International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-7

https://doi.org/10.70917/ijcisim-2026-0049
Article

A Study on Automated Recognition of Grammatical Errors in English Language Acquisition Based on BERT Modeling

Dongmei Li *

Inner Mongolia University of Technology, Hohhot 010051, Inner Mongolia, China; mabel67@126.com

Abstract: Due to the internationalization of English, many non-native English learners have developed incorrect pragmatic habits. Traditional manual correction of English pragmatic habits is both inefficient and difficult to adapt to large-scale teaching. This paper proposes a multi-task learning grammar error recognition framework based on the combination of BERT, BiLSTM, and CRF structures. In multi-task learning, the BERT language model is used for feature extraction of grammar errors, combined with BiLSTM for sequence learning and CRF for global optimal feature extraction. The three-layer structure model trained through multi-task learning can effectively identify various types of grammatical errors, achieving an average F1 score of 88.0% in evaluation metrics. The model demonstrates excellent performance across different types of grammatical error recognition, with high accuracy and robustness. Through multi-task learning, the grammatical error recognition model exhibits strong generalization capabilities, providing intelligent and personalized technical support for English education and marking a new step toward the intelligentization of future educational practices.

Keywords: english language acquisition; grammatical error identification; BERT; BiLSTM; CRF; multi-task learning

1. Introduction

As a global language, English has become increasingly important as a medium of communication and a tool for interaction with the deepening development of economic globalization. English proficiency has also emerged as a key indicator of an individual's overall literacy and capabilities. However, English grammar remains one of the most challenging aspects of learning and mastering the language. If learners are unable to identify grammatical errors in their own English texts and correct them promptly during the learning process, this can lead to illogical language expression [1-3]. This can even affect the fluency and efficiency of communication, making it more difficult to avoid communication difficulties and weakening the confidence of English learners in their ability to communicate [4-5]. For classrooms with a large volume of teacher corrections or online classrooms requiring remote instruction, the efficiency, timeliness, and accuracy of corrections are particularly lacking, necessitating the introduction of automated grammar error identification and correction methods [6-7].

Reference [8] constructed a computational neural network model based on recurrent neural networks, knowledge, and neural machine translation for English grammar correction, achieving a correction accuracy rate of 82.69% in College English Test Band 4 and Band 6 writing. Reference [9] developed an algorithm for English grammar correction supported by a classification model, primarily improving classification accuracy and the specificity of recognition rules to enhance the speed of grammar correction. Literature [10] established a star cluster oscillation optimization dense recurrent neural network model. Specifically, text data was numerically transformed, processed through various preprocessing techniques, and feature extraction was performed using term frequency-inverse document frequency to detect English grammar errors. Literature [11] utilizes a long short-term memory network to construct an English verb grammar corpus under the support of a recurrent neural network. A word embedding model is incorporated to encode the target text, thereby enabling the detection of grammatical errors. Reference [12] considers the semantic and syntactic information of sentences, sorts words by part

of speech, and performs numerical transformations on word positions using word embedding-one-hot encoding vector representations. It combines error induction techniques with a grammar classifier created using LSTM for classification and training, thereby achieving the detection of English grammatical errors.

In recent years, deep learning, especially pre-trained language models, has shown significant advantages in automatic grammar error detection in natural language processing (NLP) [13-14]. Among these, BERT possesses excellent contextual semantic understanding capabilities, making it a critical tool for grammar error detection. Compared to traditional rule-based grammar detection, BERT can leverage large-scale corpora for pre-training, enabling the network to learn language structures and patterns through data, resulting in greater flexibility and accuracy [15]. Additionally, Reference [16] incorporates a mixed attention module into the Transformer model, collects features through parallel dilated convolutions, optimizes the BERT model using term frequency-inverse document frequency, integrates the Transformer model and BERT model to design a grammar database rewriting model for handwriting recognition optimization, and finally combines English translation principles and grammar features to achieve an English translation grammar error detection system. Reference [17] integrates Seq2Seq, Transformer models, and BERT models to detect and correct grammatical errors in English essays, with the BERT model enhancing the generalization capability of the entire detection method. Reference [18] applies bidirectional long short-term memory (LSTM) networks, naive Bayes, and N-gram models to perform English grammatical error detection, error classification, error localization, and constructs a grammar error generation model and evaluation corpus, thereby forming a grammar error detection framework for machine translation models, significantly improving recognition accuracy. BERT combined with BiLSTM and CRF models can simultaneously consider grammar errors related to word order, redundancy, and omissions, thereby improving the efficiency and accuracy of grammar error detection. In summary, the use of BERT's grammar correction model not only reduces the workload of teachers but also provides students with more timely individual feedback, which is extremely beneficial for promoting learners' language proficiency and making education smarter and more humanized.

The core idea of this study is to leverage the strong semantic understanding characteristics of the BERT model and the advantages of BiLSTM+CRF in syntactic information mining to achieve a multi-task learning identification model for grammatical errors. First, BERT is used as the model's foundation for text preprocessing, extracting semantic information from the text. BERT's understanding of textual semantic information is utilized to extract contextual information from the text. Then, the BiLSTM model is employed to expand long-range dependencies within sentences, enabling the model to more effectively analyze complex grammatical errors. Finally, CRF is used for global optimization to achieve the analysis and identification of grammatical errors.

This experimental framework uses multi-task learning to simultaneously train grammar error recognition and other grammar tasks, thereby enhancing the model's robustness and generalization capabilities. Specifically, it helps the model extract an understanding of grammar rules, enabling more accurate identification of grammar errors, ensuring the accuracy of automatic grammar error recognition results, and improving the efficiency of automatic grammar error recognition to promote grammar-based intelligent assistance in English teaching.

2. Research Methods

2.1. Dataset Selection

The corpus used in this project is the Cambridge Learner Corpus published by Cambridge University Press. The statistical data of the corpus is shown in Table 1. This corpus consists of English writing samples collected from over 160 countries worldwide, used for educational purposes such as Cambridge writing courses, course exams, and classroom instruction, with a total word count exceeding 50 million words. The scale and content of the data alone are not its sole value. The corpus's comprehensive error annotations and detailed grammar error types constitute its most distinctive and valuable features. After reorganization and filtering, the corpus was categorized into six levels from A1 to C2 according to the Common European Framework of Reference for Languages (CEFR). An initial experimental corpus of 200,000 sentences was obtained through stratified sampling. During the annotation process, we assembled a team of professional linguists to implement a double-blind cross-checking and third-party arbitration system. We defined four-tiered error types, error locations, error-involved structures, and error solutions, ensuring that each grammatical error is accurately and specifically described.

Table 1. Feature of Cambridge Learner Corpus Datasets.

Feature item	Specific description
Data scale	More than 50 million words

Language proficiency level	A1-C2 (CEFR Standard)
Text type	Exam compositions, classroom assignments, daily exercises, etc
Error type quantity	77 types of grammar errors
Annotation method	Error type, location, and modification suggestions
Text source country	More than 160 countries and regions
Coverage period	From 1993 to present

In terms of text preparation, after performing necessary text segmentation and syntactic analysis, text noise (such as extraneous characters other than punctuation marks) is removed based on the patterns of syntactic errors, while punctuation marks with syntactic significance are retained in the original data. An automatic detection program is then developed to ensure that no ambiguity occurs in the data. In terms of dataset construction, the authors proposed two expansion strategies—synonym replacement and sentence structure permutation—to expand the dataset. Additionally, they manually constructed parallel texts containing typical errors as contrastive data based on rules. As a result, the dataset now contains 400,000 sentences, with half being original data and half being expanded data, significantly enhancing the model's generalization capabilities.

2.2. Model Architecture

First, we combine BERT with BiLSTM-CRF in our model, where BERT serves as the semantic understanding module and BiLSTM-CRF as the sequence labeling module. BERT first encodes the text for semantic understanding, obtaining the encoding results, while enhancing the grammatical error extraction module. Then, it uses a bidirectional LSTM to learn the contextual information of grammatical errors from the feature layer, and finally employs a conditional random field (CRF) for global optimization of the labeling. This enables our model to fully identify long-distance and local grammatical error features, while allowing timely adjustments to different features to enhance feature information.

Specifically, the model input first undergoes BERT encoding. In the BERT layer, WordPiece performs WordPiece segmentation on the sentence, and then 12 Transformer encoders encode the sentence into a 768-dimensional vector, which is a feature vector containing lexical-level information and sentence context information. This feature vector is then passed to BiLSTM, which enhances the model's modeling of context through forward and backward bidirectional propagation, making the model sensitive to errors when modeling grammatical structures. Finally, the CRF layer considers the transition probabilities between labels and performs global optimal labeling on the entire sequence, making the label sequence of the entire output result more regular.

In terms of model optimization, a hierarchical learning rate is used, with a smaller learning rate (2e-5) set for the BERT layer to retain pre-trained knowledge, and a larger learning rate (1e-3) set for the BiLSTM and CRF layers to accelerate convergence. During training, a dynamic batching mechanism is used to divide samples into groups based on the length of the input sequence, which both accelerates training and saves space. Additionally, to accommodate the need for grammatical error identification, a weighting factor is added to the loss function, assigning different levels of attention to different types of grammatical errors to help the model make more accurate judgments on critical errors.

The advantage of this structure lies in the fact that BERT's pre-training knowledge endows the model with a wealth of linguistic features, enabling it to better handle complex contexts and ambiguous recognition. The bidirectional nature of BiLSTM provides the model with more grammatical dependency information, enhancing its ability to detect local grammatical errors. The global constraints of CRF prevent the generation of unreasonable labels, thereby improving model performance [19-21]. This structure not only improves recognition accuracy but also demonstrates excellent practical generalization capabilities, achieving good detection performance for various types of grammatical error patterns.

2.3. Training Methods

While learning tasks such as part-of-speech tagging and dependency parsing, this paper proposes a multi-task joint training method to learn the task of grammar error detection. Multi-task joint learning can not only effectively mine the potential knowledge transfer between different tasks, but also improve the generalization ability of the model [22]. In the experiment, this paper uses a multi-task loss with adjustable learning weights, as shown in the following formula.

$$L = \sum_{i=1}^{N} \alpha_i L_i \tag{1}$$

In the equation, the loss value of the i th task is represented by L_i, and the corresponding α_i is the

weight coefficient.

We use the performance of each subtask on the validation set as the basis for weight updates, thereby dynamically adjusting the weights of different task optimization objectives during model training. During model training, we train on a server equipped with an NVIDIA Tesla V100 GPU, with 128GB of memory and the PyTorch 1.8.0 deep learning runtime framework. Other hyperparameters in the network training include a learning rate of 2e-5, a learning rate warm-up to 10% of the total training iterations during the initial training phase, a random dropout rate of 0.1, and an L2 regularization rate of 0.01. During model training, gradient accumulation is added to increase the batch size to 128 to enhance training stability.

During training, we used a task rotation method to balance the training tasks, i.e., alternating task optimization in each round to prevent the model from favoring any single task. For efficiency, we divided the samples into multiple groups based on sequence length, implementing dynamic batch size training to ensure GPU resource utilization while reducing memory consumption. We designed the training scheme by combining an early stopping mechanism based on the best performance over five consecutive periods and adversarial word vector perturbations (i.e., robust training), enabling the model to perform well on the main task. This also allowed the model to better understand and interpret linguistically sensitive language rules, technically enabling grammatical error correction within sentences and achieving feasible conditions for practical application.

2.4. Evaluation Indicators

In order to comprehensively reflect the performance of the BERT syntax error automatic identification model and better evaluate its applicability to different tasks and conditions, it is necessary to comprehensively use three commonly used performance indicators: precision, recall, and F1-score. By using mathematical calculations to quantify the model's identification results, we can more intuitively see how the model identifies syntax errors. Precision refers to the proportion of errors detected by the model that are actually errors. The formula is as follows.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

where TP is the number of true positives and FP is the number of false positives. Recall focuses on the proportion of actual errors identified by the model, and is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Here, FN is the number of false negatives. The F1 value combining these two metrics is defined as the harmonic mean of the two, which can be expressed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

The higher the F1 value, the better the balance between accuracy and coverage achieved by the model.

3. Research Results and Discussion

3.1. Model Performance Indicators

Based on the model constructed above, corresponding evaluation indicators were selected. In the experiment, grammatical errors were classified into four categories: word order, redundancy, omission, and choice. The accuracy, recall, and F1 values of the model were calculated for each category. The specific results are shown in Table 2. To more intuitively reflect the differences in performance across different grammatical errors, the bar chart shown in Figure 1 was used in the experiment to represent the performance of each error category across the three evaluation indicators.

Table 2. Evaluation indicators for different types of errors.

Error type	Precision/%	Recall/%	*F*1 value/%	Sample quantity
Word order error	92.5	89.8	91.1	2583
Redundancy error	88.7	85.4	87.0	1976
Missing error	86.3	83.9	85.1	3124

Selection error	90.2	87.6	88.9	2845
Average value	89.4	86.7	88.0	10528

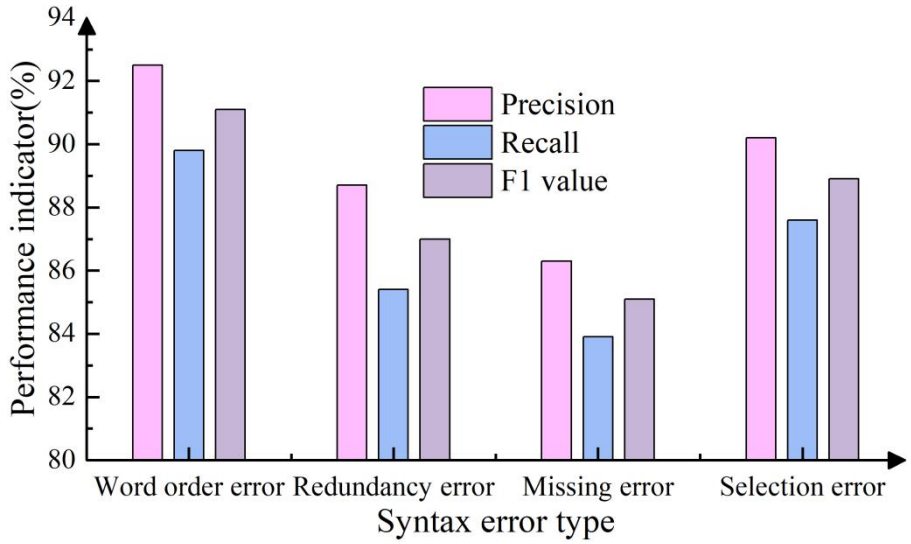

Figure 1. Different types of grammatical errors.

As shown in the figure and metrics, the model achieves an F1 score of 91.1% in deletion-type error detection, indicating that the model has a certain advantage in handling certain types of errors, particularly those involving structural sequence issues. This advantage is also reflected in the model's use of deep neural networks to establish contextual relationships and the BiLSTM module for sequence extraction and modeling. The model also performs well in detecting duplicate and omission errors, with F1 scores of 87.0% and 85.1%, respectively. This indicates that the model has a certain ability to distinguish between repeated and mismatched words. Missing errors are relatively weaker compared to the other two categories, but the accuracy rate remains high. However, the model shows some limitations when handling complex language reasoning errors involving long-distance dependencies. Looking at the average F1 score of 88.0%, this value also indicates that the model's overall judgment is relatively balanced and reasonable, making it suitable for providing reliable error correction suggestions in actual English learning and teaching. We can also observe from the actual metric evaluations that accuracy generally exceeds recall, indicating that the model adopts a conservative strategy to some extent when handling errors, sacrificing coverage but avoiding misjudgments. This may be a positive and effective outcome in practical teaching interventions.

3.2. Comparison of Different Models

The English grammar error detection system developed based on the BERT model achieved the most significant progress in testing, particularly due to the notable advantages of the BERT-BiLSTM-CRF model after integrating multi-task learning. In comparative experiments on the Cambridge Learner Corpus, this model not only outperformed baseline methods in terms of recognition accuracy but also demonstrated unmatched advantages in training speed compared to other models. This paper compares the recognition results of single-task BERT models, BiLSTM models, and the multi-task learning model proposed in this paper for different types of grammatical errors, focusing on aspects such as accuracy, speed, and practical results. The comparison results are shown in Table 3. Additionally, grammatical error types are categorized into word order errors, tense errors, article errors, preposition errors, and other errors (Error1–Error5), with the F1 score as the metric. The F1 scores of different models for various grammatical error types are shown in Figure 2.

Table 3. Performance comparison of different models.

Model	Accuracy/%	Recall/%	*F*1/%	Training time (h)	Reasoning speed (ms/sentence)
Basic BERT	82.3	79.6	80.9	24.5	156
BiLSTM	78.5	75.8	77.1	18.2	42
BERT+BiLSTM	85.7	83.2	54.4	28.3	178
This article	91.4	89.8	90.6	22.1	145

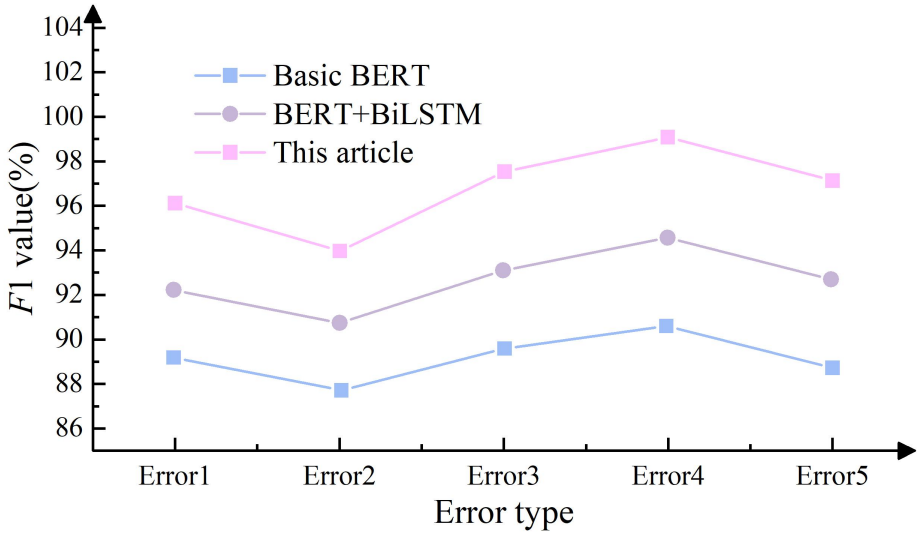

Figure 2. The *F*1 values on different error types.

The experimental results validate the effectiveness of the model proposed in this paper, with an F1 score of 90.6% for overall accuracy, which is approximately 10% higher than the original BERT model. There has also been further improvement in training speed, reducing training time by approximately 10%, and inference speed has been effectively enhanced. Looking at the recognition results for different types of grammatical errors, the F1 score for word order errors is 92.3%, significantly higher than that for tense errors (84.8%) and article errors (86.0%). The high accuracy is attributed to the fact that the BERT pre-trained model has a highly abstract understanding of syntactic information through contextual relationships, while the BiLSTM layer can capture sequence information. The combination of these two types of information enables more effective resolution of word order errors. The model also achieves high F1 scores for tense errors and article errors, which are more difficult to detect, significantly outperforming models based on traditional methods. This is due to the effective knowledge transfer within the multi-task learning framework introduced in this study, enabling the model to better identify such errors.

The model designed in this paper demonstrates good generalization performance when tested on English articles of varying difficulty levels for English learners, making it applicable to practical English teaching scenarios. Building upon the model, this paper achieves improved computational efficiency and enhanced accuracy in identifying grammatical errors by combining multi-task model learning with model improvements, providing a reliable and efficient model to help English learners improve the quality of their English writing.

4. Conclusion and Outlook

This paper proposes a multi-task learning model based on the BERT model, combined with BiLSTM and CRF, for the automatic identification of grammatical errors generated by learners during English learning. The model was experimentally tested on a large corpus of learner data, and the results indicate that the model achieves relatively excellent automatic identification results, with an average F1 score of 88.0%. When handling different types of grammatical errors (word order errors, redundant errors, and missing errors), the model achieves better automatic identification performance compared to single-task training. Additionally, due to the introduction of multi-task learning, the trained model demonstrates stronger noise resistance and better generalization capabilities compared to single-task training results.

In this paper, BERT can perform semantic understanding of text, BiLSTM can model the sequence of text, and through the CRF layer, joint global optimization is performed to better model grammatical errors. This helps teachers achieve automated grading and improve grading efficiency, as well as assists students in self-directed learning, promoting personalized learning. Additionally, the model achieves accuracy rates of 87.0% and 88.9% for redundancy and selection errors, respectively, ensuring its effectiveness in terms of generalization capability from an experimental perspective.

In the long term, although the model performs well in identifying grammatical errors, there are still some issues to be addressed in identifying more complex grammatical structures. Further improvements can be made by restructuring the model, expanding the data, and enhancing the identification rate for

complex errors. Additionally, multilingual adaptation and the promotion of applications in different languages can be explored.

References

1. Ajaj, I. E. (2022). Investigating the difficulties of learning English grammar and suggested methods to overcome them. *Journal of Tikrit University for Humanities*, 29(6), 45-58.
2. Wang, Y., Wang, Y., Dang, K., Liu, J., & Liu, Z. (2021). A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5), 1-51.
3. Aghajani, M., & Zoghipour, M. (2018). The comparative effect of online self-correction, peer-correction, and teacher correction in descriptive writing tasks on intermediate EFL learners' grammar knowledge the prospect of mobile assisted language learning (MALL). *International Journal of Applied Linguistics and English Literature*, 7(3), 14-22.
4. Akbar, H., Hakeem, Z., & Ahmad, S. (2022). English Grammar Rules as an External Barrier Faced by EFL Students for Effective Communication. *Global Language Review, VII*, 418-428.
5. Tiana, D. M., Jimmi, J., & Lestari, R. (2023). The effect of grammar mastery and self-esteem towards students' speaking skill. *Scope: Journal of English Language Teaching*, 7(2), 157-164.
6. Hashemifardnia, A., Namaziandost, E., & Sepehri, M. (2019). The effectiveness of giving grade, corrective feedback, and corrective feedback-plus-giving grade on grammatical accuracy. *International Journal of Research Studies in Language Learning*, 8(1), 15-27.
7. Gong, Z. (2024). English Grammar Auto-Correction Robot based on Grammatical Error Generation Model. *Scalable Computing: Practice and Experience*, 25(6), 5688-5700.
8. Wu, X. (2022). A computational neural network model for college English grammar correction. *Computational Intelligence and Neuroscience*, 2022(1), 9592200.
9. Zhou, S., & Liu, W. (2021). English grammar error correction algorithm based on classification model. *Complexity*, 2021(1), 6687337.
10. Wang, B. (2025). Optimization of English Grammar Error Automatic Detection Algorithm Based on Natural Language Processing. *International Journal of High Speed Electronics and Systems*, 2540270.
11. He, Z. (2021). English grammar error detection using recurrent neural networks. *Scientific Programming*, 2021(1), 7058723.
12. Agarwal, N., Wani, M. A., & Bours, P. (2020). Lex-pos feature-based grammar error detection system for the English language. *Electronics*, 9(10), 1686.
13. Wang, X., & Zhong, W. (2022). Research and implementation of English grammar check and error correction based on Deep Learning. *Scientific Programming*, 2022(1), 4082082.
14. Luhtaru, A., Korotkova, E., & Fishel, M. (2024, March). No error left behind: Multilingual grammatical error correction with pre-trained translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1209-1222).
15. Vijaya Prakash, R., Sai Teja, M., Deepthi, G., Namratha, C., Nikhil Sai, D., & Manish Raj, P. (2022). Model to Detect and Correct the Grammatical Error in a Sentence Using Pre-trained BERT. In *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2021* (pp. 251-259). *Singapore: Springer Nature Singapore.*
16. Qing, Y. (2024). Design and Application of Automatic English Translation Grammar Error Detection System based on BERT Machine Vision. *Scalable Computing: Practice and Experience*, 25(3), 2088-2102.
17. Zhu, J., Shi, X., & Zhang, S. (2021). Machine Learning-Based Grammar Error Detection Method in English Composition. *Scientific programming*, 2021(1), 4213791.
18. Chen, H. (2022). Identification of Grammatical Errors of English Language Based on Intelligent Translational Model. *Mobile Information Systems*, 2022(1), 4472190.
19. Yin, X., Huang, Y., Zhou, B., Li, A., Lan, L., & Jia, Y. (2019). Deep entity linking via eliminating semantic ambiguity with BERT. *IEEE Access*, 7, 169434-169445.
20. Chernyshov, A. (2019). BiLSTM-based approach to the natural language text dependencies analysis. *Information and Innovations*, 14(1), 44-47.
21. Liu, Y., Li, G., & Zhang, X. (2020, November). Semi-Markov CRF model based on stacked neural Bi-LSTM for sequence labeling. In *2020 IEEE 3rd international conference of safe production and informatization (IICSPI)* (pp. 19-23). *IEEE.*
22. Thung, K. H., & Wee, C. Y. (2018). A brief review on multi-task learning. *Multimedia Tools and Applications*, 77(22), 29705-29725.