**CONTEXTUAL SIMILARITY IS MORE VALUABLE THAN CHARACTER SIMILARITY: AN EMPIRICAL STUDY FOR CHINESE SPELL CHECKING**

**Ding Zhang**\(^1\)*, **Yinghui Li**\(^1\)*, **Qingyu Zhou**\(^2\), **Shirong Ma**\(^1\), **Yangning Li**\(^1,4\), **Yunbo Cao**\(^3\), **Hai-Tao Zheng**\(^1,4\)\(^\dagger\)

\(^1\)Shenzhen International Graduate School, Tsinghua University
\(^2\)OPPO Research Institute, \(^3\)Tencent Cloud Xiaowei, \(^4\)Peng Cheng Laboratory

**ABSTRACT**

Chinese Spell Checking (CSC) task aims to detect and correct Chinese spelling errors. Recently, related researches focus on introducing character similarity from confusion set to enhance the CSC models, ignoring the context of characters that contain richer information. To make better use of contextual information, we propose a novel yet effective Curriculum Learning (CL) framework for the CSC task. With the help of our model-agnostic CL framework, existing CSC models will be trained from easy to difficult as humans learn Chinese characters and achieve further performance improvements. Extensive experiments and detailed analyses on widely used SIGHAN datasets show that our method outperforms previous state-of-the-art methods. More instructively, our study empirically suggests that contextual similarity is more valuable than character similarity for the CSC task.

**Index Terms** — Natural Language Processing, Chinese Spell Checking, Contextual Similarity, Curriculum Learning

**1. INTRODUCTION**

Chinese Spell Checking (CSC) aims to detect and correct spelling errors contained in Chinese text [1, 2]. CSC is receiving more and more attention because it benefits many applications, such as essay scoring [3], OCR [4], and ASR [5]. As a fundamental NLP task, CSC is challenging because the Chinese spelling errors are mainly caused by confusing characters, i.e., phonologically/visually similar characters [6]. As shown in Table 1, “戴 (dài, wear)” and “戴 (rì, day)” are confusing due to their similar strokes.

To enable CSC models to handle confusion characters better, the pre-defined confusion set (i.e., the set of phonologically/visually similar characters) has been long regarded as a good external resource. Many previous works [6, 7] have aimed to leverage the confusion set to introduce phonological/visual similarities into the CSC models. However, these existing methods simply focus on the character similarity provided by the confusion set, but ignore the context of the characters. In fact, in a sentence with a spelling error, the context of the error position provides more useful information that facilitates the CSC process. For example, in Table 1, if the model pays attention to “帽子 (hat)” in the context, it will easily associate the wrong character “带 (dài, bring)” with the correct character “戴 (dài, wear)”. Therefore, we believe that the contextual similarity of the characters is more important for CSC than the character similarity.

In this paper, we aim to enhance CSC models by introducing contextual similarity into Chinese characters. Considering that the CSC task itself is inseparable from human learning, we hope that the model can learn like a human learns to correct spelling errors. We all know that for a student who is just beginning to learn Chinese characters, the teacher always teaches him or her from easy to difficult. Therefore, inspired by the process of humans learning Chinese characters, we also want to guide the model to learn from easy to hard. And this motivation just coincides with curriculum learning.

The core idea of curriculum learning is to train models from easy to hard [8]. And the key to curriculum learning is to design a mechanism to measure the difficulty of samples. Benefiting from this mechanism, we naturally use the contextual similarity of characters as the metric for measuring the sample’s difficulty, so as to organize the scattered training samples into ordered samples for model training. Specifically, we train the model in the order from samples with low contextual similarity to samples with high contextual similarity. Hence, the model achieves better performance than only using the traditional character similarity of confusion set. Moreover, our curriculum learning framework is model-agnostic so that it brings stable improvements for most existing CSC models.

The contributions of our work are summarized as: (1) We empirically verify that contextual similarity is more valuable than character similarity in the CSC task, which is instructive for future works. (2) We propose a simple yet effective curriculum learning framework that enhances the CSC models to explicitly focus on the contextual similarity between Chinese characters. (3) We achieve new state-of-the-art performance on SIGHAN benchmarks and conduct extensive

---

**Table 1.** Examples of Chinese spelling errors. The wrong/correct characters are in red/blue.
analyses to demonstrate the effectiveness of our proposed method.

2. RELATED WORK

2.1. Chinese Spell Checking

In the field of CSC [14], many works focus on constructing and employing confusion set to guide the models to correct the erroneous characters. SpellGCN [9] employs graph convolutional network and pre-defined confusion set [15] to generate candidate characters for the CSC task. PLOME [6] and MLM-phonetics [10] optimize the pre-defined confusion set [15] to generate candidate characters for the CSC task. PLOME [6] and MLM-phonetics [10] optimize the pre-defined confusion set [15] to generate candidate characters for the CSC task.

To the best of our knowledge, existing CSC works improve model performance by introducing character similarity provided by confusion set, but has not made the model focus on contextual similarity of Chinese characters. As a matter of fact, the context of the spelling error position is able to provide vital information for CSC task. In this paper, it is the first time that contextual similarity is applied successfully into the CSC task.

2.2. Curriculum Learning

The idea of Curriculum Learning is to train models from easy to hard, which is proposed in [16]. With the great success in CV [17, 18], Curriculum Learning has attracted many researchers to apply this strategy to all kinds of NLP [19] tasks, which include Machine Translation [20, 21], Question answering [22], Reading Comprehension [23]. For CSC, although Self-Supervised Curriculum Learning has been employed in [24], it is only integrated into a particular model. In our work, the universality of Curriculum Learning is the first time to be demonstrated for CSC. Our designed CL framework is model-agnostic for most existing CSC models. Besides, different from [24], our work focuses more on contextual similarity.

3. PROPOSED APPROACH

3.1. Study Motivation

The core of our work is how to make the CSC models explicitly pay more attention to the context of Chinese characters. Therefore, we propose to use contextual similarity as the metric to measure the difficulty of samples in curriculum learning.

Based on detailed observation, we get the following two obvious facts: (1) A sample is more difficult if it has more wrong characters. (2) A sample is more difficult if the wrong character it contains is more similar to the corresponding correct character. According to these two facts, we design a specific difficulty evaluation strategy and propose the curriculum learning framework for CSC. More specifically, as shown in Figure 1, our curriculum learning framework is divided into two parts: Difficulty Evaluation and Curriculum Arrangement, which will be described in Sections 3.2 and 3.3.

3.2. Difficulty Evaluation

In Difficulty Evaluation, we assign a difficulty score to each training sample in the whole training set $S$. For each sample, we employ an encoder $E(\cdot)$ (e.g., BERT or other CSC models), to transform the characters in the wrong sequence $s_i$ and correct sequence $t_i$ to the corresponding contextual representations $E(s_i)$ and $E(t_i)$.

After obtaining the contextual representations of the wrong/correct sentence, we use the representation corresponding to the wrong position to calculate the cosine similarity, and then the similarities corresponding to all positions with an error are summed up as the difficulty score of the sample:

$$d_i = \sum_{j \in W_i} \frac{E(s_i)_j \cdot E(t_i)_j}{||E(s_i)_j|| \cdot ||E(t_i)_j||},$$

where $d_i$ is the difficulty score of $i$-th sample, $W_i$ are the positions with error.
| Dataset  | Method                   | Detection Level | Correction Level |
|---------|--------------------------|-----------------|------------------|
|         |                          | Acc | Pre | Rec | F1   | Acc | Pre | Rec | F1   |
| SIGHAN13| SpellGCN [9]             | -   | 80.1 | 74.4 | 77.2 | 78.3 | 72.7 | 75.4 |
|         | MLM-phonetics [10]       | -   | 82.0 | 78.3 | 80.1 | 77.8 | 70.0 | 78.2 |
|         | REALISE [11]             | 82.7| 88.6 | 82.5 | 85.4 | 81.4| 87.2 | 81.2 | 84.1 |
|         | Soft-Masked BERT [7]     |   - | 81.1 | 75.7 | 78.3 | 75.1 | 70.1 | 72.5 |
|         | CL (Soft-Masked BERT)    | -   | 84.7 | 77.0 | 80.7 | 80.9 | 74.5 | 77.6 |
|         | BERT [12]                | 70.6| 98.7 | 70.6 | 82.3 | 67.8 | 98.6 | 67.8 | 80.4 |
|         | CL (BERT)                | 75.4 | 99.1 | 74.8 | 85.3 | 74.9 | 99.1 | 73.2 | 84.2 |
|         | MacBERT [13]             | 70.8| 98.7 | 70.8 | 82.5 | 68.0 | 98.6 | 67.9 | 80.4 |
|         | CL (MacBERT)             | 76.3 | 99.3 | 75.7 | 85.9 | 75.8 | 99.2 | 73.8 | 84.6 |
| SIGHAN14| SpellGCN [9]             | -   | 65.1 | 69.5 | 67.2 | 63.1 | 67.2 | 65.3 |
|         | MLM-phonetics [10]       | -   | 66.2 | 73.8 | 69.8 | 64.2 | 73.8 | 68.7 |
|         | REALISE [11]             | 78.4| 67.8 | 71.5 | 69.6 | 77.7 | 66.3 | 70.0 | 68.1 |
|         | Soft-Masked BERT [7]     | -   | 65.2 | 70.4 | 67.7 | 63.7 | 68.7 | 66.1 |
|         | CL (Soft-Masked BERT)    | -   | 68.4 | 70.9 | 69.6 | 67.8 | 69.1 | 68.4 |
|         | BERT [12]                | 72.7| 78.6 | 60.7 | 68.5 | 71.2 | 77.8 | 57.6 | 66.2 |
|         | CL (BERT)                | 75.8 | 79.2 | 61.6 | 69.3 | 74.8 | 78.5 | 60.8 | 68.5 |
|         | MacBERT [13]             | 72.9| 78.8 | 61.0 | 68.8 | 71.5 | 78.0 | 58.0 | 66.5 |
|         | CL (MacBERT)             | 76.1 | 79.7 | 62.4 | 70.0 | 75.0 | 79.0 | 61.4 | 69.1 |
| SIGHAN15| SpellGCN [9]             | -   | 74.8 | 80.7 | 77.7 | 72.1 | 77.7 | 75.9 |
|         | PLOME [6]                | -   | 77.4 | 81.5 | 79.4 | 75.3 | 79.3 | 77.2 |
|         | MLM-phonetics [10]       | -   | 77.5 | 83.1 | 80.2 | 74.9 | 80.2 | 77.5 |
|         | REALISE [11]             | 84.7| 77.3 | 81.3 | 79.3 | 84.0 | 75.9 | 79.9 | 77.8 |
|         | Soft-Masked BERT [7]     | -   | 73.7 | 73.2 | 73.5 | 66.7 | 66.2 | 66.4 |
|         | CL (Soft-Masked BERT)    | -   | 83.5 | 74.8 | 78.9 | 79.0 | 72.1 | 75.8 |
|         | BERT [12]                | 79.9| 84.1 | 72.9 | 78.1 | 77.5 | 83.1 | 68.0 | 74.8 |
|         | CL (BERT)                | 80.5 | 85.0 | 74.5 | 79.4 | 79.0 | 84.2 | 72.3 | 77.8 |
|         | MacBERT [13]             | 80.0| 84.3 | 73.1 | 78.3 | 77.7 | 83.3 | 68.2 | 75.0 |
|         | CL (MacBERT)             | 80.9 | 85.8 | 75.4 | 80.3 | 79.3 | 84.7 | 73.0 | 78.4 |

Table 2. The performance of our CL method and all baselines. “↑” indicates that our CL method is able to enhance the corresponding baseline. We underline the previous state-of-the-art performance for convenience.

### 3.3. Curriculum Arrangement

In this section, we describe an **Annealing** method to arrange all the training samples $S$ into an ordered curriculum based on the difficulty scores that are introduced in Section 3.2.

Firstly, we sort all training samples in ascending order of their difficulty scores and split them into $k$ subsets $\{S_1, S_2, ..., S_k\}$. Note that these subsets are non-overlapping for preventing over-fitting and improving the generalization performance. Then we arrange a learning curriculum which contains $k + 1$ training stages. At the $i$-th stage ($i \leq k$), we further split each of the $k$ subsets $\{S_1, S_2, ..., S_k\}$ into $k$ parts by order of difficulty. For each subset $S_j$, we obtain $\{S_{j,1}, S_{j,2}, ..., S_{j,k}\}$ and use the $i$-th part $S_{j,i}$ for this $i$-th stage, thus the final training set $C_i = \{S_{1,i}, S_{2,i}, ..., S_{k,i}\}$ is employed for the $i$-th stage. It is worth mentioning that the training set $C_i$ will be shuffled for maintaining local stochastics within $i$-th stage.

For the former $k$ stages, the model is trained on the $C_i$ for one epoch one after another to lead the model learning from easy to difficult. At the last stage (i.e., the $k + 1$-th stage), the model is trained on the whole training set $S$ for fitting the original data distribution.

### 4. EXPERIMENTS

#### 4.1. Datasets

Following previous works [7, 11], we use the same training data which contains SIGHAN 13/14/15 [15, 25, 26] and a generated pseudo dataset [5]. Additionally, to ensure fairness, models’ performance is evaluated on the same test data as our baselines, from the test datasets of SIGHAN 13/14/15.

#### 4.2. Baseline Methods

We select strong confusion set-based models as baselines: SpellGCN [9] applies GCNs to learn the character similarity from confusion set. PLOME [6] designs pre-training strategy based on the confusion set. MLM-phonetics [10] introduces phonetic similarity into masked language models from confusion set. REALISE [11] extracts and mixes semantic, phonetic, and graphic information. In addition, we select three popular CSC models to be combined with our proposed CL method: BERT [12] directly fine-tunes the BERT base on the CSC training data. Soft-Masked BERT [7] utilizes the con-
fusion set to generate sufficient training data. MacBERT [13] improves the masking strategy of BERT and adds a full connection layer to detect errors. These BERT-based CSC models are all convenient to obtain the contextual representations of Chinese characters.

### 4.3. Experimental Setup

During the training phase, we follow the hyper-parameters of Soft-Masked-Bert [7]. For Soft-Masked-BERT, we maintain a learning rate $2e^{-5}$ and fine-tune the parameters with Adam. As for MacBERT [13], the learning-rate is set to $5e^{-5}$, the batch size is set to 32. The $k$ is empirically set to in our method. During the testing phase, we evaluate the models in both detection and correction utilizing sentence-level accuracy, precision, recall, and F1 score.

### 4.4. Experimental Results

Table 2 shows the results of our CL method compared to baselines. We can see that, by reordering the training data, our method yields consistent gain with a large margin against all baselines.

| Method                             | Correction F1 | $\Delta$ |
|------------------------------------|--------------|----------|
| MacBERT                            | 75.0%        | -        |
| Only Difficulty Evaluation         | 77.2%        | +2.2%    |
| Only Curriculum Arrangement        | 76.3%        | +1.3%    |
| Using Character Similarity         | 76.5%        | +1.5%    |
| CL (MacBERT)                       | 78.4%        | +3.4%    |

Table 3. Results of ablation studies. “$\Delta$” indicates the absolute F1 improvements on correction level.

### 4.5. Ablation Study

To explore the contribution of each component in our curriculum learning framework, we conduct ablation studies with the following settings: 1) only using difficulty evaluation to sort training samples for the training process of MacBERT, 2) only using curriculum arrangement to randomly arrange training stages for the training process of MacBERT, and the training samples of each stage are randomly selected. 3) Besides, to verify the advantage of contextual similarity, we also use the confusion set-based character similarity as the difficulty metric in our CL framework.

From Table 3, we observe that both difficulty evaluation and curriculum arrangement bring improvements for MacBERT, which indicates the rationality of these two modules we design. Particularly, the greater improvements that only using difficulty evaluation brings to MacBERT than only using curriculum arrangement and using character similarity reflects the correctness of our motivation that contextual similarity is more valuable than character similarity in the CSC task.

### 4.6. Parameter Study

The key parameter in our framework is the number of subsets $k$, so it is essential to study its effects. Figure 2 illustrates the performance change of CL (MacBERT), we find that when the value of $k$ reaches a certain value, the performance of the model does not improve anymore. In fact, this phenomenon is consistent with the process of human learning. Imagine that when the courses are divided too trivially (that is, when the value of $k$ is too large), it is difficult for humans to learn effective knowledge from too many courses. Therefore, it is critical to choose the best $k$, although there are stable improvements based on MacBERT at all values of $k$.

### 4.7. Case Study

In the first example of Table 4, “母” is the wrong character. “美” is the corrected character. The two characters are only acoustically similar. Therefore, the low difficulty score evaluated by our CL (MacBERT) of this example is reasonable. In the second example, the erroneous character “造” is easily to be detected and corrected to “坐” by using the information of “走路” as well as “公交” in the context. The difficulty score of the latter example is also significantly higher than the first example. This verifies the effectiveness of our designed difficulty scoring function and suggests that the model has learned how to distinguish between highly and slightly similar spelling errors.

### 5. CONCLUSION

In this paper, we aim to exploit the contextual similarity of characters to obtain better CSC performance than character similarity contained in traditional confusion set. Additionally, we propose a simple yet effective curriculum learning framework for the CSC task. With the help of such a model-agnostic framework, most existing CSC models significantly perform better. In the future, it would be a very interesting direction to apply the core idea of our work to more scenarios such as Chinese grammatical error correction.
6. REFERENCES

[1] Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Haitao Zheng, “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, Smanara Muresan et al, Ed. 2022, pp. 3202–3213, Association for Computational Linguistics.

[2] Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen, “Linguistic rules-based corpus generation for native Chinese grammatical error correction,” in Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 576–589, Association for Computational Linguistics.

[3] Fei Dong and Yue Zhang, “Automatic features for essay scoring - an empirical study,” in EMNLP. 2016, pp. 1072–1077, The Association for Computational Linguistics.

[4] Haithem Afli, Zhengwei Qiu, Andy Way, and P´eraic Sheridan, “Using SMT for OCR error correction of historical texts,” in LREC. 2016, European Language Resources Association (ELRA).

[5] Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang, “A hybrid approach to automatic corpus generation for chinese spelling check,” in EMNLP, 2018, pp. 2517–2527, Association for Computational Linguistics.

[6] Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang, “PLOME: pre-training with misspelled knowledge for chinese spelling correction,” in ACL/IJCNLP (1). 2021, pp. 2991–3000, Association for Computational Linguistics.

[7] Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li, “Spelling error correction with soft-masked BERT,” in ACL. 2020, pp. 882–890, Association for Computational Linguistics.

[8] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe, “Curriculum learning: A survey,” arXiv preprint arXiv:2101.10382, 2021.

[9] Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi, “Spellgen: Incorporating phonological and visual similarities into language models for chinese spelling check,” in ACL. 2020, pp. 871–881, Association for Computational Linguistics.

[10] Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang, “Correcting chinese spelling errors with phonetic pre-training,” in ACL/IJCNLP (Findings). 2021, vol. ACL/IJCNLP 2021 of Findings of ACL, pp. 2250–2261, Association for Computational Linguistics.

[11] Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao, “Read, listen, and see: Leveraging multimodal information helps chinese spell checking,” in ACL/IJCNLP (Findings). 2021, vol. ACL/IJCNLP 2021 of Findings of ACL, pp. 716–728, Association for Computational Linguistics.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in NAACL-HLT (1). 2019, pp. 4171–4186, Association for Computational Linguistics.

[13] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu, “Revisiting pre-trained models for chinese natural language processing,” in EMNLP (Findings). 2020, vol. EMNLP 2020 of Findings of ACL, pp. 657–668, Association for Computational Linguistics.

[14] Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Li Yangning, Shulin Huang, Ruiyang Liu, Chao Li, Yunbo Cao, and Haitao Zheng, “Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for Chinese spell checking,” in Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 238–249, Association for Computational Linguistics.

[15] Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee, “Chinese spelling check evaluation at SIGHAN bake-off 2013,” in SIGHAN@IJCNLP, 2013, pp. 35–42, Asian Federation of Natural Language Processing.

[16] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, “Curriculum learning,” in ICML. 2009, vol. 382 of ACM International Conference Proceeding Series, pp. 41–48, ACM.

[17] Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, and Yuexian Zou, “SSVMR: saliency-based self-training for video-music retrieval,” CoRR, vol. abs/2302.09328, 2023.

[18] Hongxiang Li, Meng Cao, Xuxin Cheng, Zhihong Zhu, Yaowei Li, and Yuexian Zou, “Generating templated caption for video grounding,” CoRR, vol. abs/2301.05997, 2023.

[19] Zhihong Zhu, Weiyuan Xu, Xuxin Cheng, Tengtao Song, and Yuexian Zou, “A dynamic graph interactive framework with label-semantic injection for spoken language understanding,” CoRR, vol. abs/2211.04023, 2022.

[20] Tom Kocmi and Ondrej Bojar, “Curriculum learning and mini-batch bucketing in neural machine translation,” arXiv preprint arXiv:1707.09533, 2017.

[21] Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou, “M3ST: mix at three levels for speech translation,” CoRR, vol. abs/2212.03657, 2022.

[22] Cao Liu, Shizhu He, Kang Liu, and Jun Zhao, “Curriculum learning for natural answer generation,” in IJCAI. 2018, pp. 4223–4229, ijcai.org.

[23] Yichan Liang, Jianheng Li, and Jian Yin, “A new multi-choice reading comprehension dataset for curriculum learning,” in ACML. 2019, vol. 101 of Proceedings of Machine Learning Research, pp. 742–757, PMLR.

[24] Zifa Gan, Hongfei Xu, and Hongying Zan, “Self-supervised curriculum learning for spelling error correction,” in EMNLP (1). 2021, pp. 3487–3494, Association for Computational Linguistics.

[25] Junjie Yu and Zhenghua Li, “Chinese spelling error detection and correction based on language model, pronunciation, and shape,” CLP 2014, p. 220, 2014.

[26] Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen, “Introduction to SIGHAN 2015 bake-off for chinese spelling check,” in SIGHAN@IJCNLP. 2015, pp. 32–37, Association for Computational Linguistics.