Two Issues with Chinese Spelling Correction and A Refinement Solution

Changxuan Sun **Linlin She** **Xuesong Lu***
School of Data Science and Engineering
East China Normal University
{changxuansun@stu,linlinshe123@stu,xslu@dase}.ecnu.edu.cn

Abstract

The Chinese Spelling Correction (CSC) task aims to detect and correct misspelled characters in Chinese text, and has received lots of attention in the past few years. Most recent studies adopt a Transformer-based model and leverage different features of characters such as pronunciation, glyph and contextual information to enhance the model's ability to complete the task. Despite their state-of-the-art performance, we observe two issues that should be addressed to further advance the CSC task. First, the widely-used benchmark datasets SIGHAN13, SIGHAN14 and SIGHAN15, contain many mistakes. Hence the performance of existing models is not accurate and should be re-evaluated. Second, existing models seem to have reached a performance bottleneck, where the improvements on the SIGHAN's testing sets are increasingly smaller and unstable. To deal with the two issues, we make two contributions: (1) we manually fix the SIGHAN datasets and re-evaluate four representative CSC models using the fixed datasets; (2) we analyze the new results to identify the spelling errors that none of the four models successfully corrects, based on which we propose a simple yet effective refinement solution. Experimental results show that our solution improves the four models in all metrics by notable margins.

1 Introduction

Chinese Spelling Correction (CSC) aims to detect and correct misspelled characters in Chinese text. The task is challenging yet important, being used in various NLP applications such as search engines (Martins and Silva, 2004), optical character recognition (Afli et al., 2016) and international Chinese education (Liu et al., 2011). To solve the task, recent studies have employed Transformer (Vaswani et al., 2017) or BERT (Kenton and Toutanova, 2019) as the base model and incorporated rich semantic features of characters to promote performance (Cheng et al., 2020; Liu et al., 2021; Xu et al., 2021; Li et al., 2022a; Liu et al., 2022; Liang et al., 2023; Huang et al., 2023).

Despite the promising results, we observe two issues with the current research for CSC. First, the widely-used benchmark datasets, SIGHAN13 (Wu et al., 2013), SIGHAN14 (Yu et al., 2014) and SIGHAN15 (Tseng et al., 2015), contain many mistakes, most of which are the meaningless sentences and the spelling errors in the target sentences. The former are the common mistakes made by Chinese beginners, as the SIGHAN datasets are collected from the Chinese essay section of Test for foreigners. These mistakes make the meaning of the sentences unclear and may affect the correction of spelling errors. The latter are the spelling errors that were not identified by the Chinese teachers in the test. Specifically, it is known that SIGHAN13 contains many misuses of "的", "地" and "得" in the target sentences. These mistakes definitely affect the accuracy of the evaluation results. Surprisingly, previous studies have never attempted to fix the mistakes to better evaluate their models. Second, recent models seem to have reached a performance bottleneck on the SIGHAN's testing sets, as evidenced by the increasingly smaller and unstable improvements (i.e., a newly proposed model does not perform better in all metrics) in the evaluation metrics. For instance, SCOPE (Li et al., 2022a) performs worse than MLM-phonetics (Zhang et al., 2021) in detection recall and correction recall on SIGHAN14 and performs worse than REALISE (Xu et al., 2021) in detection precision and correction precision on SIGHAN15. Furthermore, SCOPE combined with DR-CSC (Huang et al., 2023) improves SCOPE by only around 1 point in all metrics and also performs worse than comparative models in several metrics on SIGHAN13 and SIGHAN14. While

* Xuesong Lu is the corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 196–204

these models are focused on different aspects of spelling errors, we speculate the reason is that there exist certain errors which none of them can stably correct.

To tackle the two issues, we make two contributions in this paper. First, we examine the SIGHAN datasets sentence by sentence and fix all possible mistakes. Then, we retrain four representative CSC models using the fixed datasets and re-evaluate their performance. Second, we analyze the evaluation results and identify the spelling errors that none of the models successfully corrects, based on which we propose a simple solution to refine the output of the models without training. Experimental results show that our simple solution improves the four models in all metrics by notable margins.

2 Fixing SIGHAN and Re-evaluating Four Models

Type 1: meaningless sentences	
Original	连忙我都没有时间跟父母见面！ Quickly I don't even have time to meet my parents!
Fixed	忙得我都没有时间跟父母见面！ I'm so busy that I don't even have time to meet my parents!
Type 2: spelling errors in target sentences	
Original	很多伪心的路，在我们面前挥手。 Many roads of false hearts wave in front of us.
Fixed	很多违心的路，在我们面前挥手。 Many roads against our will wave in front of us.
Type 3: unconverted traditional Chinese characters	
Original	一张又一张地念著， Read one page after another,
Fixed	一张又一张地念着， Read one page after another,

Table 1: Some examples of different mistake types and the corresponding fixes.

Two authors of the paper independently examine the SIGHAN datasets and identify the sentences with mistakes. Then they review each identified sentence and discuss whether it should be fixed and how to fix it. To ensure the accuracy of fixing, we fix the datasets in two rounds and both rounds take the same steps. First, we examine the fluency of the sentences and identify those that are meaningless. In this case, both a source sentence and the corresponding target sentence need to be fixed, and the spelling errors remain unchanged. Second, we identify the spelling errors in the target sentences. Third, we identify the traditional Chinese characters that are not converted into simplified ones by OpenCC[1] in both source and target sentences. Table 1 shows the example sentences with mistakes and the corresponding fixes. More examples are presented in Table 6 of the appendix.

Table 2 shows the statistics of fixes for the three datasets as well as the original statistics. The numbers in the parentheses are the numbers of sentences with spelling errors. Note that the rows indicated by "Fixed" show the statistics for the fixed sentences only. We observe that all three datasets have a considerable number of lines[2] fixed, with many spelling errors including the newly-identified errors indicated in the square brackets. Note that a new spelling error is identified when a spelling error in a target sentence is fixed. That is, the numbers in the square brackets are the numbers of spelling errors in the target sentences of the original SIGHAN datasets.

Training Data		#Lines	avgLength	#Errors
SIGHAN13	Original	700 (340)	41.8	343
	Fixed	247 (117)	44.5	234 [114]
SIGHAN14	Original	3437 (3358)	49.6	5122
	Fixed	1280 (1197)	55.1	2360 [273]
SIGHAN15	Original	2338 (2273)	31.3	3037
	Fixed	675 (634)	36.9	1113 [172]
Testing Data		**#Lines**	**avgLen**	**#Errors**
SIGHAN13	Original	1000 (966)	74.3	1224
	Fixed	569 (551)	79.1	1149 [407]
SIGHAN14	Original	1062 (551)	50.0	771
	Fixed	442 (305)	55.3	538 [147]
SIGHAN15	Original	1100 (569)	30.6	703
	Fixed	357 (229)	35.1	337 [67]

Table 2: Summary statistics of the original datasets and the fixed parts.

Then, we select four representative CSC models and re-evaluate them on the fixed datasets, namely, PLOME (Liu et al., 2021), REALISE (Xu et al., 2021), LEAD (Li et al., 2022b) and SCOPE (Li et al., 2022a). The four models generally have the strongest performance among existing models according to the literature, and the authors have released the source code[3] that are easily run. For each

[1] https://github.com/BYVoid/, Apache License 2.0.

[2] A line consists of a source sentence and a target sentence.

[3] PLOME: https://github.com/liushulinle/PLOME, REALISE: https://github.com/DaDaMrX/ReaLiSe, LEAD: https://github.com/geekjuruo/LEAD, SCOPE: https://github.com/jiahaozhenbang/SCOPE

Datasets & Models		Detection			Correction		
SIGHAN13		D-P	D-R	D-F	C-P	C-R	C-F
Original	PLOME	81.3	77.9	79.6	79.6	76.3	77.9
	REALISE*	88.6	82.5	85.4	87.2	81.2	84.1
	LEAD*	88.3	83.4	85.8	87.2	82.4	84.7
	SCOPE*	87.4	83.4	85.4	86.3	82.4	84.3
Retrained	PLOME	76.7	74.5	75.5	75.0	72.9	73.9
	REALISE	77.6	73.9	75.7	76.4	72.8	74.5
	LEAD	78.0	74.6	76.3	76.4	73.0	74.4
	SCOPE	65.4	61.9	63.6	63.6	60.2	61.9
Refined	PLOME	**79.9** (↑3.2)	**78.1** (↑3.6)	**79.0** (↑3.5)	**78.0** (↑3.0)	**76.2** (↑3.3)	**77.1** (↑3.2)
	REALISE	**80.6** (↑3.0)	**77.5** (↑3.6)	**79.0** (↑3.3)	**79.4** (↑3.0)	**76.3** (↑3.5)	**77.8** (↑3.3)
	LEAD	**81.5** (↑3.5)	**78.4** (↑3.8)	**79.9** (↑3.6)	**79.9** (↑3.5)	**76.8** (↑3.8)	**78.3** (↑3.9)
	SCOPE	**75.9** (↑10.5)	**74.0** (↑12.1)	**75.0** (↑11.4)	**73.9** (↑10.3)	**72.0** (↑11.8)	**72.9** (↑11.0)

Table 3: The results on SIGHAN13. The asterisk * indicates the results are copied from the original paper.

Datasets & Models		Detection			Correction		
SIGHAN14		D-P	D-R	D-F	C-P	C-R	C-F
Original	PLOME	73.5	70.0	71.7	71.5	68.0	69.7
	REALISE*	67.8	71.5	69.6	66.3	70.0	68.1
	LEAD*	70.7	71.0	70.8	69.3	69.6	69.5
	SCOPE*	70.1	73.1	71.6	68.6	71.5	70.1
Retrained	PLOME	70.0	67.5	68.7	67.5	65.2	66.3
	REALISE	74.4	67.7	70.9	72.2	65.7	68.8
	LEAD	76.6	70.0	73.1	74.7	68.3	71.4
	SCOPE	82.4	77.2	79.7	80.8	75.7	78.1
Refined	PLOME	**71.6** (↑1.6)	**69.5** (↑2.0)	**70.5** (↑1.8)	**69.5** (↑2.0)	**67.5** (↑2.3)	**68.5** (↑2.2)
	REALISE	**76.4** (↑2.0)	**70.3** (↑2.6)	**73.2** (↑2.3)	**74.5** (↑2.3)	**68.6** (↑2.9)	**71.4** (↑2.6)
	LEAD	**77.9** (↑1.3)	**72.2** (↑2.2)	**75.0** (↑1.9)	**76.5** (↑1.8)	**70.9** (↑2.6)	**73.6** (↑2.2)
	SCOPE	**83.5** (↑1.1)	**79.0** (↑1.8)	**81.2** (↑1.5)	**81.9** (↑1.1)	**77.7** (↑2.0)	**79.7** (↑1.6)

Table 4: The results on SIGHAN14. The asterisk * indicates the results are copied from the original paper.

model, we adopt the training settings in the original paper. We train each model four times with random seeds and report the average results on the testing sets. We use the widely-adopted sentence-level precision, recall and F1 (Wang et al., 2019) to evaluate the models, which are also used in their original papers. The evaluation is conducted on detection and correction sub-tasks. The results are reported in Table 3, 4 and 5, where the rows indicated by "Original" are the results on the original SIGHAN datasets, and the rows indicated by "Retrained" are the results of the models retrained using the fixed SIGHAN datasets. The "Original" results are all copied from the corresponding papers except for PLOME on SIGHAN13 and SIGHAN14. The authors have not reported the results which we have to reproduce.

Comparing the results of "Original" and "Retrained", we observe that the results are largely changed. On SIGHAN13, all results decrease drastically. This is mainly because the "Original" results are calculated after excluding "的", "地" and "得", since the targets are almost not correct, whereas the "Retrained" results are calculated on all spelling errors. This indicates the models can still not correct "的", "地" and "得" well, especially for SCOPE which has the largest performance drop. On SIGHAN14 and SIGHAN15, the results generally increase after the datasets are fixed. Based on the results, we suggest to use the fixed datasets for more accurate evaluation in the future.

Datasets & Models		Detection			Correction		
SIGHAN15		D-P	D-R	D-F	C-P	C-R	C-F
Original	PLOME*	77.4	81.5	79.4	75.3	79.3	77.2
	REALISE*	77.3	81.3	79.3	75.9	79.9	77.8
	LEAD*	79.2	82.8	80.9	77.6	81.2	79.3
	SCOPE*	81.1	84.3	82.7	79.2	82.3	80.7
Retrained	PLOME	77.7	78.9	78.3	75.6	76.8	76.2
	REALISE	86.0	82.9	84.4	84.1	81.0	82.5
	LEAD	85.4	83.3	84.3	83.5	81.4	82.4
	SCOPE	90.7	86.8	88.7	89.5	86.0	87.7
Refined	PLOME	**78.8** (↑1.1)	**79.9** (↑1.0)	**79.4** (↑1.1)	**76.4** (↑0.8)	**77.5** (↑0.7)	**77.0** (↑0.8)
	REALISE	**87.0** (↑1.0)	**84.3** (↑1.4)	**85.6** (↑1.2)	**85.2** (↑1.1)	**82.6** (↑1.6)	**83.9** (↑1.4)
	LEAD	**86.2** (↑0.8)	**84.5** (↑1.2)	**85.3** (↑1.0)	**84.2** (↑0.7)	**82.6** (↑1.2)	**83.4** (↑1.0)
	SCOPE	**91.5** (↑0.8)	**88.2** (↑1.4)	**89.8** (↑1.1)	**90.4** (↑0.9)	**87.2** (↑1.2)	**88.8** (↑1.1)

Table 5: The results on SIGHAN15. The asterisk * indicates the results are copied from the original paper.

An interesting observation is that the "Retrained" results generally show the models ranked by performance from high to low are SCOPE[4], LEAD, REALISE and PLOME, which coincides with the "Original" results. This indicates that we have correctly retrained the models and the fixed SIGHAN can reflect their performance discrepancies.

3 A Refinement Solution using ChineseBERT

We extract the sentences from the testing sets that none of the four models successfully reproduces the target sentence, and analyze the reasons of failures. We observe three main types of failures. First, the models often fail to correct the particles "的", "地" and "得" and the pronouns such as "他（们）", "她（们）", "它（们）", "那" and

[4]SCOPE seems to be much more affected by "的", "地" and "得". After excluding them, SCOPE performs better on the fixed SIGHAN13 as shown in Table 7 of the appendix.

"哪". Second, the models often fail to correct the spelling errors in special terms, including idioms, proverbs, proper nouns and other commonly-used expressions. Third, the models often make over-corrections.

Most of the above failures can be solved by inferring the correct character using the contextual information of the corresponding sentence. Based on the idea, we propose a simple refinement solution with ChineseBERT (Sun et al., 2021) on top of the output of the four models. Specifically, given a sentence output by any model, we mask the character pertaining to the above failure cases, and let ChineseBERT infer the new character without training. Then we measure the phonological distance between the masked character and the inferred character, where the distance is calculated as the edit distance between the pinyins (with tone) of the two characters. If the distance is below a threshold[5], we keep the inferred character; otherwise, we keep the masked character. The intuition is that about 83% spelling errors have similar pronunciation with the correct character (Liu et al., 2010), so if the inferred character has a very different pinyin than the masked character, it is unlikely to be the correct character. If there are multiple characters to mask in a sentence, we mask them one at a time and infer using ChineseBERT, from beginning to end. Once there is no character to mask, we stop the process and use the last output of ChineseBERT as the refined sentence. Note that if a sentence output by the above four models contains no character to mask, the sentence is the final output and the refinement process does not run.

The problem at hand is how to identify the characters to be masked. We design three strategies for the three failure types, respectively. First, we directly mask the particles "的", "地" and "得" and the pronouns "他", "她", "它", "那" and "哪". Second, for a special term with spelling errors, we notice that the *jieba*[6] tokenizer produces different tokens with and without the Hidden Markov Model (HMM). The former tends to regard it as a new word and the latter tends to tokenize it into single characters. Hence, for a sentence output by the above models, we use the two methods to tokenize it and regard the parts with different tokenization results as the special terms to mask. Note that this approach may mask phrases other than special terms if there exist spelling errors. Third, to identify over-corrections, we calculate the edit distance between the pinyins (with tone) of the changed character and the original character in the source sentence. If the distance is above 3 as discussed in the last paragraph, we regard it as a potential over-correction and mask the character.

The results are presented in Table 3, 4 and 5, indicated by "Refined". We observe that after refinement, the performances of all the four models are improved by notable margins in all metrics on the three datasets, compared to the "Retrained" results. The results show our simple solution is very effective, even without training.

4 Related Work

Recent studies mainly adopt Transformer or BERT/ChineseBERT as the base model to solve the CSC task, and incorporate rich semantic features of the Chinese language to enhance the ability of the base model. For instance, Cheng et al. (2020) and Nguyen et al. (2021) use the confusion sets[7] to exclude unlikely candidates output by BERT. More studies such as Xu et al. (2021); Huang et al. (2021); Liu et al. (2021); Li et al. (2022a,b); Liang et al. (2023); Zhang et al. (2023); Wei et al. (2023) leverage phonological and/or visual features of characters to boost the performance. Studies like Zhang et al. (2020, 2021); Li et al. (2021); Zhu et al. (2022); Huang et al. (2023) adopt the detection-correction framework to increase the accuracy of identifying potential spelling errors. Other studies learn contextual information in sentences to detect and correct spelling errors (Guo et al., 2021; Wang et al., 2021; Liu et al., 2022; Li et al., 2022c).

5 Conclusion

In this work, we discuss two issues with the Chinese Spelling Correction task: the existence of mistakes in the SIGHAN datasets and the smaller and unstable improvements of new models. We manually fix the mistakes and re-evaluate four representative CSC models on the fixed datasets. We analyze the common types of failures of the models and propose a simple yet effective refinement solution. Experimental results show our solution can stably improve the base models in all metrics. While the current refinement solution is purely rule

[5] In the experiments, we set the threshold to 3.

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

[7] The confusion sets are a collection of sets, where each set is formed with phonologically or visually similar characters.

based, in the future we will develop data-driven methods to further improve the performance.

Limitations

There are two main limitations in the current work. First, the four models evaluated in the experiments belong to the category that incorporate phonological and visual features of Chinese characters. We choose them because they are reported in their papers to have the strongest performance among existing models and the source code are well maintained and released by the authors for reproducing and training. However, we should evaluate diverse models in the future, such as those using the detection-correction framework and those incorporating the contextual information. Second, our strategy to identify the characters in special terms and over-corrections to be masked is rule based and is not very accurate. For special terms with spelling errors, the identification depends on whether the jieba tokenizer with and without HMM yield different tokenization results. For over-corrections, we empirically identify them based on the edit distance between the pinyins (with tone) of a changed character and the original character. The threshold of the distance is set empirically and the visual distance is not considered, which is also the case for deciding whether to preserve the character inferred by ChineseBERT or not at the final output. While the current refinement solution is simple yet effective, we will explore more complex methods to further improve the accuracy of identifying the characters to be masked, as well as the final performance for CSC.

Ethical Statement

The datasets and the models used in the current study are all released and authorized by the original authors for research purpose. These datasets contain neither identifying information nor any other ethical issues. The output of the models do not contain any violence, pornography or other inappropriate information. Hence, there is no ethical issue in the current study.

Acknowledgement

This work is supported by the grant from the National Natural Science Foundation of China (Grant No. 62277017).

References

Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using smt for ocr error correction of historical texts. In *10th conference on International Language Resources and Evaluation (LREC'16)*, pages 962–966. European Language Resources Association.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. Spellgcn: Incorporating phonological and visual similarities into language models for chinese spelling check. In *Proceedings of 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881.

Zhao Guo, Yuan Ni, Keqiang Wang, Wei Zhu, and Guotong Xie. 2021. Global attention decoder for chinese spelling error correction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1419–1428.

Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11514–11525.

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 (Volume 1: Long Papers)*, pages 5958–5967.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jiahao Li, Quan Wang, Zhendong Mao, Junbo Guo, Yanyan Yang, and Yongdong Zhang. 2022a. Improving chinese spelling check by character pronunciation prediction: The effects of adaptivity and granularity. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4275–4286.

Jing Li, Gaosheng Wu, Dafei Yin, Haozhao Wang, and Yonggang Wang. 2021. Dcspell: A detector-corrector framework for chinese spelling error correction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1870–1874.

Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Li Yangning, Shulin Huang, Ruiyang Liu, Chao Li, Yunbo Cao, and Haitao Zheng. 2022b. Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 238–249.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022c. The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3202–3213.

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.

C-L Liu, M-H Lai, K-W Tien, Y-H Chuang, S-H Wu, and C-Y Lee. 2011. Visually and phonologically similar characters in incorrect chinese words: Analyses, identification, and applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(2):1–39.

Chao-Lin Liu, Min-Hua Lai, Yi-Hsuan Chuang, and Chia-Ying Lee. 2010. Visually and phonologically similar characters in incorrect simplified chinese words. In *Coling 2010: Posters*, pages 739–747.

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*, pages 3008–3018.

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 (Volume 1: Long Papers)*, pages 2991–3000.

Bruno Martins and Mário J Silva. 2004. Spelling correction for search engine queries. In *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004. Proceedings 4*, pages 372–383. Springer.

Minh Nguyen, Hoang Gia Ngo, and Nancy F Chen. 2021. Domain-shift conditioning using adaptable filtering via hierarchical embeddings for robust chinese spell check. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075.

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*, pages 32–37.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Baoxin Wang, Wanxiang Che, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2021. Dynamic connected networks for chinese spelling check. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2437–2446.

Dingmin Wang, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5780–5785.

Xiao Wei, Jianbao Huang, Hang Yu, and Qian Liu. 2023. Ptcspell: Pre-trained corrector based on character shape and pinyin for chinese spelling correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6330–6343.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at sighan bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 716–728.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of sighan 2014 bake-off for chinese spelling check. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 126–132.

Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang. 2021. Correcting chinese spelling errors with phonetic pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2250–2261.

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*, pages 882–890.

Xiaotian Zhang, Yanjun Zheng, Hang Yan, and Xipeng Qiu. 2023. Investigating glyph-phonetic information for chinese spell checking: What works and what's

next? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1–13.

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*, pages 1244–1253.

Appendix

Type of mistake		Example
Meaningless sentences	Original	大家也可怕你的工厂把自然破坏， People are also fright that your factory will destroy nature
	Fixed	大家也害怕你的工厂把自然破坏， People are also afraid that your factory will destroy nature
	Original	对我来说，在教室里录影小学生非常不好。 For me, recording in the classroom is very bad primary school students.
	Fixed	对我来说，在教室里录影对小学生非常不好。 For me, recording in the classroom is very bad for primary school students.
	Original	所以我们今天免费提供饮料或点甜。 So we're offering a free drink or point sweet today
	Fixed	所以我们今天免费提供饮料或甜点。 So we're offering a free drink or dessert today
	Original	有一天，有一个人以为我偷了的车子！ One day, a man thought I had stolen car!
	Fixed	有一天，有一个人以为我偷了他的车子！ One day, a man thought I had stolen his car!
Spelling errors in target sentences	Original	拿到礼物的人不觉得使用，或一点儿都没有用处 The person who received the gift did not find it use or useful at all
	Fixed	拿到礼物的人不觉得实用，或一点儿都没有用处 The person who received the gift did not find it useful or useful at all
	Original	这件话以后对父母越来越感谢。 After this piece of sentence, I am more and more grateful to my parents.
	Fixed	这句话以后对父母越来越感谢。 After this sentence, I am more and more grateful to my parents.
	Original	这种作法并不能来解释问题。 This practise magic does not explain the problem.
	Fixed	这种做法并不能来解释问题。 This approach does not explain the problem.
Unconverted traditional Chinese characters	Original	老师一来倒楣的一定是走廊、地板和黑板 When a teacher comes, it is always the corridor, the floor, and the blackboard get dump lintel
	Fixed	老师一来倒霉的一定是走廊、地板和黑板 When a teacher comes, it is always the corridor, the floor, and the blackboard get bad luck
	Original	可是公车没有座位所以他们站著说话。 But there were no seats on the bus so they stood book and talked.
	Fixed	可是公车没有座位所以他们站着说话。 But there were no seats on the bus so they stood and talked.
	Original	因为在那里有着各式各样、琳琅满目的书笈 Because there are all kinds of a box for books, dazzling eyes
	Fixed	因为在那里有着各式各样、琳琅满目的书籍 Because there are all kinds of books, dazzling eyes

Table 6: More examples of different mistake types and the corresponding fixes.

Datasets & Models		Detection			Correction		
SIGHAN13		D-P	D-R	D-F	C-P	C-R	C-F
Retrained	PLOME	81.3	77.9	79.6	79.6	76.3	77.9
	REALISE	81.9	77.6	79.7	80.0	75.9	77.9
	LEAD	84.7	79.8	82.2	82.3	77.6	79.9
	SCOPE	81.8	78.1	80.0	80.0	76.4	78.1

Table 7: The retrained results on SIGHAN13, excluding “的”, “地” and “得”.