuse models instead.

The fact that more samples decrease the performance of models instead inspires us to think about the relationship between grammatical error correction and training datasets. More training samples do not necessarily improve the performance of models, though more unique source sentences can indeed improve the performance of models. For example, Seq2Edit-based models trained with ONE-TARGET-random strategy (581,085 samples) achieve 32.37 F₀.₅ score using 127,986 ONE-TARGET-random training samples shown in Table 5, while they achieve 31.45 F₀.₅ score using 127,986 ONE-TARGET-random training samples shown in Table 5.

5. CONCLUSION

In this paper, we observe that more correction uncertainty exists in multi-reference training samples. Though multi-reference is necessary for evaluation, it is a heavy burden for model training instead in this era. Therefore, we come up with a simple yet effective data training strategy called ONE-TARGET, which reduces the number of training samples significantly with equal or even better performance. We further demonstrate that multi-reference training samples improve the Recall of models at the cost of a great decrease in Precision, which is not recommendable for CGEC. In the future, we will explore establishing multi-reference relationships for models, enabling models to understand exactly the essence of CGEC.
6. REFERENCES

[1] Yu Wang, Yuelin Wang, Kai Dong, Jie Liu, and Zhao Liu, “A comprehensive survey of grammatical error correction,” ACM Trans. Intell. Syst. Technol., vol. 12, no. 5, dec 2021.

[2] Zhichao Li, Kevin Parnow, Masao Utiyama, Eiichiro Sumita, and Hai Zhao, “MiSS: An assistant for multi-style simultaneous translation,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 1–10, Association for Computational Linguistics.

[3] Yiting Lu, Mark Gales, and Yu Wang, “Spoken language grammatical error correction,” in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020.

[4] Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki, “Interpretable language learners using example-based grammatical error correction,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, pp. 7176–7187, Association for Computational Linguistics.

[5] Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, et al., “Linguistic rules-based corpus generation for native chinese grammatical error correction,” arXiv preprint arXiv:2210.10442, 2022.

[6] Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng, “The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking,” in Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22–27, 2022, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, Eds. 2022, pp. 3202–3213, Association for Computational Linguistics.

[7] Christopher Bryant and Hwee Tou Ng, “How far are we from fully automatic high quality grammatical error correction?,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, July 2015, pp. 697–707, Association for Computational Linguistics.

[8] Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang, “MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction,” in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, July 2022, pp. 3118–3130, Association for Computational Linguistics.

[9] Leshem Choshen and Omri Abend, “Inherent biases in reference-based evaluation for grammatical error correction,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 2018, pp. 632–642, Association for Computational Linguistics.

[10] Zewei Zhao and Houfeng Wang, “Maskgec: Improving neural grammatical error correction via dynamic masking,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 01, pp. 1226–1233, Apr. 2020.

[11] Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan, “Overview of the nlpcx 2018 shared task: Grammatical error correction,” in CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2018, pp. 439–445.

[12] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carenli, “Deduplicating training data makes language models better,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, pp. 8424–8445, Association for Computational Linguistics.

[13] Swareep Mishra and Bhavdeep Singh Sachdeva, “Do we need to create big datasets to learn a task?,” in Proceedings of SustainNLP: Workshop on Simple and Efficient Natural Language Processing, Online, Nov. 2020, pp. 169–173, Association for Computational Linguistics.

[14] Christopher Bryant, Mariano Felice, and Ted Briscoe, “Automatic annotation and evaluation of error types for grammatical error correction,” in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, July 2017, pp. 793–805, Association for Computational Linguistics.

[15] Mariano Felice, Christopher Bryant, and Ted Briscoe, “Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments,” in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, Dec. 2016, pp. 825–835, The COLING 2016 Organizing Committee.

[16] Kostiantyn Omelianchuk, Vitaliy Atrashchvych, Artem Chernodub, and Oleksandr Skurzhanskyi, “GECToR – grammatical error correction: Tag, not rewrite?,” in Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA → Online, July 2020, pp. 163–170, Association for Computational Linguistics.

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, pp. 7871–7880, Association for Computational Linguistics.

[18] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si, “Structbert: Incorporating language structures into pre-training for deep language understanding,” 2019.

[19] Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu, “CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation,” CoRR, vol. abs/2109.05729, 2021.