Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

10-2023

FlaCGEC: A Chinese grammatical error correction dataset with fine-grained linguistic annotation

Hanyue DU

Yike ZHAO

Qingyuan TIAN

Jiani WANG

Lei WANG
Singapore Management University, lei.wang.2019@phdcs.smu.edu.sg

See next page for additional authors

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Asian Studies Commons, Databases and Information Systems Commons, and the East Asian Languages and Societies Commons

Citation

DU, Hanyue; ZHAO, Yike; TIAN, Qingyuan; WANG, Jiani; WANG, Lei; LAN, Yunshi; and LU, Xuesong. FlaCGEC: A Chinese grammatical error correction dataset with fine-grained linguistic annotation. (2023). *CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, October 21-25*. 5321-5325.
Available at: https://ink.library.smu.edu.sg/sis_research/8463

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Author

Hanyue DU, Yike ZHAO, Qingyuan TIAN, Jiani WANG, Lei WANG, Yunshi LAN, and Xuesong LU

This conference proceeding article is available at Institutional Knowledge at Singapore Management University: https://ink.library.smu.edu.sg/sis_research/8463

FlaCGEC: A Chinese Grammatical Error Correction Dataset with Fine-grained Linguistic Annotation

Hanyue Du
East China Normal University
Shanghai, China
hydu@stu.ecnu.edu.cn

Yike Zhao
East China Normal University
Shanghai, China
ykzhao@stu.ecnu.edu.cn

Qingyuan Tian
East China Normal University
Shanghai, China
qytian@stu.ecnu.edu.cn

Jiani Wang
East China Normal University
Shanghai, China
jiani.wang@stu.ecnu.edu.cn

Lei Wang
Singapore Management University
Singapore
lei.wang.2019@phdcs.smu.edu.sg

Yunshi Lan*
East China Normal University
Shanghai, China
yslan@dase.ecnu.edu.cn

Xuesong Lu
East China Normal University
Shanghai, China
xslu@dase.ecnu.edu.cn

arXiv:2311.04906v1 [cs.CL] 26 Sep 2023

ABSTRACT

Chinese Grammatical Error Correction (CGEC) has been attracting growing attention from researchers recently. In spite of the fact that multiple CGEC datasets have been developed to support the research, these datasets lack the ability to provide a deep linguistic topology of grammar errors, which is critical for interpreting and diagnosing CGEC approaches. To address this limitation, we introduce FlaCGEC, which is a new CGEC dataset featured with fine-grained linguistic annotation. Specifically, we collect raw corpus from the linguistic schema defined by Chinese language experts, conduct edits on sentences via rules, and refine generated samples manually, which results in 10k sentences with 78 instantiated grammar points and 3 types of edits. We evaluate various cutting-edge CGEC methods on the proposed FlaCGEC dataset and their unremarkable results indicate that this dataset is challenging in covering a large range of grammatical errors. In addition, we also treat FlaCGEC as a diagnostic dataset for testing generalization skills and conduct a thorough evaluation of existing CGEC models.

CCS CONCEPTS

• **Computing methodologies** → **Language resources**; **Natural language processing**.

KEYWORDS

Chinese Grammatical Error Correction, Fine-grained Linguistic Annotation, Deep Learning

ACM Reference Format:
Hanyue Du, Yike Zhao, Qingyuan Tian, Jiani Wang, Lei Wang, Yunshi Lan, and Xuesong Lu. 2023. FlaCGEC: A Chinese Grammatical Error Correction Dataset with Fine-grained Linguistic Annotation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615119

*Corresponding author.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615119

Table 1: Comparison of FlaCGEC with existing CGEC datasets.

Datasets	Annotation type	Type number	Source
NLPCC [23]	Edits	4	CFL
CGED [14, 15]	Edits	4	CFL
CTC [22]	Edits	3	Native speaker
MuCGEC [21]	Edits, Linguistic	19	CFL
NaCGEC [10]	Edits, Linguistic	26	Native speaker
FCGEC [19]	Edits, Linguistic	28	Native speaker
FlaCGEC	Edits, Linguistic	210	Native speaker

1 INTRODUCTION

Writing grammatically correct Chinese sentences is difficult for learners studying Chinese as a Foreign Language (CFL) and even for native Chinese speakers due to its complex grammar rules. Chinese Grammatical Error Correction (CGEC), aiming to detect and correct all grammatical errors in a sentence and produce an error-free sentence, has attracted intensive attention from researchers for its crucial value in many natural language processing scenarios such as writing assistant and search engine [6, 9, 13]. Due to its profound significance, a surge of datasets have been observed [14, 16, 21, 23].

However, most of these datasets [14, 23] only provide correct sentences as the ground truth and have limitations in providing linguistic annotations to a CGEC method, which hinders the further improvement of a method. Recent studies [10, 19, 21] proposed CGEC datasets with grammatical error types. However, their grammatical error types follow a shallow linguistic schema. A fine-grained linguistic schema widely covering the gramamtical points is demanded, which not only increases the interpretability of CGEC tasks, but also helps diagnose the CGEC methods [11, 18].

To this end, we present **FlaCGEC**[1], a Chinese grammatical error correction dataset with fine-grained linguistic annotation. We show an overall comparison of the differences between FlaCGEC and other datasets in Table 1. We first derive a linguistic schema from the grading standards textbook, where 78 instantiated grammar points are organized in a deep hierarchical structure. For data collection, we first collect target sentences from Chinese proficiency test for the sake of obtaining a corpus with diverse grammatical points. Then we design edit rules to generate a set of enormous sentences for each target sentence and align the grammar points to corresponding edits. Eventually, we obtain erroneous sentences covering diverse grammatical errors, which will be further verified by annotators.

We reproduce the state-of-the-art CGEC models and thoroughly evaluate them on FlaCGEC. We discover there is a gap between the best baseline model and human performance, which indicates that FlaCGEC is still challenging. Furthermore, we observe a significant performance drop of models that are trained on existing datasets. This reveals the distinction of FlaCGEC. We also consider FlaCGEC as a diagnostic dataset for analyzing the existing CGEC models and discover that current models have poor generalization capability over diverse grammatical types and struggle to correct sentences with the complicated syntax as well as special usage of grammar points. We hope FlaCGEC dataset could provide a comprehensive challenge to encourage more contributions to CGEC tasks.

2 DATASET

2.1 Annotation

FlaCGEC dataset enables quantitative and comprehensive evaluation of CGEC methods. It not only provides the target sentences as the golden standards, but also annotates the error types explicitly. Following the M2 format annotation of general GEC datasets [2, 12], we annotate data with (1) the span of grammatical erroneous context (2) the error type and (3) the corresponding correction.

To ensure the annotated error types are canonical and recognized by the standard syllabus, we apply the official grammatical types in *Chinese Proficiency Grading Standards for International Chinese Language Education*[2] for annotating. This could help examine the performance of a CGEC method on specific error types and give fine-grained feedback to CGEC approaches. We show an example with M2 format annotation of FlaCGEC below.

[S] 章鱼有发达的神经系统，为人亲善。
Translation: Octopuses have powerful neural systems, regarding human kindly.
[T] 章鱼有发达的神经系统，对人亲善。
Translation: Octopuses have powerful neural systems, treating human kindly.
[A] 11 11 ||| S-Preposition[词类介词] ||| 对

where the lines preceded by [S] and [T] represent the source sentence and target sentence, respectively. [A] goes ahead of an annotation, which consists of start token index, end token index, instantiated grammar point as well as edit, and correction. The above annotation indicates that to correct the sentence, preposition

[1]Website: https://github.com/hyDududu/FlaCGEC
[2]The textbook could be found in http://www.moe.gov.cn/jyb_sjzl/ziliao/A19/202111/W020211118507389477190.pdf.

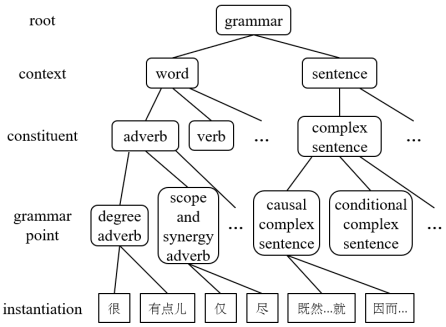

Figure 1: Hierarchical structure of linguistic schema.

"为" should be changed to "对". We allow a single source sentence to contain multiple grammatical errors as presented in Table 4.

2.2 Data Collection

Linguistic Schema. Our goal is to construct a CGEC dataset with fine-grained linguistic annotation. Unlike previous studies which manually designed linguistic schema for grammatical error types [10], we employ the schema defined in the grading standards textbook to specify the grammar points. Figure 1 demonstrates the deep hierarchical structure of our linguistic schema. Each grammar point has various words and phrases as examples. In the next section, we annotate grammatical errors using the above schema.

Target Sentence Collection. To collect target sentences, we utilize reading corpus from HSK exam [3, 20], which is an official Chinese proficiency test. HSK corpus contains passages, where the sentences generally follow the above grammatical schema. It is worth noting that previous CGED datasets [14] also employed HSK exam, but they utilize sentences wrtitten by CFL in HSK exam, which is more likely to reveal the limited grammatical errors from CFL. By contrast, our study creates more complicated sentences based on the standard reading corpus featured with varying grammatical errors.

Therefore, we extract all the passages in HSK corpus via OCR and chunk them into sentences. Then we either write regular expressions of close grammar points like *pronouns* to annotate sentences or collect the illustrative examples with annotation from textbook. These sentences are treated as the initial data pool. Next, we iteratively train a tagging model to predict the grammar points contained in a sentence by adding the predicted sentences with high confidence into the data pool as augmented data [7]. Eventually, we collect a set of target sentences from HSK corpus that are annotated with grammar points.

Source Sentence Generation. Following the traditional methods [10] that automatically generate large-scale training data containing grammatical errors, we generate erroneous sentences via the following edits:

- **Removing words** means we randomly remove words of certain grammar points from the sentences such that some grammatical components are missing. We denote this edit type as "M".
- **Substituting words** means we randomly replace words of certain grammar types in the sentences with another word of the same grammar types such that the collocation of the sentence is not appropriate. We denote this edit type as "S".

Table 2: Statistics and properties of FlaCGEC dataset.

Properties	Train	Dev	Test
#Sentences	10,804	1,334	1,325
Average source sentence length	35.09	34.76	35.83
Average target sentence length	35.59	35.29	36.34
#Edits per sentence	1.72	1.69	1.71
#Grammar points	77	69	72

- **Reordering words** means we randomly reorder the words of certain grammar points in the sentence leading to incorrect sentence syntax. We denote this edit type as "W".

Starting from the collected target sentences with annotated grammar points, we first do Chinese word segmentation via Jieba toolkit[3]. For words related to grammar points, we randomly perform one or multiple of the above edits, with a possibility of leaving the word unchanged. This results in multiple edits on a single target sentence and even multiple edits on the same words. Eventually, we collect about 12, 568 source sentences. Each source sentence is associated with its target sentence as well as the corresponding M2 annotation.

Bad Case Filtering. Random combinations of edits and grammar points result in a large number of candidate instances but some of them are improperly applied resulting in invalid source sentences. Then we employ native speakers to filter out bad cases. Bad cases are identified if: 1) too many errors exist in a single source sentence leading to confusing semantics which cannot be recovered even by native speakers; 2) the grammatical errors in sentences rarely exit in real-life scenarios and cannot be reproduced by annotators.

Specifically, we invite 13 Chinese postgraduate students to filter the bad cases. We write an annotation guideline to help the annotators better understand the annotation task. Meanwhile, we provide intensive training to them before annotation. Annotators should determine whether an example is a bad case based on the above judgment criterion and give a range of scores depending on the matching degree to the above criterias.

To ensure dataset quality, we select a senior annotator to review. For each batch of annotated data, the senior annotator randomly samples instances to review. If the annotation disagreement exceeds a threshold, the batch is reassigned to a new annotator until agreement is reached. After that, we filter out the bad cases, keep the rest and randomly split the data into training, development, and test sets with an 8 : 1 : 1 ratio and without target sentences overlapping for data split. This results in our FlaCGEC dataset.

2.3 Data Analysis

We report the statistics of FlaCGEC dataset in Table 2. We can see that FlaCGEC dataset provides a great number of erroneous sentences for training a good CGEC model. The varying error types are included in the dataset and they are evenly distributed over training, development and test sets. We display the distribution of edit types on selected grammatical types in FlaCGEC dataset in Figure 2. We sample 10 grammar points from our linguistic schema and display the distribution of edits on these grammar points in FlaCGEC dataset. We notice that there is a variance of the frequencies between grammar points (e.g., more errors exist in *passive sentences* than *exclamatory sentences*). There is a correlation between edits and grammar points (e.g., *exclamatory sentences* only have *substitution* edits while *prepositions for time* has a lack of *substitution* edits). This is because some combinations of grammar points and edits are invalid and recognized as bad cases. This could help a GEC model learn the nature of language expression [19].

[3] https://github.com/fxsjy/jieba.

Table 3: Detection and correction results evaluated on FlaCGEC test set with different training data. Δ denotes the decrease percentage of $F_{0.5}$ of the current setting based on FlaCGEC → FlaCGEC.

Train Data → Test Data	Model	Detection				Correction			
		R	*P*	$F_{0.5}$	Δ	*R*	*P*	$F_{0.5}$	Δ
FCGEC → FlaCGEC	GECToR-Chinese	9.95	34.01	22.92	65.90	5.91	14.49	11.23	56.98
	Chinese BART	10.58	19.11	16.46	48.08	9.54	12.49	11.76	31.62
	EBGEC	3.31	19.21	9.80	86.73	3.17	13.26	8.11	87.50
CTC → FlaCGEC	GECToR-Chinese	26.39	51.19	43.09	38.59	20.03	29.77	27.13	32.46
	Chinese BART	26.43	34.52	32.53	26.32	24.35	23.98	24.05	12.67
	EBGEC	3.72	33.19	12.85	82.60	3.68	24.36	11.48	82.30
FlaCGEC → FlaCGEC	GECToR-Chinese	66.62	72.95	71.59	–	50.03	47.75	48.19	–
	Chinese BART	52.50	51.84	51.97	–	43.48	35.98	37.27	–
	EBGEC	**79.56**	**72.55**	**73.85**	–	**75.33**	**62.70**	**64.87**	–
Human		78.64	86.93	85.14		63.95	73.72	71.53	

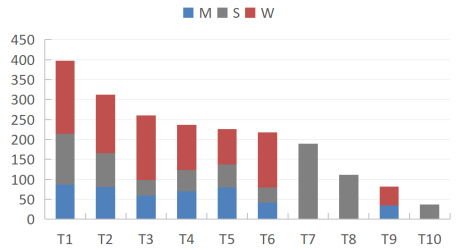

Figure 2: Distribution of edit types on selected grammar points in FlaCGEC. T1: *passive sentence*; T2: *preposition for places*; T3: *successive complex sentence*; T4: *casual complex sentence*; T5: *comparative sentence*; T6: "is"-*sentence*; T7: *interrogative sentence*; T8: *declarative sentence*; T9: *preposition for time*; T10: *exclamatory sentence*.

3 EXPERIMENTS

3.1 Experimental Setup

To test the performance of cutting-edge CGEC approaches on our FlaCGEC dataset, we adopt three mainstream CGEC models: GECToR-Chinese [21], Chinese BART [21][4] and EBGEC [8][5]. We use the public source codes of these benchmark models, maintain their official hyperparameters, and perform experiments with various settings, which are illustrated in detail in the following sections.

In terms of evaluation metrices, we refer to the CLTC2022 shared task [17][6] and employ **MaxMatch** (M_2) scorer [4] for evaluation. We treat a detection prediction as correct if the predicted start token index and end token index are identical to the ground truth. On top of that, if the edit is identical to the ground truth, the correction prediction is correct. We report standard micro Precision, Recall, and $F_{0.5}$ score to evaluate the performance.

[4] https://github.com/HillZhang1999/MuCGEC
[5] https://github.com/kanekomasahiro/eb-gec.
[6] https://github.com/blcuicall/CCL2022-CLTC

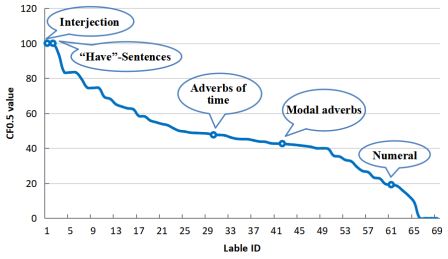

Figure 3: Correction $F_{0.5}$ of GECToR-Chinese on FlaCGEC test set over fine-grained grammatical points.

Table 4: Some error cases predicted by GECToR-Chinese. [P] **goes ahead of a predicted sentence.**

Modal adverb[情态副词]: 就
[S] 尊重和被尊重，就站在镜子前面，…
Translation: Respect and being respected, just standing in front of a mirror, ...
[T] 尊重和被尊重，就好像站在镜子前面，…
Translation: Respect and being respected, just like standing in front of a mirror, ...
[P] 尊重和被尊重，就好站在镜子前面，…
Translation: Respect and being respected, just have to stand in front of a mirror, ...
Conjunctions for connecting clauses[介词连接分句]:虽然...但是 Negative adverb[否定副词]:没有
[S] 但有没受到老板的责备，而且他心里很失落。
Translation: But did he not receive the blame from his boss, and he is upset
[T] 虽然没有受到老板的责备，但是他心里很失落。
Translation: Even though he did not receive the blame from his boss, he is upset.
[P] 虽然有受到老板的责备，但是他心里很失落。
Translation: Even though he received the blame from his boss, he is upset.

3.2 Difficulty of FlaCGEC

To understand the difficulty of the FlaCGEC dataset, we conduct a set of experiments on FlaCGEC. The results are presented in Table 3 and we have the following observations:

(1) Regarding human evaluation, we hire 4 native speakers to annotate the test set. As shown in Table 3, FlaCGEC dataset is also challenging for humans due to the wide range of grammatical errors it examines. Overall, the human evaluation yields high precision but relatively low recall. This observation is similar to prior study [19].

(2) Amongst the models trained on FlaCGEC dataset, EBGEC obtains the best result for detection and correction. We speculate that this is because EBGEC is able to correct the grammatical errors by referring to the most similar training example while the other two methods entirely rely on the contexts. However, EBGEC cannot play full of the advantage when it encounters a gap between training and test data. In comparison, GECToR-Chinese shows more robust results in the three experimental settings, such that we conduct a set of analysis based on GECToR-Chinese in the next section. Overall, there is still a significant gap of performance between the best model with humans on FlaCGEC.

(3) When we train the benchmark models on the other two datasets, we notice that their performance all drops a lot. This reflects that FlaCGEC dataset contains a great number of grammatical errors that are not involved in other datasets. Comparing FCGEC and CTC datasets, FlaCGEC is more likely to have similar grammatical errors as those in CTC dataset. We investigate and notice that FCGEC examines more on the correct syntactic construction of sentences while our dataset focuses more on the accurate discrimination of grammatical usage in sentences.

Table 5: Detection and correction $F_{0.5}$ of PLMs evaluated on sampled FlaCGEC test data under zero-shot setting.

PLMs	Detection			Correction		
	R	P	$F_{0.5}$	R	P	$F_{0.5}$
ChatGPT	39.36	23.12	25.30	25.18	11.36	12.76
GPT-3	39.89	33.19	34.34	23.02	15.31	16.41

3.3 More Analysis

Analysis of Fine-grained Performance. We display the experimental results of GECToR-Chinese over fine-grained grammar points in Figure 3. We obverse there is a variance of correction $F_{0.5}$ over fine-grained grammatical points. The best results are close to 1 while the worst results approach 0. Specifically, certain grammatical errors like *interjection*, *"have"-sentences* can be easily solved by GECToR-Chinese. But it has difficulty in correcting errors related to *numeral*. This may be because *numeral* has more flexible usage in Chinese expressions, which requires a deep understanding of the sentences and strong generalization capability of CGEC models.

Case Study. We show some error cases in Table 4. Case 1 has grammar errors related to *modal adverb*, which is not easy to be detected as a CGEC model needs to understand the context and know it is a metaphor. GECToR-Chinese succeeds to understand the sentence and accurately detects the spans to be corrected. But it fails to correct it by inserting a wrong modal adverb. In Case 2, a CGEC model must grasp sentence semantics, detect conjunction misuse, and understand the necessary contextual information. GECToR-Chinese accurately corrects the conjunction but unsuccessfully comprehends the statement and the predicted sentence presents an invalid progressive description. This indicates that CGEC models suffer more when special usage of grammar is examined.

Zero-shot Transfer on FlaCGEC. Recently, Large-scale Language Models (LLMs) [1, 5] have shown to be effective in *few-shot* or *zero-shot* scenarios. To understand whether LLMs have the capability to solve FlaCGEC dataset under zero-shot setting, we employ ChatGPT[7] and GPT-3 with the following prompt:

Prompt($\mathbf{x}$) = *I will show you a Chinese sentence with grammatical errors, please show me the correct sentence. The wrong sentence is* $\mathbf{x}$.

We evaluate the generated sentences with M_2 and present the results in Table 5. From the results, we observe that the simple prompt could lead to remarkable gains in CGEC tasks. But the results are still significantly worse than the state-of-the-art on both detection and correction, which are 73.85% and 64.87%, respectively. This is because LLMs recover basic semantics of sentences but neglect the accurate discrimination of the grammar points.

4 CONCLUSIONS

In this paper, we introduce FlaCGEC, a CGEC dataset with fine-grained linguistic annotation. We conduct a thorough evaluation of cutting-edge CGEC methods showing that our dataset is challenging and could provide environments to test diverse generalization abilities of CGEC methods. With FlaCGEC, current CGEC methods can be trained and diagnosed to improve performance continuously.

[7] https://openai.com/blog/chatgpt/

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by East China Normal University (2022ECNU—WHCCYJ-31), Natural Science Foundation of China (Project No. 61977026) and Shanghai Pujiang Talent Program (Project No. 22PJ1403000).

REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*. 1877–1901.

[2] Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 52–75.

[3] Xiliang Cui and Baoli Zhang. 2013. Design Concepts of "the Construction and Research of the Inter-language Corpus of Chinese from Global Learners". In *Language Teaching and Linguistic Studies*. 27–34.

[4] 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 the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. 22–31.

[5] 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)*. 4171–4186.

[6] 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*. 117–126.

[7] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 968–988.

[8] Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for Language Learners Using Example-Based Grammatical Error Correction. *arXiv preprint arXiv:2203.07085* (2022).

[9] 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 (Volume 1: Long Papers)*. 4973–4984.

[10] Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, et al. 2022. Linguistic Rules-Based Corpus Generation for Native Chinese Grammatical Error Correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

[11] Masato Mita and Hitomi Yanaka. 2021. Do Grammatical Error Correction Models Realize Grammatical Generalization?. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*. 4554–4561.

[12] Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*. 1–12.

[13] 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*. 163–170.

[14] 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*. 42–51.

[15] 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*. 25–35.

[16] Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 Bake-off for Chinese Spelling Check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*. 32–37.

[17] Yingying Wang, Cunliang Kong, Xin Liu, Xuezhi Fang, Yue Zhang, Nianning Liang, Tianshuo Zhou, Tianxin Liao, Liner Yang, Zhenghua Li, Gaoqi Rao, Zhenghao Liu, Chen Li, Erhong Yang, Min Zhang, and Maosong Sun. 2022. *Overview of CLTC 2022 Shared Task: Chinese Learner Text Correction.* Technical Report.

[18] Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A Comprehensive Survey of Grammatical Error Correction. *ACM Trans. Intell. Syst. Technol.* 12, 5 (2021).

[19] Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: Fine-Grained Corpus for Chinese Grammatical Error Correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

[20] Baoli Zhang. 2011. The Principles for Building the "International Corpus of Learner Chinese". In *Applied Linguistics*. 100–108.

[21] 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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3118–3130.

[22] Honghong Zhao, Baoxin Wang, Dayong Wu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2021. Overview of CTC 2021: Chinese Text Correction for Native Speakers. *arXiv preprint arXiv:2208.05681* (2021).

[23] 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*, Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan (Eds.). 439–445.