Hindawi
Scientific Programming
Volume 2021, Article ID 3429463, 10 pages
https://doi.org/10.1155/2021/3429463

Research Article

Research and Design of Automatic Scoring Algorithm for English Composition Based on Machine Learning

Yu Zhao

School of International Studies, University of Science and Technology Liaoning, Anshan City 114000, Liaoning Province, China

Correspondence should be addressed to Yu Zhao; zhaoyu963@ustl.edu.cn

Received 15 November 2021; Revised 29 November 2021; Accepted 6 December 2021; Published 23 December 2021

Academic Editor: Baiyuan Ding

With the development of artificial intelligence and big data, the concept of "Internet plus education" has gradually become popular, including automatic scoring system based on machine learning. Countries all over the world vigorously promote the deep integration of information technology and discipline teaching in various fields. English is a medium of communication in the current era of education information development trend. English composition automatic scoring mode is gradually accepted by the majority of educators and applied in the actual classroom teaching. However, the research of English composition automatic grading in teaching space is not perfect. Most systems have used traditional algorithms. Therefore, this paper constructs the automatic scoring algorithm and sentence elegance feature scoring algorithm of English composition based on machine learning, explores the influence of the algorithm on English writing teaching, and proves the correctness of the design idea and algorithm of this paper through a lot of experiments.

1. Introduction

English is one of the main international languages and the most widely used language in the world. Nowadays, the process of internationalization is accelerating. As an international lingua franca, English is a basic skill for students to go to the world [1]. Learning English can not only enrich their knowledge but also broaden their horizons and improve personal advantages. In the process of English learning, English composition writing is indispensable. Writing can not only reflect learners' grasp of basic knowledge, such as vocabulary and grammar, but also reflect learners' overall grasp of sentence structure, discourse structure, and logical reasoning ability [2]. As English composition can directly reflect the comprehensive language ability of English learners; writing ability is the key item of the test organizers to investigate the language ability of students. Thus, more and more learners pay attention to it [3]. At present, writing is still a weak link in students' English learning, and it is also a link that teaching staff need to spend a lot of time preparing. In usual learning, teachers generally manually review English compositions, and it is suitable for a small number of short ones [4]. However, when the intensity of marking is high, and the length of essays is long, this method will lead to low efficiency and high rate of misjudgments. More importantly, the subjectivity of manual assessment is too strong, and people with different knowledge and regions may have a great bias on the same essay.

In recent years, the rapid development of computer technology has led to the prosperity of other industries, which has promoted the research and application of Automated Essay Scoring (AES) technology in education [5]. On the one hand, this technology can intelligently analyze and grade compositions. Compared with manual evaluation, computer evaluation has low cost and high review efficiency. The automatic scoring system gives full play to the characteristics that computers are good at repetitive work, which can liberate teachers' physical and mental strength to a great extent and allow teachers to devote more energy to teaching and research. On the other hand, the analysis results can provide feedback regarding more evaluation information, such as spelling and grammatical errors of each composition, so that students can initially modify their own articles according to systematic suggestions. The analysis results can

also recommend excellent words and sentences and composition materials to provide more scientific writing guidance for learners. At present, the construction and evaluation of AES system are mostly based on the statistical information of the content of the paper. The research content is relatively simple, and the evaluation of the logic of the composition and the quality of words and sentences is still not in-depth and accurate. Therefore, while improving the accuracy of the predicted score, we should also comprehensively evaluate the composition content, so that the AES system can be better applied to the actual composition correction.

2. Related Work

The research on automatic scoring in education industry started earlier, and there were many systematic studies on different subjects and languages. The research on the scoring system related to composition can be traced back to the 1960s. The earliest scoring system Project Essay Grader (PEG) [6] developed by Professor Ellis Page took the features extracted from shallow linguistic methods (such as article length, word length, punctuation marks, grammar, etc.) as independent variables. The AES model is trained by multiple linear regression with the score due to the composition as the dependent variable. This method does not involve the language content and composition structure, so the evaluation results are biased. Subsequently, Landauer Thomas et al. developed an automatic Essay scoring system named Intelligent Essay Analysis (IEA) [7] based on Latent Semantic Analysis (LSA) [8]. When the model is used to score students' English compositions, the compositions to be graded and excellent examples are mapped to the vector space, and the scores of compositions are predicted by comparing the similarity values. This approach takes the overall content of the article into account and is more accurate than PEG. It can also be used to detect plagiarism in the article, which has greatly advanced the field of automatic grading.

With the deepening of the research on automatic scoring system, American Educational Examination Institute developed the E-rate system in the 1990s [9]. It used natural language processing technology and statistical methods to mark the part of speech of each word through a part-of-speech tagger and analyzed the syntactic structure of the text through syntactic analysis. These methods can be used to evaluate the quality of writing language, content, and text structure and have been applied to the automatic scoring of GMAT and GRE. Although the design of E-Rater is more comprehensive, it is not as comprehensive as PEG in language analysis and not as in-depth as IEA in content analysis, so there is a lot of room for improvement. The AES system developed by Professor Liang's team in China [10] is based on superficial linguistic features and linear regression model training to analyze the accuracy of word spelling and grammar usage in each sentence. However, it failed to give students more evaluation results in terms of discourse quality, sentence quality, and relevance. Subsequently, Wang et al. [11] improved the automatic scoring effect from the perspective of semantic dispersion and introduced the convolutional neural network training model, which showed good performance in composition prediction ability. Qiu [12] evaluated the fluency of the composition and quantified it into AES model to improve the scoring effect of the system. Liu et al. [13] considered figurative parallelism and other rhetorical devices in Chinese compositions and constructed a corpus of ancient poems to automatically identify ancient poems in compositions, which has better accuracy compared with the benchmark system. Yangwei and Huang used Auto-Encoder (AE) [14] to reconstruct linguistic features and then input the reconstructed feature vectors into SVM for regression training, which improved the performance compared with its previous reconstruction.

With the rapid development of intelligent hardware, artificial intelligence has made rapid progress, including natural language processing. Natural language processing based on deep learning can be summarized as the problem of original data feature representation in application field and the problem of selecting appropriate deep learning algorithm to construct application model. For the former problem, mature models include bag-of-words (BOW) [15] and Vector Space Model (VSM) [16]. These methods have certain defects. The word bag model, such as one-hot Encoding, will also become very large and sparse when the number of categories is large. Vector space models such as Term Frequency-Inverse Document Frequency (TF-IDF) [17] characterize text features by calculating the probability of words becoming keywords in the text. However, this method is greatly influenced by the global text set and only makes use of the statistical information of words. The location information and context information are not utilized, so the text features cannot be fully represented. Bengio et al. [18] used deep neural network to construct a language model, which could map words into a vector space of fixed dimensions, solving the problems of sparse features and large dimensions caused by one-hot coding. By training neural network language model through unsupervised learning, semantic information contained in the text could be obtained. Its disadvantage is that it involves a lot of parameter training, which leads to a long training cycle.

Mikolov et al. proposed word2vec [19], a word vector training model in 2013. This model is based on Continuous bag-of-words (CBOW) and Skip-Gram models. The former can predict the probability of the occurrence of the current word based on the semantic information before and after the word. The latter is the most widely used word vector representation model, which uses the current word to predict the probability of the occurrence of the preceding word. Mikolov then proposed the Doc2veC model based on paragraph vector, which mainly added paragraph vector to the word2vec model and also included two model structures: Distributed Memory Model and Distributed Bag-of-Words, which can represent sentences and text. Jeffrey proposed the Glove word vector representation model in 2014 [20], which accelerated the training of word vector and enriched the semantic information of word vector. The Embedding from Language Model (ELMO) was proposed by Peters in March 2018 [21], which adopted the double-layer bidirectional

LSTM structure pretraining model. Then, Word Embedding dynamically adjusts the representation of corresponding words according to the context of the words in the input sentence, which can solve the problem of polysemy. In October 2018, Jacob Devlin et al. proposed the Bidirectional Encoder Representations from Transformers (BERT) representations from Transformers language representation model [22]. The bidirectional encoder of Transformer is used to pretrain the model based on the context of all layers. Fine-tuning an output layer can create an optimal model for downstream tasks, which is the best language representation model at present.

Researchers generally believe that natural language has certain logical and recursive characteristics, and language contexts are closely related. For example, sentences in natural language are actually composed of words and phrases recursively. Therefore, recursion is an important feature of natural language. Therefore, models can be selected according to the characteristics of natural language, such as recurrent neural network, convolutional neural networks, and a series of improved models [22]. KIM proposed to use text-CNN model for sentence classification in 2014 [23] and proved that convolutional neural network plays an important role in extracting text features. In 2017, Liu et al. [24] proposed a combination model of deep neural network DC-NN, which used an improved recursive neural network to generate phrase pair semantic vectors suitable for phrase generation process and used an auto-encoder to improve the performance of phrase generation process. It performs better than baseline models on machine translation tasks. Hassan and Mahmood [25] proposed the training method of integrating CNN and RNN and obtained the convolution layer of long-term dependence through joint training, achieving an accuracy of 93.3% in emotion analysis. LSTM, first proposed by Schuster and Paliwal [26], was an improved model based on cyclic neural network, which selectively stored and forgot long-term memories through memory units. Wang et al. [27] applied the attention mechanism to convolutional neural network to extract features and integrated them into the LSTM model, which showed excellent features in Tweet sentiment analysis. Hochreiter and Schmidhuber [28] completed the multilabel classification task of books based on long-short-term memory, with good performance in various indicators. In view of the fact that LSTM model can only remember the information above and does not make full use of the information below, Schuster and Paliwal [26] proposed bidirectional LSTM model (BiLSTM), which added a reverse layer to LSTM model, and allowed it to use context information to improve model performance. Zeng et al. [29] used the bidirectional LSTM model to obtain the bidirectional semantic information of comments to achieve the task of sentiment classification of comment texts, which has a better effect compared with convolutional neural network. Although many researches have been done in the field of automatic scoring, the accuracy and comprehensiveness of scoring are not thorough enough. In this paper, composition features are extracted from multiple dimensions, and appropriate technical models are selected through specific analysis to extract features and train models, so as to obtain the best performance of the scoring model.

3. Network Framework

3.1. Basic Feature Extraction

3.1.1. Based on Linguistic Features. Feature extraction is an important process of machine learning modeling, which determines the quality of model results. Linguistic features refer to the use of statistical methods to extract shallow features of composition words and sentences without considering the meaning of the text, such as the proportion of part of speech and number and length of words and sentences. Simply speaking, linguistic features are the direct application of statistics. The early AES models, such as PEG scoring systems, basically used superficial linguistic features such as text length, number of sentences, part of speech, and other surface features to make models and then used regression equations to train the model and build AES system. These linguistic features have some effect. These linguistic features have some effect. The statistical results based on words and sentences can well reflect the complexity of words and sentences, so as to reflect the author's ability to use words and sentences. For example, the total number of words, the number of words after repetition, and the number of high-frequency words reflect the author's vocabulary. The number of clauses, phrases, etc. reflect the author's ability to use complex sentence patterns, so we use morphological and syntactic statistical features, and the details are shown in Table 1.

3.1.2. Features Based on Semantic Expression. Features of semantic expression are not only to consider statistical features but also to consider deep meanings of language texts. The common ones are to get the word vector clustering features according to composition words vector set under different clustering number distribution, to get text vector features of text vector from the text level, and to get theme distribution characteristics of the theme probability distribution of compositions through training the LDA model.

(1) Word Vector Clustering Model Based on word2vec. The word vector clustering model based on word2vec feature extraction is one of the most classic maturity models. The first step is to build a word vector by using word2vec tool based on the fixed code words, which means that each word is mapped to a fixed dimension of feature space, thus obtaining term vectors. This process needs to set the word vector length and receive a word vector corresponding to the whole composition. For the second time, k-means clustering method is used to cluster the word vectors, and "optimal fitness" is adopted to select the number of clustering clusters. The main idea of "optimal fitness" is to minimize the distance between word vectors in the cluster and maximize the distance between cluster centroids [30, 31]. The number distribution of word vectors under different clustering clusters is counted, and the returned number distribution is regarded as the clustering feature of word vectors. The

Table 1: Text feature description.

Overview of lexical features	Overview of syntactic features
The proportion of word list size to composition Length (not repeated)	Average sentence length and variance
Number of sentences whose length is greater than a fixed value (e.g., 4, 8, 12)	Number of sentences whose length is greater than a fixed value (e.g., 4, 8, 12)
Statistical characteristics such as average character length (mean word length, median, standard deviation)	Average number of verbs, nouns, modal verbs, prepositions in sentences
The proportion of nouns, adjectives, verbs and prepositions	Average number of punctuation marks in a sentence
Number of high-frequency words	The number of sentences that fully express the meaning
The size of the word list after removing the stop word	The number of clauses and the average length of clauses

clustering model trained based on this idea can show better robustness and higher performance. The core formula is as follows:

$$\alpha = \frac{1}{k}\sum_{1\le p<q\le k}^{k}\cos\left(s_p, s_q\right), \beta = \frac{1}{k}\sum_{b=1}^{k}\frac{1}{t_b}\sum_{1\le i<j\le t_b}^{t_b}\cos\left(w_i, w_j\right), f(x)\frac{\alpha}{\beta}, \tag{1}$$

where s_p, s_q are the centroid of clustering clusters p and g of composition word package. k represents clusters number of word package, α is the average distance between cluster, t_b is the first b clustering clusters to the total number of words, w_i, w_j are for the i and j word vectors, β is for all the average distance between cluster-heads clustering vector, $f(x)$ is related to the number of clustering cluster of fitness function, and adjust the k values to make the value of fitness function be the largest number. At this point, the average distance between clusters should be as large as possible, and the average distance between samples in clusters should be as small as possible. After such a series of operations, the number distribution of words belonging to each cluster in the composition can be counted, so as to characterize the word vector features of the composition.

(2) Text Vector Features Based on Doc2vec. Although word2vec has greatly promoted the development of natural language processing, the common sentence representation methods based on word2vec include joining, adding, averaging, local maximum, and minimum values. These methods use simple superposition averaging and other relatively rough methods, without considering the influence of context information on vocabulary. Mikolov et al. proposed and improved the Doc2vec text vector features proposed. Doc2vec is basically similar to the word2vec model, except that the text vector is added in the training process, so the text vector of the context can be encoded. The model also includes two structures: Distributed Memory (DM) Model and Distributed Bag-of-Words (DBOW). It also has the same functionality as the Skip-Gram model and the CBOW model. By encoding the text with considering the context, text vectors can be finally got.

(3) LDA Feature Extraction. The Latent Dirichlet Allocation (LDA) topic model was proposed by Friedman et al [32]. They believed that the topic of an article is in line with the Dirichlet distribution, so as to obtain the relationship between texts, and the VSM is compared to increase probability information. LDA model is composed of a three-layer generative Bayesian network structure [32], including documents, topics, and words. The core probability calculation is shown in Formula (4).

$$p\left(w_i \mid d_j\right) = \sum_{s=1}^{k} p\left(w_i \mid z = s\right) p\left(z = s \mid d_j\right), \tag{2}$$

where $p(w_i \mid z = s)$ represents the probability that the word w_i belongs to s topics, and $p(z = s \mid d_j)$ is the probability of the s topic in the short text d_j. Based on the LDA topic model, the topic probability distribution of the text can be obtained, which can be extracted as the topic features of the text.

3.2. AES Model Based on BiLSTM. Short-long-term memory network (LSTM) is a special kind of recurrent neural network (RNN). RNN network takes the output of the previous moment as a part of the input and inputs it into the neural network together with the external input at the moment, but there are problems of short-term memory and gradient disappearance. LSTM improves the RNN model by replacing the hidden layer nodes of RNN with Memory units and protects or controls the node states of LSTM neural network through the gate structure, so that the closed loop is formed between the hidden layer, and the weight of the hidden layer is responsible for controlling scheduling Memory. The state of the hidden layer participates in the next prediction as the memory state of the current moment, which solves the problems of short-term memory and gradient disappearance.

BiLSTM allows bidirectional information flow in LSTM. Two hidden states are used from backward to forward and from forward to backward; the internal structure is shown in Figure 1. The calculation formula of BiLSTM at the current moment is shown as follows:

$$\begin{aligned} f_t &= \sigma\left(w_f x + b_f\right), \\ i_t &= \sigma\left(w_i + b_i\right), \\ o_t &= \sigma\left(w_o + b_o\right), \\ c_t &= f_t \otimes c_{t-1} + i_t \otimes \tanh\left(w_c x + b_c\right), \\ h_t &= o_t \otimes \tanh\left(c_t\right), \end{aligned} \tag{3}$$

where (f_t, i_t, o_t) is the output value of the forgetting gate component, and c_t is the current state at moment c. The unit state matrix and neuronal paranoia of forgetting gates, memory gates, are represented by $W_i, W_f, W_o \in R$ and $b_i, b_f, b_o \in R$.

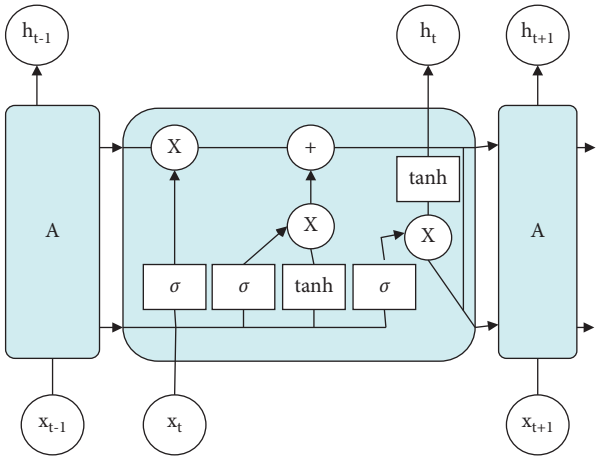

Figure 1: Internal structure of neuron.

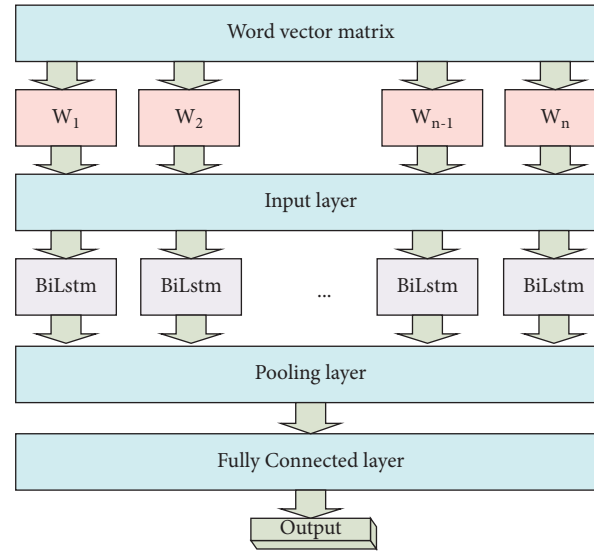

Figure 2: AES network framework based on BiLSTM.

This method of memory selection is consistent with the expression of the relation between words before and after in English composition, in order to mine the contextual timing information in the composition, so as to extract the characteristics of the logic before and after the composition. We take the word vector matrix as the feature input to represent the article and use the LSTM model, which is suitable for dealing with timing problems to build the model. Compared with other network models, LSTM can make full use of context information of composition. The structure of AES model based on LSTM is shown in Figure 2.

3.3. Beautiful Sentence Recognition Based on CNN

3.3.1. Network Overview. Composition morphology and grammar are the basic requirements of writing; really judging the level of a composition is advanced expression of beautiful sentences; these beautiful sentences greatly improve the appreciation of the composition. These sentences usually include fusion advanced vocabulary, clever sentences of English grammar, and some contain rhetoric devices. Therefore, to quantify the beautiful degree and distribution of all writing sentences and fused related characteristics at the same time by building beautifully set identification model can help build the model of AES as well as improving the score prediction efficiency. And the result will not mention mechanization.

The problem of sentence elegance recognition can be regarded as a text classification problem. The main task is to learn the text content by computer and train the classification model according to the given text label, so as to obtain the classification results of the new input text. Traditional statistics-based machine learning methods usually extract features from text manually and then train the classification model with machine learning classifier. However, it is difficult to grasp the beauty features of language and make perfect features artificially, while the biggest advantage of deep learning method is that automatic selection and combination of features can better reflect the features of text information. Traditional methods based on statistics and rules use manual sentence features, which make it difficult to extract the core information of good sentences containing advanced grammar (parallelism and inversion have fixed structures that are easy to extract linguistic features, but metaphors, personification, and other figures of speech have unique linguistic structures). The neural network model can automatically learn the semantic vector that can represent the sentence feature from a large amount of data, so as to complete the task of binary classification.

Convolutional Neural Network is a commonly used artificial Neural Network structure, which uses convolutional kernel arithmetic to obtain local information and then obtains global information through the pooling layer. Its main structure includes input layer, convolution layer, and pooling layer. In the task of automatic composition scoring, the input layer is the text representation matrix of word vector, and the convolution layer can set convolution kernels of different sizes. This method can obtain certain context timing information. In addition, compared with the traditional neural network model, CNN introduces weight sharing, which reduces the complexity of the network model and improves the training speed. In the pooling layer, the piecewise maximum pooling method is adopted, which can retain the relative location information of multiple local maximum eigenvalues. If strong features appear repeatedly, this method can also obtain feature intensity. However, absolute position information is lost, and only coarse position information is retained. After convolution and pooling, sentence level representation is obtained.

3.3.2. AES Network Model Based on CNN

(1) Input Layer. This section attempts to use convolutional neural network to extract sentence elegance features, so it is necessary to complete the transformation from word granularity features to sentence granularity features. In the input layer of convolutional neural network, the word vector is obtained by word embedding method, and the two-

dimensional vector matrix that can represent the sentence is obtained $m * k$ (m is the number of words in the sentence, and k is the length of the word vector). The specific operation is as follows: firstly, word2vec model is trained based on Wikipedia corpus, and then word vector of each word is obtained by incremental training based on the corpus obtained in this paper. The vector length is set to 128, and then word vector matrix is obtained by splicing word vectors in sentences. Since the longest sentence in the training set has 123 words, this paper adopts zero-complement processing to make the input matrix of each sentence the same size for sentences with a length less than 47, that is, 123×128, and the model structure is shown in Figure 3.

(2) Convolution Layer. Convolution is a mathematical operator that generates a third function through two functions f and g and represents the integral of the overlapping length of the product of the overlapping function values of f and g through inversion and translation. For neural networks, it is actually to search for more representative features through operation. The word vector matrix is input into the convolutional neural network, and three types of convolution kernels, 3×128, 4×128, and 5×128, are designed, respectively, in the convolution layer, each of which has 50 convolution kernels. For the operation with sentence length m, the jth convolution kernel computes convolution of the words in the window with length a, and the output result is c_j and the sentence length is expressed as follows:

$$\begin{aligned} x_{i:i+h-1} &= x_i \oplus x_{i+1} \oplus \cdots \oplus x_{i+h-1}, \\ c_j &= f\left(w * x_{i:i+h-1} + b_j\right), \end{aligned} \tag{4}$$

where x_i is the i vector in the two-dimensional matrix composed of adjacent a word vectors, w is the weight parameter of the convolution kernel, b_j is the bias element, and f is the activation function. In order to consider the contextual information of each word as much as possible, this paper extracts local features of different dynamics by changing the size of window length a. By convolution operation on the input data of convolution check, a feature graph c can be obtained, and finally 150 feature graphs can be obtained. In order to accelerate the convergence speed of network training, ReLu function is used as the activation function of each neuron.

$$\begin{aligned} c &= \left[c_1, c_2, \ldots, c_{m-a+1}\right], \\ f(x) &= \max(0, x). \end{aligned} \tag{5}$$

(3) Pooling Layer. In order to obtain more accurate feature expression, local optimal features need to be obtained from feature maps; that is, chunk-max-pooling operation is performed on feature maps extracted from the convolutional layer. Firstly, the feature graph is divided into parts, then the maximum value of each one is extracted and combined into a vector composed of maximum value, and then the maximum value vector of all feature graphs is spliced to complete the whole process of extracting a local feature from the input data. The feature maps were input to the pooling layer, and each feature map was divided into three parts by local maximum pooling operation, and the features were combined in sequence. Local maximum pooling has better performance than global maximum pooling, because it preserves relative order information as much as possible while capturing the strength of features.

(4) All Connections. The combination and artificial features vectors are joined together into one dimensional vector and connected to the whole connection layer, finally through the beautiful output node to output value. The output node is the sigmoid function. In the frame of the Keras neural network, we adopt the method of stochastic gradient descent to adjust parameters reversely through iterations. Thus, the final training meets the requirements of model structure.

4. Experimental Analyses

4.1. Experimental Preparation and Evaluation Indicators

4.1.1. The Experimental Data. In order to ensure the accuracy of the test results, the "Composition Scoring Competition" on the international data mining platform Kaggle is selected [33]. The composition data set includes 8 data subsets, and each subset has corresponding composition questions. Students write according to the requirements of the questions. Under each training data subset, there are more than one thousand compositions that students have learned and the corresponding manual grading scores, and each composition is usually between 150 and 650 words. The score range of each data subset is also different. For example, the score range of data set 3 is 0–3, while the score range of data set 7 is 0–30, and the score range of data set 8 is 0–60.

In order to further verify the reliability of the beautiful essay model, we collected the beautiful sentences of IELTS and TOEFL on Douban and "English beautiful essay" on other English websites. We cut the beautiful essays into sentences and defined the above sentences as beautiful sentences. Then, primary school English composition and junior high school English composition modules are collected from the English network. We cut the composition into sentences, respectively, and define such sentences as not beautiful sentences. Among them, 3340 original beautiful sentences and 3304 unbeautiful sentences were selected, and the sentences with more than 6 words and less than 50 words were selected as valid samples. In the end, there were 3200 beautiful sentences and 2900 unbeautiful sentences.

4.1.2. The Evaluation Indicators. Quadratic Weight Kappa (QWK) is used to evaluate the performance of the model. QWK is a consistency test method used to evaluate whether the results of the model are consistent with the actual results. Let the composition score have N natural number grades, and there are two markers (A: manual grading, and B: algorithm grading). The score of each composition E can be represented by an array (e_a, e_b), which represents the score of composition e by A and B markers, respectively. First, A histogram matrix O of order n is constructed, which represents the number of essays that marker A typed with A score of i and marker B typed with A score of j. An n-order

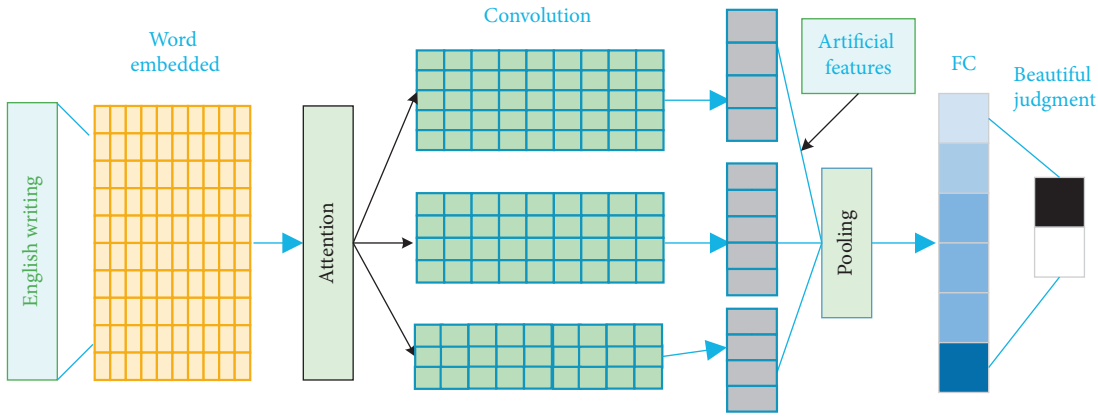

Figure 3: AES network model based on CNN.

weighting matrix w is then calculated based on the difference between the marks scored by the two markers. After calculating the cross product of the marker's scoring histogram vector to obtain the matrix E, the quadratic weighted Kappa value K is solved in the following way:

$$w_{i,j} = \frac{(i-1)^2}{(N-1)^2},$$
$$K = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}. \tag{6}$$

Then, the approximate variance stable Fischer transformation is carried out to obtain Z. Finally, after taking the mean value of Z, the inverse Fischer transformation is carried out to obtain the final average quadratic weighted Kappa value.

$$Z = \frac{1}{2}\ln\frac{1+k}{1-k},$$
$$K = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}. \tag{7}$$

The evaluation model still adopts the conventional neural network evaluation indexes: accuracy rate, recall rate, and $F1$-score.

$$AR = \frac{TP + TN}{TP + TN + FP + FN},$$
$$R = \frac{TP}{TP + FN} \times 100\%, \tag{8}$$
$$F1 = \frac{2(TP + FP) + (TP + FN)}{TP} \times 100\%.$$

4.2. The Experimental Contrast

4.2.1. Comparison of Basic Experiments. The baseline methods are GBRT [34] model based on linguistic features and LSTM model based on word2vec word vector features, denoted as A1 model and A2 model, respectively. The former is an AES model based on machine learning and manual feature extraction, while the latter uses neural network to extract features and train the model. GBRT is an integrated learning method whose base model is regression tree. Gradient descent approximation method is used to fit the residual term of the lifting tree, which is considered to be one of the machine learning algorithms with strong generalization ability. The average quadratic weighted Kappa value of GBRT model A1 based on linguistic features is 0.63, and that of LSTM model A2 based on word vector features is 0.7.

In order to verify the reliability of the machine learning English composition automatic scoring algorithm proposed in this paper, we carry out eight feature combination experiments. The purpose is to find the best combination of features suitable for the model through the combination of multiple groups of features, so that the final result is consistent with the actual situation. It is obvious from Figure 4 that the evaluation results of incorporating semantic or linguistic features into the baseline model are improved to some extent compared with the baseline model. The word vector clustering feature (C1) has the most obvious improvement effect, while the improvement effect of feature combination 2 is not obvious after the addition of theme feature (C3). The comparison of feature combination 5 and 7 shows that the model effect even decreases after the addition of theme feature, and the average quadratic weighted Kappa value is the lowest in the whole experiment. This is because this paper did not make a careful analysis and comparative experiment on the selection of the number of topics when extracting the theme features. In experimental group 5, the performance of the model improved greatly after the inclusion of linguistic features, reflecting the important role of lexical features and syntactic features in AES.

4.2.2. Experimental Comparison of Beauty Sentence Evaluation. Convolutional neural network can extract deep sentence features well, so three groups of feature type experiments are carried out (artificial feature extraction A, CNN feature extraction, and artificial + CNN feature extraction ACNN). The trailing + - indicates whether it is a beautiful sentence. From the overall trend of accuracy and recall rate in Figure 5, the experimental results of these three

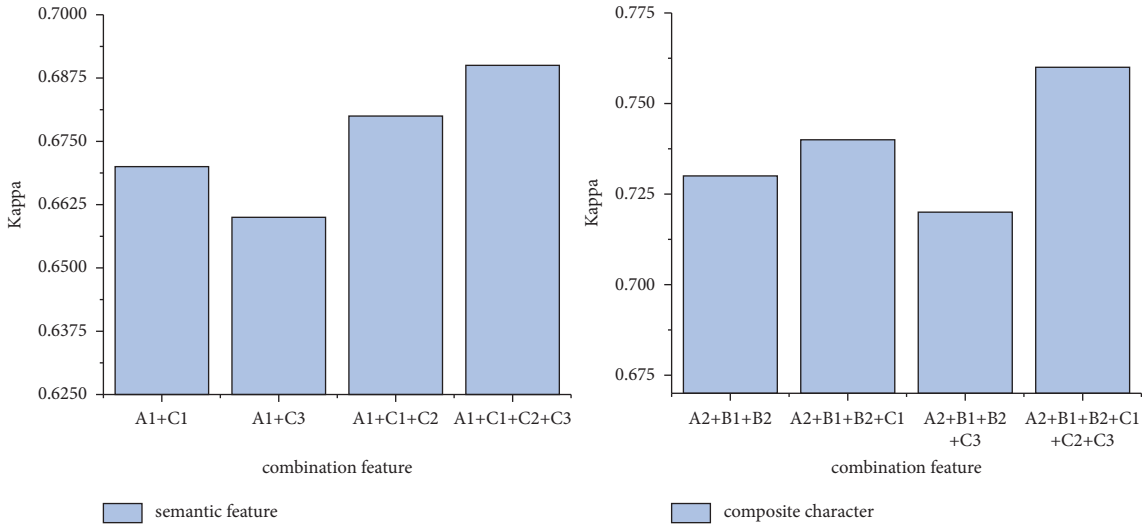

Figure 4: Feature fusion experiment. *Note*. Linguistic features: B1 lexical features and B2 syntactic features; semantic features: C1 word vector clustering feature, C2 text vector feature, and C3 topic feature.

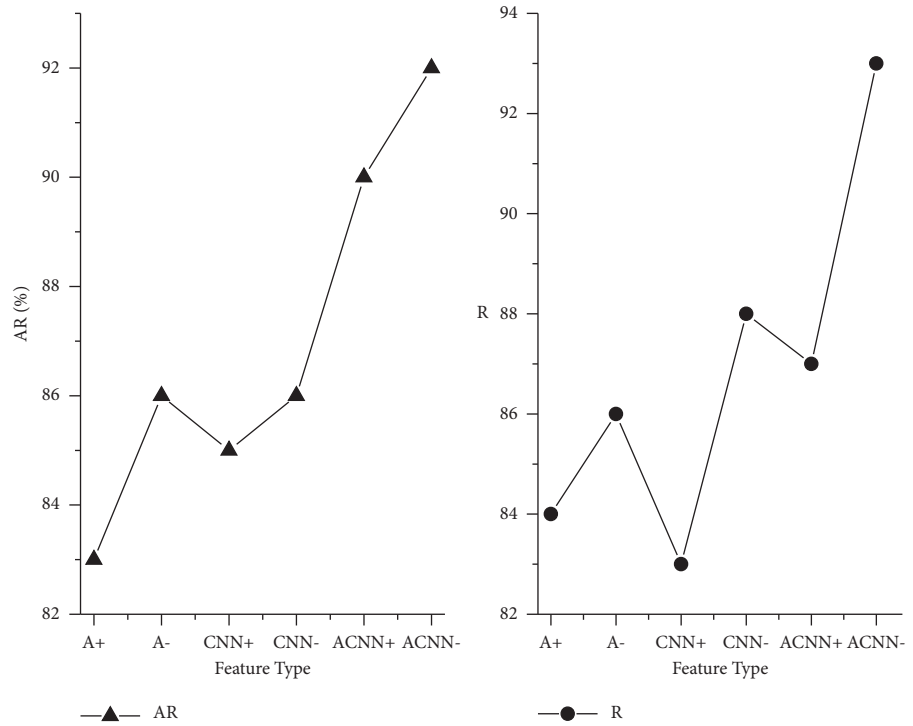

Figure 5: Sentence elegance judgment based on different feature combinations.

groups are basically similar; what is different is that the experiment accuracy rate of artificial feature extraction of beautiful sentence is the lowest, that is, only 83%. And the lowest recall rate is CNN feature extraction for the judgment of the beautiful sentences, and the experimental results and characteristics of the same type of beautiful sentence are always lower than the experimental results of unbeautiful sentences. It can be concluded that there is still a bias in the algorithm for the definition of beautiful sentences, and sometimes, there is a certain difference in people's judgment of beautiful sentences in practical application.

According to the F1 evaluation index in Figure 6, the graceful sentence is still lower than the ungraceful sentence, which further verifies our previous test and analysis. It can be seen from the experimental data that, compared with the single manual definition of features and the method of extracting sentence features only using CNN, combining the two feature extraction methods can greatly improve the classification effect of beautiful sentences and unbeautiful sentences. At the same time, the evaluation index of beautiful sentences in each category is smaller than that of nonbeautiful sentences. It can be seen that the judgment of

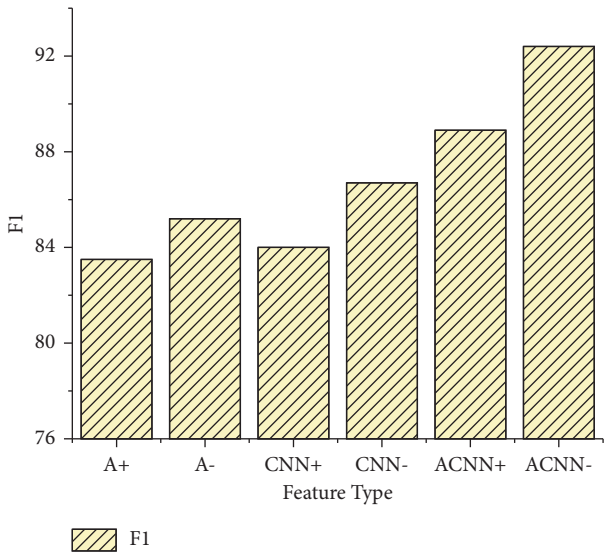

Figure 6: Sentence elegance accuracy and recall rate under different feature combinations.

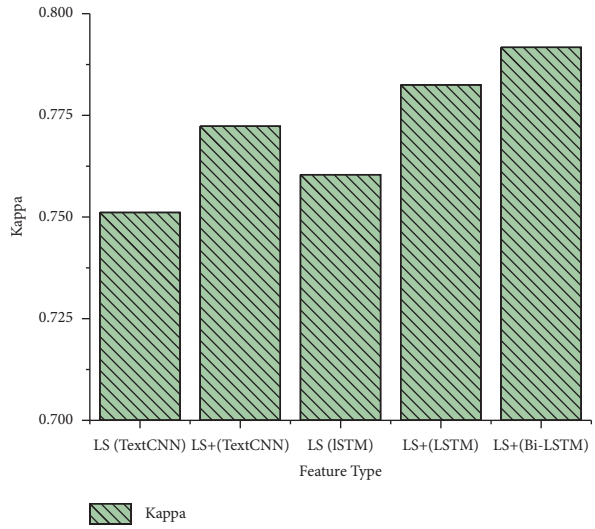

Figure 7: F1 value of sentence elegance under different feature combinations.

sentence beauty is a difficult point, which is also the key point to distinguish the writing level.

In order to further improve the experiment, we not only consider the combination of machine learning feature extraction but also try to analyze the experimental results of different network classification for different feature combinations from the perspective of network learning (LS represents the combination of linguistic and semantic features, and + represents the introduction of sentence elegance features). It can be clearly seen from the experimental results in Figure 7 that the model incorporating sentence elegance features has better performance than the text-CNN and LSTM models before inclusion. Meanwhile, when the feature types are the same, the performance of LSTM model is better than that of Text-CNN, showing the former's excellent memory learning ability in Text mining. The improved BiLSTM based on LSTM has achieved the best result of the second experiment, mainly because it can better discover the time influence of different features from different directions and get better sentence features with time validity.

5. Conclusions

Aiming at the problems existing in automatic grading of English compositions in English learning, this paper studies AES algorithm based on machine learning feature extraction. Based on linguistic features, semantic features, and their extraction methods, the paper studies the effect of linguistic features and semantic features used in AES on the model. Based on this, the BiLSTM basic automatic scoring algorithm and the improved CNN automatic scoring algorithm for beautiful sentence judgment are discussed. The experimental results show that the combination of linguistic features and semantic features is better, and the combination of a variety of features is used to analyze the performance of automatic composition scoring to select the best model through the combination of features. This paper aims to reduce the teacher's burden of assessment and uses it as the starting point, and also to improve the quality of teaching and students' English writing level; a complete set of improved English composition automatic scoring system has been designed through the study and discussion of English composition scoring. [35].

Data Availability

The datasets used in this paper are available from the author upon request.

Conflicts of Interest

The author declares no conflicts of interest regarding this work.

Acknowledgments

The study is the phased achievement of "the curriculum ideological and political demonstration class construction project" of Department of Education of Liaoning Province, China (Grant no. 2021-01380).

References

[1] F. Sharifian, "Globalization and the development of metacultural competence in English as an international language learning," *Multilingual education*, vol. 3, no. 1, pp. 1–11, 2013.

[2] Y. Guo, "Cultivation of learners' subjective awareness in English writing teaching," *Teaching and Management: Theory*, vol. 6, no. 7, pp. 100-101, 2009.

[3] Z. Zhu, "Good English composition is an important embodiment of improving students' language application ability," *North-south Bridge*, vol. 9, no. 4, p. 184, 2015.

[4] Y. Ye, "Weak links in writing teaching from the perspective of language errors in students' exercises," *Foreign Language Teaching,*, vol. 12, no. 4, pp. 77–81.

[5] D. Boulanger and V. Kumar, *Deep Learning in Automatic Paper Grading*, Springer, Berlin, Germany, 2018.
[6] E. B. Page, "The imminence of grading essays by computer," *Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966.
[7] K. Landauer Thomas, L. Darrell, and F. Peter, "The intelligent essay assessor," *IEEE Intelligent Systems & Their Applications*, vol. 15, 2000.
[8] P. W. Foltz, "Latent semantic analysis for text-based research," *Behavior Research Methods, Instruments, & Computers*, vol. 28, pp. 197–202, 1996.
[9] R. D. Manning, M. S. Cohen, and R. L. Demichiell, "An overview of current research on automated essay grading," *Journal of Information Technology Education: Research*, vol. 2, 2003.
[10] M. Liang and Q. Wen, "Review and enlightenment of automatic composition scoring system in foreign countries," *Audio-visual Teaching of Foreign Languages*, vol. 7, no. 5, p. 7, 2007.
[11] Y. Wang, Z. Li, and Y. He, "Key techniques for automatic essay scoring based on semantic dispersion of text," *Journal Of Technology*, vol. 30, no. 6, p. 9, 2016.
[12] T. Qiu, "Research and implementation of discourse fluency assessment techniques in intelligent English composition assessment," *Beijing University of Posts and Telecommunications*, vol. 1533, no. 3.
[13] M. Liu, B. Qin, and T. Liu, *Intelligent Computers and Applications*, vol. 6, no. 1, pp. 1–4, 2016.
[14] W. Yangwei and X. Huang, "Automatic scoring of English composition based on linguistic features and self-encoder," *Application of Computer Systems*, vol. 26, no. 1, p. 8, 2017.
[15] c. Zhao and y. wang, "An image optimization classification method based on word bag model," *Journal of Electronics and Information Technology*, vol. 34, no. 9, 2012.
[16] Q. Guo, Y. Li, and Q. Tang, "Research on text similarity calculation based on VSM," *Application Research of Computers*, vol. 25, no. 11, pp. 3256–3258, 2008.
[17] P. Tao and L. Lu, "PU text classification enhanced by term frequency-inverse document frequency-improved weighting," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 3, pp. 728–741, 2013.
[18] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, 2003.
[19] T. Mikolov, K. Chen, and G. Corrado, "Efficient estimation of word representations in vector space," in *Proceedings of the International Conference on Learning Representations (ICLR 2013*, Scottsdale, AZ, USA, May 2013.
[20] E. Emnlp, *Empirical Methods in Natural Language Processing*, 2014.
[21] M. Peters, M. Neumann, M. Iyyer et al., "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, June 2018.
[22] J. Devlin, M. W. Chang, and K. Lee, "BERT: pre-training of deep bidirectional, transformers for language understanding," 2018, https://arxiv.org/abs/1810.04805.
[23] Y. Kim, "Convolutional neural networks for sentence classification," 2014, https://arxiv.org/abs/1408.5882.
[24] Y. Liu, X. Qiao, and S. Zhao, "Deep Fusion of large-scale features in statistical machine translation," *Journal of Zhejiang University (Engineering Science)*, no. 1, 2017.
[25] A. Hassan and A. Mahmood, "Convolutional recurrent deep learning model for sentence classification," *IEEE Access*, vol. 6, p. 1, 2018.
[26] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[27] W. Wang, L. Wang, and Y. Chai, *Application Research of Computers*, vol. 36, no. 5, p. 5, 2019.
[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[29] Z. Zeng, L. Li, and J. Chen, "Bidirectional depth LSTM for emotion classification," *Journal of Computer Science*, vol. 45, no. 8, p. 6, 2018.
[30] H. Liu, S. Liu, and X. Zhang, "An optimized K-means text feature selection in clustering mode," *Computer Science*, vol. 38, no. 1, p. 3, 2011.
[31] Z. Zhou, "Machine learning," *China Civil And Commercial*, vol. 3, no. 21, p. 93, 2016.
[32] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, 1997.
[33] Kaggle, "The Hewlett Fewlett foundation:automated essay scoring," 2019, http://www.kaggle.com/c/asap-aes/data.2012-2-10.
[34] J. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, 2001.
[35] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Annals of Applied Statistics*, vol. 1, 2001.