Overview of NLPCC 2025 Shared Task 5: Chinese Government Text Correction with Knowledge Bases

Yang Song, Yuxiang Jia, Yuchen Yan, Jiajia Cui, Lingling Mu, and Hongfei Xu(✉)

Zhengzhou University, Henan 450001, China
{ieyxjia,iellmu}@zzu.edu.cn, {yanyuchen,jjcui}@gs.zzu.edu.cn, hfxunlp@foxmail.com

Abstract. In recent years, the field of Chinese text error correction has advanced rapidly, with machine learning-based correction algorithms significantly improving performance. However, existing research often overlooks the integration of Knowledge Bases (KBs) to guide error correction, despite their potential value in rectifying critical factual errors or dynamically adjusting correction results based on KB updates. In this paper, we develop specialized KBs and datasets for the automatic text error correction of Chinese government documents. The datasets are built upon authentic news corpora, real-world user inputs and their needs. Furthermore, we present KB-oriented metrics to evaluate text correction performance on knowledge-related terms. We test the baseline performances of several Large Language Models (LLMs), including Deepseek, Qwen, GLM, and Baichuan, for their exceptional language understanding and reasoning capabilities, and then report the performances and methods of five systems participating in the shared task.

Keywords: Chinese Text Error Correction · Knowledge Base · Large Language Model

1 Introduction

The core objective of the Chinese text error correction task is to identify and rectify errors in Chinese texts. In recent years, with the rapid advancement of Natural Language Processing (NLP), Chinese text error correction has made significant progress [2,8–13]. It is a helpful proofreading application and is also beneficial for the other applications [1,5,6,15].

However, it should be noted that while existing efforts on error correction have achieved remarkable success in detecting and correcting spelling and grammatical errors empowered by Pre-trained Language Models (PLMs) [3,4,14,16–19], they still exhibit notable deficiencies when handling factual errors involving domain-specific knowledge. Correcting such errors typically requires precise mastery and flexible application of specialized knowledge, however, current correction methods generally lack knowledge integration mechanisms. Incorporating

X.-L. Mao et al. (Eds.): NLPCC 2025, LNAI 16105, pp. 417–428, 2026.
https://doi.org/10.1007/978-981-95-3352-7_35

knowledge resources can help error correction systems correct text that contradicts objective facts. Particularly noteworthy is the dynamic update capability of Knowledge Bases (KBs), enabling continuous tracking of the latest facts and knowledge to ensure error correction results remain updated.

In this work, we construct specialized knowledge bases and datasets for Chinese government text error correction, based on authentic news corpora and real-world user inputs. The KBs developed in this study encompasses a rich variety of categories, including official information, policy/spirit names, standard expressions, fixed phrases, idioms and so on, as detailed in Fig. 2. We use match based methods to locate related KB terms in sentences, and construct training set, validation set and test set by extracting sentences with KB matches in news corpora and real user inputs with human corrections. To evaluate the performance of error correction systems regarding KB terms, we introduce four evaluation metrics for the correction with KBs: KB Accuracy (KB-Acc), KB precision (KB-P), KB Recall (KB-R), and KB F0.5 score (KB-F0.5). These metrics aim at precisely assessing the error correction capability of systems in handling KB related errors.

Given the exceptional capabilities of Large Language Models (LLMs) in semantic understanding and knowledge utilization, we select Deepseek, Qwen, GLM and Baichuan as baseline systems, and test their performances with an instruction template which integrates KB terms into the correction of the input sentences. We have organized NLPCC 2025 shared task 5 based on the resources. The shared task has attracted 10 teams from both the universities and industries, with five teams ultimately submitted system results. The results provided in Tables 4 and 6 show the challenge of the task.

2 Task Description

The Chinese Government Text Correction (CGTC) shared task aims to detect and correct spelling and grammatical errors in Chinese government texts with the help of knowledge bases. It is highly related to Chinese Spelling Error Correction (SEC) and Chinese Grammatical Error Correction (GEC) but has special characteristics with respect to the domain and the user needs. For example, in government text correction, misspelling of policy names or containing factual errors are usually considered more serious than grammatical errors.

For the input sentence to correct, we employ matching based methods (described in Sect. 3.1) to extract KB terms from the corresponding KBs. The correction model is expected to generate correct outputs based on both the input sentence and the extracted corresponding KB terms. Two main challenges of the task are that: 1) the KB terms extracted based on matching are not always helpful, and sometimes may be misleading, as shown in Fig. 1, and 2) the model has to decide whether to use the KB terms and how to leverage the KB terms for correction, while a large part of KB terms may be of low-frequency in the pre-training data and it is normally hard for models to take care of long-tail cases.

原始句子：我院以“筑牢中华民族共同体意识”工作为契机。
Original sentence: Our institute takes the work of building the consciousness of the Chinese nation community as an opportunity.
知识库条目(KB term)：铸牢中华民族共同体意识(forging the consciousness of the Chinese nation community)
纠错结果：我院以“铸牢中华民族共同体意识”工作为契机。✓
Corrected: Our institute takes the work of forging the consciousness of the Chinese nation community as an opportunity.
参考答案：我院以“铸牢中华民族共同体意识”工作为契机。
Reference: Our institute takes the work of forging the consciousness of the Chinese nation community as an opportunity.

原始句子：严格党的组织生活，抓好党员学习教育，不断提高党员素质。
Original sentence: Strictly enforce the Party's organizational life, focus on the study and education of Party members, and continuously improve the quality of Party members.
知识库条目(KB term)：党史学习教育(the study and education of Party history)
纠错结果：严格党的组织生活，抓好党史学习教育，不断提高党员素质。✗
Corrected: Strictly enforce the Party's organizational life, focus on the study and education of Party history, and continuously improve the quality of Party members.
参考答案：严格党的组织生活，抓好党员学习教育，不断提高党员素质。
Reference: Strictly enforce the Party's organizational life, focus on the study and education of Party members, and continuously improve the quality of Party members.

Fig. 1. Positive and negative examples for KB corrections.

3 Data Construction

To support the research on the task, 1) we construct knowledge bases of 10 governmental document relevant categories, 2) we build a synthetic dataset by extracting sentences which have at least one KB term matching from Chinese news corpora, and 3) we collect real user inputs, extract the sentences resulting in KB matches, and manually annotate correction results to build the development and test sets.

3.1 Knowledge Bases and Matching Methods

The constructed KBs encompass 10 categories: Event Information, Organization Information, Official Information, Geographic Information, Laws and Regulations, Policy/Spirit Names, Idioms, Fixed Phrases, Standard Expressions and Document Reference. The examples of each KB type are shown in Fig. 2, and the numbers of KB terms in each KB type are shown in Table 1.

We implement 4 matching based algorithms to extract KB terms for input sentences.

Event Information Matching Algorithm. The algorithm first finds all event names in the event information KB in the input sentence, and returns the found events and their corresponding times in the event information KB.

事件信息(Event Information)
(中国共产党第二十次全国代表大会,会议时间,2022年10月16日-2022年10月22日)
(The 20th National Congress of the Communist Party of China, Meeting Time, 2022.10.16-2022.10.22)

机构信息(Organization Information)
(中华人民共和国发展和改革委员会,发改委)
(National Development and Reform Commission of the People's Republic of China, National Development and Reform Commission)

领导人信息(Official Information)
(马朝旭,中华人民共和国外交部副部长、外交部党委委员)
(Ma Zhaoxu, Vice Minister of Foreign Affairs of the People's Republic of China、Member of the Party Committee of the Ministry of Foreign Affairs)

地域信息(Geographic Information)
(石家庄市,河北省)
(Shijiazhuang City, Hebei Province)

法律法规(Laws and Regulations)
中华人民共和国立法法
The Legislation Law of the People's Republic of China

政策精神名称(Policy/Spirit Names)
"两弹一星"精神
The Spirit of "Two Bombs, One Satellite"

成语(Idioms)
全力以赴
Spare no effort

固定用语(Fixed Phrases)
不忘初心、牢记使命
Stay true to our original aspiration and keep our mission firmly in mind.

规范表述(Standard Expressions)
党代会胜利召开
The Party Congress was successfully convened

文件引用(Document Reference)
(最高人民检察院是最高检察机关,领导地方各级人民检察院和专门人民检察院的工作,对全国人民代表大会和全国人民代表大会常务委员会负责并报告工作。)
(The Supreme People's Procuratorate is the highest procuratorial organ, leading the work of the people's procuratorates at all local levels and the specialized people's procuratorates. It is responsible to and reports work to the National People's Congress and its Standing Committee.)

Fig. 2. Knowledge base examples.

Official Information Matching Algorithm. The algorithm finds all position titles in the official information KB in the input sentence, and returns all found position titles and corresponding names in the official information KB.

Geographical Information Matching Algorithm. The algorithm finds the cities or counties in the geographical KB in the input sentence. For found counties, the algorithm tries to find the existence of any city or province name in the input sentence, and returns the city-county or province-county pairs in the geographical KB of the counties if a city/province name was found in the input sentence. For found cities, the algorithm searches all province names in the input sen-

Table 1. Statistics of knowledge bases.

KB types	Amount
Event Information	137
Organization Information	100
Official Information	4,054
Geographic Information	3,595
Laws and Regulations	1,178
Policy/Spirit Names	54
Idioms	9,341
Fixed Phrases	124
Standard Expressions	80
Document Reference	54,593

tence, and returns the province-city pairs in the geographical KB of the cities if a province name was found in the input sentence.

Matching based on Sequences' Similarity. For the KB elements inside the other KBs, we compute the sequences' similarity ratio using the python's difflib for the KB matching. We find the sub-span of the input sentence which leads to highest sequence similarity to the KB term, and returns the KB term if the similarity ratio is larger than a threshold.

Table 2. Amount of extracted sentences for training data synthesis for each KB type. A sentence may fall in more than one KB types at the same time.

KB type	Amount (number of sentences)
Event Information	15,279
Organization Information	3,005,692
Official Information	29,432
Geographic Information	835,391
Laws and Regulations	33,220
Policy Spirit Names	235,515
Idioms	2,536,394
Fixed Phrases	626,967
Standard Expressions	17,325
Document Reference	50,253

3.2 Training Set Synthesis

It is difficult to collect large-scale real text error correction datasets, especially when we require the input sentences for correction containing KB-related errors. However, the scale of the training set may be crucial for obtaining good performance with some machine learning approaches. So we provide a dataset for large-scale training set synthesis.

We first collect sentences from available online news. Specifically, we leverage the Chinese part of the newscrawl corpus [7], and crawl $522M$ deduplicated sentences from the government websites. We normalize the sentences with NFKC and convert traditional Chinese into simplified Chinese through OpenCC. Next, we use the KB matching algorithms (described in Sect. 3.1) to iterate the collected sentences. We keep the sentences with at least one KB match and augment the sentences with corresponding matched KB terms. As a result, we obtain around $644M$ sentences together with KB terms for data synthesis. The statistics over 10 KB types are as shown in the Table 2.

The extracted sentences can be categorized into two types: exact match or partial match. For exact match, the KB term directly appears in the sentence, and the data synthesis can be facilitated by introducing errors into the sentence through modifying the matched KB element, and train the model to correct the modified error sentence into the original sentence with the help of the extracted KB term. For partial match, we may assume the extracted sentences are normally correct, and train the model to re-generate the extracted sentence with both the sentence and matched KB terms as input. In this way, the model is trained to ignore surface matched but unrelated KB terms.

3.3 Validation and Test Data

To test the performance in real-world cases, we collect real user inputs for the construction of the validation and test sets. We only keep the sentences with matched KB terms, and manually annotate the correction results. The annotation team comprises native Chinese-speaking graduate students and Chinese linguistic experts working on Chinese government text correction. Each sentence is randomly assigned to 2 native Chinese-speaking graduate students and the linguistic experts decide the final correction result if the annotation results for a same sentence are different. The annotation agreement is 95.16%.

After sentence selection and annotation, we obtained 806 sentences, their corresponding matched KB terms and manually annotated correction results. The statistics w.r.t. KB types are as shown in Table 3. We randomly sampled 306 of them as the validation set and used the remained 500 instances for testing.

3.4 Evaluation Metrics

We evaluate with the traditional Precision (P), Recall (R) and F0.5 scores implemented by the ChERRANT toolkit. In addition, we also compute the Accuracy (Acc) and P/R/F0.5 scores only for the correction operations on KB terms to directly measure the correction performances regarding KBs.

Table 3. Statistics of annotated instances for validation and testing.

KB types	Amount
Event Information	11
Organization Information	198
Official Information	15
Geographic Information	43
Laws and Regulations	9
Policy/Spirit Names	72
Idioms	370
Fixed Phrases	115
Standard Expressions	104
Document Reference	56

知识库条目(KB term)：感激不尽(was endlessly grateful)

原始句子：赵某某对巡察组干部感激不已。
Original sentence: Zhao Jun was immensely grateful to the inspection team cadres.

大模型输入(模板)：
已知成语信息：感激不尽
请对句子中可能的错误进行纠正，只输出结果，不要输出其它内容：
赵军对巡察组干部感激不已。
LLM input(template):
Known idiom: was endlessly grateful
Please correct any possible errors in the sentence and only output the result without any additional content:
Zhao Jun was immensely grateful to the inspection team cadres.

大模型输出(纠错结果)：赵军对巡察组干部感激不尽。
LLM output(correction result): Zhao Jun was endlessly grateful to the inspection team cadres.

Fig. 3. An example for the LLM template.

4 LLM Baselines

We selected several representative LLMs as baselines for their exceptional language generation and comprehension capabilities, including DeepSeek, Qwen, GLM and Baichuan. We instruct the LLMs for text correction using the template in Fig. 3. Despite that we have prompted the LLMs to only generate the correction result, sometimes the model may still generate the other outputs, we extract the most similar sentence from the LLM outputs as the correction result.

We first tested the performance of LLMs of around 7B parameters, with and without KB terms in the prompt. Results in Table 4 show that: 1) the task is challenging for the tested LLMs, Qwen 3 8B with thinking enabled only achieves a highest KB F0.5 score of 12.38, and 2) providing KB terms generally leads to better performance than without KB terms.

Table 4. Results of LLMs without (w.o) or with (w) KB terms. The "-t" suffix indicates enabling thinking for Deepseek and Qwen.

LLM			ChERRANT			KB			
			P	R	F0.5	Acc	P	R	F0.5
w.o KB	Deepseek-R1	Qwen-7B-t	0.43	8.43	0.53	30.18	0.59	11.11	0.73
		Qwen-7B	0.55	9.64	0.67	31.49	0.64	11.11	0.79
	Qwen	2.5-7B	1.89	22.89	2.32	39.72	3.02	38.89	3.70
		3-8B-t	1.01	10.84	1.23	57.90	3.05	16.67	3.64
		3-8B	100.00	1.20	**5.75**	89.22	0.00	0.00	0.00
	GLM	4-9B	1.47	14.46	1.79	44.61	2.22	22.22	2.71
		4-9B-Z1	0.73	19.28	0.91	24.05	0.85	22.22	1.05
	Baichuan		2.23	14.46	2.68	66.24	7.09	27.78	**8.33**
w KB	Deepseek-R1	Qwen-7B-t	1.34	21.69	1.65	32.37	1.76	27.78	2.16
		Qwen-7B	1.53	24.10	1.88	33.02	2.69	41.67	3.31
	Qwen	2.5-7B	4.35	39.76	5.29	38.01	6.07	69.44	7.42
		3-8B-t	4.39	30.12	5.30	60.62	10.36	55.56	**12.38**
		3-8B	100.00	1.20	**5.75**	89.22	0.00	0.00	0.00
	GLM	4-9B	3.59	28.92	4.35	44.12	5.70	55.56	6.94
		4-9B-Z1	1.77	30.12	2.18	25.40	2.30	44.44	2.84
	Baichuan		3.68	16.87	4.37	59.91	6.82	33.33	8.11

We also tested the effects of LLM model sizes on performance with Deepseek and Qwen 2.5. Results in Table 5 show that larger models generally bring about higher F0.5 scores.

5 System Submissions

10 teams registered for the shared task, and 5 teams ultimately submitted the results.

CGT-Corrector. They adopt a computationally efficient approach by fine-tuning the Qwen2.5-14B-Instruct model on $20k$ representative samples carefully selected through multi-dimensional evaluation considering textual features, knowledge relevance and category distribution. By systematically introducing controlled noise patterns, they constructed an optimized training set.

KARTC. They propose a knowledge-aware error correction framework consisting of three modules. The Error Detection Module analyzes input texts to separately identify semantic errors and knowledge errors. The Candidate Sentence Generation Module, guided by error detection, generates multi-source candidate sentences in separate channels. The Over-Correction Mitigation Rewriting

Table 5. Results of LLMs with varying sizes.

LLM		ChERRANT			KB			
		P	R	F0.5	Acc	P	R	F0.5
Deepseek-R1	Qwen-1.5B-t	0.40	6.02	0.49	35.40	0.40	5.56	0.49
	Qwen-1.5B	0.61	9.64	0.75	35.57	0.78	11.11	0.96
	Qwen-7B-t	1.34	21.69	1.65	32.37	1.76	27.78	2.16
	Qwen-7B	1.53	24.10	1.88	33.02	2.69	41.67	3.31
Qwen	2.5-0.5B	1.34	15.66	1.64	46.15	2.71	25.00	3.30
	2.5-1.5B	3.36	30.12	4.09	41.78	5.79	61.11	7.07
	2.5-7B	4.35	39.76	**5.29**	38.01	6.07	69.44	**7.42**

Table 6. Results of submitted systems.

System		ChERRANT			KB			
		P	R	F0.5	Acc	P	R	F0.5
CGT-Corrector		47.76	22.38	**38.93**	84.20	46.94	41.82	**45.82**
KARTC		38.26	30.77	36.48	79.64	34.57	50.91	36.94
QEFDA	result1	13.08	11.89	12.82	77.72	17.02	14.55	16.46
	result2	17.19	15.38	16.79	74.14	15.28	20.00	16.03
	result3	8.29	12.59	8.90	65.93	6.90	14.55	7.71
CoT-LoRA		4.77	38.46	5.78	33.02	4.08	47.27	4.99
Sky		0.33	3.50	0.40	59.68	2.55	7.27	2.93

Module employs a knowledge-aware fusion rewriting mechanism based on the candidate set and original sentence, aiming at enhancing correction robustness and semantic retention while mitigating over-correction.

QEFDA. They employ Parameter-Efficient Fine-tuning (PEFT) with QLoRA to fine-tune Qwen-2.5-7B-Instruct on Lang8, HSK and the task-specific dataset which is formatted into error-correction pairs (incorrect-correct sentence pairs). To further improve generalization, they augment the task dataset through techniques including synonym replacement and random deletion.

CoT-LoRA. They extract data from the training set and reformat them into prompt-completion pairs for the LoRA fine-tuning of Qwen2.5-7B-Instruct and Yi1.5-7B-Instruct. To enhance the performance, they implement a voting mechanism to ensemble predictions from multiple fine-tuned models and select the optimal output through majority voting.

Results in Table 6 show that: 1) despite fine-tuning a larger model (Qwen2.5 14B Instruct), the data selection mechanism considering textual features, knowledge relevance and category distribution employed by CGT-Corrector seems very effective, and they obtains the highest F0.5 scores and largest improvements over

our baselines, 2) the collaborative framework proposed by KARTC also leads to large performance gains over our baselines, their method also obtains the highest recall, probably due to the use of an separate error detection module, and 3) QEFDA also achieves higher performance than our baselines, showing that fine-tuning simply on error-correction pairs and employing simple data augmentation strategies are beneficial for the task compared to instruct LLMs without fine-tuning.

6 Conclusion

In NLPCC 2025 shared task 5, we investigate the use of knowledge bases for Chinese government text correction. Specifically, we construct a number of KBs and corresponding methods to match potential KB terms in sentences. We extract a large dataset for training set synthesis, with each instance inside the dataset has resulted in at least one KB match. We also collect and annotate the development and test sets for the task based on real user inputs. For evaluation, we develop KB oriented metrics to evaluate the performance of text correction regarding KB terms. We report the performance of instructing several mainstream LLMs, with thinking enabled or disabled, and with various model sizes. The shared task has attracted registrations from 10 teams, with 5 teams ultimately submitted the final results. We report their methods and the performances of submitted systems. Data selection strategy and collaborative modeling have been proven effective in the shared task.

Acknowledgments. We appreciate our reviewer for the insightful comments and suggestions. This work is partially supported by the National Natural Science Foundation of China (Grant No. 62306284), China Postdoctoral Science Foundation (Grant No. 2023M743189), and the Natural Science Foundation of Henan Province (Grant No. 232300421386).

References

1. Afli, H., Qiu, Z., Way, A., Sheridan, P.: Using SMT for OCR error correction of historical texts. In: Calzolari, N., et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 962–966. European Language Resources Association (ELRA), Portorož, Slovenia (2016). https://aclanthology.org/L16-1153/
2. Cao, H., Yuan, L., Zhang, Y., Ng, H.T.: Unsupervised grammatical error correction rivaling supervised methods. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 3072–3088. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.185
3. Chang, H., et al.: Overview of CCL23-eval task: Chinese learner text correction. In: Sun, M., Qin, B., Qiu, X., Jiang, J., Han, X. (eds.) Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations), pp. 239–249. Chinese Information Processing Society of China, Harbin, China (2023). https://aclanthology.org/2023.ccl-3.27/

4. Gan, Z., Xu, H., Zan, H.: Self-supervised curriculum learning for spelling error correction. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.T. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3487–3494. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.281
5. Gao, J., Li, X., Micol, D., Quirk, C., Sun, X.: A large scale ranker-based system for search query spelling correction. In: Huang, C.R., Jurafsky, D. (eds.) Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 358–366. Coling 2010 Organizing Committee, Beijing, China (2010). https://aclanthology.org/C10-1041/
6. Gupta, H., Del Corro, L., Broscheit, S., Hoffart, J., Brenner, E.: Unsupervised multi-view post-OCR error correction with language models. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.T. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 8647–8652. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.680
7. Kocmi, T., et al.: Findings of the WMT24 general machine translation shared task: the LLM era is here but MT is not solved yet. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation, pp. 1–46. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.wmt-1.1
8. Li, W., Luo, W., Peng, G., Wang, H.: Explanation based in-context demonstrations retrieval for multilingual grammatical error correction. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 4881–4897. Association for Computational Linguistics, Albuquerque, New Mexico (2025). https://aclanthology.org/2025.naacl-long.251/
9. Li, W., Wang, H.: Detection-correction structure via general language model for grammatical error correction. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1748–1763. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.96
10. Liu, Y., Li, Z., Jiang, H., Zhang, B., Li, C., Zhang, J.: Towards better utilization of multi-reference training data for Chinese grammatical error correction. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024, pp. 3044–3052. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.findings-acl.180
11. Sun, C., She, L., Lu, X.: Two issues with Chinese spelling correction and a refinement solution. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 196–204. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-short.19
12. Wang, X., et al.: VisCGEC: Benchmarking the visual Chinese grammatical error correction. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5054–5068. Association for Computational Linguistics, Albuquerque, New Mexico (2025). https://aclanthology.org/2025.naacl-long.261/

13. Wang, X., Mu, L., Zhang, J., Xu, H.: Multi-pass decoding for grammatical error correction. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 9904–9916. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.553
14. Xu, L., Wu, J., Peng, J., Fu, J., Cai, M.: FCGEC: fine-grained corpus for Chinese grammatical error correction. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 1900–1918. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.137
15. Yang, Y., Wu, H., Zhao, H.: Attack named entity recognition by entity boundary interference. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 1734–1744. ELRA and ICCL, Torino, Italia (2024). https://aclanthology.org/2024.lrec-main.153/
16. Yin, X., Wan, X., Zhang, D., Yu, L., Yu, L.: Overview of the NLPCC 2023 shared task: Chinese spelling check. In: Natural Language Processing and Chinese Computing (2023). https://doi.org/10.1007/978-3-031-44699-3_30
17. Zhang, Y., et al.: MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3118–3130. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.naacl-main.227
18. Zhang, Y., et al.: Nasgec: a multi-domain Chinese grammatical error correction dataset from native speaker texts. arXiv:2305.16023 (2023)
19. Zhao, Y., Jiang, N., Sun, W., Wan, X.: Overview of the NLPCC 2018 shared task: grammatical error correction. In: Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2018. LNCS (LNAI), vol. 11109, pp. 439–445. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99501-4_41