Automatic scoring method of English composition based on language depth perception

TangDan1, Sun Yu1*
1Yunnan Normal University, Kunming, Yunnan Province, 650500, China
*Corresponding author’s e-mail: sunyu_km@hotmail.com

Abstract. AutomatedEssaySystem (AES) refers to a system for evaluating and rating compositions by using techniques in the fields of statistics, natural language processing and linguistics. The selection of composition features is one of the key issues in the study of automatic composition scoring. This paper abstracts the features of language sense from the depth of language perception, which makes up for the shortcomings of the automatic scoring system of shallow features (such as vocabulary difficulty, grammar, etc.). The features of language sense are mainly the similarity of keywords extracted from the first and last paragraphs of a composition by rake algorithm, the similarity of texts calculated by cosine similarity to the context content of a composition and the similarity detection of composition language materials by automata. Finally, from the three aspects of correlation, error and overall accuracy, it analyzes whether the deep-seated language features improved the scoring ability of AES system. The experimental results show that the added depth perception features enhance the language ability of automatic scoring and are closer to manual scoring.

1. Introduction
Writing ability is an important embodiment of students' English language ability, and also an effective way to evaluate students' English learning[1,2]. However at present, the automatic scoring method of English composition in China mainly analyzes the composition in shallow aspects such as word selection, grammar and content[3]. AES means that a computer can rate compositions on its own. In recent decades, with the rapid development of computer technology, natural language processing technology has been well developed, so scholars at home and abroad began to study the use of computers instead of labor to evaluate compositions[4]. AES can not only improve the efficiency of scoring, but also eliminate the inconsistency of composition evaluation, which is of great significance to control the scoring error.

China's research on computer-based composition automatic scoring started late. In recent years, China has developed an automatic English scoring system for colleges and universities. For example, the online writing automatic scoring represented by pigai focuses on vocabulary and grammar, but its advantages in text structure, content logic and coherence are not obvious. At present, most of the automatic scoring system is to evaluate the basic features of composition, such as grammar, lexical difficulty, syntax, etc. It is impossible to evaluate the characteristics of the composition, such as the semantic meaning, the relevance between the content and the topic, and whether the composition meets the requirements of the task[5,6]. Therefore, this paper thinks about the deep-seated language characteristics, thinks that the excellent English commentary has the composition structure that echoes the beginning and end, and will use the material accumulated daily. Therefore, the extraction of these composition features is very important for automatic computer scoring.
2. Context-related recognition

By analyzing the relevant materials of English composition and the teaching process, it is easy to conclude that good English composition usually has the composition structure that echoes at the beginning and end of the paragraph[7,8]. Nowadays, most of the compositions in English test are argumentative papers. The basic structure of argumentative papers is composed of three parts: "introduction, theory and conclusion". It is required that the first paragraph of a composition must put forward a thesis or argument, the main part of a composition should be demonstrated in layers, and the end of a composition should be summarized. There are four standards to establish a good argumentative framework: three arguments, the hierarchy, the natural connection, and the first and last echoes.

In this paper, we test the context similarity of the structure of the first and last echo of argumentation. The composition should go through word parting, wordlabeling and other pre-processing before the corresponding text to be selected for identification.

The most basic feature of the head and tail echo is the consistency of the theme, and the most intuitive expression of the consistency is the common keywords. The first paragraph will put forward arguments, and the last paragraph will summarize them. These two parts are inseparable from the core keywords of the theme. To construct a common key word bag, we can judge the relevance of the first and last two segments.

Bag of Words is a method of representing text data when modeling text by using machine learning algorithms. When text data is to be represented as a feature vector, the more common text feature representation is bag of Words. The word bag model is to put words into a bag without considering its lexical and word order (ie each word is independent)

RAKE is short for Rapid Automatic Keyword Extraction algorithm and is a domain-independent keyword extraction algorithm. It attempts to determine the key phrases in the body of the text by analyzing the frequency of occurrence of the text and its co-occurrence with other words in the text.

The keywords extracted by RAKE are not just a single word, but may be a phrase. The score of a word is related to the frequency of words, and the score composition of each phrase is accumulated by the words that make up the phrase. The fast automatic keyword extraction algorithm of nltk can be realized by python.

In natural language processing, such as dialogue system, information retrieval, machine translation, etc., it often involves how to measure the similarity between texts. In the recognition of context association, the similarity detection of the first and last two paragraphs confirms the information that has been skipped by the bag of words.

Text similarity includes the following three methods: one is the traditional method based on keyword matching, such as ma n-gram similarity; the other is mapping text to vector space, and then using cosine similarity; the third is deep learning method, such as convnet based on convolutional neural network [9].

There are many methods to calculate text similarity, such as cosine similarity, Euclidean distance, Jaccard distance, editing distance and so on[10,11]. For short texts such as English composition, cosine similarity method is more simple and effective. In this paper, the cosine similarity calculation method based on VSM is adopted [12].

3. Recognition of composition material

In addition to the content of the article, an excellent English essay should also be proficient in citing all kinds of knowledge to reflect the ability to accumulate language. The quality of the text is a comprehensive expression of what the author has learned. It has a certain degree of requirements for the application of the language materials of the students, such as celebrity quotes and beautiful sentence patterns. In this paper, the language material is called the personal language material in the concept of language sense [13]. In English papers, thesis and arguments have a large number of beautiful sentence templates.
Automatic recognition of beautiful sentences can be done with string matching but inefficient. The traditional string matching system uses character-by-character matching between pattern strings and matching strings. The efficiency of traditional string matching is very low in the task with a lot of material retrieval. This article uses AC automaton mode matching. Automata is an idealized "machine". It is only a theoretical tool for abstract analysis of problems, and does not have actual physical form. The AC automaton is a classic multi-mode string matching algorithm that can perform a scan of the main string to match the functions of multiple pattern strings. The implementation of AC automaton is to add the next array similar to KMP on the basis of Trie tree.

The construction of the AC automaton is divided into two steps: 1 constructing the Trie tree, 2. constructing the fail pointer based on the Trie tree (equivalent to the next array of KMP). Here are the words "she", "he", "say", "shr", "her" as examples, and the resulting AC automaton is shown in Fig. 1.

![Figure 1.Structure diagram of AC automaton](image)

The matching process of text string of AC automata can be divided into two cases:

1) If the current character matches, there is a path from the current node to the target character along the edge of the tree. If the current matching character is the end of a word, you can traverse to the root along the fail pointer of the current character. If there is a tag at the end of these nodes (represented by the tag here, the node is a tag at the end of a word), these nodes are all nodes that can match on. After the statistics are completed, those nodes are marked. At this time, you only need to continue matching along the path to the next node, and to move the target string pointer to the next character to continue matching.

2) If the current character does not match, the character pointed by the failure pointer to the current node will continue to match, and the matching process will end with the pointer pointing to root. Repeat either of the two processes until the pattern string reaches the end.

4. Evaluation method

Correlation degree indicates whether the two kinds of phenomena are related, the direction of correlation and the degree of correlation [15]. In this paper, Pearson correlation coefficient is used as the correlation of composition features. The correlation coefficient is used to measure the correlation (linear correlation) between two variables $X$ and $Y$, with values between -1 and 1. The formula is as shown in equation (1).

$$r(X, Y) = \frac{C(x, y)}{\sqrt{v[x] v[y]}}$$ (1)

In formula (1), $X$ and $Y$ represent the predicted score and the actual score of manual marking. In addition, $C(x, y)$ is the covariance of $X$ and $Y$, $v[x]$ is the variance of $X$, and $v[y]$ is the variance of $Y$. The closer the value of $r$ is to 0, the lower the correlation is. On the contrary, the closer the absolute value of $r$ is to 1, the stronger the correlation is. Through the correlation coefficient, we can get the correlation between the predicted score and the actual score.
The error of prediction score and manual evaluation score can be expressed by mean square deviation, as shown in formula (2).

\[ M = \frac{1}{n} \sum_{i=1}^{n} (S_i - P_i)^2 \]  

(2)

Mean square deviation is the average of the sum of the squares of the distances from the real value of each data, which is used to express the deviation degree between the prediction score and the manual evaluation score. In formula (2), n is the size of the test set, and s is the manual evaluation score, and p is the prediction score.

The overall accuracy can be expressed as (3).

\[ A = \frac{SC}{n} \times 100\% \]  

(3)

Where n is the size of the test set, SC is the sum of the number of differences between the predicted score and the manually measured score within the threshold.

5. Result analysis

This paper extracts a series of features based on depth perception of language, adds the features extracted in Table 1 to the basic features, and then analyzes the impact of these features on the automatic composition scoring.

| Features | Meaning |
|----------|---------|
| BC       | The same number of words in the first and last paragraphs of a composition |
| BS       | Similarity between the first and last paragraphs of a composition |
| CC       | Number of materials in composition |
| AF       | All features added |

Generally, feature extraction is established in the process of repeated experiments and mistakes. We divide the collected English argumentative papers into development sets and test sets according to the ratio of 3:1. Both development set and test set are corpus data. The development set is divided into training set and development test set. Training set is used for training model, and development test set is used for error analysis (error analysis is an effective method to improve feature extraction), and test set is used for final evaluation. The test results are shown in Table 2.

| Method | Relevance | Error | Overall accuracy |
|--------|-----------|-------|------------------|
| AES    | 0.25      | 5.23  | 73.2%            |
| BC     | 0.41      | 5.12  | 73.4%            |
| BS     | 0.32      | 5.13  | 73.2%            |
| CC     | 0.43      | 4.86  | 71.8%            |
| AF     | 0.52      | 4.63  | 76.1%            |

The results show that the performance of AES is improved by adding the features of language sense, and the most significant one is that the correlation is improved by 0.21 after adding the feature of the number of personal composition materials. The overall accuracy of the method is improved to 76.1%.

6. Conclusions

The proposed feature extraction method based on deep perception has a positive impact on AES. But at the same time, there are also some shortcomings, such as the scoring index may not be ideal,
resulting in a large difference between the prediction score and the manual evaluation score, thus reducing the overall accuracy. In addition, this paper is based on the composition features of English argumentative paper extraction. If it is extended to other types of automatic scoring, it may not get satisfactory results. Moreover, the number of text data trained in this paper is not large, which restricts the objectivity of automatic scoring to a certain extent.

References

[1] Wen Qiufang. Research on construct validity of composition content: an attempt to use structural equation modeling software amos5 [J]. Foreign language research, 2007, (3): 67.

[2] Ma Xinfang. An empirical analysis of the relationship between second language reading ability and writing ability [D]. Beijing: Beijing Language University, 2004.

[3] Ke Dengfeng, Peng Xingyuan. Automatic scoring of Chinese composition based on vocabulary scoring [C] / / Sun Maosong, Chen qunxiu. Research progress in Computational Linguistics in China: 11th National Conference on computational linguistics, Luoyang, 2011. Beijing: Tsinghua University Press, 2011: 588-595.

[4] Li Ting, Zhang Jingxiang. English composition scoring algorithm with centralized trend adaptive enhancement [J]. Computer engineering and application, 2018,54 (09): 151-155 + 256.

[5] Ge Shili, Chen Xiaoxiao. Problems and Countermeasures in the study of automatic scoring of College English composition [J]. Shandong foreign language teaching, 2009, (3): 21-26.

[6] Pu Jianzhong. A survey of interlanguage among Chinese Non-English Majors Based on Learner Corpus [a]. Edited by Yang Huizhong, GUI Shichun, Yang da. An analysis of Chinese learners’ English based on CLEC corpus [C]. Shanghai: Shanghai Foreign Language Education Press, 2005: 135-139.

[7] Fu Ruiji, Wang Dong, Wang Shijin, Hu Guoping, Liu Ting. Beautiful sentence recognition for automatic composition scoring [J]. Chinese Journal of information, 2018,32 (06): 88-97.

[8] Liu Mingyang, Qin Bing, Liu Ting. Automatic scoring of college entrance examination composition based on literary features [J]. Intelligent computer and application, 2016, 6 (01): 1-4 + 8.

[9] Behnke S. Hierarchical neural networks for image interpretation[J]. Lecture Notes in Computer Science, 2003, 2766(3): 1345-1346.

[10] Dong Yangyi, Li Weihua, Yu Hui. Hierarchical relation mining of Chinese text based on mixed cosine similarity [J]. Computer application research, 2017,34 (5): 1406-1409.

[11] Jiang Hua, Han Anqi, Wang Meijia, Wang Zheng, Wu Yunling. A string similarity algorithm based on improved edit distance [J]. Computer Engineering, 2014, 40 (1): 222-227.

[12] Joachims T. Text categorization with support vector machines: Learning with many relevant features[M]. Machine Learning: ECML-98. Springer Berlin Heidelberg, 1998:137-142.

[13] Zhong Qidong, Zhang Jingxiang. Chinese composition scoring algorithm based on embedded language depth perception [J / OL]. Computer engineering and application: 1-7[2019-11-07].Http://kns.cnki.net/kcms/detail/11.2127.tp.20190530.1622.032.html.

[14] Hu Lijuan. Research on large-scale multi pattern matching algorithm [D]. Xi’an University of Electronic Science and technology, 2017.

[15] Huang zhi’e, Xie Jia Li, Xun en-dong. study on feature selection of HSK automatic composition scoring [J]. Computer engineering and application, 2014,50 (06): 118-122 + 126.