Research Article

Intelligent Scoring of English Composition by Machine Learning from the Perspective of Natural Language Processing

Dan Zhang¹ and Xiaorong Yuan²

¹School of General Education, Hunan University of Information Technology, Changsha, Hunan, China
²Department of Computer Science, Hunan Urban Profession College, Changsha, Hunan, China

Correspondence should be addressed to Xiaorong Yuan; 120212202031@ncepu.edu.cn

Received 5 January 2022; Accepted 1 March 2022; Published 6 May 2022

Abstract

To eliminate the shortcomings of the current English composition scoring, this study intends to propose an automatic scoring model by artificial intelligence (AI) technology—machine learning (ML)—from the perspective of natural language processing (NLP). According to the sentence judgment criteria and scoring criteria of English composition, and with a combination of the n-gram model and decision tree (DT) algorithm as technical support, a composition scoring system is established with the content of the composition as the extraction feature. The random forest algorithm is taken as the training algorithm of the model. By using different maximum feature extraction numbers to test the model using the random forest training method, it is found that when the number of features is 11, the model can play the best prediction performance. Similarly, different learning rate values are used to verify the composition scoring rate of the model. It is found that when the learning rate is 6E-6, the model can play the best scoring performance. Therefore, it is believed that this automatic composition scoring model can be applied to English composition scoring in schools, and some theoretical concepts are put forward for intelligent education. The proposed composition scoring system by N-gram model and DT algorithm and trained by the random forest algorithm has high performance, which can be applied in the automatic scoring of students’ English composition in schools, and provides a basis for the educational applications of ML under AI.

1. Introduction

Natural language processing techniques have been used in the literature by numerous researchers to extract features from essays in the datasets. In recent years, the number of relevant studies on AI technology—machine learning (ML)—from the perspective of natural language processing (NLP). According to the sentence judgment criteria and scoring criteria of English composition, and with a combination of the n-gram model and decision tree (DT) algorithm as technical support, a composition scoring system is established with the content of the composition as the extraction feature. The random forest algorithm is taken as the training algorithm of the model. By using different maximum feature extraction numbers to test the model using the random forest training method, it is found that when the number of features is 11, the model can play the best prediction performance. Similarly, different learning rate values are used to verify the composition scoring rate of the model. It is found that when the learning rate is 6E-6, the model can play the best scoring performance. Therefore, it is believed that this automatic composition scoring model can be applied to English composition scoring in schools, and some theoretical concepts are put forward for intelligent education. The proposed composition scoring system by N-gram model and DT algorithm and trained by the random forest algorithm has high performance, which can be applied in the automatic scoring of students’ English composition in schools, and provides a basis for the educational applications of ML under AI.
review composition scoring scale, Baethge et al. summarized and found that the explanation of composition importance, the purpose of review, reference and presentation, evidence level, and results were very important for the evaluation of composition structure [7]. In the research field of automatic scoring of English composition, there have been research results and theories using various intelligent technologies. For example, Ramesh and Sanampudi found that it was a great challenge to evaluate a composition by considering the correlation between content and prompt, the development of ideas, cohesion, and coherence. However, few researchers focused on content-based assessment, while many researchers focused on style-based assessment [8]. Ghanta developed a machine learning model for performing automated essay scoring and evaluating their performance. In this research, the available essay dataset were used for testing and training the efficiency of the considered techniques. Feature extraction form essays in the datasets were performed using natural language processing techniques [9].

Shin and Gierl compared the protocols between human evaluators when studying the establishment of a composition scoring system using vector machine (VM) and convolutional neural network (CNN). The results show that the model performs better, which means that the model produces more comparable results with human evaluators [10]. Beseiso and others proposed a neural network model by a transformer. By deep learning (DL), the composition scoring results are compared with human raters. It is concluded that the proposed model is superior to the existing composition scoring methods in terms of scores. The analysis proves the applicability of the proposed model in automatic composition scoring at the level of higher education [11]. From the above research, it can be seen that the current automatic composition scoring system is mostly by the combination of neural network and DL technology. Although there are many studies, there is no breakthrough.

Nevertheless, this study intends to design and implement an automatic scoring system for English composition using innovative technology. This proposed algorithm capitalizes on the weaknesses of the existing algorithms and presents a system that improves the accuracy and efficiency in the context of natural language processing. From the perspective of NLP, this study will use a language processing model of Chinese language model n-gram, introduce methods in the scoring of Chinese composition into the scoring of English composition, and propose a scoring model for text content feature extraction combined with the DT in ML. The random forest algorithm is used to train the model, the quadratic weighted value is used to judge the feasibility of the algorithm, and then different learning rates are used to judge the learning ability most suitable for different situations of the model, and the learning situation most suitable for the model is selected to provide some opinions for solving the subjective impact of students’ composition scoring.

2. Materials and Methods

2.1. The Scoring Characteristics of English Composition. As the composition is a subject with a large proportion of subjectivity, it can reflect students’ comprehensive ability in the examination, including logical expression ability, correct writing ability of words and sentences, and their mastery of composition structure [12], which can express and convey their views according to open topics. Generally speaking, the topics of English composition have several characteristics as shown in Table 1 [13].

By summarizing the contents of Table 1, it can also be precisely concluded that when scoring English compositions, it is necessary to evaluate the content and structure of the compositions from four aspects as shown in Table 2 [14]. Therefore, the scoring of English composition is very subjective, excluding the consideration of whether the spelling of sentences and words is correct. At present, most of the scoring standards in the examination are from the perspective of linguistics and do not give a set of specific scoring rules from the aspect of content expression. Therefore, the scoring of English composition generally needs to be carried out from the scoring process of experts [15]. As shown in Figure 1, the expert scoring structure is widely used at present.

2.2. Related Concepts of NLP. In the field of NLP [16], lexical analysis is an important work, which is generally completed by using part of speech tagging technology. Analyzing the syntax of sentence patterns can analyze the number of words or phrases, and using filter words can remove redundant or inconsistent words [17]. Figure 2 illustrates the parsing process of the following sentence “my dog also loves playing ball.”

In Figure 2, ROOT represents the next sentence to be analyzed, NP and NN mean nouns and common nouns, VR indicates verbs or phrases, ADVP stands for adverb phrases, PRP denotes person, RB refers to adverbs, and VBZ and VBG point to verbs, to judge the mastery of the pattern of this sentence.

2.3. N-Gram Language Model. N-Gram is a model that statistics language through probability [18]. The analysis method is to predict and analyze the last word and sentence through the previous word and sentence. Given a sequence of “I love,” according to the prediction results of the model and the prediction results according to the maximum likelihood value, such as “you =40%” and “her = 1%,” these probabilities add up to 100%. Given a sentence pattern as \( s = w_1 w_2 \ldots w_n \), the following equation solves the probability of this sentence:

\[
p(s) = p(w_1)p(w_2 | w_1)p(w_3 | w_1 w_2) \ldots p(w_n | w_1 \ldots w_{n-1}).
\]

(1)

The parameters \( w_1, w_2, \ldots, w_n \) represent each word in the sentence. To solve the probability of the occurrence of a specific sentence pattern, \( n \) parameters must be estimated. Due to the increasing number of subsequent words, the parameters involved in the calculation will also increase, and the calculation process will be very complex. Therefore, the Markov hypothesis [19] is introduced to make the solved
words related to the previous words, and (1) can be changed as follows:

$$p(s) = \prod_{i=1}^{n} p(w_i | w_1 w_2 \ldots w_{n-1}) \approx \prod_{i=1}^{n} p(w_i | w_{i-2} w_{i-1}).$$  \hspace{1cm} (2)$$

The following equation demonstrates how to solve the parameter \(p(w_i | w_{i-2} w_{i-1})\) in (2):

$$p(w_i | w_{i-2} w_{i-1}) \approx \frac{c(w_{i-2} w_{i-1} w_i)}{c(w_{i-2} w_{i-1})},$$  \hspace{1cm} (3)$$

where \(c(w_{i-2} w_{i-1} w_i)\) represents the number of occurrences of three words \((w_{i-2} w_{i-1} w_i)\) in the corresponding corpus and \(c(w_{i-2} w_{i-1})\) indicates the number of occurrences of two words \(w_{i-2} w_{i-1}\) in the corpus. With a large enough corpus, such as the web, we can compute these counts and estimate the probability from equation (3). For the case where the probability is 0 due to the nonoccurrence of a single word, the smoothing method is generally used, that is, the method of reducing the frequency of the existing words, and then assigning the reduced probability to other nonoccurrence words. Theoretically, the more words in the calculation of the model, the better the effect. But the amount of calculation also increases gradually. The basic intuition of the n-gram model is that instead of computing the probability of a word given its entire history, we can approximate the history by just the last few words of a text.

### 2.4. Machine Learning

Before we go deep into the details of how to apply these machine learning and AI tools in NLP and text analysis, let’s clarify certain basic ideas. Machine
learning for NLP and text analytics involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of the text. The techniques can be expressed as a model that is applied to other text, also known as supervised machine learning. The most common ML methods for NLP are as follows: support vector machines, decision tree, and neural networks. ML [20] generally involves knowledge disciplines in many fields, which is work that needs experience. The “experience” here is generally expressed in the form of data. Therefore, ML is an algorithm that reflects the model through data, that is, a learning algorithm, which includes a learning algorithm called DT [21], which is summarized by name. DT is a decision algorithm similar to a tree structure. Case expression is whether an English composition is excellent or not, which is judged by a variety of factors (content, logical expression, sentence smoothness). Figure 3 signifies the structure of DT, which uses leaves and branches to represent the connection between nodes.

The basic work of the DT is to select the basic root node, which can accommodate all training samples, and then select the optimal characteristics [22]. According to the sample characteristics, the training set samples are divided into subsets, so that each subset is the optimal node under specific conditions. If it can be grouped, it can correspond to leaf nodes; if not, it needs to be further divided and divided again to form nodes. The selection of optimal features is usually referred to by information gain.

Information gain is generally referred to by entropy. Assuming that $X$ should be a random variable with limited values, if $P(X = x_i) = p_i$, $i = 1, 2, \ldots, n$, the entropy can be defined by

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i. \quad (4)$$

Since the value of entropy is only related to the $X$ distribution [23] and has nothing to do with the specific value, $X$ can be changed to $h (P)$, as signified by

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i. \quad (5)$$

Because the information gain is the reduced value of information complexity under certain conditions, if the information gain $g(C, B)$ of feature $B$ to training sample set $C$ is the information entropy $H(c)$ of $D$ minus the difference of entropy $H(C | B)$, then the following equation describes the relations between them:

$$g(C, B) = H(C) - H(C | B). \quad (6)$$

Another method for selecting the optimal features is the Iterative Dichotomiser 3 (ID3) algorithm [24], which is a process of selecting features through information gain. Table 3 lists its calculation steps.

To avoid the phenomenon that the ID3 algorithm tends to favor feature nodes with more values, the information gain is taken as the solution. The ratio of the information gain $g(C, B)$ to the entropy $H_B(C)$ of the training example set $c$ about the value of feature $B$ is the corresponding information gain ratio $g_R(C, B)$. The following equation illustrates the process:

$$g_R(C, B) = \frac{g(C, B)}{H_B(C)} \quad (7)$$

The following equation expresses the calculation of $H_B(C)$:

$$H_B(C) = -\sum_{i=1}^{n} \frac{|C_i|}{|C|} \log_2 \frac{|C_i|}{|C|}. \quad (8)$$

Parameter $n$ in (8) is the number of characteristic $B$ values.
2.5. Feature Extraction from Text Content. Text content feature extraction [25] generally uses vector space as the basic method. Vector space assumes that the composition content is only related to the occurrence frequency of specific sentence units in the text, but independent of position and order; that is, the composition content representation of vector space is calculated through the sentence unit frequency. In a paragraph of text \( d \), there are \( n \) different
feature items \( f_1, f_2, \ldots, f_n \). The text can represent the feature item set \( d(f_1, f_2, \ldots, f_n) \), \( f_i \) refers to the feature item, \( 1 \leq i \leq n \), and the feature items are irrelevance and independent of order. The calculation is implemented on the feature item weight \( w_i \) of \( f_i \) in the text, and \( f_1, f_2, \ldots, f_n \) is taken as an \( n \)-dimensional coordinate system. Then, \( w_1, w_2, \ldots, w_n \) are the corresponding dimension values, so text \( d \) is taken as an \( n \)-dimensional vector with \( w_1, w_2, \ldots, w_n \) in the \( n \)-dimensional space. The following equation defines the calculation of \( d \):

\[
\vec{V}_d = (w_1, w_2, \ldots, w_n),
\]

(9)

where \( \vec{V}_d \) indicates the vector of text \( d \) in \( n \)-dimensional space.

In vector space, the most widely used way to get the weight is through the term frequency-inverse document frequency method. Term frequency (TF) and inverse document frequency (IDF) [26] are used to judge the importance of words. If the same word appears more frequently in the same composition, it appears less frequently in other compositions. Then, the larger the IDF value, and vice versa, the smaller the IDF. In the composition \( d \), if the \( k \)th characteristic term \( t_{i,k} \) appears \( n_{i,k} \) times in the composition, its TF function can be expressed by

\[
TF_{i,k} = n_{i,k}.
\]

(10)

However, because different compositions have different lengths and numbers of words, to unify the standard, it is necessary to standardize the frequency of words. The following equation is used to realize the standardization:

\[
TF_{i,k} = \frac{n_{i,k}}{\sum_{m}^{m} n_{m,k}}.
\]

(11)

The number of IDF values indicates the frequency of occurrence of characteristic terms \( t_{i,k} \) in text \( d \). The following equation demonstrates its calculation:

\[
IDF_{i,k} = \log \left( \frac{|D|}{|\{d : t_{i,k} \in d\}|} \right).
\]

(12)

Assuming there are \( M \) compositions, where the characteristic items \( t_{i,k} \) appear in the \( m_{i,k} \) compositions, then there is

\[
IDF_{i,k} = \log \left( M \left( m_{i,k} + \alpha \right) \right).
\]

(13)

Parameter \( \alpha \) is the constant of experience, generally taken as 0.01. For the characteristic terms \( t_{i,k} \), the following is used to calculate its characteristic weight:

\[
w_{i,k} = TF_{i,k} \times IDF_{i,k}.
\]

(14)

TF-IDF method can comprehensively consider the discrimination ability and frequency of feature items, and professional vocabulary can more clearly express the theme of the composition. Figure 4 manifests the flow diagram of this feature extraction algorithm using vector space.

2.6. The Training Method for Random Forest. The random forest training method is an extended form of bootstrap aggregation (Bagging) [27]. It uses the funded sampling technology to select \( n \) samples from the training sample set in the way of taking out and putting back, to form a new sample training set to train the model, and then generates \( n \) DTs to form the random forest model and test the results of the data. In essence, it is an optimized version of the DT algorithm. Each DT in the random forest is formed by using independent samples, and then, these trees are combined concurrently. The computing power of a single DT is weak, but after a large number of DTs forming groups, each sample will choose the result with the highest probability after passing through any
Table 4 signifies the establishment steps of random forest.

Figure 5 illustrates the scoring structure for English composition using random forest training, in which the training sets are Train_{xk} and Train_{yk}, and the number of DTs in test set Test_{xN2,M} is all K.

The following equation aims at calculating the mean value of the results from the DTs in Figure 5.

$$\text{Test}_{y_i} = \frac{1}{K} \sum_{k=1}^{K} T_k(\text{Test}_i).$$

(15)

Here, K is the number of DTs.

3. Experimental Results and Discussion

3.1. Verification of the Effect of Feature Number on the Training Performance of the Model. In this link, the prediction accuracy effect of the model under the random forest training algorithm is tested by changing the maximum number of features in the model. The number of feature numbers is 5, 7, 9, 10, 11, 12, 14, 16 parameter feature numbers, which allows a single DT to use the maximum number of features. The accuracy value is judged by the quadratic weighted value. Figure 6 presents the test results.

Figure 6 reveals that with the increase in the maximum number of features, the accuracy of the sentence prediction model under the random forest training algorithm initially increases and then decreases. When the maximum number of features reaches 11, the prediction accuracy of the whole model is the highest, which is 0.8983. When the maximum number of features continues to increase, the prediction accuracy of the model declines. The reason is that when the maximum number of features is at a small value, increasing the number of features can make each DT in the model have more optional features, and the prediction ability of the DT will be enhanced. However, when the number of features of the DT reaches the maximum range, the difference of each DT will be reduced. Therefore, the prediction ability of each DT begins to decline. Therefore, to ensure the maximum prediction ability and accuracy of the model, the maximum number of features should be kept at about 11. To further enhance the prediction ability of the model, there are three strategies proposed here. They are to further enrich the sentences in the corpus, and the more training samples, the better the prediction effect of the model. Adding other language or grammar detection modules, the current prediction model cannot detect grammar deeply. Other AI technologies (neural network, neural network, DL) are introduced for further training and prediction. The use of a neural network can avoid manual feature extraction and improve the accuracy of the model from the feature extraction part.

3.2. The Scoring Rate Test of the Model for English Composition under Different Learning Rates. In this link, different learning rates are used to test the scoring effect of the model on a total of seven compositions in English composition scoring. The learning rates are 4E-5, 6E-6, and 6E-7, respectively. The secondary weighted value is still used to judge the scoring rate of the model. Figure 7 displays the test results.

Figure 7 implies that when using the scoring model to score seven English compositions under three different learning rates, through the test, when the learning rate is 6E-6, the scoring rate of the model reaches the maximum and shows the best performance, and its average scoring rate reaches 0.81. If the learning rate is smaller or greater than 6E-6 is selected for scoring, the scoring performance of the model has decreased. If the learning rate is too large, the model may not be able to converge and score English compositions. Therefore, with the learning rate of 6E-6, the scoring performance of the model can reach the maximum. The present work proposes two measures to improve the performance of this link. The first is to establish a sentence...
corpus specially applied to the model to enrich the sample size; the second is to study feature extraction in different aspects from the perspective of linguistics.

4. Conclusion

There are problems that the scoring criteria for students' English composition are inaccurate, the scoring rules are too subjective, and most of the scoring criteria take the composition content as the starting point. To put forward a scoring system for English composition, initially, the present work discusses the characteristics of grammar, sentence, and structure of English composition and expounds on the scoring criteria for English composition. By the n-gram model, combined with the DT algorithm under ML, a composition scoring system for text content feature extraction is established, and then the random forest algorithm is introduced to train the model. Then, the present work extracts a large number of non-text features that can reflect the quality of students' compositions. Simultaneously, according to Bai Jil parameter adjustment experience, a set of ML model parameter adjustment methods is summarized, so that the model parameters can be adjusted to the optimal direction, efficiently and quickly, and further improve the prediction results of the model. In the following test link, different maximum feature numbers are used to test the prediction performance of the model under the random forest training algorithm. It is found that the feature number with the value of 11 can make the model play the best prediction ability. Then, different learning rate values are used to test the performance of the English composition scoring rate of the model. The results show that when the learning rate is 6E-6, the model can play the best scoring performance. Therefore, the conclusion is that this composition scoring system by N-gram model and DT algorithm and trained by random forest algorithm has high performance, which can be applied in the automatic scoring of students' English composition in schools, and provides a certain basis for the educational applications of ML under AI. Besides, the shortcomings are in the first part of the results. At present, the model has insufficient judging ability for grammar, which needs to be further optimized. In the follow-up work, attention will be made to introducing neural networks and other technologies to improve the performance of grammar scoring.

Data Availability

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

Conflicts of Interest

The authors declare no conflicts of interest.

References


