CG-CSC: A Counterfactual Generation Method for Improving Chinese Spelling Correction

Anonymous ACL submission

Abstract

Chinese Spelling Correction (CSC) aims to detect and correct misspelled characters in Chinese text, a prerequisite for reliable downstream Natural Language Processing (NLP) applications. Although existing methods have achieved promising performance, they still suffer from spurious correlations caused by long-tailed data distributions, leading to over-correction of head-frequency mappings and under-correction of rare or unseen mappings. To address this, we propose **Counterfactual Generation for Chinese Spelling Correction** (CG-CSC), a causally grounded framework that synthesises counterfactual pairs to balance the training data distribution. Experimental results on three widely used SIGHAN benchmarks show that our method significantly improves correction performance, particularly on rare and out-of-training cases, demonstrating enhanced robustness and generalization.

1 Introduction

Chinese Spelling Correction (CSC) is a fundamental Natural Language Processing (NLP) task (Gao et al., 2010), it aims at automatically detecting and correcting misspelled characters in Chinese text, with critical applications in Optical Character Recognition (OCR) (Afli et al., 2016), Automatic Speech Recognition (ASR) (ERRATTAHI et al.), and AI-based language processing systems (Dong and Zhang, 2016). Despite recent progress driven by pre-trained language models (PLMs) and large language models (LLMs), a persistent challenge arises from spurious correlations that degrade correction robustness, which are statistical associations learned from unbalanced training data.

Current CSC models trained by supervised learning on labeled parallel datasets are fundamentally constrained by the training datasets that have an unbalanced distribution (Hong et al., 2019; Zhang et al., 2020; Huang et al., 2021; Zhu et al., 2022; Huang et al., 2023; Xu et al., 2021; Wu et al., 2024a). These datasets exhibit *long-tailed distributions*, where a small fraction of frequent "misspelled-to-correction" mappings dominate the training instances while many rare mappings are underrepresented. For instance, Figure 1(a) shows that 21% of head mappings in SIGHAN15 capture 60% of the training data, leaving the remaining 79% of mappings (including low-frequency and out-of-training (OOT) mappings highlighted in red in Figure 1(a)) undersampled. This leads to two critical issues:

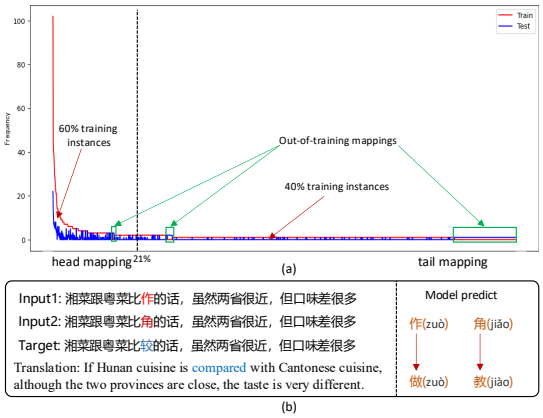

Figure 1: (a) The distribution of "misspelled-to-correction" mappings in the SIGHAN15 training/test dataset. The red box highlights the out-of-training (OOT) mappings that only appear in the test set. (b) An example of spurious correlation from the CSC model. The misspelled/ground-truth characters are highlighted in red/blue.

1. **Over-correction fueled by spurious correlations**: Models prioritize high-frequency mappings, even when semantically invalid. For example, Figure 1(b) shows a trained model correct the "作" (zuò, meaning abstract behavior or result) to "做" (zuò, meaning "concrete thing or action") or change "角" (jiǎo, meaning "role") to "教" (jiào, meaning "teach"), where "作→做" and "角→教" are high-frequency mappings.

(2) **Under-correction due to distributional bias**: Low-frequency or OOT mappings (e.g., in Figure 1(b), "作" (zuò) or "角" (jiǎo) need to be corrected as "较" (jiào, meaning "compare"), where "作→较" and "角→较" are low-frequency or OOT mappings.) are ignored because the model never observed their mappings during training. Such mappings represent unseen semantic relationships in training data, hindering generalization.

Recent LLM-based approaches, such as ARM (Liu et al., 2024), enhance the performance of supervised-learning-based CSC model (Zhang et al., 2020; Zhu et al., 2022; Li et al., 2022) by adjusting the probability distributions of character predictions using LLMs. While these methods are effective to some extent—leveraging additional contextual information to guide the corrector toward accurate predictions—they fail to address the root cause: the inherent tendency of supervised CSC models to favor high-frequency mappings statistically, rather than to infer corrections based on the causal relationship between context and character semantics.

To address these challenges, we propose a novel **C**ounterfactual **G**eneration method for CSC (**CG-CSC**). We first explain the relationship between the correction result and various dependent features from a causal perspective, and then introduce a counterfactual mechanism to mitigate these spurious correlations from unbalanced distribution training data. Specifically, we design a soft-sampler to synthesize counterfactual pairs (e.g., "作→较") by exploiting character-level features such as pronunciation similarity and glyph structure. These generated instances allow the model to awake additional mappings and thereby enhance its robustness. We evaluate our method on three widely used CSC benchmark datasets (SIGHAN13, SIGHAN14, SIGHAN15), where CG-CSC outperforms the PLM fine-tuned baseline on all metrics on all three benchmarks. Additionally, an ablation study highlights the significance of counterfactual texts on CSC performance.

2 Related Work

2.1 Causal Inference

Causal inference aims to eliminate confounders between variables to determine causal effects (Pearl, 2009). As a result, it has become an effective method for debiasing in various fields, including computer vision (Niu et al., 2021), recommendation systems (Zhang et al., 2021c), and natural language processing (Tian et al., 2022). These works are mainly inspired by counterfactual reasoning and causal intervention. For example, (Niu et al., 2021) proposed addressing language bias in visual question answering by subtracting the outcomes of a counterfactual language-only model from those of a standard language-vision model. Additionally, counterfactual reasoning is widely employed to address spurious correlations between inputs and labels in various tasks, including natural language understanding (Tian et al., 2022; Wu et al., 2024c), Named Entity Recognition (Yang et al.; Zhang et al., 2021b), Sentiment Analysis (Wu et al., 2024b). (Liu et al., 2022) proposed a method to de-confound objects from their context in object detection using backdoor adjustment. This approach involves approximating inverse probability weights to estimate the do-operator.

In this paper, we adopt a causal perspective to analyze and identify that spurious correlations in existing CSC models are a key bottleneck impeding further improvements in model performance. These spurious correlations primarily stem from unbalanced training datasets. To address this issue and enable the model to uncover the main causal effects, we propose using counterfactual generation for CSC.

2.2 Chinese Spelling Correction

Chinese spelling correction (CSC) has received extensive attention over the last two decades due to its uniqueness and challenges, and many works have significantly improved the performance (Hong et al., 2019; Zhang et al., 2021a; Huang et al., 2021; Zhu et al., 2022; Zhang et al., 2020; Huang et al., 2023). Especially, inspired by masked language model (MLM) (Devlin et al., 2019), which predicts each character to be corrected using the fusion features from contextual, phonological, and visual information. For example, (Cheng et al., 2020) reported SpellGCN, which integrates phonological and visual similarity information into character classifiers using a graph network, which then feeds the graph representation into MLM. To make full use of information from all dimensions, ReaLise is proposed by (Xu et al., 2021), they employ three distinct feature networks to capture phonetic, graphemic, and semantic features, ultimately passing the fused representation through MLM. These methods focus on constructing CSC features and feeding them into an MLM-based corrector. In

contrast, FASPell (Hong et al., 2019) leverages phonological and visual similarity features to construct a filtering model, selects the most suitable candidate Chinese characters from a pre-trained Language Model (PLM). In addition, some methods (Zhang et al., 2020, 2021a; Huang et al., 2023; Zhu et al., 2022) follow the detection-correction framework, which first uses an error detection module to detect the position information of misspelled characters and then feeds the detection results to the correction module to get predictions. Specifically, SoftMasked-BERT (Zhang et al., 2020) uses a two-stage detection and correction pipeline method, which linearly combines each token embedding with the embedding of [MASK], and predicts the error character based on a fine-tuned masked language model. MDCSpell (Zhu et al., 2022) uses parallel detection and correction feature representation modules, and the corrector receives the detector's hidden states, thus, the inference in correction incorporates the feature from both detection and correction. Different from those works, MLM-phonetics (Zhang et al., 2021a) and DR-CSC (Huang et al., 2023) further introduce phonological and visual information into the detection-correction framework.

Despite their notable success, these methods often struggle to achieve further performance gains when faced with unbalanced training data. When confronted with unbalanced training datasets, these models tend to focus disproportionately on high-frequency mappings, persistently memorizing these common patterns. As a result, they create spurious correlations between character semantic representations and predicted corrections, often neglecting corrections for low-frequency mappings. To address this issue, recent work such as ARM (Liu et al., 2024) leverages LLM to refine the correction probabilities of existing CSC models. While ARM has demonstrated some improvements, it does not fundamentally resolve the underlying challenge. In contrast, our work takes a causal perspective to analyze the limitations of current models and proposes a counterfactual generation-based soft-sampler that synthesizes balanced training data, leading to more robust and effective CSC performance.

3 Methodology

3.1 Problem Formulation

The Chinese spelling correction (CSC) task aims to detect and correct the misspelled characters in a Chinese sentence. Formally, given an input textual sequence $\boldsymbol{x} = (w_1, w_2, \cdots, w_n)$, where each w_i is a character from a predefined vocabulary $\boldsymbol{V}$, the CSC model's goal is to produce a corrected sequence $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$, with each $y_i \in \boldsymbol{V}$ representing the suggested correction for the corresponding character.

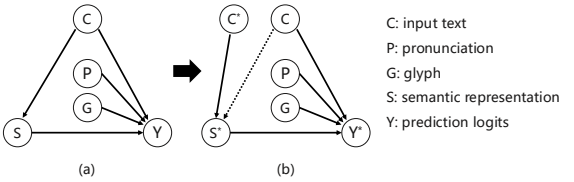

Figure 2: (a) a unified structured causal model for Chinese spelling correction. (b) causal interventions on semantic representation S.

3.2 A Causal View

In the task of CSC, the unbalanced training data introduces serious data bias, leading to spurious correlation, and then misleads the error correction of models. To understand the causal relationships, we introduce a Structural Causal Model (SCM) (Pearl, 2009) to reformulate existing methods based on sequence labeling. As Fig. 2 (a) shows, S represents the semantic representation of a character, P and G represent phonetic and glyph, Y is the predicted character, and C is the confounders, such as training data distribution bias.

To formalize, we use $\mathcal{G} = \{\mathbb{V}, \mathbb{F}, U\}$ to express SCM, where $\mathbb{V} = \{V_1, \ldots, V_n\}$ denotes the set of observables (vertices), $\mathbb{F} = \{f_1, \ldots, f_n\}$ denotes the set of functions (edges), and $U = \{U_1, \ldots, U_n\}$ is the set of exogenous variables (e.g., noise), n is the number of nodes in $\mathcal{G}$. Furthermore, we format the causal effects of each variable $\boldsymbol{H}$ in $\mathcal{G}$ on Y as linear transformations, then the prediction can be obtained by summation:

$$Y_s = \sum_{i \in \mathbb{N}} \boldsymbol{W}_{iY} \boldsymbol{H}_i \tag{1}$$

where, $\mathbb{N} = \{S, P, G\}$ denotes the parents of Y, $\boldsymbol{W}$ is learnable weight.

In CSC task, the path $S \rightarrow Y$ represents the model directly classifies the characters based on the semantic representation, and gets the predicted character. The causal path $S \leftarrow C \rightarrow Y$ denotes confounders, such as training samples distribution bias, mislead the calculation of character semantic representation, resulting in the spurious correction between character semantic representation

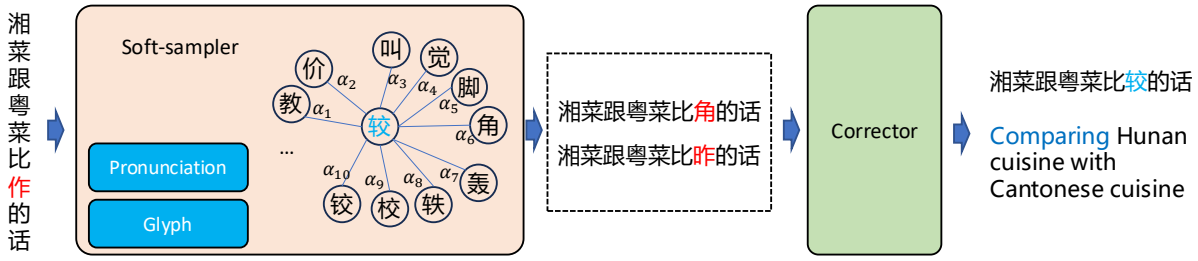

Figure 3: Overview of our CG-CSC framework. The soft-sampler is used to generate counterfactual training instances, which uses a Chinese confusion set as an external resource. The symbol α represents sampling probabilities.

and predicted correction. This spurious correction leads to over-correction, i.e., the model shows stubborn memory, such as the correction mapping 作→做. So, we expect to block the backdoor path $S \leftarrow C \rightarrow Y$, thus, we intervene on S.

3.3 Casual Intervention for CSC

In this paper, we propose the use of counterfactual generation for CSC.

Counterfactuals The concept of counterfactual reflects an imaginary secario for "what would the outcome be had the variable(s) been different". Let $Y \in \boldsymbol{V}$ denote the outcome variable, and let $S \in \boldsymbol{V} \backslash \{Y\}$ denote the variable of study. The counterfactual is obtained by setting $S = s^*$ and formally estimated as:

$$Y_{s^*}(u) = Y_{\mathcal{G}_{s^*}}(u) \tag{2}$$

where, $\mathcal{G}_{s^*}$ denotes all functions of SCM $\mathcal{G}$ assign $S = s^*$. The counterfactual Y_{s^*} of the original instance-level prediction Y_s is computed as:

$$\begin{aligned} Y_{s^*} &= f_Y(do(S = s^*), G = g, P = p) \\ &= \sum_{i \in \mathbb{N} \backslash \{S\}} \boldsymbol{W}_{iY} \boldsymbol{H}_i + \boldsymbol{W}_{SY} \boldsymbol{H}_{s^*} \end{aligned} \tag{3}$$

where, the function f_Y is used to computes Y. Thus, we only replace the character semantic representation $\boldsymbol{H}_S$ with $\boldsymbol{H}_{s^*}$.

Counterfactual generation with soft-sampler To update CSC models, we design a soft-sampler to synthesize counterfactual pairs and then train the model on the synthesized dataset. Specifically, we choose a high-quality confusion set to replace the semantic representation of the character and change the model's stubborn mapping memory. This set includes characters that are highly similar to the target character in both pronunciation and glyph. Additionally, the similarity in pronunciation and glyph has been proven effective on CSC tasks (Cheng et al., 2020; Zhu et al., 2022; Liang et al., 2023; Huang et al., 2023). Inspired by counterfactuals, we seek to determine the decisive factors influencing character corrections in the CSC task. As illustrated in Fig. 3, we replace misspelled characters in the training set sentences using a character from the selected confusion set, i.e., we feed C^* to the function of the edge $C \rightarrow S$ to get S^*. This approach aims to eliminate key clues in the semantic representation S^* of the counterfactual, thereby enhancing the model's focus on the main effect while reducing spurious correlations.

The core of the soft-sampler is to build a sampling dictionary. Specifically, it includes three calculation operations: (1) counting the frequency c_{ij} of each correction mapping in the training data set, where i and j denote the character id from the predefined vocabulary $\boldsymbol{V}$ and c_{ij} means correct character i as j. (2) Count the mappings in the confusion set, with an update frequency of $c_{ij} \propto c_{ij} + 1 * \lambda$, where taking into account the difference in the impact of similar pronunciation and similar glyphs, λ is set to 0.6 for same and similar pinyin mapping and 0.4 for same stroke mappings. (3) Calculate soft sampling probability by normalization:

$$p_{ij} = c_{ij} / \sum_{j=1} c_{ij} \tag{4}$$

Then, the sampling dictionary is used to synthesize counterfactual instances. For each training instance, we employed two parameter generators to generate two control parameters $q_1 \in [0, 1]$ and $q_2 \in (0, 1]$, which are used to determine the synthesis ratio and extract the characters to be corrected, respectively.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets To assess the effectiveness of CG-CSC for the CSC task, we perform extensive empir-

ical evaluations on three widely used datasets: SIGHAN13 (Wu et al., 2013), SIGHAN14 (Yu et al., 2014) and SIGHAN15 (Tseng et al., 2015). Specifically, we evaluate model with three test datasets from SIGHAN13, SIGHAN14 and SIGHAN15 and train CG-CSC by four datasets, which include three training data sets from SIGHAN and an additional set of training data generated by an automatic method (Wang et al., 2018). For consistency, we apply the same preprocessing procedure as (Zhu et al., 2022), which includes converting the characters in these datasets to simplified Chinese using OpenCC [1].

Evaluation Metrics In this paper, we use sentence-level metrics to evaluate the performance of the trained model on CSC. Unlike character-level metrics, sentence-level metrics impose a stricter standard: a prediction is considered correct only if all misspelled characters in the sentence have been detected or corrected. Following the previous work (Hong et al., 2019; Cheng et al., 2020), we take the commonly used sentence-level precision, recall, and F1 score measures.

4.2 Implementation details

In the implementation of CG-CSC, we use PyTorch (Paszke et al., 2019) as the underlying framework and build the model with Transformers library (Wolf et al., 2020). For model training, AdamW (Loshchilov and Hutter, 2019) is used as an optimizer with max epochs 20, the learning rate is set as 5e-5, the batch size is set to 32. The parameter q_1 is set to 0.8, indicating that the model is trained on a combination of the original dataset and an additional counterfactual dataset that is 80% the size of the original. Set q_2 to 0.6, meaning that when sampling from the confusion set, there is a 60% probability of selecting characters with similar pronunciation and a 40% probability of selecting characters with similar glyphs. To ensure a fair comparison with existing methods, we follow ARM (Liu et al., 2024) and integrate CG-CSC with three supervised-learning-based CSC models: SoftMasked-BERT (Zhang et al., 2020), MDCSpell (Zhu et al., 2022) and SCOPE (Li et al., 2022), where we use the official code [2] and parameters to implement SCOPE. All experiments are conducted on a GPU server equipped with two RTX A6000 GPUs (48 GB each).

[1] https://github.com/BYVoid/OpenCC

[2] https://github.com/jiahaozhenbang/SCOPE

4.3 Baselines

Considering that CG-CSC is essentially a fine-tuning model, we select a wide range of CSC methods based on fine-tuning as comparison models:

- SoftMasked-BERT (Zhang et al., 2020), integrates a detection network with a BERT-based correction network via a soft-masking mechanism.
- MDCSpell (Zhu et al., 2022) is a detector-corrector multi-task framework for CSC that leverages BERT to retain visual and phonological features from raw input and applies a late fusion strategy to mitigate the influence of misspellings.
- SCOPE (Li et al., 2022) introduces a dual-decoder architecture with adaptive task weighting and an iterative correction strategy, leveraging a fine-grained auxiliary CPP task.
- DR-CSC (Huang et al., 2023) includes a modular detection-and-reasoning component that decomposes CSC into detection, reasoning, and searching subtasks, enabling the integration of external knowledge and improving both performance and interpretability across various non-autoregressive models.
- SpellGCN (Cheng et al., 2020) incorporates phonological and visual similarity knowledge representation into BERT by employing a specialized graph convolutional network.
- DCN (Wang et al., 2021) aims to optimize the incoherent problem, they use a dynamically connected network to measure the degree of dependence between any two adjacent Chinese characters.
- DORM (Liang et al., 2023) allows the direct feature interaction between textual and phonetic information.
- ARM (Liu et al., 2024) denotes a plug-and-play Alignment-and-Replacement Module that leverages LLMs to enhance semantic understanding in CSC while mitigating over-correction issues, improving existing models. It has the same motivation as CG-CSC.

Dataset	Models	Detection			Correction		
		Prec.	Rec.	F1	Prec.	Rec.	F1
SIGHAN13	SpellGCN(Cheng et al., 2020)	80.1	74.4	77.2	78.3	72.7	75.4
	DCN(Wang et al., 2021)	86.8	79.6	83.0	84.7	77.7	81.0
	DORM(Liang et al., 2023)	87.9	83.7	85.8	86.8	82.7	84.7
	DR-CSC(Huang et al., 2023)	88.5	83.7	86.0	**87.7**	83.0	85.3
	SoftMasked-BERT(Zhang et al., 2020)	85.2	78.0	81.4	83.8	76.8	80.1
	+ARM(Liu et al., 2024)	85.9	79.5	82.6	84.6	78.2	81.3
	+CG-CSC	85.6	79.5	82.4	84.5	78.8	81.6
	MDCSpell(Zhu et al., 2022)	85.7	78.5	81.9	84.6	77.5	80.9
	+ARM(Liu et al., 2024)	86.4	79.5	82.8	85.5	78.6	81.9
	+CG-CSC	86.1	79.3	82.6	85.1	78.8	81.8
	SCOPE(Li et al., 2022)	87.4	83.4	85.4	86.3	82.4	84.3
	+ARM(Liu et al., 2024)	**88.7**	84.1	86.3	87.6	83.1	85.3
	+CG-CSC	88.0	**85.1**	**86.5**	86.8	**84.1**	**85.9**
SIGHAN14	SpellGCN(Cheng et al., 2020)	65.1	69.5	67.2	63.1	67.2	65.3
	DCN(Wang et al., 2021)	67.4	70.4	68.9	65.8	68.7	67.2
	DORM(Liang et al., 2023)	69.5	73.1	71.2	68.4	71.9	70.1
	DR-CSC(Huang et al., 2023)	70.2	73.2	71.7	69.3	72.3	70.7
	SoftMasked-BERT(Zhang et al., 2020)	69.6	69.6	69.6	68.5	68.5	68.5
	+ARM(Liu et al., 2024)	70.4	71.3	70.9	69.3	70.2	69.7
	+CG-CSC	70.0	72.8	71.4	68.7	71.7	70.2
	MDCSpell(Zhu et al., 2022)	66.2	66.5	66.3	64.2	64.6	64.4
	+ARM(Liu et al., 2024)	67.3	68.8	68.1	65.4	66.9	66.2
	+CG-CSC	67.1	69.1	68.1	64.8	70.1	67.3
	SCOPE(Li et al., 2022)	70.1	73.1	71.6	68.6	71.5	70.1
	+ARM(Liu et al., 2024)	**71.2**	**75.0**	**73.1**	69.2	**73.0**	71.1
	+CG-CSC	70.9	74.7	72.8	**69.6**	72.9	**71.2**
SIGHAN15	SpellGCN(Cheng et al., 2020)	74.8	80.7	77.7	72.1	77.7	75.9
	DCN(Wang et al., 2021)	77.1	80.9	79.0	74.5	78.2	76.3
	DORM(Liang et al., 2023)	77.9	84.3	81.0	76.6	82.8	79.6
	DR-CSC(Huang et al., 2023)	**82.9**	84.8	83.8	**80.3**	82.3	81.3
	SoftMasked-BERT(Zhang et al., 2020)	75.5	79.2	77.3	74.1	77.8	75.9
	+ARM(Liu et al., 2024)	76.4	80.9	78.6	74.7	79.0	76.8
	+CG-CSC	75.8	81.4	78.5	74.4	79.5	76.9
	MDCSpell(Zhu et al., 2022)	76.3	79.6	77.9	75.2	78.5	76.8
	+ARM(Liu et al., 2024)	76.4	81.3	78.8	75.2	80.0	77.5
	+CG-CSC	76.0	82.2	79.0	75.1	80.3	77.6
	SCOPE(Li et al., 2022)	81.1	84.3	82.7	79.2	82.3	80.7
	+ARM(Liu et al., 2024)	82.3	86.1	84.1	79.5	83.1	81.3
	+CG-CSC	81.6	**87.5**	**84.4**	79.0	**84.8**	**81.8**

Table 1: Experimental results on SIGHAN13, SIGHAN14 and SIGHAN15 test sets, and each model includes sentence-level precision, recall, and F1 score for both detection and correction. For a fair comparison, we select SoftMasked-BERT(Zhang et al., 2020), MDCSpell(Zhu et al., 2022) and SCOPE(Li et al., 2022) as backbone models and integrate them with CG-CSC. The highest score for each evaluation metric is highlighted in **bold**.

4.4 Main Results

Table 1 shows the experimental results. Across all three SIGHAN datasets and backbones, our proposed CG-CSC consistently achieves the best or highly competitive F1 scores for both detection and correction. These results highlight the robustness and effectiveness of counterfactual generation in enhancing CSC performance.

In the same setting (e.g., SoftMasked-BERT, MDCSpell and SCOPE as baselines and backbones), compared with "+ARM" that uses LLM to adjust the corrector prediction probability, all backbones bring more improvements in correction F1 after integrating CG-CSC. Addition, CG-CSC significantly improves detection and correction F1 compared to the three base models. Specifically, correction F1 +1.5%, +1.7%, +1.0% on SIGHAN13/14/15 with SoftMasked-BERT, +0.9%,

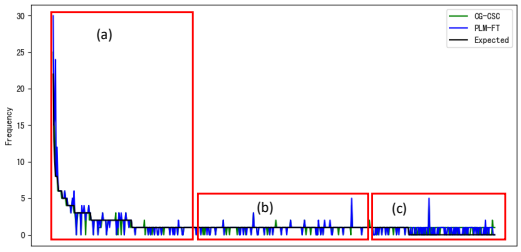

Figure 4: Correction mapping frequency comparison.

+2.9%, +0.8% with MDCSpell, and +1.6%, +1.1%, +1.1% with SCOPE, respectively. These gains validate the value of augmenting training with counterfactual samples, particularly for rare or unseen errors.

Furthermore, CG-CSC outperforms or matches these models that include the pre-trained stage, such as DORM and DR-CSC. Despite adjusts the distribution of training data and does not explicitly model CSC, CG-CSC achieves superior results, demonstrating that counterfactual generation contributes substantially to performance improvements.

Notably, while methods like DCN and ARM explicitly model the output sequence, our CG-CSC does not alter the output mechanism, yet it consistently surpasses both in F1 scores across all three datasets. This indicates that CG-CSC enhances generalization and robustness by exposing the model to a more diverse and balanced set of training signals.

Taken together, these findings confirm that CG-CSC is a highly effective and versatile training strategy that enhances various supervised-learning-based backbone models and consistently boosts CSC performance across diverse benchmarks.

4.5 Ablation Study

4.5.1 The efficiency of counterfactual generation

We use SIGHAN15 test sets to examine the effectiveness of our model in eliminating spurious correlations.

Fig. 4 shows the comparison of correction mapping frequency between the expected target, PLM-FT (ChineseBERT), and the proposed CG-CSC. In region (a) of Fig. 4, which represents high-frequency mappings, PLM-FT tends to over-correct significantly beyond the true frequency, while CG-CSC remains relatively stable. In region (b), which denotes low-frequency mappings, both models exhibit under-correction, and PLM-FT is less robust than CG-CSC, especially with two strange high-frequency mappings. In region (c), which includes characters that do not require correction but the model over-corrects. It is noteworthy that PLM-FT corrects more in this region than CG-CSC. The result shows that CG-CSC have a better ability to tackle both over-correction and under-correction problems.

4.5.2 Counterfactual generation vs. data augmentation

To distinguish the difference between the proposed CG-CSC and the data augmentation method, we adopt a random sampling strategy from predefined vocabulary $\boldsymbol{V}$ and set the same probability of 0.8 to decide whether a misspelled character is replaced by another random character. The comparison results are shown in Table 2:

Overall, Both Data Augmentation (SCOPE+DA) and the proposed Counterfactual Generation strategy (SCOPE+CG-CSC) consistently improve the baseline SCOPE model across all three SIGHAN benchmarks (SIGHAN13/14/15), in terms of both detection and correction performance. Specifically, SCOPE+CG-CSC achieves the best performance in all metrics across all datasets. For instance, on SIGHAN13, it achieves a detection F1 of 86.5% and a correction F1 of 85.9%, significantly outperforming both the original SCOPE and SCOPE+DA. On SIGHAN14, although all methods perform relatively poorly compared to other datasets, SCOPE+CG-CSC still leads with a detection F1 of 72.8% and a correction F1 of 71.2%, demonstrating its robustness even in more challenging scenarios. On SIGHAN15, SCOPE+CG-CSC again surpasses other variants, reaching a detection F1 of 84.4% and a correction F1 of 81.8%.

These results indicate that the proposed counterfactual data generation method is more effective than traditional data augmentation in improving the generalization and robustness of the CSC model.

5 Case study

To further demonstrate the effectiveness of CG-CSC over existing models, we analyze two representative cases in Table 3.

In the first case, the sentence contains a rare error where "角(jiǎo)" (meaning "role") should be corrected to "较(jiào)" (meaning "compare"). The backbone model SCOPE failed to detect or correct the misspelled character, resulting in under-correction. This is an example of under-correction caused by distributional bias (the mapping "角 → "

Dataset	Baseline	Detection			Correction		
		Prec.	Rec.	F1	Prec.	Rec.	F1
SIGHAN13	SCOPE(Li et al., 2022)	87.4	83.4	85.4	86.3	82.4	84.3
	SCOPE+DA	87.1	84.0	85.5	85.9	82.7	84.3
	SCOPE+CG-CSC	**88.0**	**85.1**	**86.5**	**86.8**	**84.1**	**85.9**
SIGHAN14	SCOPE(Li et al., 2022)	70.1	73.1	71.6	68.6	71.5	70.1
	SCOPE+DA	70.3	73.1	71.7	68.2	72.0	70.0
	SCOPE+CG-CSC	**70.9**	**74.7**	**72.8**	**69.6**	**72.9**	**71.2**
SIGHAN15	SCOPE(Li et al., 2022)	81.1	84.3	82.7	79.2	82.3	80.7
	SCOPE+DA	81.2	84.7	82.9	78.7	82.8	80.7
	SCOPE+CG-CSC	**81.6**	**87.5**	**84.4**	**79.0**	**84.8**	**81.8**

Table 2: Performance of SCOPE with Data Augmentation (SCOPE+DA) and Counterfactual Generation (SCOPE+CG-CSC) on SIGHAN13/14/15.

id	#Model/Function	Sentence
1	Input	湘菜跟粤菜比角(jiǎo)的话
	SCPOE	湘菜跟粤菜比角(jiǎo)的话
	+ARM(Liu et al., 2024)	湘菜跟粤菜比较(jiào)的话
	+CG-CSC	湘菜跟粤菜比较(jiào)的话
	Translation	Comparing Hunan cuisine with Cantonese cuisine
2	Input	团长将士兵布置(bù zhì)在城外，让他们安(ān)兵不动。
	SCPOE	团长将士兵布置(bù zhì)在城外，让他们安(ān)兵不动。
	+ARM(Liu et al., 2024)	团长将士兵布置(bù zhì)在城外，让他们按(àn)兵不动。
	+CG-CSC	团长将士兵部署(bù shǔ)在城外，让他们按(àn)兵不动。
	Translation	The captain deployed the soldiers outside the city and ordered them to stand still.

Table 3: Examples of CSC results from SCOPE+CG-CSC, in comparison with results from SCOPE and SCOPE+ARM baselines. Red and blue are used to mark misspelled and correct characters, respectively.

only appears once in training data). In contrast, both ARM and CG-CSC successfully detected and corrected the error, demonstrating their ability to handle low-frequency or out-of-training (OOT) mappings effectively.

In the second case, the input contains two spelling errors: "布置(bù zhì)" should be "部署(bù shǔ)" (meaning "deploy"), and "安(ān)" should be "按(àn)" (meaning "according to"). Both SCOPE and ARM failed to correct all two misspelled characters, most likely because the input had multiple misspelled characters, causing the model to be distracted in understanding the meaning of the input sentence. Specifically, SCOPE failed to correct "布置(bù zhì)" due to its low frequency in the training set, leading to under-correction. While ARM managed to detect and correct "安(ān)" to "按(àn)", only CG-CSC accurately corrected both errors. These results highlight the ability of CG-CSC to mitigate both over-correction and under-correction, benefiting from a more balanced training distribution achieved via counterfactual generation.

Together, these cases illustrate that CG-CSC improves robustness and generalization by exposing the model to diverse, informative correction instances beyond what is seen in naturally skewed training data.

6 Conclusion

In this paper, we examine the issue of erroneous corrections in the context of the causal perspective on the CSC task. We identify that an unbalanced distribution of training datasets can lead to spurious correlations between character semantic representations and predicted corrections. To address this, we propose CG-CSC, a method based on causal interventions for CSC. We then develop CG-CSC with counterfactual generation and conduct comprehensive experiments to validate its effectiveness. The experimental results demonstrate that CG-CSC exhibits competitive performance and robustness. Furthermore, the method based on counterfactual generation holds significant potential for similar tasks, such as grammatical error correction, and warrants further exploration in future research.

Limitations

6.1 Language Limitation

This study focuses exclusively on Chinese character spelling correction, as CSC presents unique challenges distinct from alphabetic languages like English. Specifically, (1) Chinese text lacks explicit word boundaries, and (2) the language comprises over 100,000 characters, with around 3,500 commonly used ones—many of which share similar pronunciations or visual forms. Nevertheless, we argue that long-tailed distributions are a common challenge across many NLP tasks. Investigating this issue in other language settings or NLP applications—and exploring the use of resources analogous to the Chinese confusion set—offers a promising direction for future research.

6.2 Running Efficiency

We have not specifically optimized the running efficiency of CG-CSC in our current implementation. On a single NVIDIA RTX A6000 GPU (48GB), each training process takes approximately 10 hours to complete. We believe that the training efficiency could be substantially improved by leveraging multi-GPU environments with data-parallel training, which would allow for larger batch sizes and reduced training time. Future work can explore such optimizations to make CG-CSC more scalable and training-friendly for larger datasets.

References

Haithem Afli, Zhengwei Qui, Andy Way, and Páraic Sheridan. 2016. Using smt for ocr error correction of historical texts.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proc. of EMNLP 2016*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Rahhal ERRATTAHI, Asmaa EL HANNANI, and Hassan OUAHMANE. Automatic speech recognition errors detection and correction: A review.

Jianfeng Gao, Chris Quirk, and 1 others. 2010. A large scale ranker-based system for search query spelling correction. In *The 23rd International Conference on Computational Linguistics*.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. Faspell: A fast, adaptable, simple, powerful chinese spell checker based on dae-decoder paradigm. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 160–169.

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, Singapore.

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.

Jiahao Li, Quan Wang, Zhendong Mao, Junbo Guo, Yanyan Yang, and Yongdong Zhang. 2022. 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, Abu Dhabi, United Arab Emirates. 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)*, pages 13509–13521, Toronto, Canada.

Changchun Liu, Kai Zhang, Junzhe Jiang, Zirui Liu, Hanqing Tao, Min Gao, and Enhong Chen. 2024. ARM: An alignment-and-replacement module for Chinese spelling check based on LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10156–10168, Miami, Florida, USA. Association for Computational Linguistics.

Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, and Tao Yang. 2022. Contextual debiasing for visual recognition with causal mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12755–12765.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing nlu models via causal intervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11376–11384.

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.

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, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2527.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Haiming Wu, Hanqing Zhang, Richeng Xuan, and Dawei Song. 2024a. Bi-DCSpell: A bi-directional detector-corrector interactive framework for chinese spelling check. In *Findings of ACL: EMNLP 2024*, pages 3974–3984, Miami, Florida, USA. Association for Computational Linguistics.

Jialong Wu, Linhai Zhang, Deyu Zhou, and Guoqiang Xu. 2024b. DINER: Debiasing aspect-based sentiment analysis with multi-variable causal inference. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3504–3518, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024c. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

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, He-Yan 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.

Zhen Yang, Yongbin Liu, and Chunping Ouyang. Causal intervention-based few-shot named entity recognition. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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. 2021a. 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.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021b. De-biasing distantly supervised named entity recognition via causal intervention. 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 4803–4813.

Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021c. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 11–20.

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.

A Example Appendix

This is an appendix.