Research Article

An Automatic Error Correction Method for English Composition Grammar Based on Multilayer Perceptron

Juan Wang [iD] **and Feng Gu**

School of Foreign Languages, Chaohu University, Hefei 238000, Anhui, China

Correspondence should be addressed to Juan Wang; 057063@chu.edu.cn

In order to improve the timeliness of English grammar error correction and the recall rate of English grammar error correction, this paper proposes an automatic error correction method for English composition grammar based on a multilayer perceptron. On the basis of preprocessing the English composition corpus data, this paper extracts the grammatical features in the English composition corpus and constructs a grammatical feature set. We take the feature set as the input information of the multilayer perceptron and realize feature classification through network learning and training. The grammatical error items in the English composition are detected according to the similarity, and the error correction is completed by setting the penalty parameter and reducing the deviation parameter. The experimental results show that the syntax error detection time of this method is less than 6 minutes, the recall rate is higher than 90%, and the detection error rate is lower than 6%. The method improves the timeliness of grammatical error correction and improves the efficiency of error correction.

1. Introduction

English, as the universal language in the world, plays a vital role in the development of globalization. English grammar is a series of language rules that are systematically summarized after the study of the English language [1, 2]. The essence of English grammar lies in mastering the use of language. However, for domestic English learners, grammar is often a difficult point in their English learning. English grammar is a set of language rules that have been systematically summarized after studying English. The essence of English grammar is the use of language. However, English learners often encounter grammar problems when learning English. The importance of grammar comes from the benefits of mastering it. The first and most closely related to us is to solve the problems of reading and writing [3]. However, due to the limited teaching resources, learners often cannot get correct revision opinions and examples in time for the grammar problems they encounter. If there is a grammatical error correction method, which enables learners to point out grammatical errors in the process of using English to communicate and read and write in a timely mannerand give corresponding feedback suggestions, it can greatly reduce learners' learning difficulty and significantly improve their English proficiency.

Therefore, a series of grammar correction methods have been designed. For example, a method for correcting English compositions based on the network platform and the mobile network platform is designed in reference [4]. Through three rounds of action research, this paper finds out the differences in the process of writing reviews and the effectiveness of improving students' English writing. Furthermore, it discusses the practical methods of composition error correction in the network environment and the mobile network environment. In reference [5], a method of grammar error correction based on deep learning technology is designed. This method starts with requirement analysis, first introduces the deep learning technology model based on seq2seq and corpus, then analyzes the grammar error correction model based on seq2seq, and finally introduces the architecture design of the grammar error correction algorithm model and the operation framework and main principles of

core modules. This method shows that more and more attention has been paid to the application of artificial intelligence in grammar error correction. The formation of this method can not only effectively reduce the teacher's homework but also promote students' autonomous learning. A Chinese grammar error correction method based on the transformer enhancement architecture is designed in Ref [6]. This method regards the task of grammar error correction as a translation task. The transformer model based on the multiattention mechanism is used as the error correction model.

However, in practical application, it is found that the traditional grammar error correction method is applied to English composition, but it has some problems, such as long grammar error detection time, low text recall rate, and low grammar error detection rate. To solve this problem, this study designs a new automatic error correction method for English composition grammar based on the multilayer perceptron.

2. Design of the Automatic Error Correction Method for English Composition Grammar

2.1. English Composition Corpus Data Pretreatment.
In order to improve the validity of the corpus, numerical and normalized processing should be performed on the collected corpus data before extracting grammatical features. The English composition corpus selected in this article is Stanford Tokenizer 1.

2.1.1. Numerical Processing.
As grammatical data are nonnumerical attributes, it is difficult to carry out distance calculation, so it is necessary to convert the attributes of all dimensions of grammatical data into numerical values. The frequency presented by attributes of different dimensions is used to replace the initial attributes to realize numerical transformation so as to prevent the unequal distance between the values of the same attributes in the transformation process, resulting in errors in the classification process [7].

2.1.2. Normalized Processing.
There is a large gap between values of different dimensions in the grammatical data. In order to use data of different dimensions more effectively, normalized processing should be carried out on data of different dimensions. The normalization process is as follows:

$$X' = \frac{X - N_{\min}}{N_{\max} - N_{\min}} \times 100, \qquad (1)$$

where $N_{\max}$ and $N_{\min}$ represent the highest and lowest data in a dimension, respectively, and X represents the English composition corpus data to be normalized.

2.2. Extracting Grammatical Features from the English Composition Corpus.
In this study, the TF-IDF algorithm was used to extract grammatical features from the preprocessed English composition corpus data to form a sample

set of grammatical features and then the grammatical features were extracted by weight of text proximity and word occurrence frequency. The TF-IDF algorithm can be said to be a statistical algorithm that uses a keyword to evaluate the importance of an article or a document. As the frequency of occurrence in the corpus decreases inversely proportional, the TF-IDF algorithm is cited by major search engine platforms and is also used as a metric or rating basis for evaluating the relevance of keywords. Where the weight of words can be obtained by the product of IDF and TF. On this basis, the best grammar of words can be extracted. The operation process is as follows:

$$\omega_{s,b} = IDF \times TF, \qquad (2)$$

where b represents the English composition document number, s represents a certain word, and ω represents the weight of the word.

The task of IDF is to improve the criticality of words that appear less frequently and the difference in texts [8], and its calculation formula is as follows:

$$IDF = \lg \frac{M}{m_i}, \qquad (3)$$

where m_i represents the number of words i in all documents and M represents the number of documents in the English composition text.

TF stands for feature frequency, and its expression is as follows:

$$TF = \frac{\lg\left(S_F\left(s_l, d\right) + 1\right)}{\lg S}, \qquad (4)$$

where, after document processing, S represents the total number of words in the document b and $S_F\left(s_k, d\right)$ represents the number of words s in the document b.

Combined with the above content, the grammatical feature set of the English composition corpus is constructed as follows:

$$J = \frac{M \times (TF + IDF)\omega_{s,b}}{S}. \qquad (5)$$

With the support of the corpus feature database, it is generally necessary to go through several steps of sentence breaking, word splitting, part-of-speech recognition, and sentence analysis when correcting a complete English composition text.

2.2.1. Sentence Interruption.
Sentence breaking is to cut a large section of theEnglish text into separate sentences one by one. For example, punctuation marks can be used to complete sentence breaking, but such a sentence breaking is very prone to error [9]. For example, by means of the full stop in English, when a sentence is broken, the computer cannot tell the full stop is an abbreviated "." or the actual period is used to terminate a sentence. To this end, the University of Pennsylvania developed the Natural Language Toolkit (NLTK) to assist in the sentence breaking step. The tool is widely used in artificial intelligence, information retrieval, and machine learning.

2.2.2. Break Up the Words.

Words are the basic unit of English. After the completion of clauses, the segmentation of words in each sentence is a crucial step to complete English learning with the help of machine learning [10]. Although the composition of English sentences between the vocabularies is generally separated by space and a convenient sentence; however, two words such as "New York" and two words such as "it's" are separated by space, which is difficult to accomplish the task of word segmentation perfectly. For this reason, researchers have also proposed a series of reasonably usable word segmentation models based on word segmentation rules and regular expressions.

At the same time, the NLTK supports users to customize the dataset and also has a good word segmentation function through the customization of word segmentation scenes. In this paper, the NLTK is used for word segmentation.

2.2.3. Part-of-Speech Recognition and Sentence Analysis.

Due to the differences in the language environment, most English words have multiple parts of speech when translated into Chinese, and each part of speech may correspond to multiple meanings [11]. After separating the words in the sentence one by one, in order to analyze the sentence and achieve the goal of grammar error correction, we should first identify the part of speech of each word in the existence of the sentence as a verb, noun, or other part of speech. Based on the result of part-of-speech recognition, sentence analysis can be realized by combining the position of the word in the sentence and the rules of English grammar. The correctness of part-of-speech analysis plays a decisive role in the realization of grammar error correction. Currently, the more commonly used tools for part-of-speech analysis include Stanford POS tagger based on the statistical probability method, OpenNLP language processing kit developed by Apache, Trigram "n" Tags based on HMM and viterbi algorithms, and spaCy based on *Python* and Cython.

2.3. Feature Classification Based on Multilayer Perceptron.

Based on the above extraction of grammatical features from the English composition corpus, this study classifies the features based on the multilayer perceptron.

The neural network (NN) is one of the greatest inventions in the field of intelligent computing, which imitates neurons of the human brain and is mainly used to solve classification problems and make data predictions [12]. The concept of neural networks was first proposed in 1943. Since then, neural networks have developed rapidly and different types of neural networks have been proposed one after another. The MLP NN (multilayer perceptron NN) neural network is one of the most famous classifiers in this field and has been widely studied and applied [13].

As shown in Figure 1, the multilayer perceptron contains many network layers that are sequentially connected and the full length of the network layer is called the depth of the model.

In Figure 1, the first layer is the input layer, the last layer is the output layer, and the other network layers are hidden layers. If the perceptron network has no hidden layer, it is a
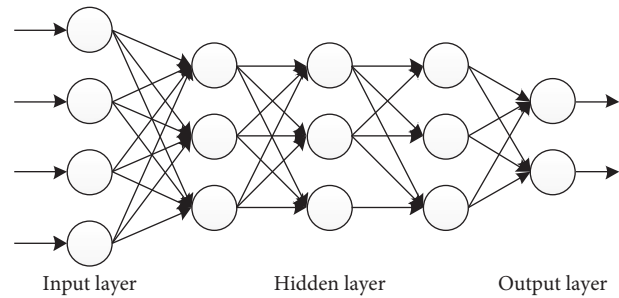

Input layer   Hidden layer   Output layer

FIGURE 1: Structure diagram of the multilayer perceptron.

single-layer perceptron network. The single-layer perceptron network cannot classify linear nonseparable problems, so the number of hidden layers needs to be increased. The multilayer perceptron with multiple hidden layers can better simulate the structure of human brain neurons, extract the features of the input data layer by layer, and achieve more complex classification with better classification performance [14].

Each node of the perceptron network is called a neuron. The neuron usually receives multiple different inputs and generates a single output to transmit to multiple neurons at the next layer of the network. Its structural model is shown in Figure 2.

In Figure 2, x represents the signals transmitted from the first layer of the perceptron network, and there are n signals in total. ω represents the weight of the connection between a layer of neurons on the perceptron network and this neuron; b represents the offset value of this neuron; f stands for activation function; y represents the output of this neuron, and the relation between input and output is as follows:

$$y = f\left(b + \sum_{i=1}^{n}(x_i\omega_i)\right). \qquad (6)$$

There are many options for activation functions. Sigmoid activation function is usually used in the early shallow neural network. Due to the possibility of gradient disappearance and other problems in the deep neural network, ReLU activation function is usually used at present and its expression is as follows:

$$f(z) = \max\{0, z\}, \qquad (7)$$

where z represents the activation function input and $f(z)$ represents the activation function output. In the network output layer, softmax activation function is often used to classify the output and tanH and other activation functions with better performance can also be used to normalize.

Network training is to make the network output as close to the ideal value as possible by changing each connection weight in the network, which is usually called label [15]. The error is obtained by comparing the output with the ideal value, which is generally called the loss function. The loss function takes the partial derivative of each weight to obtain the update gradient of the weight and updates the
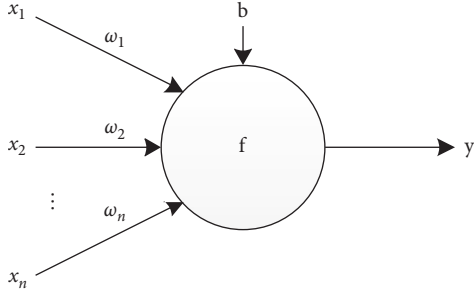
FIGURE 2: Neuronal structural model.

weight according to the proportion of the learning rate. However, updating weights for each sample training will slow down the network training speed. The simple solution is to sum the loss function of a batch of training samples and then update the weight, which is the batch gradient descent method. The parameters to be learned can be expressed as follows:

$$\delta = f\left(e_1 + \omega_2\left(e_2 + \omega_1\right)\right), \tag{8}$$

where e_1 and e_2 represent error vectors and ω_1 and ω_2 represent weight coefficient matrices.

Based on the above analysis, the grammatical feature set J in the English composition corpus obtained in Section 2.2 is classified and processed, which is sent to the multilayer perceptron as the input data. The classification result Q obtained is as follows:

$$Q = J\frac{\sqrt{\left(\delta \times M\right)^2}}{S \times f}. \tag{9}$$

For the classification set Q of English composition grammatical features obtained, the basic K-means is used to cluster it and the classification result is $Q \longrightarrow \{Q_1, Q_2, \ldots, Q_k\}$.

2.4. English Composition Grammatical Feature Error Detection.

According to the classification result Q_k of grammatical features in English composition, the evaluation function can be used to evaluate its separability between classes and cohesiveness within classes. The mean square error can be selected as the evaluation index, and its calculation formula is expressed as follows:

$$MSE = \sqrt{\sum_{j=1}^{M} \frac{d\left(s_j - c_k\right)^2}{M - 1}}, \tag{10}$$

where M represents the number of samples, d represents the processing dimension, and c_k represents the clustering center of the k-th class and the component of the s_j sample. Calculate the similarity between all samples in the grammatical feature classification set and each cluster center, and divide each sample into the class where the cluster center with the highest similarity is located. The similarity calculation formula is expressed as follows:

$$Sim = \frac{1}{\sqrt{\sum_{p=1}^{d} \left(Q_k^p - c_k^p\right)^2}}, \tag{11}$$

where the pth component of sample Q_k in the grammatical feature classification set is represented by Q_k^p. The pth component of the cluster center c_k is denoted by c_k^p. Based on the clustering results obtained, the evaluation index was calculated, and then the clustering center was updated. The above process was repeated, and then, the evaluation index obtained from the clustering results of the previous round was compared with the similarity index obtained from this round of clustering. The feature mean square error threshold is set as τ. If the difference between the two similarity indexes is greater than τ, it indicates that there are grammatical errors in the text. Through repeated iterative operations, the set $\{E_1, E_2, \ldots, E_k\}$ of grammatical errors in English composition is obtained.

2.5. English Composition Grammar Error Correction Model Design.

The grammar error set $\{E_1, E_2, \ldots, E_k\}$ obtained above is divided into several subsets $\{e_1', e_2', \ldots, e_k'\}$, and a grammar error correction model is constructed based on the input information. The specific process is shown in Figure 3.

The specific construction process of the model is as follows:

Step 1. Input English composition grammar error subset $\{e_1', e_2', \ldots, e_k'\}$, parameter disturbance quantity ∂, and English composition grammar feature classification set $\{Q_1, Q_2, \ldots, Q_k\}$.

Step 2. On the basis of the error subset, it is projected onto the set of grammatical feature classification of English composition, and the data with high similarity to the error subset are selected to form the subset $\{A_1, A_2, \ldots, A_k\}$ to be corrected.

Step 3. The corrected subset is analyzed, the low-deviation region of the sample is calculated, and the disturbance quantity ∂ and the parameter λ to reduce the multilayer perceptron deviation are selected.

Step 4. Assume that ∂_h and λ_h represent the hth penalty parameter and the deviation reduction parameter, and add these two parameters to the subset to be corrected and output the correction result after correcting the wrong grammar, thus completing the automatic error correction processing of English composition grammar.

3. Experiment and Result Analysis

In order to verify the practical application performance of the proposed method based on the multilayer perceptron, the following experiments are designed.

3.1. The Experiment to Prepare.

The experiment takes an English corpus database as an example and randomly selects some English texts as experimental subjects. There are
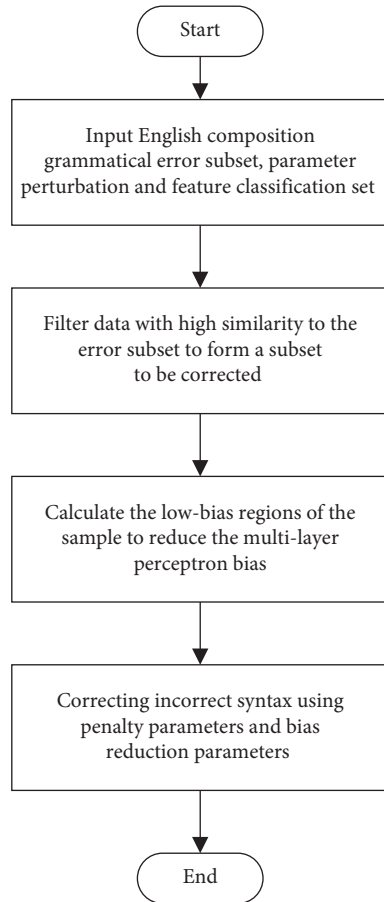
FIGURE 3: English composition grammar error correction process.

10 types of grammatical mistranslations, including verb-object errors, abbreviations errors, rhetorical errors, voice errors, word order errors, lexical errors, missing words, subject-predicate errors, and multiwords. The experimental corpus was randomly divided into 5 datasets (A, B, C, D, and E), and the basic information of each data set was as follows.

The total number of samples in dataset A is 8000, and the number of grammatical mistranslations is 360, including the categories of grammatical mistranslations including abbreviations, multiwords, voice errors, verb errors, missing words, subject-predicate errors, and lexical errors.

The total number of samples in dataset B is 1000, and the number of grammatical mistranslations is 70. The mistranslations include subject-predicate errors, verb errors, word order, missing words, and voice errors.

The total sample number of dataset C is 7000, and the sample number of grammatical mistranslations is 480. The mistranslations include missing words, multiwords, word order, lexical errors, singular and plural noun errors, subject-verb errors, verb errors, and abbreviation errors.

The total number of samples in dataset D is 3000, and the number of grammatical mistranslations is 200, including the mistranslations of verb errors, abbreviation errors, rhetorical errors, voice errors, word order, lexical errors, missing words, and multiple words.

The total sample number of the E dataset is 2000, and the sample number of grammatical mistranslations is 500. The mistranslations include subject-predicate errors, verb errors, abbreviation errors, rhetoric errors, voice errors, missing words, multiwords, word order, lexical errors, and singular or plural noun errors.

In order to avoid the uniformity of experimental results, the methods in reference [4] (action research of English composition error correction method based on the network platform and mobile network platform) and reference [5] (analysis of the grammar error correction algorithm based on deep learning technology) were taken as the control group and the grammatical error detection time, text recall rate, and grammatical error detection rate were taken as indexes, respectively, to complete the performance verification together with the paper.

3.1.1. Syntax Error Detection Time. The syntax error detection time is the length of time spent detecting syntax errors. The error correction time of different methods is calculated by the computer. The shorter the time for checking syntax errors, the better the timeliness of the method.

3.1.2. Text Recall Ratio. The text recall ratio refers to the ratio of the amount of relevant information found in the English corpus to the total amount. Recall is a measure of the success of a retrieval method in detecting errors in a corpus. That is, the percentage of errors detected in English versus all related sentences. The higher the recall, the better the effect of the method.

3.1.3. Syntax Error Detection Rate. The syntax error detection rate refers to the probability that syntax is detected. The percentage of grammatical sentences that were detected incorrectly versus all the sentences used in the experiment. The lower the error rate, the better the performance of the method.

3.2. Results and Analysis. First, the syntax error detection time of different methods is tested, and the results are shown in Figure 4.

By analyzing the results shown in Figure 4, it can be seen that with the increase in the number of experiments, the detection time of grammatical errors of different methods also changes accordingly. The syntax error detection time of the method in reference [4] varies between 9 min and 12 min, while that of the method in reference [5] varies between 6 min and 9 min. In contrast, the syntax error detection time of the method in this paper is less and its value is less than 6 min. It can be seen that the proposed method has higher timeliness in detecting grammatical errors. Because this paper preextracts the grammatical features in the English composition corpus. The syntactic feature set we constructed shortens the time for subsequent machine error correction using a multilayer perceptron.
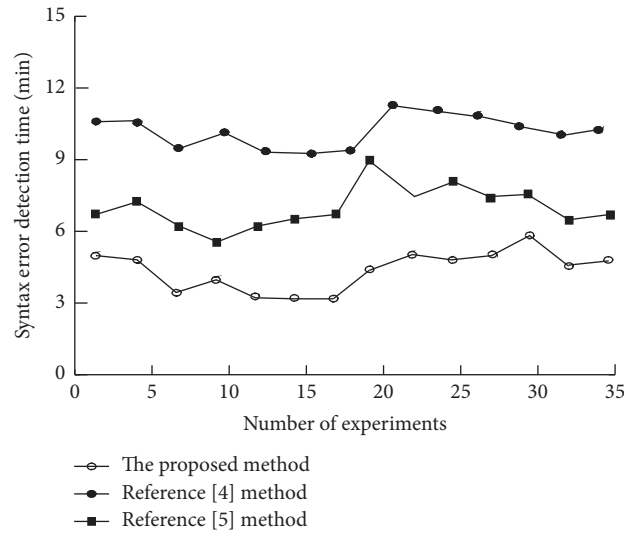
FIGURE 4: Comparison of the syntax error detection time between different methods.

TABLE 1: Recall rate of the English composition text by different methods.

The experimental time (min)	English composition text recall rate (%)		
	The proposed method	Reference [4] method	Reference [5] method
5	96.2	85.0	77.3
10	92.5	80.1	77.1
15	94.7	84.2	74.9
20	90.3	87.3	80.5
25	93.0	83.9	78.3
30	95.7	88.0	75.5

TABLE 2: Comparison of syntax error detection rates between different methods.

Amount of information to be processed (GB)	Syntax error detection rate (%)		
	The proposed method	Reference [4] method	Reference [5] method
200	3.7	16.1	20.4
400	4.3	15.0	20.9
600	4.9	18.0	19.5
800	3.6	15.2	21.1
1000	5.5	17.8	20.3

On this basis, at the same time, the recall rate of the English composition text is tested with different methods, and the results are shown in Table 1.

By observing the results in Table 1, it can be seen that when the experiment time is 10 minutes, the recall rate of the English composition text in this method is 92.5%, 80.1%, and 77.1%, respectively, of the English composition text in reference [4] and reference [5]. When the experiment time is 20 minutes, the recall rate of the English composition text of this method is 90.3%, 87.3% of the English composition text recall rate in reference [4] method, and 80.5% of the English composition text recall rate in reference [5] method. When the experiment time is 30 minutes, the recall rate of theEnglish composition text in this method is 95.7%, 88.0%, and 75.5%,respectively, in reference [4] method and reference [5] method. In contrast, the recall rate of the English composition text in this method is higher in the same period

of time, indicating that this method is more reliable. In this paper, the extracted grammatical features are input into the multilayer perceptron and the features are classified through network learning and training. Thus, the recall rate of grammar composition text is improved.

Finally, the syntax error detection rates of different methods are compared,and the results obtained are shown in Table 2.

According to the results shown in Table 2, when the amount of information to be processed is 200 GB, the detection rate of grammatical errors in this method is 3.7%, 16.1%, and 20.4%, respectively, of the method in reference [4]. When the amount of information to be processed is 600 GB, the detection rate of the proposed method is 4.9%, 18.0%, and 19.5%, respectively, of the method in reference [4] and reference [5]. When the amount of information to be processed is 1000 GB, the syntax error detection rate of the

proposed method is 5.5%, 17.8%, and 20.3%, respectively, of the method in reference [4]. In contrast, when the amount of information to be processed is the same, the syntax error detection rate of the proposed method is lower, indicating that the proposed method is more effective. Because we take the feature set as the input information of the multilayer perceptron. Then, the grammatical error items in the English composition are detected according to the similarity, and the error is corrected by setting the penalty parameter and reducing the deviation parameter. Such an approach can reduce the error detection rate of the subgram.

4. Conclusions

In China, English education has been given great importance by the educational circles and students' English level has been improving steadily. If students cannot get accurate and detailed feedback in time, the learning efficiency will be greatly reduced and the learning effect will be affected. In order to improve the timeliness of grammatical error detection and the recall rate and reduce the rate of grammatical error detection, this study extracts grammatical features from the English composition data. On the basis of feature classification, grammatical errors in English compositions are detected according to similarity and then corrected by setting penalty parameters and reducing deviation parameters. In the experiment, this method has obtained a good application effect and can lay the foundation for the enhancement English composition writing level. The experimental results show that the syntax error detection time of this method is less than 6 minutes, the recall rate is higher than 90%, and the detection error rate is lower than 6%. To a great extent, the grammatical error correction technology is limited to the size of the parallel corpus that can be obtained. In the future research, we can combine the other subfields of NLP to produce a high-quality and large-scale parallel corpus.

Data Availability

The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] L. Hu, Y. Tang, X. Wu, and J. Zeng, "Considering optimization of English grammar error correction based on neural network," *Neural Computing & Applications*, vol. 34, no. 5, pp. 3323–3335, 2021.

[2] S. Ailani, A. Dalvi, and I. Siddavatam, "Grammatical error correction (GEC): research approaches till now," *International Journal of Computer Application*, vol. 178, no. 40, pp. 1–3, 2019.

[3] J. Lichtarge, C. Alberti, S. Kumar, N. Shazeer, N. Parmar, and S. Tong, "Corpora generation for grammatical error correction," *CoRR*, vol. 19, no. 4, pp. 587–590, 2019.

[4] J. Li, "Action research of English composition error correction method based on network platform and mobile network platform," *Journal of Anhui Vocational College of Electrontcs and Information Technology*, vol. 17, no. 4, pp. 46–50, 2018.

[5] Y. Jing, "Analysis of grammar error correction algorithm based on deep learning technology," *Information & Technology*, vol. 44, no. 9, p. 143, 2020.

[6] C. Wang, L. Yang, Y. Wang, Y. Du, and E. Yang, "Chinese grammatical error correction method based on transformer enhanced architecture," *Journal of Chinese Information Processing*, vol. 34, no. 6, pp. 106–114, 2020.

[7] S. Liu and J. Wei, "Multilayer perceptron classification algorithm based on spectral clustering and simultaneous two sample representation," *Computer Science*, vol. 46, no. S2, pp. 194–198, 2019.

[8] A. Zabolotskikh, A. Zabolotskikh, T. Dugina, and D. Tavberidze, "Creating individual learning paths in the Moodle plugin for undergraduate students to study English grammar," *Education and Information Technologies*, vol. 26, no. 1, pp. 617–637, 2020.

[9] Z. Zhao and H. Wang, "MaskGEC: improving neural grammatical error correction via dynamic masking," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 1226–1233, 2020.

[10] J. Lang and F. Qiu, "Hybrid data assimilation based on multilayer perceptron," *Journal of Physics: Conference Series*, vol. 1948, no. 1, Article ID 012161, 2021.

[11] S. Kiyono, J. Suzuki, T. Mizumoto, and K. Inui, "Massive exploration of pseudo data for grammatical error correction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 7, pp. 2134–2145, 2020.

[12] F. Arias del Campo, M. C. Guevara Neri, O. O. Vergara Villegas, V. G. Cruz Sánchez, H. d. J. Ochoa Domínguez, and V. García Jiménez, "Auto-adaptive multilayer perceptron for univariate time series classification," *Expert Systems with Applications*, vol. 181, no. 14, Article ID 115147, 2021.

[13] M. Wanni, "Research on English grammar recognition system based on combination of genetic algorithm and KNN algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 6, pp. 7291–7302, 2020.

[14] Y.-O. Hong, "A study on the change of Chinese grammar by English grammar," *The Journal of Chinese Language, Literature and Translation*, vol. 45, no. 3, pp. 121–144, 2019.

[15] B. Wang, K. Hirota, K. Hirota, C. Liu, Y. Dai, and Z. Jia, "An approach to NMT Re-ranking using sequence-labeling for grammatical error correction," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 24, no. 4, pp. 557–567, 2020.