Combining Image and Text for Chinese Spelling Correction

Shitong Qin
stqin.research@outlook.com

Leqi Sha
leqisha@gmail.com

Jia Li
lijia@gmail.com

Abstract

Correcting spelling mistakes is a complex task that presents significant challenges in obtaining satisfactory solutions. In this study, we focus on Chinese spelling error correction (CSC). The state-of-the-art method utilizes the BERT architecture to detect and correct errors in a single sentence, and the joint training of error detection and error correction effectively reduces cascading errors. However, deploying such models in real scenarios often encounters mis- and omissions caused by OCR errors. To address this issue, we propose an innovative approach to combine image and text information, exploiting visual cues to improve the performance of the entire error correction pipeline in scenarios where OCR is inaccurate. Our evaluation demonstrates the potential of this approach in real-world scenarios to address both miscorrections and missed corrections, leading to more precise and comprehensive error detection and correction.

1 Introduction

Spelling check, a fundamental task in all written languages, involves detecting and correcting spelling errors made by humans. It aims to identify and rectify spelling errors in text, whether at the word-level or character-level (Yu and Li (2014); Yu et al. (2014); Xiong et al. (2015)). Its importance extends to various natural language applications, including search (Martins and Silva (2004)), optical character recognition (OCR) (Wang et al. (2018)), and essay scoring (Burstein and Chodorow (1999)). In this study, our focus lies in character-level Chinese spelling error correction (CSC). Presents unique and formidable challenges due to the distinctive characteristics of the Chinese language (Zhang et al. (2012),Zhang and Zhao (2011)). These characteristics make CSC significantly different and more intricate compared to English or other alphabetical languages, demanding specialized approaches and techniques.

Current models for Chinese spelling correction encompass various architectures, including Seq2Seq(Wu and Wu (2022)), Graph-based(Cheng et al. (2020)), and Transformer-Encoder models(Wang et al. (2020)). The Seq2Seq approach requires the construction of a source and target, where the source contains text with potential spelling errors and the target consists of text without errors. This architecture offers the advantage of covering all correction scenarios, such as insertion, deletion, and substitution. On the other hand, the native Transformer-Encoder structure has limitations in terms of input and output text length, requiring them to be equal. As a result, it can only handle deletion and substitution cases, with deletion being represented using special tokens. In contrast, graph-based models like SpellGCN(Cheng et al. (2020)) tackle the problem differently. By constructing a graph over the characters and training SpellGCN to map this graph into interdependent character classifiers, it introduces a novel approach. In comparison to other models, this proposed method demonstrates significant performance improvements.

Chinese spelling checks play a vital role in ensuring the accuracy and readability of written texts. While existing Chinese spelling correction models achieve impressive performance, their practical application is often hampered by the performance of optical character recognition (OCR) systems (Shi et al. (2016)). The integration of an OCR system with spell correction is critical for real-world scenarios, as OCR introduces recognition errors, which can adversely affect the overall performance of the correction system.

Limitations of Chinese spelling correction models in real-world scenarios:

- Traditional Chinese spelling correction models mainly focus on text input and assume that OCR recognition is error-free. However, OCR systems often make mistakes, and OCR systems may mistakenly identify modified char-

acters as similar-looking characters, causing misspellings to propagate through the error correction pipeline. Therefore, mitigating the impact of OCR recognition errors is crucial to improving the overall performance of error correction pipelines.

- In the case of smeared characters, the visual features from images tend to be much more prominent than the textual features. However, previous models only considered the text modality and overlooked the image modality, resulting in the loss of valuable information.

- Optimizing OCR systems can be a costly endeavor, requiring a significant amount of annotated data for training. Conversely, the annotation data for spelling correction models can be obtained through data augmentation techniques, significantly reducing the associated cost. This advantage allows for a more scalable and cost-effective approach to obtaining annotated samples for training spelling correction models compared to the laborious and resource-intensive process of improving OCR systems.

In this paper, we comprehensively analyze the limitations of existing Chinese spelling correction models when combined with OCR systems. We emphasize the importance of considering OCR-induced errors and the potential benefits of increasing the input of images. Furthermore, we describe a method for the ensemble of novel image-text fusion models for Chinese spelling correction. Through extensive experiments on large-scale datasets, we demonstrate the effectiveness of our method compared to traditional spelling correction models. Our research contributes to advancing the field of Chinese spelling correction by addressing the challenges posed by OCR errors and effectively utilizing textual and visual information.

The rest of the paper is organized as follows: Section 2 provides an overview of related work on Chinese spelling correction and OCR integration. Section 3 presents our proposed method, detailing the integration between an OCR system and an image-to-text fusion model. Section 4 presents the experimental results and analysis. Finally, in Section 5, we conclude the paper by summarizing our contributions and emphasizing the importance of the proposed method.

2 Related Work

In recent years, the field of Chinese spelling correction has witnessed significant advancements. Several studies have been conducted to address the challenges and improve the performance of spelling correction models. In this section, we provide an overview of the relevant literature in Chinese spelling correction, focusing on key contributions and approaches. (Sun et al., 2023) proposed an error-guided correction model that leverages BERT's power, introducing a zero-shot error detection method to focus on potential errors and avoid unnecessary modifications. (Huang et al., 2014) demonstrated the effectiveness of a tri-gram language modeling approach with dynamic programming and additive smoothing for Chinese spelling check. (Li et al., 2021) improved CSC by identifying weak spots, generating valuable training instances, and applying a task-specific pre-training strategy, achieving state-of-the-art performance. (Wang et al., 2023) addressed limitations in previous CSC methods with an improved non-autoregressive spelling correction model, achieving significant gains in name recall and stable improvement across different bias list name coverage ratios.

As far as we know, there is no model that specifically combines image information and text information for error correction, but in other fields, such as text classification, multimodality has been widely used. Miller et al. (2020) proposed a novel approach that integrates natural language understanding and image features with associated metadata to improve image classification accuracy, achieving a 1.56 reduction in the Top 5 error rate compared to benchmark results. Gallo et al. (2020) explore multimodal classification using textual information and visual representations, applying novel fusion techniques and stacking methods to improve performance on the challenging UPMC Food-101 dataset, surpassing state-of-the-art results.

These studies represent a range of approaches in Chinese spelling correction, highlighting the utilization of rule-based methods, statistical language models, deep learning models, hybrid approaches, attention mechanisms, transformer models, contextual information, and reinforcement learning techniques. While these prior works have contributed significantly to the field, the integration of OCR systems and the incorporation of image-based inputs remain relatively unexplored areas, which we address in this paper.

3 Our Approach

3.1 Problem Formulation

The problem at hand is Chinese spelling correction, which aims to identify and rectify errors present in Chinese text. Given a text sequence $X = \{x_1, x_2, ..., x_n\}$ consisting of n characters, our objective is to generate a target character sequence $Y = \{y_1, y_2, ..., y_n\}$. This problem can be framed as a conditional generation task, where the goal is to model and maximize the conditional probability $p(Y|X)$.

3.2 Model

We propose a novel neural network architecture to address the challenges of erroneous corrections and omissions caused by OCR errors in real-world text correction applications. The overall structure of our model is similar to the Soft-Mask Bert model(Zhang et al. (2020)), as illustrated in Figure 1. Through analyzing examples of erroneous corrections and omissions resulting from OCR errors, we have observed that these errors commonly occur on characters with smudges or scratches. OCR systems often misinterpret these characters as visually similar alternatives, as depicted in Figure 2.

Previous spelling correction models solely relied on the recognized text outputted by OCR systems, assuming the correctness of the recognized characters. These models predominantly relied on contextual semantics to identify potential misspelled characters. By incorporating the images containing the OCR-recognized characters as input to the model, we can effectively leverage visual information to discern whether a character requires correction. The presence of smudges or scratches in the images, which are typically associated with OCR-recognized errors, facilitates the differentiation between characters that need correction and those that do not.

In Figure 2, the character "口" with a deletion mark was mistakenly recognized as the character "日" by OCR. When inputted into the text correction model, this "日" character may either be deleted or replaced with another character. However, the correct action should be to delete this character. If only the text information containing the "日" character is available, the correction model can only rely on the context of the "日" character to determine whether to delete it, which increases the difficulty of model learning.

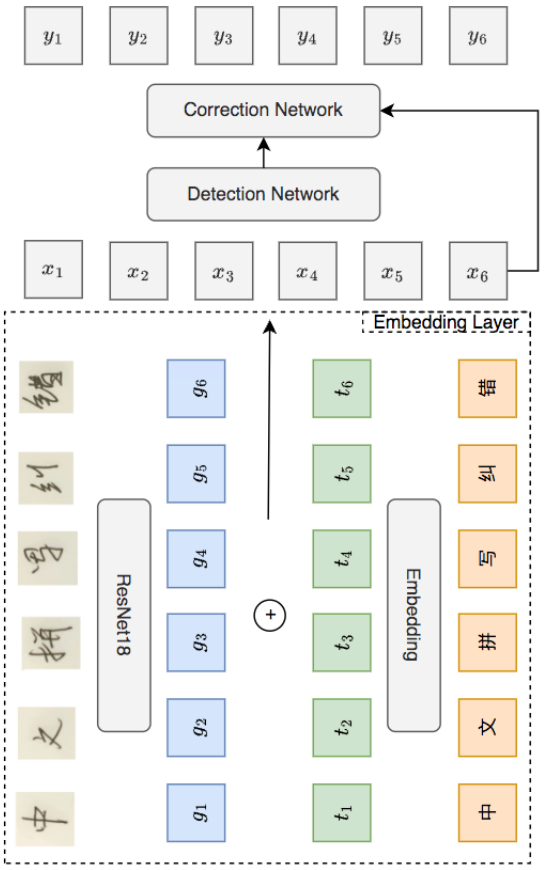

Figure 1: Architecture of Model

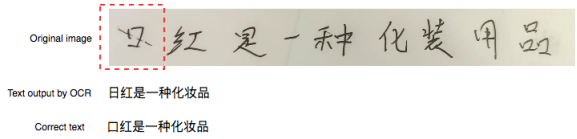

Figure 2: Segmenting a text line into regions containing individual characters using OCR.

Compared to ensuring that the OCR model correctly recognizes each individual character, accurately segmenting the image regions of each character is relatively simpler. We tested multiple existing OCR [1] systems, and they were able to effectively segment the image regions containing the text. This provides a guarantee for constructing our model. The result of using an OCR system to segment a line of text is shown in the figure 3.

After obtaining the text and its corresponding images, we encode the text using an embedding matrix, while the images are encoded using the ResNet18 architecture (He et al. (2016)) . Subsequently, we add the encoded representations of the text and images together, serving as the input for

[1] https://ai.youdao.com/product-ocr-print.s

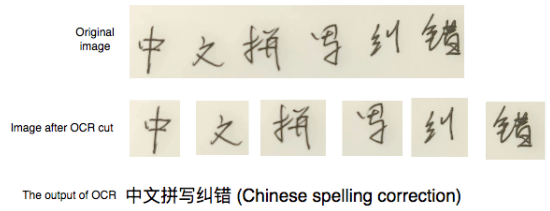

Figure 3: Segmenting a text line into regions containing individual characters using OCR.

the subsequent modules of our model.

By combining the contextual information from the text and the visual cues from the images, our proposed method aims to improve the accuracy and robustness of Chinese spelling correction, particularly in real-world scenarios affected by OCR errors.

3.3 Embedding

We describe the specific components and processes within our proposed method for combining image and text in Chinese spelling correction.

- Image Encoding, The images containing the OCR-recognized characters are passed through the ResNet18 architecture for image encoding. ResNet18 is a deep convolutional neural network that extracts high-level features from images. It enables the model to capture visual patterns and characteristics associated with character recognition errors, such as smudges or scratches. The output of the ResNet18 network provides a compact and meaningful representation of the visual information present in the images.

$$g_i = ResNet18(Image_i), i \in \{1, 2, ..., n\} \quad (1)$$

- Text Encoding, The text is encoded using an embedding matrix. The embedding matrix maps each character in the text to a dense vector representation. This vector representation captures the semantic and syntactic information of the characters. By encoding the text, we preserve the contextual information necessary for accurate spelling correction.

$$t_i = Embed(Token_i), i \in \{1, 2, ..., n\} \quad (2)$$

- Fusion of Image and Text, To fuse the encoded image and text representations, we employ an element-wise addition operation. This operation combines the visual and textual features, allowing the model to leverage both sources of information for improved spelling correction. The fused representation provides a comprehensive input that captures both the semantic context of the text and the visual cues from the corresponding images.

$$x_i = t_i + g_i, i \in \{1, 2, ..., n\} \quad (3)$$

By integrating image and text information, our method enhances the performance of Chinese spelling correction systems, especially in scenarios affected by OCR errors. The fusion of visual and textual features enables the model to make more informed corrections and address OCR-induced errors effectively.

3.4 Detection & Correction Network

Unlike Soft-Masked BERT, in the error detection network, we use a 3-layer Transformer-Encoder Layer instead of GRU. Using the Transformer-Encoder Layer allows for better extraction of contextual information.

$$\begin{gathered} Output_{multihead} = MultiHead([x_1, x_2, ...x_n]) \\ Output_{ffn} = FFN(Output_{multihead}) \\ P_{Detection} = Softmax(Output_{ffn}) \end{gathered} \quad (4)$$

Where $\vec{x} = [x_1, x_2, ..., x_n]$ represents the vector after merging the image and text information, $P_{Detection}$ corresponds to the probability of whether there is an error at each position in the sentence.

The correction network of this model is similar to the correction network of Soft-Masked BERT, with slight differences in the input. To enhance the original information from both the image and text, the input of the correction network consists of three parts: a weighted sum considering error detection probabilities, text embedding vectors, and image representation vectors corresponding to each token.

$$\begin{gathered} \vec{x}_{correct} = P_{Detection} * e_{mask} + \\ (1 - P_{Detection}) * \vec{x} \\ \vec{x}_{correct_{new}} = \vec{g} + \vec{x}_{correct} + \vec{t} \\ Output_{multihead} = MultiHead(\vec{x}_{correct_{new}}) \\ Output_{ffn} = FFN(Output_{multihead}) \end{gathered} \quad (5)$$

Where e_{mask} is the vector value corresponding to the special token, [MASK], in the embedding matrix, and $\vec{x}_{correct_{new}}$ is the input of the correction network.

4 Experimental Results

4.1 Datasets

Spelling correction is a unique task in natural language processing (NLP) that offers the advantage of being able to construct training sets through error generation, resulting in an infinite source of parallel corpora. In order to simulate real-world scenarios, we deliberately introduced various types of errors based on the following categories:

- Correct text recognition + altered text image: We applied deliberate modifications such as smudges, alterations, or additions to the correctly recognized text in order to simulate the presence of such errors in real-life situations.
- Correct text recognition + slight scratches on the text image: By introducing subtle scratches or imperfections to the handwritten text, we aimed to replicate the challenges faced in scenarios where the quality of the input image is compromised.
- Correct text recognition + tilted text image: We purposely rotated the text at different angles to mimic the distortion that can occur when capturing handwritten text from various orientations or perspectives.
- Correct text recognition + missing strokes in the text image: To simulate cases where certain strokes are missing in the handwritten text, we deliberately omitted specific parts of the characters, making the recognition task more challenging.
- Incorrect text recognition (similar characters) + correct text image: We intentionally introduced errors by replacing visually similar characters, aiming to simulate the common mistakes made due to the resemblance between certain characters.
- Incorrect text recognition (similar characters) + altered text image: Similar to the previous category, we combined the use of visually similar characters with deliberate modifications to the handwritten text images, further increasing the difficulty of the recognition task.
- Incorrect text recognition (similar characters) + slight scratches on the text image: By introducing slight scratches to the visually similar characters, we aimed to replicate situations where the recognition accuracy is affected by minor imperfections.
- Incorrect text recognition (similar characters) + tilted text image: We applied angular distortion to the visually similar characters, simulating scenarios where recognition errors occur due to variations in writing styles or angles.
- Incorrect text recognition (similar characters) + missing strokes in the text image: This category involved both visually similar characters and deliberate omission of specific strokes, creating a more challenging recognition task.
- Incorrect text recognition (randomly chosen word from the dictionary) + correct text image: To represent random recognition errors, we randomly replaced characters in the correctly written text with other characters randomly selected from the dictionary.
- Incorrect text recognition (randomly chosen word from the dictionary) + altered text image: Similar to the previous category, we combined randomly selected characters with deliberate modifications to the handwritten text images.
- Incorrect text recognition (randomly chosen word from the dictionary) + slight scratches on the text image: By introducing slight scratches to the randomly selected characters, we aimed to simulate recognition errors caused by minor imperfections in the input images.
- Incorrect text recognition (randomly chosen word from the dictionary) + tilted text image: We applied angular distortion to the randomly selected characters, simulating situations where recognition errors arise due to variations in writing styles or angles.
- Incorrect text recognition (randomly chosen word from the dictionary) + missing strokes in the text image: This category combined randomly selected characters with deliberate omission of specific strokes, increasing the complexity of the recognition task.

By generating these diverse types of errors, we were able to construct a varied training set that closely resembled real-world spelling errors. This approach not only enhances the performance and

robustness of the spelling correction model but also allows for a more comprehensive evaluation.

Using the above data construction methods, we construct 3300W parallel corpus using primary and secondary school Chinese as the original data set, Chinese textbooks and novels.

4.2 Baselines

To demonstrate the advantages of our model in real correction scenarios, several baseline models were selected for comparison. Soft-Masked Bert (Zhang et al. (2020)) proposes a novel neural architecture, which consists of a network for error detection and a network for error correction based on BERT. Experimental results show that our method outperforms baselines, including the one solely based on BERT, in Chinese spelling error correction. Chunk-based CSC (Bao et al. (2020)) extends confusion sets, proposes a chunk-based framework, and adopts a global optimization strategy. Experimental results demonstrate the proposed approach's superior performance on benchmark datasets and an optical character recognition dataset.

4.3 Experiment Setting

For model training, we employed the AdamW optimizer (Loshchilov and Hutter (2017)) with a batch size of 32 and a learning rate of 5e-5, conducting the optimization process for 6 epochs. To ensure robustness, we performed the experiments for 4 runs and reported the averaged metric.

4.4 Main Results

Table 1 presents the evaluation results of three models on the test set in terms of accuracy (Acc.), precision (Pre.), recall (Rec.), and F0.5 score. Among the models, Soft-Masked Bert achieved an accuracy of 80.1, precision of 60.2, recall of 61.3, and an F0.5 score of 60.42. The Chunk-based CSC model attained slightly lower performance with an accuracy of 72.1, precision of 52.9, recall of 53.4, and an F0.5 score of 53.0.In comparison, our proposed model demonstrated superior performance across all metrics. With an accuracy of 82.9, precision of 63.2, recall of 62.7, and an F0.5 score of 63.1, our model outperformed both Soft-Masked Bert and Chunk-based CSC.The higher accuracy of our model suggests its ability to correctly detect errors in the text, while the higher precision indicates a lower rate of false positives. Moreover, the higher recall value implies that our model can effectively capture a larger proportion of actual errors. The higher F0.5 score, which balances precision and recall with a higher weight on precision, further confirms the overall superiority of our model.These results highlight the effectiveness of our proposed approach in error detection. The improvements in precision, recall, and F0.5 score demonstrate the enhanced performance of our model compared to the existing approaches, making it a promising solution for spelling error detection tasks.

Models	**Acc.**	**Pre.**	**Rec.**	**F0.5**
Soft-Masked Bert	80.1	60.2	61.3	60.42
Chunk-based CSC	72.1	52.9	53.4	53.0
Ours	**82.9**	**63.2**	**62.7**	**63.1**

Table 1: The accuracy, precision, recall, and F1 score for error detection on the test set.

Models	**Acc.**	**Pre.**	**Rec.**	**F0.5**
Soft-Masked Bert	77.9	76.2	71.3	75.17
Chunk-based CSC	70.2	69.3	64.5	68.28
Ours	**81.2**	**80.9**	**76.7**	**80.02**

Table 2: The accuracy, precision, recall, and F1 score for error correction on the test set.

Table 2 displays the evaluation results of three models on the test set in terms of accuracy (Acc.), precision (Pre.), recall (Rec.), and F0.5 score for error correction. our proposed model, denoted as "Ours," demonstrated superior performance across all metrics. With an accuracy of 81.2, precision of 80.9, recall of 76.7, and an F0.5 score of 80.02, our model outperformed both Soft-Masked Bert and Chunk-based CSC in terms of error correction. By combining image and text information, our model benefits from a more comprehensive understanding of the context and visual cues, enabling it to make more accurate corrections. The higher accuracy, precision, recall, and F0.5 score of our model indicate its effectiveness in error correction tasks.

At the same time, we also discussed the influence of different types of data enhancement methods on the indicators on the evaluation set. As can be seen from the table,

5 Conclusion

We developed a spelling error correction model that leverages image and text information, trained on diverse error scenarios. Our model outperformed existing approaches in accuracy, precision, recall, and F0.5 score on the test set. This highlights the significance of incorporating visual cues and con-

sidering real-life error scenarios in spelling error correction. Our research contributes to advancing automated error correction technologies with potential applications across domains.

References

Zuyi Bao, Chen Li, and Rui Wang. 2020. Chunk-based chinese spelling check with global optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2031–2040.

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative english speakers. In *Computer mediated language assessment and evaluation in natural language processing*.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. Spellgcn: Incorporating phonological and visual similarities into language models for chinese spelling check. *arXiv preprint arXiv:2004.14166*.

Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. 2020. Image and text fusion for upmc food-101 using bert and cnns. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Qiang Huang, Peijie Huang, Xinrui Zhang, Weijian Xie, Kaiduo Hong, Bingzhou Chen, and Lei Huang. 2014. Chinese spelling check system based on tri-gram model. In *Proceedings of the third CIPS-SIGHAN joint conference on Chinese language processing*, pages 173–178.

Chong Li, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2021. Exploration and exploitation: Two ways to improve chinese spelling correction models. *arXiv preprint arXiv:2105.14813*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Bruno Martins and Mário J Silva. 2004. Spelling correction for search engine queries. In *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004. Proceedings 4*, pages 372–383. Springer.

Stuart J Miller, Justin Howard, Paul Adams, Mel Schwan, and Robert Slater. 2020. Multi-modal classification using images and text. *SMU Data Science Review*, 3(3):6.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Rui Sun, Xiuyu Wu, and Yunfang Wu. 2023. An error-guided correction model for chinese spelling error correction. *arXiv preprint arXiv:2301.06323*.

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

Hongfei Wang, Michiki Kurosawa, Satoru Katsumata, and Mamoru Komachi. 2020. Chinese grammatical correction using bert-based pre-trained model. *arXiv preprint arXiv:2011.02093*.

Xiaoqiang Wang, Yanqing Liu, Jinyu Li, and Sheng Zhao. 2023. Improving contextual spelling correction by external acoustics attention and semantic aware data augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xiuyu Wu and Yunfang Wu. 2022. From spelling to grammar: A new framework for chinese grammatical error correction. *arXiv preprint arXiv:2211.01625*.

Jinhua Xiong, Qiao Zhang, Shuiyuan Zhang, Jianpeng Hou, and Xueqi Cheng. 2015. Hanspeller: a unified framework for chinese spelling correction. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 1, June 2015-Special Issue on Chinese as a Foreign Language*.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 220–223.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of sighan 2014 bake-off for chinese spelling check. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 126–132.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked bert. *arXiv preprint arXiv:2005.07421*.

Xiaotian Zhang, Chunyang Wu, and Hai Zhao. 2012. Chinese coreference resolution via ordered filtering. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 95–99.

Xiaotian Zhang and Hai Zhao. 2011. Unsupervised chinese phrase parsing based on tree pattern mining. *Proc. of the 11th Confernece of China Computational Linguistics*.