Research on the Automatic Scoring Method of English Essay based on the Improved K-Nearest Neighbor Algorithm

Hao Jiang[1, a], Yaru Jin[2, b]

[1 , 2] School of Computer Science and Engineering, Southeast University, Nanjing, 211189, China

[a]email: hjiang@seu.edu.cn, [b]email:yaru_jin0065@163.com

Keywords: Automatic Essay Scoring; LSA; Information Gain; K-Nearest Neighbor

Abstract. Compared with the traditional manual marking, automatic English essay scoring can improve the consistency, objectivity, and efficiency of the scoring process. In this essay, the relevant attributes of English composition is extracted, and the improved KNN algorithm is used to score the English essay. The experimental results show that the automatic scoring in which the improved KNN method combines with feature selection has smaller error, compared with manual scoring, and the accuracy of scoring has been improved significantly.

Introduction

The automatic essay scoring systems available on abroad adopt different core technology, which cause to there is a huge difference in their ability to analyze composition. The Project Essay Grade (PEG) was the first automatic essay scoring system implemented [1], which mainly analyze the linguistic feature of the text blocks. Unlike PEG, the Intelligent Essay Assessor (IEA) is developed on the basis of latent semantic analysis, it is not only able to assess the composition of semantics, but also to evaluate the content of the essay. The Bayesian Essay Test Scoring System (BETSY) is based on the text classification and it is directed by probability theory and it categorizes text based on the training corpus [2][3].The mainstream systems put into use in domestic are Bingo Intelligent English Essay Reviewing System and Jukuu Correcting Network System, they are both able to score essay on the linguistic and content aspects, but they put more emphasis on the words and grammar, so it cannot give an accurate assessment about content and unique individual expression.

In this paper, we obtain the feature items of essays on content and linguistic firstly. And then select the most important features according to their information gain to measure the distance between essays. Finally, we put greater weight to the nearer neighbors to calculate the score of essay. The experimental results show that the improved KNN method combines with feature selection has a higher accuracy in automatic scoring.

Content Processing Model

In general, the vector space model can be used to represent text. The text D_i can be represented by t dimensional vector: $D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$, wherein the d_{ij} is the times of the jth feature item appearing in the text D_i. The similarity between texts can be obtained by calculating the cosine similarity of the feature vectors, and the degree of similarity of the texts depends on the number of words they sharing. However, there is synonymous and polysemy phnomenon of words, making this method of calculating document similarity not accurate enough.

Latent Semantic Analysis (LSA) was proposed in 1990 by Scott Deerwester, Susan T. Dumais, it is a new indexing and retrieval methods. It can analyze large amounts of text sets using statistical methods to extract and express the semantics of words. Its result vector no longer reflects the frequency and the distribution of words, but enhanced semantic relations [4].

The detailed procedure of LSA algorithm is shown below:

Step 1 (Analyzing document sets and establishing Term-Document matrix): This paper uses χ^2(chi-square test) to select feature words. The evaluation function is used to assess each word in the initial vector to obtain an evaluation score, and then all of the words are sorted on descending order according to their evaluation scores, the first k words are fetched as the feature words. We can construct a Term-Document matrix X utilizing the text and feature words.

Step 2 (The singular value decomposition (SVD) to Term-Document matrix): Singular value decomposition theory is the mathematical basis for latent semantic analysis, it can decompose any matrix into the product of other three matrices: $X = T_0 S_0 {D_0}^T$, Wherein T_0 and D_0 are orthogonal matrix which satisfy the criteria:${T_0}^T T_0 = T_0 {T_0}^T = I_t$, ${D_0}^T D_0 = D_0 {D_0}^T = I_d$, I_t and I_d are identity matrixes, and their order are t and d, S_0 is square matrix, and assuming its nonzero diagonal elements arranged in descending order.

Step 3 (Reducing the dimensionality of the matrix X): For the matrix S_0 gained from the SVD, we fetch its first k diagonal elements to form a new diagonal matrix S, and fetch the first k diagonal elements from matrix T_0 and D_0 to form new diagonal matrix T and D corresponding. Thus there is an approximate matrix $\hat{X}$ of X, $\hat{X} = TSD^T$, wherein the order of the matrix $\hat{X}$ is k, and it is a best approximation matrix to matrix X.

Step 4 (Calculating the Correlation Matrix): The correlation between texts can be obtained by calculating their cosine similarity according the Term-Document matrix.

This paper uses the Correlation Matrix obtained by LSA method to measure the intrinsic correlation degree of context in a same essay. The correlation degree of context includes two aspects, which are between paragraphs and within paragraph.

The inherent correlation degree R of the whole composition can be worked out as follows:

$$R = R_e + \sum_{i=1}^{n} r_{p_i} = \sum_{i=1}^{n-1} t_i\, h_{i+1} + \sum_{i=1}^{n} \frac{1}{m} \sum_{k=1}^{m-1} r_{k,\ k+1} \tag{1}$$

In formula (1), R_e represents the correlation values between paragraphs, $t_i h_{i+1}$ represents the correlation between the tail sentence of ith paragraph and the head sentence of $(i+1)$th paragraph and r_{p_i} represents the correlation of the ith paragraph, $r_{k,\ k+1}$ represents the correlation between the kth sentence and $(k+1)$th sentence of ith paragraph, m represents the number of sentences of ith paragraph, n represents the number of paragraphs of text.

Through the above calculation, the correlation matrix of an essay can be transformed to a value, which is used to measure the inherent correlation degree of the composition.

Language Processing Model

The main measures in language are word and sentence. For an untreated composition, it needs to be split into individual sentences first of all. And then, the split sentences are separated into words through Lucene Analyzer module, after process of removing stop words, the separated words are queried in WordNet and their usage are counted. On the other hand, the separated sentences are built into syntax tree through Stanford parser module, and then the syntax tree is part-of-speech tagged, finally, the modal verbs and phrases are counted and the information about the breadth and depth of the syntax tree is counted. Figure 1 is a flow diagram of the language processing.

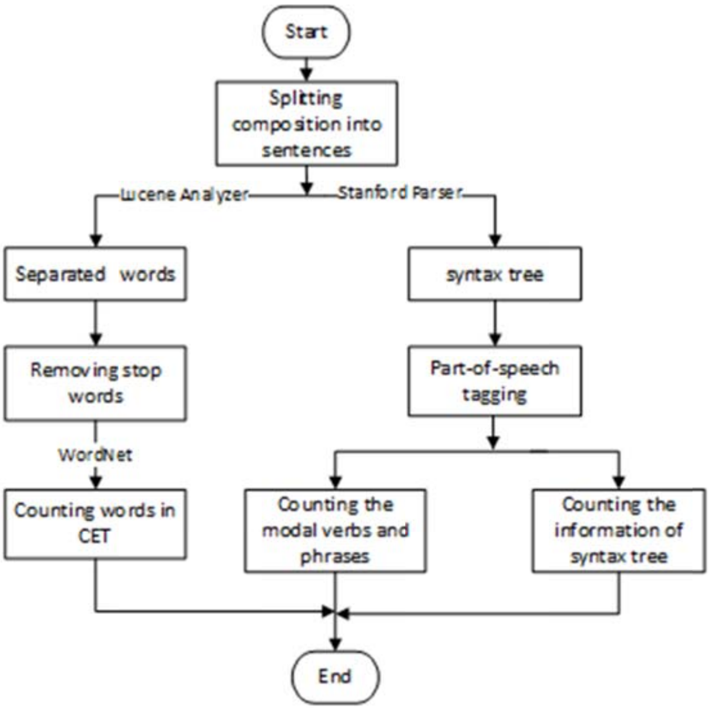

Fig.1. The flow diagram of the language process

Improvement of KNN

The k-nearest neighbor (KNN) is a basic classification and regression method [5]. In the KNN algorithm, the score of the test sample is equal to the average score of the k nearest training samples. The distance between two compositions is defined as formula (2), in which,$a_r(x)$ denotes the value of the rth feature of composition x, we can see that the weight of each feature is equal, and the default value is 1. But in reality, the impact to result may be different for each feature, and the distance between neighbors will be dominated by irrelevant features.

$$d(x_i，x_j) = \sqrt{\sum_{r=1}^{n}\left(a_r(x_i) - a_r(x_j)\right)^2} \tag{2}$$

This paper introduces a decision tree feature selection method - information gain to calculate the weight of features [5]. The information gain indicates the reduction of uncertainty in classifying samples set D due to learning the information about attribute A. The bigger the information gain is, the more important the feature is.The amount of information needed while identifying the category of tuples in D can be calculated as follows:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{3}$$

Wherein, p_i is the probability which a tuple of D belongs to the classC_i, estimated with $|C_{i,\ D}|/|D|$. Supposing there are v distinct value in attribute A: $\{a_1，a_2，\ ...，a_v\}$, and we will divide the sample set D into v subsets according to the value of attribute A, so D will be divided into v subsets:$\{D_1，D_2，\ ...，D_v\}$, wherein D_j represents the element in D having the value a_j on attribute A. Ideally, the tuples with same value a_j would be divided into the same category. However, in the actual division, the tuples with same value a_j may be divided into different categories, in this case, the division need to continue. When the second division happens, the amount of additional information which we need to ensure the classification accurate can be calculated according to formula (4), in which item $|D_j|/|D|$ acts as the weight of the jth partition of the division.

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{4}$$

The information gain of attribute A is calculated as follows:

$$g(D，A) = Info(D) - Info_A(D) \tag{5}$$

The distance calculation equation is formula (6) after the introduction of the information gain.

$$sim(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(v_r(a_r(x_i) - a_r(x_j)))^2} \quad (6)$$

The item x_i represents a marked composition, x_j represents an essay need to score, and v_r represents the weight of the rth attribute. We put larger weights to the attributes with larger information gain.

In KNN method, the training sample have equal impact on the classification result, while in reality, the training sample will have greater impact if the distance between the test sample and training samples is smaller, especially in the case of the distribution of the data set is uneven, and the training sample which far from the test samples is likely to lead to error in the predictions.

Therefore, one of the obvious improvements on the KNN method is weighting the neighbors according to distance, the closer neighbor of test sample x_q will be endowed with greater weight. We can calculate the classification result of the test sample x_q as formula (7):

$$f'(x_q) \leftarrow \sum_{i=1}^{k}\left(\frac{w_i f(x_i)}{\sum_{i=1}^{k} w_i}\right), \quad w_i = \frac{1}{sim(x_i, x_j)^2} \quad (7)$$

Experiment and Result Analysis

The training data set of this experiment is set for 40 GRE essay and the test data is 30 GRE essay, and 4 closest training sample to the test sample are selected, then the score of the test sample will be calculated using the improved KNN method.

Table1. The information gain (IG) of 10 attributes of essays

attributes	Words in CET 4	Words in CET 6	Modal verbs	Sentences without errors	conjunction
IG	0.001	0.013	0.005	0.216	0.009

attributes	Average length of sentences	Depth of sentence structure	Breadth of sentence structure	Length of essay	Intrinsic correlation
IG	0.154	0.154	0.120	0.044	0.296

In this paper, we extract the features of the marked English essay first, and then calculate the information gain (IG) of each attribute. The information gains of 10 attributes about the 20 composition selected randomly are shown in Table1. As can be seen from the table, the weights of the number of hitting CET4 words and modal verb are small, because they are distributed evenly among the classes; and the weight of intrinsic correlation degree is relatively large. Therefore, LSA algorithm is very effective to measure the internal information of composition; while the weight of others properties are relatively average.

We use the residual sum of squares to measure the error between machine scoring and artificial scoring, and the residual sum of squares is calculated as formula (8), wherein, x_i indicates the score of the ith essay in machine scoring, and y_i in manual scoring.

$$S = \sum_{i=1}^{n}(x_i - y_i)^2 \quad (8)$$

Analysis 1 The comparisons among machine scoring using improved KNN with feature selection, machine scoring using original KNN with feature selection and manual scoring.

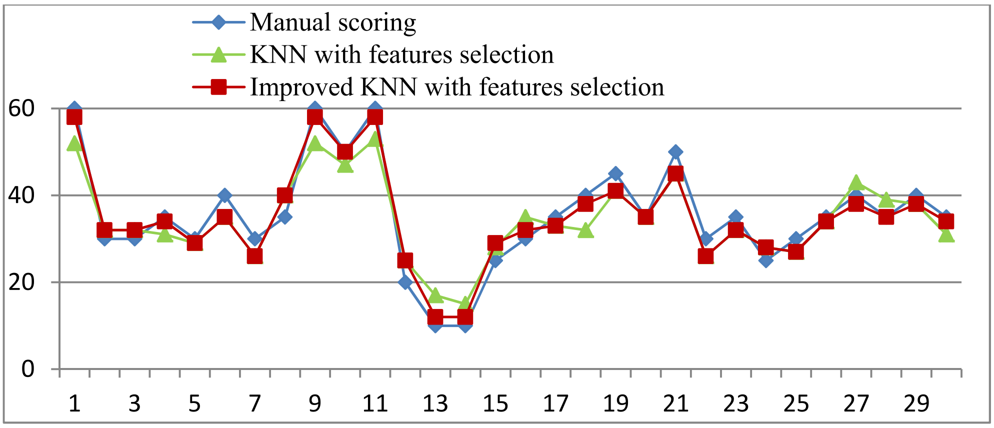

Fig.2. Manual scoring and machine scoring with feature selection

As can be seen from Figure2, the results of machine scoring with improved KNN method can fit the results of artificial scoring better. And the residual sum of squares between the manual scoring and the original KNN is 608, while the residual sum of squares between the manual scoring and the improved KNN is 243, decreased 60.03% comparing to the former. Thus, the improved KNN has superior effect on improving the accuracy of scoring.

Analysis 2 The comparisons among machine scoring using improved KNN with feature selection, machine scoring using improved KNN without feature selection and manual scoring.

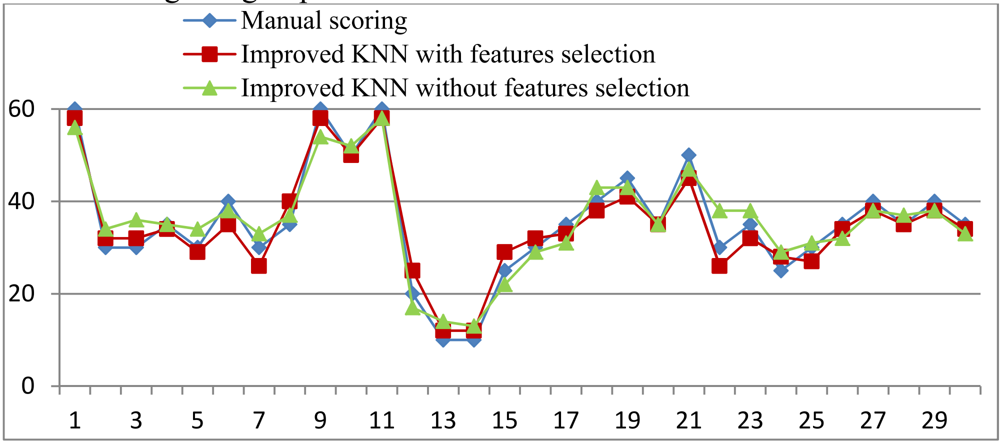

Fig.3. Manual scoring and machine scoring with improved KNN

As can be seen from Figure3, the error at the peak and bottom has decreased obviously compared to Figure2. And the residual sum of squares between the manual scoring and the original KNN is 342, while the residual sum of squares between the manual scoring and the improved KNN is 243, decreased 28.94% comparing to the former. Therefore, the feature selection not only can play the role of dimensionality reduction, but also contribute to improve the accuracy of the scoring at the same time.

Conclusion

In this paper, we adopt latent semantic analysis combined with improved KNN method to score English essay. The Term-Document matrix is mapped into a low-dimensional space by single value decomposition of latent semantic analysis, thus we analyze the correlation of texts on the semantic layer. Considering the shortcomings of the KNN method, this paper improves the method on two aspects: firstly, the distance between test sample and training samples is calculated by weighting the features of essays according their information gain, the other is to weight the scores of k-nearest training samples according their distance from the test sample. The results of the experiment show that the improved k-nearest neighbor with feature selection has smaller residual sum of squares and the effect of the scoring has been improved significantly, which demonstrates the superior performance of this method.

Acknowledgement

This work was supported by Collaborative Innovation Center of Wireless Communications Technology and the National Natural Science Foundation of China (Grant No. 6157060279).

Reference

[1] Page E B. Project Essay Grade: PEG [C]. In: Automated essay scoring: A cross-disciplinary perspective. Mahwah, United States, 2003，43-54.

[2] Rudner L M & T Liang. Automated Essay Scoring Using Bayes' Theorem [J].The Journal of Technology, Learning and Assessment, 2002，1(2):3-21.

[3] Valenti S, F Neri & A Cucchiarelli. An Overview of Current Research on Automated Essay Grading [J]. Journal of Information Technology Education, 2003，2:319-330.

[4] LIN Hong Fei. The Mechanism of Text Title Classification Based on Example [J]. Journal of Computer Research & Development, 2001,38(9):1132-1136.

[5] James G, D Witten, T Hastie & R Tibshirani. An Introduction to Statistical Learning with Application in R[M]. New York: Springer; 2013.