electronics

MDPI

Article

PMDRSpell: Dynamic Residual Chinese Spelling Check Method Guided by Phonological and Morphological Knowledge

Guanguang Chang [1], Yangsen Zhang [1,*], Youren Yu [2] and Jiayuan Song [3]

[1] Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing 100192, China; 2022020945@bistu.edu.cn
[2] Algorithm Center, China Electronics Cloud Computing Technology Cooperation, Beijing 100141, China; yuyouren@cestc.cn
[3] School of Automation, Beijing Information Science and Technology University, Beijing 100192, China; 2024020510@bistu.edu.cn
* Correspondence: zys@bistu.edu.cn

Abstract: Since the errors in Chinese Spell Correction (CSC) involve phonetically or morphologically confusing Chinese characters, mainstream models have made numerous attempts to fuse phonological and morphological knowledge. We observe that in erroneous sentences where the vast majority of Chinese characters are correctly written, mainstream models may unintentionally increase the difficulty of predicting these correct characters when integrating multi-modal knowledge across all characters. Additionally, these models often overlook the potential relationship between the phonological and morphological modalities of a Chinese character when utilizing multi-modal information. In this paper, we propose an end-to-end model called PMDRSpell, which models erroneous Chinese characters in sentences using their multi-modal knowledge and reduces the use of multi-modal information for correct Chinese characters. Additionally, it uncovers the relationship between phonological and morphological features based on the characteristics of phonograms, enhancing the similarity between similar Chinese characters. Specifically, coarse-grained and hierarchical detection is first employed to localize and mask error locations within sentences, using the original embedding information as residual features. Next, correlation information in the phonological and morphological modalities of the erroneous characters is extracted to construct new representational features, which are then used to update the erroneous Chinese character information within the residual features. Finally, the masked sentences are predicted using the MLM model and classified to generate correct sentences by combining the residual features with the updated multi-modal information. Our model effectively reduces the interference from correct Chinese characters during the inspection process and leverages multi-modal information to accurately correct incorrect Chinese characters. In our comparison experiments with recent state-of-the-art models, PMDRSpell outperforms the optimal baseline in terms of error-corrected F1 scores for Sighan14 and Sighan15 by 1.2 and 1.0 percentage points, respectively.

Keywords: natural language processing; Chinese spelling check; dynamic residual; multi-modal

check for updates

Citation: Chang, G.; Zhang, Y.; Yu, Y.; Song, J. PMDRSpell: Dynamic Residual Chinese Spelling Check Method Guided by Phonological and Morphological Knowledge. *Electronics* **2024**, *13*, 4989. https://doi.org/10.3390/electronics13244989

Academic Editor: Arkaitz Zubiaga

Received: 3 November 2024
Revised: 12 December 2024
Accepted: 16 December 2024
Published: 18 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Chinese Spelling Check (CSC), as a basic task in natural language processing, aims at detecting and correcting the spelling errors of Chinese characters that appear in the text due to the confusing pronunciation or shape of some Chinese characters, such as Table 1, which shows that the word "字 (character)" in the word "金字塔 (pyramid)" is mistakenly written as "子 (son)", and the word "覆 (cover)" in the word "颠覆 (subversion)" is mistakenly written as "复 (repeat)". These spelling errors are mainly due to human writing errors or machine recognition errors, such as automatic speech recognition (ASR) or optical character recognition (OCR) [1]. CSC can improve the quality of Chinese information, which is important for information retrieval (IR), automatic speech recognition (ASR) and other fields.

Table 1. Examples of Chinese spelling errors. Word labeled in red indicates the wrong word in the sentence, and word labeled in blue indicates its corresponding correct word.

Err.	古埃及金子塔。
Cor.	古埃及金字塔。
Trans.	Ancient egyptian pyramids.
Err.	时代力量正在颠复着家电行业。
Cor.	时代力量正在颠覆着家电行业。
Trans.	The power of the times is subverting the home appliance industry.

Current mainstream Chinese Spelling Correction (CSC) models employ sequence annotation methods and leverage multi-modal information of Chinese characters (including morphological, acoustic, and semantic modalities) to enhance error correction efficacy. Most models integrate the multi-modal information of each Chinese character [2–5], inadvertently increasing interference with correct character information and creating a performance bottleneck. It is important to note that in CSC tasks, correct characters constitute the majority, while incorrect characters are relatively few. Ideally, multi-modal information should target incorrect characters more effectively while minimizing interference with correct ones.

Furthermore, mainstream multi-modal models often overlook the correlation between phonological and morphological modalities. Some studies [6,7] even disregard morphological modality knowledge, focusing instead on Pinyin confusion errors, which is not a generalized approach. In domains with frequent OCR usage, glyph confusion errors are predominant. Typically, a Chinese character's shape guides its sound, and the sound reflects the shape, mirroring the evolution of Chinese characters and the learning process for children. This intrinsic connection is evident when people can often approximate the pronunciation of unfamiliar characters.

To address these issues, we propose PMDRSpell, a dynamic residual generalized spell-checking model based on multi-modal information fusion, built on Soft-Masked Bert [8]. By selectively utilizing the visual and acoustic modal knowledge of erroneous parts within the residual network, we correct input information while reducing the interference of multi-modal information on the correct parts. Recognizing that pre-trained language models (PLMs) acquire substantial knowledge during the pre-training phase, we suggest an efficient method to harness this latent knowledge by masking the locations of "non-multiple word errors" [9] in the sentence. Additionally, we design a parallel cross-attention fusion mechanism called Paca, which emphasizes the correlation between the phonological and morphological modalities of the same erroneous Chinese character, thereby enhancing the PLM model's efficiency in leveraging multi-modal information. Ultimately, PMDRSpell outperforms mainstream methods on generalized benchmarks.

In summary, this work has the following three main contributions: (1) A coarse-grained and hierarchical error detection model is proposed. (2) A novel parallel cross-attention fusion mechanism is proposed. (3) A novel dynamic residual mechanism is proposed for the CSC task.

2. Related Work

The task of CSC was initially accomplished using rule and statistics-based methods [10–12]. However, these methods cannot dig deeper into the semantics and detect the hidden errors.

In recent years, Chinese text spelling correction models have been dominated by pre-trained sequence labeling models. In the field of general text spell checking, Zhang et al. [8] proposed Soft-Masked BERT, a proofreading model that assigns weights to masks based on the probability of word errors at each position, using these weights for prediction. Since most Chinese spelling errors involve characters that are similar in shape or sound, Huang et al. [3] introduced the PHMO model, which extracts character shape, sound, and semantic features, and then adaptively fuses them using VGG19 [13], Tacotron2 [14], and BERT [15], respectively. Wang et al. [16] proposed the Dynamic Connected Network,

which generates candidate Chinese characters through a Pinyin Enhanced Candidate Generator and models dependencies between adjacent characters using an Attention Network. Li et al. [17] developed the LEAD framework, which learns heterogeneous knowledge of phonetics, visuals, and meanings from a lexicon. Lv et al. [5] introduced ECSpell, which integrates multi-modal knowledge and employs an error-consistent masking strategy to construct pre-training data. Song et al. [6] introduced a retrieval framework for RSpell, which improves the model's proofreading of specialized domain terms by retrieving relevant terms and inputting them along with sentences into the CSC model. Another study by Liu proposed ReLM [18]. ReLM transforms the mainstream CSC sequence labeling task into a paraphrasing task to better leverage the pre-trained knowledge of PLM models, which focuses more on extracting semantic information.

However, mainstream models that fuse multi-modal information also significantly impact the correct Chinese characters and do not adequately explore the correlations between modalities. This work focuses on addressing these issues and making improvements.

3. Approach

3.1. Problem Formulation

CSC is a sequence generation task in which the sequence lengths of the required inputs and outputs are the same. In general, there are not many errors in a sentence, so Chinese spelling correction only needs to change very few Chinese characters, while the others remain unchanged. One of the characteristics of the CSC task is that most of the erroneous characters are morphologically or phonetically close characters. Therefore, the model for CSC needs to consider more information about the phonology and morphology of the erroneous characters in the correction process.

3.2. Model

As shown in Figure 1, the model is composed of three core modules: the Error Detection Module, the Multi-modal Dynamic Residual Module, and the Error Correction Module. The process begins with the Error Detection Module, which identifies and localizes erroneous regions in the input text containing Chinese characters. These identified regions are then processed by the Multi-modal Dynamic Residual Module, which extracts and fuses multi-modal information to enhance the understanding of the context and nature of the errors. Finally, the Error Correction Module combines the outputs of the previous two modules to predict and correct the identified errors, producing a refined and accurate output.

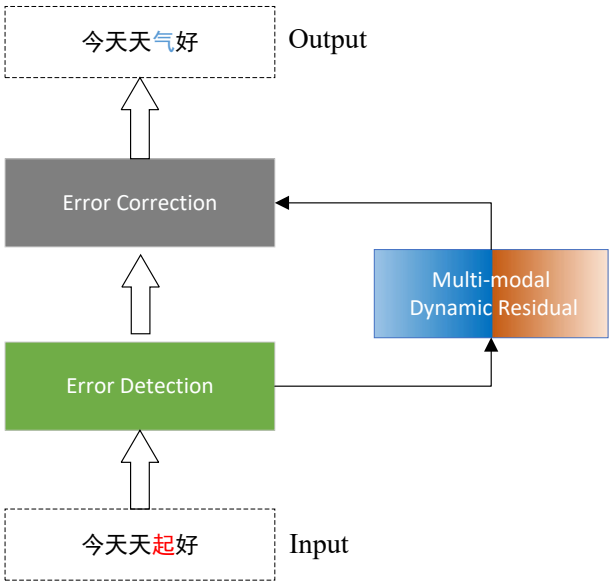


Figure 1. Overall flow chart. The corrected sentence is "It's a beautiful day".

The specific process is shown in Figure 2. First, in the error detection module, sentences containing errors are input into the error detection network, where the locations of the erroneous words are identified and appropriately masked. Next, in the multi-modal dynamic residual module, the glyph and phonological features of the erroneous Chinese characters are extracted. The parallel cross-attention fusion mechanism is used to explore the correlations between the modalities, and these are fused with the original word embedding to obtain new representational features for the erroneous characters, which are then used to update the residual matrix. Finally, the error correction network generates corrected decoding vectors, which are fused with the updated residual matrix. A classifier then selects the words with the highest probability from the word list to produce the final corrected sentence.

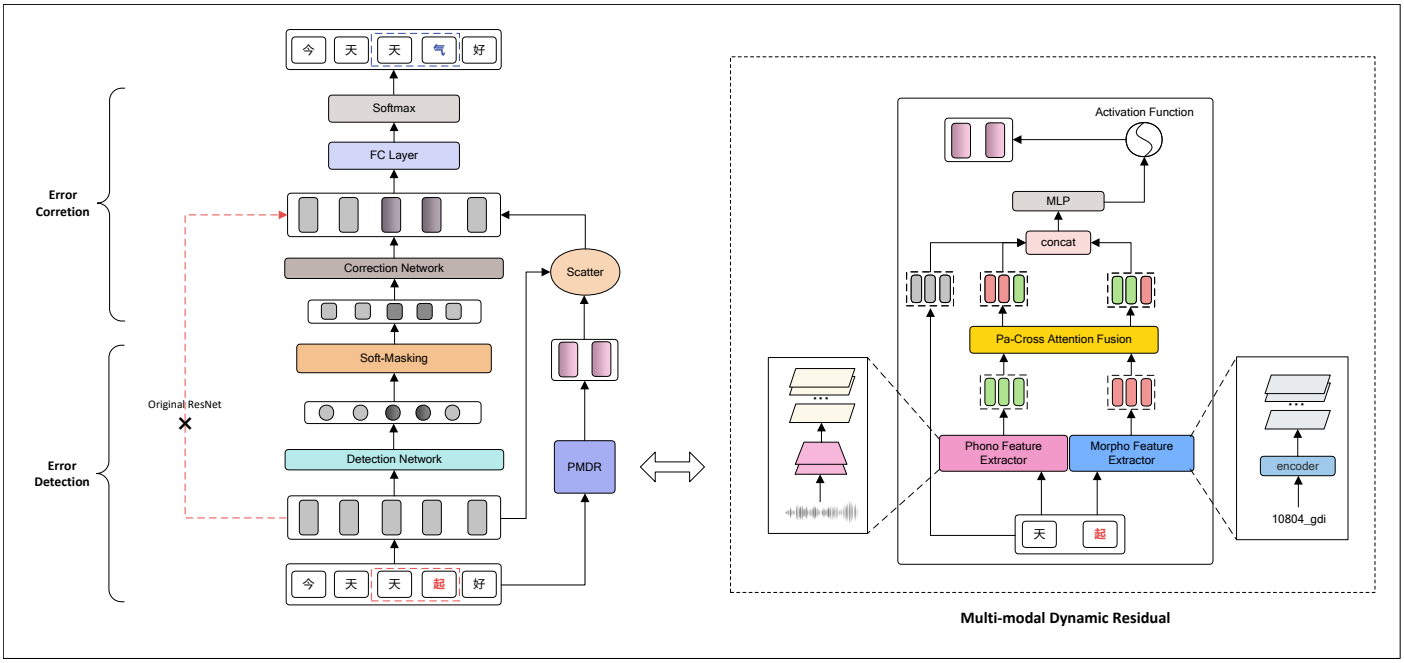


Figure 2. PMDRSpell architecture diagram. × indicate that it is not really used in the model. "Pa-Cross Attention Fusion" denotes a cross-fusion module for morphological features and phonological features. The corrected sentence is "It's a beautiful day".

Next, we will describe the details of this model.

3.3. Error Detection Module

Previous models use fine-grained detection methods, which do not fully utilize the information of Chinese words. Zhang et al. [9] proposed the concept of "non-multiple-word error", which refers to the fact that a Chinese multiple-word (the number of Chinese characters is greater than or equal to 2) has an error due to an alias substitution error of one or more Chinese characters, resulting in an error in the word itself. Inspired by the concept of "non-multiple word error", we use the Chinese word segmentation tool called "Hanlp" to segment the samples and detect the position of "non-multiple word error" in the sentence using a coarse-grained method.

As shown in Figure 3, the Chinese segmentation result of the original sentence "今天天气好 (It's a beautiful day.)" is "今天 (today)/天气 (weather)/好 (good)", which miswrites "天气 (weather)" as "天起 (from the sky)". In our work, the position of the error word "天起 (from the sky)" is expected to be detected instead of just the error character "起 (rise)". At the same time, we use ON-LSTM [19] as the error detection model, which utilizes hierarchical structure information to improve the detection of "non-multiple word errors". Different from the classical LSTM structure, ON-LSTM proposes a hierarchical concept, where the low and high hierarchical levels represent the structures with smaller

and larger granularity in the text, respectively. In Chinese text, character is the structure with the lowest hierarchical level, followed by words, and then phrases, and so on. Neurons with high hierarchical levels usually represent long-term or global information and are updated less frequently; neurons with low hierarchical levels usually represent short-term or local information and are updated more frequently.

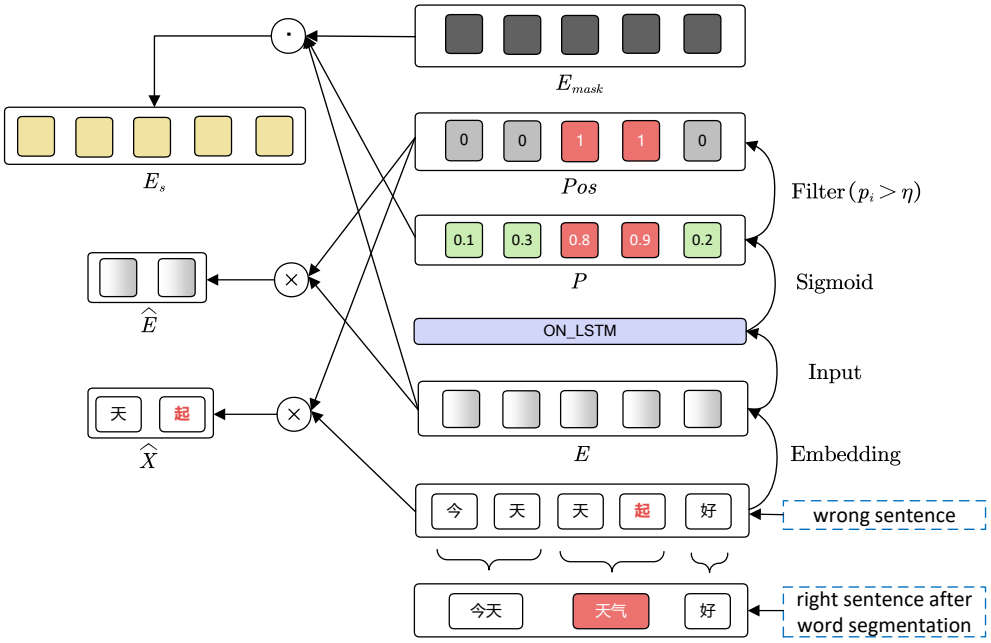


Figure 3. Process of error word selection. $\otimes$ denotes the mask operation and $\odot$ denotes the soft-masking operation. The process progressively obtains the mask of the misspelled word from the bottom up and then obtains the misspelled word in the sentence and its embedding.

We define the input sentence as $X = [x_1, x_2, \ldots, x_n]$ where x_i denotes the i_{th} Chinese character in the sentence and n denotes the sequence length; the output state of the last layer of ON-LSTM as $H^d = \left[h_1^d, h_2^d, \ldots h_n^d\right]$ where $h_i^d \in \mathbb{R}^d$ and d is the dimension uniformly specified in this model; the labeling data of the detection network as $G = [g_1, g_2, \ldots g_n]$ where the labeling data means that the $g_i = 1$ Chinese character is wrong and $g_i = 0$ means right, and the error probability is $P = [p_1, p_2, \ldots, p_n]$, where $p_i \in \mathbb{R}$ is the error probability of the i_{th} Chinese character, which could be calculated as shown in (1):

$$p_i = P_d(g_i = 1 \mid X) = sigmoid\left(W_d h_i^d + b_d\right) \tag{1}$$

where $P_d(g_i = 1|X)$ denotes the conditional probability of the word in the i_{th} position is predicted to be wrong when the input sentence is X; $W_d \in \mathbb{R}^d$ and $b_d \in \mathbb{R}$ denote the parameters to be learned. Drawing on the work of Soft-Masked Bert [8], we use soft-masking here to dynamically mask words at different positions according to their error probabilities to alleviate the problem of correction failures due to detection errors. Specifically, we linearly compute the error probability with the word embedding to obtain the soft-masked embedding. We define the embedding of the input sentence as $E = [e_1, e_2, \ldots, e_n]$, where $e_i \in \mathbb{R}^d$. The soft-masked embedding is defined as $E^s = \left[e_1^s, e_2^s, \ldots, e_n^s\right]$, where $e_i^s \in \mathbb{R}^d$ for the i_{th} word is computed as shown in (2):

$$e_i^s = p_i \cdot e_{mask} + (1 - p_i) \cdot e_i \tag{2}$$

where $e_{mask} \in \mathbb{R}^d$ is in the mask embedding $E_{mask} = [e_{mask}, e_{mask}, \ldots, e_{mask}]$. Firstly, the error probability of each Chinese character position is obtained in the error detection

module, and then we use the error probability to obtain the suspected character position: $Pos = [pos_1, pos_2, pos_3, \ldots, pos_n]$ where $pos_i = 1$ indicates that the i_{th} character may be incorrect and $pos_i = 0$ indicates that the i_{th} character may be right. The embedded feature $\hat{E} \in \mathbb{R}^{n \times d}$ and the Chinese characters $\hat{X}$ corresponding to this position are screened using the error probability threshold η. The specific computation is as shown in (3) and (4):

$$pos_i = Filter(p_i > \eta) \tag{3}$$

$$\hat{E} = Mask(Pos, E);\ \hat{X} = Mask(Pos, X) \tag{4}$$

where *Filter* denotes the filtering operation that picks out locations where the probability of error exceeds the threshold, η denotes the error probability threshold, *Mask* denotes the masking operation, and E and X denote the input text and its embedding.

3.4. Multi-Modal Dynamic Residual Module

3.4.1. Phonological Feature Extraction

Unlike the traditional pinyin of Chinese characters, real audio can express the phonological features of Chinese characters more realistically. Therefore, we use the audio spectrum of Chinese characters to generate the phonological features of Chinese characters in the phonological feature extraction part. In order to facilitate the extraction of phonological features for Chinese characters in subsequent work, we use PaddleSpeech [20], which can convert Chinese text to realistic audio, to pre-construct a mapping table of Chinese characters' audio encoding, and then the error words located by the error detection module are converted into audio encoding by this mapping table. Finally, we use the convolution module of wav2vec2.0 [21] to extract the phonological features of the error words $\hat{X}$, which are defined as $F_a \in \mathbb{R}^{n \times d}$.

3.4.2. Morphological Feature Extraction

Chinese characters can be characterized using four-corner coding [22] and five-stroke font [23] to represent their glyphs. Four-corner coding encodes a Chinese character by the upper-left, upper-right, lower-left, and lower-right strokes of the character, and five-stroke font encodes a Chinese character by the radicals and strokes of the character. As shown in Table 2, the similarity of the four-corner codes of similar characters is usually high, and the four-corner codes of "已" and "己" are even identical. However, the similarity of their five-stroke character patterns is usually low, and the five-stroke character patterns of the similar characters, such as "蓁" and "集", are even completely different. Therefore, it is not enough to use either the four-corner code or the five-stroke font alone to represent the morphological features, and the shorter sequences will lead to an insufficient characterization ability of the extracted glyph features.

Therefore, we splice the four-corner coding and the five-stroke font together as the glyph sequence of the Chinese character. Similar to the coding table for phonological features, we pre-construct a Chinese character glyph coding table, convert the error words localized by the error detection module into glyph codes through this coding table, and then input them into the transformer encoder to obtain the morphological features of the error words $\hat{X}$, which are defined as $F_f \in \mathbb{R}^{n \times d}$.

Table 2. Cases of glyph sequences of similar Chinese characters, including the four-corner sequence, the five-stroke sequence, and the combined sequence of the two. This work uses the edit distance between strings to indicate the similarity between them. Outliers in the similarity are shown in **bold**.

Word	Four-Corner Sequence		Five-Stroke Sequence		Merged Sequence	
	Coding	Similarity	Coding	Similarity	Coding	Similarity
体	25230	60%	wsgg	50%	25230_wsgg	60%
休	24290		wsy		24290_wsy	

Table 2. *Cont.*

Word	Four-Corner Sequence		Five-Stroke Sequence		Merged Sequence	
	Coding	Similarity	Coding	Similarity	Coding	Similarity
已 己	17717 17717	**100%**	nngn nnnn	75%	17717_nngn 17717_nnnn	90%
雅 雁	70215 71215	80%	ahty dwwy	25%	70215_ahty 71215_dwwy	60%
蒸 集	44904 20904	60%	adwt wysu	**0%**	44904_adwt 20904_wysu	40%

3.4.3. Parallel Cross-Attention Fusion

Phonograms are the most dominant type of Chinese characters, which are composed of a phonetic radical that provides phonetic clues to the whole character and a morphological radical that provides semantic clues to the whole character, and these components exhibit specific positional and functional patterns [24]. Phonetic radicals are extremely important for the cognition of phonograms. As shown in Figure 4, from the perspective of regularity, the pronunciation of some Chinese characters is identical or very similar to their phonetic radicals, e.g., the pronunciation of "功" is the same as its phonetic radical "工"; from the perspective of consistency, there are strong similarities in the pronunciation of Chinese characters of the same phonetic radical family, such as 工/gong/, 红/hong/, 虹/hong/, 江/jiang/ (the word in '//' is the pinyin of the Chinese character), in which many of them are pronounced differently from their phonetic radicals "工", but have very similar phonetics. It is inferred that Chinese character features are indicative of character sounds and vice versa, yet most existing models have overlooked this relationship.

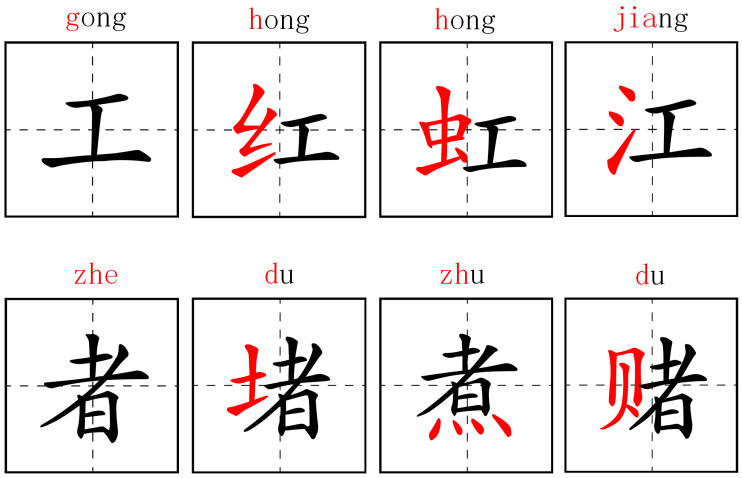


Figure 4. Examples of phonograms (each line represents a phonetic radical family). In pinyin, red symbols indicate different syllables, while black symbols indicate similar syllables. In Chinese characters, red symbols indicate the morphological radicals, black symbols indicate the phonetic radicals.

Therefore, this work proposes a novel parallel cross-attention fusion mechanism called Paca, which utilizes the attention mechanism to explore the correlation between phonological and morphological features and extracts more complete and deeper Chinese character features, thus enhancing the effectiveness of text error correction.

Each block in the Transformer model contains a multi-head attention module, where the correlation between characters at different positions can be learned and captured [25]. The computational process is shown in (5)∼(7):

$$head_p = Attention\left(QW_p^Q, KW_p^K, VW_p^V\right) \quad (5)$$

$$MultiHead(Q, K, V) = Concat(head_1, \cdots, head_h)W^O \tag{6}$$

$$MH_{a\&n}(Q, K, V) = Norm(MultiHead(Q, K, V)) + Q \tag{7}$$

where p and h denote the current number of heads and the total number of heads of multi-head attention, respectively; $head_p$ denotes the output of the p_{th} attention head; Q, K, V denote the query matrix, key matrix, and value matrix; W_p^Q, W_p^K, W_p^V denote the weight matrices to be learned; *Concat* denotes the splicing operation; *Norm* denotes the normalization operation; *Attention*, *MultiHead*, $MH_{a\&n}$ represent the attention layer, multi-head attention layer, and multi-head attention layer after residuals and normalization, respectively. The cross-attention extends the self-attention mechanism by computing cross-modal attention weights, where the features of one modality serve as queries and the features of another modality act as keys and values. This approach enables the establishment of associations between different modalities, thereby facilitating the interaction and fusion of information across modalities.

Prior to modal fusion, we first combine the phonological features and morphological features to obtain a joint representation $C \in \mathbb{R}^{n \times d}$. This joint representation is then used for subsequent cross-modal retrieval and fusion using the cross-attention mechanism. The process of fusing the joint representation C is detailed in Equation (8).

$$C = tanh\left(F_a W F_f\right) \tag{8}$$

where *tanh* is a nonlinear activation function, and $W \in \mathbb{R}^{d \times n}$ is the weight parameter to be learned. We constructed a parallel bottom-up network stacked with several base attention modules, where each base module consists of a multi-head attention and residual normalization layer.

As shown in Figure 5, we devised a parallel cross-attention fusion module named Paca, which harnesses a multi-layer attention mechanism to unearth the similarities between different modalities. Paca encompasses two parallel feature fusion channels, namely the phonological features fusion channel (left side of the figure) and the morphological features fusion channel (right side of the figure), where the parameters of the same layer in different channels are shared.

Specifically, in the fusion channel for phonological features, the extracted phonological features are first encoded using the first module, which is the self-attention module, and then the joint representation C is input to the next layer's attention module together with the output of this layer, where the joint representation C serves as the K and V of the next attention module, and the output of the upper module serves as the Q. The same is true for the morphological features. After several iterations, the phonological features fused with the morphological features $F_a^{'} \in \mathbb{R}^{n \times d}$ and the morphological features fused with the phonological features $F_f^{'} \in \mathbb{R}^{n \times d}$ are finally obtained. The specific ways are shown in (9)∼(11):

$$\begin{cases} Q_i^a, K_i^a, V_i^a = MH_{a\&n}\left(Q_{i-1}^a, K_{i-1}^a, V_{i-1}^a\right), C, C & 1 < i \leq N \\ Q_i^f, K_i^f, V_i^f = MH_{a\&n}\left(Q_{i-1}^f, K_{i-1}^f, V_{i-1}^f\right), C, C & 1 < i \leq N \end{cases} \tag{9}$$

$$\begin{cases} Q_i^a = K_i^a = V_i^a = F_a & i = 1 \\ Q_i^f = K_i^f = V_i^f = F_f & i = 1 \end{cases} \tag{10}$$

$$F_a^{'} = MH_{a\&n}(Q_N^a, K_N^a, V_N^a); F_f^{'} = MH_{a\&n}\left(Q_N^f, K_N^f, V_N^f\right) \tag{11}$$

where i denotes the current number of layers; N denotes the total number of layers in the fusion channel; Q_i^a, K_i^a, V_i^a and Q_{i-1}^a, K_{i-1}^a, V_{i-1}^a denote the query, key, and value matrix of the i_{th} and $(i-1)_{th}$ layer in the phonological fusion channel, respectively. Similarly, Q_i^f, K_i^f, V_i^f and Q_{i-1}^f, K_{i-1}^f, V_{i-1}^f denote the query, key, and value matrix of the i_{th} and $(i-1)_{th}$ layer in the morphological fusion channel, respectively. After multiple layers of

parallel cross-attention fusion, phonological features are integrated into the space of morphological features. Specifically, each morphological feature is paired with a phonological feature through a cross-attention layer. Similarly, phonological features are integrated into the space of morphological features, with each phonological feature being associated with the morphological feature. Throughout the multi-layer cross-parallel fusion, irrelevant information is gradually diminished, while the correlation between different modalities is enhanced, resulting in a richer feature representation.

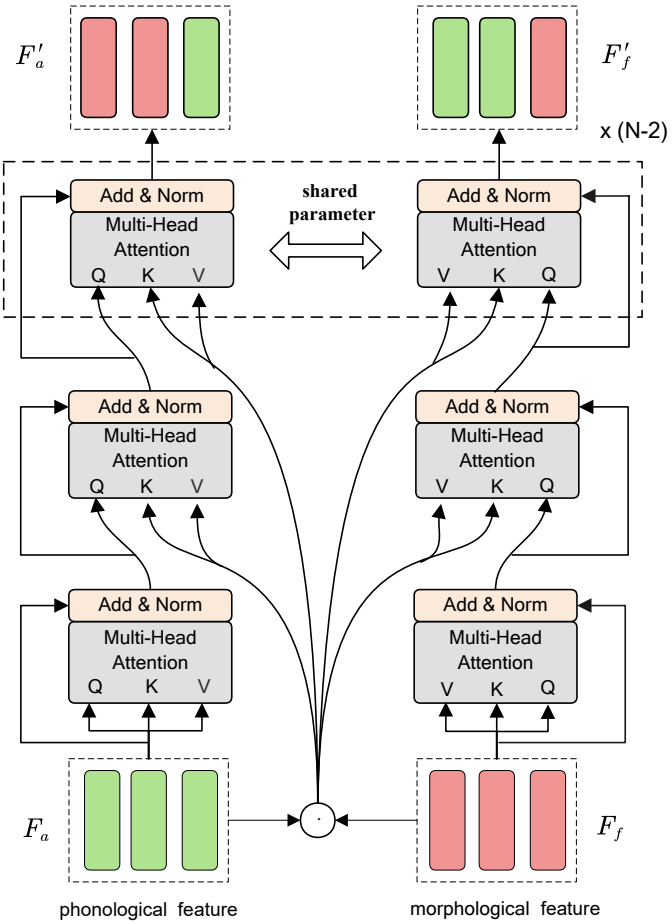


Figure 5. Parallel cross-attention fusion module (Paca). ⊙ denotes the matrix multiplication. "Add & Norm" denotes the residual and normalisation layers.

3.4.4. Dynamic Residual

Mainstream multi-modal Chinese Spelling Correction (CSC) models extensively leverage the multi-modal knowledge associated with each Chinese character within a sentence. Nevertheless, CSC erroneous sentences typically contain a few incorrect characters that closely resemble correct ones amid a majority of correctly written characters. Recognizing this, it becomes apparent that focusing more on the multi-modal information of erroneous characters and reducing the emphasis on correct characters is imperative.

An additional concern arises from Soft-Masked BERT's disregard for the multi-modal information of Chinese characters. Instead, it relies solely on the embedding information of incorrect characters within its residual structure, potentially leading to model misguidance and hindered convergence.

To address these shortcomings, we propose a dynamic residual mechanism built upon the foundation of the Soft-Masked BERT model [8]. This mechanism dynamically adjusts the residual embedding matrix by incorporating the multi-modal information of erroneous characters during model updates, as shown in Figure 6, to achieve precise utilization of multi-modal information.

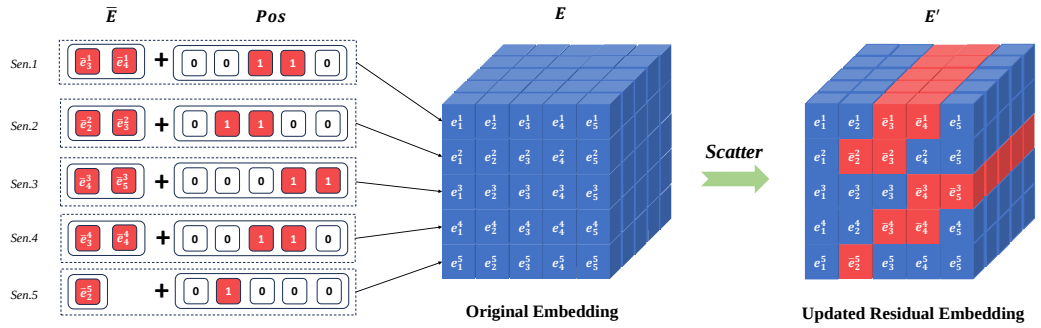


Figure 6. Dynamic residual mechanism. In the updated residual embedding, those labeled in red are the parts that are dynamically corrected by the mechanism, and those labeled in blue are the unaltered parts. *Sen.x* denotes the x_{th} statement and *bs* denotes the size of the batch.

After obtaining phonological features $F_f^{'}$ and morphological features by Paca, we perform a further fusion and adaptive adjustment to derive the ultimate multi-modal feature $\bar{E} \in \mathbb{R}^{n \times d}$. The specific way is shown in (12):

$$\bar{E} = tanh\left(\left[\hat{E}, F_a^{'}, F_f^{'}\right] W^{'} + b^{'}\right) \tag{12}$$

where $\left[\hat{E}, F_a^{'}, F_f^{'}\right]$ denotes the splice of $\hat{E}, F_a^{'}, F_f^{'}$ along the second dimension; $W^{'} \in \mathbb{R}^{3d \times d}$ and $b^{'} \in \mathbb{R}^{d}$ are parameters to be learned. As shown in Figure 6, we adjust the original embedding using the ultimate multi-modal features to obtain the updated residual embedding, which we define as $E^{'}$. The specific update method is shown in (13):

$$E^{'} = Scatter\left(E, \bar{E}, Pos\right) \tag{13}$$

where *Scatter* denotes the mapping function. In the process of correcting the residual matrix with multi-modal information, the embeddings of the correct Chinese characters remain unchanged, while the embeddings of the incorrect characters are replaced with their multi-modal fusion information. This approach allows the CSC model to preserve the correct information in the sentence during the error correction process and utilize the multi-modal information of the incorrect characters to infer their correct form, thereby facilitating the model's convergence.

3.4.5. Error Correction Module

This work uses Bert [15] as an error correction model. We input the soft-masked hidden states E^s into Bert and define the output of Bert as $H^c = \left[h_1^c, h_2^c, h_3^c, \ldots, h_n^c\right]$, where $h_n^c \in \mathbb{R}^d$. Finally, after processing linearly using dynamic the residual mechanism, the word with the highest probability at each position in the word list is then selected in the manner shown in (14) and (15):

$$h_i^{'} = h_i^c + e_i^{'} \tag{14}$$

$$P_c(y_i = j \mid X) = softmax\left(W_c h_i^{'} + b_c\right)[j] \tag{15}$$

where $h_i^{'} \in \mathbb{R}^d$ denotes the hidden state obtained by linear combination with the dynamic residual mechanism; $e_i^{'} \in \mathbb{R}^d$ denotes the fused features on updated residual embedding; $P_c(y_i = j|X)$ denotes the conditional probability of selecting the j_{th} character in the character list when the input sentence is X; $softmax$ denotes the softmax function; $W_c \in \mathbb{R}^d$ and $b_c \in \mathbb{R}$ are parameters to be learned.

4. Experiments

4.1. Dataset and Metrics

Dataset This work uses sighan13 [26], sighan14 [1], sighan15 [27], and wang271k [28] as the training set for model training and the test sets of sighan13, sighan14, and sighan15 for model evaluation. We refer to the work of Cheng [29] to convert traditional Chinese characters to simplified ones to facilitate the experimental follow-up.

Metrics In the evaluation metrics section, we use sentence-level metrics as the metrics for the error detection and error correction sections, which require that all errors in a sentence be detected or corrected to be considered successful. Meanwhile, accuracy, precision, recall, and F1 score are used as the common metrics for both parts of the evaluation.

4.2. Baselines

We compare some state-of-the-art models in recent years with our model. The selected models are shown below:

- Soft-Masked Bert [8]: Soft-Masked Bert completes error correction by masking different locations according to their error probabilities after error detection.
- PHMO [3]: PHMO fuses multi-modal features by extracting pinyin, glyph and semantic features separately and using adaptive gates.
- LEAD [17]: the LEAD framework enables the CSC model to learn heterogeneous knowledge from the lexicon in terms of speech, vision and meaning.
- MDCSpell [4]: MDCSpell is a framework that uses BERT to capture multi-modal features and enhance the detector-corrector.
- ECSpell [5]: ECSpell employs an error-consistent masking strategy to create data for pre-training and enhance the model.
- DORM [7]: DORM separates the two types of features, text and speech, to allow direct interaction between text and speech information.
- ReLM [18]: ReLM transforms the mainstream CSC sequence annotation task into a rephrasing task, leveraging semantic understanding by extracting pre-trained knowledge from the pre-trained model to accomplish the CSC task.

4.3. Implementation Details

PMDRSpell is implemented on the foundation of Pytorch's transformers. We initialize the weights of the error correction model using bert-base-chinese from Hugging Face. The convolutional layers of wav2vec2.0 [21] serve as the phonological feature extractor, with chinese-wav2vec2-base used to initialize its weights. For extracting morphological features of Chinese characters, we employ a four-layer Transformer encoder, initialized with the first four layers of BERT. The model is trained for 10 epochs on the SIGHAN dataset using the AdamW optimizer with a learning rate of 1.0×10^{-4}. The batch size is set to 48 for both training and prediction processes, and the learning rate is linearly decayed during training. Meanwhile, all the Chinese characters are converted to audio using PaddleSpeech [20], and each audio is sliced into a Chinese character audio coding table at a sampling rate of 1600 Hz (1600 sampling rate can limit the length of audio coding of a single Chinese character to less than 1000, and it is easier to sense the similarity between the rhythms of the sounds at a smaller sampling rate), and then we splice the four-corner coding of each Chinese character into the five-stroke sequences with the length of 10 to build a table of Chinese character glyph codes. The error probability threshold η to 0.6, the number of Paca's layers N to 4, and the rest of the parameters are used as default parameters. All experiments were conducted on an NVIDIA A100 with 40G RAM.

4.4. Main Results

As shown in Table 3, compared to the best metrics in the baseline models, our proposed model leads the F1 score on the SIGHAN14 and SIGHAN15 datasets by 1.2 and 1.0 percentage points, respectively, thus proving the effectiveness of the present model.

Table 3. Comparison results of the models on sighan. We show the results for the accuracy (Acc), precision (Pre), recall (Rec), and F1 score (F1), where the optimal metrics are in **bold**. "#" denotes the model is additionally pretrained.

DataSet	Method	Detection Level				Correction Level			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Sighan13	PHMO #(2021) [3]	77.1	**99.5**	76.8	**86.7**	75.4	**99.5**	75.1	**85.6**
	LEAD(2022) [17]	**-**	88.3	**83.4**	85.8	-	87.2	82.4	84.7
	MDCSpell(2022) [4]	-	89.1	78.3	83.4	-	87.5	76.8	81.8
	DORM #(2023) [7]	**-**	87.9	83.7	85.8	-	86.8	**82.7**	84.7
	ReLM (2024) [18]	**-**	-	-	-	-	74.4	73.1	73.7
	PMDRSpell (ours)	**86.6**	93.6	76.7	84.3	**86**	93.5	75.4	83.5
Sighan14	PHMO #(2021) [3]	78.5	**85.3**	67.6	75.5	76.9	**84.7**	64.3	73.1
	LEAD(2022) [17]	**-**	70.7	71	70.8	-	69.3	69.6	69.5
	MDCSpell(2022) [4]	**-**	70.2	68.8	69.5	-	69	67.7	68.3
	DORM #(2023) [7]	**-**	69.5	**73.1**	71.2	-	68.4	**71.9**	70.1
	ReLM (2024) [18]	**-**	-	**-**	-	-	69.9	70.2	70
	PMDRSpell (ours)	**80.6**	80.4	72.3	**76.1**	**79.5**	79.8	69.5	**74.3**
Sighan15	Soft-Masked Bert(2020) [8]	80.9	73.7	73.2	73.5	77.4	66.7	66.2	66.4
	PHMO #(2021) [3]	82.6	**90.1**	72.7	80.5	80.9	**89.6**	69.2	78.1
	LEAD(2022) [17]	-	79.2	82.8	80.9	-	77.6	81.2	79.3
	MDCSpell(2022) [4]	-	80.8	80.6	80.7	-	78.4	78.2	78.3
	ECSpell #(2023) [5]	**86.3**	81.1	83	81	**85.6**	77.5	81.7	79.5
	DORM #(2023) [7]	**-**	77.9	**84.3**	81	-	76.6	**82.8**	79.6
	ReLM (2024) [18]	**-**	-	**-**	-	-	70	72.2	71.1
	PMDRSpell (ours)	82.9	**85**	79.1	**82**	81.6	84.7	76.3	**80.6**

Specifically, compared to the same type of multi-modal model PHMO [3], we are ahead in error correction accuracy by 10.6, 3.4, and 0.7 percentage points, and in recall by 0.3, 5.2, and 7.1 percentage points, respectively. Although this model employs an adaptive approach to fusing multi-modal knowledge, its fusion of entire sentences increases the impact on correct Chinese characters. Additionally, its linear fusion of multi-modal features does not effectively utilize the correlation between phonological and morphological features, resulting in overall effectiveness that is not as good as the model proposed in this paper. Moreover, while our model's overall framework is inspired by the design of the Soft-Masked BERT [8], both employing the architecture of error detection, soft masking, and error correction, our model outperforms it by fourteen points in the F1 metrics of SIGHAN15. This improvement is attributed to Soft-Masked BERT's limited utilization of multi-modal knowledge for error correction. Instead, its approach suppresses the model's reasoning by using the embedding information of incorrect Chinese characters within the residual structure.

4.5. Ablation Study

In this section, we perform sequential abatement of the innovative parts of this model and experimentally compare its measured effects on three public datasets. As shown in Table 4, when we ablate away each part of the model sequentially, its effect decreases to some extent. Specifically, when the coarse-grained and hierarchical error detection module is eliminated, the F1 Score of error detection and correction decreases by one percentage point on average, which indicates that the use of "non-multiple word errors" and hierarchical information to locate the word errors can improve the detection and correction effects of the model to a certain extent; when the dynamic residual module is ablated, the error correction F1 score is directly reduced by four percentage points, which also proves that our proposed dynamic residual mechanism can effectively prevent the model from being misled by the risk of misinformation; when the Paca module is ablated, the F1 Score is reduced by two percentage points; when the phonological features are ablated, the effect

is reduced by three percentage points on average, and when the morphological features are ablated, the effect is reduced by two percentage points on average. The different effects of the two ablations also reflect sideways that the proportion of phonological proximity errors in the dataset is higher than the proportion of morphological proximity errors, which reinforces Liu's work [30] that the proportion of phonological similarity errors is the largest in Chinese. Therefore, both the way of multi-modal information fusion and the fused information can be beneficial to the model and can effectively improve the model.

Table 4. Ablation results on three datasets. We conducted the ablation of the following modules in turn: replacing the coarse-grained and hierarchical detection method with a simple fine-grained method using bi-directional LSTM model (-Coarse and Hierarchy), replacing the dynamic residual mechanism with a static residual approach that incorporates multi-modal information for all Chinese characters in the incorrect sentence (-Dynamic Residual), removing the parallel cross-attention fusion module (-Paca), removing the phonological features (-Phonology), and removing the morphological features (-Morphology). Bold indicates the optimal result in the comparison data.

DataSet	Method	Detection Level				Correction Level			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Sighan13	PMDRSpell	**86.6**	93.6	**76.7**	**84.3**	**86**	93.5	**75.4**	**83.5**
	-Coarse & Hierarchy	85.7	93	75.2	83.2	84.9	92.8	73.5	82
	-Dynamic Residual	81.7	91.4	67.4	77.6	80.3	91	64.5	75.5
	-Paca	84.3	93	72.1	81.2	83.6	92.8	70.6	80.2
	-Phonology	83.7	91.9	71.7	80.57	83	91.7	70.1	79.5
	-Morphology	85.8	**93.9**	74.6	83.1	84.7	**93.8**	72.2	81.6
Sighan14	PMDRSpell	80.6	80.4	**72.3**	**76.1**	**79.5**	79.8	**69.5**	**74.3**
	-Coarse & Hierarchy	79	79.2	69	73.7	78.2	78.7	67.2	72.5
	-Dynamic Residual	79	79.1	69.3	73.9	77.9	78.4	66.6	72
	-Paca	**80.8**	**81.1**	69.5	74.9	78.7	**80.4**	66.4	72.8
	-Phonology	78.1	77.1	69.2	73	76.7	76.3	66.1	70.8
	-Morphology	79.1	78.9	69.9	74.1	77.7	78	66.6	71.9
Sighan15	PMDRSpell	**82.9**	85	79.1	**82**	**81.6**	**84.7**	**76.3**	**80.6**
	-Coarse & Hierarchy	82.6	84.1	**79.5**	81.7	80.9	83.5	75.9	79.5
	-Dynamic Residual	82.1	**85.2**	76.8	80.8	80	84.5	72.6	78.1
	-Paca	82.3	84.7	78	81.2	80	83.9	73.3	78.3
	-Phonology	81.6	83	78.5	80.7	79.3	82.1	73.9	77.8
	-Morphology	81.3	82.3	78.7	80.5	79	81.5	74.1	77.6

Overall, it can be concluded that our proposed model can effectively improve the effectiveness of the model. Whether it is the coarse-grained and hierarchical error detection module, the dynamic residual module, or the multi-modal information extraction and fusion approach that we propose can enhance the effectiveness of the model to a certain extent.

4.6. Effect of Hyper Parameters

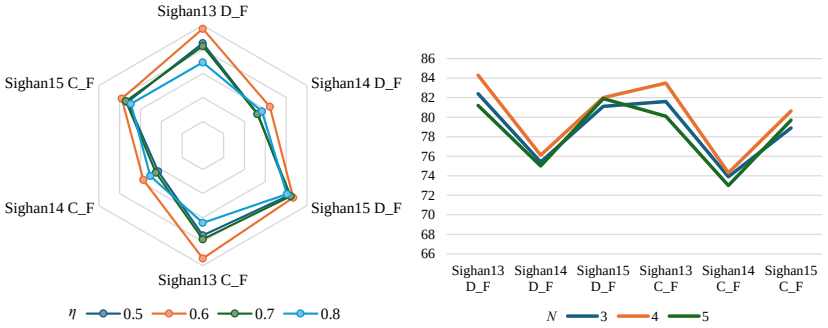
In this work, the error probability η and the number of layers N of the Paca module tend to affect the model's understanding and utilization of multi-modal information, and this section explores the impact of these hyperparameters on the model's effectiveness.

Theoretically, the value of η has bounds. A large η reduces the dynamic residual's role as its relative impact shrinks. A small η overly modifies word embeddings, disrupting semantic and syntactic info and hindering model inference. Experimentally, in Figure 7a, η = 0.6 gives the best performance on three public datasets, confirming the theory as this η balances residual and embedding preservation.

Theoretically, with layer number N, fewer layers mean less than ideal feature fusion and incomplete representation. More layers cause over-fusion, adding complexity and distortion that challenges model fitting. Experimentally, Figure 7b shows the best performance

when the Paca module's layer number N is 4, matching the theory as four layers balance fusion and over-fusion, optimizing overall performance.

In conclusion, whether it is the error probability η or the number of layers N of the Paca module, the experimental conclusions show that these parameters work better when they are of moderate size, with η of 0.6 and N of 4 being the best.



(**a**) (**b**)

Figure 7. The effect of different hyperparameters on the datasets sighan13, 14, 15. "D_F" and "C_F" denote the error detection F1 score and error correction F1 score, respectively. (**a**) Effect of different thresholds, (**b**) Effect of the number of Paca fusion layers.

5. Discussion

5.1. Language Adaptability and Extensibility

Chinese spelling errors are significantly different from spelling errors in other languages. In English, for example, Chinese spelling errors mainly involve the confusion of similar characters or homophones in Chinese characters (such as "木" and "本"), while English spelling errors usually involve errors in the letters inside words (such as "receive" misspelled as "recieve"). Since Chinese does not have explicit word boundaries, spelling errors directly affect semantic expression, and corrections need to be made by combining context and multi-modal information to infer the correct form. In contrast, English spelling errors can usually be corrected more easily through grammar and dictionary lookups.

The difficulty of Chinese spelling errors lies in the implicit word boundaries, the correlation between the glyphs of phonetic characters and audio modes, and the complex multi-modal knowledge and context matching need to be considered when correcting. For other languages, such as Japanese, there are also similar phonetic phenomena to Chinese. Therefore, this model not only performs well in Chinese spelling error proofreading but also provides a reference for other languages, especially languages with similar morphology (such as Japanese). It should be noted that the characteristics of spelling errors in each language may be different, so whether the model is suitable for other languages still needs to be properly adjusted and studied according to its language characteristics.

5.2. Time Complexity

In the analysis of time complexity, this paper mainly analyzes several core sub-models, namely ON-LSTM, Bert, and Paca. The models used in the feature extraction module are sub-layers of Bert and wav2vec2.0, and their time complexity is lower than that of the Bert model. Therefore, they will not be discussed separately here.

First, ON-LSTM incorporates the hierarchical structure information of statements in LSTM, and its time complexity is in the same order of magnitude as that of LSTM, which is $O(n * h^2)$, where n is the length of the sequence and h is the dimension of the hidden layer. The time complexity of BERT and Paca are in the same order of magnitude as $O(L * n^2 * d)$ and $O(N * n^2 * d)$, where L is the number of layers in BERT, N is the total number of layers of the cross-attention mechanism in Paca, n is the length of the input sequence, and d is the dimension of the model. In terms of magnitude, the time complexity of ON-LSTM is lower compared to Bert and Paca, so the total time complexity of PMDRSpell is of the same magnitude as Bert. At the same time, PMDRSpell does not require additional pre-training,

and the training cost is lower compared to similar models such as PHMO and DORM, which require additional pre-training.

Overall, although PMDRSpell introduced additional modules, the computational complexity did not exceed that of Bert in essence, so it did not bring a lot of additional overhead.

5.3. Practical Application

In practical use, this model can be embedded in text automation software to play the role of text spelling correction.

The most typical example is the use of OCR and ASR software. Both OCR (optical character recognition) and ASR (automatic speech recognition) workflows involve multiple steps, where OCR includes image preprocessing, character recognition, and post-processing, while ASR includes audio acquisition, signal processing, automatic speech recognition, and post-processing. In these two workflows, the text proofing model can be integrated between character, automatic speech recognition, and post-processing as a corrective link. For OCR, it can correct spelling errors caused by image quality or font problems after character recognition; for ASR, the model can automatically correct homophones and pronunciation errors after automatic speech recognition. By placing the model in these two links, the accuracy of the system can be effectively improved.

5.4. Case Study

To demonstrate the practical effectiveness of this model, we analyze specific cases.

As shown in Figure 8, “滚滚长虹东逝水(The rolling Yangtze River passes eastward.)” incorrectly writes “江 (river)” as “虹 (rainbow)”. The characters “虹” (hóng, 虫+工) and “江” (jiāng, 氵+工) belong to the same radical family “工 (work)” and exhibit strong similarities in syllables as well as morphological consistency. Our model can focus more on the correct “江 (river)” rather than “蜈 (centipede)” and “洪 (flood)”, which have similar pronunciation or morphology to “虹 (rainbow)”. However, When the Paca module was melted away, the model focused a lot of attention on the homophonic “洪 (flood)” rather than the “江 (river)” of the same vocal family, resulting in correction failure. Similarly, “他亲眼目睹了这件事的发生 (He witnessed it happen.)” misspells “睹 (see)” as “赌 (bet)”. Both “赌” (dǔ, 贝+者) and “睹” (dǔ, 目+者) share the same phonetic radical “者 (person)”. When the Paca module was melted away, it was difficult for the model to capture the information of the same phonetic family, resulting in too much attention being attracted by the Chinese character “读 (read)” with the same pronunciation, and it could not quickly converge to the correct answer “睹 (see)”. By effectively utilizing phonograms, we improve the accuracy of the model.

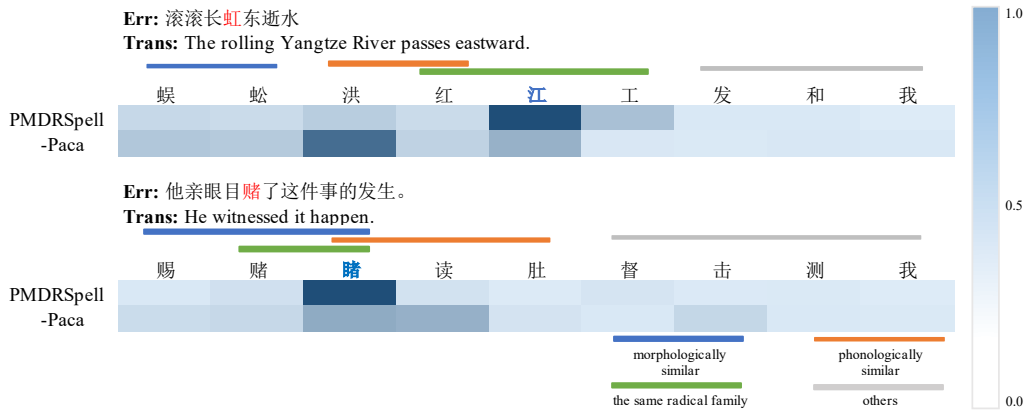


Figure 8. Heat map visualization of probability. The darker the blue, the higher the prediction probability of the corresponding Chinese character given by the model. “-Paca” means that there is no cross-fusion of multi-modal information and information mining of Chinese characters of the same phonetic family.

In addition, for other semantic class spelling errors, due to Bert's strong semantic understanding ability in the pre-trained MLM stage, the downstream error correction model can adaptively correct the semantic class errors, avoiding the problem of multi-modal information failure.

6. Conclusions

This work proposes a multi-modal dynamic residual mechanism, which leverages multi-modal information from errors to dynamically correct its residual network structure. This approach enhances the model's inference of erroneous Chinese characters and reduces the impact of multi-modal information on correct characters compared to mainstream multi-modal models. Additionally, we introduce a hierarchical detection method for "non-multiple word errors" that focuses on errors within entire words. We also designed a parallel cross-attention fusion mechanism, Paca, to explore the connections between glyph and phonological features of phonograms, which constitute the majority of Chinese characters. This mechanism improves the model's effectiveness by emphasizing the correlation between different modalities of phonograms, facilitating further in-depth research on the Chinese Spell Correction (CSC) task.

In future work, we will explore integrating field-specific external knowledge bases into our model to enhance the model's error correction capabilities in specialized areas, thereby improving its practical utility.

Author Contributions: Conceptualization, G.C. and Y.Z.; methodology, J.S.; software, G.C. and Y.Y.; validation, G.C. and J.S.; formal analysis, G.C. and Y.Y.; investigation, J.S.; resources, Y.Z.; data curation, J.S.; writing—original draft preparation, G.C.; writing—review and editing, Y.Z. and Y.Y.; visualization, J.S.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Beijing Natural Science Foundation (grant number L233008).

Data Availability Statement: The Sighan dataset (sighan13; sighan14; sighan15) used in this paper is available from the following website: http://ir.itc.ntnu.edu.tw/lre/sighan7csc.html (accessed on 14 October 2013); http://ir.itc.ntnu.edu.tw/lre/clp14csc.html (accessed on 20 October 2014); http://ir.itc.ntnu.edu.tw/lre/sighan8csc.html (accessed on 30 July 2015). The Wang271K is available from the following website: https://github.com/wdimmy/Automatic-Corpus-Generation/blob/master/corpus/train.sgml (accessed on 30 May 2019).

Conflicts of Interest: The authors G.C., Y.Z. and J.S. are affiliated with Beijing Information Science and Technology University, while the author Y.Y. is an employee of the China Electronics Cloud Computing Technology Cooperation. The participation of Y.Y. in this research was solely for academic collaboration. This work was fully funded by Beijing Information Science and Technology University. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CSC	Chinese Spelling Check
OCR	Optical Character Recognition
ASR	Automatic Speech Recognition
IR	Information Retrieval
Paca	Parallel Cross-attention Fusion

References

1. 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, Wuhan, China, 20–21 October 2014 ; pp. 126–132.
2. Xu, H.D.; Li, Z.; Zhou, Q.; Li, C.; Wang, Z.; Cao, Y.; Huang, H.; Mao, X.L. Read, Listen, and See: Leveraging Multimodal Information Helps Chinese Spell Checking. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 716–728.

3. Huang, L.; Li, J.; Jiang, W.; Zhang, Z.; Chen, M.; Wang, S.; Xiao, J. PHMOSpell: Phonological and morphological knowledge guided Chinese spelling check. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 5958–5967.
4. Zhu, C.; Ying, Z.; Zhang, B.; Mao, F. MDCSpell: A multi-task detector-corrector framework for Chinese spelling correction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 1244–1253.
5. Lv, Q.; Cao, Z.; Geng, L.; Ai, C.; Yan, X.; Fu, G. General and domain-adaptive chinese spelling check with error-consistent pretraining. *ACM Trans. Asian-Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–18. [CrossRef]
6. Song, S.; Lv, Q.; Geng, L.; Cao, Z.; Fu, G. RSpell: Retrieval-augmented Framework for Domain Adaptive Chinese Spelling Check. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Foshan, China, 12–15 October 2023; Springer: Berlin/Heidelberg, Germany; pp. 551–562.
7. Liang, Z.; Quan, X.; Wang, Q. Disentangled phonetic representation for chinese spelling correction. *arXiv* **2023**, arXiv:2305.14783.
8. Zhang, S.; Huang, H.; Liu, J.; Li, H. Spelling Error Correction with Soft-Masked BERT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 882–890.
9. Zhang, Y.; Cao, Y.; Yu, S. A hybrid model of combining rule-based and statistics-based approaches for automatic detecting errors in Chinese text. *J. Chin. Inf. Process.* **2006**, *20*, 1–7+55.
10. Zhang, S.; Xiong, J.; Hou, J.; Zhang, Q.; Cheng, X. HANSpeller++: A unified framework for Chinese spelling correction. In Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, Beijing, China, 30–31 July 2015; pp. 38–45.
11. Wang, Y.R.; Liao, Y.F. Word vector/conditional random field-based Chinese spelling error detection for SIGHAN-2015 evaluation. In Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, Beijing, China, 30–31 July 2015; pp. 46–49.
12. Liu, M.; Jian, P.; Huang, H.Y. Introduction to BIT Chinese spelling correction system at CLP 2014 bake-off. In Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, Wuhan, China, 20–21 October 2014; pp. 179–185.
13. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
14. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
16. Wang, B.; Che, W.; Wu, D.; Wang, S.; Hu, G.; Liu, T. Dynamic connected networks for Chinese spelling check. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, 1–6 August 2021; pp. 2437–2446.
17. Li, Y.; Ma, S.; Zhou, Q.; Li, Z.; Yangning, L.; Huang, S.; Liu, R.; Li, C.; Cao, Y.; Zheng, H. Learning from the Dictionary: Heterogeneous Knowledge Guided Fine-tuning for Chinese Spell Checking. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 238–249.
18. Liu, L.; Wu, H.; Zhao, H. Chinese Spelling Correction as Rephrasing Language Model. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 18662–18670.
19. Shen, Y.; Tan, S.; Sordoni, A.; Courville, A.C. Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
20. Zhang, H.; Yuan, T.; Chen, J.; Li, X.; Zheng, R.; Huang, Y.; Chen, X.; Gong, E.; Chen, Z.; Hu, X.; et al. PaddleSpeech: An Easy-to-Use All-in-One Speech Toolkit. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, Hybrid, Seattle, WA, USA/Online, 10–15 July 2022; pp. 114–123.
21. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
22. Zhou, Y.; Sun, Z.; Wu, X.; Yu, K. Chinese Spelling Correction Model Based on Gated Feature Fusion. *J. Beijing Univ. Posts Telecommun.* **2023**, *46*, 91.
23. Wang, M.; Fang, Y.; Chen, W. E-Commerce Event Detection with Chinese Character Glyph Based ELMo. *J. Chin. Inf. Process.* **2021**, *35*, 94–102.
24. Jin, S.; Xin, C. Consistency of Phonetic Radical Families in Phonograms and the Teaching of Chinese as a Foreign Language. *J. Chin. Linguist.* **2015** , 19–30+95.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
26. Wu, S.H.; Liu, C.L.; Lee, L.H. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, Nagoya, Japan, 14–18 October 2013; pp. 35–42.
27. 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, Beijing, China, 30–31 July 2015; pp. 32–37.

28. 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, Brussels, Belgium, 31 October–4 November 2018; pp. 2517–2527.
29. Cheng, X.; Xu, W.; Chen, K.; Jiang, S.; Wang, F.; Wang, T.; Chu, W.; Qi, Y. SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 871–881.
30. Liu, C.L.; Lai, M.H.; Tien, K.W.; Chuang, Y.H.; Wu, S.H.; Lee, C.Y. Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications. *ACM Trans. Asian Lang. Inf. Process. (TALIP)* **2011**, *10*, 1–39. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.