Natural language processing for automatic error detection in Chinese language learning

Zhengxin Li

School of Public Teaching Department,
Fujian Vocational College of Agriculture,
Fuzhou, 350303, China
Email: lizhengxin@fjny.edu.cn

Rongzhen Wu*

School of Information Engineering,
Fujian Vocational College of Agriculture,
Fuzhou, 350303, China
Email: wurongzhen@fjny.edu.cn
*Corresponding author

Abstract: With the rapid development of global Chinese language education, the demand for efficient and accurate automated teaching assistance tools is growing. Traditional manual grading methods are often time-consuming and yield inconsistent results, highlighting the necessity for intelligent technological solutions. This paper explores the application of natural language processing techniques in automatic error detection for Chinese as a second language. It proposes a method based on pre-trained language models and evaluates it using a publicly available corpus of Chinese learner compositions. Experimental results demonstrate the strong performance of the proposed method in identifying grammatical and lexical errors, achieving detection accuracy exceeding 80% for major error categories. This represents a significant improvement over baseline systems (over 25% increase). This technology shows great potential as an efficient teaching support tool, enabling more effective and consistent feedback mechanisms within intelligent educational environments.

Keywords: natural language processing; NLP; Chinese language teaching; automatic bias detection; applicability analysis.

Reference to this paper should be made as follows: Li, Z. and Wu, R. (2026) 'Natural language processing for automatic error detection in Chinese language learning', *Int. J. Information and Communication Technology*, Vol. 27, No. 8, pp.36–52.

Biographical notes: Zhengxin Li is an Associate Professor in the School of Public Teaching Department at Fujian Vocational College of Agriculture, China. She received her Master's degree from Fujian Normal University, China, in 2009. Her research interests include traditional culture, literature and Chinese teaching.

Copyright © The Author(s) 2026. Published by Inderscience Publishers Ltd. This is an Open Access Article distributed under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

Rongzhen Wu is a Professor at the School of Information Engineering at Fujian Vocational College of Agriculture, China. She received her Master's degree from Huazhong University of Science and Technology, China, in 2010. Her research interests include big data technology, deep learning and pre-training language model optimisation.

1 Introduction

Against the backdrop of the global 'Chinese language fever' continuing to intensify and increasingly frequent cultural exchanges, Teaching Chinese as a Second Language (TCSL) faces unprecedented opportunities and challenges. This trend is further propelled by China's sustained socioeconomic growth and expanding global cultural influence, which have enhanced the practical utility of Chinese proficiency in international trade, diplomacy, and cultural sectors. When confronted with a massive number of learners, the core component of traditional teaching models – the identification and correction of language errors – is under immense pressure. Manual homework grading by teachers is not only inefficient but also struggles to maintain consistent standards, failing to provide learners with the immediate, frequent feedback that is crucial for language acquisition. Consequently, developing efficient and precise automated error detection technology has become an urgent necessity to alleviate teaching burdens, enhance instructional quality, and advance the intelligent development of international Chinese education. This requirement transcends mere technical execution, carrying substantial practical significance for advancing theories of language acquisition and transforming pedagogical approaches. The integration of intelligent tutoring systems (ITS) into language education has been shown to potentially alleviate teacher workload and provide personalised learning pathways (Nye, 2015). Unlike conventional automated tools that often follow static rules, ITS utilise adaptive learning algorithms to dynamically tailor instruction and provide personalised feedback based on individual learner performance.

Language error analysis stands as a core research topic in second language acquisition studies. Since Selinker proposed the 'interlanguage' theory (Selinker, 2015), researchers have systematically categorised and traced the origins of learners' systematic errors across phonology, vocabulary, grammar, and pragmatics. In the field of Chinese language teaching, scholars have conducted in-depth descriptions and analyses of common error types. Within the domain of teaching Chinese as a foreign language, scholars have deeply depicted and analysed the common types of bias, such as the confusion between 'le' after the verb and 'le' at the end of the sentence, the absence of specific complements (e.g., resultant and tendency complements), the avoidance and misuse of the word 'put', the improper collocation of quantifiers and nouns, and word order errors caused by negative transfer from the learner's native language. Among these, the 'ba' and 'bei' constructions are particularly challenging as they involve unique syntactic role permutations and specific semantic-pragmatic constraints that are often subject to negative transfer from learners' first languages. The avoidance and misuse of the word '', the improper collocation of quantifiers and nouns, and the disordered word (Tsai and Chu, 2017). These findings provide valuable theoretical insights into learners' language development processes and lay a solid linguistic foundation for constructing automated detection tools. However, traditional error analysis heavily relies on expert subjective judgment, making

it difficult to apply at scale in practical teaching scenarios. The lag in analysis results also limits its ability to provide immediate instructional guidance during the teaching process.

To overcome these limitations, natural language processing (NLP) techniques have naturally been introduced into this field. Automated grammatical error detection and correction techniques first achieved notable progress within English language education, advancing from initial rule-based and statistical methods to the contemporary deep learning framework exemplified by pre-trained models like bidirectional encoder representations from transformers (BERT), generative pre-trained transformer (GPT), and T5. Their detection performance on English texts has reached near-practical levels. However, directly applying these techniques to Chinese error detection presents unique challenges. As an analytic language, Chinese lacks morphological inflection, relying primarily on word order and function words to express syntactic relationships. This renders many morphological features effective in Indo-European languages ineffective. Simultaneously, the accuracy of Chinese word segmentation directly impacts downstream task performance, while errors in learner texts further disrupt segmentation, creating a vicious cycle (Rao et al., 2020). Moreover, many Chinese errors – particularly those involving semantic and discourse coherence – heavily depend on contextual cues, demanding advanced linguistic comprehension capabilities from models. Although some studies have attempted to apply sequence labelling and sequence-to-sequence generation models to the Chinese grammar error detection and correction (GEC) task with some success, current research primarily focuses on improving overall model performance metrics without sufficiently examining the applicability of the techniques themselves.

The concept of applicability refers to investigating the extent to which current advanced NLP technologies can reliably serve Chinese language teaching practices. Existing research has yet to systematically address a series of critical questions: Do these technologies exhibit significant differences in detecting various types of linguistic errors? How do they perform when handling high-frequency, rule-based errors (e.g., misuse of ‘le’) versus low-frequency errors requiring common sense and contextual understanding (e.g., collocation of culturally specific terms)? Where exactly lie the boundaries of their technical strengths and limitations? Current research indicates that most work remains technology-driven, often prioritising higher F1 scores on specific test sets over meticulous evaluation and attribution of effectiveness from a pedagogical perspective (Fleckenstein et al., 2023). This disconnect between technology and real-world applications hinders frontline educators’ understanding and trust in model outputs, thereby obstructing the effective translation of technology into practice. Therefore, filling this research gap by systematically analysing the applicability of NLP technologies in Chinese bias detection tasks – clarifying their capabilities and current limitations – is crucial for driving genuine technological implementation and achieving deep integration with teaching processes. This study, grounded in this premise, aims to construct a systematic analytical framework for a comprehensive and in-depth examination of NLP technology applicability.

2 Related work

2.1 Theoretical research on error analysis in Chinese as a second language

Research on error analysis within the domain of Chinese as a Second Language (CSL) acquisition is well-established and theoretically grounded, with its core objective being the systematic description, classification, and explanation of systematic errors in learners' language production to reveal the developmental patterns of the interlanguage. Early studies, heavily influenced by contrastive analysis and interlanguage theory, centred on predicting and elucidating errors by examining contrasts between the learner's first language and the target language. As research progressed, scholars increasingly recognised that error generation results from the combined influence of multiple factors, including interlingual transfer (negative transfer from the native language), intralingual transfer (overgeneralisation of target language rules), learning strategies, and communicative strategies. Regarding specific error typologies, researchers have conducted extensive and detailed descriptive work. Research on learner language has led to systematic categorisation of lexical, grammatical, and cultural errors, as well as in-depth analysis of acquisition challenges related to specific Chinese sentence patterns such as the 'ba' and 'bei' constructions. In recent years, research perspectives have become increasingly diverse, expanding from traditional morphological and syntactic analysis to encompass discourse coherence, pragmatic functions, and even sociocultural dimensions. These linguistic theoretical achievements provide an indispensable theoretical framework for constructing a hierarchical and operationally feasible error classification system suitable for automated detection. They also constitute the ontological knowledge that any technological application must adhere to. Computational linguistics approaches greatly benefit from such rigorous linguistic theoretical frameworks (Bender, 2013).

2.2 The technological evolution of automatic GEC

Automatic GEC stands as a crucial application domain within NLP. Its technological evolution distinctly mirrors the broader paradigm shifts in NLP. Early research primarily relied on expert-handcrafted rules, which could precisely capture specific, highly regular errors. However, their weaknesses were evident: high labour costs, extremely low coverage, and difficulty in maintenance (Heidorn, 2000). With the advancement of machine learning, the research focus shifted toward statistical approaches. These methods transformed the GEC task into classification or translation problems, leveraging statistical patterns learned from large corpora to detect and correct errors. Examples include confusion-set-based spell checking, error correction using noise channel models that treat incorrect sentences as noisy versions of correct ones, and statistical machine translation (SMT) frameworks that 'translate' erroneous sentences into correct ones. These data-driven approaches significantly enhance system coverage and adaptability, yet their performance heavily relies on feature engineering quality and training data scale. In recent years, deep learning has revolutionised this field. Sequence-to-Sequence (Seq2Seq) models based on recurrent neural networks (RNNs) and attention mechanisms have become mainstream, enabling end-to-end learning of complex mappings from errors to correct forms. Notably, the emergence of pre-trained language models (e.g., BERT, GPT and T5) has enabled models to acquire deep linguistic representations through pre-training on massive unlabeled text corpora. Subsequent fine-tuning on relatively

small amounts of GEC annotated data achieves state-of-the-art performance. This evolution from rule-based to neural and pre-trained models signifies a paradigm shift towards data-driven, generalisable approaches in NLP-based educational applications (Bryant et al., 2023). This represents a paradigm shift from earlier task-specific models, which required extensive feature engineering for each new application, to models that leverage generalised linguistic knowledge acquired through pre-training on massive corpora. This paradigm not only substantially improves performance but also enhances the ability to handle complex errors and context-dependent relationships.

Table 1 A comparison of studies related to automatic Chinese grammar error detection

Methodology	*Error coverage*	*Performance*	*Core limitations/focus*
Rule-based & Statistical ML	Limited specific error types	High precision, low recall	Poor generalisation
Neural Seq2Seq models	Diverse but noisy labels	Moderate F1-score	Unstable on complex syntax
Pre-trained model fine-tuning	Lexical & grammatical errors	High overall F1-score	Lacks granular error analysis
Multi-model comparative analysis	Comprehensive taxonomic errors	Fine-grained F1 by type	Systematic applicability evaluation

2.3 Research on automatic grammatical error detection for Chinese

Although the application of GEC technology to Chinese began relatively late, it has attracted increasing attention from scholars. This growing interest is paralleled by a global surge in research on NLP for under-resourced languages and specific linguistic phenomena (Leacock et al., 2014). Owing to the distinct linguistic features of Chinese, related research faces a series of distinct challenges, foremost among which is the problem of word segmentation. The Chinese writing system lacks spaces between words, making word segmentation an essential preprocessing step for nearly all Chinese NLP tasks. However, grammatical errors in learner texts can directly interfere with segmenters, causing error propagation that subsequently impacts bias detection performance (Rao et al., 2020). Early Chinese GEC research similarly followed a path from rule-based to statistical approaches (Shu et al., 2017). For instance, some studies attempted to construct rule libraries targeting common errors or employed classifier-based methods (e.g., support vector machine) to detect specific error types. With the rise of deep learning, researchers began adopting RNN and Transformer-based Seq2Seq models trained on crowdsourced learning platform data such as Lang-8. In recent years, pre-trained models have become the mainstream approach, with multiple studies confirming the effectiveness of fine-tuned models like BERT on Chinese GEC tasks (Chen and Zhang, 2022). The Chinese grammatical error diagnosis (CGED) shared task organised by the Natural Language Processing Techniques for Educational Applications (NLPTEA) workshop has provided a unified evaluation platform and dataset for this field, significantly advancing technical progress. However, a review of existing research reveals that the vast majority of work still focuses on improving single technical metrics like overall accuracy and F1 scores, with models typically operating as ‘black boxes.’ There is a lack of detailed analysis on how models perform differently across various types of errors. Furthermore, these studies fail to evaluate the reliability of

these technologies in real-world teaching scenarios or identify their fundamental limitations from the perspective of practical Chinese language instruction. This disconnect between technology and application makes it difficult for frontline educators to understand and trust the outputs of automated tools, hindering their effective integration and deployment in actual teaching environments.

3 Methodology

3.1 Problem formulation and task definition

This research frames the automatic detection of grammatical errors in Chinese as a task of sequence labelling. Given an input sequence $\mathbf{X} = (x_1, x_2,\ldots,x_n)$ composed of n characters or lexical units, where x_i represents the i^{th} unit in the sequence, the model aims to output a corresponding label sequence $\mathbf{Y} = (y_1, y_2,\ldots, y_n)$. The labels y_i are drawn from a predefined label set $\mathcal{L}$, which adopts the classical 'BIO' annotation scheme (Ramshaw and Marcus, 1999) and is extended to accommodate Chinese error types. The BIO scheme was selected for its widespread adoption in sequence labelling tasks and its efficiency in precisely demarcating the boundaries and types of errors within a sequence. Specifically, $\mathcal{L}$ includes start (B–) and internal (I–) tags denoting 'correct' and various errors, such as B-WO (lexical error start), I-GR (grammatical error internal), C (correct), etc. This formal approach enables the model not only to detect the presence of errors but also to precisely pinpoint their scope and type.

From a probabilistic standpoint, we model this task by learning a conditional distribution $P(\mathbf{Y} \mid \mathbf{X}; \Theta)$, with Θ denoting the model parameters. The most probable label sequence $\hat{\mathbf{Y}}$ for an input $\mathbf{X}$ is derived by maximising this conditional probability:

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}\in\mathcal{Y}^n} P(\mathbf{Y} \mid \mathbf{X};\Theta) \tag{1}$$

where $\mathcal{Y}^n$ denotes the set of all possible n-length label sequences.

3.2 Sorting target types and spatial parameters

This study utilises the Hanyu Shuiping Kaoshi (HSK) dynamic writing corpus (Zhang, 2023) as its experimental data source. This corpus is recognised as an authoritative, a large-scale, publicly accessible dataset within the research domain of CSL acquisition. This corpus was selected for its large scale, diversity of learners across different native languages and proficiency levels, and its meticulously annotated errors, making it an authoritative benchmark in CSL research. Its materials are derived from writing tasks in the HSK (Chinese Proficiency Test), encompassing texts from learners with diverse native language backgrounds and varying levels of Chinese proficiency. The corpus features meticulous manual error annotation, ensuring high reliability and validity.

Preprocessing is a critical step in ensuring data quality. First, we clean the raw text by removing irrelevant tags and formatting information. Subsequently, we employ the Natural Language Processing & Information Retrieval (NLPIR) Chinese Word Segmentation System, developed by the Institute of Computing Technology, Chinese Academy of Sciences, for text processing. Chinese Academy of Sciences to segment

correct sentences (Zhang et al., 2023). The NLPIR system was chosen for its strong academic reputation, proven high performance in benchmark evaluations, and demonstrated suitability for segmenting educational and learner-generated text. For sentences containing errors, we employ an iterative alignment strategy: first segmenting the corrected sentences, then mapping the segmented results back to the original erroneous sentences using sequence alignment algorithms (e.g., minimum edit distance). This approach minimises the interference of errors during the segmentation phase.

The evaluation of a word segmentation system relies on the metrics of precision, recall, and F1-score. Let S_{gold} denote the manually annotated ground truth segmentation results, and S_{pred} denote the segmenter's output results:

Accuracy rate P measures the proportion of correctly predicted words out of all predicted words:

$$P = \frac{\left|S_{gold} \cap S_{pred}\right|}{\left|S_{pred}\right|} \tag{2}$$

Recall rate R measures the proportion of correctly predicted words out of all ground truth words:

$$R = \frac{\left|S_{gold} \cap S_{pred}\right|}{\left|S_{gold}\right|} \tag{3}$$

The F1-score represents the harmonic mean of precision and recall, calculated as:

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{4}$$

where $|\cdot|$ denotes the number of elements in a set. In this study, we ensured that the F1 score for the word segmentation stage exceeded 98%, thereby establishing a reliable foundation for subsequent tasks (Aromataris and Pearson, 2014). High-quality tokenisation is universally recognised as a critical preprocessing step that directly influences the performance of downstream NLP models.

Finally, we convert the processed text and labels into an input format acceptable to the model, transforming characters or words into corresponding embedding vectors. Let the vocabulary size be $|V|$ and the embedding dimension be d_{model}. By looking up the embedding matrix $\mathbf{E} \in \mathbb{R}^{|V| \times d_{model}}$, the input sequence $\mathbf{X}$ is transformed into a sequence of embedding vectors $\mathbf{H}^{(0)} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$.

3.3 Bias classification system

Based on theories of CSL acquisition and drawing upon the existing annotation system of the HSK corpus, we have constructed a hierarchical error classification framework designed to establish a mapping between technical labels and linguistic theories. This system comprises four major categories and 11 subcategories:

1 Morphological errors (ME): Includes misuse of nouns, verbs, adjectives, adverbs, measure words, and particles.

2 Syntactic errors (SE): Includes missing constituents, redundant constituents, incorrect word order, and mixed sentence structures.

3 Semantic errors (SemE): Primarily refers to inappropriate word combinations and illogical semantic relationships.

4 Discursive errors (DE): Primarily refers to coherence issues caused by the misuse of conjunctions.

This classification system provides the theoretical foundation for subsequent fine-grained performance analysis (Goo, 2010).

3.4 Model selection and architecture

To comprehensively evaluate the applicability of different technical approaches, we selected two representative models for comparative analysis.

First is the baseline model conditional random fields (CRFs), a discriminative probabilistic graphical model well-suited for sequence labelling tasks. CRF was selected as a baseline for its established effectiveness in sequence labelling tasks and its capability to model dependencies between adjacent labels, providing a robust and interpretable benchmark. They capture dependencies between input sequences and output labels by defining feature functions. Given an input sequence **X** and a label sequence **Y**, their conditional probability is defined as:

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp\left(\sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n} f_k \left(y_{i-1}, y_i, \mathbf{X}, i \right) \right) \tag{5}$$

where $Z(\mathbf{X}) = \sum_{\mathbf{Y}'} \exp\left(\sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n} f_k \left(y'_{i-1}, y'_i, \mathbf{X}, i \right) \right)$ is the normalisation factor (Partition Function). f_k is the k^{th} feature function, measuring the association between the adjacent labels (y_{i-1}, y_i) and the entire input sequence **X** at position i. λ_k is the weight parameter for the corresponding feature function, learned through training (Sutton and McCallum, 2012).

We combine unigram features (e.g., current character, part-of-speech), bigram features (e.g., adjacent character combinations), and lexicon features (presence in the negative word dictionary) to construct the feature template for the CRF.

The second model represents a state-of-the-art approach, implemented as a BERT-based sequence labelling system. This model uses a pre-trained BERT encoder as its foundation, augmented with a linear classification layer for predicting labels. BERT itself utilises the Transformer architecture (Vaswani et al., 2017) to capture deep contextual semantic information.

Input representation in BERT is formed by summing token embeddings, segment embeddings, and position embeddings:

$$Z(\mathbf{X}) = \sum_{\mathbf{Y}'} \exp\left(\sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n} f_k \left(y'_{i-1}, y'_i, \mathbf{X}, i \right) \right) \tag{6}$$

- Transformer encoder: BERT is composed of L layers of identical Transformer blocks stacked together. The computation for each layer l comprises two sub-layers: multi-head self-attention (MHA) and feed-forward network (FFN).
- Multi-head self-attention mechanism: First, the output $\mathbf{H}^{(l-1)}$ from the previous layer is projected onto the query, key, and value spaces via a linear transformation:

$$\mathbf{H}^{(0)} = \mathbf{E}_{token} + \mathbf{E}_{segment} + \mathbf{E}_{position} \tag{7}$$

  where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{d_{model} \times d_k}$ are the learnable parameter matrices for the h^{th} attention head, where $d_k = d_{model} / H$ and H denotes the number of attention heads.
- Then compute the scaled pointwise attention:

$$\mathbf{Q}_h = \mathbf{H}^{(l-1)} \mathbf{W}_h^Q \tag{8}$$

- Concatenate the outputs of all heads and perform linear projection again to obtain the final output of the MHA layer:

$$\mathbf{K}_h = \mathbf{H}^{(l-1)} \mathbf{W}_h^K \tag{9}$$

  where $\mathbf{W}^O \in \mathbb{R}^{d_{model} \times d_{model}}$.
- Feedforward neural network with residual connections: The output of the MHA layer undergoes residual connections and layer normalisation (LayerNorm, LN) before being fed into the FFN:

$$\mathbf{V}_h = \mathbf{H}^{(l-1)} \mathbf{W}_h^V \tag{10}$$

$$\text{head}_h = \text{Softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_k}} \right) \mathbf{V}_h \tag{11}$$

$$\text{MHA}\left(\mathbf{H}^{(l-1)}\right) = \text{Concat}\left(\text{head}_1, ..., \text{head}_H\right) \mathbf{W}^O \tag{12}$$

  where $\mathbf{W}1 \in \mathbb{R}^{d_{model} \times d_{ff}}$, $\mathbf{W}2 \in \mathbb{R}^{d_{ff} \times d_{model}}$, d_{ff} denotes the dimension of the FFN intermediate layer, and GeLU represents the Gaussian Error Linear Unit activation function (Lin et al., 2022).

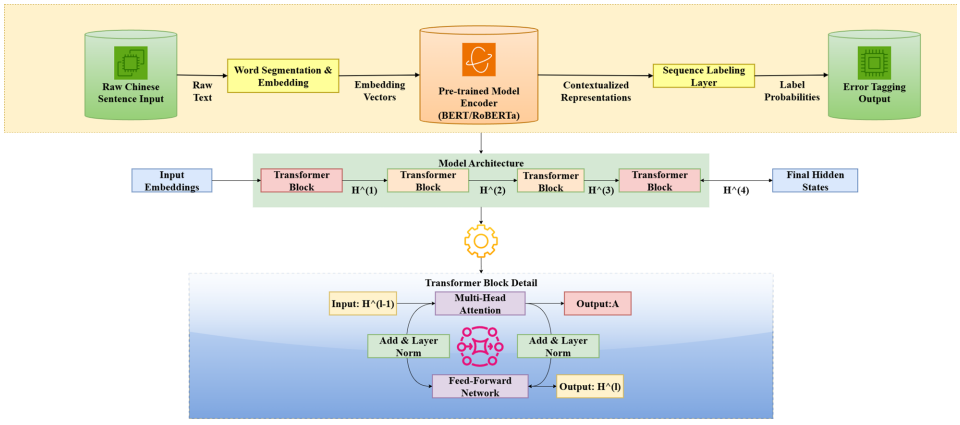
Finally, we pass the output state $\mathbf{H}^{(L)}$ from the last layer of BERT through a fully connected layer to compute scores for all labels at each position i:

$$\mathbf{Z}^{(l)} = \text{LN}\left(\mathbf{H}^{(l-1)} + \text{MHA}\left(\mathbf{H}^{(l-1)}\right)\right) \tag{13}$$

where $\mathbf{W}_{cls} \in \mathbb{R}^{d_{model} \times |\mathcal{L}|}$ and $\mathbf{b}_{cls} \in \mathbb{R}^{|\mathcal{L}|}$. Finally, the Softmax function is applied to obtain the probability distribution:

$$\text{FFN}\left(\mathbf{Z}^{(l)}\right) = \text{GeLU}\left(\mathbf{Z}^{(l)} \mathbf{W}_1 + \mathbf{b}_1\right) \mathbf{W}_2 + \mathbf{b}_2 \tag{14}$$

Figure 1 Schematic diagram of the pre-trained model-based automatic error detection for Chinese (see online version for colours)



3.5 Experimental setup and evaluation metrics

Experimental configuration: The dataset was randomly partitioned into training, development, and test sets with a ratio of 8:1:1. The development set served for hyperparameter optimisation and early stopping. Regarding the BERT model, we adopted the Chinese pre-trained model BERT-wwm-ext (Cui et al., 2021) released by the HIT-iFlytek Joint Laboratory as the base model and fine-tuned it using the AdamW optimiser (Xie et al., 2020). The parameter update rules are as follows:

$$\mathbf{H}^{(l)} = \mathrm{LN}\left(\mathbf{Z}^{(l)} + \mathrm{FFN}\left(\mathbf{Z}^{(l)}\right)\right) \tag{15}$$

$$\mathbf{s}_i = \mathbf{h}_i^{(L)}\mathbf{W}_{cls} + \mathbf{b}_{cls} \tag{16}$$

$$P\left(y_i \middle| \mathbf{X}\right) = \mathrm{Softmax}\left(\mathbf{s}_i\right) = \frac{\exp\left(s_{i,y_i}\right)}{\sum_{c\in\mathcal{L}} \exp\left(s_{i,c}\right)} \tag{17}$$

$$\mathcal{L} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n} \log P\left(y_i^{(j)} \middle| \mathbf{X}^{(j)}; \Theta\right) \tag{18}$$

where $\mathbf{g}_t$ is the gradient at step t, $\mathbf{m}_t$ and $\mathbf{v}_t$ are the first-order and second-order momentum estimates, β_1 and β_2 are momentum hyperparameters, η is the learning rate, and ϵ is a small constant added for numerical stability. The learning rate employs a linear warmup strategy, with the warmup steps constituting 10% of the total training steps.

- *Loss function:* Model training is optimised by minimising the cross-entropy loss function:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \frac{\widehat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \tag{19}$$

where N is the total number of training samples, and $y_i^{(j)}$ is the true label at position i for the j^{th} sample.

- Evaluation metrics: We employ precision (P), recall (R), and F1-score ($F1$) as core evaluation metrics, whose definitions are provided in Section 3.2. All evaluations are conducted at the **token level** (character or word).

4 Experiments and results analysis

4.1 Overall performance comparison

Evaluating NLP systems for pedagogical purposes requires careful consideration of metrics that align with educational outcomes (Leacock et al., 2014). To comprehensively evaluate the applicability of NLP techniques for the task of automatic detection of Chinese language bias, we evaluated the comprehensive performance of the models outlined in chapter 3 using the test set. The models included in the comparison are:

1 CRF baseline model: Utilises a CRF with feature templates comprising unigram, bigram, and lexical features.

2 BERT: Employs the BERT-base architecture, where the output of the CLS token undergoes sequence labelling via a classification layer.

3 RoBERTa-wwm-ext serves as our optimal model, which builds upon the Chinese RoBERTa architecture with Whole Word Masking and has been further enhanced through extended pre-training on expanded corpora. During pre-training, it incorporates improvements such as dynamic masking and removal of the next-sentence prediction task to achieve stronger language representation capabilities. RoBERTa-wwm-ext was identified as optimal due to its Whole Word Masking pre-training strategy and extended training on large corpora, which enhance its ability to capture deep contextualised representations of Chinese words and syntax.

All models underwent training and hyperparameter tuning using identical training and development sets, with final evaluation conducted on a common test set. Token-level precision (P), recall (R), and F1-score ($F1$) were employed as the primary evaluation metrics. To better align with educational applications, we report Strict F1, which counts a prediction as correct only when both the predicted error span and error type perfectly match the human annotation.

Table 2 Overall performance comparison of different models on the test set

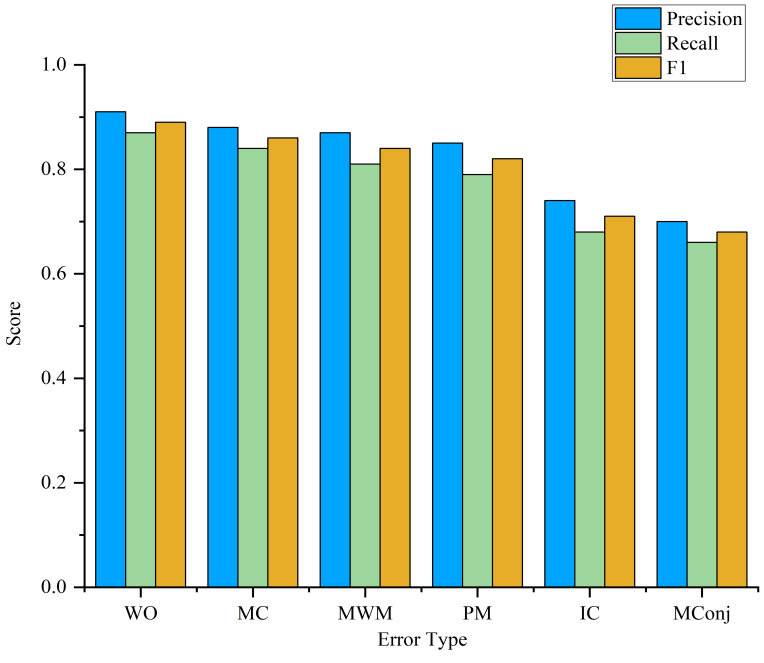
Mould	*Precision*	*Recall*	*F1 score*
CRF	0.712	0.602	0.653
BERT	0.783	0.801	0.792
RoBERTa-wwm-ext	0.804	0.839	0.821

- Results and analysis: The overall performance comparison is presented in Table 1. As shown, deep learning approaches utilising pre-trained language models (e.g., BERT and RoBERTa) substantially exceed traditional machine learning techniques

like CRF in every metric. Specifically, the RoBERTa-wwm-ext model attained the highest performance, achieving an F1 score of 0.821. This represents an absolute improvement of nearly 17 percentage points compared to the CRF baseline model (F1=0.653). This demonstrates that the deep contextual lexical and syntactic knowledge captured by pre-trained models provides an overwhelming advantage for understanding and detecting linguistic biases in Chinese.

Notably, the RoBERTa model demonstrates exceptional performance in recall (0.839), indicating its ability to detect more genuine errors overlooked by the CRF model. However, its precision (0.804) is slightly lower than its recall, suggesting a tendency toward over-correction – where the model may misclassify some correct expressions as errors. This phenomenon warrants particular attention in practical applications, as frequent false positives can erode trust in the system among both teachers and learners. The precision-recall trade-off is a well-known challenge in automated error detection systems designed for language learning. The CRF model exhibits the opposite characteristics, with its precision (0.712) exceeding its recall (0.602), indicating a relatively conservative approach that results in more missed errors. Therefore, interpreting model performance requires a balanced view that considers both overall metrics and their breakdown across different error categories (Loewen et al., 2009). The observed performance gap between error types underscores the necessity of moving beyond aggregate metrics towards more nuanced, error-specific evaluations (Wang et al., 2021).

Figure 2 Performance of the RoBERTa model on different error types (see online version for colours)



Notes: WO: word order, MC: missing constituent, MWM: measure word misuse, PM: particle misuse, IC: inappropriate collocation, MConj: misused conjunction

4.2 Performance analysis by type

The overall performance masks the model's divergent behaviour across different types of bias. To explore the applicability boundaries of NLP technology, we further analysed the fine-grained performance of the best model (RoBERTa-wwm-ext) across our defined bias classification system.

- Results and analysis: The results are shown in Figure 2 (grouped bar chart). It is clearly observable that the model exhibits significant imbalance in its ability to detect different types of bias.
    1 SE showed the best performance: the model achieved near-excellent results on 'incorrect word order' (F1 = 0.89) and 'missing constituents' (F1 = 0.86). These errors typically possess relatively clear contextual syntactic clues, such as misplaced subject-verb-object structures or missing core verbs, which pre-trained models can effectively leverage using their learned syntactic constraints.
    2 ME showed robust performance: The model also achieved high accuracy for 'measure word misuse' (F1 = 0.84) and 'particle misuse' (e.g., 'le', 'zhe', 'guo') (F1 = 0.82). This stems from the strong habitual collocation relationships between measure words and nouns, or verbs and particles in Chinese, which models readily capture from large-scale corpora.
    3 SemE and DE errors pose the primary challenge: Model performance notably declines on 'inappropriate word combinations' (F1 = 0.71) and 'misused conjunctions' (F1 = 0.68). These errors heavily depend on deeper semantic comprehension and discourse logical reasoning, rather than merely local syntactic patterns. For instance, determining whether 'spreading knowledge' or 'spreading news' is more natural, or whether the contrastive logic of 'although...but...' is appropriately applied, requires models to possess near-human common sense and reasoning capabilities – a major bottleneck in current technology (Davis and Marcus, 2015). Semantic and discourse-level understanding remains a formidable challenge for even the most advanced NLP models (Lake and Murphy, 2023).

This analysis holds significant pedagogical implications. It demonstrates that current technology is best suited as an 'auxiliary screening tool' for grammatical and formal errors, greatly reducing teachers' workload in such repetitive tasks. However, for assessing higher-order language competencies involving semantics and discourse, teachers' professional judgment remains indispensable. Outputs from technological tools in such scenarios should serve only as supplementary references.

4.3 Case studies of errors

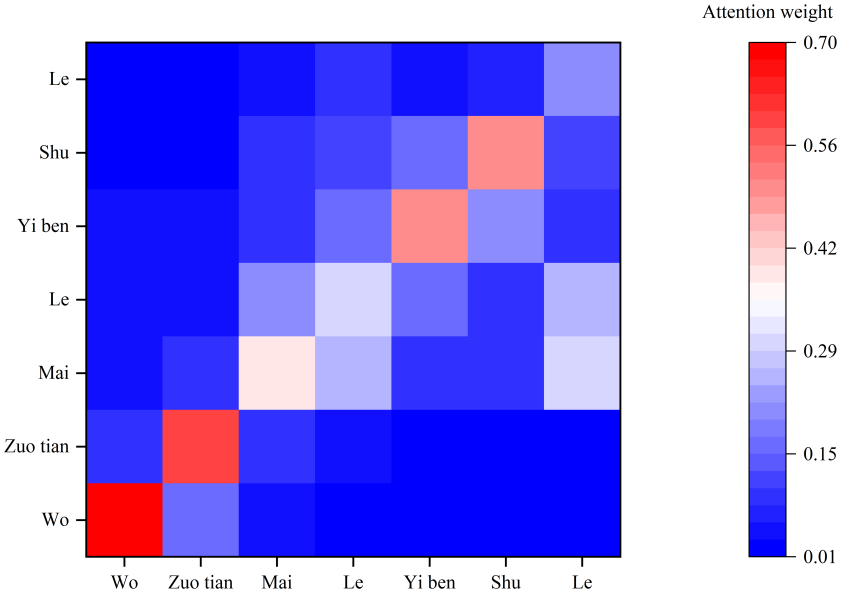
To qualitatively assess the model's behaviour, we conducted an in-depth error case study. Figure 3 visualises the model's attention distribution in a typical case.

In a successful case, the second 'le' in the original sentence 'wo zuo tian mai le yi ben shu le' was manually annotated as redundant and requiring deletion. The model correctly identified it as a 'B-RD' (beginning of redundant error). This sentence exemplifies a typical misuse of 'le', and the model likely mastered the grammatical rule that a final 'le'

is generally unnecessary after 'verb + 'le' + object.' The heatmap in Figure 3 shows that when judging the second 'le', the model assigned higher attention weights to the verb 'mai' and the first 'le', indicating its decision was based on understanding the sentence's overall grammatical structure.

In the failed example, the original sentence 'ta de guan dian fei chang jian gu, shuo fu le suo you ren' contained the error 'jian gu' being flagged as a mismatched collocation and should be corrected to 'jian ding'. However, the model failed to identify this error. 'Guan dian jian gu' constitutes a semantic collocation error. While 'jian gu' and 'jian ding' may correspond to the same adjective in English, in Chinese they respectively modify concrete objects and abstract concepts. This missed detection indicates the model still lacks sufficient deep semantic knowledge, unable to fully grasp Chinese collocation restrictions and semantic nuance details.

Figure 3 Heat map of the model's attention when judging the sentence-final 'le' (see online version for colours)



4.4 Melting experiment

To assess the impact of individual components on model performance, an ablation analysis was performed using the RoBERTa-wwm-ext framework. We tested the following variants: a model without Whole-Word Masking (WWM) pretraining, and a model that removed the top-level CRF layer and performed pointwise prediction using only Softmax.

The ablation results indicate that different components significantly impact model performance. After removing the WWM strategy, the model's overall F1 score decreased by 1.2 percentage points (from 0.821 to 0.809). This suggests that masking entire words during Chinese pre-training enables more effective learning of complete lexical representations, thereby enhancing performance on downstream lexical and syntactic

error detection tasks. On the other hand, removing the CRF layer caused the F1 score to drop by 0.8 percentage points (to 0.813). The CRF layer enhances the global consistency of label sequences by modelling transfer constraints between labels (e.g., 'I-ERROR' cannot follow 'C'), but its gain is relatively limited. This suggests that the pre-trained model already possesses strong sequence modelling capabilities. Ablation studies are a standard methodology in machine learning for quantifying the contribution of individual model components (Dabre et al., 2020). The effectiveness of these efforts depends critically on adopting a human-centred perspective in artificial intelligence, where technology is designed to augment and empower educators, not to replace them (Topali et al., 2025).

5 Conclusions

This study systematically evaluates the applicability of NLP techniques for automatic error detection in CSL. Experiments conducted on large-scale public corpora demonstrate that pre-trained language models achieve near-practical performance in detecting formal and local grammatical errors. However, they exhibit notable limitations in identifying errors that require deeper semantic understanding or discourse-level reasoning. These findings clearly indicate that current NLP technologies are best utilised as auxiliary tools for handling high-frequency normative errors rather than as replacements for human assessment.

The theoretical contributions of this research encompass the creation of a detailed evaluation framework that combines principles from computational linguistics and second language acquisition theory, which enables nuanced analysis of model performance across error categories. Furthermore, emphasising model interpretability – through visualisation mechanisms such as attention heatmaps – facilitates greater transparency and trust among educators, supporting the adoption of such technologies in real-world teaching contexts. From a practical perspective, we recommend a human-machine collaborative approach wherein automated systems provide initial feedback on routine errors within digital learning platforms, thereby allowing instructors to focus on cultivating higher-level language skills. Developers are encouraged to design interpretable models incorporating linguistic knowledge and to create efficient interfaces for teacher oversight and feedback integration.

Future research should focus on overcoming limitations in semantic and discourse-related error detection, potentially through incorporating external knowledge resources or advanced reasoning models. Investigating few-shot learning and domain adaptation methods could enhance personalised support for learners from diverse native language backgrounds. Moreover, long-term empirical studies in authentic classroom settings are essential to validate the effectiveness and practicality of human-AI collaborative grading models. Continued interdisciplinary collaboration among linguistics, education, and computer science remains crucial for developing robust and pedagogically sound NLP applications for language education.

Declarations

All authors declare that they have no conflicts of interest.

References

Aromataris, E. and Pearson, A. (2014) 'The systematic review: an overview', *AJN The American Journal of Nursing*, Vol. 114, No. 3, pp.53–58.

Bender, E.M. (2013) 'Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax', *Synthesis Lectures on Human Language Technologies*, Vol. 15, No. 2, pp.1–178.

Bryant, C., Yuan, Z., Qorib, M.R., Cao, H., Ng, H.T. and Briscoe, T. (2023) 'Grammatical error correction: a survey of the state of the art', *Computational Linguistics*, Vol. 49, No. 3, pp.643–701.

Chen, B. and Zhang, J. (2022) 'Pre-training-based grammatical error correction model for the written language of Chinese hearing impaired students', *IEEE Access*, Vol. 10, pp.35061–35072.

Cui, Y., Che, W., Liu, T., Qin, B. and Yang, Z. (2021) 'Pre-training with whole word masking for chinese bert', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp.3504–3514.

Dabre, R., Chu, C. and Kunchukuttan, A. (2020) 'A survey of multilingual neural machine translation', *ACM Computing Surveys (CSUR)*, Vol. 53, No. 5, pp.1–38.

Davis, E. and Marcus, G. (2015) 'Commonsense reasoning and commonsense knowledge in artificial intelligence', *Communications of the ACM*, Vol. 58, No. 9, pp.92–103.

Fleckenstein, J., Liebenow, L.W. and Meyer, J. (2023) 'Automated feedback and writing: a multi-level meta-analysis of effects on students' performance', *Frontiers in Artificial Intelligence*, Vol. 6, p.1162454.

Goo, J. (2010) 'Structure of service level agreements (SLA) in IT outsourcing: the construct and its measurement', *Information Systems Frontiers*, Vol. 12, No. 2, pp.185–205.

Heidorn, G. (2000) 'Intelligent writing assistance', *Handbook of Natural Language Processing*, Vol. 20, No. 1, pp.1–184.

Lake, B.M. and Murphy, G.L. (2023) 'Word meaning in minds and machines', *Psychological Review*, Vol. 130, No. 2, p.401.

Leacock, C., Chodorow, M., Gamon, M. and Tetreault, J. (2014) 'Automated grammatical error detection for language learners', *Synthesis Lectures on Human Language Technologies*, Vol. 7, No. 1, pp.1–170.

Lin, T., Wang, Y., Liu, X. and Qiu, X. (2022) 'A survey of transformers', *AI Open*, Vol. 3, pp.111–132.

Loewen, S., Li, S., Fei, F., Thompson, A., Nakatsukasa, K., Ahn, S. and Chen, X. (2009) 'Second language learners' beliefs about grammar instruction and error correction', *The Modern Language Journal*, Vol. 93, No. 1, pp.91–104.

Nye, B.D. (2015) 'Intelligent tutoring systems by and for the developing world: a review of trends and approaches for educational technology in a global context', *International Journal of Artificial Intelligence in Education*, Vol. 25, No. 2, pp.177–203.

Ramshaw, L.A. and Marcus, M.P. (1999) 'Text chunking using transformation-based learning', *Natural Language Processing Using Very Large Corpora*, Vol. 11, No. 3, pp.157–176.

Rao, G., Yang, E. and Zhang, B. (2020) 'Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis', *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, Vol. 6, No. 20, pp.25–35.

Selinker, L. (2015) 'Interlanguage', *Error Analysis*, Vol. 2, No. 3, pp.31–54.

Shu, X., Wang, J., Shen, X. and Qu, A. (2017) 'Word segmentation in Chinese language processing', *Statistics and its Interface*, Vol. 10, No. 2, pp.165–173.

Sutton, C. and McCallum, A. (2012) 'An introduction to conditional random fields', *Foundations and Trends® in Machine Learning*, Vol. 4, No. 4, pp.267–373.

Topali, P., Ortega-Arranz, A., Rodríguez-Triana, M.J., Er, E., Khalil, M. and Akçapınar, G. (2025) 'Designing human-centered learning analytics and artificial intelligence in education solutions: a systematic literature review', *Behaviour & Information Technology*, Vol. 44, No. 5, pp.1071–1098.

Tsai, P-S. and Chu, W-H. (2017) 'The use of discourse markers among mandarin Chinese teachers, and Chinese as a second language and Chinese as a foreign language learners', *Applied Linguistics*, Vol. 38, No. 5, pp.638–665.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, Vol. 30, No. 1, pp.5998–6008.

Wang, Y., Wang, Y., Dang, K., Liu, J. and Liu, Z. (2021) 'A comprehensive survey of grammatical error correction', *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 12, No. 5, pp.1–51.

Xie, Z., Sato, I. and Sugiyama, M. (2020) 'Stable weight decay regularization', *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119, No. 2, pp.10534–10543.

Zhang, B. (2023) 'Design principles and functionality of Chinese interlanguage corpora: a case study of the HSK dynamic composition corpus 2.0', *Learner Corpora: Construction and Explorations in Chinese and Related Languages*, Vol. 15, No. 3, pp.11–32.

Zhang, H-P., Liu, Q., Cheng, X., Zhang, H. and Yu, H-K. (2003) 'Chinese lexical analysis using hierarchical hidden Markov model', *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Vol. 17, No. 4, pp.63–70.

Notes

At the authors' request, the funding information was removed from the published version of this paper in September 2026.