Research progress on Chinese and English text error correction

Yao Wang*

International School, Beijing University of Posts and Telecommunications, Beijing, 100000, China

Abstract. Text error correction is an essential task in natural language processing (NLP) that focuses on automatically identifying and correcting errors in written text. With the increasing amount of digital text in both Chinese and English, errors such as typos, grammatical mistakes, and contextual ambiguities have become more prominent, affecting readability and communication. Over the years, various models and methodologies have been developed to tackle these challenges, evolving from traditional rule-based systems to more advanced statistical and machine learning-based approaches. This paper presents an overview of the current research status in Chinese and English text error corrections. By analyzing several papers involving Chinese and English error correction models, the article points out the types of errors such as spelling, grammatical and semantic errors included in text error correction and their basic elements. It discusses the advantages and limitations of the latest approaches including rule-based models, statistical and deep learning methods. The development potential of deep learning and big modeling techniques in the field of text error correction is shown. These findings help to advance automated error correction systems and their facilitation in the real world.

1 Introduction

In the age of big data and the mobile internet, digital text has become an integral part of daily life, replacing traditional paper-based documents. This shift towards electronic information dissemination has significantly increased the volume of textual data produced and consumed in fields like education, media, business, and personal communication. The prevalence of digital text also highlights the growing issue of text errors, including typographical, grammatical, and semantic mistakes, which can severely affect the accuracy and clarity of the information being conveyed. In this context, the demand for efficient and accurate text error correction systems has become more urgent than ever.

Errors in electronic text are not only common but can propagate rapidly across platforms, especially in environments where large-scale communication takes place, such as social media, email, and digital publishing. Human error correction, which traditionally involves manual proofreading and editing, is time-consuming and prone to mistakes, particularly as the amount of text data continues to explode. In high-stakes environments such as academic publishing, legal documentation, and business correspondence, even minor errors can lead to

* Corresponding author: wangyao1008@bupt.edu.cn

misunderstandings or reputational damage. Therefore, developing automated error correction systems that can efficiently identify and rectify text errors is of paramount importance.

Manual text error correction, though accurate, is slow, labour-intensive, and prone to errors, especially with the growing volume of digital text. As digital content expands, manual proofreading has become impractical. To address this, automated error correction systems have emerged as a more scalable and efficient solution, powered by advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP). In particular, deep learning models excel at processing large datasets and improving their ability to generalize across different types of errors, surpassing the limitations of traditional rule-based and statistical approaches. These systems also prove highly effective in multilingual contexts, where language-specific errors present unique challenges.

Given the massive amount of digital text generated daily, the importance of automated text error correction cannot be overstated. It is essential for ensuring accuracy on a large scale, as reliance on human proofreaders alone is no longer sufficient to meet the growing demands of the digital age. Researchers have responded by developing increasingly sophisticated models, ranging from traditional rule-based systems to advanced AI-driven techniques. Current research underscores both the strengths and limitations of these methodologies, providing insights into how error correction systems can be further refined to improve their efficiency and reliability.

This paper offers a comprehensive review of the current state of Chinese and English text error correction research, examining key methodologies and their effectiveness in addressing spelling, grammatical, and semantic errors. By exploring the evolution of error correction models, from rule-based and statistical approaches to cutting-edge deep learning techniques, this paper aims to shed light on the achievements and challenges that researchers face in this rapidly developing field.

2 Text error correction methods

2.1 Essential elements and methods of error correction in English and Chinese

In Natural Language Processing, English text error correction involves three main types of problems: spelling errors, grammatical errors and semantic errors. Table 1 shows the example of errors in English.

Table 1. example of errors in English.

errors	Error Case	Correct Case
spelling	hte	the
grammatical	She goes to school yesterday	She went to school yesterday
semantic	He gave a piece of cake advice	He gave a piece of valuable advice

In NLP, cognitive Chinese text error correction needs to clarify the following core elements. There are three common types of errors in Chinese text: spelling errors, grammatical errors, and semantic errors. Spelling errors refer to the problem of incorrectly written or misspelled words involving characters in the text. Grammatical errors refer to misspellings, formatting errors, etc., while semantic errors refer to errors in the meaning of the text, such as the use of incorrect vocabulary or unclearly expressed sentences. Table 2 shows the example of errors in Chinese.

Table 2. example of errors in Chinese.

errors	Error Case	Correct Case
spelling	她的确是个才华横**熔**的艺术家。	她的确是个才华横**溢**的艺术家。
grammatical	他**对去**中国感到非常激动。	他**对去中国的事情**感到非常激动。
semantic	在面试的时候，他穿了一件正式的**睡衣**。	在面试的时候，他穿了一件正式的**西服**。

Error correction methods for text correction can be broadly categorized into three types, which include the Rule Engine-Based Approach, Statistical Learning-Based Approach and Deep Learning-Based Approach.

The rule engine approach relies on predefined rules to detect and correct spelling and grammatical errors. The method is mainly based on dictionary matching and grammar rules, but it is difficult to deal with complex context-dependent problems.

In statistical learning and Deep learning approaches based on large-scale corpora, errors are predicted and corrected by statistical patterns, as well as employs neural network models to learn error patterns from large-scale textual data. In recent years, techniques based on pre-trained language models have made significant progress in English text error correction. This approach performs well in error-frequent linguistic phenomena but may have limitations for rare errors.

Chinese error correction tasks can be categorized into three primary areas. Chinese Spelling Check (CSC) primarily addresses word substitution errors, including errors related to input-output mismatches and sentence-level issues. This task has been extensively studied over time, with a primary focus on datasets such as those provided by the SIGHAN evaluation task. Chinese Grammatical Error Correction (CGEC) is more complex, involving operations such as adding or deleting words, often leading to non-equal-length corrections. This task typically relies on predefined templates and large-scale models to handle the intricate structure of grammatical errors.

Error correction systems are generally divided into two main components. The first is Error Detection, which focuses on identifying the location of the error. The second is Error Correction, where for the detected erroneous words, candidate corrections are generated based on features such as phonetic similarity and character shape. These candidate corrections are then ranked using models such as language models, and the optimal correction is selected based on the ranking.

3 Rule Engine-Based Approach

For both English and Chinese, rule-based text error correction algorithms rely on predefined linguistic rules and patterns to recognize and correct errors in text. These patterns capture a variety of linguistic features including grammar, spelling and syntax. In such systems, error detection is achieved by recognizing deviations from established grammatical or lexical norms, while correction is generated through rule-based substitutions or modifications. For example, in English, such approaches may focus on correcting subject-verb agreement or the use of articles by applying syntactic rules, whereas in Chinese, they utilize predefined grammatical frameworks to address errors such as improper word order or missing constituents. Such approaches provide transparent and interpretable corrections, but relying only on simple contextual libraries can be difficult when dealing with more complex and context-dependent errors.

Sidorov et al. introduced a rule-based system for automatic grammar correction that leverages syntactic N-grams for detecting and correcting grammatical errors, specifically for English learners (L2). The system applies predefined syntactic rules to analyze text structure

and identify errors in syntax and grammar. This approach focuses on frequent English errors such as subject-verb agreement and article use, relying on a syntactic N-gram model to track error patterns. By analyzing the relationships between syntactic units rather than individual words, the system can provide more context-sensitive corrections, especially for English learners who tend to make consistent, rule-governed errors [1].

In contrast, Ma et al. proposed a rule-based corpus generation method aimed at correcting Chinese grammatical errors. The model is based on linguistic rules to detect six types of errors in Chinese, including structural confusion, improper word order, and missing components. This system generates error-correction pairs using predefined grammatical rules, allowing it to address frequent grammar issues in Chinese text. The model is particularly effective in handling sentence structure issues, such as misplaced modifiers and incorrect collocations, and has been instrumental in creating a benchmark dataset for further testing and improvement [2].

4 Statistical Learning-Based Approach

A statistical learning-based approach based on statistical learning infers errors in text through a large amount of labeled data and utilizes the adaptive nature of neural network models to improve error correction performance. By combining data-driven statistical features, the model can automatically learn common error patterns in language usage, leading to effective error detection and correction.

Chollampatt et al. present a model based on a multilayer convolutional encoder-decoder neural network that dramatically improves the ability to automatically correct spelling and collocation errors by utilizing character N-gram information to initialize embeddings. The approach utilizes finer-grained feature representations (e.g., character-level N-grams) and thus excels in capturing local contextual information. Evaluation results show that the model significantly outperforms previous neural network-based approaches as well as traditional statistical machine translation systems on both the CoNLL-2014 and JFLEG benchmarking datasets, demonstrating high accuracy, especially in the grammar correction task [3].

In addition, the Grammatical Error Correction Transformer with Representations (GECToR) model proposed by Omelianchuk et al. achieves efficient grammatical error correction utilizing a transformer encoder. The model is first pre-trained on synthetic data and then fine-tuned on a parallel corpus containing both erroneous and correct text. The GECToR model is designed with a series of token-level custom transformation operations, which can accurately map input tokens to target corrected tokens. The model has F0.5 scores of 65.3/66.5 and 72.4/73.6 on the CoNLL-2014 and BEA-2019 test sets, respectively, which far exceeds many Transformer-based sequence-to-sequence models [4]. From the results, the inference speed of GECToR is 10 times faster than the standard sequence-to-sequence GEC system, which makes it more efficient and scalable in practical applications.

The Soft-Mask BERT model, proposed by Zhang et al. from ByteDance in 2020, divides the Chinese spelling error correction task into two parts: error detection (detection network) and error correction (correction network) [5]. In the error detection part, the model detects each input character by BiGRU, outputs the error probability of each character, and derives the soft-masked embedding as the input vector for the error correction part. Specifically, for each character, the BiGRU model calculates its error probability, and the higher the value, the more likely the character is wrong, and the closer the soft-masked vector is to the mask vector, and vice versa, the closer it is to the input vector. This operation effectively reduces the phenomenon of excessive error correction in the BERT model.

In the error correction part, the soft-masked vectors are input into BERT for multi-category error correction classification. Firstly, the input sequence is contextualised by the BERT model to obtain the contextual representation vector of each character. To enhance the

semantic information, the model employs residual concatenation, where the context vectors output from BERT are added with the initial input vectors to form a new representation. Finally, after a fully connected classification layer, Soft-Mask BERT performs error correction prediction for each character and outputs the final corrected sequence.

The model is trained in an end-to-end manner and is jointly evaluated by the two loss functions for error detection and error correction. Soft-Mask BERT is tested on the News Title dataset and SIGHAN dataset. In the News Title dataset, the Soft-Mask BERT model found and corrected more than 54% of the errors with an accuracy of more than 55%. Compared to BERT-Finetune, Soft-Mask BERT can make more efficient use of global context information for more rational error correction. This design balances error detection and correction in the error correction task, significantly improving the accuracy and efficiency of error correction.

5 Deep Learning-Based Approach

Deep learning-based text error correction methods leverage large-scale datasets and powerful neural networks to automatically detect and correct errors. These models typically rely on neural architectures, such as transformers and recurrent neural networks (RNNs), which are pre-trained on vast amounts of data to capture rich linguistic patterns. In English, models like BERT and GECToR use context-aware embeddings to correct grammatical errors, such as subject-verb agreement or incorrect word usage, by learning from diverse text corpora. Similarly, in Chinese, deep learning approaches like Soft-Masked BERT and FASPell utilize advanced neural techniques to address spelling, grammatical, and semantic errors, incorporating character shape and phonetic similarity to enhance correction accuracy. These models excel at handling complex, context-dependent errors and continue to evolve with the integration of cutting-edge architectures like transformers and sequence-to-sequence models.

The BERT model, proposed by Devlin et al. as one of the most widely used pre-training models, generates highly contextually relevant representations through a deep bi-directional Transformer encoder. In error correction tasks, BERT can utilize its bidirectional semantic understanding to detect and correct spelling, syntactic and semantic errors. Its performance in error correction tasks, especially in complex contextual understanding and linguistic rule capture, significantly improves the overall accuracy of the error correction system. The BERT model provides strong foundational support for other deep learning-based error correction models and offers a more effective solution for automated grammatical and linguistic error correction tasks [6].

The approach relies on rules and patterns learned by the model during pre-training and exhibits high robustness and generalization capabilities. This rule-based approach can show advantages in dealing with more complex syntactic structures and long-distance dependencies and thus has a wide range of applications in dealing with diverse textual error correction tasks.

The FASPell method, proposed by Hong et al. in 2019 [7], introduces a novel approach to Chinese spelling error correction utilizing a BERT-based denoising autoencoder (DAE) alongside a confidence-similarity decoder (CSD). The DAE generates correction candidates dynamically, replacing traditional confusion sets by leveraging BERT's contextual embedding capability. The CSD then ranks these candidates based on confidence scores and character similarity, which are derived from phonetic and visual features. This method improves overall correction accuracy by combining character-sound and character-form information, though it struggles with multi-word and under-detected errors due to non-end-to-end training and the manual fitting of the CSD threshold.

In 2020, SpellGCN, proposed by Cheng et al., further refined Chinese spelling correction through the use of Graph Convolutional Networks (GCNs). This model integrates character

phonological and visual similarities into BERT's output representations. By embedding these features into semantic word vectors, SpellGCN enhances the model's ability to correct confusing characters effectively. Despite its strong performance in public datasets, the model's accuracy is limited when handling complex, multi-word errors, and its effectiveness is constrained by the breadth of the confusion set [8].

Nguyen et al. introduced a more adaptable approach to Chinese spell-checking with Domain-shift Conditioning using Adaptable Filtering via Hierarchical Embeddings (HeadFilt). HeadFilt dynamically adjusts its filters, moving away from fixed confusion sets to better accommodate the variability and heterogeneity of error domains. By addressing sparse error distributions, this model demonstrated state-of-the-art results on competitive datasets, offering a robust solution for spelling correction [9].

Additionally, the Spelling Correction as a Foreign Language model proposed by Zhou et al. frames spelling correction as a machine translation task. This approach employs a multilayer encoder-decoder architecture with attention mechanisms, where the encoder processes contextual information and the decoder predicts the correct form by aligning the encoded states with attention. This method effectively handles more complex spelling errors, performing sequence-level error correction akin to translation tasks [10].

6 Conclusion

The importance of Chinese and English text error correction research lies in its ability to enhance the accuracy and clarity of written communication. As digital text production continues to expand globally, especially with the rise of the internet and mobile technology, the need for automated error correction systems becomes more urgent. Manual correction is inefficient and error-prone, making automated methods essential for handling the vast amounts of text generated daily. This paper has provided a review of the current methodologies used in text error correction. By examining rule-based, statistical, and deep learning approaches, as well as machine translation techniques, the study highlights the various ways in which spelling, grammatical, and semantic errors are addressed in both Chinese and English texts. The overview illustrates the strengths and weaknesses of these methods, showing that while significant progress has been made, limitations remain, particularly with context-sensitive or rare errors. Looking forward, there is substantial potential for deep learning models and large-scale language models to further improve error correction systems. These models, with their ability to learn from massive datasets and generalize across diverse linguistic structures, are likely to overcome some of the existing challenges. However, ongoing research is needed to refine these technologies and ensure they are adaptable across different error types and language domains. In conclusion, this paper underscores the growing importance of text error correction research. Reviewing the field's current progress provides a foundation for future developments that will contribute to more accurate, scalable, and efficient error correction systems, meeting the demands of an increasingly digitized world.

References

1. G. Sidorov, A. Gupta, M. Tozer, et al., Rule-based system for automatic grammar correction using syntactic n-grams for English language learning (L2), in Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, 96-101 (2013).
2. S. Ma, et al., Linguistic rules-based corpus generation for native Chinese grammatical error correction. arXiv:2210.10442 (2022).

3. S. Chollampatt, H.-T. Ng, A multilayer convolutional encoder-decoder neural network for grammatical error correction, in Proceedings of the AAAI Conference on Artificial Intelligence, 32, 1 (2018).
4. K. Omelianchuk, V. Atrasevych, A. Chernodub, et al., GECToR – Grammatical error correction: Tag, not rewrite. arXiv preprint arXiv:2005.12592 (2020).
5. S. Zhang, H. Huang, J. Liu, et al., Spelling error correction with soft-masked BERT. arXiv preprint arXiv:2005.07421 (2020).
6. J. Devlin, BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
7. Y. Hong, X. Yu, N. He, et al., FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm, in Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019), 160-169 (2019).
8. 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).
9. M. Nguyen, G. H. Ngo, N. F. Chen, Domain-shift conditioning using adaptable filtering via hierarchical embeddings for robust Chinese spell check. IEEE/ACM Trans. Audio, Speech, Lang. Process. **29**, 2027-2036 (2021).
10. Y. Zhou, U. Porwal, R. Konow, Spelling correction as a foreign language. arXiv preprint arXiv:1705.07371 (2019).