

Research Article

Research on Computer Intelligent Proofreading System for English Translation Based on Deep Learning

Nannan Ma 🕩

School of Foreign Languages, Shenyang Aerospace University, Shenyang 110001, China

Correspondence should be addressed to Nannan Ma; manannan@stu.sau.edu.cn

Received 18 May 2022; Accepted 8 June 2022; Published 21 June 2022

Academic Editor: Chia-Huei Wu

Copyright © 2022 Nannan Ma. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

English, as a global language, is widely used in daily life, which leads to higher requirements for people's language proficiency. Grammar, as an essential part of English, plays a significant role in language acquisition. Although the needs for English learning have become greater, the resources of English teaching are quite limited. Under this circumstance, it is imperative to apply information technology to alleviate the burden of lacking sufficient educational resources. The intelligent grammar correction system is in special needs. In this study therefore, the deep learning-based CNN and RNN bidirectional propagation model were utilized to construct a feedback filtering function, screen the suggested text, and select the high-quality suggested text for the retraining of the error correction model. The proofreading optimization algorithm and feedback filtering algorithm were used in practice to implement each functional module. The core function service module was designed with a cluster deployment scheme with high system performance and response speed to ensure system stability and availability. Experimental results demonstrate that the correct rate gradually increased with increasing the number of tests and finally stabilized at about 86%. The overall average correcting rate obtained from the present model was verified to be superior to that from the GRU model and the MGB model.

1. Introduction

With the sustainable development of the society, global communications are becoming more frequent than ever. As a result, English translation proofreading technique can enable people to proofread errors in the texts of diverse fields concerning government reports, media announcements, academic papers, etc. Misunderstandings caused by language are the major hurdle in international communications and negotiations. Consequently, the automatic text proofreading has been an urgent problem to solve. The intelligent translation proofreading system can not only reduce the burden of teachers but also allow English learners to quickly get feedback on their English grammar learning, which may motivate the enthusiasm of students for independent learning.

Previously, the processing of text, audio and video, and images was limited by computer processing capabilities. With the development of technology, scientific breakthroughs have shed some light on solving this problem to develop natural language processing. New algorithm theories and technologies such as statistical machine learning and deep learning have been introduced into natural language processing, making it possible to implement an intelligent English translation proofreading system [1, 2].

There are certain commonalities between syntax proofreading and machine translation. For example, they both first need encode the sequence to obtain the semantic vector and then decode the vector into another sequence. Therefore, the grammar proofreading task proposed in this study can be also regarded as a small language translation task. According to the characteristics of English grammar error proofreading task, a neural machine translation-based grammar proofreading method is proposed in this study to further analyze the problems of existing English grammar proofreading methods. To make full use of this new model, the following aspects need to be considered. First, the training data must be annotated with other factors, which would involve running automated tools on the corpus, since manually annotated corpora of the texts are few and expensive

to produce [3]. Second, a word matching mechanism needs to be built for all sentences in parallel training, and a word-layer alignment is used for word alignment. Furthermore, each mapping step constitutes an integral part of the overall model, which, from a training perspective, requires learning translations and generating tables from wordaligned parallel corpora and defining scoring methods that help users navigate between fuzzy mappings. The core of this algorithm is the convolutional sequence-to-sequence model [4]. In the training process of the model, the feedback filtering algorithm is used to train the wrong sentence generation model to increase the scale of the training corpus. In view of the role of the transfer learning method in the translation of small languages, the grammar proofreading model is initialized by the parameters of the translation model to improve the performance of model invariant checking. In the model application, sentences are first proofread for word errors and then input into the model, together with reordering the results of the cluster search.

Deep learning technology is one of the most popular directions in the field of machine learning in recent years. Compared with conventional machine learning algorithms, deep learning technology has lower requirements for data preprocessing and feature engineering and has stronger model generalization ability. Deep learning has made breakthroughs in sentiment analysis, summary generation, speech recognition, image processing, and other fields, so this study will focus on solving the problem of English translation proofreading by deep learning technology. Based on the proposed error correction algorithm, the experimental verification of the algorithm model is performed. Taking into account the application of the algorithm, an English translation proofreading system is constructed, and the error correction algorithm and feedback filtering algorithm have been used in the preliminary study [5]. The system module division and distributed architecture design have been performed to improve the system's horizontal expansion capability and module upgrade capability.

In this study, we proposed a deep learning-based feedback filtering optimization algorithm. The evaluation indicators such as proofreading recall rate (CR), proofreading precision rate (CP), and proofreading F1 value as well as hyperparameters (A and B values) are introduced into the English translation training model. Furthermore, the proposed English translation proofreading model is compared with the common GRU and MGB models to analyze their advantages and disadvantages.

2. Related Work

Similar to machine translation, the translation proofreading system also need to learn many lessons from itself. At the beginning of the development of translation proofreading, common rule-based grammar checker determines the syntactical correctness of a sentence, which is mostly used in word processors and compilers. However, it is basically impossible to fully check errors by exhausting grammar rules due to the complexity and change of English grammar. Later, a statistical machine translation technology was introduced into the translation proofreading system, based on data-driven and statistical methods. This technology has received increasing attention because of its potential in promoting the development of translation proofreading [6].

Yuan et al. proposed a relatively novel solution, i.e., a syntax tree for grammar proofreading [7]. However, it does not work effectively when the sentence construct the syntactic tree correctly. Yang H and Yang Y suggested triggering rules to detect grammatical errors, based on rules with weights. Written in Java and C++, it only provides external interfaces without open source [8]. Tian insisted that the DFA-based fault-tolerant matching algorithm of English words was not suitable for Chinese characters [9]. The Chinese fuzzy matching faces two important problems: the large-scale Chinese thesaurus and the difficulty in determining the length of the fuzzy matching window. Using "meta distance" to replace the concept of edit distance of English letters, a fast Chinese fuzzy matching method was proposed to proofread Chinese sentences entered by the Wubi input method. The recall rate was approximately 79%, together with the recall accuracy rate of about 64% and the proofreading accuracy rate of about 75%. Ren selected a fixedlength window of a sentence for fuzzy matching, so as to construct a fuzzy word graph and solve the shortest path for word error proofreading by a shortest path problem based on segmentation method [10].

The conditional random field (CRF) is a model based on conditional probability distribution, which has been used to correct some grammatical and semantic errors of specific words. Chen J Y et al. thought that sentences were composed of elements in the trigram model, which can be converted into a weighted finite state transition machine. This is decoded by beam search and A* algorithm. Accordingly, the wrong sentence is a grammar model unit, which dynamically divides the sentence into all possible grammar model combinations. Thus, an actual optimal path was obtained for the proofreading result [11].

In fact, proofreading of papers has achieved good results among students. The challenge is the construction of the state transition machine and the decoding process of the sentence. Lei and Xing proposed a context-based proofreading method by selecting a word in the confusion set that fits well with the context for proofreading [12]. The combined use of binary and ternary connections was found to be better than using a single model. Through the manual annotated sentence test, the recall rate of error detection was about 74.9%, together with the error detection accuracy rate of about 75.8% and the correction rate of about 70%. Huo constructed a lattice word graph of single-word words and confusing word sets that do not exist in the vocabulary by means of word tagging and segmentation for sentences. The optimal path was addressed to obtain a candidate sentence using the forward algorithm. Based on the statistical machine translation model, the wrong sentence is regarded as the source language, and the correct sentence is regarded as the target language. Finally, the proofreading candidate results were synthesized by the above two combined models [13].

In addition, support vector machine (SVM) was a very powerful classification method that is used to score the proofreading candidates and select the final result. Thus, SVM can significantly improve the proofreading accuracy, but the effect of statistical translation model was invisible. Zhang and Cao considered each word as a potential typo, the confusion set vectors of each word in the sentence were combined to form a candidate matrix, and the Chinese characters that can be formed into words between adjacent vectors were combined into words, by searching in the matrix. This method is only for Pinyin input method proofreading, and the search space of the optimal path is very large, which is prone to false alarms. The test sentences come from some artificially generated errors and some real errors in online news. The recall rate of errors is 87.2%, the accuracy of error reporting is 75.0%, and the correct rate of first candidate correction is 59.9% [14].

Due to the huge flexibility and uncertainty of natural language itself, English, as a representative, is of numerous vocabulary, complex grammar, and wide range of usage scenarios, which makes it more difficult for computers to automatically check errors and proofread. Natural language consists of words, including grammatical and semantic knowledge. Translation proofreading needs to extract grammatical information of sentences, correct grammar, and eliminate ambiguity. Due to the large data difference between Chinese and English, the performance of information extraction is inevitably affected. Moreover, it is quite difficult to obtain some satisfactory results by the existing machine translation technology. Therefore, this article will study deep learning technology and use it to solve the problem of English translation and proofreading.

3. Deep Learning-Related Technologies

3.1. RNN Model. Recurrent neural network (RNN) is a classic deep learning model different from traditional neural networks, since the former integrates the previously learned status through a recurrent approach. RNN usually performed together with convolutional neural network (CNN) in terms of fine-tuning. It has been demonstrated that CNN showed great potential in the ImageNet image recognition competition, and AlphaGo was also built on the basis of the product neural network [15]. RNN was a neural network with memory, which can record the network data information from the start time to the current time, i.e., the output of the neuron was determined by the current input and historical input data.

When RNN bidirectional modeling processed sequence information, such as dialogue generation scenarios, only knowing the current and previous input was not sufficient. If you can know the subsequent information, it can help to generate better answer sentences. The traditional RNN model was flawed in this regard, because its memory behavior was only to store the information before the current moment rather than the later information, as displayed Figure 1.

The bidirectional RNN was improved for the situation where the context cannot be obtained. The bidirectional recurrent neural network fed the input into the hidden layer of the neural network in two ways, forward and backward,



FIGURE 1: RNN bidirectional structure model diagram.

respectively, and connected to the output layer at the same time, so as to jointly determine the output [16, 17]. The calculation formula is shown in

$$h_t = f(h_{t-1}, x_t).$$
 (1)

After acquiring the hidden state at each moment, the semantic vector was obtained. The semantic vector contained the basic information of the input sequence and extracted important features, defined as below:

$$C = q(h_1, h_2, h_3, \dots, h_t).$$
(2)

The decoding stage, which can be understood as the inverse process of the encoding stage, predicted the next output word according to the semantic vector and the generated output sequence, and the process is shown in

$$y_t = \arg \max P(y_t) = \prod_t p(y_u | y_1, y_2, \dots, y_{t-1}).$$
 (3)

The decoding stage also has some calculation rules for the hidden state, and the hidden state of the decoding stage is shown in

$$S'_{t} = f(s_{t-1}, t_{t-1}, C).$$
(4)

The long-term dependency problem referred to the problem of gradient disappearance and gradient explosion when using conventional RNN for training when the input sequence was long. The reverse decoding method based on the hidden state increased the amount of input information during decoding [18, 19].

3.2. Convolutional Neural Network (CNN) Model. A convolutional neural network is a deep neural network with local connections, and its structural model is shown in Figure 2.

Convolutional neural networks were mainly used in the fields of image processing, and they were far superior to traditional machine learning algorithm models in tasks such as image classification and recognition. In recent years, they have also been widely used in the fields of text and speech processing [20]. Convolutional layers generally consisted of multiple convolutional filters, different convolutional filters



FIGURE 2: Convolutional neural network model.

can learn different features from the input, and deep convolutional layers can further learn complex features based on shallow learning. The purpose of the convolution operation was to learn local features from the input, and the calculation formula is shown in

$$c_i = f() \sum_{i=0}^{h-1} w_i + b.$$
 (5)

A specific feature extraction function was achieved by learning and optimizing the weight parameters of the filter by back-propagation. After multiple convolution operations, the entire sentence was processed, and the feature matrix was obtained as below:

$$C = [c_1, c_2, \cdots, c_{h+1}].$$
 (6)

After the input features were obtained through the convolutional layer, it was pooled and processed to further reduce the network training parameters and control the degree of overfitting to a certain extent the convolutional layer. The maximum value in the pooling window was selected as the sampling value, and all the values in the pooling window were added and averaged, and the average value was used as the sampling value, defined as below:

$$\widehat{C} = \sum c_{t-1}^n + w_b. \tag{7}$$

Each neuron in the fully connected layer was connected to all neurons in the previous layer, and the features of the output of the previous layer were synthesized and flattened into a one-dimensional output.

Due to the local operation of the convolutional layer and the dimensionality reduction of the pooling layer, the convolutional neural network had fewer weight parameters than the fully connected feedforward neural network, which is more suitable for building large-scale complex network structures with fast training speed. Also, it was quite suitable



FIGURE 3: Word graph node data structure.

TABLE 1: Data set division.

	Training set	Validation set	Test set
Proportion	75%	15%	10%
Number of samples	22500	4500	3000

for parallel training, which makes it an attractive deep neural network structure.

4. English Translation Proofreading Algorithm Optimization

4.1. Shortest Path Word Segmentation Optimization. On the basis of the internal structure and activation function of neural network, we proposed an optimization method using a new type of activation function in training grammar classification models, namely, an adaptive and extensible linear correction unit. The convolutional neural network after updating the activation function was compared with the traditional neural network from several perspectives, such as the stability of the function, the verification accuracy rate during the training process, and the training speed [21]. The shortest path word segmentation algorithm was divided into two types according to the weight of word graph edges: minimizing the number of words and maximizing sentence probability, as shown in Figure 3.

Since the system handled English automatic word segmentation and named entity recognition under a unified framework, it can be tested for English automatic word segmentation and named entity recognition, respectively. For English automatic word segmentation, a corpus was used in this study as a test set. For each test set, it was classified into two categories: open test and closed test [22]. In order to train and test the performance of the neural network model, the data set was divided into training set, development set, and test set. Each subtask contained 30,000 data samples, and the division ratios are shown in Table 1. The model was trained in one batch using an activation optimization algorithm, and the training data consisted of a batch-sized sample from the bAbI data set. Compared with stochastic gradient descent, independent adaptive learning rates were designed to calculate the first-order moment estimation and second-order moment estimation of the gradient, which is suitable for solving large-scale data and parameters. Accordingly, the model parameters were adjusted through the back-propagation algorithm, based on the loss function value until the loss function converges to the minimum value. The specific model training parameters after parameter tuning are shown in Table 2.

The early stop mechanism was used in model training. When the correct rate in the training set and the validation set remained constant in the set round interval, the training was stopped. This can effectively prevent the occurrence of overfitting and reduce the meaningless training time. The changes in the loss function value and the correct rate as a function of training round are shown in Figures 4 and 5, respectively.

From Figures 4 and 5, it can be seen that the loss value of the validation set of the model on the training set during the whole training process of the model was floating at 250, and the correct rate of the model on the training set and validation set was maintained at about 10%, indicating that there was no activation function. The gradient descent method does not effectively update the network weight parameters, the loss function value cannot be effectively reduced until it reaches the minimum value, and the neural network model cannot learn effective answer information. The reason for this is that there was no sentence attention weight calculation and answer information obtained from the sentence of the information text, but the question sentence was directly used for answer prediction, resulting in inefficient model training and low answer prediction accuracy.

4.2. Feedback Filtering Algorithm and Experimental Verification. Grammar correction was also similar to translation, and the results obtained by the error correction model were not necessarily completely correct, especially for complex grammar errors. These statement suggestions were filtered to identify which were the correct statements to modify and which were the wrong statements. This was related to the process of system relearning and retraining, which belongs to the key module of the system. This screening process was actually a question of which was more credible between the revised sentence given by the system and the revised suggested sentence given by the user and then the sentence needed to be scored, and those with high scores can be considered to be sentences with a higher probability of having no grammatical errors.

The confusion set used in the proofreading algorithm adopted the feedback filtering algorithm confusion set, which disassembled the English sentences into groups of radicals, and encoded these groups to encode English sentences. This English sentence encoding method was characterized by the appearance characteristics of English sentences to a certain extent, and English sentences with the same block were often similar in appearance. Different

TABLE 2: Hyperparameters for model training.

Parameter	Describe	Value
Vector	Length of word vector	55
Input	The size of the input layer	75
Batch	The number of samples in a batch of training	300
Max_epoch	Largest training round	256
Dropout	Dropout probability	0.9
Learning	Learning rate	0.001
L2	L2 regularization parameter	0.001
Max_gradient	Threshold for gradient pruning	11



FIGURE 4: Loss function transformation of the model without activation function.



FIGURE 5: Correct rate transformation of the model without activation function.

from the former, the four-corner code was a coding method for judging the glyph from a macro perspective, by encoding the abstract shapes of the upper-left, lower-left, upper-right, and lower-right corners of an English sentence. The length of the four-corner code was fixed at five digits, and there was an additional code in addition to the four-corner number. In addition to the four-corner code, stroke order coding was also a coding method to describe the glyphs of English sentences, which represented the writing strokes and stroke order of English sentences. In this study, the quadrangle code and stroke order code were used to further filter words with similar glyphs, and the size of the confusion set was simplified. In order to facilitate horizontal comparison with other systems, the evaluation indicators included proofreading recall rate (CR), proofreading precision rate (CP), and proofreading F1 value. Hyperparameters A and B control the size of the discount factor when calculating the word graph node weights and final reordering, respectively. The algorithm tends to select confusing words that are close to the original string as proofreading candidates. By scanning the values of A and B, a combination with the highest F1 was selected. The experimental results are shown in Figures 6 and 7.

As can be seen from Figures 6 and 7, it is best to select about 3.5 for A and about 5 for B.

5. Analysis of Results

In the previous section, the proofreading algorithm optimization of English translation was completed. In this section, the basic GRU model and the MGB model were used to test the proposed model in this study, and the model with the best performance was selected after multiple parameter tuning. The English proofreading module provided two external interfaces, English proofreading and model training. The English proofreading interface completed the function of English proofreading of the input English sentences. The main workflow was to use sentence segmentation and use the trained error correction model to correct errors and then summarize the error correction results for return. The model training interface completed the model retraining and module upgrade functions. First, the model training was completed. After the error correction effect is evaluated, if the error correction effect of the latest model is better than that of the existing model, the model will be replaced to complete the system upgrade. The experimental test is carried out on the test set data, and the test results are shown in Figure 8.

Through experiments, it can be seen that the correct rate of the method proposed in this paper gradually increases after the number of tests and finally stabilizes at about 86%. The overall average correct rate is better than the GRU model and the MGB model. The proofreading knowledge acquisition module was used to establish various knowledge bases required by the automatic proofreading module. These knowledge bases represented the characteristics of a language from different angles and degrees. In order to accurately represent the rules and characteristics of language, it is necessary to collect large-scale and extensive relevant corpora.

When the distribution of unlabeled data and test data is identical, the added unlabeled data will greatly improve the classification performance; otherwise, negative migration will occur, reducing the information extraction performance. Therefore, the approach of our model relies heavily on the



FIGURE 6: Experiment of selecting value A.



FIGURE 7: Selection experiment of numerical value B.



FIGURE 8: Proofreading experiments on the correctness of different models.

distribution of unlabeled data. In the cross-language information extraction, the language gap between the source language and the target language affects the information extraction performance. Meanwhile, proofreading errors generated by machine translation will further degrade the system performance. In this study, we proposed a dualview cross-language information extraction method based on the noise reduction autoencoder. In the reconstruction process of the noise reduction autoencoder, noise was appropriately introduced to improve the robustness and antinoise capability of the information extraction system.

6. Conclusions

In order to make full use of the algorithm model, an English translation proofreading system was designed and implemented, which is convenient for English learners. First, the requirements of the English translation and proofreading system were analyzed to extract the system function points. Then, the modules were divided to clarify the responsibilities of each module. In the optimization algorithm, a feedback filtering algorithm was proposed to filter the user modification suggestions, screen out high-quality suggested texts, and improve the quality of model retraining corpus, thereby improving the effect of English translation and proofreading. In view of the strong dependence of the fusion decoding of a single confusion network on the reference sentence rather than the ordering ability, we proposed to build multiple confusion networks to perform consistent decoding with a rescoring method. After the number of tests, the accuracy rate gradually increased and finally stabilized at about 86%. The overall average accuracy rate was better than that of the GRU model and the MGB model.

For the work that has been done in this study, the effect of English translation and proofreading needs to be improved. More experiments should be carried out on the current algorithm model, the network structure and parameters should be adjusted, and the algorithm model should be improved by referring to the work of other researchers. The English translation proofreading system currently has a relatively simple function and needs to enrich the system functions. The retraining triggering method of the English translation proofreading model is relatively simple at present.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] J. Min, "Research on the application of computer intelligent proofreading system in college English teaching," *Journal of Physics: Conference Series*, vol. 1915, no. 3, article 032078, 2021.
- [2] R. Hu, "Research on intelligent automatic translation system in Chinese and English based on integration technology,"

7

International Journal of Multimedia and Ubiquitous Engineering, vol. 11, no. 12, pp. 115–126, 2016.

- [3] L. Wu and L. Wu, "Research on business English translation framework based on speech recognition and wireless communication," *Mobile Information Systems*, vol. 2021, no. 4, p. 11, 2021.
- [4] C. Shi, "Research on intelligent language translation system based on deep learning algorithm," *Soft Computing*, 2021.
- [5] R. Wang, "Research on intelligent English translation method based on the improved attention mechanism model," *Scientific Programming*, vol. 2021, no. 8, p. 8, 2021.
- [6] B. Ren, "An intelligent Chinese text proofreading method based on deep learning," *Modern Science & Technology of Telecommunications*, vol. 4, no. 2, pp. 66–70, 2017.
- [7] Z. Yuan, C. Jin, and Z. Chen, "Research on language analysis of English translation system based on fuzzy algorithm," *Journal* of Intelligent and Fuzzy Systems, vol. 40, no. 3, pp. 1–9, 2020.
- [8] H. Yang and Y. Yang, "Design of English translation computer intelligent scoring system based on natural language processing," *Journal of Physics: Conference Series*, vol. 1648, no. 2, article 022084, 2020.
- [9] X. Tian, "Research on English translation of Chinese college students based on computer scoring system," *Journal of Physics: Conference Series*, vol. 1992, no. 3, article 032021, 2021.
- [10] B. Ren, "Research on intelligent English teaching system based on the intelligent network system," *Automation & Instrumentation*, vol. 9, no. 3, pp. 4–13, 2016.
- [11] J. Y. Chen, Z. Wang, J. Y. Chen et al., "Design and research on intelligent teaching system based on deep learning," *Computer Science*, vol. 2, no. 21, pp. 61–67, 2019.
- X. Lei and J. Y. Xing, "Intelligent English automatic translation system based on phrase translation combination," *Automation* & *Instrumentation*, vol. 4, no. 8, pp. 58–64, 2018.
- [13] Y. Huo, "Analysis of intelligent evaluation algorithm based on English diagnostic system," *Cluster Computing*, vol. 22, no. S6, pp. 13821–13826, 2019.
- [14] X. Zhang and Z. Cao, "A framework of an intelligent education system for higher education based on deep learning," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 7, p. 233, 2021.
- [15] C. Zhao, Y. Wang, and X. Zhang, "Research and design of intelligent question answering system based on deep learning," *China Computer & Communication*, vol. 7, no. 2, pp. 31–38, 2019.
- [16] L. I. Chun-Sheng, L. U. Peng-Fei, and K. J. Zhang, "Research on mechanism of intelligent question answering system based on sentence similarity computation," *Computer Technology and Development*, vol. 7, no. 2, pp. 73–79, 2018.
- [17] K. Qian and Z. Hou, "Intelligent evaluation of the interior sound quality of electric vehicles," *Applied Acoustics*, vol. 173, article 107684, 2021.
- [18] T. Zhang, "Deep learning classification model for English translation styles introducing attention mechanism," *Mathematical Problems in Engineering*, vol. 2022, 10 pages, 2022.
- [19] Q. Lin, "Intelligent text scoring system based on deep learning," *Journal of Physics Conference Series*, vol. 1486, no. 3, article 032052, 2020.
- [20] H. Tang, "Research on teaching quality evaluation method of network course based on intelligent learning," *International Journal of Continuing Engineering Education and Life-Long Learning*, vol. 30, no. 1, p. 1, 2020.

- [21] H. Li, S. Zhang, J. Shi, and Y. Hu, "Research and design of intelligent learning system based on recommendation technology," *Control & Intelligent Systems*, vol. 47, no. 1, pp. 43–49, 2019.
- [22] J. Wang, "Research on English language accurate translation method based on semantic analysis," *Modern Electronics Technique*, vol. 3, no. 2, pp. 48–53, 2018.