Driving Chinese Spelling Correction from a Fine-Grained Perspective

Anonymous ACL submission

Abstract

This paper explores the task: Chinese spelling correction (CSC), from a fine-grained perspective by recognizing that existing evaluations lack nuanced typology for the spelling errors. This deficiency can create a misleading impression of models' performance, incurring an "invisible" bottleneck hindering the advancement of CSC research. In this paper, we categorize spelling errors into six types and conduct a fine-grained evaluation across a wide variety of models, including tagging models, ReLM, and LLMs. As a result, we pinpoint the underlying weaknesses of existing state-of-the-art models - utilizing contextual clues and handling co-existence of multiple typos. However, these two types of errors suffer from very low occurrence in conventional training corpus. Therefore, we introduce new error generation methods to artificially augment their occurrence. Armed with augmented data, we eventually enhance the overall performance of prior CSC models by boosting their performance on specific errors. We hope that this work could provide fresh insight for future CSC research.

1 Introduction

This paper studies the evaluation principle for Chinese spelling correction (CSC), a fundamental task in natural language processing to rectify all potential spelling errors in a Chinese sentence. Evaluation plays a critical role in CSC research, where the researchers are allowed to understand the way models behave and guide for further solutions. Due to the profoundness of Chinese language, there are diverse misspelling variations in real human corpora. However, existing benchmarks (Tseng et al., 2015; Lv et al., 2023; Wu et al., 2023b) are constrained to producing an overall score for all kinds of spelling errors, providing a coarse reflection of model performances. This deficiency incurs an "invisible" barrier that bottlenecks the progress of CSC research. In this paper, we propose a fine-grained evaluation principle, named **FiBench-CSC**, in the hope of navigating the follow-up research in the CSC community.

Figure 1: Samples of different types of spelling errors.

We categorize the spelling errors in a Chinese sentence to five distinct types. Figure 1 offers an illustration of them. We first categorize the errors by the way they are misspelled. **Phonological error** and **morphological error** are the two most common error types, stemming from the pinyin and stroke similarities inherent in Chinese characters (Liu et al., 2010). The former is caused by users' keyboard input or audio speech recognition, while the latter is caused by handwriting. In Figure 1, "舍" are "赶" are the phonological and morphological counterparts of "舌" and "赴" respectively. These two types of errors are rich in the confusion sets, which are used to generate synthetic errors on top of monolingual sentences. We group the remaining errors not conforming to any of these two types into **non-similarity error**.

Second, we categorize the errors by the number of them within a single sentence, i.e. **single error** and **multi-typo error**. The latter refers to cases where are more than one typos in one sentence. Co-existence of multiple typos may largely distort the context and create intricacy for correc-

tion. For example in Figure 1, there are two typos at the same time, where "饮食" is misspelled to "饮事" and "消化" is misspelled to "消话". The correction of the latter typo necessitates the correct understanding of the former phrase "饮食规律", which is disturbed by the typo "饮事".

Third, we introduce **contextual error**. This type of errors locally manifests as correct phrase within the sentence. However, their correction strongly relies on contextual clues. For example in Figure 1, "语言" (lingual) is misspelled to "预演" (preview), both of which are legitimate Chinese words. Only if referring to the subsequent context of "多语言服务" (multilingual services), can one figure out the final answer. The edit pairs of contextual errors are diverse case by case and may hardly occur in the confusion sets. Given that many CSC models are developed with signals from confusion sets, correction of these errors can be a challenging task, for which merely memorizing edit pairs from the training corpus is insufficient.

In FiBench, each dataset is reorganized into six subsets, each associated with one specific error type, thus allowing for an ever fine-grained insight into models' strengths and shortcomings. Our paper unfolds as below. In §2, we first conduct a comprehensive FiBench evaluation on a broad range of CSC models. While state-of-the-art counterparts show adeptness in using phonological and morphological clues, we pinpoint contextual and multi-typo errors that notably struggle with. However, these two errors can hardly exist in conventinal confusion sets. In §3, we introduce two error generation methods to automatically synthesize contextual and multi-typo errors given arbitrary sentences. In §4, we refine the training process of recent models based on newly generated error signals. Furthermore, we an in-depth analysis targeting the two challenging error types.

2 FiBench

In this section, we scrutinize the performances exhibited by existing CSC models from a fine-grained perspective. The findings in this section serve as the foundation for the subsequent methods and experiments in the paper.

2.1 Categorization Principle

Phonological & Morphological & Non-similarity We obtain the phonological errors and morphological errors by checking if the edit pair in the sentences exists in the associated confusion set, while categorizing the rest into non-similarity errors. The confusion sets employed in our study are released by Liu et al. (2022).

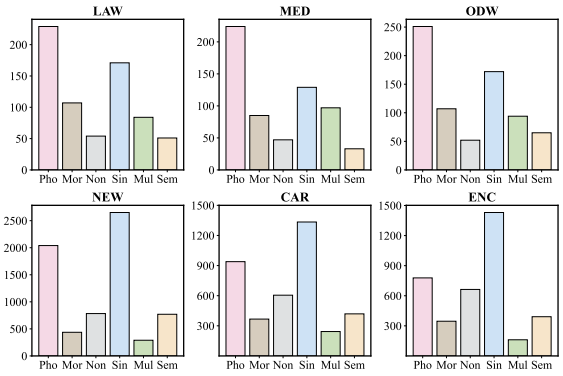

Figure 2: Statistics of error types in six chosen domains.

Contextual The process we obtain the contextual errors is two-step. First, we check if the edit pair in the sentence can form a correct word or phrase in the local context. The logic behind this is that if the error cannot make a correct expression, it can be easily detected without referring to the context. This step can greatly increase the efficiency of the process. Second, we conduct manual screening to eventually obtain the contextual errors.

Single & Multi We obtain the single and multi-typo errors simply by counting the number of errors in the sentence.

2.2 Datasets

We conduct experiments on two public datasets, ECSpell (Lv et al., 2023) and LEMON (Wu et al., 2023b). **ECSpell** is a small-scale CSC benchmark with three domains: LAW (law) with 1,960 training and 500 test samples, MED (medical treatment) with 3,000 training and 500 test samples, and ODW (official document writing) with 1,728 training and 500 test samples. **LEMON** is an open-domain CSC benchmark with a diverse set of real-life spelling errors. In our experiments, we choose the three biggest domains as representatives: NEW (news title) with 5,887 test samples, CAR (car) with 3,245 samples, and ENC (encyclopedia) with 3,274 test samples.

Figure 2 eventually demonstrates the statistics of six error types in ECSpell and LEMON. There exists an overlap of samples among each error subset.

2.3 Models and Methods

We span a broad range of models/methods.

BERT The pre-trained BERT (Devlin et al., 2019) is the fundamental architecture to perform the CSC task in the way of sequence tagging.

Soft-Masked BERT Zhang et al. (2020) apply a GRU network as the additional detector and mask the detected errors in the sentence using soft-masking technique to encourage the correction.

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

CRASpell The correction-making is likely to be biased when there is one more error in the context. Liu et al. (2022) augment the original sentence by introducing an additional error and optimizing a smoothness loss (Jiang et al., 2020; Wu et al., 2023a) on it.

PLOME PLOME (Liu et al., 2021) is an enhanced pre-trained CSC model based on BERT. It is incorporated with phonological and morphological clues in its word embeddings.

Masked-Fine-Tuning Above models learn CSC by sequence tagging. We apply the masked-fine-tuning technique (MFT) to boost the tagging process (Wu et al., 2023b), which is designed to enhance the language modeling aspect of CSC learning.

ReLM Rephrasing Language Model (ReLM) (Liu et al., 2024) is a non-autoregressive language model, which regards CSC as sentence rephrasing on top of entire semantics. A CSC model is thus transferred to a pure language model.

LLM Similar to ReLM, CSC is a sentence rephrasing task for large language models (LLMs), while they rephrase the sentences in an autoregressive manner. However, we find that generative models suffer from the overly-paraphrase issue. The prompt we use is `Detect whether there are any misspelled words in the sentence. If there are any, please correct them.` The important trick here is that the model won't do anything on the input sentence if there are no errors detected, which we find useful for reducing the above issue. We adopt Baichuan2-7b (Yang et al., 2023) in our experiments. We also instruct GPT4 (OpenAI, 2023) to perform this task through in-context learning with 5 shots. For each sentence, the in-context samples are uniformly chosen from sentences with the same error type in the training set. The prompt we use is `Please correct the spelling errors in the given sentence, ensuring that the modified sentence is the same length as the original one. If there are no errors in the sentence, please copy it exactly as it is.`

Tagging vs. Rephrasing It is worth noting that current CSC models can be categorized into tagging models and rephrasing models, by their training objectives. The former corresponds to BERT, Soft-Masked BERT, MDCSpell, CRASpell, and PLOME, while the latter corresponds to ReLM and a series of autoregressive models. In the following, we will explore their differences.

2.4 Training Setup

On ECSpell, we fine-tune each model separately on the three domains for 5,000 steps with the batch size selected from {32, 128} and learning rate from {2e-5, 5e-5}. Especially for fine-tuning Baichuan2, we set the learning rate to 3e-4 and use LoRA (Hu et al., 2022a) with $r = 8$ and $\alpha = 32$ to improve efficiency. On LEMON, we adopt the pre-trained models open-sourced by Wu et al. (2023b). Each model is trained on 34 million synthetic pair-wise sentences from `wiki2019zh` and `news2016zh`. We evaluate each pre-trained model in zero-shot learning on each LEMON domain.

2.5 Evaluation Result

Table 1 reports the performances of a line of CSC models on ECSpell and LEMON.

Most models show nice adeptness in addressing phonological and morphological errors. We can also see that these two types of errors are less challenging for models under zero-shot learning, compared to the other ones. In addition, we find that PLOME achieves great success on associated phonological and morphological errors, which undergoes additional pre-training guided by confusion sets. It indicates that the similarity signals like pronunciations and shapes are rich in the training corpus for CSC models to fit the error model (Wu et al., 2023b).

A large performance disparity emerges when models moving from addressing single errors to multi-typo errors. For multi-typo errors, we find distinct trends between fine-tuned models and the zero-shot models. Among the fine-tuned models, enhanced tagging models like Soft-Masked BERT, MDCSpell, and CRASpell, even after undergoing MFT, only achieve limited performance

		Phono.	Morpho.	Non-s.	Single	Multi	Contextual	FPR	Overall
EC-LAW	BERT	52.9	58.5	17.2	56.2	15.4	28.5	10.6	36.9
	PLOME†	91.7	79.0	54.7	82.1	47.3	19.9	5.2	69.6
	$BERT_{MFT}$	93.3	93.2	89.9	92.9	57.1	59.1	13.8	75.6
	$Soft\text{-}Masked_{MFT}$	95.6	93.2	94.1	94.5	69.8	75.5	15.5	80.0
	$MDCSpell_{MFT}$	95.9	95.6	96.1	96.7	70.3	74.7	14.6	82.0
	$CRASpell_{MFT}$	96.8	95.1	94.2	96.7	66.6	76.7	9.3	82.8
	ReLM	99.1	99.0	**98.1**	**99.1**	**96.4**	87.9	10.2	89.4
	ReLM†	**99.9**	**99.5**	96.2	98.8	**96.4**	**98.0**	5.3	**95.6**
	Baichuan2	93.6	92.3	94.3	92.4	85.7	80.8	**2.0**	92.8
	GPT4 (5-shot)	80.7	71.2	72.0	70.7	52.2	43.7	6.5	67.9
EC-MED	BERT	33.7	54.2	37.6	39.6	11.6	42.6	9.8	25.1
	PLOME†	82.5	76.5	61.7	72.3	45.0	12.7	7.3	60.0
	$BERT_{MFT}$	75.5	88.6	72.4	78.2	37.8	66.6	11.6	55.9
	$Soft\text{-}Masked_{MFT}$	88.1	89.9	84.7	86.1	51.0	66.6	12.4	65.6
	$MDCSpell_{MFT}$	86.0	93.3	81.9	87.0	57.9	74.1	12.0	69.0
	$CRASpell_{MFT}$	86.1	90.9	86.9	82.8	63.1	67.7	8.7	72.5
	ReLM	92.9	97.0	93.6	94.0	67.3	67.7	9.8	80.9
	ReLM†	**98.4**	**97.3**	**97.6**	**98.3**	**90.3**	74.9	8.7	**89.9**
	Baichuan2	90.8	91.6	86.6	86.6	77.7	**80.0**	**5.1**	79.8
	GPT4 (5-shot)	57.5	65.5	68.0	66.9	53.7	56.2	24.0	56.4
EC-ODW	BERT	27.4	41.6	27.4	35.2	10.3	29.7	14.1	24.8
	PLOME†	87.6	80.4	71.4	60.2	57.5	31.2	3.9	67.2
	$BERT_{MFT}$	76.7	83.8	69.5	73.2	42.9	51.2	17.0	57.6
	$Soft\text{-}Masked_{MFT}$	86.0	92.7	72.9	80.6	53.6	58.8	14.9	66.4
	$MDCSpell_{MFT}$	86.5	93.2	75.5	81.5	55.2	62.2	16.6	67.0
	$CRASpell_{MFT}$	89.6	90.2	78.5	83.4	56.9	66.1	6.4	74.9
	ReLM	93.9	93.7	84.5	86.3	82.1	79.6	10.2	81.0
	ReLM†	**97.1**	**97.1**	88.6	**92.4**	**91.3**	**89.4**	2.1	**92.3**
	Baichuan2	89.8	94.3	**92.1**	85.6	87.2	88.8	2.5	87.5
	GPT4 (5-shot)	87.1	83.9	75.5	76.6	71.6	61.8	**1.7**	72.5
LE-NEW	$BERT_{MFT}$†	71.3	72.0	45.0	63.9	11.3	49.3	9.4	56.0
	$Soft\text{-}Masked_{MFT}$†	71.8	72.1	42.8	64.0	10.8	50.4	10.5	55.6
	$MDCSpell_{MFT}$†	**74.9**	73.2	37.7	65.6	11.0	**53.0**	9.1	57.3
	$CRASpell_{MFT}$†	72.9	73.8	38.9	64.4	5.6	50.7	10.5	55.4
	ReLM†	**74.9**	**75.8**	**44.0**	**67.0**	10.2	52.2	**8.3**	**58.8**
	GPT4 (5-shot)	40.6	47.8	38.5	37.1	**23.7**	46.1	41.9	34.8
LE-ENC	$BERT_{MFT}$†	62.4	62.1	35.5	53.9	5.7	42.1	13.8	45.2
	$Soft\text{-}Masked_{MFT}$†	59.3	62.1	33.9	52.8	5.6	39.4	14.7	44.1
	$MDCSpell_{MFT}$†	**63.8**	66.7	33.7	54.7	7.3	41.4	13.8	46.1
	$CRASpell_{MFT}$†	62.8	**68.1**	39.2	**56.8**	4.9	**43.3**	14.3	47.6
	ReLM†	63.1	63.4	**41.4**	56.5	3.3	39.8	**12.7**	**47.6**
	GPT4 (5-shot)	40.4	51.3	40.4	39.1	**25.2**	32.7	33.1	40.5
LE-CAR	$BERT_{MFT}$†	74.1	65.9	45.3	64.5	4.2	**47.5**	12.2	51.9
	$Soft\text{-}Masked_{MFT}$†	73.6	67.4	47.1	64.5	7.6	46.8	12.3	52.2
	$MDCSpell_{MFT}$†	74.8	70.3	38.3	64.0	8.1	43.4	**11.9**	51.9
	$CRASpell_{MFT}$†	74.6	**71.8**	42.7	64.7	5.9	45.5	13.2	51.9
	ReLM†	**76.8**	66.3	**45.0**	**65.7**	9.7	44.7	**11.9**	**53.5**
	GPT4 (5-shot)	39.8	43.3	39.0	36.1	**19.2**	32.1	31.5	35.9

Table 1: Fine-grained performances on ECSpel (EC-x) and LEMON (LE-x). We report the F1 score for each error type, the false negative rate (FPR) on non-error sentences, and the overall F1 score on all sentences. "Non-s." refers to the non-similarity errors. † refers to the pre-trained model on additional CSC data. The subscription MFT indicates that the model is trained using masked-fine-tuning.

on multi-typo errors. This suggests that additional errors in the context can significantly compromise the model's decision-making. In contrast, the fine-tuned rephrasing models, ReLM and Baichuan2, demonstrated remarkable adeptness in handling multiple typos within a single sentence, indicating that the rephrasing process can effectively bypass the negative impact of multiple errors co-existing simultaneously. However, under zero-shot learning, the performance of all models on multi-typo errors

deteriorates substantially, including ReLM, which is considered more powerful in language modeling. This indicates a potential issue in the training process that researchers might overlook constructing samples that contain multi-typo errors, resulting in models' inability during testing.

Contextual errors pose a consistent challenge in every scenario. For fine-tuned models, we find that rephrasing models generally perform better, such as ReLM and Baichuan2, while MFT appears to be particularly important for tagging models to deal with contextual errors. We also see that the guidance of confusion sets is ineffective in addressing such errors. This is because that the edit pairs of contextual errors are almost nonexistent in them. However, for zero-shot models, all of them struggle with contextual errors. Correspondingly, their performance on non-similarity errors also encounters a big decline. Surprisingly, ReLM rephrases the sentence based on the entire semantics (Liu et al., 2024), which is supposed to naturally excel in handling contextual errors. However, we find that zero-shot ReLM offers no advantage over other models. We conjecture that adaptation to contextual errors heavily relies on domain-specific signals, allowing models to better understand the context within the domain. Unfortunately, it is very hard to achieve in open-domain CSC. This indicates that open-domain CSC is the biggest challenge faced by the community.

We realize that **existing CSC models exhibit deficiencies in addressing these two types of errors, bottlenecking their overall performance in practical applications.** Therefore, there emerges a very need for the development of a comprehensive benchmark specifically targeting the performance of CSC models in addressing multi-typo and contextual errors. This forms the focus of our follow-up section, where we will introduce such a benchmark.

3 Error Generation

In this section, we discuss the error generation method to automatically generate contextual errors with the assistance of the powerful lexical processing capability of LLMs, as well as the synthesis method to generate multi-typo errors.

3.1 Contextual Error

We design a three-step pipeline. Given a sentence, we first tokenize it into words using the segmentation tool and randomly select one of them as the target word. We prompt GPT4 to substitute the target word for a new word. The prompt for substitution is shown in Figure 3. In this prompt, we instruct GPT4 to follow two primary principles: 1. the new word is still a legitimate Chinese word; 2. the new word will introduce an unnatural semantics to the entire sentence.

Substitution:
You are a native Chinese speaker to modify the given sentence following the requirements below.

1. Change the word in "<>" to a new word using the same number of characters.
2. The new word in "<>" is correct within the local context.
3. The new word in "<>" should induce a wrong or strange meaning of the new sentence.
4. Do not change the other words outside of "<>".

Input:
比赛至今他从未出现，可见他是一个<鸽子>。
Response 1:
比赛至今他从未出现，可见他是一个<毒瘤>。
Response 2:
比赛至今他从未出现，可见他是一个<歌者>。

Verification:
You are a skilled Chinese writer. People admire you. I will give a pair of sentences, please help me decide the following situations:

1. two sentences are in the same meaning, and they are both grammatically and contextually correct.
2. two sentences are in different meanings, but they are both grammatically and contextually correct.
3. two sentences are supposed to be in the same meaning, but either is not grammatically and contextually correct.

Input 1:
比赛至今他从未出现，可见他是一个<鸽子>。
比赛至今他从未出现，可见他是一个<毒瘤>。
Response 1:
2
Input 2:
比赛至今他从未出现，可见他是一个<鸽子>。
比赛至今他从未出现，可见他是一个<歌者>。
Response 2:
3

Figure 3: Prompts we use to generate contextual errors.

The first step is a tough task even for GPT4. It is likely to solely paraphrase the given sentence or introduce another word, potentially retaining correctness while altering the original meaning. If either of two situations occurs, we will acquire an invalid sentence pair. To address this, we design the second step to verify the validity of the output sentence from the first step. As detailed in Figure 3, we further prompt GPT4 to identify the relationship between the output sentence in the first step and the original one. Only if both sentences convey the same meaning and one contains grammatical and contextual error, do we keep this sentence pair.

LLMs like GPT4 lean to make somewhat unstable responses. To ensure reliability, we eventually

employ a ruled-based filter to verify if the new word can form a legitimate expression if checking its existence in a word vocabulary.

3.2 Multi-typo Error

We construct a distribution to synthesize multiple typos in one sentence. Each typo can be any of a contextual error, phonological error, or morphological error. The last two errors are sampled from the associated confusion sets, while the contextual errors are generated using the prior method. Given an arbitrary sentence, we introduce N typos in it. N follows the p-Binomial distribution $\sim \text{Binomial}(n, p)$, where n is the number of characters in the sentence. When N is determined, specifically, we uniformly sample N positions in the sentence and replace each of them with: 1. a phonologically similar character 60% of the time; 2. a morphologically similar character 30% of the time; 3. a character/word making a contextual error 10% of the time. This is due to the empirical fact that contextual errors occur at a lower frequency in real-world sentences.

The expectation of N is np, meaning that there will be one typo for every one hundred characters in the sentence, if p is set to 1%.

4 Data Augmentation

In this section, we refine the existing datasets using the error generation methods introduced in § 3. Based on the augmented data, we introduce several effective training strategies to facilitate stronger CSC models.

4.1 Strategy

We have observed that models fine-tuned on ECSpell exhibit a greater susceptibility to contextual errors, yet better performance in the face of multi-typo errors (from Table 1). This is attributed to the fact that contextual errors have a significantly lower occurrence in the training set, even lower than that of multi-typo errors (from Figure 2). Therefore, we randomly sample a proportion of the target sentences in the training set and generate new contextual errors on them. Given that contextual errors occur less frequently in natural language, excessive introduction of them may compromise the model's overall performance. Hence, we **complement the training data with 100 new samples with contextual errors** for each domain ($\sim$ 5% of original training samples). Additionally, in § 3, we have conjectured that adaption to contextual errors strongly depends on domain-specific signals. We prepare another 100 samples with contextual errors for comparison, where the target sentences are sourced from Chinese wikipedia.

For open-domain CSC, models are pre-trained a large scale of pair-wise sentences without being fine-tuned on specific training sets. We thus employ two strategies, **continue-training** and **few-shot learning**. Instead of undergoing a new round of complete pre-training, we choose to continually train the model on refined sentences. Specifically, we refine the synthetic pair-wise sentences from `wiki2019zh` (each already with one typo) by imposing random additional typos to them, and train the prior model for another one epoch. Since the sentence initially contains a typo, we set p for the Binomial distribution to a lower value 0.001. Another more efficient approach is to construct a few samples with highly concentrated errors to allow the model to quickly adapt to the associated error types. We set p to 0.1 and generate 100 samples with multi-typo errors. However, our experience suggests that this rapid method can trade off the performance on the rest error types.

Lastly, we acknowledge that addressing contextual errors is a remaining challenging for open-domain CSC. Unfortunately, LEMON doesn't offer domain-specific training sets for us to conduct further experiments. In future work, we will try to address this issue. In our experiments, we verify our method for contextual errors through ECSpell.

4.2 Result

Given that ReLM is the newest state-of-the-art counterparts, our experiments are based on ReLM. The upper part of Table 2 showcases the effectiveness of incorporating new contextual errors. Significant performance improvement can be observed in the domains of MED and ODW. For instance, on MED, the performance on contextual errors increases from 75.8 to 87.7, which further results in the improvement of the overall performance. On the other hand, we find that constructing contextual errors using the general corpus doesn't yield significant benefit. It indicates that the exploitation of contextual information is consistent with our prior hypothesis in § 3.

From the lower part of Table 2, we find that continue-training enhances the certain aspects of the model in a more stable manner. For multi-typo errors, the resultant ReLM gains a significant

	LAW		MED		ODW	
	Con	All	Con	All	Con	All
ReLM	98.0	96.0	75.8	89.2	91.1	91.6
ReLM$^{\clubsuit\text{domain}}$	**100.0**	**96.4**	**87.7**	**90.7**	**95.9**	**92.1**
ReLM$^{\clubsuit\text{wiki}}$	97.1	95.0	78.2	90.0	91.9	90.5
	NEW		**ENC**		**CAR**	
	Mul	All	Mul	All	Mul	All
ReLM	10.2	**58.8**	3.3	47.6	9.7	53.5
ReLM$^{\clubsuit\text{CT}}$	**18.7**	58.6	12.9	**48.3**	**22.0**	**54.3**
ReLM$^{\clubsuit\text{FS}}$	16.8	53.9	**14.6**	43.5	20.0	50.6

Table 2: Results after simple data augmentation. "CT" refers to continue-training and "FS" refers to few-shot.

		Loc-F1	F1	Sensitivity
LAW	BERT	45.0	36.9	18.7
	BERT$_{\text{MFT}}$	60.6	75.6	67.0
	Soft-Mask$_{\text{MFT}}$	59.0	80.0	69.6
	MDCSpell$_{\text{MFT}}$	66.4	82.0	**70.9**
	ReLM	73.8	**96.0**	62.1
	ReLM$^{\clubsuit}$	70.7	**96.4**	72.6
	Baichuan2	37.1	92.8	64.8

Table 3: Analysis of contextual errors. We report the local-F1 (Loc-F1), overall F1 (F1), semantic sensitivity (Sensitivity) of several representative models on EC-LAW.

boost from 10.2 to 18.7 on NEW, 3.3 to 11.9 on MED, and 9.7 to 19.0 on ODW respectively. In contrast, the improvement brought by few-shot learning seems even more significant. However, we find that it rapidly deteriorates the overall performance. In our experiments, each model has been fine-tuned for only 3 epochs on few-shot samples. This is due to the fact that few-shot samples may significantly distort the natural data distribution. The recent BERT-based CSC models are not strong enough to overcome such a negative impact. Therefore, it won't be a feasible approach for general scenarios.

4.3 Analysis of Contextual Errors

The possible explanation behind this is the models' weak awareness of the context. **To track how context impacts the model's prediction,** we design a further experiment. First, we truncate the source sentence by solely keeping three neighboring words around the error characters. We calculate the F1 score on these truncated samples, denoted as *local-F1*. Second, we pick out the samples on which the model makes a wrong prediction with only local context. Then, we recover the full context for these samples and calculate the ratio that the model's prediction changes. We define this indicator as *Semantic Sensitivity*, which measures the sensitivity of a CSC model to the context change.

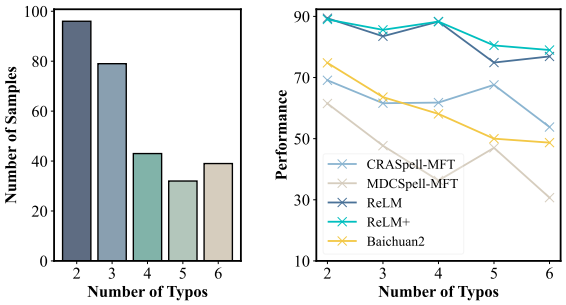

Figure 4: Left: Statistics of the number of typos in each example. Right: Variation of performances (F1) with the increasing number of typos. We choose ODW as the representative domain.

By comparing the first two columns of Table 3, we surprisingly find that BERT even achieves a better F1 score with only the information of local context. In contrast, the local-F1 of ReLM and Baichuan2 lag far behind their overall F1, where the full context is visible, suggesting their reliance on the entirety of contextual information for optimal predictions. Referring to the last column, we find that BERT is almost insensitive to context change and in only 18.7% of cases, the context recovery would impact its prediction. It underscores a significant drawback in tagging models, predominantly focusing on local edit pairs, i.e. the error model, thus having a poor utilization of semantics. We find that the ReLM model undergoing data augmentation exhibits a higher sensitivity to the context.

4.4 Analysis of Multi-typo Errors

For multi-typo errors, CSC models can be vulnerable to contextual noise while making the correction (Zhu et al., 2022; Liu et al., 2022). **Furthermore, we look deeper into the impact of the number of typos co-existed in the sentence** by grouping the multi-typo errors by their numbers. The results are depicted in Figure 4. Intuitively, all models experience a decline in performance when the number of typos rises. However, ReLM is able to maintain a nice and stable performance, outstripping all other tagging models by a big margin. This finding is consistent with that in Sec. 3. Among tagging models, CRASpell outperforms other counterparts, especially when the number of typos is above four, suggesting that optimizing the smoothness loss during training effectively allows the model to adapt

Case 1: synthetic contextual error
雷击债券余额不超过公司净资产的百分之十。[SRC] 累计债券余额不超过公司净资产的百分之十。[TRG]
Case 2: synthetic multi-typo error
知识单权权利人在许诺合同中进行价格歧视。[SRC] 知识产权权利人在许可合同中进行价格歧视。[TRG]
Bad Case 1: exploiting contextual clues
首先要简单的修剪美貌四周的碎毛。[SRC] 首先要简单的修剪眉毛四周的碎毛。[TRG] 首先要简单的修剪美貌四周的碎毛。[Original] 首先要简单的修剪眉毛四周的碎毛。[Augmented]
Bad Case 2: addressing multi-typo error
契而不舌的艰苦追求,使我们国内领先。[SRC] 锲而不舍的艰苦追求,使我们国内领先。[TRG] 契而不舍的艰苦追求,使我们国内领先。[Original] 锲而不舍的艰苦追求,使我们国内领先。[Augmented]

Table 4: Samples of contextual errors and multi-typo errors generated by our two error generation methods.

to multi-typo errors.

4.5 Case Study

We further offer a closer look on concrete cases. The case study comprises two parts. We first demonstrate the newly generated sample (TRG) given SRC by our methods. In case 1 (*The cumulative bond balance shall not exceed ten percent of the company's net assets*), we synthesize the contextual error "雷击" (lightning) → "累计" (accumulative). The correction of this error necessitates the model not only to seek clues from the context but also consider phonological similarity. Case 2 (*Intellectual property rights holders engage in price discrimination in licensing contracts*) contains two typos, where the correction of the second error "许可" (license contract) → "许诺" (promise contract) is strongly dependent on the correction of the first one "知识单权" → "知识产权" (intellectual property rights).

In the second part, we demonstrate the two cases where the model could successfully address them after undergoing data augmentation. In bad case 1 (*First, trim the stray hairs around the eyebrows*), the original ReLM fails to detect the contextual error "眉毛" → "美貌". After fine-tuning on augmented contextual errors, the augmented ReLM can successfully address it. In bad case 2 (*Persistent and strenuous efforts have made us a leader in the domestic market*), the augmented ReLM successfully detects the two typos.

5 Related Work

A large body of research in CSC focuses on introducing external clues, e.g. phonological and morphological similarity (Wang et al., 2019; Liu et al., 2021; Huang et al., 2021; Sun et al., 2023; Liang et al., 2023), negative samples (Li et al., 2022b), retrieval (Song et al., 2023), auxiliary objectives (Liu et al., 2021; Li et al., 2022a). Another line of work focuses on disentangling the detection and correction module (Zhang et al., 2020; Zhu et al., 2022; Huang et al., 2023). In contrast to these efforts, our work centers on the foundation principles for CSC.

Foundation Study for CSC and Benchmark Foundation study plays an essential role in the research of CSC. Wu et al. (2023b) explore the two underlying sub-models behind a general CSC model, the error model and language model. Liu et al. (2024) discuss the primary training objective for the CSC task. This paper focuses on the fundamental evaluation principle and offers an ever fine-grained perspective. Benchmarking is equally important. Recently, many attempts at benchmarks offer new standards for CSC research, e.g. IME (Hu et al., 2022b) for errors stemming from pinyin similarity, ECSpell for multi-domain (Lv et al., 2023), MCSC for medical-specialist (Jiang et al., 2022), LEMON for open-domain CSC (Wu et al., 2023b). A similar effort is Hu et al. (2022b), which synthesizes a large number of errors by approximating the real error distribution. Yet, diverging from their path, this paper focuses on the refinement of existing benchmarks with synthetic data. It potentially skews the real error distribution because we argue that it is those lower-frequency errors that pose the bottleneck of CSC models.

6 Conclusion

This paper identifies and categorizes spelling errors in Chinese into various types. We conduct a fine-grained evaluation across a broad spectrum of CSC models. The nuanced assessment offers a clear view of each model's strengths and weaknesses, which is crucial for their practical application and future enhancement. Additionally, we introduce automatic error generation methods specifically designed for contextual errors and multi-typo errors where current models show notable vulnerability. We also study the impact of context and number of typos using the augmented datasets.

7 Limitations

Our evaluation covers the most representative CSC methods in recent years while does not encompass all possible ones. Future work can further improve the landscape of FiBench. Besides, the experimental results demonstrate the potential of LLMs in certain aspects, such as tackling multi-typo errors and processing contextual signals. However, our paper primarily focuses on BERT-based models, without deeper exploration of LLMs. In the other hand, our study uncovers several valuable future directions. Open-domain CSC emerges as a notable challenge with sparse exploration. Firstly, we long for effective methods for handling **negative transfer between error types and domains**. Secondly, we long for **greater diversity in the training corpus** to enhance the base models. In this paper, we only consider the models trained from the source of wikipedia.

References

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yong Hu, Fandong Meng, and Jie Zhou. 2022b. CSCD-IME: correcting spelling errors generated by pinyin IME. *CoRR*, abs/2211.08788.

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, Singapore, December 6-10, 2023*, pages 11514–11525. 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.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2177–2190. Association for Computational Linguistics.

Wangjie Jiang, Zhihao Ye, Zijing Ou, Ruihui Zhao, Jianguang Zheng, Yi Liu, Bang Liu, Siheng Li, Yujiu Yang, and Yefeng Zheng. 2022. Mcscset: A specialist-annotated dataset for medical-domain chinese spelling correction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 4084–4088. ACM.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4275–4286. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022b. 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, Dublin, Ireland, May 22-27, 2022*, pages 3202–3213. 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.

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, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 739–747. Chinese Information Processing Society of China.

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.

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.

Qi Lv, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. 2023. General and domain-adaptive chinese spelling check with error-consistent pretraining. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(5):124:1–124:18.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Siqi Song, Qi Lv, Lei Geng, Ziqiang Cao, and Guohong Fu. 2023. Rspell: Retrieval-augmented framework for domain adaptive chinese spelling check. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part I*, volume 14302 of *Lecture Notes in Computer Science*, pages 551–562. Springer.

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

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, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for chinese spelling check. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5780–5785. Association for Computational Linguistics.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Pengjun Xie, Fei Huang, and Min Zhang. 2023a. Adversarial self-attention for language understanding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13727–13735. AAAI Press.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023b. 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.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

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.

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.