Linguistic Rules-Based Corpus Generation for Native Chinese Grammatical Error Correction

Shirong Ma[1]*, **Yinghui Li**[1]*, **Rongyi Sun**[1], **Qingyu Zhou**[2], **Shuling Huang**[1], **Ding Zhang**[1], **Yangning Li**[1], **Ruiyang Liu**[4], **Zhongli Li**[2], **Yunbo Cao**[2], **Hai-Tao Zheng**[1,3]†, **Ying Shen**[5]†

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Tencent Cloud Xiaowei, [3]Peng Cheng Laboratory
[4]Department of Computer Science and Technology, Tsinghua University
[5]School of Intelligent Systems Engineering, Sun-Yat Sen University
{masr21,liyinghu20}@mails.tsinghua.edu.cn

Abstract

Chinese Grammatical Error Correction (CGEC) is both a challenging NLP task and a common application in human daily life. Recently, many data-driven approaches are proposed for the development of CGEC research. However, there are two major limitations in the CGEC field: First, the lack of high-quality annotated training corpora prevents the performance of existing CGEC models from being significantly improved. Second, the grammatical errors in widely used test sets are not made by native Chinese speakers, resulting in a significant gap between the CGEC models and the real application. In this paper, we propose a linguistic rules-based approach to construct large-scale CGEC training corpora with automatically generated grammatical errors. Additionally, we present a challenging CGEC benchmark derived entirely from errors made by native Chinese speakers in real-world scenarios. Extensive experiments[1] and detailed analyses not only demonstrate that the training data constructed by our method effectively improves the performance of CGEC models, but also reflect that our benchmark is an excellent resource for further development of the CGEC field.

1 Introduction

> In the field of theoretical linguistics, the native speaker is the authority of the grammar.
>
> *- Noam Chomsky*

Chinese Grammatical Error Correction (CGEC) aims to automatically correct grammatical errors that violate language rules and converts the noisy input texts to clean output texts (Wang et al., 2020c). In recent years, CGEC has attracted more and more attention from NLP researchers due to its broader applications in all kinds of daily scenarios and downstream tasks (Duan and Hsu, 2011; Kubis et al., 2019; Omelianchuk et al., 2020).

With the progress of deep learning, data-driven methods based on neural networks, e.g., Transformer (Vaswani et al., 2017), have become the mainstream for CGEC (Zhao and Wang, 2020; Tang et al., 2021; Zhang et al., 2022; Li et al., 2022a). However, we argue that there are still two problems in CGEC: (1) **For model training**, owing to the limited number of real sentences containing grammatical errors, the long-term lack of high-quality annotated training corpora hinders many data-driven models from exercising their capabilities on the CGEC task. (2) **For model evaluation**, the widely used benchmarks such as NLPCC (Zhao et al., 2018) and CGED (Rao et al., 2018, 2020) are all derived from the grammatical errors made by foreign Chinese learners (i.e., L2 learners) in their process of learning Chinese, the gap between the language usage habits of L2 learners and Chinese native speakers makes the performance of the CGEC models in real scenarios unpredictable.

As illustrated in Table 1, the samples of NLPCC and CGED are both from L2 learners, so their sentence structures are relatively short and simple. More crucially, the grammatical errors in these samples are very obvious and naive. On the other hand, in the third example, the erroneous sentence is fluent on the whole, which shows that the grammatical errors made by native speakers are more subtle, that is, they are actually in line with the habit of speaking in people's daily communication, but they do not conform to linguistic norms. *Therefore, in the broader scenarios of Chinese usage besides foreigners learning Chinese, we believe that wrong sentences made by native speakers are more valuable and can better evaluate the model performance than errors made by L2 learners.*

To alleviate the dilemma of missing large-

* indicates equal contribution. Work is done during Yinghui's internship at Tencent Cloud Xiaowei.

† Corresponding author: Hai-Tao Zheng and Ying Shen. (E-mail: zheng.haitao@sz.tsinghua.edu.cn, sheny76@mail.sysu.edu.cn)

[1] Our dataset and source codes are available at https://github.com/masr2000/CLG-CGEC.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 576–589
December 7-11, 2022 ©2022 Association for Computational Linguistics

NLPCC	**Incorrect:** 那个消息给我打冲击。 **Translation:** The news gave me a hit shock. **Correct:** 那个消息给我很大冲击。 **Translation:** The news gave me a big shock. **Source: L2 learners**
CGED	**Incorrect:** 他非常被日本的风景吸引了。 **Translation:** He was very attracted by the Japanese landscape. **Correct:** 他深深地被日本的风景吸引了。 **Translation:** He was deeply attracted by the Japanese landscape. **Source: L2 learners**
Native Error	**Incorrect:** 站在即将到来的2017年的起跑线上，我们不由得情不自禁地感到自豪和喜悦。 **Translation:** Standing at the starting line of upcoming 2017, we cannot have to help feeling proud and joyful. **Correct:** 站在2017年的起跑线上，我们情不自禁地感到自豪和喜悦。 **Translation:** Standing at the starting line of 2017, we cannot help feeling proud and joyful. **Source: Native speakers**

Table 1: Examples of grammatical errors from NLPCC, CGED and Chinese native speakers respectively.

scale training data, we propose CLG, a novel approach based on Chinese linguistic rules that automatically constructs high-quality ungrammatical sentences from grammatical corpus. Specifically, according to authoritative linguistic books (Huang and Liao, 2011; Shao, 2016), we divide Chinese grammatical errors into 6 categories, and design detailed grammatical rules to generate corresponding erroneous sentences according to the characteristics of their respective errors. Our divided 6 error types are: *Structural Confusion*, *Improper Logicality*, *Missing Component*, *Redundant Component*, *Improper Collocation*, and *Improper Word Order*. Different from traditional data augmentation, ungrammatical sentences generated by CLG are more closely matching actual errors that Chinese native speakers would make. Benefiting from our proposed CLG, high-quality and large-scale training samples are automatically constructed with annotated error types. Moreover, **to fill the gap between existing benchmarks and practical applications**, we collect a test dataset containing grammatical errors made by native Chinese speakers in real scenarios, named NaCGEC, which will be a more challenging benchmark and a meaningful resource to facilitate further development of CGEC.

We conduct extensive experiments to demonstrate the effectiveness of CLG and the challenge of NaCGEC. Quantitative experiments show that the model trained on our generated corpus performs better than that trained on traditional CGEC datasets. And compared with general data augmentation methods, the training data obtained by CLG brings larger performance improvements. In addition, qualitative analyses illustrate that it is more difficult for well-educated Chinese native speakers to identify grammatical errors in NaCGEC than in previous existing benchmarks, which indicates that errors in NaCGEC are closer to the real mistakes that native speakers would make in their daily life. We believe that our proposed corpus generation approach and benchmark can greatly contribute to the development of CGEC methods.

2 Related Work

2.1 CGEC Resources

Compared with English Grammatical Error Correction (EGEC), data resources for CGEC are still lacking. The NLPCC (Zhao et al., 2018) provides a test set containing 2K sentences and a large-scale dataset collected from the Lang-8 website for training model. The CGED (Rao et al., 2018, 2020) is an evaluation dataset focusing on error diagnosis which contains 5K sentences from HSK corpus (Cui and Zhang, 2011; Zhang and Cui, 2013). The entire HSK corpus can be also utilized for model training. The YACLC (Wang et al., 2021) collects and annotates 32K sentences from Lang-8 to construct a CGEC dataset. The latest MuCGEC (Zhang et al., 2022) selects and re-annotates sentences from the NLPCC, CGED, and Lang-8 corpora to obtain a multi-reference evaluation dataset with 7K sentences. To the best of our knowledge, no existing resources focus on the grammatical errors made by Chinese native speakers, all of these above-mentioned datasets originate from errors made by L2 learners.

2.2 CGEC Methods

CGEC can be considered as a seq2seq task. Some existing works employ CNN-based (Ren et al.,

2018) or RNN-based (Zhou et al., 2018) models to resolve the CGEC task. Most later work (Wang et al., 2020a; Tang et al., 2021; Zhao and Wang, 2020) employs the Transformer (Vaswani et al., 2017) model which has been a great success in Machine Translation. Those studies also propose some data augmentation approaches to extend the training data for improving the model performance. Recently, researchers start to treat CGEC as a seq2edit task that iteratively predicts the modification label for each position of the sentence. Similar to the GECToR (Omelianchuk et al., 2020), Liang et al. (2020) utilize a seq2edit model for CGEC. Zhang et al. (2022) directly adopt GECToR for CGEC and enhances it by using pretrained language models. TtT (Li and Shi, 2021) proposes a non-autoregressive CGEC approach by employing the BERT (Devlin et al., 2019) encoder with CRF (Lafferty et al., 2001). Li et al. (2022a) proposes a sequence-to-action model to resolve the CGEC task, which combines the advantages of both seq2seq and seq2edit approaches. Unlike previous works, we first focus on the linguistic rules of Chinese grammar and exploit them to automatically obtain high-quality training corpora to improve the performance of CGEC models.

3 Automatic Corpus Generation and Benchmark Construction

3.1 Schema Definition

According to the authoritative linguistic books (Huang and Liao, 2011; Shao, 2016), Chinese grammatical errors are categorized into 7 types: *Structural Confusion*, *Improper Logicality*, *Missing Component*, *Redundant Component*, *Improper Collocation*, *Improper Word Order* and *Ambiguity*. It is worth noting that the errors of ambiguity are often caused by the lack of context information. So if we want the model to correct such errors, we must have enough additional knowledge besides grammar, which is beyond the essence of the CGEC task. Therefore, we do not consider this type of error. In addition, there is a class of common errors, i.e., spelling errors which are mainly caused by various confusing characters with similar strokes/pronunciations (Li et al., 2022b). But Wang et al. have comprehensively studied how to automatically generate large-scale training data containing spelling errors. How to automatically generate high-quality training data is one of our core contributions, so we also don't need to focus on spelling errors in our study.

From the linguistic point of view, the schema of these 6 error types is explained as follows:

(1) **Structural Confusion** (结构混乱) means to mix two or more different syntactic structures in one sentence, which results in confusing sentence structure.

(2) **Improper Logicality** (不合逻辑) represents that the meaning of a sentence is inconsistent or does not conform to objective reasoning.

(3) **Missing Component** (成分残缺) means that the sentence structure is incomplete and some grammatical components are missing.

(4) **Redundant Component** (成分冗余) refers to an addition of unnecessary words or phrases to a well-structured sentence.

(5) **Improper Collocation** (搭配不当) is that the collocation between some components of a sentence does not conform to the structural rules or grammatical conventions of Chinese.

(6) **Improper Word Order** (语序不当) mainly refers to the ungrammatical order of words or clauses in a sentence.

Some example sentences containing various grammatical errors and the corresponding corrections are presented in Table 2. These examples are all selected from our proposed benchmark which will be mentioned in Section 3.4.

Next, we will further introduce the process of our proposed CLG and the details of NaCGEC.

3.2 Correct Sentences Collection

To achieve perfect correct/ungrammatical sentence pairs for model training, we must make sure that the correct sentences are free of any grammatical errors as much as possible. However, collecting and annotating large-scale correct sentences are extremely time-consuming and expensive processes. To address this issue, we first accumulate massive amounts of high-quality Chinese sentences from public datasets (Xu, 2019) such as the Chinese People's Daily corpus, Chinese machine translation dataset, and Chinese wiki corpus as the raw corpora. Then we randomly selected 1,000 sentences from the raw corpora for human judgment by the annotators. The result that more than 97% of the sentences are grammatically correct shows that, in

Structural Confusion	**Incorrect:** 食用水果前应该洗净削皮较为安全。 **Translation:** Fruit should be washed and peeled before eating is safer. **Correct:** 食用水果前应该洗净削皮。 **Translation:** Fruit should be washed and peeled before eating.
Improper Logicality	**Incorrect:** 集团向社会各界人士、沿途村庄百姓表示歉意。 **Translation:** The group apologizes to people from all walks of life and villagers along the way. **Correct:** 集团向社会各界人士表示歉意。 **Translation:** The group apologizes to people from all walks of life.
Missing Component	**Incorrect:** 该节目成功地实现了收视冠军。 **Translation:** The program has successfully achieved a ratings champion. **Correct:** 该节目成功地实现了收视冠军的目标。 **Translation:** The program has successfully achieved the goal of being a ratings champion.
Redundant Component	**Incorrect:** 昨天是转会截止日期的最后一天。 **Translation:** Yesterday was the last day of the transfer deadline. **Correct:** 昨天是转会的最后一天。 **Translation:** Yesterday was the last day of the transfer.
Improper Collocation	**Incorrect:** 丝绸之路开拓了千古传诵的壮美篇章。 **Translation:** The Silk Road has opened a magnificent chapter that has been passed down through the ages. **Correct:** 丝绸之路谱写了千古传诵的壮美篇章。 **Translation:** The Silk Road has written a magnificent chapter that has been passed down through the ages.
Improper Word Order	**Incorrect:** 学校三个月内要求每名学生完成20个小时的义工服务。 **Translation:** The school in three months requires each student to complete 20 hours of volunteer service. **Correct:** 学校要求每名学生三个月内完成20个小时的义工服务。 **Translation:** The school requires each student to complete 20 hours of volunteer service in three months.

Table 2: Examples of sentences with various types of grammatical errors.

a statistical sense, about 97% of the sentences in the raw corpus do not contain grammatical errors. Further, to ensure the quality of our finally selected sentences, we only select sentences with perplexity values in the top 90% (reverse order) as our original correct corpus. Finally, we acquire about 590K correct Chinese sentences from public corpora and guarantee their quality to a certain extent.

3.3 Ungrammatical Sentence Generation

After collecting correct sentences, we use the THULAC toolkit (Sun et al., 2016) for word segmentation and POS tagging, so that the six types of sentence components (subject, predicate, object, attribute, adverbial, complement) can be identified. Then for each category of grammatical errors, we design methods for constructing incorrect sentences from correct sentences according to grammatical rules and language conventions as follows:

(1) For **Structural Confusion**, we collect common correct sentence structures and mix two structures in a sentence to obtain an erroneous sentence that is also smooth and confusing.

(2) For **Improper Logicality**, we collect some patterns with logical errors and match a pattern to a correct sentence to modify it to a wrong one.

(3) For **Missing Component**, we pick a component from sentences to remove, or add words to the sentence to cover up a sentence component.

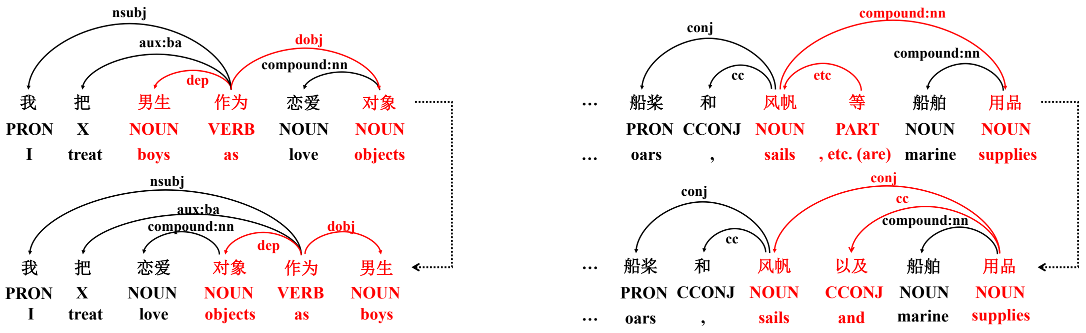
(4) For **Redundant Component**, we select a notional word or meaningful connective in a sentence, and then insert a synonym before or after it to obtain an extra component.

(5) For **Improper Collocation**, we summarize abundant common collocations between sentence components, and we match a collocation that appears in the sentence and randomly replace it with a wrong one.

(6) For **Improper Word Order**, we switch the order of some important sentence components to make the sentence ungrammatical (e.g., reversing the order of related words or the order of an attribute and a head word).

Due to the page limitation, in Figure 1, we only present examples of generating sentences containing two types of errors, “Improper Word Order” and “Improper Logicality”, and the other examples are shown in Appendix A.1.

Based on the above methods, we further design fine-grained grammatical error rules which are presented in Appendix A.2 and develop the


nsubj
aux:ba
dep
dobj
compound:nn
我 把 男生 作为 恋爱 对象
PRON X NOUN VERB NOUN NOUN
I treat boys as love objects
我 把 恋爱 对象 作为 男生
PRON X NOUN NOUN VERB NOUN
I treat love objects as boys
conj
cc
etc
compound:nn
… 船桨 和 风帆 等 船舶 用品
PRON CCONJ NOUN PART NOUN NOUN
… oars , sails , etc. (are) marine supplies
… 船桨 和 风帆 以及 船舶 用品
PRON CCONJ NOUN CCONJ NOUN NOUN
… oars , sails and marine supplies

(a) Improper Word Order

(b) Improper Logicality

Figure 1: Examples of generating ungrammatical sentences containing different types of errors. In the figures, NOUN represents "名词(noun)", VERB represents "动词(verb)", PRON represents "代词(pronoun)", CCONJ represents "连词(conjunction)", X represents "助动词(auxiliary verb)".

corresponding algorithms[2] to change correct sentences into ungrammatical sentences. Through this paradigm of automatically generating erroneous sentences of corresponding grammatical error types based on correct sentences, we obtain not only sentence pairs that can be used for model training, but also the labels of the grammatical error types of the corresponding sentence pairs. Moreover, note that the rules mentioned in the above methods can be applied in combination to construct more complex erroneous instances which have multi grammatical errors in one sentence.

3.4 Benchmark Collection and Annotation

Another core contribution of our work is that we construct the test dataset NaCGEC in which the grammatical errors are all made by native Chinese speakers in real-world scenarios. In this part, we describe how we collect and annotate NaCGEC.

Our benchmark focuses the native Chinese speakers' texts. To ensure that the language characteristics of NaCGEC are consistent with native Chinese speakers, we collect data from the following three real scenarios:

(1) **Entrance examinations for secondary schools and universities**: In the various entrance examinations for Chinese students, there is a special type of question, that is, grammatical error diagnosis questions. These questions provide us with natural and real corpus with grammatical errors. Therefore, we collect the grammatical error questions of the middle school and university entrance examinations and the questions in the supporting teaching materials for the past 10 years for the construction of our test corpus.

(2) **Recruitment examinations for government departments**: Similar to the entrance examinations for students, in China, the exams for civil servants employed by government agencies also include similar grammar questions. Different from the entrance exams, the question text of the civil service recruitment exams is closer to the domain of government official documents, while the questions of the entrance exam are closer to the domain of education. Likewise, we also collect texts of real test questions on civil service examinations and their auxiliary materials over the past 10 years.

(3) **Various Chinese news sites**: To guarantee the data size and domain diversity of our benchmark, we hired more than 20 data collectors and asked them to scour various mainstream Chinese news sites for sentences with grammatical errors. To ensure that they can find ungrammatical sentences more accurately, before they officially start their work, we have systematically taught them grammar knowledge, and asked them to pass the test of grammar error diagnosis questions set by us.

It is worth mentioning that the data collection process lasted for more than 6 months. After obtaining ungrammatical corpus, we employ highly educated university students majoring in Chinese linguistics, who are familiar with Chinese grammatical rules, to check the collected corpus and annotate the corresponding error types according

[2] The implementation of the specific algorithm can refer to the supplementary materials we submitted.

to our defined schema. The details of the annotation process are shown in Appendix A.3. We finally obtain 6,767 ungrammatical sentences and corresponding correction results to compose the NaCGEC benchmark.

4 Data Analyses

4.1 Corpus Statistics

Table 3 reports the data statistics such as the number of sentences, average length, and average edit distance (Levenshtein, 1966) of the corpus we constructed, including the automatically generated training data and the manually acquired test data.

	CLG-Train	NaCGEC-Test
Number of Sentences	591,404	6,767
Erroneous Sentences	591,404	6,496
Number of References	591,404	7,793
Average Length (Char.)	40.31	56.54
Edit Distance (Char.)	2.18	4.19
References / Sentence	1.00	1.20

Table 3: We report the statistics of the training corpus generated by CLG and the test data of NaCGEC.

	Replace	Insert	Delete	Total
Structural Confusion	0.44	0.87	1.55	2.86
Improper Logicality	0.60	0.90	2.06	3.56
Missing Component	0.10	2.10	0.42	2.62
Redundant Component	0.09	0.06	2.00	2.15
Improper Collocation	1.28	0.79	0.79	2.86
Improper Word Order	0.46	5.37	5.42	11.24

Table 4: Average edit distance for sentences containing various types of grammatical errors. Edit operations include Replace, Insert and Delete.

In addition, we perform further analysis on the test samples in NaCGEC. Specifically, we analyze the character-level edit distance required to correct sentences for different types of errors. Note that a sentence with multiple errors will be counted in the corresponding error types repeatedly. From Table 4, we see that the average edit distance for sentences with most types is between 2 and 4, while that for sentences with "Improper Word Order" is as high as 11.24. It indicates that such error requires significant adjustments to sentence structure, which is a challenge for CGEC models.

4.2 Human Evaluation

To verify that our NaCGEC more closely matches the language style of Chinese native speakers and the grammatical errors that native speakers would make, we conduct a human evaluation experiment.

Specifically, we randomly sample 300 correct sentences and 300 incorrect sentences from the test sets of NLPCC, CGED, and NaCGEC respectively. We carefully check these sentences to ensure that the correct and incorrect labels of them are correct from the perspective of native Chinese speakers.

Then we invite 3 Chinese native speakers to do the following two annotation task for these sentences: (1) Annotators should determine whether a sentence contains grammatical errors or not. (2) Annotators should determine whether the language style of a sentence matches that of a native speaker, giving a score of 0, 1, or 2 depending on the degree of match. We compare the results of whether the sentences contain grammatical errors judged by the annotators with the corresponding ground truth, and calculate Precision, Recall, and F_1 of binary classification made by annotators. The results of 3 annotators are averaged and reported in Table 5.

	Pre	Rec	F_1	Score
NLPCC	78.57	82.76	80.61	0.92
CGED	95.00	90.48	92.68	0.85
NaCGEC	72.86	68.00	70.34	1.78

Table 5: Results of human evaluation experiment. Precision, Recall, and F_1 are metrics of annotators distinguishing whether sentences contain grammatical errors, and the Score represents the extent to which annotators judge that sentences conform to the language habits of native speakers (0, 1, 2).

The results demonstrate that: (1) Compared to NLPCC and CGED datasets, the test dataset of NaCGEC obtains a significantly higher language style score, suggesting that the sentences in NaCGEC are more in line with the language habits of native speakers. (2) Annotators who are Chinese native speakers have more difficulty distinguishing whether the sentences from our benchmark contain grammatical errors, indicating that the grammatical errors in our benchmark are more likely to be made by native speakers in their everyday writing.

5 Experiments

5.1 Experimental Setup

We evaluate the performance of two mainstream CGEC methods on our benchmark, including a seq2seq model and a seq2edit model:

		Pre	Rec	$F_{0.5}$
Transformer (Vaswani et al., 2017)	Data Aug.(1000K)	3.50	1.49	2.76
	Lang8(1220K)	8.22	1.04	3.44
	Lang8+HSK(1377K)	5.91	0.79	2.57
	CLG(591K)	17.19	6.20	12.69
	Lang8+CLG(1811K)	26.75	5.89	15.66
GECToR-Chinese (Zhang et al., 2022)	Data Aug.(1000K)	4.35	1.85	3.42
	Lang8(1220K)	20.77	6.97	14.88
	Lang8+HSK(1377K)	22.01	8.73	16.88
	CLG(591K)	23.25	11.03	19.04
	Lang8+CLG(1811K)	27.71	12.19	22.09

Table 6: Experimental results on NaCGEC benchmark. The number in parentheses represents the number of sentences in the corresponding dataset.

Transformer (Vaswani et al., 2017) is a widely used model with the encoder-decoder structure and it is usually utilized to resolve seq2seq tasks. Following previous work (Fu et al., 2018; Wang et al., 2020a; Tang et al., 2021) which treats CGEC as a monolingual translation task and employs the Transformer on the NLPCC, we implement and train a Transformer as our seq2seq model.

GECToR-Chinese (Zhang et al., 2022) is a seq2edit model, which treats CGEC as a sequence labeling task and predicts the modification label, including insertion, deletion, and substitution (Malmi et al., 2019), for each token of the sentence. The model adopts GECToR (Omelianchuk et al., 2020) and enhances it by applying a pretrained language model as the encoder.

Following the NLPCC shared task (Zhao et al., 2018), we employ word-level **MaxMatch (M_2)** Scorer (Dahlmeier and Ng, 2012) for evaluation, which computes Precision, Recall, and $F_{0.5}$ between the gold edit set and the system edit set.

We train the model using the Lang8 (Zhao et al., 2018) dataset which is the official training dataset of NLPCC, HSK (Cui and Zhang, 2011; Zhang and Cui, 2013) dataset, and the training dataset generated by CLG. In addition, we construct a training dataset by data augmentation (Wang et al., 2020a; Tang et al., 2021) for the comparison between our proposed CLG and traditional data augmentation methods. After the training stage, we test the trained models on NaCGEC benchmark. Additionally, the implementation details of our experiments are presented in Appendix A.4.

5.2 Experimental Results

The experimental results shown in Table 6 demonstrate that: (1) Existing models do not perform well on our benchmark, with $F_{0.5}$ being below 25, revealing that NaCGEC benchmark is challenging. (2) For both Transformer and GECToR-Chinese, models trained on CLG training dataset outperform models trained on other datasets, among which Transformer trained on CLG performs far better than that trained on other datasets. It reflects that the training data generated by our proposed CLG effectively assists models to correct grammatical errors made by Chinese native speakers. (3) CLG and Lang8 dataset can be jointly employed to train the model and further improve the performance of the model, which demonstrates that the dataset constructed by CLG is compatible with existing datasets without conflicts.

	Pre	Rec	$F_{0.5}$	$\Delta F_{0.5}$
No Pretrained	34.17	13.41	26.09	-
Data Aug. (1000K)	41.49	14.48	30.21	+4.12
Data Aug. (1600K)	42.30	15.94	31.79	+5.70
CLG (591K)	38.24	16.64	30.36	+4.27
CLG + Data Aug. (1591K)	41.73	17.02	32.34	+6.25

Table 7: Experimental results of Transformer on NLPCC benchmark.

5.3 Analysis of Generalization Ability

Furthermore, to verify the generalizability of the data generated by CLG, we conduct an experiment on the NLPCC dataset. To be specific, before training with the Lang8 dataset, we pretrain the model using CLG training corpus. The compared baseline models include: the Transformer without pretraining and the Transformer pretrained on a corpus constructed by traditional data augmentation approach. Following previous work (Wang et al., 2020a; Tang et al., 2021), we implement the following data augmentation approach. After a correct sentence is segmented into words, the following operations are performed for each word in the sentence according to different probabilities: 70% of no modification, 10% of inserting a random word before this word, 10% of replacing the word with a random word, and 10% of deleting this word.

The results in Table 7 show that: (1) Pretraining the model with CLG improves the model's performance on the NLPCC test dataset, and the improvement is greater than pretraining the model with a larger corpus constructed by traditional data augmentation, which reflects the superiority of our CLG over traditional data augmentation methods on the CGEC task. (2) Compared with directly increasing the size of corpus constructed by data augmentation, our CLG and data augmentation approach can be combined to achieve better results. To summarize, the training data constructed by our proposed CLG brings significant and stable improvements to the model on both NaCGEC, which focuses on native speaker errors, and NLPCC, which focuses on foreign learner errors. This phenomenon proves the good generalization ability of the data constructed by CLG.

5.4 Case Study and Fine-grained Analysis

Incorrect	其对象主要是面向低收入家庭 Its target is mainly for low-income families
Correct	其对象主要是低收入家庭 Its target is mainly low-income families
Model (Lang8)	其对象主要是面向低收入家庭
Model (CLG)	其对象主要是低收入家庭
Type	Structural Confusion
Incorrect	请把这件事你不要放在心上 I hope this matter you do not keep in mind
Correct	请你不要把这件事放在心上 I hope you do not keep this matter in mind
Model (Lang8)	请把这件事放在心上 I hope you keep this matter in mind
Model (CLG)	请不要把这件事你放在心上
Type	Improper Word Order

Table 8: Cases of models trained on different datasets correcting ungrammatical sentences. The red/blue/orange color in the first column represents that the prediction result of the corresponding model is wrong/correct/approximately correct.

	Pre	Rec	$F_{0.5}$
Structural Confusion	37.14	23.53	33.28
Improper Logicality	31.63	19.17	27.99
Missing Component	9.48	4.00	7.44
Redundant Component	27.75	15.90	24.15
Improper Collocation	7.82	3.30	6.13
Improper Word Order	19.82	5.65	13.20

Table 9: Model performance on different types of grammatical errors.

Table 8 illustrates two cases of models trained on different datasets correcting erroneous sentences in NaCGEC. The first sentence contains an error of "Structural Confusion", where "对象...面向(target...for)" is a mixture of two different sentence patterns. The model trained on the Lang8 dataset cannot recognize such an error, while the model trained on CLG perceive the error and correct the sentence properly. The second sentence involves an error of "Improper Word Order", reversing the order of "把这件事(this matter)" and "你不要(you do not)". The error is so difficult for models that the model trained on Lang8 removes "你不要", which makes the corrected sentence grammatical but entirely changes the original sentence meaning. Meanwhile, the model trained on CLG does not correct it completely, but identifies the sentence including an error of "Improper Word Order" and makes a partially appropriate correction. From these two cases, it can be further revealed that: (1) The distribution of grammatical errors in the Lang8 dataset differs significantly from the distribution of grammatical errors that occur in the everyday writing of native speakers. (2) The model trained with CLG can better identify and correct the grammatical errors of native speakers.

Additionally, we also analyze the performance of the model on sentences with fine-grained grammatical error types. The results of Table 9 illustrate that the model performs particularly poorly in dealing with three types of error, i.e., "Missing Component", "Improper Collocation" and "Improper Word Order", which reveals that existing methods have insufficient understanding of common grammatical collocations and insufficient knowledge of macroscopic sentence structure. *We suggest that future research for the CGEC task should be more refined and pay more attention to the connections among each individual component of sentences and the overall structure of sentences.*

6 Conclusion

In this paper, we propose the CLG based on linguistic rules to automatically generate ungrammatical sentences from correct texts for obtaining large-scale CGEC training corpus. Besides, we collect sentences with grammatical errors written by Chinese native speakers and construct a more challenging CGEC benchmark NaCGEC. Experimental results and detailed analyses indicate that NaCGEC is consistent with the language habits of Chinese native speakers, and the training data generated by CLG effectively improves the model performance on NaCGEC and other mainstream benchmarks. We hope our work provides better resources and a new direction for future CGEC research.

7 Limitations

7.1 Limitation of Language

This work focuses on the Chinese Grammatical Error Correction (CGEC) task. The grammatical error classification and ungrammatical sentence generation approach CLG involved in this paper are based on Chinese linguistic rules and cannot be directly extended to other languages such as English. However, we believe that a linguistic perspective can also be introduced in GEC tasks of other languages for more in-depth analysis and exploration.

7.2 Limitation of Experiments

We conduct experiments on CGEC employing two mainstream approaches, including Transformer and GECToR-Chinese models. Limited by hardware resources, the total number of parameters of the two models we construct are around 54M and 108M respectively, and we do not try to build larger models or utilize model ensemble. Future work can construct larger models and use more tricks to improve the performance on our benchmark.

8 Ethical Considerations

In this paper, we propose CLG that automatically generate large-scale and high-quality CGEC training data based on our designed linguistic rules. Additionally, we present a human-annotated benchmark, NaCGEC, which focuses on the grammatical errors made by native Chinese speakers. We describe the details of our proposed corpus generation method and our constructed benchmark in the main text. It is worth noting that in the process of generating training data and labeling test data, all the original corpora used are from publicly available datasets or resources on the legitimate websites, and no sensitive data is involved. Additionally, for CGEC itself, it is a task that has been concerned for a long time and has a wide range of application scenarios. Our work focuses on an important issue that has been overlooked in previous CGEC work, namely grammatical errors made by native Chinese speakers. Therefore, NaCGEC is likely to directly facilitate the research of CGEC, and further enhance the ability of future CGEC models to deal with hard errors made by native speakers.

Acknowledgement

This research is supported by National Natural Science Foundation of China (Grant No.62276154 and 62011540405), Beijing Academy of Artificial Intelligence (BAAI), the Natural Science Foundation of Guangdong Province (Grant No. 2021A1515012640), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (Grant No. HW2021008).

References

Xiliang Cui and Bao-lin Zhang. 2011. The principles for building the "international corpus of learner chinese". *Applied Linguistics*, 2:100–108.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Huizhong Duan and Bo-June Paul Hsu. 2011. Online spelling correction for query completion. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 117–126. ACM.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kai Fu, Jin Huang, and Yitao Duan. 2018. Youdao's winning solution to the NLPCC-2018 task 2 challenge: A neural machine translation approach to chinese grammatical error correction. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part I*, volume 11108 of *Lecture Notes in Computer Science*, pages 341–350. Springer.

Borong Huang and Xudong Liao. 2011. *Modern Chinese (Updated Fifth Edition)*. Higher Education Press, Beijing, China.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Marek Kubis, Zygmunt Vetulani, Mikolaj Wypych, and Tomasz Zietkiewicz. 2019. Open challenge for correcting errors of speech recognition systems. In *Human Language Technology. Challenges for Computer Science and Linguistics - 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17-19, 2019, Revised Selected Papers*, volume 13212 of *Lecture Notes in Computer Science*, pages 322–337. Springer.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, volume 10, page 707.

Jiquan Li, Junliang Guo, Yongxin Zhu, Xin Sheng, Deqiang Jiang, Bo Ren, and Linli Xu. 2022a. Sequence-to-action: Grammatical error correction with action guided sequence generation. *CoRR*, abs/2205.10884.

Piji Li and Shuming Shi. 2021. Tail-to-tail non-autoregressive sequence prediction for Chinese grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4973–4984, Online. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022b. 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*, pages 3202–3213, Dublin, Ireland. Association for Computational Linguistics.

Deng Liang, Chen Zheng, Lei Guo, Xin Cui, Xiuzhang Xiong, Hengqiao Rong, and Jinpeng Dong. 2020. BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 57–66, Suzhou, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51, Melbourne, Australia. Association for Computational Linguistics.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis. In *Proceedings of the*

6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 25–35, Suzhou, China. Association for Computational Linguistics.

Hongkai Ren, Liner Yang, and Endong Xun. 2018. A sequence to sequence learning for chinese grammatical error correction. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, volume 11109 of *Lecture Notes in Computer Science*, pages 401–410. Springer.

Jingmin Shao. 2016. *General Theory of Modern Chinese (Third Edition)*. Shanghai Educational Publishing House, Shanghai, China.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. Thulac: An efficient lexical analyzer for chinese.

Zecheng Tang, Yixin Ji, Yibo Zhao, and Junhui Li. 2021. Chinese grammatical error correction enhanced by data augmentation from word and character levels. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics, Hohhot, China*, pages 13–15.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Chencheng Wang, Liner Yang, Yingying Wang, Yongping Du, and Erhong Yang. 2020a. Chinese grammatical error correction method based on transformer enhanced architecture. *Journal of Chinese Information Processing*, 34(6):106.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2527, Brussels, Belgium. Association for Computational Linguistics.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020b. Structbert: Incorporating language structures into pre-training for deep language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yingying Wang, Cunliang Kong, Liner Yang, Yijun Wang, Xiaorong Lu, Renfen Hu, Shan He, Zhenghao Liu, Yun Chen, Erhong Yang, and Maosong Sun. 2021. YACLC: A chinese learner corpus with multidimensional annotation. *CoRR*, abs/2112.15043.

Yu Wang, Yuelin Wang, Jie Liu, and Zhuo Liu. 2020c. A comprehensive survey of grammar error correction. *CoRR*, abs/2005.06600.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Bao-lin Zhang and Xiliang Cui. 2013. Design concepts of “the construction and research of the interlanguage corpus of chinese from global learners”. *Language Teaching and Linguistic Study*, 5:27–34.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. *CoRR*, abs/2204.10994.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, volume 11109 of *Lecture Notes in Computer Science*, pages 439–445. Springer.

Zewei Zhao and Houfeng Wang. 2020. Maskgec: Improving neural grammatical error correction via dynamic masking. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1226–1233. AAAI Press.

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. 2018. Chinese grammatical error correction using statistical and neural models. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, volume 11109 of *Lecture Notes in Computer Science*, pages 117–128. Springer.

A Appendix

A.1 Other Examples of Ungrammatical Sentence Generation

Figure 2 presents some examples of generating ungrammatical sentences that are not mentioned in the body of this paper. In the figures, NOUN represents "名词(noun)", VERB represents "动词(verb)", PRON represents "代词(pronoun)", CCONJ represents "连词(conjunction)", ADP represents "介词(preposition)", X represents "助动词(auxiliary verb)", PART represents "助词(particle/auxiliary word)", NUM represents "数量词(numeral)".

A.2 Fine-grained Grammatical Error Rules

As described in Section 3.3, we design rules for generating sentences containing corresponding errors for each of the 6 types of grammatical errors. Moreover, each of these 6 types of errors can be divided into several fine-grained error types. As shown in Figure 3, these grammatical errors can be separated into a total of 26 fine-grained types of error. In fact, for each fine-grained type of errors, we consider their characteristics and design the corresponding rules. These fine-grained grammatical error rules are explained as follows:

(1) For **Structural Confusion**, there are 3 fine-grained error types.

- **Mixed Patterns** (句式杂糅) means confusion in sentence structure caused by mixing different expressions of the same meaning together.
- **Mixed Subjects** (中途易辙) means a sentence has two subjects but without any connectives.
- **Mixed Sentences** (两句合一) refers to the inappropriate merging of two sentences, that is, the structure of one sentence is complete, but the last part of the sentence is used as the beginning of the other sentence.

(2) For **Improper Logicality**, there are 5 fine-grained error types.

- **Measure Word** (数词不当) refers to the exact and approximate measure words are used in a sentence simultaneously.
- **Unreasonable** (不合事理) means the sentence doesn't conform to formal or moral logic such as conceptual judgments.
- **Improper Negation** (否定失当) means that the sentence uses multiple negation, making the sentence meaning reversed.
- **Reverse Host-guest** (主客倒置) means that the position of the host and guest described in a sentence are reversed.
- **Imposing Cause and Effect** (强加因果) means that the sentence forces a connection between two events that are not directly causally linked.

(3) For **Missing Component**, there are 4 fine-grained error types.

- **Lack Subject** (主语残缺) means that the sentence does not have a subject.
- **Lack Predicate** (谓语残缺) means that the sentence does not have a predicate.
- **Lack Object** (宾语残缺) means that the sentence does not have an object.
- **Lack Modifier** (修饰成分残缺) refers to sentences that lack necessary modifiers.

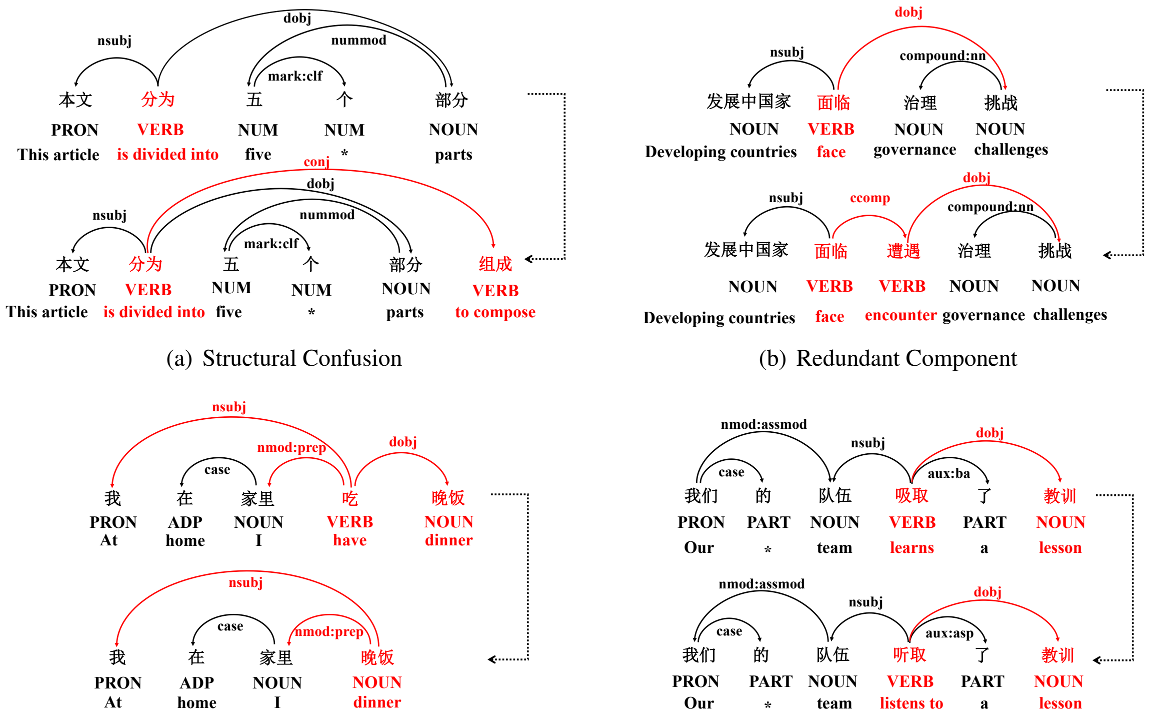
(4) For **Redundant Component**, there are 2 fine-grained error types.

- **Multi Words** (堆砌词语) refers to the use of two or more words in a sentence where one word would have been sufficient, resulting in redundancy.
- **Multi Meanings** (语义重复) means that two or more words of the same meaning are used in a sentence or that the latter word already contains the meaning of the former word.

(5) For **Improper Collocation**, there are 5 fine-grained error types.

- **Subject-predicate** (主谓搭配不当) means the inappropriate pairing of subject and predicate in a sentence.
- **Predicate-object** (动宾搭配不当) and **Subject-object** (主宾搭配不当) are almost the same as **Subject-predicate**.
- **Modifier-head Word** (修饰语中心语搭配不当) means the modifier does not correctly modify the head word in a sentence.
- **Connectives** (关联词语搭配不当) refers to the incorrect collocation of connectives used in a sentence.

(6) For **Improper Word Order**, there are 7 fine-grained error types.


nsubj
dobj
nummod
mark:clf
本文 分为 五 个 部分
PRON VERB NUM NUM NOUN
This article is divided into five * parts
conj
dobj
nsubj
nummod
mark:clf
本文 分为 五 个 部分 组成
PRON VERB NUM NUM NOUN VERB
This article is divided into five * parts to compose
dobj
nsubj
compound:nn
发展中国家 面临 治理 挑战
NOUN VERB NOUN NOUN
Developing countries face governance challenges
dobj
nsubj
ccomp
compound:nn
发展中国家 面临 遭遇 治理 挑战
NOUN VERB VERB NOUN NOUN
Developing countries face encounter governance challenges
nsubj
nmod:prep
case
dobj
我 在 家里 吃 晚饭
PRON ADP NOUN VERB NOUN
At home I have dinner
nsubj
case
nmod:prep
我 在 家里 晚饭
PRON ADP NOUN NOUN
At home I dinner
nmod:assmod
nsubj
dobj
case
aux:ba
我们 的 队伍 吸取 了 教训
PRON PART NOUN VERB PART NOUN
Our * team learns a lesson
nmod:assmod
nsubj
dobj
case
aux:asp
我们 的 队伍 听取 了 教训
PRON PART NOUN VERB PART NOUN
Our * team listens to a lesson

(a) Structural Confusion

(b) Redundant Component

(c) Missing Component

(d) Improper Collocation

Figure 2: Other Examples of generating ungrammatical sentences containing different types of errors.

- **Multi Attributives** (多重定语) means that the multi-attributives' order is wrong.
- **Multi Adverbials** (多重状语) means that the order of multi-adverbials is incorrect.
- **Attributive-head Word** (定语与中心语次序不当) means that the positions of attributive and head word are reversed.
- **Prepositions** (介词短语位置不当) refers to wrong positions of prepositions.
- **Connectives-subject** (关联词位置不当) refers to the improper location of connectives in clauses.
- **Associated Words** (副词位置不当) means that the associated words or phrases are not in the right order.
- **Adverbial-attributives** (定语状语次序不当) means that attributives and adverbs misuse each other.

A.3 Details of NaCGEC Annotation Process

The resources of NaCGEC mainly come from three aspects, namely entrance examinations, recruitment examinations, and various Chinese news sites. Note that the data from entrance/recruitment examination scenarios include specific error types, correct sentences and even analysis of the reasons for the errors, because these data are all from real exam questions. Therefore, our annotation work is mainly carried out on data from news websites. To be specific, our annotation team consists of 5 annotators and 1 senior annotation referee. Our annotation workflow mainly consists of two parts:

(1) To get the effective grammatical error labels, we ask each annotator to give each sample the type of grammatical error he think the sample has. So for each sample, we will get 5 preliminary annotation results. We will select the one with the most occurrences among these five results as the final labeling result for this sample. If there are multiple error types with the same number of occurrences, the senior referee will decide the final error type for this sample based on his profound knowledge.

(2) To get the reasonable reference correct sentences, we require each annotator to rewrite every sentence and give as many rewrites as they feel feasible. Finally, we will select the rewriting results that appear twice or more as the reference standard sentences for the erroneous sentence. If the rewriting results given by the 5 annotators differ from each other for

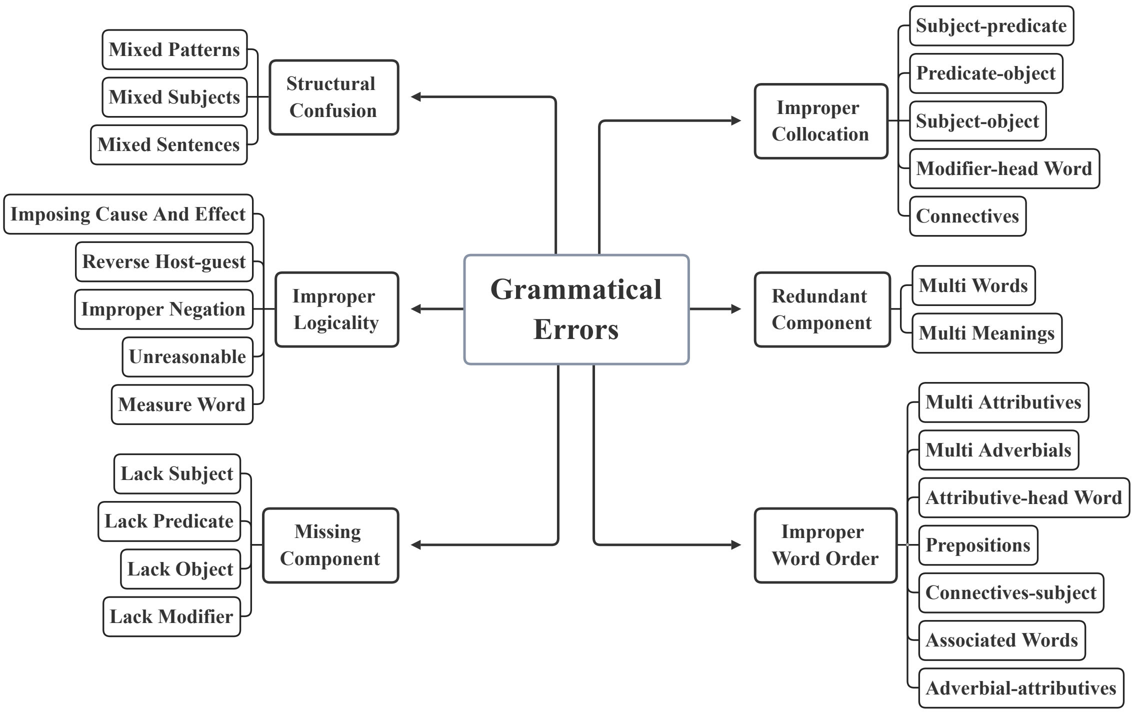


Figure 3: An overview of the fine-grained classification of different types of grammatical errors.

a sample, we will ignore this sample directly after the senior judge has reviewed it.

Additionally, considering that the process of labeling grammatical error types is essentially a multi-label classification process, we use the Fleiss' kappa (Fleiss, 1971) to verify the annotator agreement of labeling grammatical error types. The final Fleiss' kappa score is 0.823, which indicates that our annotation results can be regarded as "almost perfect agreement" (Landis and Koch, 1977).

A.4 Implementation Details

For the Transformer, we implement the code using PyTorch (Paszke et al., 2019). The word embedding matrix shares weight between the source side and the target side, and the embedding dimension is set to 512. Both the encoder and decoder of the Transformer contain 6 layers, and each layer contains 8 attention heads. The maximum length of the input is set to 200, and sentences longer than 200 are truncated. We train the model with the AdamW (Loshchilov and Hutter, 2017) optimizer for 20 epochs. The learning rate and dropout rate are set to 5e-4 and 0.1 respectively. In the inference phase, we use Beam Search as the decoding method, with the number of beams being 8.

For the GECToR-Chinese, we utilize the code provided by the original authors (Zhang et al., 2022). Due to the limitation of hardware resources, we select MacBERT(Base) (Cui et al., 2020) as its encoder, unlike the original code using StructBERT (Wang et al., 2020b). The maximum sentence length is also set to 200. Adam (Kingma and Ba, 2015) is employed as the model optimizer. The training process is divided into two stages: in the first stage the parameters of the encoder are frozen and the model is trained for 2 epochs with a learning rate set to 1e-3, and in the second stage the full parameters are tuned and the model is trained for 20 epochs with a learning rate set to 2e-5.