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 first categorize spelling errors into six types and conduct a *fine-grained evaluation* across a wide variety of models, including BERT-based models and LLMs. Thus, we are able to pinpoint the underlying weaknesses of existing state-of-the-art models - utilizing contextual clues and handling co-existence of multiple typos, associated to *contextual errors* and *multi-typo errors*. However, these errors suffer from low occurrence in conventional training corpus. Therefore, we introduce new error generation methods to synthesize their occurrence. Eventually, these augmented data can be leveraged to enhance the training process of CSC models. We hope this work could provide fresh insight for future CSC research.

Phonological Error:
舍 (>> 舌) 尖上的中国。
A bite (>> house) of China.

Morphological Error:
全速以赶 (>> 赴)，事半功倍，以静制动。
Full speed after (>> ahead) , twice result with half effort...

Non-similarity Error:
关于游戏引擎，下列说法正敏 (>> 确) 的是。
The following statements are sensitive (>> true) about game engines.

Multi-typo Error:
请平时注意休息，饮事 (>> 食) 规律，吃清淡，易消话 (>> 化) 的东西。
Please... make (>> eat) regular... something easy to talk (>> digest).

Semantic Error:
多预演 (>> 语言) 版本，根据用户群体特点和需求，提供多语言服务。
The multipreview (>> lingual) version offers multilingual services...

Figure 1: Samples of different types of spelling errors.

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 models' 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.

We categorize the spelling errors in a Chinese sentence to six distinct types. Figure 1 offers an illustration of five 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 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. 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 to **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 there is more than one typo in one sentence. Co-existence of multiple typos may largely distort the context and create intricacy for correction. 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 a correct phrase within the sentence. However, their correction strongly relies on utilizing 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 vary case by case and may not be found in the confusion sets. Given that many CSC models are constructed based on confusion sets, correction of contextual errors can be a challenging task, requiring much more than memorizing edit pairs from the training corpus.

In FiBench, we reorganize the target dataset 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 conduct a comprehensive FiBench evaluation choosing 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 they notably struggle with. However, the contextual errors are sparse in conventional confusion sets. In §3, we introduce new methods for error generation to synthesize the contextual and multi-typo errors given arbitrary sentences with the assistance of LLMs. In §4, we harness the new synthetic sentences to refine the training of CSC models, and witness a blazer to state-of-the-art performance by boosting the target efficacy in specific errors.

2 FiBench

In this section, we scrutinize existing benchmarks 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).

Contextual To obtain the contextual errors, we check if the edit pair in the sentence can form a legitimate word within the locality by referring a fixed vocabulary. The logic behind is that if the error cannot form a correct phrase, it can be easily detected without referring to the context.

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

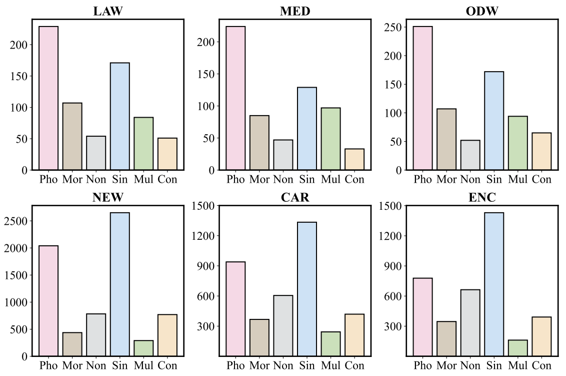

Figure 2: Statistics of error types in six chosen domains.

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 specific 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 across multiple domains. In our experiments, we choose three domains as representative: NEW (news title) with 5,887 test samples, CAR (car) with 3,245 test samples, and ENC (encyclopedia) with 3,274 test samples.

Figure 2 eventually demonstrates the statistics of six error types in ECSpell and LEMON. From our categorization principle, there will be overlap of samples among each error subset.

2.3 Models and Methods

We span a broad range of CSC methods including BERT-based models and LLMs.

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 softly 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 Liu et al. (2022) augment the original sentence by introducing an additional typo in the context and optimizing a smoothness loss (Jiang et al., 2020; Wu et al., 2023a) on it.

Masked-Fine-Tuning Above counterparts model 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.

LLM Similar to ReLM, CSC is a sentence rephrasing task for large language models (LLMs), where they rephrase the sentence in an autoregressive manner. However, we find that generative models suffer from the over-paraphrase issue. To address this, we use the prompt `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 find that applying masked-fine-tuning technique can boost the performance of Baichuan2-7b. We also instruct GPT4 (OpenAI, 2023) and Qwen2-72b (Bai et al., 2023; qwe, 2024) to perform this task through in-context learning with 5 shots. For each sentence, the in-context samples are uniformly chosen from sentences into 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`. We post-process the output of the LLMs to obtain the corrected sentence.

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

2.4 Training Setup

For all the experiments of BERT-based models, 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`. On ECSpell, we further 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 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.

Models show nice adeptness in addressing phonological and morphological errors. From results on ECSpell, We find that current state-of-the-art models perform perfectly (f1 more than 0.95) on phonological and morphological errors after domain-specific finetuning. We can also see that these two types of errors are less challenging for models under zero-shot learning, compared to the other types. It indicates that the similarity clues like pronunciations and shapes are rich in the training corpus for CSC models to fit the error model (Wu et al., 2023b).

A performance disparity emerges when models moving from addressing a single typo to multiple typos. For multi-typo errors, we find distinct trends between fine-tuned models and zero-shot models. Among the fine-tuned models, performances of all BERT-based models drops slightly when moving from addressing a single typo to multiple typos. This indicates that domain-specific fine-tuning can help train a better language modeling, making multi-typo errors less challenging. 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 conventional training process that researchers might overlook constructing

		Phono.	Morpho.	Non-s.	Single	Multi	Contextual	Overall
EC-LAW	BERT_{MFT}	99.1	99.0	97.1	98.2	93.4	94.9	94.0
	$\text{Soft-Masked}_{\text{MFT}}$	99.7	99.0	**99.9**	**99.4**	**97.0**	97.0	96.0
	$\text{MDCSpell}_{\text{MFT}}$	99.1	99.9	**99.9**	99.1	**97.0**	94.9	**97.1**
	$\text{CRASpell}_{\text{MFT}}$	99.3	99.0	99.0	98.5	95.2	97.0	95.6
	ReLM	**99.9**	**99.5**	96.2	98.8	96.4	**98.0**	95.6
	Baichuan2	93.6	92.3	94.3	92.4	85.7	80.8	92.8
	Qwen2-72b (5-shot)	85.7	85.9	74.0	84.7	62.6	59.1	72.7
	GPT4 (5-shot)	77.7	82.6	80.5	80.2	56.1	56.2	76.6
EC-MED	BERT_{MFT}	**99.7**	**99.4**	**98.6**	97.6	77.8	78.1	86.5
	$\text{Soft-Masked}_{\text{MFT}}$	98.8	97.0	94.3	95.2	87.9	**86.1**	87.4
	$\text{MDCSpell}_{\text{MFT}}$	98.6	99.4	93.3	96.4	87.0	84.3	88.7
	$\text{CRASpell}_{\text{MFT}}$	98.2	98.2	96.7	96.4	**92.6**	83.0	89.6
	ReLM	98.4	97.3	97.6	**98.3**	90.3	74.9	**89.9**
	Baichuan2	90.8	91.6	86.6	86.6	77.7	80.0	79.8
	Qwen2-72b (5-shot)	73.2	78.5	80.4	77.8	63.9	58.4	59.7
	GPT4 (5-shot)	74.5	80.4	74.9	77.1	62.1	59.9	66.4
EC-ODW	BERT_{MFT}	**97.1**	96.2	87.7	90.8	83.4	83.4	87.3
	$\text{Soft-Masked}_{\text{MFT}}$	96.3	**97.1**	85.7	90.7	89.7	86.1	88.4
	$\text{MDCSpell}_{\text{MFT}}$	96.7	96.2	**90.7**	**92.4**	89.2	87.0	90.4
	$\text{CRASpell}_{\text{MFT}}$	96.9	96.2	86.5	90.4	**92.3**	**90.3**	89.5
	ReLM	**97.1**	**97.1**	88.6	**92.4**	91.3	89.4	**91.6**
	Baichuan2	89.8	94.3	**92.1**	85.6	87.2	88.8	87.5
	Qwen2-72b (5-shot)	94.9	93.3	80.3	87.7	81.9	80.6	81.8
	GPT4 (5-shot)	87.1	83.9	75.5	76.6	71.6	61.8	73.3
LE-NEW	$\text{BERT}_{\text{MFT}}{}^{\dagger}$	71.3	72.0	45.0	63.9	11.3	49.3	56.0
	$\text{Soft-Masked}_{\text{MFT}}{}^{\dagger}$	71.8	72.1	42.8	64.0	10.8	50.4	55.6
	$\text{MDCSpell}_{\text{MFT}}{}^{\dagger}$	**74.9**	73.2	37.7	65.6	11.0	53.0	57.3
	$\text{CRASpell}_{\text{MFT}}{}^{\dagger}$	72.9	73.8	38.9	64.4	5.6	50.7	55.4
	$\text{ReLM}^{\dagger}$	**74.9**	**75.8**	44.0	**67.0**	10.2	52.2	58.8
	Qwen2-72b (5-shot)	64.4	69.2	**48.3**	60.0	**42.7**	55.3	57.4
	GPT4 (5-shot)	69.1	70.5	50.5	64.7	41.8	**67.7**	**63.4**
LE-ENC	$\text{BERT}_{\text{MFT}}{}^{\dagger}$	62.4	62.1	35.5	53.9	5.7	42.1	45.2
	$\text{Soft-Masked}_{\text{MFT}}{}^{\dagger}$	59.3	62.1	33.9	52.8	5.6	39.4	44.1
	$\text{MDCSpell}_{\text{MFT}}{}^{\dagger}$	**63.8**	66.7	33.7	54.7	7.3	41.4	46.1
	$\text{CRASpell}_{\text{MFT}}{}^{\dagger}$	62.8	68.1	39.2	56.8	4.9	43.3	47.6
	$\text{ReLM}^{\dagger}$	63.1	63.4	41.4	56.5	3.3	39.8	47.6
	Qwen2-72b (5-shot)	55.8	67.0	46.8	54.5	**36.7**	47.1	48.3
	GPT4 (5-shot)	61.1	**75.1**	**56.6**	**66.1**	35.4	**61.0**	**60.6**
LE-CAR	$\text{BERT}_{\text{MFT}}{}^{\dagger}$	74.1	65.9	45.3	64.5	4.2	47.5	51.9
	$\text{Soft-Masked}_{\text{MFT}}{}^{\dagger}$	73.6	67.4	47.1	64.5	7.6	46.8	52.2
	$\text{MDCSpell}_{\text{MFT}}{}^{\dagger}$	74.8	70.3	38.3	64.0	8.1	43.4	51.9
	$\text{CRASpell}_{\text{MFT}}{}^{\dagger}$	74.6	**71.8**	42.7	64.7	5.9	45.5	51.9
	$\text{ReLM}^{\dagger}$	**76.8**	66.3	45.0	**65.7**	9.7	44.7	53.5
	Qwen2-72b (5-shot)	55.7	61.7	40.2	49.5	30.4	44.6	45.5
	GPT4 (5-shot)	65.0	61.3	**52.0**	61.7	**33.2**	**50.1**	**56.5**

Table 1: Fine-grained performances on ECSpel (EC-x) and LEMON (LE-x). We report the F1 score for each error type and the overall F1 score on all sentences. "Non-s." refers to the non-similarity errors. † refers to the zero-shot performance of the corresponding models. The subscription MFT indicates that the model is trained using masked-fine-tuning.

samples that contain multi-typo errors, resulting in models' inability during testing.

Contextual errors pose a consistent challenge in every scenario. For finetuned models, contextual errors remain challenging, particularly in the domain of medical treatment (MED). On average, the F1 performance on contextual errors drops by 7.1 points compared to the overall F1 score across five BERT-based methods. However, for zero-shot models, all of them struggle with contextual errors. Correspondingly, their performance on non-similarity errors also encounters a big decline. The poor performance in handling non-similarity errors and contextual errors from LEMON highlights the importance of domain-specific knowledge and features for spelling correction. This indicates that open-domain CSC is the greatest challenge currently faced by the community.

LLMs show potential in open-domain CSC, but there is room for improvement in handling phonological errors. We find that the few-shot performances of Qwen-72b and GPT-4 on ECSpell are weaker than those of fine-tuned BERT-based models. However, on LEMON, an open-domain benchmark, their performances surpass those of the BERT-based models, particularly in handling multi-typo and contextual errors. This is mainly due to their strong reasoning ability and the extensive knowledge acquired during pre-training. Nonetheless, their performance on phonological typos is weaker than that of BERT-based models, which are trained on 34 million synthesized examples using a confusion set. **This fine-grained comparison suggests directions for further open-domain CSC research.** For LLMs, incorporating phonological similarity could enhance their performance in CSC. Additionally, equipping BERT-based models with more knowledge is crucial, and data augmentation using LLMs can be a potential solution.

Based on Fibench, we have the following conclusions. Firstly, the performance of CSC models fine-tuned on domain-specific data is quite high. However, open domain CSC, which is more representative of real-world applications, remains challenging and warrants further study. Secondly, **existing CSC models exhibit deficiencies in addressing two specific types of errors, bottlenecking their overall performance in practical applications.** However, sentences that comprise contextual and multi-typo errors are rare in typical training sets. Therefore, there emerges a very need for methods to generate them artificially, which forms the follow-up section.

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 by checking its existence in a word vocabulary.

	Pin	Mor	Non-sim.	Sin.	Multi.	Context.
LAW	46	4	141	74	58	132
MED	52	16	138	52	76	128
ODW	65	15	156	79	78	157

Table 2: Statistics of the generated contextual errors.

From Table 2, we can find that the generated contextual errors contain more non-similarity and multi-typo examples, which are also more challenging for CSC models. This demonstrates that our error generation method can produce additional training examples specifically designed to address the weaknesses of current CSC models.

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 $\sim$ 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.

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. 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 on 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 multi-typo error type. 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.

	LAW		**MED**		**ODW**	
	Con	All	Con	All	Con	All
ReLM	98.0	95.6	74.9	89.9	89.4.	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
BERT	94.9	94.0	78.1	86.5	83.4	87.3
BERT$^{\clubsuit\text{domain}}$	**95.9**	**95.5**	**89.2**	**89.5**	**85.7**	**90.1**
BERT$^{\clubsuit\text{wiki}}$	93.0	94.9	86.1	88.9	77.7	88.3
	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$^{\clubsuit\text{FS}}$	15.7	56.6	**14.1**	46.2	15.4	52.1

Table 3: Results after data augmentation. "CT" refers to continue-training and "FS" refers to few-shot.

4.2 Result

In this section, we conduct experiments on masked-fine-tuned BERT and ReLM, which are tagging and rephrasing models respectively. The upper part of Table 3 showcases the effectiveness of incorporating new contextual errors. Significant performance

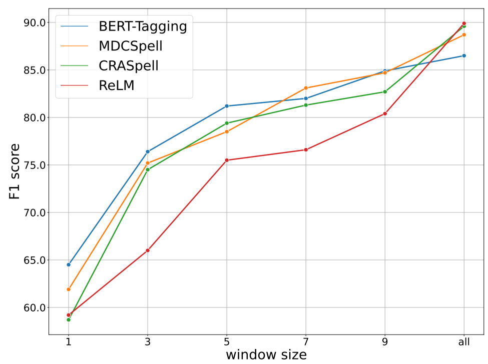

Figure 4: The variation of F1 score with the local context size. We choose EC-Med as the representative domain

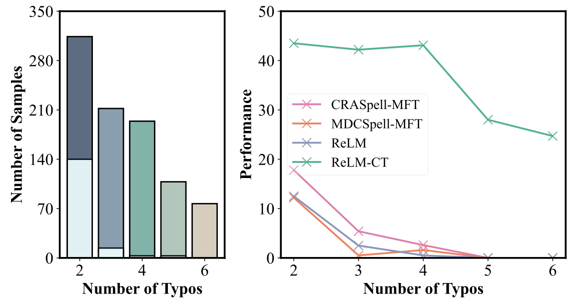

Figure 5: Left: Statistics of the number of typos in each example. Right: Variation of performances (F1) with the increasing number of typos. We choose LE-ENC as the representative domain.

improvement can be observed in the domains of MED and ODW. For instance, on MED, the performance on contextual errors of ReLM increases from 74.9 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 3, 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 boost from 10.2 to 18.7 on NEW, 3.3 to 12.9 on ENC, and 9.7 to 22.0 on CAR respectively. The improvement brought by few-shot learning is also notable. 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.

5 Further Analysis

5.1 Analysis of Contextual Errors

As discussed in Section 2, contextual errors present significant challenges for CSC models. To analyze the impact of context on model predictions, we truncate the local phrases surrounding the typo and examine how varying the truncation window size affects CSC models' performance. Specifically, we symmetrically truncate the source sentence by retaining only the $2n-1$ neighboring words around the erroneous characters, then calculate the F1 score for these truncated samples.

From Figure 4, we find unsurprisingly that performance of all the models improve with the growth of context size. Meanwhile, ReLM, which significantly outperforms the baseline model BERT-Tagging, performs worse than BERT-Tagging when the context size is below 9. This indicates that ReLM, with its rephrasing training objective, is more dependent on the entire sentence for making corrections rather than relying on the local words.

5.2 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. Considering that multi-typo errors with more than two typos are sparse in the test set of ECSpell, we supplemented the test set with additional examples generated using the method described in Section 3 to investigate the influence of the number of typos in a single sentence.

The results are depicted in Figure 5. Intuitively, all models experience a decline in performance when the number of typos rises. Among tagging models, CRASpell outperforms other counterparts, suggesting that optimizing the smoothness loss during training effectively allows the model to adapt to multi-typo errors. We also find that continue-training with more multi-typo errors can significantly improve the performance on multi-typo errors. The F1 score of ReLM keeps above 0.4 with less than 4 typos in one sentence, which demonstrates the effectiveness of our data augmentation method.

5.3 Case Study

We further offer a closer look on concrete cases. The case study comprises two parts. We first

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: Case study.

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.

6 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.

7 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 in both finetuning and open-domain benchmarks. The nuanced assessment offers a clear view of each model's strengths and weaknesses, especially for LLMs, 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 demonstrate that continue-training on these augmented examples can enhance the corresponding aspect of CSC models. We also study the impact of context and number of typos to CSC models.

8 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 open-domain benchmark and 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 aim to study how to complement the strengths of BERT-based models in phonetic similarity, generation stability, and efficiency with the powerful semantic and knowledge capabilities of large language models (LLMs), achieving a synergy of their respective advantages. Lastly, 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

2024. Qwen2 technical report.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

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.