The Research on CET Automated Essay Scoring Based on Data Mining

Hao Jiang[1], Guoqiang Huang[1], and Jiangang Liu[2]

[1] School of Computer Science and Engineering, Southeast University, 211189, Nanjing, Jiangsu, China
hjiang@seu.edu.cn, huangguoqiang0823@163.com
[2] School of Foreign Languages, Southeast University, 211189 Nanjing, Jiangsu, China
jhonliunj@21cn.net

Abstract. At present, the studies in foreign on natural language processing for automated essay scoring are in full swing. However, these studies are aimed at native English speakers, and it is essentially different from the focus on domestic CET essay scoring. With large-scale popularization of the CET automated essay scoring by the Ministry of Education, the problem on essay scoring is becoming the bottleneck of improving efficiency and large-scale popularization. In this paper, from the perspective of data mining, using classification algorithm KNN-based-association, essay is evaluated in both content and language way. Compared with manual scoring, we analyze their difference.

Keywords: Semantic Analyze Vector Space KNN.

1 Introduction

At present, AES (Automated Essay Scoring) studies in foreign have been becoming a hot issue in natural language processing. And now CET using computer is carried out in various universities by Ministry of Education, and English composition, as a large-scale compulsory language test items, can measure the candidate's mastery of language and comprehensive competence. With large-scale popularization of the CET using computer, the scoring of English writing has been becoming the bottleneck of improving the efficiency of large-scale scoring.

To begin with we will provide a brief background on automated essay scoring, then using classification algorithm KNN-based-association, essay is evaluated in both content and language way. Compared with manual scoring, we analyze their difference from the accuracy and efficiency and conclude our work.

2 Analysis of Automated Essay Scoring

Compared with AES, manual scoring is interfered by various factors, which will directly affect the quality of essay scoring. AES not only ensures the efficiency, fairness and justice, but also provides feedback on composition to the candidates in a timely manner to improve the English learning process.

M. Zhou and H. Tan (Eds.): CSE 2011, Part II, CCIS 202, pp. 100–105, 2011.

The scope of this research lies in both content and language, it contains pretreatment on the composition, the feature processing, language analysis, and makes the corresponding related categories forecast, and computes the results of the weighted score, and implements the AES system ultimately.

2.1 Analysis of Essay Content Scoring

There are three method models to score on the essay content: VSM (Vector Space Model); LSA (Latent Semantic Analyze); concept vector space model based on WordNet semantic dictionary. Modeling of the three methods respectively, compare the pros and cons of various methods by the experimental results.

VSM(Vector Space Model): For each of a document, use a vector to express, turn unstructured text into vector form, and use the mathematical approach to deal with text content. However, VSM has a drawback: it assumes it is independence between the text keywords, in practice, it will produce errors for synonyms and polysemy.

LSA (Latent Semantic Analyze): analyze large sets of text using statistical methods, analyze the potential text semantic structure between words, extract and show the semantics of the words, the latent semantic contains all of the linguistic context information in composition. This model uses a dual-mode factor analysis, based on singular value decomposition [3]. Singular value decomposition is based on the matrix having different entities in the ranks, such as feature item-document matrix, feature items and documents will be expressed as an alternative dimension of a vector space, dot product or cosine value of the vector indicates their similarity. This matrix will be broken down into three specific forms. For example, a $t*d$ feature item-document matrix X, can be broken down into three other matrices:

$$X = T_0 S_0 D_0^{\mathrm{T}} \tag{1}$$

This is called the singular value decomposition of X. T_0 and D_0 are called left singular vector and right singular vector matrix, S_0 contains the diagonal matrix of singular values. By singular value decomposition, LSA can effectively eliminate noise, reduce the impact by the ambiguity and synonymous of words, characterize the true meaning of text better.

Concept vector space model based on WordNet semantic dictionary: WordNet will ignore some of the smaller function words set in English. According to the meaning of the word, form a semantic network, including the synonymous relationship and upper-lower relationship. The model selects the synonym set (Synsets) as a feature item, expresses composition from the semantic aspects. In addition, there also need to use weight function, the product of the concept of frequency and general degree, to enhance the characterization capabilities of feature items.

2.2 Analysis of Essay Language Scoring

For essay language scoring, considering every point of text, such as vocabulary, grammar, sentence structure, select some characteristics variables; conduct a comprehensive analysis of these characteristic variables to measure the quality of the composition language, and score according to each feature item.

Vocabulary. According to the outline CET glossary, give a statistical analysis on new words and the number of the correct spelling of words appeared in the composition, the analysis result could become a feature that can measure the ability of using vocabulary.

Grammar. Select the number of sentences having no syntax errors in the composition as a feature, to measure candidates' mastery of English grammar.

Sentence Structure. Select the number of the different types of phrases in the composition, to measure the candidates' ability of grasping the sentence structure. Select the number of modal verbs to measure the candidates' ability of grasping advanced grammar such as subjunctive.

Apart from the above point of view, we can also choose the number of conjunction and the average sentence length appearing in the composition as a measure of continuity of the sentence. These variables are gathered together, can serve as a standard that can measure the quality of the composition language of candidates.

3 Design and Implementation of AES System

3.1 Generation of Feature Vectors of Essay Content

(1) Do a pretreatment of the composition, it contains isolating from words essay and skipping irrelevant words, such as "a", "the", "is", then turn essay text into the form that computer can handle.

(2) Reduce the dimension of essay: Depending on the feature extraction method (i.e., TFIDF, mutual information, information gain, chi-square test), construct different evaluation functions, conduct an independent assessment for each feature of the initial vector, to obtain an evaluation score. Then according to the size of assessment scores, all the features are sequenced and we select a specified number of feature subset.

(3) Generate feature vector of essay: After extracting feature words, give the weight to the training composition sets and the composition waiting for scoring, generate a feature vector for each composition. Among them, the feature vector has a weight for each feature, to measure the scoring contribution of the feature on the essay content.

a_{ii} is the weight of the *i-characteristic* of *j-chapter* essay, this weight function consists of two parts: ①the local weight function *L(i,j)* , is the weight of feature word *i* in document *j*. It selects the logarithm of document frequency of the feature word *i*:

$$L(i, j) = \log(tf(i, j) + 1) \tag{2}$$

tf(i,j) is the occurrences of the feature word *i* in the *j-chapter* essay.

②the global weight function *G(i)*, is the weight of feature word *i* in the whole essay collection and is the entropy that determined by the distribution of the feature items in all the composition collections:

$$G(i) = H(d \mid i) = -\sum_j p(i, j) \log p(i, j) \tag{3}$$

H(d|i) is the conditional entropy for a given feature word *i. p (i, j)* is the probability of the feature item *i* appearing in the *j-chapter* essay. *W(i,j)* represents the importance of the feature word *i* in the composition:

$$W(i,j) = L(i,j) / G(i) = \log(tf(i,j)+1) / -\sum_{j=1}^{n} p(i,j) \log p(i,j) \tag{4}$$

(4) The compositions are expressed as a vector of semantic space, using LSA and SVD (Singular Value Decomposition) model approach.

3.2 Essay Content Scoring Based on the Distance-Weighted KNN Algorithm

When a essay sample set is unbalanced, scoring error will be relatively large through using the KNN algorithm. Therefore, we select a distance-weighted KNN algorithm, this algorithm weight the contribution of the nearest k neighbors, and according to the relative distance of the query point, assign greater weight to closer neighbors.

Assume that all the compositions are expressed as the point of *n*-dimensional space ξ^n . The arbitrary composition x is expressed as the following feature vector: $< a_1(x), a_2(x), \cdots a_n(x) >$, $a_r(x)$ represents the weight of the *i-characteristic* in composition *x*, thus the distance between two instances x_i and x_j is defined as $d(x_i, x_j)$:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \tag{5}$$

In the KNN learning algorithm, objective function value can be a real value. Function argument is the feature vector of composition content, essay content scoring by teacher is the objective function value here, The return value $f'(x_q)$ is the estimate of $f(x_q)$, it is the average of the nearest k training samples away from *x*.

Training Algorithm:

```
For each training sample;
 (x for the composition of the feature vector, f (x)
     for writing scores);
 Add this sample to list training_examples;
```

Classification Algorithm:

```
Given a query instance to be classified x_q;
    (The feature vector that waiting for scoring);
Select the nearest k samples in the list
    training_examples away from x_q , and expressed by
x_1 K x_k
return f'(x_q) ← Σ_{i=1}^{k} w_i f(x_i) / Σ_{i=1}^{k} w_i   w_i = 1 / d(x_q, x_i)^2
```

When the query point $x_q = x_i$, thus $d(x_q, x_i)^2 = 0$.

In this case, assume $f^{'}(x_q) = f(x_i)$.

3.3 Select Variables for Essay Language Scoring

First of all, separate sentences, and do linguistic analysis. And do syntactic analysis and syntax analysis for each sentence, using open source software--Stanford Parser. First of all, we construct a parser, enter the essay text of all students, parse each sentence, to get a tree structure called grammatical structure. At last, extract language characteristic variables from the tree structure, and make a statistic, put data into KNN algorithm. And score on the essay language.

4 Experiment and Discussions

Based on the "Spoken and Written English Corpus of Chinese students" by Qiufang Wen, Lifei Wang and Maocheng Liang, selected a theme topic for the Western fast food composition. In this thesis, of which 200 were selected as training sample set, 70 as the test set. By doing experiments, do a comparative analysis the pros and cons of the various scoring models compared with the teacher scoring results, the results are given in the Fig.1.

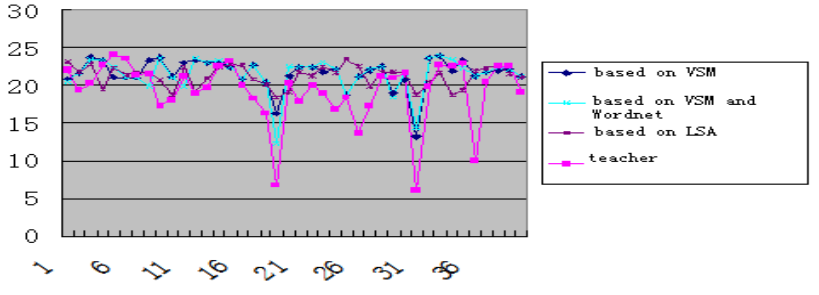

Fig. 1. The results using three models contrast with the teacher scoring

About the errors of the AES with feature selection and teacher scoring, the results are given in Table 1.

Table 1. Comparison of three models

Methods Model	Based on VSM	Based on VSM and WordNet	Based on LSA
Error	411.4	364.4	282.0

Under the condition of a more balanced sample set, we used LSA, selected 30 features using the method of chi-square test, and used the KNN algorithm, to select the minimum error score compared to the essay content waiting for scoring. On whether the feature selection, the error of the scoring results are given in Table 2.

Table 2. The results with/without feature selection

Feature selection	Yes	No
Error under the condition of the average sample set	276.1	343.6

5 Conclusions and Future Work

The results of experiments on English text showed that: in all feature selection methods, Chi-square test and the Information Gain obtained the best results. Term frequency inverse document frequency method has the same performance with the former two, the mutual information method has the worst performance [8-10].

And the result showed that the essay scoring based on LSA was closer to the teacher scoring about essay content, and improved the accuracy of the final score result with the feature selection.

Future work will concentrate more on improvements of KNN algorithm and natural language processing technology for further optimizing the system accuracy and efficiency. Also, we want to improve the deep excavation quality of the composition.

References

1. Dikli, S.: Automated Essay Scoring. Turkish Online Journal of Distance Education 7(1), 49–62 (2006)
2. Landauer, T.K., Laham, D., Foltz, P.W.: Automated Essay Scoring: A Cross Disciplinary Perspective. In: Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor, Mahwah, United States, pp. 87–112 (2003)
3. Singular value decomposition, http://en.wikipedia.org/wiki/Singular_value_decomposition
4. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D.: Miller,k.: Introduction to WordNet: An on-line Lexical Database. In: Five Papers on WordNet, CSL Report, Cognitive Science Laboratory. Princeton University, Princeton (1993)
5. Rogati, M., Yang, Y.: High-Performing Feature Selection for Text Classification. In: Proceedings of the 11th ACM Conference on Information and Knowledge Management, McLean, America, pp. 659–611 (2002)
6. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval 1(1), 76–88 (1999)
7. Wen, Q., Wang, L., Liang, M.: Spoken and Written English Corpus of Chinese students 2.0. Foreign Language Teaching and Research Press
8. Yang, Y., Pedersen, J.P.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of 14th International Conference on Machine Learning, Nashville, United States, pp. 412–420 (1997)
9. Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods. In: Proceeding of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval, Berkeley, United States, pp. 67–68 (1999)
10. Yang, Y.: An Evalutation of Statistical Approaches to Text Categorization. Information Retrieval 1(1), 76–88 (1999)