IEICE TRANS. INF. & SYST., VOL.E107–D, NO.10 OCTOBER 2024

PAPER

# MISpeller: Multimodal Information Enhancement for Chinese Spelling Correction

**Jiakai LI**[†,††], *Nonmember*, **Jianyong DUAN**[†,††a)], *Member*, **Hao WANG**[†,††], **Li HE**[†,††], *and* **Qing ZHANG**[†,††], *Nonmembers*

**SUMMARY** Chinese spelling correction is a foundational task in natural language processing that aims to detect and correct spelling errors in text. Most spelling corrections in Chinese used multimodal information to model the relationship between incorrect and correct characters. However, feature information mismatch occured during fusion result from the different sources of features, causing the importance relationships between different modalities to be ignored, which in turn restricted the model from learning in an efficient manner. To this end, this paper proposes a multimodal language model-based Chinese spelling corrector, named as MISpeller. The method, based on ChineseBERT as the basic model, allows the comprehensive capture and fusion of character semantic information, phonetic information and graphic information in a single model without the need to construct additional neural networks, and realises the phenomenon of unequal fusion of multi-feature information. In addition, in order to solve the overcorrection issues, the replication mechanism is further introduced, and the replication factor is used as the dynamic weight to efficiently fuse the multimodal information. The model is able to control the proportion of original characters and predicted characters according to different input texts, and it can learn more specifically where errors occur. Experiments conducted on the SIGHAN benchmark show that the proposed model achieves the state-of-the-art performance of the F1 score at the correction level by an average of 4.36%, which validates the effectiveness of the model.

***key words:*** *Chinese spelling correction, multimodal information fusion, feature divergence, replication factor*

## 1. Introduction

Chinese Spelling Correction (CSC) is a fundamental task in Natural Language Processing (NLP), aiming to detect and correct spelling errors in Chinese text. CSC frequently serves as a downstream auxiliary task, ensuring the quality of search engine query texts [1] and academic papers. It plays a pivotal role in addressing spelling errors induced by phonetic and graphic similarities, [2], [3] which predominantly manifest as common issues in Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) systems.

Compared to other languages, Chinese exhibits distinctive characteristics. Firstly, the Chinese language boasts an extensive corpus, with a vast lexicon comprising at least 3,500 commonly used characters, potentially resulting in a wide search space and an uneven distribution of errors. Secondly, through historical evolution, Chinese has developed a comprehensive native pronunciation system, often represented by pinyin, and writing standards. However, this circumstance frequently engenders two notable issues, namely, homophonic characters (homophones) where the same character has multiple identical pronunciations and visually similar characters (homographs) where multiple characters share similar shapes. According to Liu et al.'s research, approximately 83% and 48% of errors in Chinese text can be attributed to phonetic and graphic similarity of characters, respectively [4]. The first sentence in Table 1 provides an example of character phonetic similarity, where the character "学" is misspelled as "雪". In the second sentence, "干" is spelled as "千", exemplifying an instance resulting from graphic similarity of character shapes.

**Table 1** Examples of incorrect character recognition due to phonetic and graphic similarity.

| Type | Sentence | Correction |
|---|---|---|
| Phonological | 今天我要去上雪(xue3) | 学(xue2) |
| Graphic | 一会儿我要去千(qian1)活 | 干(gan4) |

In recent years, pre-trained language models (LLMs) have achieved significant success in the realm of NLP. With the emergence of the Transformer architecture, pre-trained language models, exemplified by BERT, have witnessed an unprecedented evolution [5], [6]. The utilization of pre-trained language models in Chinese spelling correction tasks has emerged as a mainstream methodology. Notable instances include FASpell, Softmasked-BERT, SpellGCN and PLOME [7]–[10]. Concurrently, certain researchers have directed their attention towards the characteristics of Chinese character phonetics and graphic forms, with the aim of augmenting the model's correction capabilities by amalgamating auditory and graphic information [9], [11], [12]. Nevertheless, despite the commendable progress made by antecedent work in this domain, the integration of semantic, phonetic and graphic features—which represent distinct modalities of characters—may encounter challenges stemming from feature disparities during the fusion process. This has led prior works to appear as though they leverage multimodal information for error correction, yet they remain at a superficial level of utilizing multimodal information. Effectively engaging in deep modeling of multimodal information is of paramount importance for the CSC task. In fact, contemporary state-of-the-art methods in the field of CSC prominently feature the explicit or implicit utilization of multimodal information. Implicit utilization encompasses considerations of phonetic

Manuscript received December 19, 2023.
Manuscript revised April 10, 2024.
Manuscript publicized June 7, 2024.

[†]School of Information Science and Technology, North China University of Technology, Beijing, 100144, P.R. China.
[††]CNONIX National Standard Application and Promotion Lab, North China University of Technology, Beijing, 100144, P.R. China.
a) E-mail: duanjy@ncut.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2023EDP7269

similarity between predicted characters and target characters, such as establishing specific training objectives that target both glyptic and phonetic aspects can provide additional multimodal features that facilitate subsequent fusion for error correction purposes, employing Gated Recurrent Units (GRU) to extract phonetic and graphic features from characters, predict target character pronunciation in a coarse-grained, non-adaptive manner, or augmenting the decoding probabilities of characters with similar pronunciation [11], [13]. However, most of these multimodal features are obtained through additional neural networks, which may result in feature disparities. Furthermore, in the fusion process, a simple addition approach may not fully utilise the positive impact of multimodal information on error correction. Explicit employment involves directly encoding individual character phonetics into pronunciation feature vectors as input, or employing a modeling approach that incorporates semantic, phonetic and graphic information of input characters. They introduced a gating mechanism to selectively blend these modalities in order to predict the final result [12], [14]. While Li et al. had made improvements by introducing fine-grained auxiliary tasks based on Liu et al. compared to previous coarse-grained investigations, they primarily focused on the correction performance by utilizing character phonetic feature information, often overlooking the role of the detection module [9], [15]. This further underscores the potential for enhancement in the realm of CSC within the multimodal domain.

To address the aforementioned issues, this paper introduces an efficient Chinese spelling correction system that seamlessly integrates multimodal information, called MISpeller. This spelling correction system is built upon ChineseBERT as its core encoder, integrating semantic, glyptic and phonetic features in a meaningful way for error correction within a single encoder, without relying on external neural networks, requires a complex architectural design, a novel and systematic exploration that has not been previously undertaken in the extant literature. The ChineseBERT pre-trained model combines graphic and phonetic information, two essential features of Chinese characters [16]. Firstly, the model uses an encoder based on ChineseBERT to acquire multimodal information of characters, including semantics, graphic and phonetic features, in a unified manner, which is then directly output as internal features. On the one hand, this approach resolves feature disparities and exhibits higher efficiency and enhanced integration compared to previous methods that rely on the acquisition of external features followed by simple addition or channel fusion. As illustrated in Fig. 1, some previous research endeavors sought to incorporate phonetic and graphic information by integrating external networks within pre-trained language models to generate candidate characters. For instance, Cheng et al. introduced the SpellGCN model, which leveraged BERT to encode node features for each character and subsequently employed graph convolutional neural networks to independently learn the shape and phonetic similarity relationships between characters within a confusion set [7]. However, it

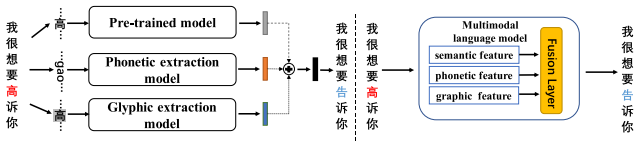


**Fig. 1** A structural comparison of CSC Models.

becomes evident that models trained within this framework exclusively utilize the phonetic and graphic features acquired from the CSC corpus, thereby resulting in a feature mismatch with the semantic information garnered by pre-trained models. In contrast, ChineseBERT, as a pre-trained multimodal language model (as depicted in the right half of Fig. 1), can encode output vectors that encompass the semantic, graphic and phonetic features of characters simultaneously. This characteristic mitigates the need to reintegrate differing features acquired from distinct models and effectively resolves the issue of feature disparity. On the other hand, the exclusive use of ChineseBERT as the backbone of the multimodal Chinese spelling correction system simplifies the overall architecture of the CSC model.

Furthermore, due to a higher error rate stemming from phonetic similarity within the multimodal information, this paper adopts a fine-grained approach in the extraction of phonetic features. In contrast to prior research where the entire phonetic pronunciation of predicted characters was considered, the approach presented in this paper further utilizes fine-grained phonetic features, specifically character initials, finals and tones, to assist in predictions. For instance, in the Chinese language, the pronunciation of a character, such as “高” (gāo), consists of three components: the initial consonant, the final vowel, and the tone. This can be further divided into “g”, “a” and “1”. This fine-grained representation of phonetics not only better encapsulates the intrinsic patterns of Chinese pronunciation but also encodes phonetic similarities among characters. For example, characters like “高” (gāo) and “告” (gào) may not exhibit overall phonetic similarity, but they share the same initial and final sounds, differing only in tone.

Lastly, to address the issue of overcorrection, this paper introduces a replication mechanism, which incentivizes the model to preferentially select the original character when it is effective in both predicting the character and aligning with the input character. Additionally, the model presented in this paper is based on a detection-correction architecture. The detector component aligns closely with the detection task and is implemented using ELECTRA, a model that differs from token prediction tasks [17]. Instead, it trains a discriminator to discern whether each token in the input is generated by the generator. This approach aligns more effectively with actual CSC detection scenarios, further mitigating the occurrence of suboptimal correction outcomes resulting from incorrect information provided by the detection module. Across three manually annotated dataets, SIGHAN 2013, SIGHAN 2014 and SIGHAN 2015, MISpeller exhibits a mean improvement of 3.3%, 4.28% and 5.5% in F1 score in terms of correction

levels compared to current state-of-the-art methods. These results serve as validation of the effectiveness and advancement of the proposed model. The contributions of this paper are summarized as follows:

- The possibility of deep-level multimodal information modeling for Chinese characters was investigated, leading to the proposal of MISpeller, a Chinese spelling correction system that seamlessly integrates multimodal information. Constructed on the foundation of ChineseBERT and ELECTRA, our method can effectively fuses genuine multimodal information within a single model in a single pass, thereby overcoming problems arising from disparate feature representations.
- In the realm of character phonetic features, a fine-grained extraction approach was employed, greatly enhancing the effective modeling of character pronunciation features. To combat the problem of overcorrection, a novel replication mechanism was introduced, motivating the model to choose between the original and predicted characters.
- The model achieved robust generalization performance on the SIGHAN CSC benchmark.

## 2. Related Work

The CSC task constitutes a fundamental task within the domain of NLP. Simultaneously, the inherent intricacies of the Chinese language, such as polysemy and homophony, give this task an exceptional level of complexity. The objective of the CSC task is to detect and correct spelling errors within Chinese sentences. Owing to its practical applicability, particularly in the realm of search queries, CSC technology has garnered substantial attention in recent years. Presently, CSC primarily encompasses three major paradigms: rule-based dictionary methods, statistical language model-based approaches, and deep learning-based methodologies.

Early endeavors in CSC followed a pipeline approach, consisting of error detection, candidate recall and candidate ranking stages. Mangu et el. introduced a rule-based approach for automatically acquiring linguistic knowledge from a small set of comprehensible rules [18]. Chang et al. proposed a method for automatic correction of erroneous Chinese characters by integrating rules with a linear regression model [19]. Subsequently, with the advancement of language models, Liu et al. introduced a hybrid Chinese character correction system based on segmentation language models and statistical machine translation [20]. Yu et al. incorporated character glyph and phonetic features into language models and achieved detection and correction by evaluating sentence or phrase complexity [21]. Due to the ongoing advancements in deep learning, recent research in CSC has shifted its focus towards Deep Neural Networks (DNN). An increasing number of scholars are integrating deep learning methodologies into the domain of CSC. Within the realm of deep learning-based methodologies, two major branches have emerged: those based on the Encoder-Decoder model framework and those rooted in pre-trained language models. The limitation of language models lies in their inability to utilize the input source sentence as an external condition. To address this issue, Wang et al. treated the CSC task as a sequence labeling problem and employed bidirectional LSTMs to predict the correct characters [22]. Subsequently, an increasing body of research has viewed the CSC task as a translation task, namely, an Encoder-Decoder sequence labeling approach. These methods take text containing erroneous characters as input, and through an encoding-decoding process, produce prediction targets of the same length as the source sentence. Afli et al. integrated machine translation principles with the CSC task, resulting in a nearly 13% improvement in model performance compared to the initial baseline [23]. Additionally, Ji et al., Ge et al. and Wang et al. had also conducted research in this direction [24]–[26]. Various pre-trained language models, with BERT as a representative example have proliferated rapidly since the introduction of the Transformer architecture by Vaswani et al. in 2017 [5], [6]. Large-scale pre-trained language models, characterized by their billion-scale parameters and exceptional generalization capabilities, have emerged as a prominent focal point in current research within this field [27]. It is noteworthy that LLMs are an evolutionary development based on the Transformer framework. Zhang et al. employed BERT as a corrector to rectify error positions predicted by Gated Recurrent Units (GRU), thereby introducing a soft masking framework known as Soft-masked BERT [10]. Meanwhile, Hong et al. conducted fine-tuning of BERT using various masking strategies, selecting the optimal candidates as the final correction results [8]. While BERT can be applied to the CSC task with favorable outcomes, the challenge of learning the similarity between target characters and their original counterparts remains non-trivial. Consequently, researchers have endeavored to embed multimodal information of characters into correction models to facilitate the correction process. Wei et al. introduced two novel pre-training objectives for the correction model, designed to separately capture character phonetic and graphic information, which are subsequently integrated with semantic information in later stages to facilitate the correction process [13]. Li et al. introduced a speech prediction auxiliary task that, in conjunction with fine-grained character phonetic features, enabled adaptive weighting for correction [15]. Guo et al. conducted pre-training of BERT using a crafted confusion set that included characters with similar phonetic and graphic features, thus enabling BERT to effectively leverage character phonetic and graphic attributes for correction [28]. In a similar vein, Wang et al. employed a pinyin-enhanced candidate generator based on character pronunciation features and incorporated an attention mechanism to model dependencies between adjacent tokens [29]. Zhang et al. performed joint fine-tuning of detection and correction modules, both pre-trained and featuring phonetic attributes, to enhance the correction process [30]. Liu et al. employed GRU to extract phonetic and graphic characteristics of characters and predicted target character pronunciations in a coarse-grained,

non-adaptive manner [9]. Xu et al. employed a modeling approach that incorporates semantic, phonetic, and graphic information of input characters. They introduced a gating mechanism to selectively blend these modalities in order to predict the final correction [12]. Cheng et al. utilized Graph Convolutional Networks (GCNs) to acquire knowledge regarding the phonetic and graphic similarities among characters [7]. Subsequently, they integrated graphical representations with BERT outputs for the ultimate prediction. In a comprehensive review, Zhang et al. summarized the approaches employed in recent years that leverage multimodal information to assist in CSC. Additionally, they introduced a novel evaluation metric, CCCR, aimed at exploring the balance of character graphic and phonetic features within pre-trained language models [31].

The premise of this study is consistent with that of Xu et al., which involves the use of multimodal information for correction [12]. However, the specific implementation details diverge significantly, which will be elaborated upon in the third section.

## 3. Approach

This section introduces our approach MISpeller for the CSC task. Below, we first define the problem formulation and then describe our approach in detail.

### 3.1 Task Formulation

The task of CSC is to automatically detect and correct text containing potentially incorrect characters. Given a source sentence $X = \{x_1, x_2, \cdots, x_n\}$ containing n characters with spelling errors, the CSC model takes $X$ as input, performs character-level error detection and correction, and produces the correct target sentence $Y = \{y_1, y_2, \cdots, y_n\}$. Since $X$ and $Y$ are of equal length, this task can be viewed as a sequence labelling problem, i.e. modelling the probability $p(Y|X)$. Typically, most of the characters in the sentence are to be copied as they are, with only a small subset containing spelling errors $x_i \in X$ that bear some similarity to their correct counterparts $y_i \in Y$.

### 3.2 MISpeller Architecture

The basic concept of the MISpeller architecture is to train a singular encoder, capable of autonomously capturing and fusing character multimodal information within the model itself, without reliance on external neural networks, thereby resolving the issue of disparate feature fusion resultant from the utilization of distinct models for external knowledge acquisition. The overall structure, as shown in Fig. 2, is constructed based on the prevalent detection-correction framework [10], [32]. The detector is used to detect the positional information of erroneous characters, and the corrector uses this detection information to perform the correction.

Specifically, the first step is to generate embedding vectors for input to the encoder of correction network. In this process, the ELECTRA detector and the corrector share character embedding vectors. Subsequently, within the singular encoder, the model automatically captures the multimodal information of characters, integrating semantic, phonetic and morphological features into a $3D$-dimensional vector. Finally, this vector is mapped to a fused vector of dimensionality $D$ at the output layer for further processing. At the same time, the detector produces vectors containing position error information. These vectors are fed to the correction module for rectification and to the classification layer for prediction of the positions of erroneous characters. Finally, a copying module calculates a copying factor to linearly fuse the original character embeddings, the multimodal hidden layer vectors from the encoder of correction network, and the detection information vector, resulting in the generation of the final encoded vector. This encoded vector is then used as input to the classification layer, initialised with the transposed vocabulary, to produce the prediction result.

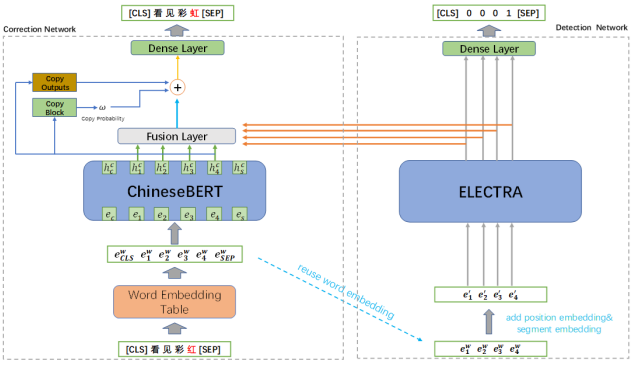


**Fig. 2** The framework of the MISpeller. Note: the generator G in the ELECTRA detection network is only involved during the training phase and is omitted during the inference period.

#### 3.2.1 Detection Network

The objective of the detection network is to identify the positions of erroneous characters in the source sentence, typically formulated as a binary sequence labeling task. In this paper, ELECTRA, which aligns more closely with the detection task, is employed as the foundational framework for the detector. ELECTRA is trained by implementing two neural networks, namely the generator G and the discriminator D, each composed of encoders (e.g., Transformer encoders). The generator G is utilized to predict the masked portions in the input text while directly copying the remaining segments from the source text. The discriminator D, on the other hand, leverages the output of the generator as input to determine, for each token, whether it was generated by the generator G. The encoder in question works by mapping a sequence to the input token $x = (x_1, \cdots, x_n)$, thereby transforming it into a series of contextual vector representations denoted $h(x) = (h_1, \cdots, h_n)$. For an input text of length $n$, after a preprocessing step, its feature sequence $E = (e_1, \cdots, e_n)$ is obtained. In this context, $e_i$ represents the feature vector of character $x_i$, which is a cumulative vector sum consisting

of semantic embedding, position embedding and segment embedding. This feature sequence is then used as input to the ELECTRA generator G, and an external softmax layer is used to calculate the probability that a given token $x_t$ is the output:

$$p_G(x_t|x) = exp(e(x_t)^T h_G(x)_t) / \sum_{x'} exp(e(x')^T h_G(x)_t) \quad (1)$$

where $p_G(x_t|x)$ denotes a conditional probability, signifying the likelihood of a particular token being associated with the original character being masked. $e$ represents the character feature vector.

For a given position $t$, the discriminator D assesses the correctness of token $x_t$ concerning the original character. We represent the sequence of the last hidden layer states of the discriminator as $h_D = (h_D^1, \ldots, h_D^n)$. On the one hand, the output vector $h_D$ is fed into the error correction network for integration, and on the other hand, $h_D$ is propagated to subsequent fully connected layers to generate a probability distribution:

$$D(x,t) = Sigmoid\left(w^T h_D(x)_t\right) \quad (2)$$

where $D(x,t)$ represents the probability of character error at position $t$, with 1 denoting an incorrect character and 0 signifying a correct one. $w^T$ represents the transposed learnable binary classification parameters of the fully connected layer.

### 3.2.2 Correction Network

The correction network is a multi-class task with ChineseBERT as its backbone, used to determine the replacement of misspelled characters with the correct ones. Chinese characters, being logographic, encapsulate critical information in their graphic and phonetic features. In this context, Sun et al. introduced the groundbreaking Chinese pre-trained language model, ChineseBERT [16]. By combining two crucial features of Chinese characters, namely their graphic and phonetic representations, ChineseBERT outperforms some downstream tasks in Chinese language processing compared to pre-trained models that focus only on semantic features. However, the model falls short of fully capturing multimodal information and autonomously performing CSC tasks. Therefore, as shown in Fig. 3, this paper utilizes this framework as the foundation of the correction module. The corrector supplements the semantic features missing in ChineseBERT for the first time, thus creating a model with true internal fusion of multimodal information capability. This aspect has not been addressed in previous research. On the one hand, the encoder of correction network first captures the semantic, phonetic and graphic features of the character input vector and concatenates them together, mapping them to a common dimension through a fully connected layer to create a fused feature vector. This fused feature vector, along with positional encoding vectors, constitutes the input to BERT. Simultaneously, both whole-word masking and character-level masking are employed during pretraining. By capturing and integrating the multimodal information of characters within a single model without the need for additional networks, this approach significantly alleviates ambiguities between different feature dimensions. On the other hand, with regard to phonetic feature extraction, the model adopts a fine-grained approach. Specifically, it extracts the initial, final and tone as phonetic features of characters, as opposed to prior methods that captured entire pinyin representations. This fine-grained phonetic feature extraction enables a more robust modeling of the phonetic similarity between the target character and incorrect characters. Initially, following data preprocessing, the requisite embedding vectors are fed into the correction network, which shares semantic feature vectors with the input of the detection network. The sequence of hidden layer states from the final layer output of the encoder of correction network is denoted as $h_C = (h_C^1, \ldots, h_C^n)$. Subsequently, the hidden layer state sequences of the detection and correction networks are fused. In this study, the dimension of the last hidden layer of both networks is set to be the same, allowing for the straightforward addition of their outputs to obtain the fused representation:

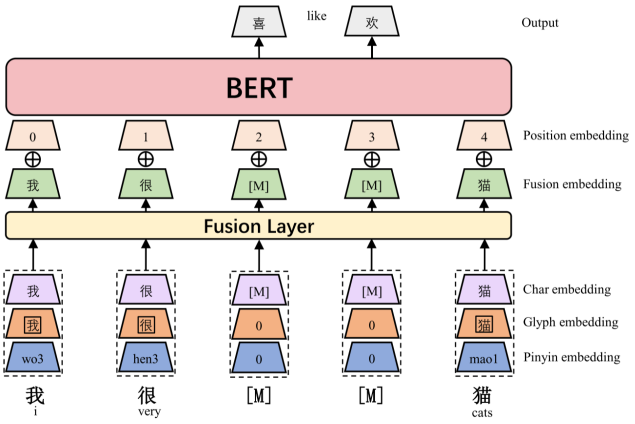


**Fig. 3** An overview of ChineseBERT.

$$h = h_D + h_C \quad (3)$$

where $h_D$ signifies the sequence of hidden states from the final layer of the ELECTRA detection network, and $h_C$ represents the sequence of hidden states from the final layer of the ChineseBERT correction network.

To address the issue of overcorrection, this paper introduces a replication mechanism that uses a replication factor to encourage the model to favour the retention of the original character when both the predicted character and the original character are viable. The replication factor, denoted as $\omega \in R$, is computed by a double-layer feedforward neural network with layer normalization. The specific calculation process is outlined in the following formula:

$$h_{copy} = W_{ch} f_{In}(h_C) + b_{ch} \quad (4)$$

$$h'_{copy} = f_{In}(f_{act}(f_{copy})) \quad (5)$$

$$\omega = Sigmoid(W_c h'_{copy}) \tag{6}$$

where $h_C$ represents the last-layer hidden state representation of the correction network, $W_{ch} \in R^{768\times d_c}$, $b_{ch} \in R^{d_c}$, $W_c \in R^{d_c\times 1}$ denote the model parameters, $f_{act}$ corresponds to the GeLU activation function, and $f_{In}$ represents the layer normalization function. The final output of the correction network is as follows:

$$h' = h_C * \omega + h * (1 - \omega) \tag{7}$$

In previous CSC models, only $h$ was used as the final output, leading to the unnecessary problem of overcorrection in the model [7], [9], [10]. In contrast, by incorporating the probability of retaining the original letter in the final output vector, MISpeller allows for a higher probability of selecting the input character when it is valid but not necessarily optimal. Lastly, rather than viewing the correction task as a multi-classification problem with a randomly initialised projection layer, it is treated as a similarity task. This perspective is based on the premise that if a character at a given position is correct, the final encoding vector produced by the detection and correction network should closely resemble the word embedding of the input character. Conversely, if the character at that position is incorrect, the final encoding vector is expected to show similarity to the word embedding of the candidate correction character. The formal formulation of this classification task is as follows:

$$P(y_i|x) = Softmax(Wh') \tag{8}$$

where $W$ represents the learnable parameters of the fully connected classification layer. In particular, the transpose of the word embeddings is used to initialise the $W$ of the classification layer, as opposed to random initialisation. This choice stems from the fact that word embeddings serve as mappings from word IDs to word vectors, making their parameters a natural fit for the transposed configuration of the classification layer. This initialisation strategy speeds up training and imparts greater stability to the model.

### 3.3 Learning

The detection task is seen as a binary classification cross-entropy problem, where the objective is to determine whether a character is correct or erroneous. At the same time, the correction task assumes the form of a multi-class classification problem, where the focus is on identifying the identity of the correct character. The loss functions are formalised as follows:

$$L^d = -\sum_{i=1}^{n} log\ D(x,t) \tag{9}$$

$$L^c = -\sum_{i=1}^{n} log\ P(y_i|x) \tag{10}$$

where $L^d$ and $L^c$ are the loss functions for the training of the detection network and the correction network, respectively. Ultimately, these individual loss functions are linearly combined to form the overall loss function:

$$L = \alpha L^c + (1 - \alpha) L^d \tag{11}$$

where $\alpha \in [0, 1]$ is a hyperparameter, and experimental validation has determined its value to be 0.8. The objective is to minimize the overall loss function, thereby facilitating end-to-end training of the comprehensive model.

### 3.4 Inference

As noted by Liu et al. and Devlin et al., the performance of correction models on multi-error texts has left much to be desired [5], [33]. To address this issue and further mitigate the problem of overcorrection, this paper introduces a cyclical correction strategy during inference. Specifically, during the inference phase, character detection and correction are performed iteratively for each input sentence, rather than attempting to detect and correct all errors in a single pass. It is specified that during each iteration only errors in a specified window size around the correction positions from the previous iteration are considered for correction. If a given position is corrected in each iteration, it is returned to its original character without further correction. The number of iterations is empirically set at 3 and the window size is fixed at 5. By running multiple rounds of correction, selecting a certain number of low-probability words after each prediction as targets for the next round of correction, and incorporating correction information from the previous round into the next, richer contextual information is made available, thereby improving the quality of the correction process.

## 4. Experiments

### 4.1 Datasets

Based on previous work in the field of CSC, this paper utilizes the official SIGHAN training data, as well as 271K synthetic data generated by Wang et al. to constitute the training dataet [22], [34]–[36]. To evaluate the performance of the proposed model, three test dataets from the SIGHAN 2013, SIGHAN 2014 and SIGHAN 2015 benchmarks are used as the test sets. The statistics for the training and test sets are presented in Table 2. As the original SIGHAN dataets are in traditional Chinese characters, the OpenCC[†] tool is used to convert the characters in the datasets to simplified Chinese, in line with previous research. In addition, pypinyin[††] is applied to obtain the pinyin representation for each character and segment it into initial, final and tone according to a pre-defined vocabulary of initials and finals provided by Xu et al. [12]. Furthermore, due to the poor quality of the manually annotated corpora in the SIGHAN 2013 test dataet, a significant number of instances involving the mixed use of

[†]https://github.com/BYVoid/OpenCC
[††]https://pypi.org/project/pypinyin

**Table 2** Statistics information of the used data resources. The number in brackets in the #Line column indicates the number of sentences with errors.

| Training Data | #Line | Avg.Length | #Errors |
|---|---|---|---|
| SIGHAN 2013 | 350 | 49.2 | 350 |
| SIGHAN 2014 | 6526 | 49.7 | 10087 |
| SIGHAN 2015 | 3174 | 30.0 | 4237 |
| Wang271K | 271329 | 44.4 | 271329 |
| Test Data | #Line | Avg.Length | #Errors |
| SIGHAN 2013 | 1000 | 74.1 | 996 |
| SIGHAN 2014 | 1062 | 50.1 | 529 |
| SIGHAN 2015 | 1100 | 30.5 | 550 |

the "的", "地" and "得" particles remain unannotated, resulting in suboptimal scores for high-performing models on this dataet. To mitigate this problem, this paper also employs the post-processing method proposed by Xu et al., which involves removing all detected and corrected instances of "的", "地" and "得" from the model output [12].

### 4.2 Evaluation Metrics

In the context of previous evaluations in the field of CSC, the most commonly used evaluation metrics are precision (P), recall (R) and F1 score, and these metrics are calculated at two different levels: character-level scores and sentence-level scores. Character-level evaluation involves evaluating detection and correction on a character-by-character basis, while sentence-level evaluation considers a sentence to be correct only if all errors within it are successfully detected and corrected. This paper uses character-level P, R and F1 as evaluation metrics for the following reasons:

- Character-level features represent the minimal evaluation unit in CSC, thus character-level metrics can reflect the ability of a model in finer gained.
- The CSC test corpus contains a significantly larger number of characters than sentences, making character-level statistics a more reliable measure due to its scale.

### 4.3 Implementation Details

In this paper, the pre-trained ChineseBERT is used as the correction network. In order to speed up the convergence of the model, the input embedding layer of the correction network is utilized within the detection module to initialise the weights of ELECTRA. Throughout the specific training process, where the dimensions of the hidden layer feature vectors are set to 768. The learning rate is set to 2e-5 with linear decay, a dropout rate of 0.1 is applied, the batch size is set to 32, training spans 20 epochs with print intervals of every 20 steps, and the optimisation algorithm is Adam.

### 4.4 Baselines

We compare our model with the following methods:

- SpellGCN (Cheng et al., 2020): A pre-defined character confusion set is applied by Graph Convolutional Networks (GCN) to a Bert-based correction model [7].
- GAD (Guo et al., 2021): This approach learns the global relationships between potential correct input characters and candidates for potential erroneous characters [28].
- DCN (Wang et al., 2021): By employing a distinctive dynamic connected network, $K^n$ path are generated in the output phase of the model (where $K$ is the number of candidate words and $n$ is the sentence length). Subsequently, the dynamic connected network is used for scoring to select the optimal path. An attention mechanism is introduced to model the dependency relationships between adjacent characters [29].
- REALISE (Xu et al., 2021): This approach encodes both phonetic and glyphic information based on semantic cues and ultimately introduces gating mechanisms to selectively fuse semantic, phonetic and glyphic information for predictive output [12].
- MDCSpell (Zhu et al., 2022): This method designs a multi-task framework, where BERT is used as a corrector to capture the visual and phonological features of the characters and integrated character position information from a detector to reduce error interference [37].

**Table 3** Experimental results of each model on the CSC13, CSC14 and CSC15 test sets.

| Dataset | Model | Detection-level | | | Correction-level | | |
|---|---|---|---|---|---|---|---|
| | | D-P | D-R | D-F | C-P | C-R | C-F |
| SIGHAN 2015 | SpellGCN | 88.9 | 87.7 | **88.3** | 95.7 | 83.9 | 89.4 |
| | GAD | 88.6 | **87.8** | 88.2 | 96.3 | 84.6 | 90.1 |
| | DCN | 87.1 | 80.6 | 83.7 | 85.7 | 80.1 | 82.8 |
| | REALISE | **89.5** | 85.5 | 87.5 | 89.2 | 84.3 | 86.7 |
| | MDCSpell | 86.9 | 87.3 | 87.1 | 87.1 | 85.0 | 86.0 |
| | **MISpeller(ours)** | 89.1 | 86.6 | 87.8 | **98.1** | **87.5** | **92.5** |
| SIGHAN 2014 | SpellGCN | 83.6 | 78.6 | 81.0 | 97.2 | 76.4 | 85.5 |
| | GAD | **85.1** | **80.9** | **82.9** | 98.0 | 79.2 | 87.6 |
| | DCN | 77.4 | 78.4 | 77.9 | 88.8 | 77.7 | 82.9 |
| | REALISE | 80.4 | 77.8 | 79.1 | 86.7 | 80.0 | 83.2 |
| | MDCSpell | 82.6 | 78.0 | 80.2 | 86.9 | 81.1 | 83.9 |
| | **MISpeller(ours)** | 83.9 | 80.7 | 82.3 | **98.2** | **81.2** | **88.9** |
| SIGHAN 2013 | SpellGCN | 82.6 | 88.9 | 85.7 | 98.4 | 88.4 | 93.1 |
| | GAD | 85.8 | **89.5** | 87.6 | 99.0 | 88.6 | 93.5 |
| | DCN | 88.1 | 85.9 | 87.0 | 89.9 | 88.2 | 89.0 |
| | REALISE | **90.6** | 86.5 | 88.5 | 90.2 | 87.2 | 88.7 |
| | MDCSpell | 90.5 | 86.2 | 88.3 | 97.5 | 86.5 | 91.7 |
| | **MISpeller(ours)** | 90.1 | 89.1 | **89.6** | **99.1** | **90.3** | **94.5** |

### 4.5 Main Results

Table 3 shows the evaluation results of the proposed model and five baseline models for character-level detection and correction performance on the SIGHAN 2013, SIGHAN 2014 and SIGHAN 2015 test sets. The results in the table are highlighted in bold to indicate the best performance results.

As shown in Table 3, when tested with MISpeller, the model achieves significant performance gain over the other baselines, especially in the F1 score of the correction module. This clearly demonstrates the effectiveness of the proposed approach. Moreover, compared to the latest and most competitive MDCSpell baseline, the method significantly improves the F1 score of the correction module by 2.8%, 5.0% and 6.5%, respectively, indicating that the ChineseBERT-based correction module effectively mitigates the problem of disparate multi-feature information during the multimodal

fusion process, thus validating the effectiveness of the correction network in the proposed multitask architecture. At the same time, it can be observed that the precision and recall results of the model are significantly better than the previous baseline models. Compared to the best performance of the baseline model, the model shows a precision improvement of 0.1%, 0.2% and 1.8%, respectively, and a recall improvement of 1.7%, 0.1% and 2.5%, respectively. This improvement is mainly due to the improved use of detection information by the correction module, which effectively reduces the number of false positives. It also highlights the ability of the replication mechanism to alleviate overcorrection problems, resulting in a significant recall improvement for the model.

It is worth noting that while the approach has achieved significant performance in terms of precision and F1 score, the recall of the detection model on the test sets shows a certain gap compared to the other baseline models. The reason for this disparity can be attributed to the fact that, unlike SpellGCN and GAD, the approach does not incorporate external knowledge such as confusion set [7], [28]. However, artificially constructed confusion set inherently impose limitations by restricting error correction cues, thereby limiting the model's ability to generalise. At the same time, other base models such as REALISE require additional pre-training due to the need to integrate different modules. Instead, the competitiveness of the model proposed in this paper, even without the use of external knowledge and the escalation of training costs, indirectly underlines the effectiveness of the method.

### 4.6 Ablation Study

In this section, this paper investigates the effect of the hyperparameter $\alpha$ in the loss function on the performance of the model, as well as the encoder and the replication mechanism to the MISpeller. This ablation study performs character-level evaluations of the model using the SIGHAN 2015 test set.

In the multi-task learning, the effect of the hyperparameter $\alpha$ within the loss function on the model performance is illustrated in Fig. 4. Experimental results show that as the value of $\alpha$ increases, the F1 score of the correction module also shows continuous improvement. However, once $\alpha$ exceeds the threshold of 0.8, the F1 score experiences a sharp decline. Consequently, setting $\alpha$ to 0.8 achieves an overall optimal correction F1 score. A closer look reveals that this result Is well-founded. This is because, compared to the detection task, the correction task poses greater convergence challenges, requiring higher weights during the learning process. At the same time, an excessively high $\alpha$ reduces the contribution of the detection model during learning, thus limiting its expected performance. Ultimately, a relative higher $\alpha$ achieves the overall best balance between the learning of these two tasks.

As MISpeller uses a different and potentially more powerful encoder (i.e., ChineseBERT) compared to the baselines, we further conduct experiments to eliminate the effects of different encoders and focus solely on the disambiguation

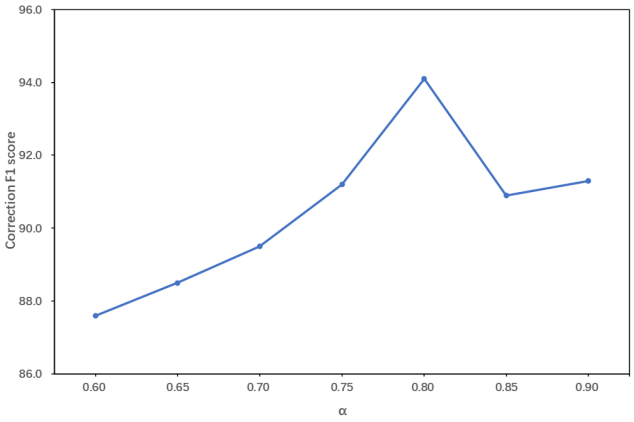


**Fig. 4** The impact of the choice of the hyperparameter $\alpha$ in the loss function.

**Table 4** Ablation results on the test set of SIGHAN15.

| Model | $P_c$ | $R_c$ | $F1_c$ |
|---|---|---|---|
| MISpeller | **98.1** | **87.5** | **92.5** |
| REALISE | 89.2 | 84.3 | 86.7 |
| MISpeller(REALISE) | **91.2** | **85.1** | **88.0** |
| BERT-base | 80.6 | 77.3 | 78.9 |
| MISpeller(BERT-base) | **82.3** | **78.1** | **80.1** |
| MISpeller(-Copy) | 93.4 | 87.2 | 90.2 |

task of multimodal information fusion, which is the main contribution of this work. This paper conducts ablation studies on the SIGHAN 2015 test set with the following settings to investigate and maintain consistency with other parameters:

- Replacing the ChineseBERT encoder with the REALISE encoder [12].
- Replacing the ChineseBERT encoder with BERT-base [5].
- Removing the replication mechanism module.

The experimental results are shown in Table 4, exclusively showcase the precision, recall, and F1 score for correction. The principal aim of this paper is to leverage multimodal language models to uncover relationships between characters, thereby addressing the issue of disparate multi-feature fusion information. The Table 4 indicates that any change in the individual modules leads to varying degrees of degradation in the performance of the correction network. In particular, when the encoder is replaced by BERT-base, the evaluation metrics of the correction model show a significant decrease, indicating the positive role of the multimodal information of the characters in the performance of the model. Replacing the encoder with the REALISE encoder leads to a decrease in the precision, recall and F1 score of the correction module by 6.9%, 2.4% and 4.5% respectively. This suggests that the acquisition of multiple features from a multimodal language model at the same time, compared to additional neural network-derived features, mitigates the problem of feature information inequality during fusion, allowing the model to effectively blend semantic, phonetic and glyphic information, which demonstrates the effective-

ness of the method proposed in this paper. Meanwhile, when the encoder component is replaced with REALISE, although there is a decrease in performance compared to MISpeller, it consistently outperformed the baseline model REALISE. The precision, recall and F1 scores of the correction module increased by 2%, 0.8% and 1.3%, respectively. These results verify the effectiveness of our approach irrespective of the encoder. Removing the replication mechanism module results in a decrease of 4.7%, 0.3% and 2.3% in the precision, recall and F1 score of the correction module, respectively. This demonstrates the effectiveness of the proposed replication mechanism in mitigating overcorrection problems, with a smaller decrease in performance compared to MISpeller (REALISE), indirectly indicating the greater contribution of multimodal information to the improvement in error correction performance. In summary, any modification to the modules within the model has an impact on the performance of MISpeller, underlining the importance of the different modules proposed in this paper.

### 4.7 Analysis of Multimodal Information Fusion

In this subsection, we further analyze the impact of the different fusion methods of character multimodal information (semantic, phonetic, and graphical) on the performance of Chinese spelling check. Previous research has shown that the integration of semantic, phonetic and graphical features of characters can improve the correction performance of models. However, whether multimodal information is effectively integrated has not been sufficiently considered. This is because the fusion of different features may lead to information inequality, which could potentially limit the auxiliary role of multimodal information in spelling correction. Prior studies using multimodal information for spelling correction have mainly relied on additional neural networks to extract phonetic and graphical features of characters before fusion, of which REALISE is a typical representative. Therefore, we use REALISE as the baseline model for this experiment and implement the following settings: a) masking the extraction of phonetic features from the correction network encoder and using an external speech extractor GRU network to capture phonetic features, followed by simple addition fusion with other features(-Phonetic with adding). b) masking the extraction of graphical features from the correction network encoder and using ResNet for external pre-fusion capture, followed by selective modality fusion with other features(-Graphic with selective modality). c) using only one font type for the construction of graphical input in the correction network encoder and applying static weight linear fusion outside the model(-Multi-Fonts with static weight).

Table 5 shows the average scores of the correction network across three SIGHAN test sets. The main objective of this study is to improve the correction performance by reducing the phenomenon of feature information inequality during multimodal fusion. From the experimental results above, it can be observed that the transition from the intra-model fusion proposed in this paper to the inter-model fusion of different models, whether at the level of phonetic or graphical features, leads to different degrees of performance degradation. The reason is that the aforementioned three traditional fusion methods all belong to the external fusion of the model, which may lead to an unequal distribution of multi-feature information during fusion, thereby reducing the efficiency of the use of the model for multimodal information. This not only indicates the ability of the proposed method to disambiguate multiple features effectively and to use multimodal information more efficiently for correction, but also reaffirms the positive role of multimodal information in correction, thereby validating the rationality and efficacy of the proposed approach. Furthermore, although the performance of the model weakens after changing the fusion approach, it consistently outperforms the baseline model REALISE. This is because REALISE uses multimodal information for correction, but uses different models to extract features for external fusion, which reduces fusion efficiency and limits model performance.

**Table 5** The average experimental results of different feature fusion configurations in MISpeller on the SIGHAN test sets.

| Model | $P_c$ | $R_c$ | $F1_c$ |
|---|---|---|---|
| REALISE | 88.7 | 83.8 | 86.2 |
| MISpeller | 98.4 | 86.3 | 91.9 |
| -Phonetic with adding | 97.5 | 85.0 | 90.8 |
| -Graphic with selective modality | 97.8 | 85.2 | 91.0 |
| -Multi-Fonts with static weight | 97.6 | 85.1 | 90.9 |

## 5. Conclusion

This paper proposes MISpeller, which is a Chinese spelling correction model based on a multimodal language model. The architecture automatically captures and combines contextual semantics, phonetic and graphic features, resulting in a unified output. And it uses an end-to-end training approach within a detection-correction framework. Previous research has emphasized the importance of contextual semantics, phonetic cues and glyphic information. In contrast to existing studies that use auxiliary networks to capture phonetic and glyphic features, this paper leverages the ChineseBERT multimodal language model to generate these features simultaneously. This approach effectively alleviates the problem of unequal feature fusion while promoting parameter efficiency, thereby reducing training costs. Furthermore, to address the issue of overcorrection by the fine-tuned ChineseBERT in multi-typo texts, this paper introduces a replication mechanism. It applies a replication factor as a weight to fuse multimodal information, which encourages the model to prefer to choose the input character when the miscorrected and input character are both valid according to the given context. Experimental results on publicly available datasets show that MISpeller outperforms all compared models and has superior generalization capabilities in CSC. In future research, the key to improving model performance will be the deep and fine-grained extraction of multimodal information and its efficient fusion. At the same time, the

search for detectors better suited to CSC discrimination scenarios to improve error detection capability of the model, and the use of state-of-the-art large language model for the construction of corpora to address the scarcity of CSC datasets are crucial points for research in this area.

## Acknowledgments

This work is supported by the Humanities and Social Science Foundation of the Ministry of Education (Grant No. 62476007).

### References

[1] J. Duan, T. Ji, and H. Wang, "Error correction for search engine by mining bad case," IEICE Trans. Inf. & Syst., vol.E101-D, no.7, pp.1938–1945, 2018.
[2] T.T.H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of post-ocr processing approaches," ACM Computing Surveys (CSUR), vol.54, no.6, pp.1–37, 2021.
[3] S. Park, D. Shin, S. Paik, S. Choi, A. Kazakova, and J. Lee, "Improving distinction between asr errors and speech disfluencies with feature space interpolation," arXiv preprint arXiv:2108.01812, 2021.
[4] C.L. Liu, M.H. Lai, Y.H. Chuang, and C.Y. Lee, "Visually and phonologically similar characters in incorrect simplified chinese words," COLING (Computational Linguistics), pp.739–747, 2010.
[5] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol.30, 2017.
[7] X. Cheng, W. Xu, K. Chen, S. Jiang, F. Wang, T. Wang, W. Chu, and Y. Qi, "Spellgcn: Incorporating phonological and visual similarities into language models for chinese spelling check," arXiv preprint arXiv:2004.14166, 2020.
[8] Y. Hong, X. Yu, N. He, N. Liu, and J. Liu, "Faspell: A fast, adaptable, simple, powerful chinese spell checker based on dae-decoder paradigm," Proc. 5th Workshop on Noisy User-generated Text (W-NUT 2019), pp.160–169, 2019.
[9] S. Liu, T. Yang, T. Yue, F. Zhang, and D. Wang, "Plome: Pre-training with misspelled knowledge for chinese spelling correction," Proc. 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, 2021.
[10] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling error correction with soft-masked bert," arXiv preprint arXiv:2005.07421, 2020.
[11] T. Ji, H. Yan, and X. Qiu, "Spellbert: A lightweight pretrained model for chinese spelling check," Proc. 2021 conference on empirical methods in natural language processing, pp.3544–3551, 2021.
[12] H.-D. Xu, Z. Li, Q. Zhou, C. Li, Z. Wang, Y. Cao, H. Huang, and X.-L. Mao, "Read, listen, and see: Leveraging multimodal information helps chinese spell checking," arXiv preprint arXiv:2105.12306, 2021.
[13] X. Wei, J. Huang, H. Yu, and Q. Liu, "Ptcspell: Pre-trained corrector based on character shape and pinyin for chinese spelling correction," Findings of the Association for Computational Linguistics: ACL 2023, pp.6330–6343, 2023.
[14] L. Huang, J. Li, W. Jiang, Z. Zhang, M. Chen, S. Wang, and J. Xiao, "Phmospell: Phonological and morphological knowledge guided chinese spelling check," Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp.5958–5967, 2021.
[15] J. Li, Q. Wang, Z. Mao, J. Guo, Y. Yang, and Y. Zhang, "Improving chinese spelling check by character pronunciation prediction: The effects of adaptivity and granularity," arXiv preprint arXiv:2210.10996, 2022.
[16] Z. Sun, X. Li, X. Sun, Y. Meng, X. Ao, Q. He, F. Wu, and J. Li, "Chinesebert: Chinese pretraining enhanced by glyph and pinyin information," arXiv preprint arXiv:2106.16038, 2021.
[17] K. Clark, M.T. Luong, Q.V. Le, and C.D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.
[18] L. Mangu and E. Brill, "Automatic rule acquisition for spelling correction," ICML, pp.187–194, Citeseer, 1997.
[19] T.-H. Chang, H.-C. Chen, and C.-H. Yang, "Introduction to a proofreading tool for chinese spelling check task of sighan-8," Proc. Eighth SIGHAN Workshop on Chinese Language Processing, pp.50–55, 2015.
[20] X. Liu, K. Cheng, Y. Luo, K. Duh, and Y. Matsumoto, "A hybrid chinese spelling correction using language model and statistical machine translation with reranking," Proc. seventh SIGHAN workshop on chinese language processing, pp.54–58, 2013.
[21] J. Yu and Z. Li, "Chinese spelling error detection and correction based on language model, pronunciation, and shape," Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp.220–223, 2014.
[22] D. Wang, Y. Song, J. Li, J. Han, and H. Zhang, "A hybrid approach to automatic corpus generation for chinese spelling check," Proc. 2018 Conference on Empirical Methods in Natural Language Processing, pp.2517–2527, 2018.
[23] H. Afli, Z. Qui, A. Way, and P. Sheridan, "Using smt for ocr error correction of historical texts," https://doras.dcu.ie/23226/1/Using%20SMT%20for%20OCR%20Error%20Correction%20of%20Historical%20Texts.pdf, 2016.
[24] T. Ge, F. Wei, and M. Zhou, "Fluency boost learning and inference for neural grammatical error correction," Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1055–1065, 2018.
[25] J. Ji, Q. Wang, K. Toutanova, Y. Gong, S. Truong, and J. Gao, "A nested attention neural hybrid model for grammatical error correction," arXiv preprint arXiv:1707.02026, 2017.
[26] D. Wang, Y. Tay, and L. Zhong, "Confusionset-guided pointer networks for chinese spelling check," Proc. 57th Annual Meeting of the Association for Computational Linguistics, pp.5780–5785, 2019.
[27] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," arXiv preprint arXiv:2304.13712, 2023.
[28] Z. Guo, Y. Ni, K. Wang, W. Zhu, and G. Xie, "Global attention decoder for chinese spelling error correction," Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp.1419–1428, 2021.
[29] B. Wang, W. Che, D. Wu, S. Wang, G. Hu, and T. Liu, "Dynamic connected networks for chinese spelling check," Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp.2437–2446, 2021.
[30] R. Zhang, C. Pang, C. Zhang, S. Wang, Z. He, Y. Sun, H. Wu, and H. Wang, "Correcting chinese spelling errors with phonetic pre-training," Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp.2250–2261, 2021.
[31] X. Zhang, Y. Zheng, H. Yan, and X. Qiu, "Investigating glyph phonetic information for chinese spell checking: What works and what's next," arXiv preprint arXiv:2212.04068, 2022.
[32] J. Li, G. Wu, D. Yin, H. Wang, and Y. Wang, "Dcspell: A detector-corrector framework for chinese spelling error correction," Proc. 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.1870–1874, 2021.
[33] S. Liu, S. Song, T. Yue, T. Yang, H. Cai, T. Yu, and S. Sun, "Craspell: A contextual typo robust approach to improve chinese spelling correction," Findings of the Association for Computational Linguistics:

ACL 2022, pp.3008–3018, 2022.
[34] Y.-H. Tseng, L.-H. Lee, L.-P. Chang, and H.-H. Chen, "Introduction to sighan 2015 bake-off for chinese spelling check," Proc. Eighth SIGHAN Workshop on Chinese Language Processing, pp.32–37, 2015.
[35] S.H. Wu, C.L. Liu, and L.H. Lee, "Chinese spelling check evaluation at sighan bake-off 2013," Proc. Seventh SIGHAN Workshop on Chinese Language Processing, pp.35–42, 2013.
[36] L.-C. Yu, L.-H. Lee, Y.-H. Tseng, and H.-H. Chen, "Overview of sighan 2014 bake-off for chinese spelling check," Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp.126–132, 2014.
[37] C. Zhu, Z. Ying, B. Zhang, and F. Mao, "Mdcspell: A multi-task detector-corrector framework for chinese spelling correction," Findings of the Association for Computational Linguistics: ACL 2022, pp.1244–1253, 2022.

**Jiakai Li** is a master's student, born in 1998, now he studies at North China University of Technology. His major research field including natural language processing and artificial intelligence.

**Jianyong Duan** is a professor, born in 1978.10. He graduated from department of computer science, Shanghai Jiao Tong University by 2007.12. His major research field including natural language processing and artificial intellegence.

**Hao Wang** received the Ph.D. degree in Computer Application Technology from Tsinghua University in 2013. He is now an associate professor in College of Informatics, North China University of Technology. His research interests include machine learning and data analysis.

**Li He** is an associate professor, graduated from Yanshan University in 2002 with a master's degree. Now she works in the Department of Computer Science, North China University of Technology. The main research interests include data warehouse and data mining, large database processing.

**Qing Zhang** is currently an Assistant Professor in the Department of Computer Science at North China University of Technology (NCUT). Before joining NCUT, he has served as a principal researcher in the field of conversational AI at Huawei Co., Ltd. He got Ph.D from Ministry Of Education Key Lab of Computational Linguistics, School of Computer Science at Peking University (PKU). His research focuses on natural language processing and machine learning.