Local Attention Augmentation for Chinese Spelling Correction

Shuo Wang[1,2], Chaodong Tong[1,2], Kun Peng[1,2], and Lei Jiang[1,2(✉)]

[1] Institute of Information Engineering, Chinese Academy of Sciences, BeiJing, China
[2] School of Cyberspace Security, University of Chinese Academy of Sciences, BeiJing, China
{wangshuo4035,pengkun,tongchaodong,jianglei(✉)}@iie.ac.cn

Abstract. Chinese spelling correction (CSC) is an important task in the field of natural language processing (NLP). While existing state-of-the-art methods primarily leverage pre-trained language models and incorporate external knowledge sources such as confusion sets, they often fall short in fully leveraging local information that surrounds erroneous words. In our research, we aim to bridge a crucial gap by introducing a novel CSC model that is enhanced with a Gaussian attention mechanism. This integration allows the model to adeptly grasp and utilize both contextual and local information. The model incorporates a Gaussian attention mechanism, which results in attention weights around erroneous words following a Gaussian distribution. This enables the model to place more emphasis on the information from neighboring words. Additionally, the attention weights are dynamically adjusted using learnable hyperparameters, allowing the model to adaptively allocate attention to different parts of the input sequence. In the end, we adopt a homophonic substitution masking strategy and fine-tune the BERT model on a large-scale CSC corpus. Experimental results show that our proposed method achieve a new state-of-the-art performance on the SIGHAN benchmarks.

Keywords: Gaussian Attention · Spelling Check · Chinese Spelling Correction.

1 Introduction

Spelling correction is an important task that detects and corrects language errors made by humans. It finds wide applications in various fields, including text editors [1], machine translation [2, 3], search engines [4–6], etc. Spelling errors in Chinese can lead to ambiguities or misunderstandings, ultimately affecting the readability and reliability of the content. Addressing spelling mistakes in Chinese presents unique challenges due to the language's complex character combinations

(✉)Corresponding Author.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_44

and syllable structures. These complexities further compounds the challenge of detecting and correcting errors. Moreover, CSC requires a deep understanding of contextual information for accurate corrections, owing to the nuanced and context-dependent nature of Chinese. CSC involves two types of errors: phonological similarity errors and visual similarity errors. Phonological similarity errors are more prevalent, accounting for 83% of Chinese spelling error [7]. An example in Table 1 demonstrate phonological similarity errors in CSC. In the erroneous sentences, the word "绵" (soft) is misused, while the corrected sentences use the correct word "面" (surface), which has the same pronunciation but a different meaning. From the example, it can be observed that the error character "绵" (soft) has a stronger correlation with the surrounding characters "海" (sea) and "漂浮" (float) than the distant character "报告" (report). The example in Table 1 shows that characters closer to the mistaken character have a stronger association than those further. Therefore, the local information plays a crucial role in the CSC.

Table 1. An example of Chinese Spelling Correction.

***Wrong*:**	bao gao cheng quan qiu hai mian piao fu chao wu wan yi jian su liao la ji 报 告 称 全 球 海 绵 漂 浮 超 5 万 亿 件 塑 料 垃 圾 The report states that there are over 500 million pieces of plastic waste floating in the world's sponge
***Correct*:**	bao gao cheng quan qiu hai mian piao fu chao wu wan yi jian su liao la ji 报 告 称 全 球 海 面 漂 浮 超 5 万 亿 件 塑 料 垃 圾 The report states that there are over 500 million pieces of plastic waste floating in the world's oceans

Previous research has proposed various methods, primarily aimed at leveraging the contextual information within texts, to address challenges in CSC. Previous works proposed language models for CSC, using n-gram techniques to define corresponding rules for error detection and correction[7–10]. In recent years, BERT has been widely applied in the field of CSC. Soft-Masked BERT [11] proposed a novel masking strategy for CSC. Xu et al. (2020) [12] introduced a Chinese spell checker called REALISE, which directly leverages the multimodal information of Chinese characters. Additionally, FASPell [13], SpellGCN [14], DCN [15], PLOME [16], and others have also proposed BERT-based methods for the CSC task.

While recent research in CSC has mainly focused on utilizing contextual information from language models or neural network models, the local information of individual characters has often been overlooked. However, local information can play a crucial role in spelling correction tasks, especially in cases where errors occur at the character level. In this paper, we propose a novel CSC method architecture based on local attention enhancement. The CSC model utilizes the Gaussian attention mechanism, where the attention weights follow a Gaussian distribution, to integrate contextual and local information. The contributions of this paper are as follows: (1) We propose a CSC model that enhances local attention. We combine Gaussian attention with the neural network model (BiGRU) to effectively capture the contextual relationships between Chinese characters and focus on the specific regions where errors are likely to occur, thereby improving the accuracy of error correction. (2) In the Gaussian attention layer, we use learnable hyperparameters to adjust the attention weights of the characters in the sentence. Additionally, we adopt a homophonic substitution masking strategy and fine-tune the BERT model on a large-scale CSC corpus. (3) Our method achieves the best results and significant improvements compared to other methods in both character-level and sentence-level experiments. In terms of sentence-level precision, we achieve a substantial increase from 67.4% to 87.9% compared to the baseline in the SIGHAN2014 dataset.

2 Related Work

There is a significant need for efficient and accurate CSC methods to meet the diverse requirements of different applications. These methods should be capable of identifying and correcting a wide range of spelling errors, including typographical mistakes and word order errors. Additionally, they need to interpret contextual information to provide relevant correction suggestions. In recent years, there has been a surge in research methods focused on CSC. These innovations and advancements are continually evolving, reflecting the increasing complexity and demands of processing Chinese text in various technological applications.

The evolution of CSC methods marks substantial progress in NLP. Initially, the focus was on rule-based methods, exemplified by the approach [17]. Their technique involved segmenting sentences, identifying errors, and correcting them through dictionary lookups, which laid the foundation for future work in this area.

The field then shifted towards statistical and machine learning approaches. Researchers pioneered the use of n-gram language models for CSC [8, 10, 17]. This marked a significant step forward, leveraging statistical models to better grasp the intricacies of the Chinese language. More recently, the advent of neural networks has revolutionized CSC. Notable neural network architectures such as GRU, LSTM, and Transformer have been extensively applied. Soft-Masked BERT [11] and PLOME [16] utilized GRU models for error detection, demonstrating the efficacy of these models in identifying spelling errors [11, 16]. Similarly, other works successfully employed LSTM models, showcasing their potential in correcting Chinese spelling errors [15, 18].

The current state-of-the-art in Chinese Spelling Correction (CSC), primarily employing large-scale language models like BERT, marks a significant improvement over previous methods. These models tokenize input text and map each character to an embedding vector, leveraging contextual information to learn rich, nuanced representations. However, a notable limitation, particularly in the context of CSC, is their inadequate consideration of the relative importance of surrounding characters in relation to an erroneous character. Presently, pre-training of deep bidirectional transformers for language understanding (BERT)-

DOI: 10.1007/978-3-031-63759-9_44

[19] models have emerged as the most popular choice in this domain. They have been effectively used in various systems like FASPell [13], Confusionset [20], SpellGCN [14], MLM-phonetics [21], PLOME [16], and uChecker [22]. These implementations demonstrate exceptional performance in CSC, capitalizing on the strengths of BERT models to understand and correct complex language structures and contextual nuances inherent in the Chinese language. The continuous evolution in this field underscores the dynamic nature of language processing technologies and their growing sophistication in handling the unique challenges presented by the Chinese language.

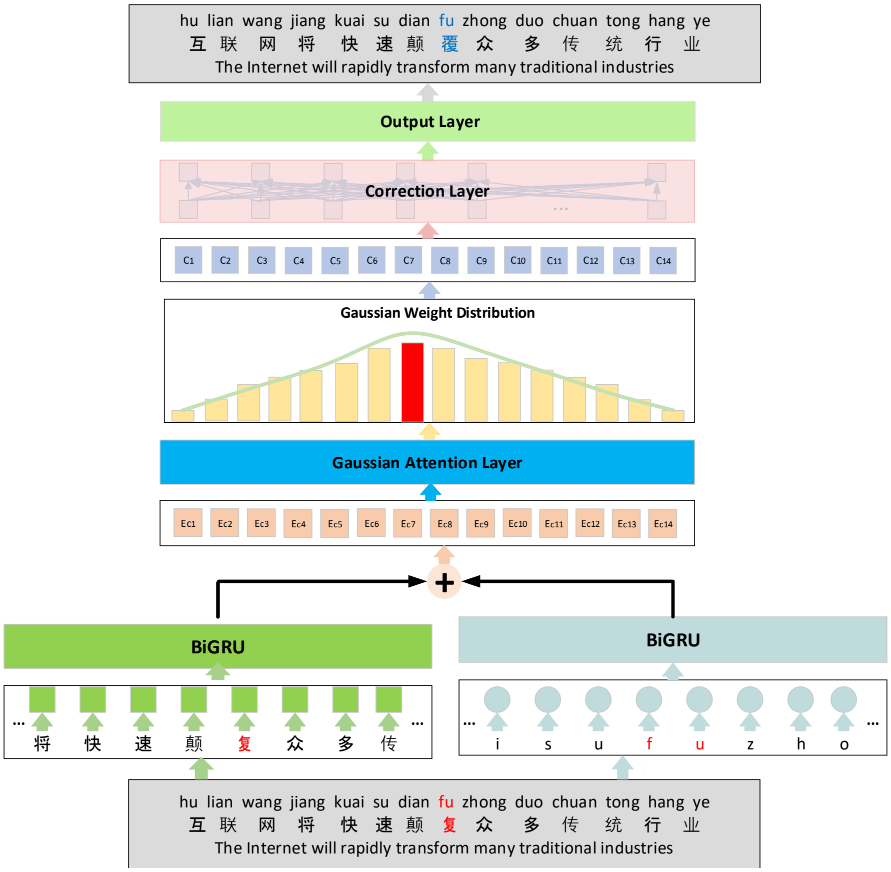
hu lian wang jiang kuai su dian fu zhong duo chuan tong hang ye
互 联 网 将 快 速 颠 覆 众 多 传 统 行 业
The Internet will rapidly transform many traditional industries
Output Layer
Correction Layer
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14
Gaussian Weight Distribution
Gaussian Attention Layer
Ec1 Ec2 Ec3 Ec4 Ec5 Ec6 Ec7 Ec8 Ec9 Ec10 Ec11 Ec12 Ec13 Ec14
+
BiGRU
BiGRU
… 将 快 速 颠 复 众 多 传 …
… i s u f u z h o …
hu lian wang jiang kuai su dian fu zhong duo chuan tong hang ye
互 联 网 将 快 速 颠 复 众 多 传 统 行 业
The Internet will rapidly transform many traditional industries

Fig. 1. Architecture of the CSC model. In the figure, E_{ci} represents the vector representation combining character and phonetic information, while C_i represents the vector representation of aggregated Gaussian weights.

3 Approach

3.1 Problem and Motivation

In the Chinese spelling correction task, we input a Chinese sentence comprising n characters, denoted as $X = (x_1, x_2, \ldots, x_n)$. The core of this task involves using a model to detect and identify spelling errors within the sentence. Once errors are identified, the model undertakes necessary corrections to generate a correct sentence without spelling errors, represented as $Y = (y_1, y_2, \ldots, y_n)$. It is important to note that in this task, the number of characters remains the same in both the input and output sentences. This task extends beyond mere replacement of a small number of characters; it critically involves understanding the relationships between characters at different positions. The essence of the challenge lies in recognizing and interpreting the interconnectedness among characters, which plays a pivotal role in accurately deciphering and correcting text.

In the task of CSC, relying solely on character-level analysis is insufficient to capture the deep semantics and contextual relationships between words. Although traditional recurrent neural networks like BiGRU can understand context, they fall short in accurately assessing the importance of each position in the sequence. To address this, we introduce the Gaussian attention mechanism, which enhances focus on crucial information through dynamic weight allocation, combined with the contextual understanding capabilities of BiGRU. Moreover, by conducting experiments on a vast array of CSC datasets, we have validated our hypothesis. This research not only confirms the effectiveness of integrating Gaussian attention with BiGRU for Chinese spelling correction but also sets the stage for future explorations in this domain.

3.2 Model

We propose a novel neural network model: the Gaussian Attention-based CSC model. As depicted in Fig. 1, our model comprises two components: the detection network and the correction network. In the detection network, embeddings of character and phonetic information are inputted into a BiGRU layer. Subsequently, the fused word representation, which amalgemates character and phonetic data, is fed into a Gaussian attention layer. Moving to the correction network, the word representation, synergized with Gaussian attention weights, is channeled into a correction layer. Ultimately, the refined output is produced post-correction. For this task, we employ the BERT model, modifying its masking strategy to better suit CSC.

The detection network is engineered to identify potential spelling errors in the input text. We use a BiGRU model to capture the contextual information of the characters. Additionally, we incorporate a Gaussian attention mechanism, focusing on the characters surrounding the erroneous one. This enhancement enables more effective detection of Chinese spelling errors.

DOI: 10.1007/978-3-031-63759-9_44

In the correction network, the BERT model generates suggestions for correcting detected errors. We adapt the BERT model's masking strategy to prioritize error characters during the correction process, enabling the model to provide more accurate and relevant suggestions for Chinese spelling errors.

3.3 Masking Strategy

The BERT model, known for its impressive performance across various tasks, is extensively applied in diverse domains. Its masking strategy, employed during training, involves randomly replacing a portion of the input sequence with the [MASK] token. The model then learns to predict these masked tokens. Typically, about 15% of the tokens in an input sequence are randomly substituted with the [MASK] token.

In our approach, we adjust the masking strategy of BERT to better suit CSC. Instead of randomly replacing 15% of the tokens with the [MASK] token, we replace 80% of the [MASK] tokens with characters that have the same pronunciation, and the remaining 20% with random characters. This modified masking strategy is more aligned with the requirements of the CSC task.

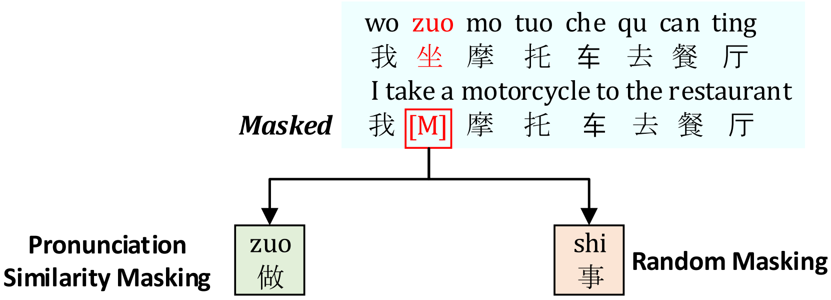
wo zuo mo tuo che qu can ting
我 坐 摩 托 车 去 餐 厅
I take a motorcycle to the restaurant
Masked
我 [M] 摩 托 车 去 餐 厅
Pronunciation Similarity Masking
zuo
做
shi
事
Random Masking

Fig. 2. An example of different masking strategies.

As shown in Fig. 2, for example, the character "坐 (sit)" would be masked using the original masking strategy with a random character "事 (matter)". In our approach, we replace it with a character that has the same pronunciation, such as "做 (do)". We maintain a separate file that contains characters with the same pronunciation for easy lookup and substitution by the model.

By adjusting the masking strategy to consider characters with the same pronunciation, our approach aims to improve the relevance and accuracy of the correction suggestions provided by the BERT model for CSC tasks.

3.4 Detection Network

In the detection module of Chinese Spelling Correction (CSC), our integration of the Gaussian Attention Layer with Bidirectional Gated Recurrent Units (BiGRU) not only significantly enhances the model's performance but also ensures

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_44

that the model fully utilizes both contextual and local information within the text. This innovative blend capitalizes on the strengths of BiGRU and Gaussian attention mechanisms to achieve this dual focus.

BIGRU Layer BiGRU is a bidirectional version of the Gated Recurrent Unit (GRU), consisting of two GRU layers. One processing the forward information in the time sequence, and the other processing the backward information. This allows the network to simultaneously consider past and future context information. In the context of spelling correction tasks, BiGRU effectively captures the contextual relationships between Chinese characters, helping the model understand and detect potential spelling errors.

The vector $\overrightarrow{h_t}$ represents the hidden state of the GRU in the forward direction, and $\overleftarrow{h_t}$ represents the hidden state of the GRU in the backward direction, as shown in the equations:

$$\overrightarrow{h_t} = GRU(\overrightarrow{h_{t-1}}, x_t) \tag{1}$$

$$\overleftarrow{h_t} = GRU(\overleftarrow{h_{t+1}}, x_t) \tag{2}$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{3}$$

Gaussian Attention Layer The Gaussian attention layer is a specially designed attention mechanism, aimed at enhancing the model's focus on the information surrounding specific characters in the text. It uses a Gaussian distribution to simulate the relative importance between characters, enabling the model to pay more attention to characters adjacent to erroneous ones. Applying the Gaussian attention layer to spelling detection can effectively capture long-range dependencies, allowing the model to focus more on the information surrounding the error character and improve the detection performance.

Let's assume that a_t represents the attention weight of character t. Then, we can define it as:

$$a_t = \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \tag{4}$$

Where μ and σ are the parameters of the Gaussian distribution, and they are learned by the network. Where μ and σ are learnable hyperparameters used to dynamically adjust the mean and standard deviation of the attention weights. By using the Gaussian distribution form, the attention weight a_t is computed as the exponential function of the distance between character position t and the hyperparameters μ and σ. The normalization operation ensures that the sum of attention weights for all positions is equal to 1, making it a probability distribution.

The Gaussian attention layer finds a particularly compelling application in the realm of CSC. It empowers the model to focus keenly on the contextual information surrounding potential errors, leading to significant improvements in detection performance. This enhanced focus stems from the ability of the Gaussian attention layer to grasp intricate semantic relationships and spelling patterns within the surrounding characters.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_44

The weighted hidden state with attention is represented as:

$$c_t = \sum_i a_i \cdot h_i \tag{5}$$

By combining the BiGRU's ability to capture contextual information with the focusing ability of the Gaussian attention mechanism, we aim to enhance the detection module of CSC, enabling the model to better understand and correct spelling errors.

3.5 Correction Network

In our correction module, we employ the BERT model, which is fundamentally based on the Transformer encoder architecture. BERT utilizes a self-attention mechanism to comprehend the contextual relationships within the text. Its bidirectional nature is especially crucial for understanding the context surrounding a given character in Chinese text, as the meaning of characters often relies on their neighboring characters. BERT is employed to identify and correct erroneous characters, and through fine-tuning, it learns to predict the most appropriate characters within a given context.

The self-attention mechanism used in the BERT model is represented by the following formula:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Here, Q, K, V respectively represent the Query, Key, and Value, and d_k denotes the dimensionality of the key.

For each input character x_i, BERT generates a contextually informed representation h_i:

$$h_i = BERT(x_i) \tag{7}$$

This representation h_i contains the contextual information of the character x_i.

For characters identified as erroneous, a prediction is made using the Masked Language Model (MLM) task:

$$\hat{x}_i = argmax_x P(x|h_i) \tag{8}$$

In this equation, $\hat{x}_i$ is the character predicted by the model to be the correct replacement.

4 Experiments

In this section, we will introduce the dataset, experimental setup, and the performance of our model on the dataset. We evaluated the performance of our model using the SIGHAN dataset, which is a benchmark dataset for CSC task.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_44

4.1 Datasets

For the CSC task, we utilized the SIGHAN dataset, a well-established benchmark for evaluating different methods. While relatively small in size, it offers valuable insights due to its diverse composition. SIGHAN comprises three key subsets: SIGHAN2013 [23], containing around 700 sentences with over 300 errors, and SIGHAN2014 [24] and SIGHAN2015 [25], each consisting of a few thousand sentences.The SIGHAN2013 subset is composed of a variety of text genres, including news articles, academic papers, and social media posts. This diversity helps to ensure that the dataset is representative of real-world spelling errors. The SIGHAN2014 and SIGHAN2015 subsets are both composed of news articles. This focus on a single genre allows for a more detailed analysis of the types of spelling errors that occur in this particular context.

For the training data, in our previous work employed a comprehensive dataset that strategically blends automatically annotated data with rigorously curated benchmarks. To ensure a dataset both ample and diverse, we leveraged an expansive collection of Chinese text automatically annotated with spelling errors. This dataset artfully simulates real-world errors by incorporating both visually similar and phonologically similar character substitutions. Visually similar characters mirror OCR-like errors, arising from visual resemblances between characters. Phonologically similar characters echo ASR-like errors, stemming from similarities in pronunciation [26]. To complement the breadth of the automated dataset, we judiciously incorporated training data from three smaller-sized, yet meticulously hand-annotated datasets: SIGHAN2013, SIGHAN2014, and SIGHAN2015. These datasets, widely recognized for their meticulous quality, offer a valuable source of precisely labeled spelling errors, fortifying the model's ability to discern and correct fine-grained distinctions.

Table 2. Experimental Data Statistics Information.

	Corpus Name	**#Sentences**	**Avg.Length**	**#Errors**
Training Set	Wang et al.(2018)	271,329	44.4	382,704
	SIGHAN2013	700	41.8	343
	SIGHAN2014	3,437	49.6	5,122
	SIGHAN2015	2,338	31.3	3,037
	Total	277,804	41.8	391,206
Test Set	SIGHAN2013	1,000	74.3	1,224
	SIGHAN2014	1,062	50.0	771
	SIGHAN2015	1,100	30.6	703
	Total	3,162	50.9	2,698

For the testing data, we assessed our model's performance using the SIGHAN-2013, SIGHAN2014, and SIGHAN2015 datasets. As these datasets originally contain traditional Chinese characters, we modified them to utilize simplified Chinese characters for validation and evaluation purposes. We also made necessary adjustments to the format of the datasets to ensure compatibility. The modified datasets consist of 1000 sentences for SIGHAN2013, 1062 sentences for SIGHAN2014, and 1100 sentences for SIGHAN2015.The specific data details are shown in Table 2.

4.2 Setup

We randomly divided the training data into a training set and a validation set, with the training set accounting for 90% of the data and the validation set accounting for 10%.

In real-world scenarios, 83% of Chinese spelling errors are caused by phonetic mistakes. Therefore, we randomly replaced 15% of the characters in the text with artificially generated errors. Among these errors, 80% were replaced with characters that have similar pronunciation, while 20% were replaced with random characters.

For the model's hyperparameter settings, we utilized the default parameters of the BERT model. Additionally, we fine-tuned the model using an optimizer. We set the learning rate of the model to 1e-4, the batch size to 16, and the loss weight to 0.5.

4.3 Baselines

We compared our method with existing methods in the field:

LMC [10] seamlessly blends bi-gram and trigram language models with Chinese word segmentation, leveraging dynamic programming and smoothing techniques to combat data sparsity and achieve robust accuracy. FASPell [13] based on a new paradigm which consists of a denoising autoencoder (DAE) and a decoder. Confusionset [20] utilizes the off-the-shelf confusionset for guiding the character generation. The Seq2Seq model jointly learns to copy a correct character from an input sentence through a pointer network, or generate a character from the confusionset rather than the entire vocabulary. SpellGCN [14]proposes to incorporate phonological and visual similarity knowledge into language models for CSC via a specialized graph convolutional network (SpellGCN). The model builds a graph over the characters, and SpellGCN is learned to map this graph into a set of inter-dependent character classifiers. MLM-phonetics [21] is a groundbreaking end-to-end CSC model that seamlessly integrates phonetic features within a unified framework for joint error detection and correction. PLOME [16] is a pre-trained masked language model with misspelled knowledge for CSC, which jointly learns how to understand language and correct spelling errors. To this end, PLOME masks the chosen tokens with similar characters according to a confusion set rather than the fixed token "[MASK]" as in BERT. DCN [15] generates the candidate Chinese characters via a Pinyin Enhanced Candidate Generator and then utilizes an attention-based network to model the dependencies between two adjacent Chinese characters. uChecker [22] is a Confusionset-guided masking strategy to fine-train the masked language model to further improve the performance of unsupervised detection and correction.

Table 3. The char-level performance of different models on SIGHAN test sets.

Dataset	Model	Detection Level			Correction Level		
		Prec.	Rec.	F1.	Prec.	Rec.	F1.
SIGHAN2013	LMC(Xie et al.,2015)	79.8	50.0	61.5	77.6	22.7	35.1
	FASPell(Hong et al.,2019)	76.2	63.2	69.1	73.1	60.5	66.2
	Confusionset(Wang et al.,2019)	66.8	73.1	69.8	71.5	59.5	69.9
	SpellGCN(Cheng et al.,2020)	80.1	74.4	77.2	78.3	72.7	75.4
	MLM-phonetics(Zhang et al.,2021)	**82.0**	78.3	80.1	79.5	77.0	78.2
	our method	78.7	**91.6**	**84.7**	**98.7**	**95.9**	**97.3**
SIGHAN2014	LMC(Xie et al.,2015)	56.4	34.8	43.0	71.1	50.2	58.8
	FASPell(Hong et al.,2019)	61.0	53.5	57.0	59.4	52.0	55.4
	Confusionset(Wang et al.,2019)	63.2	**82.5**	71.6	79.3	68.9	73.7
	SpellGCN(Cheng et al.,2020)	65.1	69.5	67.2	63.1	67.2	65.3
	MLM-phonetics(Zhang et al.,2021)	66.2	73.8	69.8	64.2	73.8	68.7
	our method	**80.7**	76.1	**78.4**	**98.3**	**88.1**	**92.9**
SIGHAN2015	LMC(Xie et al.,2015)	83.8	26.2	40.0	67.6	31.8	43.2
	FASPell(Hong et al.,2019)	67.6	60.0	63.5	66.6	59.1	62.6
	Confusionset(Wang et al.,2019)	66.8	73.1	69.8	71.5	59.5	69.9
	SpellGCN(Cheng et al.,2020)	74.8	80.7	77.7	72.1	77.7	75.9
	MLM-phonetics(Zhang et al.,2021)	77.5	**83.1**	80.2	74.9	80.2	77.5
	PLOME(Liu et al.,2021)	77.4	81.5	79.4	75.3	79.3	77.2
	uChecker(Piji Li.,2022)	85.6	79.7	82.6	91.6	84.8	88.1
	our method	**85.7**	83.0	**84.3**	**95.9**	**90.1**	**92.9**

4.4 Results and Analysis

To evaluate the methods, we compared the previous methods and our method based on accuracy, recall, and F1 score at both the character-level and sentence-level. We assessed these metrics for both detection and correction tasks.

As shown in Table 3, we compared the performance of different methods on the character-level evaluation using the SIGHAN test set. We evaluated these methods on the SIGHAN2013, SIGHAN2014, and SIGHAN2015 datasets. In Table 3, we can observe that on the SIGHAN2013 dataset, our method performs slightly lower in terms of precision at the detection level compared to MLM-phonetics, but outperforms other methods in all other evaluation metrics. On the SIGHAN2014 and SIGHAN2015 datasets, compared to other methods, our method surpasses others in terms of precision at both the detection and correction levels.

As shown in Table 4, we present the performance of different methods on the sentence-level evaluation using the SIGHAN test set. Similarly, we compared these methods on the SIGHAN2013, SIGHAN2014, and SIGHAN2015 datasets. In Table 4, we can observe that our method outperforms other methods in both detection and correction levels on the SIGHAN2013, SIGHAN2014, and SIGHAN2015 datasets.

By conducting validation on the character-level and sentence-level using the three SIGHAN datasets, the experimental results demonstrate the effectiveness

Table 4. The sentence-level performance of different models on SIGHAN test sets.

Dataset	Model	Detection Level			Correction Level		
		Prec.	Rec.	F1.	Prec.	Rec.	F1.
SIGHAN2013	LMC(Xie et al.,2015)	(-)	(-)	(-)	(-)	(-)	(-)
	FASPell(Hong et al.,2019)	76.2	63.2	69.1	73.1	60.5	66.2
	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	74.7	**77.7**	81.0
	our method	**96.0**	**95.2**	**96.2**	**98.6**	71.7	**83.0**
SIGHAN2014	LMC(Xie et al.,2015)	(-)	(-)	(-)	(-)	(-)	(-)
	FASPell(Hong et al.,2019)	61.0	53.5	57.0	59.4	52.0	55.4
	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
	our method	**86.9**	**85.6**	**86.2**	**83.6**	**76.9**	**80.1**
SIGHAN2015	LMC(Xie et al.,2015)	(-)	(-)	(-)	(-)	(-)	(-)
	FASPell(Hong et al.,2019)	67.6	60.0	63.5	66.6	59.1	62.6
	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
	PLOME(Liu et al.,2021)	77.4	81.5	79.4	75.3	79.3	77.2
	uChecker(Piji Li.,2022)	75.4	72.0	73.7	70.6	67.3	68.9
	our method	**89.8**	**89.3**	**89.5**	**87.8**	**82.0**	**84.8**

of our method, which is based on Gaussian distribution-enhanced local attention, for CSC. As expected, the local information around the error context in CSC carries more weight for error detection and correction than distant information. This fully validates the feasibility of our approach.

4.5 Ablation Study

In this subsection, we analyze the impact of several factors on the model, including Gaussian attention and masking strategy. We evaluate the effects of different factors through ablation experiments.

We conducted ablation experiments by individually removing different components to study their respective impacts on the model. First, we removed the Gaussian attention layer and compared the performance of the model without Gaussian attention. Next, we removed the masking strategy part of BERT and only used random masking. Finally, we performed experiments using only the BERT model. The experimental results, as shown in Table 5, indicate that the models without the Gaussian attention layer (Our-G) and the masking strategy (Our-M) were affected to varying degrees compared to our model. Furthermore, comparing with the BERT-only model highlights the effectiveness of our approach for the CSC task.

5 Conclusion

In this study, we introduce a novel Gaussian-based local attention enhancement approach specifically tailored for CSC tasks. Our method innovatively applies

Table 5. The ablation result of our method.

Dataset	Method	Detection Level				Correction Level			
		Acc	Prec.	Rec.	F1.	Acc	Pre.	Rec.	F1.
SIGHAN2013	Bert	**77.0**	74.2	83.2	78.6	75.2	83.0	75.2	78.9
	Ours-G	65.7	76.2	63.2	68.9	76.5	78.4	70.6	73.7
	Ours-M	68.2	76.6	82.5	79.6	82.6	76.3	72.8	74.4
	Ours	72.1	**78.7**	**91.6**	**84.7**	**93.8**	**98.7**	**95.9**	**97.3**
SIGHAN2014	Bert	75.7	64.5	68.6	66.5	74.6	62.4	66.3	64.3
	Ours-G	72.3	76.2	70.5	73.2	75.2	85.4	80.9	82.9
	Ours-M	74.5	75.1	69.4	76.1	83.9	83.2	81.5	69.1
	Ours	**77.4**	**80.7**	**76.1**	**78.4**	**91.7**	**98.3**	**88.1**	**92.9**
SIGHAN2015	Bert	82.4	74.2	78.0	76.1	81.0	71.6	75.3	73.4
	Ours-G	80.5	78.5	71.7	74.7	77.8	76.2	80.3	78.4
	Ours-M	81.2	82.3	80.7	81.5	84.2	75.8	71.5	73.8
	Ours	**83.3**	**85.7**	**83.0**	**84.3**	**92.2**	**95.9**	**90.1**	**92.9**

Gaussian attention to assign varying weights to characters, particularly emphasizing the information surrounding misspelled characters. This focus aids in extracting more pertinent information, crucial for both detecting and correcting errors. Furthermore, our approach includes adaptable hyperparameters, within the Gaussian attention layer, leading to significant performance enhancements. To refine our model further,we adopt a homophonic substitution masking strategy and fine-tune the BERT model on a large-scale CSC corpus. The efficacy of our proposed method is underscored by experimental results obtained from the SIGHAN benchmarks, where it surpasses previous methodologies. These findings confirm the effectiveness and potential of our Gaussian-based local attention enhancement method in addressing the challenges of CSC tasks.

Acknowledgments. This work is supported by Xinjiang Uygur Autonomous Region Key Researchand Development Program (No.2022B03010).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hládek, D., Staš, J., Pleva, M.: Survey of automatic spelling correction. Electronics **9**(10), 1670 (2020)
2. Eger, S., vor der Brück, T., Mehler, A.: A comparison of four character-level string-to-string translation models for (OCR) spelling error correction. The Prague Bulletin of Mathematical Linguistics **105**(1), 77 (2016)
3. Zhou, Y., Porwal, U., Konow, R.: Spelling correction as a foreign language. arXiv preprint arXiv:1705.07371 (2017)
4. Martins, B., & Silva, M. J. (2004). Spelling correction for search engine queries. Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004. Proceedings 4, 372-383.

5. Gao, J., Quirk, C.: A large scale ranker-based system for search query spelling correction. In: The 23rd International Conference on Computational Linguistics (2010)
6. Ye, D., Tian, B., Fan, J., Liu, J., Zhou, T., Chen, X., Li, M., & Ma, J. (2023). Improving query correction using pre-train language model in search engines. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (pp. 2999-3008).
7. Liu, C.-L., Lai, M.-H., Chuang, Y.-H., Lee, C.-Y.: Visually and phonologically similar characters in incorrect simplified Chinese words. In: Coling 2010: Posters, pp. 739–747 (2010)
8. Wu, J.-C., Chiu, H.-W., Chang, J.S.: Integrating dictionary and web N-grams for Chinese spell checking. International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 4, December 2013-Special Issue on Selected Papers from ROCLING XXV (2013)
9. Yu, J., Li, Z.: Chinese spelling error detection and correction based on language model, pronunciation, and shape. In: Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 220–223 (2014)
10. Xie, W., Huang, P., Zhang, X., Hong, K., Huang, Q., Chen, B., & Huang, L. (2015). Chinese spelling check system based on n-gram model. In Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (pp. 128-136).
11. Zhang, S., Huang, H., Liu, J., Li, H.: Spelling error correction with soft-masked BERT. arXiv preprint arXiv:2005.07421 (2020)
12. Xu, H.-D., Li, Z., Zhou, Q., Li, C., Wang, Z., Cao, Y., Huang, H., Mao, X.: Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. arXiv preprint arXiv:2105.12306 (2021)
13. Hong, Y., Yu, X., He, N., Liu, N., & Liu, J. (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) (pp. 160-169).
14. Cheng, X., Xu, W., Chen, K., Jiang, S., Wang, F., Wang, T., Chu, W., & Qi, Y. (2020). Spellgcn: Incorporating phonological and visual similarities into language models for chinese spelling check. arXiv preprint arXiv:2004.14166.
15. Wang, H., Wang, B., Duan, J., Zhang, J.: Chinese spelling error detection using a fusion lattice LSTM. Transactions on Asian and Low-Resource Language Information Processing **20**(2), 1–11 (2021)
16. Liu, S., Yang, T., Yue, T., Zhang, F., & Wang, D. (2021). PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 2991-3000).
17. Yeh, J.-F., Lu, Y.-Y., Lee, C.-H., Yu, Y.-H., Chen, Y.-T.: Chinese word spelling correction based on rule induction. In: Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 139–145 (2014)
18. Duan, J., Wang, B., Tan, Z., Wei, X., Wang, H.: Chinese spelling check via bidirectional lstm-crf. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), pp. 1333–1336 (2019)
19. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

20. Wang, D., Tay, Y., & Zhong, L. (2019). Confusionset-guided pointer networks for Chinese spelling check. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 5780-5785).
21. Zhang, R., Pang, C., Zhang, C., Wang, S., He, Z., Sun, Y., Wu, H., & Wang, H. (2021). Correcting Chinese spelling errors with phonetic pre-training. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2250-2261.
22. Li, P. (2022). uChecker: masked pretrained language models as unsupervised Chinese spelling checkers. arXiv preprint arXiv:2209.07068.
23. Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese spelling check evaluation at SIGHAN bake-off 2013. Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 35-42.
24. Yu, L.-C., Lee, L.-H., Tseng, Y.-H., Chen, H.-H.: Overview of SIGHAN 2014 bake-off for Chinese spelling check. In: Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 126–132 (2014)
25. Tseng, Y.-H., Lee, L.-H., Chang, L.-P., Chen, H.-H.: Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In: Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, pp. 32–37 (2015)
26. Wang, D., Song, Y., Li, J., Han, J., Zhang, H.: A hybrid approach to automatic corpus generation for Chinese spelling check. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2517–2527 (2018)

DOI: 10.1007/978-3-031-63759-9_44