NaSGEC: a Multi-Domain Chinese Grammatical Error Correction Dataset from Native Speaker Texts

Yue Zhang[1], **Bo Zhang**[2], **Haochen Jiang**[1], **Zhenghua Li**[1*]
Chen Li[2], **Fei Huang**[2], **Min Zhang**[1]
[1]Institute of Artificial Intelligence, School of Computer Science and Technology,
Soochow University, China; [2]DAMO Academy, Alibaba Group, China
[1]{yzhang21,hcj22}@stu.suda.edu.cn, [1]{zhli13,minzhang}@suda.edu.cn
[2]{klayzhang.zb,puji.lc,f.huang}@alibaba-inc.com

Abstract

We introduce NaSGEC, a new dataset to facilitate research on Chinese grammatical error correction (CGEC) for native speaker texts from multiple domains. Previous CGEC research primarily focuses on correcting texts from a single domain, especially learner essays. To broaden the target domain, we annotate multiple references for 12,500 sentences from three native domains, i.e., social media, scientific writing, and examination. We provide solid benchmark results for NaSGEC by employing cutting-edge CGEC models and different training data. We further perform detailed analyses of the connections and gaps between our domains from both empirical and statistical views. We hope this work can inspire future studies on an important but under-explored direction—cross-domain GEC.

1 Introduction

Grammatical error correction (GEC) aims to remove all underlying textual errors in a given sentence without changing its meaning (Bryant et al., 2022). During the past decade, GEC has attracted a lot of research interest and has been integrated into many real-life applications like writing assistants.

A significant effort has been undertaken to build high-quality datasets for research on GEC. Most GEC datasets are for English (Yannakoudakis et al., 2011; Dahlmeier et al., 2013; Napoles et al., 2017; Bryant et al., 2019), which mainly collect sentences from learner essays. For Chinese GEC (CGEC), datasets are relatively scarce. Similar to English GEC, most of them are built from essays written by learners, including NLPCC18 (Zhao et al., 2018), CGED (Rao et al., 2018, 2020), YACLC (Wang et al., 2021), and MuCGEC (Zhang et al., 2022a).

Besides learner GEC, there is also great demand for correcting errors made by native speakers. For English GEC, researchers have already constructed several native datasets, e.g., GMEG (Napoles et al., 2019) and CWEB (Flachs et al., 2020). For CGEC, such research has just begun. CCTC (Wang et al., 2022) is the first native CGEC dataset composed of web documents written by natives. Another recent work, FCGEC (Xu et al., 2022), collects sentences from the questions in Chinese examinations.

	目前现有的汉语依存书库规模较小。
Source	The scale of current existing Chinese dependency library is relatively small.
	~~目前~~现有的汉语依存树库规模较小。
Ref. 1	The scale of ~~current~~ existing Chinese dependency treebank is relatively small.
	目前~~现有~~的汉语依存树库规模较小。
Ref. 2	The scale of current ~~existing~~ Chinese dependency treebank is relatively small.

Table 1: A native CGEC example with two references from the THESIS domain of NaSGEC.

Among all the above datasets, only GMEG (Napoles et al., 2019) targets texts from multiple domains. The lack of multi-domain datasets inevitably introduces biases in the construction and evaluation of CGEC approaches. First, cutting-edge CGEC approaches (Li et al., 2022a; Zhang et al., 2022b; Wu and Wu, 2022) are all evaluated under the in-domain setting, where the training and test sets are from the same domain. It remains unclear how well those approaches generalize to out-of-domain inputs, which is important for practical application. Second, all CGEC approaches are only evaluated in a single domain, basically learner essays. This can be misleading since an approach that outperforms others in one domain may actually perform poorly in another.

To alleviate these problems, this work proposes **NaSGEC** (pronounced as /ˈneɪsgek/), a multi-domain dataset from **na**tive **s**peaker texts for Chinese **GEC**. NaSGEC comprises 12,500 sentences from 3 native text domains: social media platform

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9935–9951
July 9-14, 2023

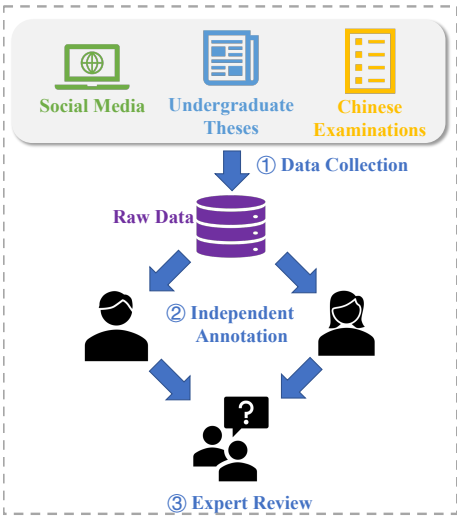


Figure 1: The construction procedure of NaSGEC.

(MEDIA), undergraduate theses (THESIS), and Chinese examinations (EXAM). These domains are closely related to real-life GEC application scenarios, i.e., writing aid, paper proofreading, and Chinese teaching. Based on detailed data analysis (see Section 3), we demonstrate that they have diverse writing styles and error distributions, thus posing great challenges for existing models and will be an ideal testbed for domain adaptation techniques. Furthermore, there are usually different correction methods for an error, as shown in Table 1. Hence, we assign each sentence to two annotators for annotation and one expert for double-checking, leading to multiple high-quality references.

Using NaSGEC, we conduct extensive experiments. We evaluate the performance of the state-of-the-art (SOTA) CGEC model on NaSGEC with different kinds of training data. We first train the model on commonly-used human-annotated training sets. Since these training sets are collected from learner texts while NaSGEC is a native dataset, we also generate synthetic training data from native texts. The multi-domain property of NaSGEC enables us to shed light on the domain problem in CGEC. We conduct domain transfer experiments and design three indicators for evaluating domain differences. In summary, our main contributions can be concluded as follows:

(1) We propose NaSGEC, a multi-domain CGEC dataset from native speaker texts, which contains 12.5k sentences with multiple references. We also conduct detailed data analysis on it.

(2) We launch benchmark experiments on NaSGEC with SOTA CGEC models and different training data. We find models have their own advantages in specific domains, suggesting that the multi-domain NaSGEC can support a more comprehensive evaluation.

(3) Based on NaSGEC, we perform preliminary domain transfer experiments and analysis. We find using small-scale in-domain data for fine-tuning can significantly boost model performance. We also analyze the similarity between domains by comparing cross-domain transfer performance. We devise several indicators of domain shifts to gain more insights. To further improve model performance in a specific domain, we propose a simple domain-aware data augmentation method.

(4) We systematically compare NaSGEC to previously released CGEC datasets, including both learner and native ones.

All codes and models have been released at https://github.com/HillZhang1999/NaSGEC. We will also release the dataset after improving it according to reviewers' comments.

2 Construction of NaSGEC

This section describes the construction process of NaSGEC in detail. As shown in Figure 1, we first collect raw sentences from three domains. Then, each sentence is assigned to two annotators for independent annotation. To guarantee data quality, an expert will carefully review the annotation results.

2.1 Data Collection

NaSGEC collects data from 3 native Chinese text domains, which cover both formal and informal writing styles and errors of different difficulties.

The **MEDIA** domain contains 4k sentences from articles posted on the *Wechat public account platform*[1], which is one of the most popular social media platforms in China. Articles in this platform covers rich topics. We also notice that the sentences in it are mostly informal and often expressed in a spoken-language tone. During our preliminary annotation, we found that errors in this domain are extremely sparse, so direct annotation would result in high costs to acquire enough erroneous sentences. Therefore, we turn to select sentences by voting with multiple competitive CGEC

[1] https://mp.weixin.qq.com/

Dataset	Writer	#Sent.	#Err. Sent. (Perc.)	Avg. Length	Avg. Edits	Avg. Refs	Avg. NEs	Type-Token
NLPCC18 (Zhao et al., 2018)	Learner	2,000	1,983 (99.2%)	29.7	2.0	1.1	0.39	0.43
MuCGEC (Zhang et al., 2022a)	Learner	7,063	6,544 (92.7%)	38.5	3.2	2.3	0.38	0.42
CCTC (Wang et al., 2022)	Native	25,207	2,331 (9.3%)	41.8	1.0	1.0	0.68	0.53
FCGEC (Xu et al., 2022)	Native	41,340	22,517 (54.6%)	53.1	1.5	1.7	1.91	0.49
NaSGEC (MEDIA)	Native	4,000	2,605 (65.2%)	49.0	1.8	1.4	0.79	0.55
NaSGEC (THESIS)	Native	1,500	1,050 (70.0%)	60.5	1.9	1.5	0.67	0.45
NaSGEC (EXAM)	Native	7,000	4,849 (69.3%)	55.9	1.4	1.7	1.00	0.51
NaSGEC	Native	12,500	8,504 (68.0%)	54.3	1.6	1.6	0.89	0.52

Table 2: Dataset statistics, including the writer, the number of sentences (#Sent.), the number and percentage of erroneous sentences (#Err. Sent. (Perc.)), the average length (characters) of sentences (Avg. Length), the average number of edits per reference (Avg. Edits), the average number of references (Avg. Refs), the average number of named entities per sentence (Avg. NEs, extracted by the LTP toolkit (Che et al., 2010)), the average ratio of vocabulary size by the total number of tokens (Type-token, calculated following Flachs et al. (2020)).

models. Specifically, we utilize large-scale pseudo training data to train three seq2seq-based models and three seq2edit-based models. Then, we only choose candidate sentences corrected by more than half of those models for annotation. We crawl 1M candidate sentences from the Wechat public account platform, and accumulate about 120k potentially wrong sentences from them with the above-mentioned method. Finally, we randomly pick 4k sentences for annotation.

The **THESIS** domain consists of 1.5k sentences from *undergraduate theses*. We first collect 120 dissertations written by Chinese undergraduates majoring in computer science, with about 40k sentences in total. Intuitively, texts in this domain are usually formal and contain technical terms. Similar to MEDIA, errors in THESIS are also very sparse. To save costs, we adopt the same method as in MEDIA to select sentences for annotation.

The **EXAM** domain contains 7k sentences from the *ungrammatical sentence judgment questions in Chinese examinations.* Such questions are elaborately designed by experts and ask students to choose 1-3 ungrammatical sentences from 4 candidates. We crawl them from a public educational website[2], as well as their answers and analyses.

2.2 Annotation Workflow

For groundwork, we extend the annotation guidelines of MuCGEC (Zhang et al., 2022a) to accommodate errors made by native speakers. We subsequently use them to instruct our annotators and gradually improve them according to annotator feedback before starting the annotation process. For example, we define how to distinguish dialect from errors after discussing with annotators.

[2] http://www.gzywtk.com/

During annotation, we ask our annotators to directly rewrite the whole sentence to craft a grammatical and fluent version of it with its intended meaning. The so-called *direct rewriting* annotation paradigm has proven efficient and effective in GEC (Sakaguchi et al., 2016; Napoles et al., 2017).

Since multiple acceptable correction ways usually exist, we assign each sentence to two random annotators for independent annotation. Following Zhang et al. (2022a), we ask each annotator to submit the best reference in his/her mind to improve the annotation efficiency. Then, an expert reviewer will check these two submissions in a double-blind manner. Besides directly rejecting incorrect submissions, the reviewer also needs to supplement other correct references missed by annotators. If annotators make wrong submissions, they are required to learn from their mistakes for self-improvement. The learning method is re-typing one of the correct references determined by reviewers. All annotations are conducted with the support of our developed online annotation platform, which is presented in Appendix A. We select and show some typical annotation examples in Appendix F.

2.3 Annotation Process

We hired 13 well-educated native undergraduates familiar with Chinese grammar as our annotators. 2 graduate students, who participated in the compilation of guidelines, served as the reviewers. Annotators received detailed instructions before annotating; those with low annotation quality were warned during annotating. We established a chat group to allow annotators to ask questions. All annotators and reviewers were paid properly. The whole annotation process took more than 4 months.

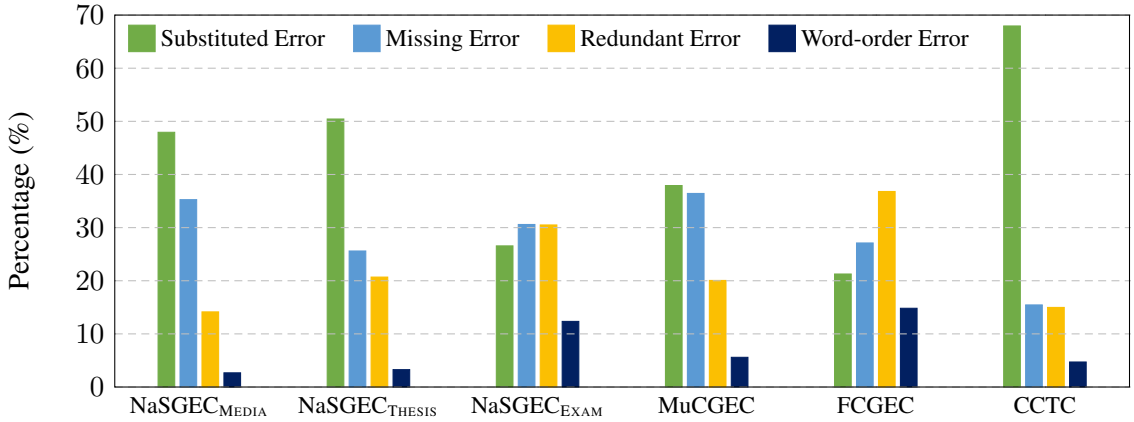


Figure 2: The distributions of 4 kinds of error in 3 domains of NaSGEC and other CGEC datasets.

3 Analysis of NaSGEC

Overall statistics. We list detailed statistics of NaSGEC and other existing datasets for comparison in Table 2. We use the tool[3] released with MuCGEC (Zhang et al., 2022a) to extract the edits of references and original sentences. Such edits are span-level edits merged from character-based ones based on pre-defined linguistic rules.

Within NaSGEC, the average length of sentences varies across domains. The sentences in THESIS are the longest, probably because students tend to write long sentences in dissertations to explain technical concepts more clearly. Regarding the average number of edits and references, we observe that erroneous sentences in EXAM need the fewest edits to correct but have the most correction ways. The reason may be that each erroneous sentence in EXAM typically only has one complicated error to challenge students, which is often varied in its valid corrections. As reflected by the type-token ratio (Richards, 1987), MEDIA has the greatest lexical variety, intuitively due to the diversity of its topics. All the above analysis indicates systematical discrepancies across NaSGEC's domains.

We also present the statistics of two mainstream learner datasets, i.e., NLPCC18 (Zhao et al., 2018) and MuCGEC (Zhang et al., 2022a). Compared with those learner datasets, sentences in NaSGEC are significantly longer but contain much fewer edits, as natives make mistakes far less frequently than learners and seldom make obvious mistakes. Besides, sentences in NaSGEC also have more name entities and a higher level of lexical variety, showing that natives have a larger vocabulary.

[3] https://github.com/HillZhang1999/MuCGEC/tree/main/scorers/ChERRANT

Moreover, we also compare two newly published native datasets, CCTC (Wang et al., 2022) and FCGEC (Xu et al., 2022). The salient feature of CCTC is its low error density. Only 9.3% of sentences in CCTC contain errors, and each erroneous sentence just has one error (reflected by Avg. Edits). As for FCGEC, it is quite similar to the EXAM domain of NaSGEC, which is unsurprising since they share the same provenance.

Error type distributions. We use the tool provided by MuCGEC to classify extracted edits into 4 error types according to their correction operations. Figure 2 shows the distributions of these error types in NaSGEC and other datasets for comparison.

Within NaSGEC, the most frequent error type in MEDIA and THEIS is substituted errors. After further decomposition, we find that the majority of substituted errors in these 2 domains are caused by spelling or misuse of punctuation, as native speakers usually make such minor mistakes due to carelessness when typing essays or papers. The MEDIA domain also has a considerable proportion of missing errors, mainly caused by missing punctuation. Such errors often occur in informal texts, as the absence of punctuation generally does not affect the understanding of the sentence. Compared with the other domains, EXAM has a more even type distribution, where the proportion of substituted, missing, and redundant errors is quite close.

Like MEDIA and THESIS domains of NaSGEC, the learner dataset MuCGEC also has a high proportion of substituted and missing errors. After a deeper look into samples, we find that learners are more prone to misuse verbs or nouns due to lexical or grammatical unfamiliarity, and they also tend to miss more specific words instead of punctuation.

	MEDIA			THESIS			EXAM			Average		
	P	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**
Real Learner	35.96	29.15	34.35	24.16	34.06	25.65	23.01	11.31	19.06	27.71	24.84	27.08
Pseudo Native	**53.39**	29.17	**45.79**	30.86	33.52	31.15	9.78	2.60	6.30	31.34	21.76	**28.80**
Pseudo Native ⇒ Real Learner	38.37	**31.16**	36.67	25.67	**35.09**	27.13	**24.48**	**11.59**	**20.02**	29.51	**25.95**	28.72
Real Learner ⇒ Pseudo Native	51.90	26.20	43.39	**31.61**	31.97	**31.87**	10.77	2.52	6.51	**31.43**	20.23	28.29

Table 3: Benchmark results on NaSGEC. "Pseudo Native ⇒ Real Learner" means that we first train the model on pseudo native data, then on real learner data. The same goes for "Real Learner ⇒ Pseudo Native".

Among all datasets, CCTC has the most unbalanced distribution: the substituted errors account for nearly 70%, and we find most of them are caused by spelling. Although both come from Chinese examinations, FCGEC and NaSGEC-EXAM still have some discrepancies, such as FCGEC contains more redundant errors, which may be due to different annotation guidelines and data sources.

Annotation Accuracy. We measure each annotator's accuracy by comparing all his/her submissions against the golden references determined by reviewers. Overall, the average annotation accuracy is 77.46%. Such a low figure clearly indicates the difficulty of the CGEC task. Moreover, it also highlights the importance of our review mechanism: about a quarter of references in our dataset will be problematic without our strict expert checking.

4 Benchmark Experiments on NaSGEC

This section provides benchmark results for NaSGEC with a current SOTA CGEC model. Following previous work, we train the model on human-annotated training data from learner texts. However, there exists a gap between learner training data and our native dataset. So we also use synthetic native training data to mitigate the gap.

4.1 Experimental Setup

Model. Our benchmark models are based on BART (Lewis et al., 2020), a pre-trained Seq2Seq model that has recently achieved SOTA performance on mainstream CGEC datasets (Zhang et al., 2022b; Wu and Wu, 2022)[4]. We provide the implementation and training details in Appendix B.

Evaluation metric. We use the character-based metric proposed by Zhang et al. (2022a). Concretely, we align the system output and golden reference with the input sentence to extract two groups of character-based edits. Then, we merge them into spans based on rules and compare them to calculate the precision (P), recall (R), and $F_{0.5}$ score. In the GEC community, there is a consensus that a good system should correct errors accurately to ensure a positive user experience. Therefore, most work uses $F_{0.5}$, which places more emphasis on precision by weighting precision twice as recall. We do not use previous word-based metrics since we find they will introduce uncertainty into evaluation due to word segmentation errors.

[4]We also experiment with another competitive CGEC paradigm (Seq2Edit) and report results in Appendix C.

4.2 Training Data

Real learner training data. There are two public available large-scale human-annotated CGEC training datasets, which refer to HSK (Zhang, 2009) and Lang8 (Zhao et al., 2018). Both of them focus on errors occurring in learner essays. Lang8 has about 1.2M sentence pairs, and HSK contains about 150k. We combine them together for training and randomly select 5k of them as the dev set following previous work (Zhang et al., 2022a).

Pseudo native training data. So far, there has been no large-scale training data for errors made by native speakers. As manual annotation is expensive, we create synthetic native training data based on heuristic rules. We first extract 100M clean sentences from the WuDaoCorpora (Yuan et al., 2021), which is mainly composed of articles crawled from native websites. Then, we inject errors into clean sentences by randomly replacing, inserting, deleting and swapping tokens. To better generate spelling errors, we also utilize confusion sets. The proportion of each error is set empirically. More details can be found in Appendix D.

4.3 Experimental Results

Table 3 shows all experimental results. We evaluate models on the whole data of each domain.

In the MEDIA and THESIS domains, the pseudo native training data significantly outperforms the real learner data, although the former is automatically crafted. This shows the text domain of train-

		MEDIA	THESIS	EXAM
Train	#Sent.	2,000	800	4,000
	#Err. Sent.	1,235	757	3,716
	#Ref.	2,568	1,083	5,818
Dev	#Sent.	500	200	1,000
	#Err. Sent.	312	141	723
	#Ref.	895	269	1,464
Test	#Sent.	1,500	500	2,000
	#Err. Sent.	912	313	1,402
	#Ref.	1,926	694	2,900

Table 4: Data split statistics of NaSGEC.

ing data can greatly influence model performance.

In the EXAM domain, the real learner training data instead outperforms the pseudo native data substantially. We speculate the reason is that most errors in the EXAM domain are carefully designed to be difficult, which can hardly be simulated by simple rules but may occur in learner essays.

We also combine both data to make full use of them. We train our model on one kind of data until it converges, then continue to train it on another. As shown in the last two rows of Table 3, the data combinations lead to minor performance improvements in two domains, i.e., THESIS and EXAM.

Finally, the best $F_{0.5}$ scores are 45.79, 31.87, and 20.02 for the MEDIA, THESIS, and EXAM domains, respectively, achieved by 3 different models. It is worth noting that, although all models only have slight differences regarding overall average performance (the largest gap is just 1.72 $F_{0.5}$), they exhibit quite divergent behaviors in different domains (up to 13.72 $F_{0.5}$ gap). This clearly demonstrates the value of NaSGEC as a multi-domain dataset to support a more comprehensive model evaluation.

5 Domain Analysis Within NaSGEC

In this section, we conduct domain transfer experiments on NaSGEC by splitting data and performing fine-tuning. We devise indicators of GEC domain shifts to gain more insights into the connections and differences between our domains. To further improve model performance in specific domains, we also propose a simple domain-aware data augmentation method.

5.1 Domain Transfer Experiments

We perform domain transfer experiments by fine-tuning the baseline on training data from different domains. To facilitate fine-tuning, we split data into training/dev/test sets. The split statistics are listed in Table 4. For each domain, we select the best model in it according to Table 3 as its baseline. After fine-tuning, we evaluate and compare all three fine-tuned models on this domain's test set. All experimental results are presented in Table 5. We also perform error type analysis in Appendix E.

Test → / Train ↓	MEDIA P/R/$F_{0.5}$	THESIS P/R/$F_{0.5}$	EXAM P/R/$F_{0.5}$
Baseline	53.77/28.24/45.54	28.39/33.15/29.23	21.88/9.83/17.57
MEDIA	**61.35/42.72/56.43**	31.96/42.29/33.60	20.85/7.17/15.09
THESIS	52.65/33.40/47.21	**34.96/43.96/36.45**	20.61/8.54/16.07
EXAM	49.16/24.74/41.06	27.93/31.58/28.59	**48.29/24.23/40.29**

Table 5: Results of transfer experiments on NaSGEC.

In-domain results. For in-domain results (fine-tune on one domain and evaluate on the same domain), we have the following observations.

First, the best performance in each domain is achieved by fine-tuning baselines on training data from the same domain, showing that in-domain data benefits more than out-of-domain data. For example, although THESIS-train is much smaller than training sets in other domains, the THESIS-tuned model still performs best on THESIS-test.

Second, fine-tuning models on little in-domain data can bring very significant performance improvements. Specifically, in-domain fine-tuning leads to 10.89, 7.22, and 22.72 $F_{0.5}$ improvements in MEDIA, Thesis, and EXAM, respectively.

Out-of-domain results. For out-of-domain results (fine-tune on one domain and evaluate on another), we have the following observations.

First, in the MEDIA domain, fine-tuning the baseline with THESIS-train can lead to performance gain and vice versa, which indicates that the MEDIA and THESIS domains are relatively similar.

Second, in the EXAM domain, fine-tuning with MEDIA-train and THESIS-train both hurt the performance of the baseline. In turn, fine-tuning with EXAM-train reduces the baseline performance in MEDIA and THESIS. This point to an obvious difference between EXAM and the other 2 domains.

Summary. Overall, fine-tuning models on training data from different domains leads to considerable performance changes, emphasizing the importance of *domain* in GEC. This also encourages us to study domain adaptation for GEC in the future.

5.2 Indicators of Domain Shifts

The domain transfer experiments reveal that there exist appreciable domain shifts in GEC. To better

Target → / Source ↓	MEDIA-test			THESIS-test			EXAM-test		
	VO (%)	TDS	EPO (%)	VO (%)	TDS	EPO (%)	VO (%)	TDS	EPO (%)
MEDIA-train	**65.03**	**0.001**	**25.84**	63.13	0.050	31.75	63.10	0.184	5.07
THESIS-train	56.47	0.025	22.77	**75.73**	**0.009**	**33.05**	65.61	0.161	5.94
EXAM-train	62.97	0.210	6.94	66.33	0.139	10.29	**68.30**	**0.001**	**14.89**

Table 6: Vocabulary Overlap (VO), Type Distribution Similarity (TDS), and Error Pattern Overlap (EPO) between training and test sets from different domains of NaSGEC. Specifically, VO and EPO are averaged over 3 calculations.

understand domain shifts in GEC, we further devise 3 indicators from a statistical perspective:

- **Vocabulary Overlap (VO)** is defined as the ratio of the vocabulary of the target domain covered by the source domain. Higher VO represents better vocabulary coverage. Since larger data usually covers vocabulary better, we sample 1,000 tokens from each domain when calculating VO to make it comparable.
- **Type Distribution Similarity (TDS)** is measured as the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the error type distributions of two domains. The lower TDS indicates closer error type distributions. We extract and classify errors with the tool from MuCGEC (Zhang et al., 2022a).
- **Error Pattern Overlap (EPO)** is computed as the ratio of the error patterns in the target domain occurring in the source domain. We define an error pattern as a mapping from the erroneous span to the corresponding correct span. To eliminate the influence of data sizes, we randomly extract 300 edits from each domain to calculate EPO.

We treat all 3 training sets as the source domains and all 3 test sets as the target domains. Then, we count the above indicators between them, as shown in Table 6. With the help of these indicators, we revisit the results of domain transfer experiments and gain more insights, as shown below.

Explanation for in-domain results. In the previous section, we observe that using in-domain data for fine-tuning consistently outperforms out-of-domain data. Here, we find that the in-domain training sets best cover the vocabulary of the test sets, as reflected by VO. After looking at TDS and EPO, we also find that in-domain training sets have the error distributions most similar to the test sets, in terms of both error types and patterns. These results show that different domains have their own characteristics in word selection and error distribution, which explains why using in-domain data contributes more than our-of-domain data.

	MEDIA			THESIS		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Pretrained Baseline	53.77	28.24	45.54	28.39	33.15	29.23
+ style adaptation	54.31	29.79	46.63	29.09	34.91	30.09
+ error adaptation	54.64	32.04	47.88	29.77	37.79	31.09
+ both	57.29	32.41	**49.66**	31.17	43.17	**33.00**
Finetuned Baseline	61.35	42.72	56.43	34.96	43.96	36.45
+ style adaptation	61.49	43.08	56.65	35.27	44.71	36.83
+ error adaptation	61.72	43.65	57.00	35.12	45.30	36.77
+ both	62.02	43.92	**57.30**	36.01	46.24	**37.68**

Table 7: Results of domain-aware data augmentation.

Explanation for out-of-domain results. Previously, we also observe that the MEDIA and THESIS domains can benefit each other via fine-tuning, while the EXAM domain is unable to help or get help from other domains. From Table 6, we find that TDS/EPO is relatively low/high between MEDIA and THESIS, exhibiting that these two domains have similar error distributions. The reason can be that they are both built from realistic writing scenes, although MEDIA is informal writing while THESIS is formal writing.

As indicated by high TDS and low EPO compared to other domains, EXAM has the most distinct error distribution. The possible reason is that errors in EXAM are deliberately designed to challenge native students and seldom occur in natives' daily writing. Such differences in error distribution can be strong evidence to explain the out-of-domain transfer phenomena.

5.3 Domain-aware Data Augmentation

As previously mentioned, the writing style and error distribution of the training data have a significant impact on the model's performance in a specific domain. Hence, we propose a simple domain-aware data augmentation method by adapting the two aspects of pseudo data to the target domain.

We first perform the *style adaptation*, which means using the raw data with a writing style simi-

Target → / Source ↓	MuCGEC VO (%)	MuCGEC TDS	MuCGEC EPO (%)	CCTC VO (%)	CCTC TDS	CCTC EPO (%)	FCGEC VO (%)	FCGEC TDS	FCGEC EPO (%)
MEDIA	**72.50**	**0.031**	5.79	**64.43**	**0.065**	**42.26**	64.93	0.276	3.98
THESIS	70.20	0.045	6.43	54.67	0.129	40.07	60.43	0.229	5.99
EXAM	70.03	0.078	**7.31**	57.83	0.427	8.47	**68.47**	**0.010**	**13.26**

Table 8: Vocabulary Overlap (VO), Type Distribution Similarity (TDS), and Error Pattern Overlap (EPO) from domains of NaSGEC to existing CGEC datasets. Specifically, VO and EPO are averaged over 3 calculations.

Test → / Train ↓	MuCGEC P/R/$F_{0.5}$	CCTC P/R/$F_{0.5}$	FCGEC P/R/$F_{0.5}$
Baseline	53.84/29.77/**46.34**	19.41/45.99/21.94	33.50/10.93/23.71
MEDIA	52.67/21.88/41.10	20.88/55.40/**23.85**	32.07/5.12/15.62
THESIS	60.61/21.09/44.09	17.98/55.73/20.80	34.10/8.15/20.83
EXAM	57.06/25.41/45.68	16.73/45.34/19.15	50.00/32.32/**45.07**

Table 9: Results of transfer experiments from domains of NaSGEC to existing CGEC datasets.

lar to the target domain for augmentation. For the MEDIA domain, we collect 100k raw sentences from the *Wechat public account platform*. For the THESIS domain, we collect 100k raw sentences from academic papers in the Chinese Scientific Literature (CSL) dataset (Li et al., 2022b). We exclude EXAM since it is difficult to gather sufficient raw data that comes from the same source.

We then conduct the *error adaptation*. We inject 4 kinds of errors (missing, substituted, redundant, and word-order errors) to the raw sentence by rules and carefully control the error type distribution to simulate the target domain.

The experimental results are shown in Table 7. The domain-aware data augmentation (+ both) leads to significant performance gains, even with the in-domain real training data (*Finetuned Baseline*). Only using either *style adaptation* (+ style adaptation, without adjusting error type distribution) or *error adaptation* (+ error adaptation, using 100k data from a general domain, i.e., WuDaoCorpora (Yuan et al., 2021)) still improves performance compared to the baseline, while the improvement is more marginal than simultaneously using both of them. Overall, this is a straightforward attempt, and we hope future work could study more methods for GEC domain adaptation based on NaSGEC.

6 Comparison with Existing Datasets

In this section, we compare NaSGEC with existing CGEC datasets, including both native and learner datasets, by analysis of domain shift indicators (Table 8) and domain transfer experiments (Table 9). Specifically, the baseline in Table 9 is trained with real learner data for MuCGEC and FCGEC, and pseudo native data for CCTC.

NaSGEC vs. Existing learner datasets. Most existing CGEC datasets are for learners. We select MuCGEC (Zhang et al., 2022a) from them for comparison, because it actually covers several previous learner datasets, e.g., NLPCC18 (Zhao et al., 2018) and CGED (Rao et al., 2018, 2020).

From domain shift indicators in Table 8, we have two observations. First, VO is always high from our domains to MuCGEC, implying our data cover the vocabulary of MuCGEC well. This may be because learners tend to use more common words. Second, all our domains get a mediocre level of TDS and EPO, revealing that errors made by native speakers differ from those made by learners. This illustrates why directly fine-tuning models on native data can not further boost performance on learner data.

From domain transfer experiments in Table 9, we can see fine-tuning on domains of NaSGEC always results in performance degradation on MuCGEC, among them EXAM brings the least decline.

We encourage future work to explore better ways to transfer between native and learner domains, which will allow us to apply the rich experience of learner GEC to under-explored native GEC.

NaSGEC vs. Existing native datasets. There are two existing native CGEC datasets, i.e., CCTC (Wang et al., 2022) and FCGEC (Xu et al., 2022).

As shown in Table 8, CCTC is most like the MEDIA domain of NaSGEC, possibly because they are both collected from natives' informal writing. EPO from MEDIA and THESIS to CCTC is higher than 40%, even exceeding their in-domain overlap ratios. As mentioned in Section 3, CCTC has a very high proportion of spelling errors. Spelling errors in Chinese, such as misusing "的/地/得", have fixed patterns and thus can be easily covered. In contrast, our data contains more long-tail and challenging grammatical errors.

Looking at transfer experiments, the recall of the baseline in CCTC greatly increased when fine-

tuned on MEDIA and THESIS, but the precision keeps low. After carefully examining, we think this is due to the difference in error density. As shown in Table 2, about 65.2% and 70.0% of sentences in MEDIA and THESIS have errors, while the number in CCTC is just 9.3%. Therefore, fine-tuning the baseline on our data will make it correct errors more aggressively, which causes poor precision in low error-density domains. In view of this, we hope future work can study how to transfer GEC models across domains with different error densities.

For FCGEC, fine-tuning the model on the EXAM domain of NaSGEC leads to a huge improvement of over 22 $F_{0.5}$ scores, indicating they are highly compatible. The indicator results also confirm this point. We hope they can be two complementary resources to facilitate CGEC for Chinese teaching.

7 Related Work

Dataset. Most GEC datasets are built for English. Early English GEC datasets, such as FCE (Yannakoudakis et al., 2011), NUCLE (Dahlmeier et al., 2013), and JFLEG (Napoles et al., 2017), are built from student essays written by non-native English learners. After realizing the flaw of the limited text domain, researchers propose GMEG (Napoles et al., 2019) and CWEB (Flachs et al., 2020), two new datasets that broaden the target domain of English GEC to native speakers' daily writing.

Early CGEC work also primarily constructs datasets from learner essays, including NLPCC18 (Zhao et al., 2018), CGED (Rao et al., 2018, 2020), YACLC (Wang et al., 2021), and MuCGEC (Zhang et al., 2022a). Concurrently with our work, some newly released CGEC datasets take native writing domains into account. CCTC (Wang et al., 2022) annotates 1,500 web documents written by native speakers from the WuDaoCorpora (Yuan et al., 2021). FCGEC (Xu et al., 2022) mainly consists of sentences from multi-choice questions in Chinese examinations. Another work, NaCGEC (Ma et al., 2022), collects data from Chinese examinations and news sites.

To the best of our knowledge, NaSGEC is the first CGEC dataset that annotates texts from multiple native domains under a unified scheme, which enables us to perform domain-wise experiments and analysis in CGEC for the first time.

Domain Adaptation. Domain adaptation has been extensively studied in various NLP tasks (Ramponi and Plank, 2020), such as machine translation (Chu and Wang, 2018; Jiang et al., 2020; Pham et al., 2021), syntax parsing (Li et al., 2019; Yang et al., 2022), and information extraction (Chen and Qian, 2021; Lekhtman et al., 2021).

Compared with other fields, research on domain adaptation for GEC is under-explored. Existing studies lie in adapting GEC models to a specific first language or proficiency level of the second language learners (Chollampatt et al., 2016; Nadejde and Tetreault, 2019). In this work, we build a multi-domain CGEC dataset from different writing scenarios and conduct basic cross-domain experiments, which can promote related research. We believe this is a valuable research direction for GEC even in the Large Language Model era (Fang et al., 2023; Coyne and Sakaguchi, 2023; Wu et al., 2023; Zhang et al., 2023).

8 Conclusion

This paper presents NaSGEC, a new multi-domain native CGEC dataset, which consists of 12,500 sentences from three representative native domains. We clearly describe the construction process and perform detailed data analysis. We conduct benchmark experiments with the SOTA BART-based CGEC model and two kinds of training data. We also launch domain transfer experiments and devise domain shift indicators, in order to have a clearer understanding of our domains. We hope NaSGEC can spur future work on cross-domain GEC evaluation, domain adaptation for GEC, and more.

Limitations

We think the limitations of our work are three-fold.

(1) As discussed in Section 2.1, we employ existing CGEC models to select sentences for annotation when building the MEDIA and THESIS domains of NaSGEC. Although this reduces annotation costs, it inevitably introduces biases into our dataset. For instance, the proportion of complex syntax- or semantic-related errors may be lower than that in reality, since existing CGEC models fail to identify them. Note that although we manage to mitigate such biases by voting with multiple models, this issue still exists. Future work should explore how to automatically mine erroneous sentences from a low error-density domain with minimal biases.

(2) The current size of our dataset is relatively

small. We will continuously collect more data from more diverse domains. Compared with other domains, THESIS has a much smaller data size (1.5k), as authorized papers are hard to obtain. In the future, we plan to cooperate with universities and thus accumulate more authorized data to enrich this domain.

(3) Based on our multi-domain NaSGEC, we have reported and analyzed cross-domain performance preliminarily. However, besides fine-tuning with small-scale data in the target domain, many other potentially helpful domain adaptation techniques can be tried. We believe cross-domain GEC is a valuable research topic and encourage future work to study it with NaSGEC.

Ethics Statement

Data license. For the EXAM and MEDIA domains of NaSGEC, we only collect sentences from public corpora or websites. For the THESIS domain, we have obtained permission from the authors of dissertations.

Annotation payment. During annotation, all annotators/reviewers were paid according to their finished task numbers and quality. The average salaries for annotators and reviewers are about 25 and 34 RMB per hour, respectively.

Acknowledgements

We thank all anonymous reviewers and the meta reviewer for their insightful comments, which will definitely help us improve this work in the future. This work was supported by the National Natural Science Foundation of China (Grant No. 62176173) and Alibaba Group through Alibaba Innovative Research Program, and also supported by Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. In *Proceedings of EMNLP-IJCNLP*, pages 4260–4270.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of BEA@ACL*, pages 52–75.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2022. Grammatical error correction: A survey of the state of the art. *arXiv preprint arXiv:2211.05166*.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese language technology platform. In *Proceedings of COLING*, pages 13–16.

Zhuang Chen and Tieyun Qian. 2021. Bridge-based active domain adaptation for aspect term extraction. In *Proceedings of ACL*, pages 317–327.

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of EMNLP*, pages 1901–1911.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of COLING*, pages 1304–1319.

Steven Coyne and Keisuke Sakaguchi. 2023. An analysis of gpt-3's performance in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The nus corpus of learner English. In *Proceedings of BEA@NAACL-HLT*, pages 22–31.

Yong Dai, Linyang Li, Cong Zhou, Zhangyin Feng, Enbo Zhao, Xipeng Qiu, Piji Li, and Duyu Tang. 2022. "Is whole word masking always better for Chinese BERT?": Probing on Chinese grammatical error correction. In *Proceedings of ACL (Short, Findings)*, pages 1–8.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. Grammatical error correction in low error density domains: a new benchmark and analyses. In *Proceedings of EMNLP*, pages 8467–8478.

Haoming Jiang, Chen Liang, Chong Wang, and Tuo Zhao. 2020. Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing. In *Proceedings of ACL*, pages 1823–1834.

Diederik P Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Entony Lekhtman, Yftah Ziser, and Roi Reichart. 2021. DILBERT: Customized pre-training for domain adaptation with category shift, with an application to aspect extraction. In *Proceedings of EMNLP*, pages 219–230.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880.

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. In *Proceedings of AAAI*, pages 10974–10982.

Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Liu Weijie, Mao Weiquan, and Zhang Hui. 2022b. CSL: A Large-scale Chinese Scientific Literature Dataset. In *Proceedings of COLING*, pages 3917–3923.

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of ACL*, pages 2386–2395.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Dingchao Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native Chinese grammatical error correction. In *Proceedings of EMNLP (Findings)*, pages 576–589.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: high-precision text editing. In *Proceedings of EMNLP-IJCNLP*, pages 5054–5065.

Maria Nadejde and Joel R. Tetreault. 2019. Personalizing grammatical error correction: Adaptation to proficiency level and L1. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP*, pages 27–33.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling robust grammatical error correction in new domains: data sets, metrics, and analyses. *TACL*, 7:551–566.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: a fluency corpus and benchmark for grammatical error correction. In *Proceedings of EACL*, pages 229–234.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR–grammatical error correction: tag, not rewrite. In *Proceedings of BEA@ACL*, pages 163–170.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT(Demo)*, pages 48–53.

Minh Quang Pham, Josep Maria Crego, and François Yvon. 2021. Revisiting multi-domain machine translation. *TACL*, 9:17–35.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP - A survey. In *Proceedings of COLING*, pages 6838–6855.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings of NLPTEA@ACL*, pages 42–51.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis. In *Proceedings of NLPTEA@AACL*, pages 25–35.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of Child Language*, 14(2):201–209.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: fluency instead of grammaticality. *TACL*, 4:169–182.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: a pre-trained unbalanced transformer for both Chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of ICCV*, pages 2818–2826.

Baoxin Wang, Xingyi Duan, Dayong Wu, Wanxiang Che, Zhigang Chen, and Guoping Hu. 2022. CCTC: A cross-sentence Chinese text correction dataset for native speakers. In *Proceedings of COLING*, pages 3331–3341.

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. *arXiv preprint arXiv:2112.15043*.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

Xiuyu Wu and Yunfang Wu. 2022. From spelling to grammar: A new framework for Chinese grammatical error correction. *ArXiv*, abs/2211.01625.

Lvxiaowei Xu, Jian-Cheng Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: Fine-grained corpus for Chinese grammatical error correction. In *Proceedings of EMNLP (Findings)*, pages 1900–1918.

Sen Yang, Leyang Cui, Ruoxi Ning, Di Wu, and Yue Zhang. 2022. Challenges to open-domain constituency parsing. In *Proceedings of ACL (Findings)*, pages 112–127.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of ACL*, pages 180–189.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. *AI Open*, 2:65–68.

Baolin Zhang. 2009. Features and functions of the HSK dynamic composition corpus (in Chinese). *International Chinese Language Education*, 4:71–79.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *arXiv preprint arXiv:2305.13225*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022a. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of NAACL-HLT*, pages 3118–3130.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022b. SynGEC: Syntax-enhanced grammatical error correction with a tailored gec-oriented parser. In *Proceedings of EMNLP*, pages 2518–2531.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of NAACL-HLT*, pages 156–165.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: grammatical error correction. In *CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC)*, pages 439–445.

A Annotation Tool

We present the annotation interface of our annotation tool in Figure 3. Given a potentially erroneous sentence, the annotator can rewrite it in a text box if he/she finds this sentence contains errors. If the sentence is correct, the annotator can directly click the `Error Free` button and submit.

Specifically, when annotating the MEDIA and THESIS domains, we provide annotators with the context of each sentence. Because sentences in these domains are extracted from complete essays or dissertations, they may need cross-sentence information to correct. We ask our annotators to mark such sentences with the `Need Context` button to facilitate future study in document-level CGEC.

Figure 3: Our annotation interface.

Figure 4 shows our review interface. The reviewer can choose whether accept each submission by clicking the check box before it. Considering other valid answers may be missed by annotators, the reviewer can also click the `Add` button to input a new correction for supplementary.

Figure 4: Our review interface.

B Experimental Details

We use the `fairseq` toolkit[5] (Ott et al., 2019) to build our benchmark models. Our model is based on the large variant of the Chinese BART (Shao et al., 2021)[6], which has about 400M parameters. Following Zhang et al. (2022b), we extend the original vocabulary of the Chinese BART to cover some

[5] https://github.com/facebookresearch/fairseq

[6] https://huggingface.co/fnlp/bart-large-chinese

Hyper-parameter	Value
Training	
Pretrained Language model	Chinese-BART-large (Shao et al., 2021)
Update steps	200,000
Devices	8 Tesla V100 GPU (32GB)
Batch size per GPU	8096 tokens
Optimizer	Adam (Kingma and Ba, 2014) ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning rate	3×10^{-5}
Warmup updates	4000
Max length	128
Loss function	Label smoothed cross entropy (label-smoothing=0.1) (Szegedy et al., 2016)
Dropout	0.3
Dropout-src	0.2
Fine-tuning	
Devices	1 Tesla V100 GPU (32GB)
Max epochs	100
Learning rate	1×10^{-5}
Batch size per GPU	1024 tokens
Generation	
Beam size	12
Max input length	128

Table 10: Our hyper-parameter settings.

common but missed Chinese characters and punctuation, e.g., Chinese quotation marks, which they find can greatly improve model performance.

We list detailed experimental hyper-parameter settings in Table 10. The total training time for using real learner data (about 1.35M sentence pairs) is about 10 hours. The total training time for using pseudo native data (about 100M sentence pairs) is about 7 days. Due to the limitation of time and computation resources, the benchmark results in Table 3 are reported over a single run. The fine-tuning time is about 20 minutes. All fine-tuning results in Table 5 and Table 9 are averaged over 3 runs with distinct random seeds.

C Results of the Seq2Edit Model

Besides Seq2Seq-based models like BART (Lewis et al., 2020), there is another competitive CGEC paradigm called Seq2Edit. The Seq2Edit-based models first predict a sequence of edits, and then apply them to the erroneous sentence to conduct corrections (Malmi et al., 2019; Awasthi et al., 2019). Recently, Zhang et al. (2022a) adapt GECToR (Omelianchuk et al., 2020), a widely-used Seq2Edit model in English, to Chinese and find it can achieve promising performance. Hence, we follow their efforts and test the ability of Chinese GECToR on NaSGEC, as shown in Table 11. Both BART and GECToR are trained on real learner training data described in Section 4.2.

MEDIA	P	R	$F_{0.5}$
BART	**35.96**	**29.15**	**34.35**
GECToR	33.36	19.85	29.36
THESIS	**P**	**R**	$F_{0.5}$
BART	24.16	**34.06**	25.65
GECToR	**42.29**	18.20	**33.44**
EXAM	**P**	**R**	$F_{0.5}$
BART	**23.01**	**11.31**	**19.06**
GECToR	20.93	8.80	16.41

Table 11: Experimental results of the Seq2Edit-based model (GECToR) compared with the Seq2Seq-based model (BART) on NaSGEC.

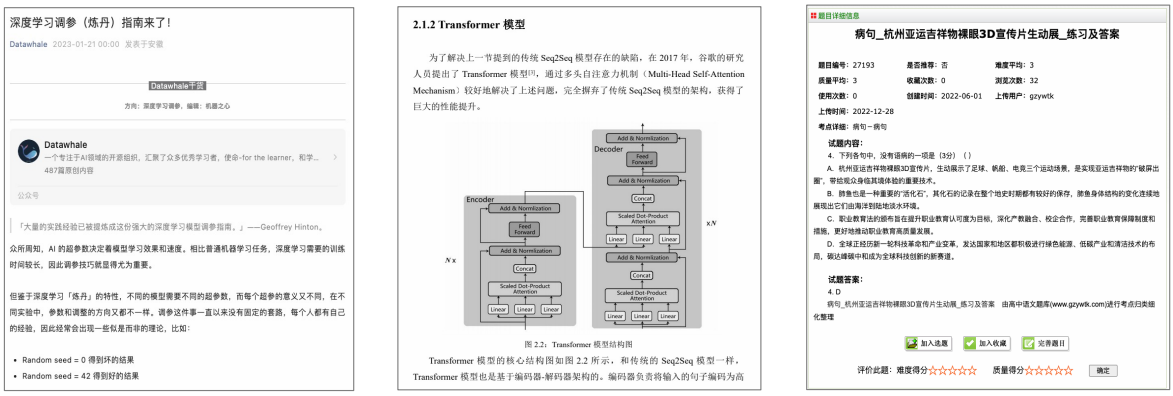
We can see that, in MEDIA and EXAM, Seq2Seq outperforms Seq2Edit substantially. However, in THESIS, Seq2Edit performs significantly better. We attribute this to Seq2Edit's natural ability to copy. Seq2Edit can directly copy tokens from the source sentence by predicting the `Keep` tag. In THESIS, there are many English words and technical terms, which Seq2Seq tends to mis-correct while Seq2Edit keeps unchanged. So Seq2Edit achieves a much higher precision in this domain. In view of this, we plan to enhance our BART-based benchmark models with the copy mechanism (Zhao et al., 2019) or other approaches in the future.

D Pseudo Data Generation

We use rule-based corruption to generate large-scale synthetic training data from clean native corpora. Specifically, we randomly select 100M sentences from the WuDao corpora (Yuan et al., 2021)[7] as the seed corpus, which is mainly composed of website articles written by native speakers. We select tokens for corruption with a probability of 0.05 and perform the following operations with corresponding probabilities (in parentheses):

- **Replacement** (0.55): We replace the current token with another token in its confusion set (0.5) or a random token from the whole vocabulary (0.5).
- **Insertion** (0.2): We insert the same token (0.5) or a random token from the whole vocabulary (0.5) before the current token

[7] https://data.wudaoai.cn/home



(a) Social media platform (MEDIA) (b) Undergraduate thesis (THESIS) (c) Ungrammatical sentence judgement question (EXAM)

Figure 5: The screenshots of data sources for our 3 domains.

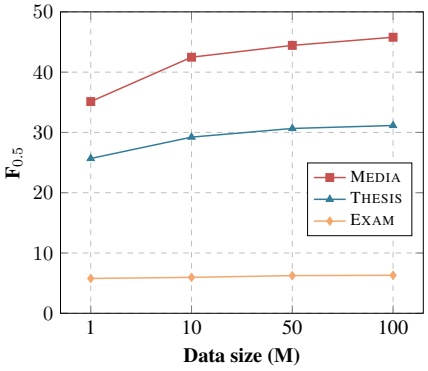


Figure 6: Impact of pseudo data size in different domains of NaSGEC.

- **Deletion** (0.2): We delete the current token.
- **Swap** (0.05): We swap the current token and the token after it.

Following Dai et al. (2022), we inject noises from both character and word granularity to achieve better performance, which means each sentence is segmented into either the word (0.5) or character (0.5) sequence before corruption. The word-level and character-level confusion sets are built considering phonetics and glyphs.

We also show the effect of the size of pseudo data in Figure 6. When the data size increases, the model performance continuously improves in the MEDIA and THESIS domains, whereas the model performance in the EXAM domain keeps low.

E Error Type Performance

In Table 12, we evaluate the error type performance of each domain's best model on NaSGEC. The best model denotes the fine-tuned model achieving the highest $F_{0.5}$ score in Table 5.

	MEDIA P/R/$F_{0.5}$	THESIS P/R/$F_{0.5}$	EXAM P/R/$F_{0.5}$
S	59.91/51.66/58.06	29.79/60.64/33.17	25.38/15.07/22.33
M	67.56/32.54/55.59	47.37/15.38/33.46	44.21/19.62/35.35
R	59.41/42.44/55.01	65.71/34.85/55.83	66.10/41.42/59.06
W	40.00/12.77/28.04	42.25/12.75/28.88	29.74/9.46/20.82

Table 12: The fine-grained performance of each domain's best model regarding error types. S: Substituted errors, M: Missing errors, R: Redundant errors, W: Word-order errors.

In all domains, models repair redundant errors consistently well, as their corrections do not need to generate new content and are the easiest and most deterministic. In contrast, models encounter difficulties in handling word-order errors universally since such errors require long-range structural knowledge to correct.

In terms of substituted and missing errors, models exhibit divergent behaviours. The performance on substituted errors in MEDIA is very high, probably because they are often spelling and punctuation errors. However, as another realistic writing scene, THESIS has a much inferior performance on substituted errors due to the low correction precision. After studying cases, we find THESIS contains many English words (e.g., LSTM) and technical terms (e.g., 支持向量机, *supporting vector machine*), which usually cause miscorrection. Besides, the performance on substituted errors in EXAM is also quite low, owing to their complexity.

Considering missing errors, the model performs much better in MEDIA than others. As discussed before, we observe that a large proportion of missing errors in MEDIA is caused by missing punctuation, which well-trained models can easily handle.

Domain: MEDIA	
Source	30日下午齐鲁晚报的一名读者报料称，南湖镇两个女孩泥水，正在医院抢救。 On the afternoon of the 30th a reader of the Qilu Evening News reported that two girls in Nanhu Town muddy water, and were being rescued in the hospital.
Ref. 1	30日下午，齐鲁晚报的一名读者报料称，南湖镇两个女孩溺水，正在医院抢救。 On the afternoon of the 30th, a reader of the Qilu Evening News reported that two girls in Nanhu Town drowned, and were being rescued in the hospital.
Source	应当注意的是，重音切记过多。过多则显示不了孰轻孰重。 It is worth noting that too much stress should be remembered. Too much stress can not show which is more important.
Ref. 1	应当注意的是，重音切忌过多。过多则显示不了孰轻孰重。 It is worth noting that too much stress should be avoided. Too much stress can not show which is more important.
Ref. 2	应当注意的是，重音切记不要过多。过多则显示不了孰轻孰重。 It is worth noting that avoiding too much stress should be remembered. Too much stress can not show which is more important.
Domain: THESIS	
Source	目前应用最为广泛的词干提取方法为波特词干算法（Poter-Stemmer），它基于后缀进行玻璃。 At present, the most widely used stemming method is the Poter-Stemmer algorithm, which is based on the suffix for glass.
Ref. 1	目前应用最为广泛的词干提取方法为波特词干算法（Poter-Stemmer），它基于后缀进行剥离。 At present, the most widely used stemming method is the Poter-Stemmer algorithm, which is based on the suffix for stripping.
Source	word2vec的基本结构是一个输入隐藏输出的三层神经网络。 The basic structure of word2vec is a three-layer neural network with input hidden output.
Ref. 1	word2vec的基本结构是一个包含输入层、隐藏层和输出层的三层神经网络。 The basic structure of word2vec is a three-layer neural network including the input layer, hidden layer and output layer.
Ref. 2	word2vec的基本结构是一个由输入层、隐藏层和输出层组成的三层神经网络。 The basic structure of word2vec is a three-layer neural network composed of the input layer, hidden layer and output layer.
Domain: EXAM	
Source	止咳祛痰片，它里面的主要成分是远志、桔梗、贝母、氯化铵等配制而成的。 Zhike Qutan Tablet, the main components of which are mainly compounded of Milkwort, Platycodon grandiflorum, Fritillaria, Ammonium chloride, etc.
Ref. 1	止咳祛痰片，它里面的主要成分是远志、桔梗、贝母、氯化铵等~~配制而成的~~。 Zhike Qutan Tablet, the main components of which are ~~mainly compounded of~~ Milkwort, Platycodon grandiflorum, Fritillaria, Ammonium chloride, etc.
Ref. 2	止咳祛痰片，它~~里面的主要成分~~是远志、桔梗、贝母、氯化铵等配制而成的。 Zhike Qutan Tablet, ~~the main components of~~ which are mainly compounded of Milkwort, Platycodon grandiflorum, Fritillaria, Ammonium chloride, etc.
Source	同学们临走时总是忘记关灯。从这一件平凡的小事中，却说明了一个大问题。 The students always forget to turn off the lights when they leave. From this trivial matter, shows a big problem.
Ref. 1	同学们临走时总是忘记关灯。从这一件平凡的小事中，我们却发现了一个大问题。 The students always forget to turn off the lights when they leave. From this trivial matter, we found a big problem.
Ref. 2	同学们临走时总是忘记关灯。~~从~~这一件平凡的小事~~中，~~却说明了一个大问题。 The students always forget to turn off the lights when they leave. ~~From~~ This trivial matter shows a big problem.

Table 13: Annotation examples in NaSGEC. "Source" denotes the source sentence, "Ref" denotes the reference.

F Annotation Examples

We show some real annotation examples from NaSGEC in Table 13. We also present screenshots of all data sources of our domains in Figure 5.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation Section

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3,4,5,6

☑ B1. Did you cite the creators of artifacts you used?
Section 3,4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Ethics Statement Section

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Statement Section

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics Statement Section

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 1,2,3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3,4

C ☑ Did you run computational experiments?

Section 5,6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix B

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3,4,5,6

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 1,2,3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3, Ethics Statement Section

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics Statement Section

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 2