Development of Text Error Correction Techniques in Chinese Corpus

Tongming Liu
School of Math, University of Leeds, Leeds, LS2 9JT, U.K.

Keywords: Chinese Text Error Correction, Chinese Spelling Correction, Chinese Grammar Error Correction.

Abstract: As the number of Chinese learners around the world continues to increase, the demand for high-quality copywriting continues to increase. High-accuracy and efficient Chinese text error correction models have increasingly important research significance. Chinese text error correction technology that has emerged in recent years can be divided into two directions: Chinese spelling correction and Chinese grammar error correction. These latest methods and models include PLOME, PHMOSpell, MaskGEC, etc., which exceed previous models in performance. At the same time, there are still certain issues in the field of Chinese text error correction that need to be addressed, like generalization abilities and over-correction. This article aims to present a comprehensive overview of recent advancements in Chinese text error correction methods and models. It will address the current challenges in this research field, providing valuable insights for scholars interested in Chinese text error correction technology. By shedding light on the development status and key issues, this article seeks to facilitate the advancement of this domain.

1 INTRODUCTION

With China's rapid economic growth and rising comprehensive national power, more and more foreigners are interested in Chinese culture and have begun to learn Chinese. By the end of 2020, Chinese language has been taught in more than 180 countries, and Chinese has been incorporated into the national education system in more than 70 countries. Now more than 20 million people in the world are learning Chinese, and the Chinese language has been used by nearly 200 million people. Teaching Chinese has developed into a well-established international education sector, particularly in nations bordering China. With the use of intelligent Chinese text error correction software, Chinese teachers and students can learn the language more quickly and easily, increase the effectiveness of their learning, and have less work to do.

In China, with the growing demand for spiritual pursuits, there is a need for a large number of high-quality text creations in China. It is frequently required to proofread manuscripts multiple times in order to guarantee their quality. The most fundamental of them is to review the text for errors and omissions and to promptly fix any that you find. However, traditional manual proofreading methods are time-consuming and demanding for proofreaders, and the cost of manual proofreading is increasing. In addition, sometimes the proofreader may not have enough time to finish the paper, which will make proofreading more difficult and higher error rate. It gets harder to guarantee the quality of proofreading as the number of proofreads rises along with the number of errors that happen during proofreading. If there are errors or missing words in the body of the text, it will adversely affect corporate promotions, business transactions, marketing copy, etc., which will in turn reduce the credibility of the article. To address this issue, machine automation can be used to evaluate the text, identify any errors or omissions, and provide recommendations for proofreaders. This will increase the effectiveness of the proofreading process overall and guarantee the article's quality.

There are many uses for Chinese text error correction. For instance, in the legal sector, where it is applied to written materials like interrogation transcripts, judgements, and other written materials, it can significantly increase the public prosecutor's office's productivity and uphold the supremacy of the law; media industry, the use of text error correction technology to complete the proofreading of news, subtitles, and other types of text, to avoid spelling, grammatical, punctuation and other errors, but also to improve the credibility of the media.

Liu, T.
Development of Text Error Correction Techniques in Chinese Corpus.
DOI: 10.5220/0012842300004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 364-370
ISBN: 978-989-758-690-3

This article will introduce several Chinese text error correction methods and corresponding models in recent years, and discuss the current problems in this field, aiming to help relevant scholars interested in Chinese text error correction technology quickly understand the development status of this field. and issues to promote the development of this field.

2 RESEARCH STATUS OF CHINESE TEXT ERROR CORRECTION TECHNOLOGY

Chinese spelling correction (CSC) at the character level and Chinese grammatical error correction (CGEC) at the phrase level make up the two main focuses of current research on Chinese Text Error Correction. An outline of pertinent research in these two directions will be given in this article. In addition, some scholars focus on the quite common semantic errors in Chinese and propose the Chinese Semantic Error Diagnosis (CSED) method, which will also be introduced in this article.

2.1 CSC

In natural language processing, CSC is a crucial activity that seeks to identify and fix possible spelling mistakes in Chinese text.

Traditionally, CSC tasks are approached as sequence labeling tasks, in which a model is trained to predict the correctness of each character in given sentences. Most of the more mainstream methods in recent years use fine-tuning sentence pairs and leveraging pre-trained mask language models (such as BERT, a bidirectional encoder representation from Transformers) to learn contextual representations of Chinese characters and words (Cheng et al, 2020, Wang et al, 2021, Li et al, 2021). However, these BERT-based models often use the fixed token "[MASK]" to represent misspelled characters 0, which means the model may not handle multiple consecutive errors well. Because BERT predicts the correct character at the [MASK] position based on the context of a single character or word, and continuous errors may destroy contextual information and affect error correction.

To overcome this limitation, a new method called Pre-training of Misspelling Knowledge for Chinese Spelling Correction (PLOME) is proposed 0. This method uses similar characters to mask selected tags based on the confusion set, rather than using fixed tags such as "[MASK]". This enables the model to capture misspelled knowledge more efficiently. PLOME significantly improves its performance in CSC tasks by introducing pronunciation prediction, which teaches the phonetic understanding of spelling errors. Moreover, the phonetic and visual similarity knowledge that is crucial to CSC is integrated into PLOME by using the Gate Recurrent Unit (GRU) network, which models the similarity between arbitrary characters based on the phonetic and stroke information of the characters, which not only improves the ability of the model to detect spelling errors (Figure 1). Similarly, in order to integrate information from both speech and visual modalities in CSC tasks, another end-to-end trainable model named PHMOSpell is proposed (Figure 2). The model extracts pinyin and glyph representations of Chinese characters from auditory and visual forms, respectively, by combining verbal and visual information. It then uses a well-designed adaptive gating mechanism to integrate these representations into a language model that has already been trained 0.

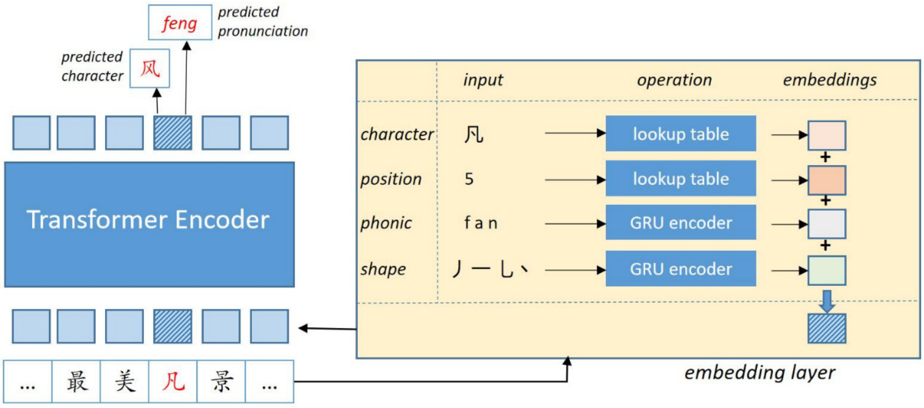

Figure 1: The framework of PLOME model 0.

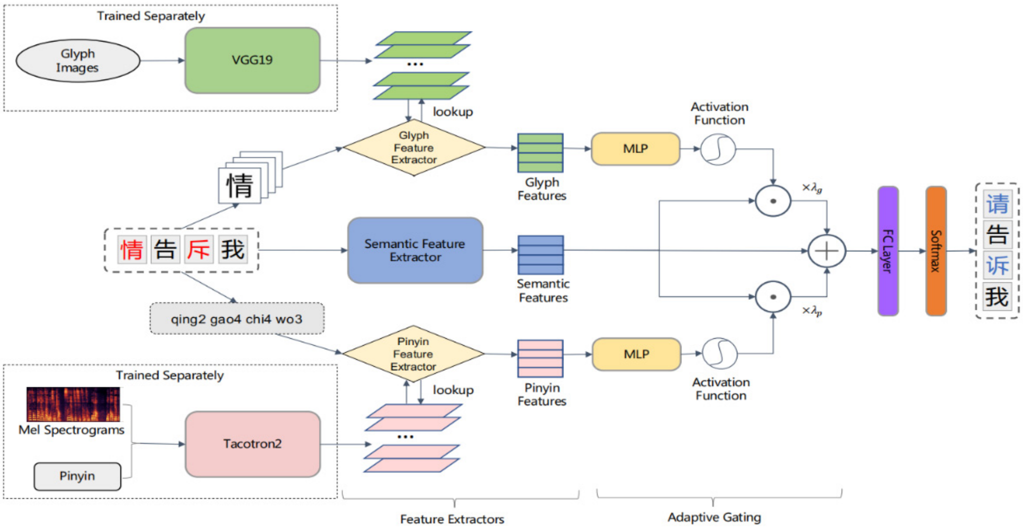

Figure 2: The architecture of PHMOSpell model 0.

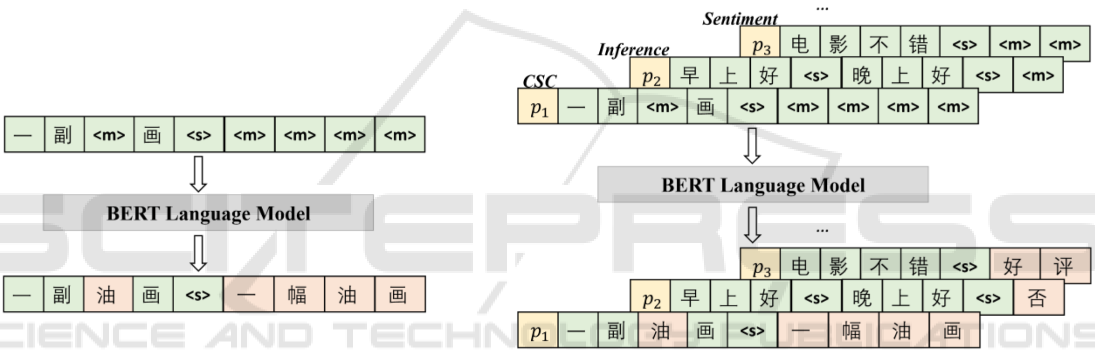
一 副 <m> 画 <s> <m> <m> <m> <m>
BERT Language Model
一 副 油 画 <s> 一 幅 油 画
Sentiment
Inference
CSC
p_3 电 影 不 错 <s> <m> <m>
p_2 早 上 好 <s> 晚 上 好 <s> <m>
p_1 一 副 <m> 画 <s> <m> <m> <m> <m>
BERT Language Model
p_3 电 影 不 错 <s> 好 评
p_2 早 上 好 <s> 晚 上 好 <s> 否
p_1 一 副 油 画 <s> 一 幅 油 画

Figure 3: Paradigm of ReLM in single-task (left) and multi-task (right) setting 0.

In CSC tasks, if only the error pattern is focused on, the correction process will rely too much on the error itself instead of considering the overall semantics of the sentence. To address the issue of excessive emphasis on error patterns while disregarding the sentence's overall semantics during the repair process, researchers have suggested a new training paradigm known as Reworded Language Model (ReLM) (Figure 3). Instead of character-by-character labelling, ReLM trains the model to reformulate the entire sentence by filling in additional slots 0. This approach is more consistent with human reasoning, as individuals tend to reformulate sentences semantically rather than just correct errors. Judging from the results, ReLM lessens the model's excessive dependence on mistakes, strengthens its capacity for generalisation, encourages the prospect of multitask learning, and for transfer.

Another problem that CSC must face is the problem of overcorrection. We want models to accurately identify errors in sentences, but sometimes models are trained to be overly correct, causing correct characters to be incorrectly changed. In order to solve the problem of over-correction, scholars have proposed a new postprocessing model called EDMSpell (Figure 4) (Sheng et al, 2023). Included in the model to postprocessthe correction findings are two checkers: the sentence-level error checker (SEC) and the character-level error checker (CEC). The post-processing module's fundamental idea is to ascertain whether the original sentence and the modified text are correct, then base the final conclusion on this judgement procedure. This can effectively filter out over-correction and reduce the number of correct characters being mistakenly corrected.

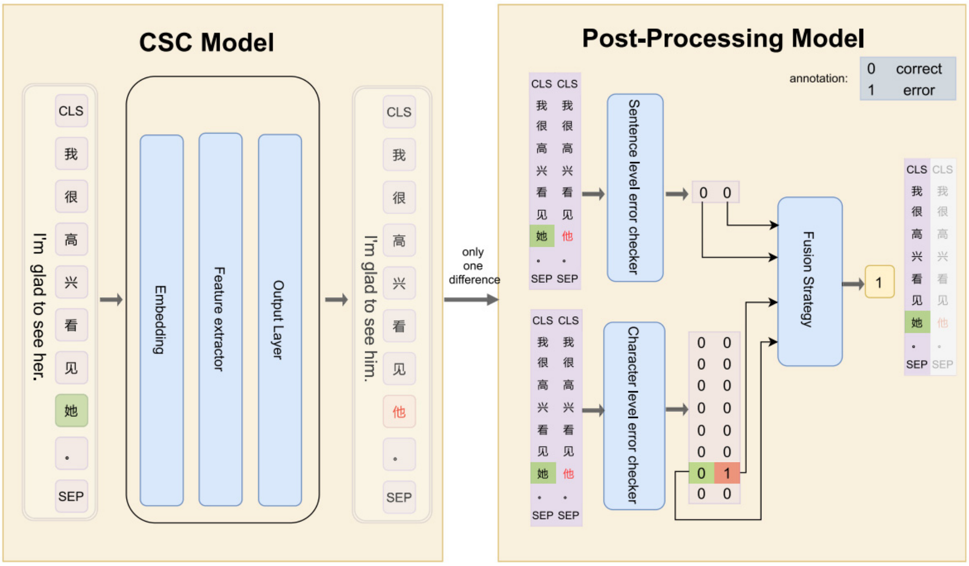

Figure 4: The architecture of EDMSpell model (Sheng et al, 2023).

2.2 CGEC

Recently, there has been a lot of focus on CGEC, a crucial problem in Natural Language Processing (NLP). GEC aims to automatically identify and fix grammatical mistakes in Chinese sentences, enhancing their overall coherence and fluency. Early research mainly focused on how to diagnose errors in Chinese corpora, and there was less research on methods to correct errors. Scholars typically approach the diagnosis of Chinese grammatical faults as sequence annotation errors, and they primarily use conditional random field (CRF) models and long short-term memory (LSTM) models to construct their systems. The Chinese grammar error diagnosis task shared by NLPCC in 2018 has greatly promoted the development of this field. Early Chinese GEC methods relied heavily on rule-based approaches, where hand-crafted grammar rules were used to identify and correct errors. The intricacy and diversity of Chinese grammar limits the usefulness of these approaches, notwithstanding their relative success. Because of this, scientists are starting to investigate data-driven strategies that make use of neural networks and machine learning.

With the emergence of deep learning technology, sequence-to-sequence (seq2seq) neural machine translation (NMT) modelshave become a popular choice for GEC in China. When translating a sentence with grammatical errors, the sentence with the errors is called the source sentence; the sentence with the errors rectified is called the target sentence (Yuan & Briscoe, 2016). Translating two languages with little to no word overlap between the source and target languages is the aim of the translation process.

The typical iterative sequence labelling strategy involves an iterative inference phase that makes the model ignore the results of prior correction rounds and concentrate only on the current phrase's error repair findings. The training portion of the method employs sentences just once. To solve this issue, relevant academics proposed the sequence labelling and iterative training based Chinese grammatical error correcting approach (CGEC-IT) (Figure 5) (Kuang et al, 2022). In the iterative training phase, this methodology employs CRF to improve the model's attention to the overall labelling outcomes, and it uses focal loss to address the text error correction problem of class imbalance. It also dynamically creates target labels for each round. The outcomes of the experiments demonstrate that this approach outperforms earlier research in terms of F0.5 score on NLPCC 2018 Task 2, confirming the usefulness of iterative training for the Chinese GEC model.

Compared with English GEC research, Chinese GEC research is relatively lacking in training data. The dynamic masking method gave rise to the creation of the MaskGEC model, which addresses the issue of limited training data (Figure 6). The MaskGEC model enhances the performance of the neural network GEC model by introducing dynamic masking technology (Zhao & Wang, 2020). During the training phase, this technique dynamically adds random masks to the original source phrases to produce more varied error-corrected sentence pairs,

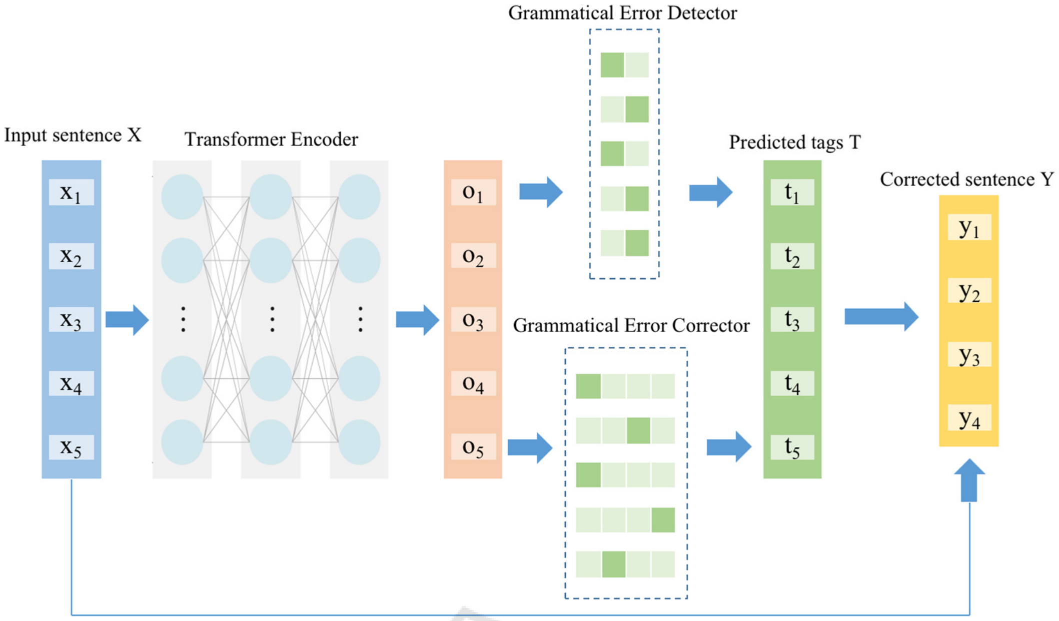

Figure 5: The framework of CGEC-IT (Kuang et al, 2022).

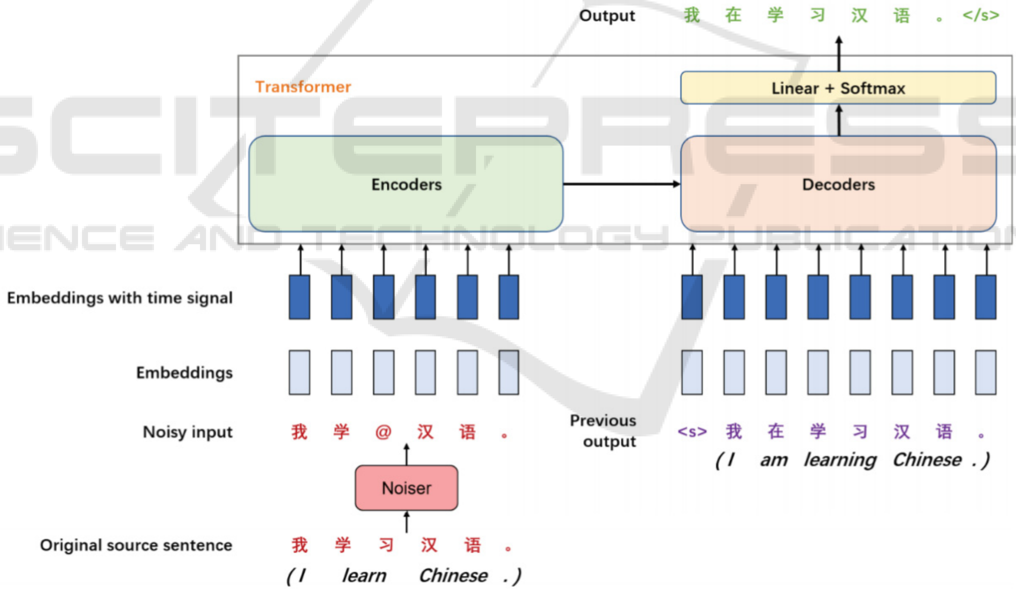

Figure 6: An illustration of the training process of MaskGEC (Zhao & Wang, 2020).

enhancing the model's capacity for generalization.

The training of Chinese GEC models sometimes faces the problem of multiple reference samples. The training strategy "ONE TARGET" proposed by relevant scholars in 2022 explores the impact of multiple reference sample settings on the training of Chinese grammatical error correction models (Ye et al, 2022). During training, a multi-reference sample setup introduces more uncertainty, thereby confusing the model. By comparing the impact of multiple reference samples and a single reference sample on model performance, the study found that training with a single reference sample is more effective, can increase the model's attention to important data, and improve CGEC performance. The "ONE TARGET" strategy selects the most suitable reference text as a training sample and filters out the remaining reference texts to improve training efficiency and model performance. By removing superfluous reference text from the dataset, this method not only expedites the training process but also enhances the model's accuracy and effectiveness. Empirical findings confirm the efficacy of the "ONE TARGET" technique by demonstrating that it is possible to attain optimal performance on the MuCGEC data set.

2.3 Chinese Semantic Error Diagnosis (CSED)

Chinese semantic error diagnosis (CSED) technology is an emerging field within the broader field of Chinese text error correction. CSED has received relatively little attention due to the lack of relevant datasets and the inherent complexity of semantic errors (Sun et al, 2023). Existing datasets (e.g., CTC and MuCGEC) contain only a limited number of semantic errors, which makes developing comprehensive models of CSED challenging. To fill this research gap, researchers developed the CSED corpus. It consists of two datasets: CSED-Recognition (CSED-R) and CSED-Correction (CSED-C).

The researchers proposed a grammar-aware model specifically tailored for the CSED task. These models take into account the syntax and structure of Chinese, allowing them to better diagnose and correct semantic errors. Experimental results demonstrate the effectiveness of adopting a syntax-aware approach to solve CSED challenges (Wang & Xie, 2022).

3 PROBLEMS OF CHINESE TEXT ERROR CORRECTION TECHNOLOGY

The two main methods for correcting text errors in Chinese, CGEC and CSC, still have several problems that need to be fixed. This chapter will provide examples of these issues.

3.1 Problems in CSC

First of all, most of the current Chinese spelling correction technologies introduce pre-training models such as BERT. However, these technologies generally have a problem, that is, their ability to understand complex contexts and handle subtle semantic differences is quite limited. For example, in terms of handling word-level errors, it is difficult for existing technologies to accurately handle polysemy and homonym errors.

Secondly, there are some new Internet words and slang in the current Chinese context, and the continuous emergence of these words greatly increases the complexity of the spelling correction task. Existing techniques tend to have low adaptability when processing texts containing such words.

Furthermore, current error correction technology has the problem of reduced efficiency and accuracy when processing long texts. This problem is particularly serious when global context information needs to be used for error correction judgment.

Also, although large-scale pre-trained models have powerful performance, it is precisely because of the size of the model and the demand for computing resources that some spelling correction technologies are difficult to use in resource-constrained environments.

Lastly, in order to increase the model's accuracy, current Chinese spelling correction techniques mostly rely on a sizable Chinese corpus as training data. However, obtaining high-quality training data is often problematic for texts in certain domains or languages with limited resources. Most of the existing corpora are provided by Internet users, so researchers need to face data inconsistencies and mismatches between wrong sentences and correct sentences. The problem of lack of high-quality training data limits the generalization ability and application scope of existing models.

3.2 Problems in CGEC

Some of the problems existing in current Chinese grammar correction technology are similar to those in spelling correction technology, such as the lack of high-quality and large-scale corpora, reliance on powerful computing resources, and low efficiency when processing long texts. Of course, this technology also has some unique problems.

First of all, some errors involved in CGEC exist in sentence structure, which is caused by the language characteristics of Chinese. Multiple word order errors will result in sentence structure issues that are challenging for existing models to effectively correct, especially at the subtle grammatical and semantic levels. This is because the correction of structural errors is typically not unique and it is challenging for the model to accurately capture all errors.

Secondly, CGEC often needs to consider the context in terms of part of speech, especially in the choice of verbs, which puts higher requirements on the model's context understanding ability (Li et al, 2019). Most Chinese verbs consist of only one character, making it even more difficult to identify and select the appropriate verb.

Lastly, there are still issues with the present CGEC technology, such as how to balance the minimal edit distance principle's requirement for sentence fluency while maintaining error correction accuracy, and how to better optimise the CGEC model for increased efficiency and accuracy. challenges. At the same time, because the evaluation criteria for CGEC tasks are not always consistent, especially when it comes to sentence structure adjustments, the establishment of evaluation criteria is also a difficult problem.

4 CONCLUSION

This article addresses the open issues in the field of Chinese text error correction and presents the most recent advances in models and techniques. Generally speaking, Chinese text error correction technology is developing rapidly in the two directions of spelling error correction and grammatical error correction. New models with better performance are constantly being produced. Today's CSC models can already integrate audio and visual information for error correction, and can consider semantic information to a certain extent. The great variety and unpredictability of Chinese, along with the scarcity of high-quality datasets, continue to pose hurdles to the performance of today's CGEC models on select public datasets.

Subsequent studies pertaining to Chinese text error correction will probably continue to concentrate on the two methodologies of CSC and CGEC. Researchers will continue to conduct further research and improvements on the subtle understanding, adaptability and generalization capabilities of the CSC model and CGEC model. In addition, for CGEC technology, it is quite necessary to have better training data sets in the future. In addition, there are relatively few error correction techniques for Chinese semantics, and allocating funds for this method's study could encourage the advancement of Chinese text error correcting methods.

REFERENCES

X. Cheng, W. Xu, K. Chen, et al. Spellgcn: Incorporating phonological and visual similarities into language models for chinese spelling check. arXiv preprint arXiv:2004.14166, (2020).

B. Wang, W. Che, D.Wu, et al. Dynamic connected networks for Chinese spelling check, Findings of the Association for Computational Linguistics : 2437-2446 (2021).

C. Li C, C. Zhang, Zheng, et al. Exploration and exploitation: Two ways to improve Chinese spelling correction models. arXiv preprint arXiv:2105.14813, (2021).

S. Zhang, H.Huang, J. Liu, et al. Spelling error correction with soft-masked BERT. arXiv preprint arXiv:2005.07421, (2020).

S. Liu S, T. Yang, T. Yue , et al. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction, Inter. Joint Conf. Natur. Lang.Proc. 2991-3000 (2021).

L. Huang, J. Li , W. Jiang, et al. PHMOSpell: Phonological and morphological knowledge guided Chinese spelling check, Inter. Joint Conf. Natur. Lang.Proc. 5958-5967, (2021).

L. Liu, H. Wu, H. Zhao. Chinese Spelling Correction as Rephrasing Language Model. arXiv preprint arXiv:2308.08796, (2023).

L. Sheng, Z. Xu, X. Li, et al., EDMSpell: Incorporating the error discriminator mechanism into Chinese spelling correction for the overcorrection problem. J. King Saud University-Comput. Inf. Sci., 35, art. no. 101573 (2023).

Z. Yuan, T. Briscoe, Grammatical error correction using neural machine translation. Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 380-386 (2016).

H. Kuang, K. Wu, X. Ma, et al., A Chinese Grammatical Error Correction Method Based on Iterative Training and Sequence Tagging. Appl. Sci., 12, art. no. 4364 (2022).

Z. Zhao, H. Wang, Maskgec: Improving neural grammatical error correction via dynamic masking. Proc. AAAI Conf. Artif. Intell., 34, 1226-1233 (2020).

J. Ye, Y. Li, S. Ma, et al., Focus is what you need for Chinese grammatical error correction. CoRR, abs/2210.12692 (2022).

B. Sun, B. Wang, Y. Wang, et al., CSED: A Chinese Semantic Error Diagnosis Corpus. arXiv preprint arXiv:2305.05183 (2023).

F. Wang, Z. Xie, An Adversarial Multi-task Learning Method for Chinese Text Correction with Semantic Detection. Int. Conf. Artif. Neural Networks, 159-173 (2022).

S. Li, J. Zhao, G. Shi, et al., Chinese grammatical error correction based on convolutional sequence to sequence model. IEEE Access, 7, 72905-72913 (2019).