Hindawi
Complexity
Volume 2021, Article ID 5544257, 13 pages
https://doi.org/10.1155/2021/5544257

Research Article

Automatic Integrated Scoring Model for English Composition Oriented to Part-Of-Speech Tagging

Fei Chen

Foreign Language Teaching Department, Teaching Center for General Courses, Chengdu Medical College, Chengdu 610500, China

Correspondence should be addressed to Fei Chen; 1000233@cmc.edu.cn

Received 6 January 2021; Accepted 27 April 2021; Published 5 May 2021

Academic Editor: Wei Wang

Copyright © 2021 Fei Chen. 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.

Part-of-speech tagging for English composition is the basis for automatic correction of English composition. The performance of the part-of-speech tagging system directly affects the performance of the marking and analysis of the correction system. Therefore, this paper proposes an automatic scoring model for English composition based on article part-of-speech tagging. First, use the convolutional neural network to extract the word information from the character level and use this part of the information in the coarse-grained learning layer. Secondly, the word-level vector is introduced, and the residual network is used to establish an information path to integrate the coarse-grained annotation and word vector information. Then, the model relies on the recurrent neural network to extract the overall information of the sequence data to obtain accurate annotation results. Then, the features of the text content are extracted, and the automatic scoring model of English composition is constructed by means of model fusion. Finally, this paper uses the English composition scoring competition data set on the international data mining competition platform Kaggle to verify the effect of the model.

1. Introduction

Grading plays the most central role in the entire teaching process, directly reflecting the current learning situation of the teaching object, indirectly reflecting the learning ability of the teaching object, and reflecting the quality of teaching to a certain extent [1, 2]. In traditional exam-oriented education, grading plays a pivotal role, directly involving talent selection. It is more difficult to be objective in the process of scoring the composition. Composition is a narrative method that expresses the meaning of a theme through words after consideration of people's thinking and language organization. The composition examines a person's language, logic, and other abilities, which is the subjective and concentrated expression of the author's overall thinking ability [3, 4]. It is not difficult to find that the scoring standards and detailed rules are descriptive except for strict and clear standards for word count and style. Therefore, the scoring process also involves the scorers' subjective cognition and understanding of the scoring standards, and even the thinking and identification of the author's own thoughts and language. The subjective factors and descriptive scoring criteria of the scorers make the scoring process relatively ambiguous. Therefore, it is difficult to be relatively objective in composition scoring through manual scoring [5, 6].

In view of the various problems existing in traditional English writing teaching, the English automatic scoring system came into being [7, 8]. Especially in recent years, with the continuous in-depth research of NLP technology, ML, IR technology, etc., some experts and scholars at home and abroad have also applied these technologies to the automatic composition scoring system, making the automatic scoring system gradually powerful and the credit of the score. The degree and validity have been greatly improved. The most representative automatic scoring systems abroad include Project Essay Grade (PEG) [9], Intelligent Essay Assessor (IEA) [10], and E-rater scoring system [11]. The excellent automatic scoring system combined with the function of text error correction [12, 13] can reduce human workload and greatly save human and material resources. The research on

the automatic scoring method of English composition has always been a challenging and constantly improving task.

The performance of the part-of-speech tagging system directly affects the scoring and analysis performance of the correction system. At the same time, part-of-speech tagging will also affect the system's syntactic analysis, spelling error detection, text fluency, and article scoring modules. Therefore, this paper proposes an automatic scoring model for English composition based on article part-of-speech tagging. First, use the convolutional neural network to extract the word information from the character level and use this part of the information in the coarse-grained learning layer. Secondly, the word layer vector is introduced, the residual network is used to establish the information path, and the coarse-grained labeling information is combined with the word vector information. Third, the model relies on the recurrent neural network to extract the overall information of the sequence data to obtain accurate annotation results. Finally, the features of the text content are extracted. Taking into account the timeliness of the feedback results and the accuracy of prediction, this paper chooses to use a simple and efficient Bagging method to integrate the three models (Random Forest, GBDT, and XGBoost) that automatically score English compositions, and each model is multiplied by a certain weight. Then add up to get the final output to realize the construction of the automatic scoring model of English composition.

2. Related Works

Natural Language Processing (NLP) is an important direction in computer science and artificial intelligence. It studies all kinds of theories and methods that can realize the effective communication between humans and computers by natural language. Natural language processing is mainly used in machine translation, public opinion monitoring, automatic summary, opinion extraction, text classification, question answering, text semantic comparison, speech recognition, Chinese OCR, and other aspects. Machine learning is a multidisciplinary interdisciplinary major covering probability theory, statistics, approximate theory, and complex algorithms. It uses computers as a tool and is committed to simulating human learning in real time and divides the existing content into knowledge structures to effectively improve learning efficiency. Automatic English essay grading uses natural language processing techniques to allow a computer system to give appropriate scores for the target essay. Therefore, this paper adopts natural language processing technology and machine-learning technology to study English writing.

Automatic composition scoring is an automated scoring of composition using information technology. There are two typical application scenarios for general composition automatic scoring technology:

(1) Carry out automatic grading work in standardized grade examination as an auxiliary tool for manual grading and give grading suggestions

(2) It acts as a teaching tool in the process of language teaching and provides meaningful appraisal and comments for students' compositions

Its purpose is to solve the various drawbacks of manual scoring mentioned above. The main essay automatic scoring systems or technologies at home and abroad are as follows:

2.1. Project Essay Grader (PEG). PEG [14, 15] extracts some simple and easy-to-extract features of the article to quantify the article, using features such as article length, word length, and punctuation.

The PEG system uses machine-learning methods to complete the scoring process, including two stages of training and scoring. That is the training and application of the model. The training sample of the PEG system consists of 100 to 400 essays scored by experts. Feature extraction is performed on these essays to obtain the weight of each feature to build a model. In the application stage, PEG performs feature extraction on the articles to be evaluated and brings them into the trained regression model to obtain the final score. The PEG system can finally achieve an R-value of 0.87 inconsistency, which can be said to be a good simulation and close to the real scoring result.

However, because PEG uses a large number of simple indirect features to characterize the article, it cannot abstract the article from the more in-depth semantic features of the article, so the PEG system is easy to deceive, as long as it simply satisfies the simple indirection extracted by the PEG system. Features can get a good score on the PEG system.

2.2. Intelligent Essay Assessor (IEA). The IEA system [16, 17] can analyze the specific meaning of words and phrases in the text. Moreover, developers believe that the meaning of an article is largely determined by the words used in the article. That is to say, the article changes the meaning of the article itself through word changes. LSA believes that these two phenomena are generally prominent in articles. One is the synonymous phenomenon where the same meaning is described by different words, and the other is the ambiguity phenomenon where the same word describes different meanings. Therefore, LSA believes that a word has multiple candidate semantic spaces, and the true meaning of the word is difficult to determine.

With the representation of the article, the article can be classified effectively. If an article needs to be scored, then the IEA needs enough articles in the same category that have a score, and the final score is based on the correlation between the article to be scored and the article that has been scored. In certain areas, the accuracy of the IEA score can reach above 0.85.

The IEA system analyses the article from the perspective of semantics, but the final method of expressing the article regards the article as a disordered combination of words, ignoring the textual structure characteristics of the article, the connection between words and sentences, sentences and sentences. The statement is obviously one-sided.

2.3. E-Rater System. E-rater [18, 19] is the first system to be applied to a wide range of standardized grade examinations. The core technology of E-rater includes two directions: artificial intelligence and natural language processing. Artificial intelligence means that the system uses an excellent machine-learning model to simulate artificial scoring, while natural language processing technology provides support for the extraction and analysis of feature variables in the model.

E-rater contains three core modules, syntax, and discourse and analysis modules. Syntactic analysis can extract the structural features of sentences and analyze the grammatical phenomena in the article, such as the analysis of clauses. Discourse analysis is to divide the article through obviously related words and get the discourse structure of the article. Thematic analysis is by characterizing the vocabulary in the article. All these three types of features dialect an article as a feature variable and then train the regression model to get a scoring model. The accuracy of the E-rater can reach an astonishing 0.97.

2.4. Intellimetric Automatic Scoring System. Intellimetric [20, 21] combines the strengths of artificial intelligence, natural language processing, and statistical technology and is a learning machine that can internalize the collective wisdom of expert raters.

The development of the whole system is a process of artificial scoring simulation. Intellimetric tries to restore the various steps of manual scoring to approach the final score on the score. The whole system constructs a feature-screening module, which tries to candidate features during the training process, finally determines the effective features, and makes the existing features more effective and accurate.

The candidate feature set contains more than 300 article-related features, and the screening module performs feature screening among these candidates. The final scoring model can reach an accuracy rate of over 0.97.

2.5. Bayesian Essay Test Scoring System (BESTY). BESTY [22, 23] uses the Bayesian classification model as the basic machine-learning model and grades the composition through classification. BESTY claims to use the most core features of the existing mature automatic scoring system as its own feature variables to abstract articles and finally achieve article scoring, here for the corresponding level classification task.

2.6. Domestic Research on Automatic Scoring of Composition. Bridgeman and Ramineni [24] used a cross-validation method to analyze a total of 320 articles, borrowed from the characteristics of PEG and other systems, and included a large number of shallow language features of articles, constructed an automatic scoring model for English composition, and finally achieved the accuracy rate exceeds 0.84. Roscoe et al. [25] used linear regression to construct a composition scoring system in which Chinese is the second foreign language based on more than 1,000 samples and achieved a score correlation of 0.6. Li et al. [26] conducted related research on automatic composition scoring with semantic analysis as the core. After scoring the semantics of the article, it is found that the semantic score has a correlation of 0.5 with the final manual score of the composition.

3. Automatic Scoring Model for English Composition Based on Part-of-Speech Tagging

3.1. Part-of-Speech Tagging. In the field of natural language processing, part-of-speech tagging technology is a very important part, which can help us obtain the part-of-speech of each word in the sentence so that we can count the part-of-speech characteristics of the text. The syntactic analysis allows us to obtain the syntactic structure of a sentence and analyze the number of clauses, gerund phrases, etc., in the sentence. Stop word filtering technology can help us to remove stop words that are not helpful to the semantic information of the content, to reduce the interference of the text content with real semantic information.

Parts-of-speech tagging is the process of determining the grammatical category of each word in a given sentence, determining its part of speech, and adding a tag [27]. Part-of-speech tagging is a very basic work. It can describe the role of a word in the context. It is the basis of grammatical analysis and semantic analysis. Therefore, part-of-speech tagging is also a very important task. The quality of the tagging results will be directly affect the performance of the entire system. We call the tool used to complete the part-of-speech tagging work the part-of-speech tagger, and the set of tags used for specific tasks is called the tag set.

The part-of-speech tagging for English composition is the basis for automatic correction. The performance of the part-of-speech tagging system directly affects the performance of the marking and analysis of the correction system. It is especially important for grammatical error detection. Because the grammatical rules in grammatical error detection are mainly determined by the part of speech and the word itself, at the same time, part-of-speech tagging will also affect the system's syntactic analysis, spelling error detection, text fluency, and article scoring modules.

As shown in Figure 1, the model first uses a convolutional neural network to extract word information from the character level and uses this part of the information in the coarse-grained learning layer. Then, the word-level vector is introduced, and the residual network is used to establish an information path to integrate the coarse-grained annotation and word vector information. Finally, the model relies on the cyclic neural network to extract the overall information of the sequence data to obtain accurate annotation results.

3.1.1. Word Feature Extraction. When using neural networks to process text data, you first need to digitize or vectorize words. Many network structures map words into a data vector. Among them, the network structures of skip-gram and continuous bag-of-words are simple and efficient.

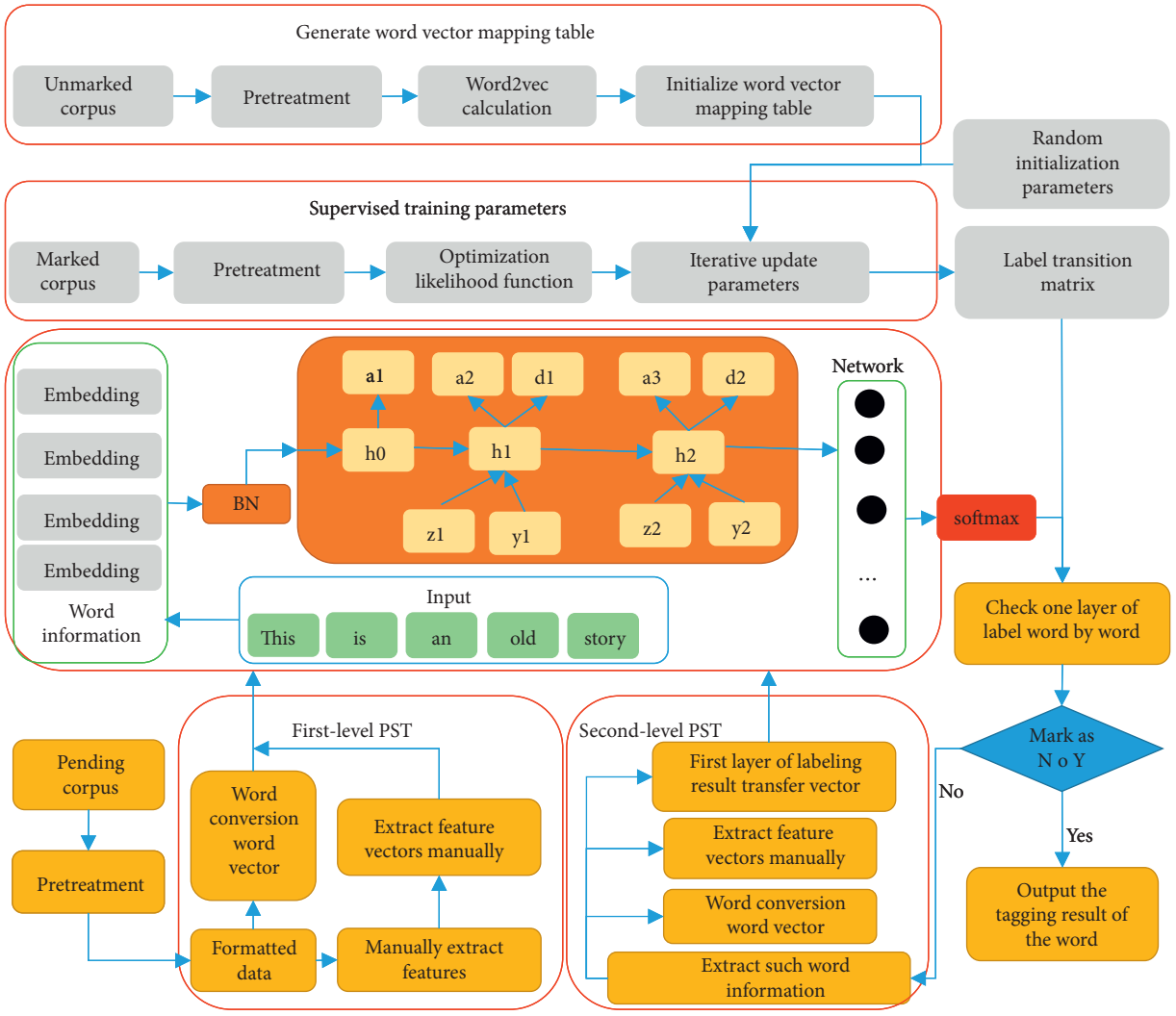


Figure 1: Part-of-speech tagging model based on recurrent neural network.

This paper designs a CRNN to vectorize words from the character level. Because a character is the smallest unit that makes up a word, and the total number of characters is a limited set, representing the word from the character level can fundamentally solve the problem of unregistered words. As shown in Figure 2, taking the input "this" as an example, first, the word is decomposed according to characters, and each character is mapped into a vector. Then combine the vectors of each character to get the word matrix. Finally, through the process of convolution, pooling, and recurrent neural network feature extraction, the final CRNN word vector is obtained.

3.1.2. Coarse Learning. From the perspective of deep learning, the underlying prior knowledge factors that can explain data changes are often shared across two or more tasks; at the same time, because of parameter sharing, the statistical strength of parameters can be greatly improved, and the generalization can be improved.

The labeling model in this paper divides the labeling process into shallow and deep multitask learning processes. First, roughly label the data. Then further divide the labels of the same category. This division method can effectively mark the part-of-speech information of English composition and other corpora. The model in this paper classifies the annotation tags in a coarse-grained manner.

3.1.3. Establishment of Information Channel. The model in this paper divides the labeling into two parts. First, perform rough labeling, and then use the roughly labeled information for fine-grained labeling. In the fine-grained annotation, the original input information is extracted and filtered, and some features are not used in the final fine-grained annotation. Moreover, as the depth of the network increases, the difficulty of training the network increases.

The residual network establishes an information path by setting a threshold function, allowing information to be transmitted across the network layer. Therefore, this article combines the residual network to divide the network level and establish the information path in different network layers. Compared with the direct accumulation of the two layers, the established path plays a major role in the

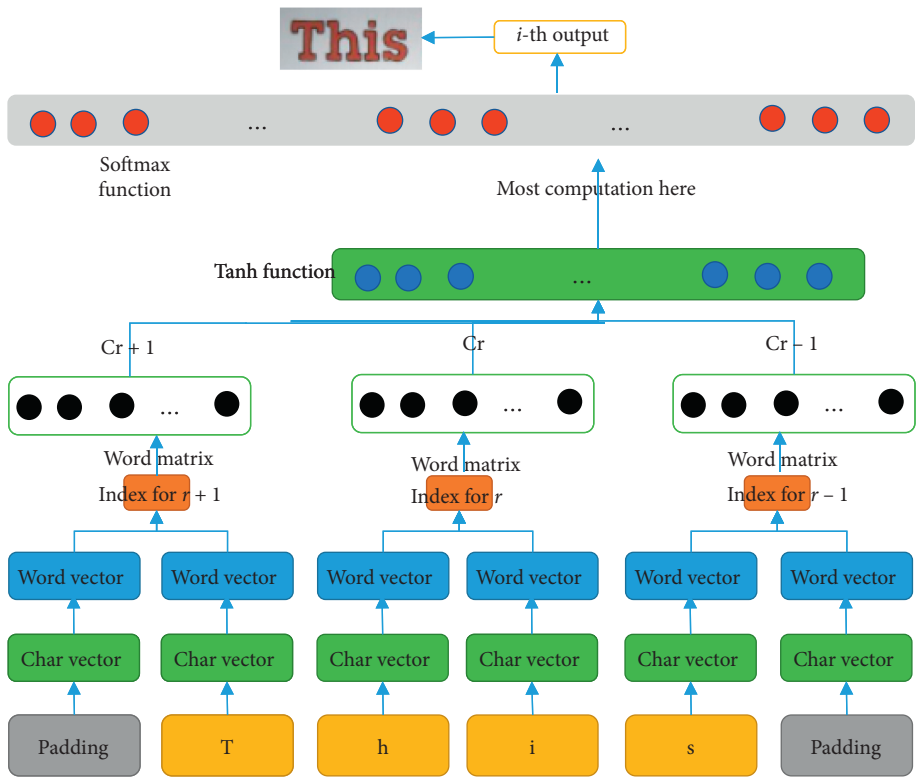


Figure 2: CRNN network structure.

propagation of the gradient, thereby reducing the training complexity of the model.

3.1.4. Batch Normalization. In the training of the deep network, the input of each layer of the network will change the data distribution due to the change of the previous layer of network parameters. This requires that the model must use a small learning rate for network training, and the parameters need to be initialized well. Nevertheless, doing so will make the training process slow and complicated. In the model in this paper, Batch Normalization is introduced to batch normalize data.

In the sequence-labeling model in this article, the input of the second layer of BLSTM consists of three parts:

(1) The output of the first layer of BLSTM

(2) Vector information at the word level of the original data

(3) Character-level vector information extracted by CRNN

When combining these three parts, you need to perform the Batch Normalization operation separately to standardize the data distribution. As shown in Figure 3, the network model uses CRNN to extract character-level vector information from the input sentence and then obtains the rough label information in the rough labeling part. The original word vector and the character extracted by CRNN form the input of the second layer of BLSTM. Compared with direct splicing as input, the model in this paper performs Batch Normalization operations on these three parts, respectively.

3.2. Automatic Scoring Model for English Composition. The overall design of the English composition scoring model is shown in Figure 4. It is mainly composed of four parts, namely the subtitle degree feature generation module, the content text feature generation module, the nontext feature generation module and the machine-learning model prediction module.

3.2.1. Deduction Degree Feature Generation Module. For *i*-th feature item *fi* of text *A*, if *fi* appears *m* times in text *A*, its *AF* value is as follows:

$$AF_i = m. \tag{1}$$

Taking into account the difference in length of articles, in order to facilitate comparison between articles, it is generally necessary to standardize the word frequency:

$$AF_i = \frac{m}{\sum_{i=1} m_i}. \tag{2}$$

AF_F value is the frequency of feature item *fi* appearing in global text GA, namely,

$$AF\_F_i = \lg\left(\frac{GA}{\{A: f_i \in A\}}\right). \tag{3}$$

For the feature item *fi*, the corresponding feature item weight is as follows:

$$w_i = AF_i * AF\_F_i. \tag{4}$$

TF-IDF weight comprehensively considers the distinguishing ability and frequency of feature items.

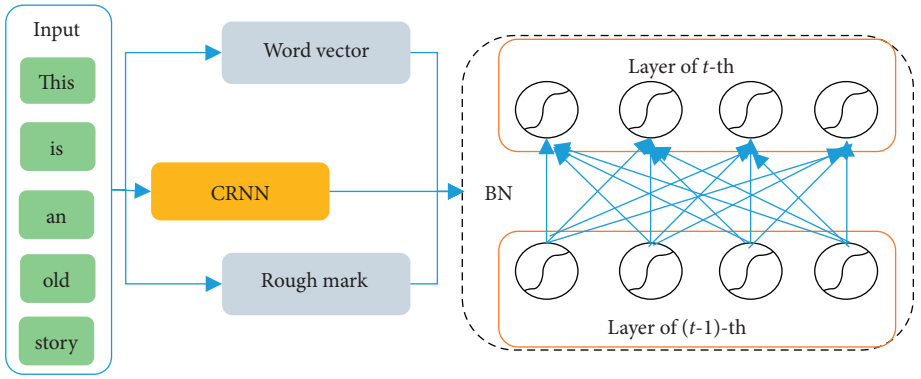


Figure 3: Batch normalization.

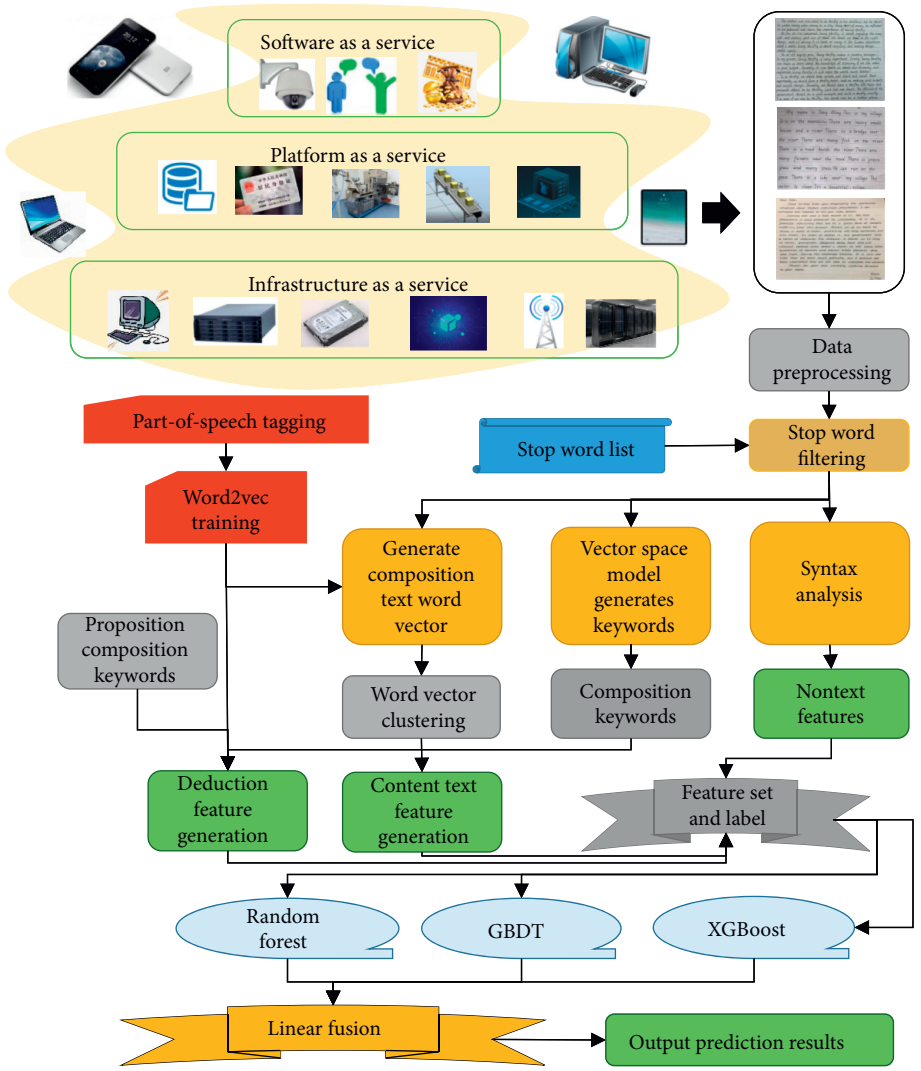


Figure 4: Automatic scoring model for English composition.

3.2.2. Content Text Feature Generation Module. For the generation of content text features, this article first uses the Wikipedia English corpus to train the model on word2vec. Then input the composition text to get the corresponding word vector set of the composition text. Then randomly select the cluster centres. In addition, iteratively calculate the category to which each word belongs and adjust the cluster centre until convergence. After clustering these word vectors, the number of words in the word category after the word vector clustering, the

size of the vocabulary and the distribution of words are calculated as features.

The model is trained on word2veC using the English corpus of Wikipedia and the number of clusters is set to k. The algorithm flow is as follows:

(1) Use the model to generate the word vector of the text, and set the word vector set of the text as $T = [t_1, \ldots, t_n]$, where t_i is the word vector of the text word.

(2) Randomly select c cluster centres $J_1, J_2, \ldots, J_c \in R^n$.

(3) For each $i \in [1, n]$, calculate the category to which t_i belongs:

$$class_i = \min\left\|t_i - J_i\right\|^2. \tag{5}$$

(4) For each $j \in [1, c]$, adjust the cluster centre J_i:

$$J_i = \frac{\sum_i [class_i = j] * t_i}{\sum_i [class_i = j]}. \tag{6}$$

(5) Judge whether the cluster centre does not change anymore, and output $J = [J_1, \ldots J_c]$, otherwise return to step 2).

3.2.3. Nontext Feature Generation Module. For the generation of nontext features, the text attributes of words can be obtained through syntactic analysis, and the number of words in the composition text, the number of text words after removing repeated words, the average length and variance of words, the number of nouns, the number of verbs, and adjectives can be counted. Nontext features are divided into two levels of words and sentences, mainly including lexical features and syntactic features.

3.2.4. Machine Learning Model Prediction Module. Model fusion is a manifestation of ensemble learning, and it is a very common technique for improving performance in various data mining competitions. Usually, the results can be improved in various machine-learning tasks.

Taking into account the timeliness of the feedback results and the accuracy of prediction, this paper chooses to use a simple and efficient Bagging method to integrate the three models that automatically score English compositions. Each model is multiplied by a certain weight. Then accumulated to get the final output. This weight is to select the optimal parameter as its corresponding weight after debugging through the offline test set. Assuming that the random forest predicted value is a, the GBDT model predicted value is b, and the XGBoost model predicted value is c, the final predicted value of the model is as follows:

$$fuse = a * w_1 + b * w_2 + c * w_3. \tag{7}$$

The weights w_1, w_2, and w_3 are obtained through offline test set training, and their sum is one.

4. Results and Discussion

4.1. Data Set. The data set of this research is publicly available on Kaggle, which is a public platform for machine-learning competitions. We can register an account for free to download the training data of the competitions held by it. This data set is the English composition of first language learners in grades 7–10 and contains eight subsets. Each subset has independent data, independent topics, and different average article lengths.

As shown in Table 1, the article types are mainly discussion, narrative, explanation, and question answering. Essays, narratives, or expository essays require the author's article to describe a story or news. While answering questions requires the author to read a paragraph of material first and then write an article based on the questions and requirements given at the end of the reading material. The themes of the eight data subsets are different. Among them, subset 1 asks to talk about the impact of computers on life. Subset 2 is about whether the library needs to review the content of the book. Subset 3–6 is to read the material first and then write the essay according to the prompts. Subset 7 requires writing a story about patience. Subset 8 shows that laughter is an important element in interpersonal relationships, and an article about laughter is required.

4.2. Parameter Settings. Before the neural network is trained, the hyperparameters in the neural network are initialized. First, both the character vector and the word vector need to be initialized. In addition, in order to prevent overfitting, it is also necessary to set up dropout and control the learning rate. For the initialization of the word vector, we choose to use the GloVe vector table with better performance, which is obtained from text training containing 6 billion words from websites such as Wikipedia. The character-level vectored representation uses average distribution for random initialization. That is, each dimension of each character vector is a value between zero and one.

When training the neural network, add a dropout layer to the input and output layers of the recurrent neural network to control network training to prevent overfitting. The ratio of dropout is set to 0.5. In the experiment, the same applies to whether to use dropout. In the model, some hyperparameters are set as shown in Table 2. Among them, the dimension of the hidden layer is set to 200. The model is trained using the Adam optimization algorithm, and each batch is set to 10. Initialize the learning rate to 0.01. In this paper, a Bayesian optimization algorithm is used to optimize the super parameters. First, assume a search function based on the prior distribution. Then, each time the result sampling point is used to test the objective function, this information is used to update the prior distribution of the objective function. Finally, the algorithm tests the point where the global maximum value given by the oil posterior distribution is likely to occur. In this case, the parameter that satisfies the condition is the optimal parameter.

Table 1: Data set description.

Data subset	Article type	Article author grade	Full marks	Number of articles
1	Essays, narratives, explanatory essays	8	12	2001
2	Essays, narratives, explanatory essays	10	6	1678
3	Answer questions based on source article	10	3	1988
4	Answer questions based on source article	10	3	1654
5	Answer questions based on source article	8	4	1768
6	Answer questions based on source article	10	4	1802
7	Essays, narratives, explanatory essays	7	30	1569
8	Essays, narratives, explanatory essays	10	60	723

Table 2: Hyperparameter settings.

Dropout rate	0.5
Batch size	10
Initial learning rate	0.01
Decay rare	0.2
Dimension	100

4.3. Result Analysis of Part-of-Speech Tagging. In this paper, we conduct comparative experiments based on the labeling model of recurrent neural networks. As shown in Figures 5 and 6, analyze and discuss the effectiveness of each structure in the network structure, and analyze the effectiveness of the network in different labeling tasks. At the same time, according to the comparison of part-of-speech tagging on different corpora, the versatility of the analysis model when processing different corpora is shown in Figure 7. The results of this paper are compared with the literature [28], the part-of-speech tagging algorithm of the maximum entropy model [29], the part-of-speech tagging algorithm of the hidden horse model [30], and the part-of-speech tagging algorithm based on SVM [31].

As shown in Figures 5 and 6, the basic network is designed to use word-level vector information to connect two layers of BLSTM and introduce a residual network structure between the two layers of BLSTM. Based on the basic network, after introducing a coarse-labeled supervision layer, in a single training, two-parameter update processes will be included, and the introduced coarse-grained labeling will supervise the network, which improves the accuracy of the labeling. However, in the above network, the input of the second layer BLSTM consists of two parts, and the data distribution of the two parts is not uniform. After the introduction of Batch Normalization, the standardization of the two parts of the input is realized. At this time, the network structure (BRCBN) has improved the accuracy of labeling. The first three network models all have the problem of unregistered words. During training, the word vectors of unregistered words have been in an untrained state. The labeling results of these words have nothing to do with the network structure and tend to be randomized. Immediately after the introduction of CRNN in the experiment, the words were vectored from the character level, and these problems were solved by learning the relations of the composition of the words. The accuracy of this network (BCRCBN) in the part-of-speech tagging experiment reached 0.976, and the $F1$ value in the named entity recognition experiment reached 0.913.

When traditional network models are annotated in a special corpus, the accuracy of the annotation is often insufficient because of the corpus. For example, using the Senna model [32], which first uses a convolutional neural network for character information extraction, and then uses a feedforward neural network for labeling. When the composition corpus is part-of-speech labeling, the labeling accuracy is 0.953. Akhil et al. [27] used a multilayer neural network to divide the labeling process into two steps. In addition, in the last layer, CRF is used for labeling. When the written English composition corpus is labeled with a part of speech, the labeling accuracy reaches 0.956. Compared with previous work, the accuracy of the model in this paper reaches 0.976, as shown in Figure 7. Therefore, even if the corpus containing English grammatical errors is labeled, the model in this paper can still maintain a high labeling accuracy.

As the complexity of the model increases, the model in this paper introduces dropout to solve the overfitting problem, as shown in Figure 8(a). In the experiment, compared to the results without dropout, dropout can significantly alleviate overfitting the problem. This is because dropout randomly makes the weights of some hidden layer nodes in the network not work and limits the weights to achieve a regular effect.

In terms of word vector selection, the model in this paper uses GloVe's 50-dimensional, 100-dimensional, and 300-dimensional word vector for comparison experiments, and compares with the random initialization method, as shown in Figure 8(b) below show, finally choose to use a 100-dimensional vector to initialize the word vector in the model.

4.4. Result Analysis of the Automatic Scoring Model. Using nontext features, this paper separately trains the model on eight composition subsets and predicts the test set scores and calculates the corresponding twice-weighted Kappa value. The experimental results are shown in Figure 9.

It can be seen from Figure 9 that, on all the composition data sets, the random forest has the largest second-weighted Kappa value, followed by XGBoost, and the gradient boosting tree has the lowest result. This is because each composition subset only has more than 1,000 essays, and all composition subsets add up to more than 10,000 samples,

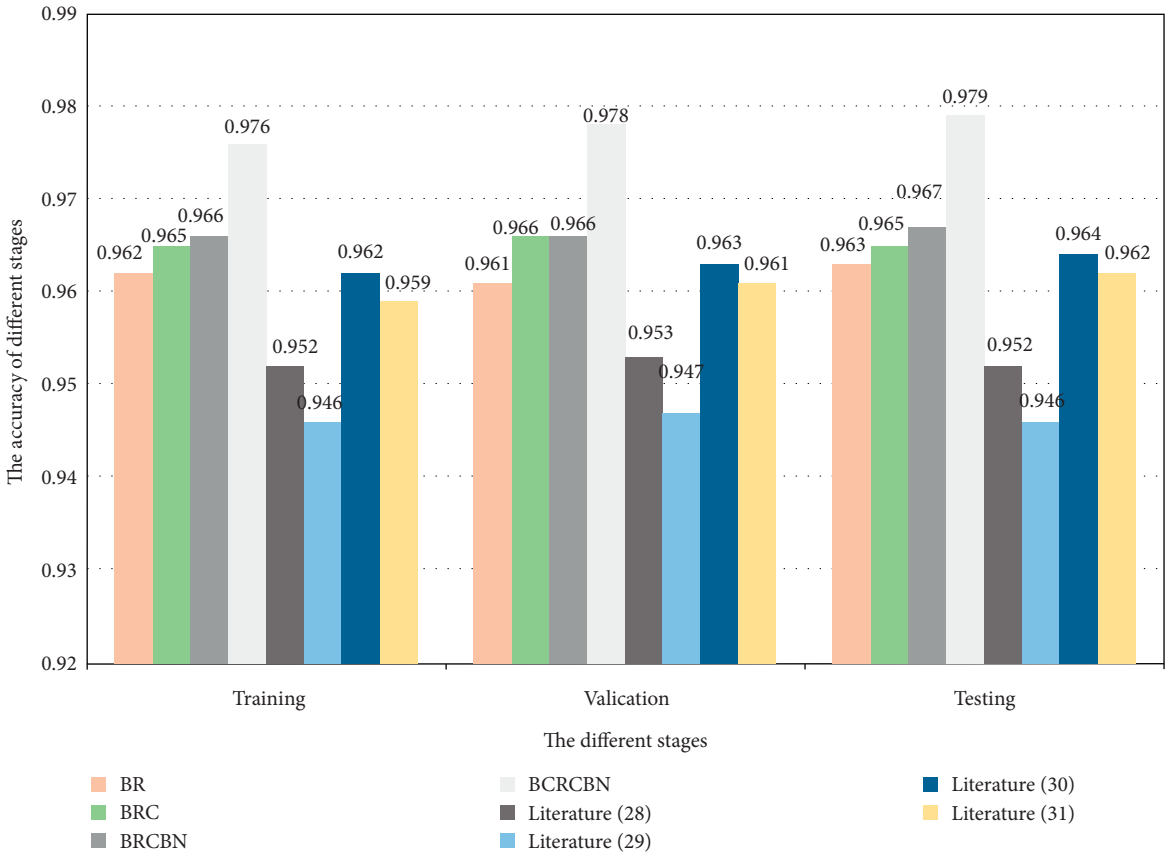


Figure 5: The accuracy of part-of-speech tagging in WSJ corpus for different network structures.

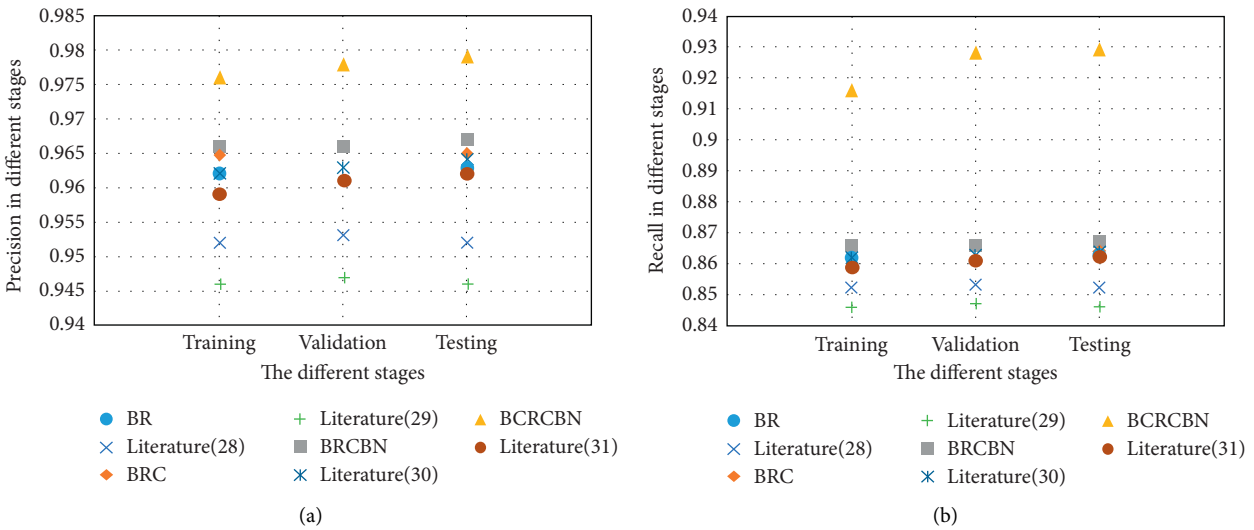
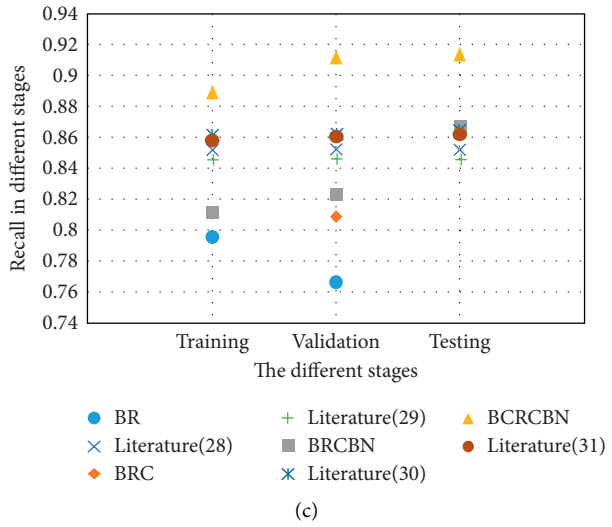


Figure 6: Continued.



(c)

Figure 6: Results of named entity recognition in WSJ corpus for different network structures.

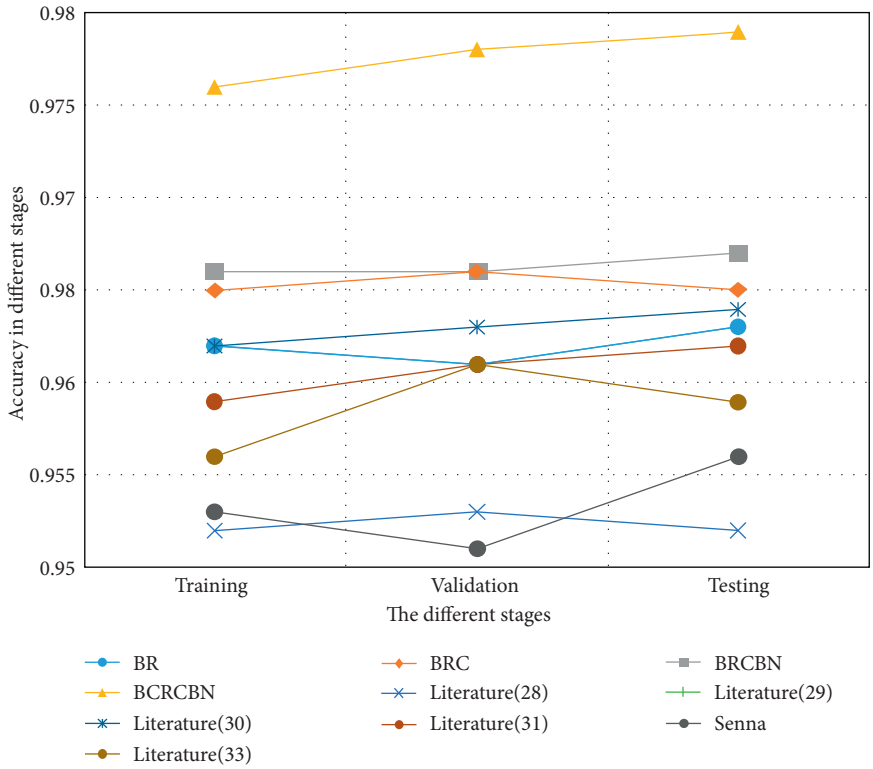


Figure 7: The accuracy of the model in Pigai corpus for part-of-speech tagging.

and the data involved in training the model is still not sufficient. Two models, such as gradient boosting tree and XGBoost, are implemented based on the Boosting method, so when the amount of data is small, the model is prone to overfitting the data. That is, the model is relatively complicated, and too much consideration of the individuality of the sample will cover the commonality of the sample, resulting in a poor prediction effect. The random forest is based on the Bagging method, which uses the results generated by multiple decision trees to generate prediction results in the form of voting or averaging. In this way, even when the amount of training sample data is small, it can still effectively avoid overfitting and reduce variance. Therefore, random forests can show better prediction effects when the amount of data is small. We believe that the XGBoost model increases the regularization term of the cost function to

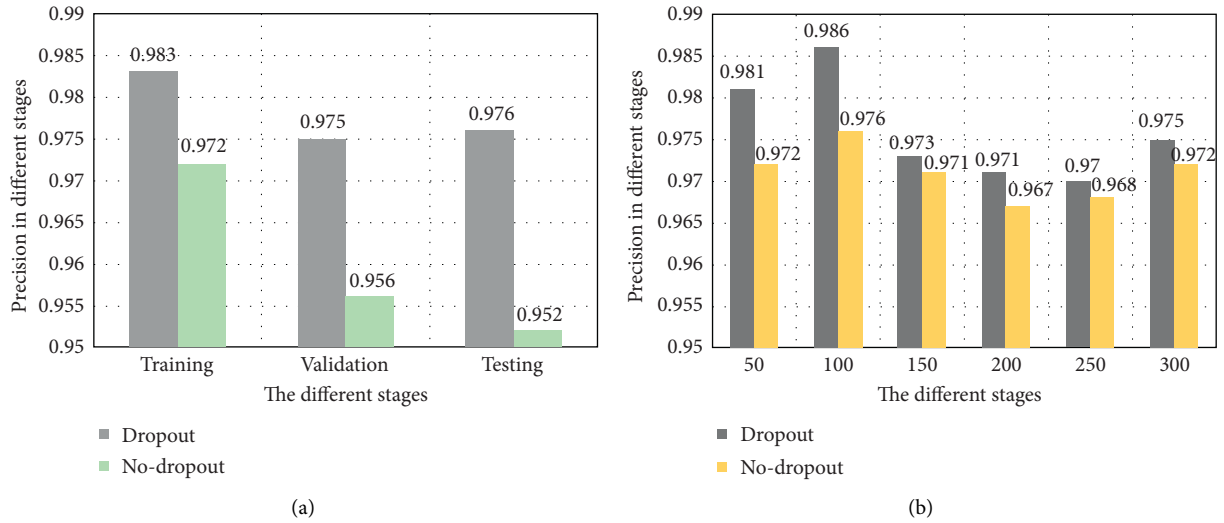


Figure 8: The influence of Dropout and dimension. (a) The role of Dropout. (b) The role of GloVe dimension.

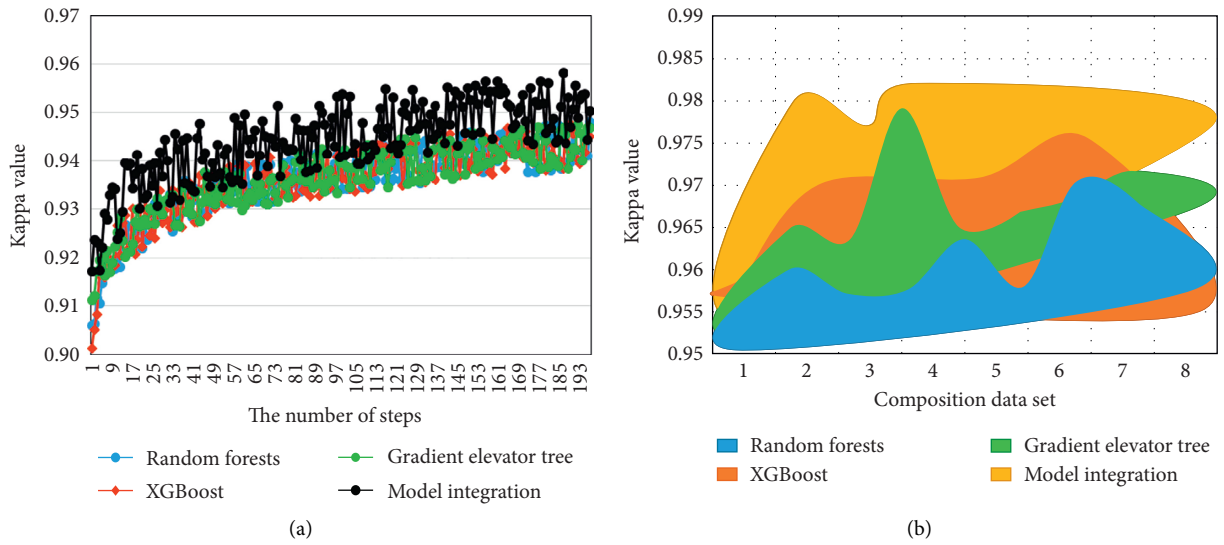


Figure 9: Results of nontext features in each essay set. (a) In all composition data sets. (b) In different composition data sets.

control the complexity of the model and limit the number of leaf nodes of the tree, which can effectively prevent overfitting. In addition, the XGBoost model also draws on the random forest column sampling method, which can also effectively prevent overfitting. Therefore, although it is also a model based on the Boosting idea, the effect of the XGBoost model is better than the gradient boosting tree model. It can be seen that in combinatorial subsets 4 and 8, the quadratic weighted Kappa value of the XGBoost model is slightly higher than that of random forest. In combination 8, the quadratic weighted Kappa values of the three models are relatively low. The author analyses that this is related to the larger scoring range of the composition set 8. When the scoring range is larger, the corresponding error will be magnified.

5. Conclusion

This paper proposes an automatic scoring model for English composition. The method uses a convolutional neural network to extract word information from character level and uses features for coarse-grained learning layer. Then, word-level vectors are introduced to integrate coarse-grained annotations with word vector information. Then, RNN is used to extract the overall information of the sequence data. Considering the timeliness of feedback results

and the accuracy of prediction, this paper chooses a simple and efficient Bagging method for linear fusion of random forest, GBDT, and XGBoost. Each model is multiplied by a certain weight and then added up to get the final output so as to realize the construction of the English composition automatic scoring model. The experimental results show that the automatic scoring model proposed in this paper has achieved a good accuracy of POS tagging, which reaches 0.976.

Although this paper has achieved better experimental results, but the open dataset is still used. But the public dataset contains a limited amount of data. Our next research plan is to build our own database and train the algorithm in multiple databases to enhance the robustness of the algorithm.

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 that there are no conflicts of interest.

References

[1] W. Qiu, S. Li, X. Cui et al., "Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 450, pp. 86–103, 2018.

[2] M. Buell, M. Han, and C. Vukelich, "Factors affecting variance in Classroom Assessment Scoring System scores: season, context, and classroom composition," *Early Child Development and Care*, vol. 187, no. 11, pp. 1635–1648, 2017.

[3] M. Foroutan, D. D. Bhuva, R. Lyu et al., "Single sample scoring of molecular phenotypes," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–10, 2018.

[4] M. Saltzberg, "Theories of labour: physically scoring brecht's mother courage and her children with stanislavski, viewpoints, and composition," *Stanislavski Studies*, vol. 8, no. 2, pp. 237–245, 2020.

[5] A. Mahshanian and M. Shahnazari, "The effect of raters fatigue on scoring EFL writing tasks," *Indonesian Journal of Applied Linguistics*, vol. 10, no. 1, pp. 1–13, 2020.

[6] M. Waris, K. Ahmad, M. Kabir, and M. Hayat, "Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix," *Neurocomputing*, vol. 199, pp. 154–162, 2016.

[7] Y. Yang, "Teaching Chinese college ESL writing: a genre-based approach," *English Language Teaching*, vol. 9, no. 9, pp. 36–44, 2016.

[8] S. H. Alharbi, "Principled eclecticism: approach and application in teaching writing to ESL/EFL students," *English Language Teaching*, vol. 10, no. 2, pp. 33–39, 2017.

[9] J. Wilson and J. Rodrigues, "Classification accuracy and efficiency of writing screening using automated essay scoring," *Journal of School Psychology*, vol. 82, pp. 123–140, 2020.

[10] X. Lu, "An empirical study on the artificial intelligence writing evaluation system in China CET," *Big Data*, vol. 7, no. 2, pp. 121–129, 2019.

[11] A. Cahill, M. Chodorow, and M. Flor, "Developing an e-rater advisory to detect babel-generated essays," *The Journal of Writing Analytics*, vol. 2, no. 1, pp. 203–224, 2018.

[12] J. S. Velázquez-Blázquez, J. M. Bolarín, F. Cavas-Martínez, and J. L. Alió, "EMKLAS: a new automatic scoring system for early and mild keratoconus detection," *Translational Vision Science & Technology*, vol. 9, no. 2, p. 30, 2020.

[13] S. Fu, H. Gu, and B. Yang, "The affordances of AI-enabled automatic scoring applications on learners' continuous learning intention: an empirical study in China," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1674–1692, 2020.

[14] H. K. Janda, A. Pawar, S. Du, and V. Mago, "Syntactic, semantic and sentiment analysis: the joint effect on automated essay evaluation," *IEEE Access*, vol. 7, pp. 108486–108503, 2019.

[15] K. Zupanc and Z. Bosnić, "Automated essay evaluation with semantic analysis," *Knowledge-Based Systems*, vol. 120, pp. 118–132, 2017.

[16] J. Wang, "A comparative study on the washback effects of teacher feedback plus intelligent feedback versus teacher feedback on English writing teaching in higher vocational college," *Theory and Practice in Language Studies*, vol. 9, no. 12, pp. 1555–1561, 2019.

[17] A. Hassan, A. Riad, and A. Shehab, "An automated essay grading framework based on neural networks (dept.E)," *Mansoura Engineering Journal*, vol. 33, no. 2, pp. 10–21, 2020.

[18] J. Chen, J. H. Fife, I. I. Bejar, and A. A. Rupp, "Buildinge-rater scoring models using machine learning methods," *ETS Research Report Series*, vol. 2016, no. 1, pp. 1–12, 2016.

[19] M. Zhang, J. Chen, and C. Ruan, "Evaluating the advisory flags and machine scoring difficulty in thee-rater automated scoring engine," *ETS Research Report Series*, vol. 2016, no. 2, pp. 1–14, 2016.

[20] W. M. Wang, Z. Li, J. W. Wang, and Z. H. Zheng, "How far we can go with extractive text summarization? Heuristic methods to obtain near upper bounds," *Expert Systems with Applications*, vol. 90, pp. 439–463, 2017.

[21] N. Ghadimi, A. Akbarimajd, H. Shayeghi, and O. Abedinia, "Two stage forecast engine with feature selection technique and improved meta-heuristic algorithm for electricity load forecasting," *Energy*, vol. 161, pp. 130–142, 2018.

[22] E.-J. Wagenmakers, M. Marsman, T. Jamil et al., "Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications," *Psychonomic Bulletin & Review*, vol. 25, no. 1, pp. 35–57, 2018.

[23] A. Van Moere and S. Hanlon, "A Bayesian approach to improving measurement precision over multiple test occasions," *Language Testing*, vol. 37, no. 4, pp. 482–502, 2020.

[24] B. Bridgeman and C. Ramineni, "Design and evaluation of automated writing evaluation models: relationships with writing in naturalistic settings," *Assessing Writing*, vol. 34, pp. 62–71, 2017.

[25] R. D. Roscoe, J. Wilson, A. C. Johnson, and C. R. Mayra, "Presentation, expectations, and experience: sources of student perceptions of automated writing evaluation," *Computers in Human Behavior*, vol. 70, pp. 207–221, 2017.

[26] Y. Li, L. Cao, J. Zhu, and J. Luo, "Mining fashion outfit composition using an end-to-end deep learning approach on set data," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1946–1955, 2017.

[27] K. K. Akhil, R. Rajimol, and V. S. Anoop, "Parts-of-Speech tagging for Malayalam using deep learning techniques,"

International Journal of Information Technology, vol. 12, no. 3, pp. 741–748, 2020.

[28] L. Nanni, S. Brahnam, and S. Brahnam, "Set of approaches based on position specific scoring matrix and amino acid sequence for primary category enzyme classification," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 38–52, 2020.

[29] J. Awwalu, S. E.-Y. Abdullahi, and A. E. Evwiekpaefe, "Parts of speech tagging: a review of techniques," *Fudma Journal of Sciences*, vol. 4, no. 2, pp. 712–721, 2020.

[30] V.-T. Bui, P. T. Nguyen, P.-T. Nguyen, V.-L. Pham, and T.-Q. Ngo, "A neural network model for efficient antonymy-synonymy classification by exploiting Co-occurrence contexts and word-structure patterns," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 1, pp. 156–166, 2020.

[31] S. M. Sarsam, H. Al-Samarraie, and A. Al-Sadi, "Disease discovery-based emotion lexicon: a heuristic approach to characterise sicknesses in microblogs," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, pp. 1–10, 2020.

[32] Z. Yuan, C. Zhang, X. Peng et al., "Intestinal microbiota characteristics of mice treated with Folium senna decoction gavage combined with restraint and tail pinch stress," *Biotech*, vol. 10, pp. 1–11, 2020.