Evolving Chinese Spelling Correction with Corrector-Verifier Collaboration

Linfeng Liu[1,2*], **Hongqiu Wu**[1*], **Hai Zhao**[1],
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University,
[2]SJTU Paris Elite Institute of Technology
Correspondence: zhaohai@cs.sjtu.edu.cn

Abstract

Recent methods address Chinese Spelling Correction (CSC) with either BERT-based models or large language models (LLMs) independently. However, both of them face challenges. BERT-based models are efficient for this task but struggle with limited generalizability to error patterns, thus failing in open-domain CSC. LLMs are advantageous in their extensive knowledge but fall into low efficiency in character-level editing. To address this dilemma, we propose *Automatic Corrector Iteration (ACI)*, a novel model collaboration pipeline to iteratively optimize a BERT-based corrector. This pipeline is free of human annotation, by leveraging the knowledge and reasoning ability of an LLM verifier to provide useful signals for the corrector. Experimental results demonstrate that our pipeline consistently improves the model performance across iterations and significantly outperforms existing data augmentation methods, achieving comparable performance with human annotation.

1 Introduction

Chinese Spelling Correction (CSC) aims at correcting erroneous characters in Chinese sentences (Yu and Li, 2014; Xiong et al., 2015). A recent line of work develops large language models (LLMs) for CSC (Li et al., 2024; Zhou et al., 2024) while some others continue to elaborate BERT-based models (Wu et al., 2023; Hu et al., 2024; Liu et al., 2024; Zhu et al., 2022; Sheng and Xu, 2024).

These works reveal that both BERT-based models and LLMs exhibit distinct advantages and limitations in addressing CSC. BERT-based corrector naturally adapts CSC task with its masked language modeling and sequence tagging character, **effectively handling phonological and visual similarity errors** (Liu et al., 2025). However, the scarcity of high-quality and real-world training data is a big issue. These models suffer from biased error patterns learned on synthetic data, leading to **over-correction issues and inadequate handling of semantic errors** (Liu et al., 2025; Wu et al., 2023; Hu et al., 2024; Liu et al., 2022; Jiang et al., 2024). While LLMs demonstrate significant advantages in generating semantically coherent text and leveraging knowledge, their effectiveness in CSC has not substantially surpassed BERT-based models (Zhou et al., 2024; Zhang et al., 2023; Li et al., 2023a). This mainly attributes to the autoregressive nature of LLMs, which constrains their ability to capture character-level mappings between the original sentence and correction, leading to **challenges in addressing phonological errors and maintaining length consistency** of the output. Additionally, LLM's high computational costs and latency restrict its large-scale application on CSC.

To address these challenges, we propose Automatic Corrector Iteration (ACI), an iterative corrector optimization pipeline using a BERT-based model as corrector and LLM as verifier. ACI leverages the **complementary strengths of LLM and BERT-based corrector** to tackle open-domain CSC. In each iteration, the BERT-based corrector identifies and corrects potential errors in monolingual data. An LLM then verifies the corrections and provides alternative suggestions when needed. The generated parallel data is subsequently used to train the corrector itself, forming a self-evolving cycle.

ACI has several advantages compared to previous data augmentation methods. (1) Compared to synthetic errors generated by rules, ACI seeks to mine the real-world spelling errors from the corpus, preventing the model from learning biased error patterns. Furthermore, ACI offers the false positive samples identified by the verifier to mitigate the over-correction issue. (2) Our method uses BERT-based models to correct and LLM to verify, leveraging LLM's extensive knowledge of

[*] These authors contributed equally.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29022–29028
November 4-9, 2025 ©2025 Association for Computational Linguistics

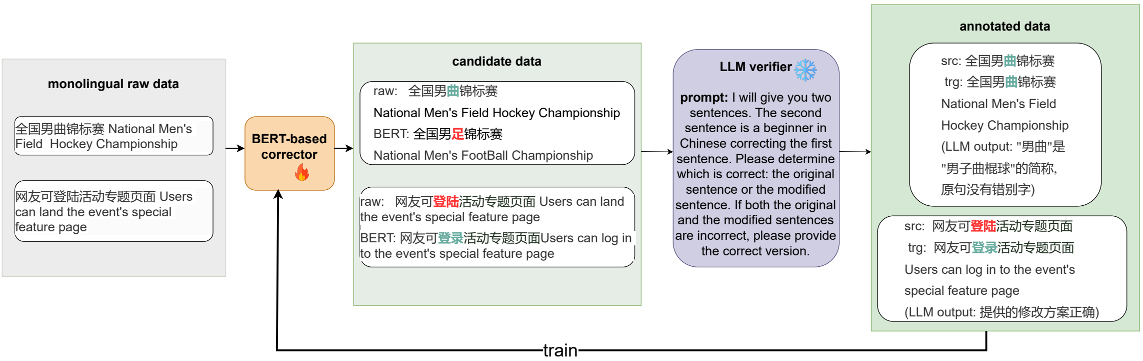
monolingual raw data
全国男曲锦标赛 National Men's Field Hockey Championship
网友可登陆活动专题页面 Users can land the event's special feature page
BERT-based corrector
candidate data
raw: 全国男曲锦标赛
National Men's Field Hockey Championship
BERT: 全国男足锦标赛
National Men's FootBall Championship
raw: 网友可登陆活动专题页面 Users can land the event's special feature page
BERT: 网友可登录活动专题页面Users can log in to the event's special feature page
LLM verifier
prompt: I will give you two sentences. The second sentence is a beginner in Chinese correcting the first sentence. Please determine which is correct: the original sentence or the modified sentence. If both the original and the modified sentences are incorrect, please provide the correct version.
annotated data
src: 全国男曲锦标赛
trg: 全国男曲锦标赛
National Men's Field Hockey Championship
(LLM output: "男曲"是"男子曲棍球"的简称, 原句没有错别字)
src: 网友可登陆活动专题页面
trg: 网友可登录活动专题页面
Users can log in to the event's special feature page
(LLM output: 提供的修改方案正确)
train

Figure 1: ACI pipeline. A BERT-based corrector recalls the candidate sentences, which are verified by an LLM.

changeable Chinese expressions with various styles and vast named entities while circumventing the limitations of autoregressive models in character-level mapping. (3) ACI is totally free of human annotation. Our empirical results show that the model performance can scale with increasing data volume.

2 Automatic Corrector Iteration

ACI collaborates two models, a BERT-based corrector, e.g. ReLM (Liu et al., 2024) and an LLM verifier, e.g. Qwen (Yang et al., 2024)). Figure 1 illustrates the ACI pipeline with its four key steps.

Preprocess monolingual data The input for an ACI iteration is a monolingual corpus. We first segment the corpus into sentences and filter out unnecessary sentences, e.g. ones containing too many non-Chinese characters.

Recall We then send the preprocessed sentences to the BERT-based corrector batch by batch. The corrector detects and corrects the potential errors in them. We then recall the sentences that are edited by the corrector as the candidate sentences. The sentences where there are no errors identified are excluded.

Verify The candidate sentences are verified by the LLM verifier whether the candidate is better compared to the raw sentence. There are three situations: (1) If LLM thinks the original sentence is better, the original sentence will be preserved as ground truth, serving as false positive samples to prevent over-correction; (2) If LLM thinks the candidate correction is better, the LLM will confirm and retain this correction; (3) For cases where both the original and corrected versions are considered as incorrect by the LLM, the LLM provides alternative corrections.

Update The verification results yield training data containing real-world spelling errors, which is used to train the corrector. This update enhances both the performance and generalizability of the corrector. The next iteration then proceeds with the updated corrector recalling candidate sentences from a new monolingual corpus.

3 Experimental Results

3.1 Experimental Setup

Dataset A line of studies has reported issues with SIGHAN (Tseng et al., 2015), such as annotation error and incoherent style with native speakers (Wu et al., 2023; Hu et al., 2024; Li et al., 2024). Following recent works, we use two CSC benchmarks, LEMON (Wu et al., 2023) and CSCD-NS (Hu et al., 2024). (1) LEMON is a large-scale multi-domain CSC dataset. (2) CSCD-NS superior in annotation quality and focus on spelling errors stemming from pinyin input methods.

Corrector Models ACI is agnostic to the type of the corrector. We evaluate it using three different BERT-based models as the corrector. Following (Wu et al., 2023), all three models are pre-trained on 34M synthetic data using confusion set.

• **BERT** Following (Devlin et al., 2019), we fine-tune the BERT model as sequence tagging to perform CSC.

• **ReLM** Liu et al. (2024) regards CSC as sentence rephrasing. The correction is made on top of the entire semantics. ReLM is a non-autoregressive language model.

• **MDCSpell** Zhu et al. (2022) design a paralleled detector-corrector network to enhance the correction. The new detector network is initialized using another BERT encoder.

		GAM	CAR	NOV	ENC	NEW	COT	MEC	CSCD-NS	avg.
BERT	*pretrained*	32.8	52.0	35.8	45.2	56.0	63.7	50.7	49.4	48.2
	synthetic	31.9	53.5	35.0	50.6	58.5	64.8	55.1	62.2	51.5
	synthetic+human	47.3	**60.9**	43.3	61.5	**64.1**	68.8	59.3	**77.1**	**60.3**
	LLM-annotator	32.4	51.4	40.8	56.9	48.4	68.9	55.2	56.0	51.3
	ACI-1	36.0	55.4	40.0	53.3	58.2	66.0	54.3	55.5	52.4
	ACI-2	46.4	59.6	44.0	61.1	62.1	69.4	60.5	67.9	58.9
	ACI-3	**47.4**	59.7	**45.4**	**62.2**	63.2	**70.8**	**66.5**	65.5	60.1
ReLM	*pretrained*	34.6	53.6	38.0	47.6	58.8	67.7	53.8	44.4	49.7
	synthetic	38.2	54.6	37.1	53.1	59.5	66.9	57.8	61.9	53.6
	synthetic+human	50.4	**61.2**	43.7	61.1	**64.8**	68.2	58.9	**77.4**	60.7
	LLM-annotator	38.2	53.4	37.2	56.4	53.1	67.8	53.2	48.2	50.9
	ACI-1	38.0	56.7	39.5	53.4	59.1	67.7	57.0	51.5	52.9
	ACI-2	**52.4**	58.3	43.1	62.1	63.1	68.8	61.4	68.7	59.7
	ACI-3	50.5	60.4	**45.5**	**63.4**	63.4	**70.9**	**66.1**	69.1	**61.2**
MDCSpell	*pretrained*	31.4	51.9	37.4	46.1	57.5	64.8	52.9	51.2	49.1
	synthetic	30.5	52.7	36.4	52.1	58.1	64.7	55.5	62.0	51.5
	synthetic+human	**50.7**	**61.2**	44.1	61.9	**65.6**	69.6	60.5	**77.0**	**61.3**
	LLM-annotator	33.7	53.7	38.5	56.5	52.2	65.8	54.6	56.5	51.4
	ACI-1	37.1	56.0	41.5	54.0	59.2	69.0	57.1	56.8	53.8
	ACI-2	50.1	58.5	42.7	61.8	62.7	71.4	63.0	67.4	59.7
	ACI-3	50.1	57.9	**44.4**	**62.0**	63.8	**72.6**	**65.0**	64.6	60.1

Table 1: Performance of different data engineering methods. ACI-1 signifies the first iteration of the ACI pipeline.

ACI Settings We use Qwen2-72b (Yang et al., 2024) as the verifier. We iterate the ACI pipeline for three times using three public Chinese corpora: *thucnews*[*], LCSTS (Hu et al., 2015), and *baike2018qa*[†], which are all without annotation. We train the BERT-based corrector using batch size 512 and learning rate 1e-5 for 10k steps. We use the LEMON development set for validation and the sentence-level F1 score as the metric.

Baselines We compare ACI with two data engineering methods to train CSC models.

- **IME-based synthetic + Human annotated** The two-stage training of first using synthetic and then using human-annotated data is the widely-used and the most useful method. We generate the synthetic data using IME (Hu et al., 2024). This method improves the quality of the synthetic data compared to traditional using the confusion set. We first train the model on IME-based synthetic data and then train it on human annotated data. The data we use is the training set of CSCD-NS, which is in high quality.
- **LLM as Annotator** We first use BERT-based corrector to recall potentially erroneous sentences and then use Qwen2-72b (Yang et al., 2024) to directly correct the recalled sentences. We integrate 3 in-context learning samples into the prompt: *Please correct the spelling mistakes in the sentence, ensuring that the modified sentence has the same number of characters as the original. Note that only typos need to be replaced, and please do not rephrase or rewrite the sentence.* The corpus we use is *thucnews*.

Synthetic	***Human***	***LLM***	**ACI-1**	**ACI-2**	**ACI-3**
2.02M	30k	87k	100k	110k	180k

Table 2: Statistics of training data for ACI and baselines. "LLM" refers to the LLM as Annotator method. ACI-x refers to a specific iteration.

The numbers of data used for training in each iteration of ACI and the baselines are in Table 2.

3.2 Main Results

Table 1 shows that ACI outperforms the straightforward LLM-based annotation in the first iteration across all three BERT-based models. This superior performance can be attributed to the higher quality training data generated by ACI compared to direct LLM annotation. By employing LLM as a validator for BERT-based correction results rather than for direct correction, ACI mitigates the negative impact of LLM's autoregressive nature.

However, Table 1 shows that while ACI and *synthetic+human* show comparable performance across various domains in LEMON, *synthetic+human* exhibits better performance on CSCD-NS. This performance gap can be attributed to the domain alignment between the human-

[*] http://thuctc.thunlp.org/

[†] https://github.com/brightmart/nlp_chinese_corpus

	GAM	CAR	NOV	ENC	NEW	COT	MEC	avg.
ReLM-ACI	**50.5**	**60.4**	45.5	**63.4**	**63.4**	**70.9**	66.1	**60.0**
BERT-ACI	47.4	59.7	45.4	62.2	63.2	70.8	**66.5**	59.3
GPT4-10shot	36.3	54.4	**45.6**	55.1	56.1	62.8	56.3	52.3
Qwen2-72b-5shot	45.5	/	/	48.3	55.3	/	/	/
Qwen1.5-14b finetuned	38.0	57.5	43.9	56.4	64.4	60.4	65.3	55.1

Table 3: Performance of different data engineering methods. ACI-1 signifies the first iteration of the ACI pipeline.

	LLM	BERT	ACI (72b)	ACI (7b)
iteration-1	69.1	57.8	**69.9**	65.1
iteration-2	69.8	64.0	**71.8**	65.1
iteration-3	69.6	70.7	**73.6**	68.4

Table 4: The accuracy of the recalled data. **LLM**: direct annotation with Qwen2-72b; **BERT**: results recalled by BERT in ACI's first step; **ACI(72b)** and **ACI(7b)**: the final annotation results of ACI with Qwen2-72b and Qwen2-7b respectively.

annotated training data and CSCD-NS test data, suggesting the potential benefit of incorporating human-annotated and domain-specific data in certain scenarios.

Meanwhile, for open-domain CSC, we show in Table 3 that BERT-based methods trained with ACI significantly outperform LLMs utilizing either in-context learning or fine-tuning. In Table 3 *Qwen1.5-14b finetuned* combines the human annotated data from CSCD-NS and 271K pseudo-data generated by ASR or OCR as the training data, and uses character-level tokenization.

3.3 Annotation Accuracy

The verification quality is a crucial factor of ACI. To evaluate the quality of generated training data, we probe the annotation accuracy of different approaches on CSCD-NS development set. For ACI pipeline, we analyze two key metrics: the accuracy of BERT-recalled candidates and the accuracy of final LLM-verified results. We compare the annotation results with the gold labels from CSCD-NS.

As shown in Table 4, ACI with Qwen2-72b achieves consistently higher accuracy compared to ACI with Qwen2-7b across all iterations. This substantial performance difference validates the necessity and of utilizing the larger 72b model. Moreover, ACI demonstrates superior accuracy compared to direct annotation, reaching 73.6 in iteration-3 versus 69.6 for direct annotation, which further corroborates our previous findings that the ACI pipeline generates higher-quality training data than direct LLM annotation.

Interestingly, Table 4 reveals that both the accuracy of BERT-recalled corrections and ACI-generated training data improve consistently across iterations. The accuracy gap between these two stages gradually narrows from 12.1% in the first iteration to 2.9% in the third. This convergence explains the diminishing performance gains observed in Table 1, where the improvement in CSC performance becomes less pronounced in the third iteration.

4 Related Works

Existing studies tackle CSC either with BERT-based models or LLMs independently. The BERT-based models focus on employing features of Chinese, e.g. phonological similarity (Liu et al., 2021; Huang et al., 2021; Sun et al., 2023; Liang et al., 2023), or disentangling the detection and correction module (Zhang et al., 2020). A line of works also propose different data augmentation methods to construct pseudo data to address the scarcity of CSC data (Wang et al., 2018; Hu et al., 2024; Sheng and Xu, 2024). LLM-based methods focus on adapting the LLMs better for CSC by adjusting the tokenizer (Li et al., 2024), introducing a minimal distortion model (Zhou et al., 2024). Our work differs from these by iteratively using the LLM's knowledge to refine the BERT-based model. Compared to recent studies on leveraging LLMs for data annotation and small model enhancement (Chen and Varoquaux, 2024; Tan et al., 2024), where the focus has been on extracting knowledge and rationales from LLMs to improve learner performance (Chung et al., 2023; Li et al., 2023b). However, our approach is tailored for the CSC task by introducing a novel iterative pipeline, leveraging the complementary strengths of BERT and LLM. Instead of direct LLM annotation, we employ BERT-based models for initial error detection and correction, followed by LLM validation and feedback.

5 Conclusion

In this paper, we propose ACI, an iterative and human-annotation-free training pipeline for CSC. ACI cooperate BERT-based corrector and LLM to iteratively generate training data and optimize the BERT-based corrector, leveraging the complementary strength of BERT's sequence tagging feature and LLM's extensive knowledge and text generation capacity. Experiments demonstrate that ACI can improve with iterations and significantly outperforms existing data augmentation approaches, achieving comparable performance with models trained on human annotated data.

6 Limitations

This paper employs Qwen2-72b as a verifier in the ACI pipeline. Although effective, the high computational cost of such a large model may limit the iteration efficiency. Future work could explore fine-tuning smaller LLMs as alternative verifiers to improve the pipeline's efficiency while maintaining its effectiveness. Additionally, the ACI pipeline could be further enhanced by incorporating effective mechanisms from recent agent research, such as reflection and voting mechanisms. These mechanisms have shown promising results in improving decision quality and could potentially boost the accuracy of the generated training data in each iteration.

References

Lihu Chen and Gaël Varoquaux. 2024. What is the role of small models in the LLM era: A survey. *CoRR*, abs/2409.06857.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 575–593. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1967–1972. The Association for Computational Linguistics.

Yong Hu, Fandong Meng, and Jie Zhou. 2024. CSCD-NS: a chinese spelling check dataset for native speakers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 146–159. Association for Computational Linguistics.

Li Huang, Junjie Li, Weiwei Jiang, Zhiyu Zhang, Minchuan Chen, Shaojun Wang, and Jing Xiao. 2021. Phmospell: Phonological and morphological knowledge guided chinese spelling check. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5958–5967. Association for Computational Linguistics.

Lai Jiang, Hongqiu Wu, Hai Zhao, and Min Zhang. 2024. Chinese spelling corrector is just a language learner. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6933–6943. Association for Computational Linguistics.

Kunting Li, Yong Hu, Liang He, Fandong Meng, and Jie Zhou. 2024. C-LLM: learn to check chinese spelling errors character by character. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5944–5957. Association for Computational Linguistics.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023a. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10443–10461. Association for Computational Linguistics.

Zihong Liang, Xiaojun Quan, and Qifan Wang. 2023. Disentangled phonetic representation for chinese spelling correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13509–13521. Association for Computational Linguistics.

Linfeng Liu, Hongqiu Wu, and Hai Zhao. 2024. Chinese spelling correction as rephrasing language model. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18662–18670. AAAI Press.

Linfeng Liu, Hongqiu Wu, and Hai Zhao. 2025. Driving chinese spelling correction from a fine-grained perspective. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10727–10737. Association for Computational Linguistics.

Shulin Liu, Shengkang Song, Tianchi Yue, Tao Yang, Huihui Cai, Tinghao Yu, and Shengli Sun. 2022. Craspell: A contextual typo robust approach to improve chinese spelling correction. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3008–3018. Association for Computational Linguistics.

Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. 2021. PLOME: pre-training with misspelled knowledge for chinese spelling correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2991–3000. Association for Computational Linguistics.

Lei Sheng and Shuai-Shuai Xu. 2024. Edacsc: Two easy data augmentation methods for chinese spelling correction. *CoRR*, abs/2409.05105.

Rui Sun, Xiuyu Wu, and Yunfang Wu. 2023. An error-guided correction model for chinese spelling error correction. *CoRR*, abs/2301.06323.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 930–957. Association for Computational Linguistics.

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, SIGHAN@IJCNLP 2015, Beijing, China, July 30-31, 2015*, pages 32–37. Association for Computational Linguistics.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2517–2527. Association for Computational Linguistics.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for chinese spelling correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10743–10756. Association for Computational Linguistics.

Jinhua Xiong, Qiao Zhang, Shuiyuan Zhang, Jianpeng Hou, and Xueqi Cheng. 2015. Hanspeller: A unified framework for chinese spelling correction. *Int. J. Comput. Linguistics Chin. Lang. Process.*, 20(1).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, Wuhan, China, October 20-21, 2014*, pages 220–223. Association for Computational Linguistics.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 882–890. Association for Computational Linguistics.

Xiaowu Zhang, Xiaotian Zhang, Cheng Yang, Hang Yan, and Xipeng Qiu. 2023. Does correction remain A problem for large language models? *CoRR*, abs/2308.01776.

Houquan Zhou, Zhenghua Li, Bo Zhang, Chen Li, Shaopeng Lai, Ji Zhang, Fei Huang, and Min Zhang. 2024. A simple yet effective training-free prompt-free approach to chinese spelling correction based on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA,* *November 12-16, 2024*, pages 17446–17467. Association for Computational Linguistics.

Chenxi Zhu, Ziqiang Ying, Boyu Zhang, and Feng Mao. 2022. Mdcspell: A multi-task detector-corrector framework for chinese spelling correction. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1244–1253. Association for Computational Linguistics.