Automatic Classification with SVM and F-VSM on Elementary Chinese Composition

Weiping Liu, Calvin C. Y. Liao, Wan-Chen Chang, Hercy N. H. Cheng, and Sannyuya Liu

***Abstract*—Currently, automated evaluation of Chinese composition still has limitations. Moreover, the human evaluation is possible subjective, time-consuming and laborious. Hence, to develop automatic evaluation of Chinese composition is very meaningful and potential. In this study, we adopted two methods: support vector machine (SVM) and feature vector space model (F-VSM) to evaluate 4193 Chinese compositions collected from 1st to 6th grade at an elementary school in Wuhan. This study integrated natural language processing techniques to extract features, and uses SVM and F-VSM to classify the composition level. We investigated 45 linguistic features and divided into four aspects: text structure, syntactic complexity, word complexity and lexical diversity. The result indicated that both SVM and F-VSM have good classification effect, and F-VSM effect is better than SVM.**

***Index Terms*—F-VSM, linguistic features, natural language processing, SVM.**

I. Introduction

In recent years, with the rapid development of computer science, the continuous progress of natural language technology, and machine learning progress, automatic composition assessment has become an inevitable trend of development [1]. It has been found that there have been early studies on automatic assessment of English compositions abroad. Representative studies include PEG (Project Essay Grade), IEA (Intelligent Essay Assessor) and E-rater. PEG was developed in 1966 by Ellis Page of the University of Duke, which was one of the earliest automated composition assessment systems [2]. PEG is mainly from the linguistic surface features of multiple regression analysis. IEA is an automatic scoring system based on latent semantic analysis, developed by Thomas Landauer of the University of Colorado [3]. IEA constructs the semantic space of the composition mainly through the latent semantic analysis model, and evaluates the similarity of the composition with the artificial scores. E-rater was developed by the US Educational Testing Service in the 1990s, with the aim of assessing the quality of writing in the GMAT exam [4]. E-rater uses the methods of statistics, vector space model and natural language processing technology to evaluate the quality of writing from the aspects of language, content and text structure. In addition, in recent years, there are some researches on neural network automatic scoring, comparison of automatic scoring and manual scoring, and automatic scoring with some tools (eg, Coh-Metrix and WAT) have been published [5]-[7]. However, there is a lack of research on automatic assessment of Chinese compositions in China. In 2006, Yanan Li studied the Chinese automatic scoring as a second language test [8], but the subject is not Chinese. Yiwei Cao and Chen Yang used latent semantic analysis techniques to study the automatic scoring of Chinese compositions in 2007 [9]. Zhie Huang studied the feature selection of automatic composition evaluation in 2014 [10]. It selected 19 features of high correlation with the quality of composition from the aspects of words, grammar, segmentation and literary expression. However, it is not enough to evaluate the composition only by using latent semantic analysis technology. More features should be considered and other methods can be used to improve the effect of automatic evaluation.

To summarize, automatic composition assessment is a difficult task, if you want to achieve high reliability, features of composition quality needs to consider many aspects, according to statistics, natural language processing, machine learning and other analysis methods, so that can be more comprehensive assessment of the quality of composition. In addition, due to the language differences between Chinese and English, so it is different in the selection of the quality features of the composition. But, automated composition assessment has the following advantages: first, compared to the manual evaluation objective, evaluation results are not affected by human factors; second, high efficiency, fast and timely scoring machine; third, low cost, evaluation of machine can save a lot of manpower. In a word, the study of automatic composition evaluation is of great significance. Therefore, this study will research the automatic evaluation of Chinese composition of primary school from the linguistic features and other aspects features, combining natural language processing, support vector machine and feature vector space model.

II. Feature Selection

This study evaluates the level of the Chinese composition from the four aspects of text structure, syntactic complexity, word complexity and lexical diversity, and uses natural

Manuscript received June 20, 2017; revised October 25, 2017. The National Social Science Fund Project of China (grant number: 14BGL131) and National Engineering Research Center for E-learning, Central China Normal University for financial support (grant numbers: CCNU16A02022, CCNU15A06073).

Weiping Liu, Calvin C. Y. Liao, Hercy N. H. Cheng, and Sannyuya Liu are with the National Engineering Research Center for e-Learning, Central China Normal University, Wuhan, China (e-mail: 1398518181@qq.com, CalvnCYLiao@gmail.com, HercyCheng.tw@gmail.com, lsy5918@gmail.com).

Wan-Chen Chang is with the Graduate Institute of Learning and Instruction, National Central University, Taoyuan, Taipei (e-mail: altheawcc@gmail.com).

doi: 10.18178/ijiet.2018.8.5.1057

language processing technology to extract 45 features. Then, some features will be described in detail from the four aspects.

A. Text Structure

The text structure mainly includes 5 features such as the length of the composition, the number of punctuation, the total sentence, the number of short sentence and the average sentence length. The length of the composition refers to the number of punctuation and words included in the composition. The number of total sentences in the composition is a comma, semicolon, colon, dash, period, question mark, exclamation mark, and ending of the ellipsis; the average sentence length is the number of sentences in the composition refers to the average number of words in each sentence, that is, the total number of words divided by the number of short sentence.

B. Syntactic Complexity

The syntactic complexity has 14 features, which mainly includes subject-verb, verb-object, indirect-object, fronting-object, double, attribute, adverbial, complement, coordinate, preposition-object, left adjunct, right adjunct, independent structure, head. For example, "I have a very beautiful mother, and she loves me very much." ‘I->have’ is subject-verb, ‘loves->me’ is verb-object, ‘a->mother’ is attribute, ‘very ->beautiful’ is adverbial, and so on. These are common grammatical relations between sentences.

C. Word Complexity

The word complexity mainly includes the total number of words, the total number of different words, the number of high frequency words, the number of intermediate frequency words, low frequency words, very low frequency words, the total number of terms, the total number of different terms, high frequency terms, intermediate frequency terms, low frequency terms, very low frequency terms of 12 features. The frequency of words and terms was divided according to "National Language Committee of modern Chinese corpus word frequency statistics table". The high frequency word corpus of Chinese words in the cumulative frequency reached 90% before; the intermediate frequency word refers to the Chinese corpus of words in a cumulative frequency reached between 90% to 95%; low frequency word refers to Chinese corpus of words between the cumulative frequency reached 95% to 97%; extremely low frequency word refers to the Chinese corpus of words in a cumulative frequency more than 97%. The high frequency terms refers to the emergence of Chinese corpus of terms in the cumulative frequency reached 80% before; the intermediate frequency term refers to the cumulative frequency of terms that appear in modern Chinese corpora up to 80%-85%; the low-frequency term refers to the cumulative frequency of terms appearing in the modern Chinese corpus up to 85%-89%; the extremely low frequency term refers to the cumulative frequency of terms in modern Chinese corpus up to 89%-91%.

D. Lexical Diversity

The lexical diversity mainly includes 14 features such as the number of content words, the number of function words and the number of logical words. Content words refer to words with practical meanings, including: nouns, verbs, adjectives, numerals, quantifiers, pronouns. A word that has no full meaning, but has grammatical or functional meaning is called function word, it mainly consists of adverbs, prepositions, conjunctions, auxiliary words and interjection. The number of logical words refers to the number of logical related words used in composition.

III. Methods

A. Data Normalization

After the feature is extracted, data normalization is needed, so that the influence of some feature size differences can be avoided. There are many data normalization methods, and the appropriate normalization method is chosen according to the actual requirements. This study uses the “Min-Max Normalization” method, after normalization, all features will be distributed between 0 and 1 [11]. The theory of the “Min-Max Normalization” method is shown in (1). *Min* represents the maximum in a class of samples, *max* represents the minimum in a class of samples, and x^* represents the normalized value of the sample.

$$x^* = \frac{x - \min}{\max - \min} \tag{1}$$

B. Support Vector Machine

Support vector machine (SVM), proposed by Vapnik in 1996, it is a classification algorithm based on statistical learning theory [12]. SVM algorithm is based on the theory of VC theory and structural risk minimization of statistical learning theory [13]. Based on the complexity of the model (i.e., the learning accuracy of the specific training sample) and the learning ability (i.e., no error Identify the ability of any sample) to find the best compromise in order to achieve better generalization capabilities. The basic idea of SVM is to find the optimal hyperplane between class and class, and to separate the two at the same time to satisfy the classification interval. The objective functions and constraints for proper classification are shown in (2) and (3):

$$\min\left(\frac{1}{2}\|w\|\right) \tag{2}$$

$$y_i\left(w \bullet x_i + b\right) \geq 1, \quad i = 1, 2, \ldots l \tag{3}$$

SVM has many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition problems, and has overcome the problems of “dimension disaster” and “overfitting” to a large extent. Studies have shown that SVM for text classification has a very good effect [14]. Moreover, the text feature belongs to the nonlinear structure, so this study chooses SVM algorithm to do the composition classification. In this study, the SVM classification process also uses the latent semantic analysis to reduce the dimension and vector processing, the classification process shown in Fig. 1.

C. Feature Vector Space Model

The traditional vector space model constructs a vector space based on a specific corpus, and maps the text into a

vector, each article is represented by an n-dimensional vector [15]. Based on the vector space model, this study proposes a feature vector space model (F-VSM) based on linguistic and multidimensional features of the composition.

$$t = \{ f_1, f_2, \cdots f_n \} \tag{4}$$

$$g_i = \{ f_{i1}, f_{i2}, \cdots f_{in} \}, \quad i = 1, 2, \ldots 6 \tag{5}$$

$$sim_i = \frac{g_i \bullet t}{\| g_i \| \times \| t \|} = \frac{\sum_{i=1}^{6} (g_i \times t)}{\sqrt{\sum_{j=1}^{n} g_{ij}^{\ 2} \times \sum_{j=1}^{n} t_j^{\ 2}}} \tag{6}$$

The main theory is to extract 45 features from each composition, and map the 45 features into a 45 dimensional vector, all of which form a feature vector space. F-VSM algorithm implementation process is shown in Fig. 2.

Equation (4) represents the normalized test set feature vectors, f_i represents the *i-th* feature. In (5), g_i represents the standard feature vectors of grade *i*, and the feature vectors of all the anthologies of the essay in the *i* grade is normalized by the data, representing the average of the essay under the grade *i*. Equation (6) is the cosine similarity of the feature vectors of the test set and the *i* grade.

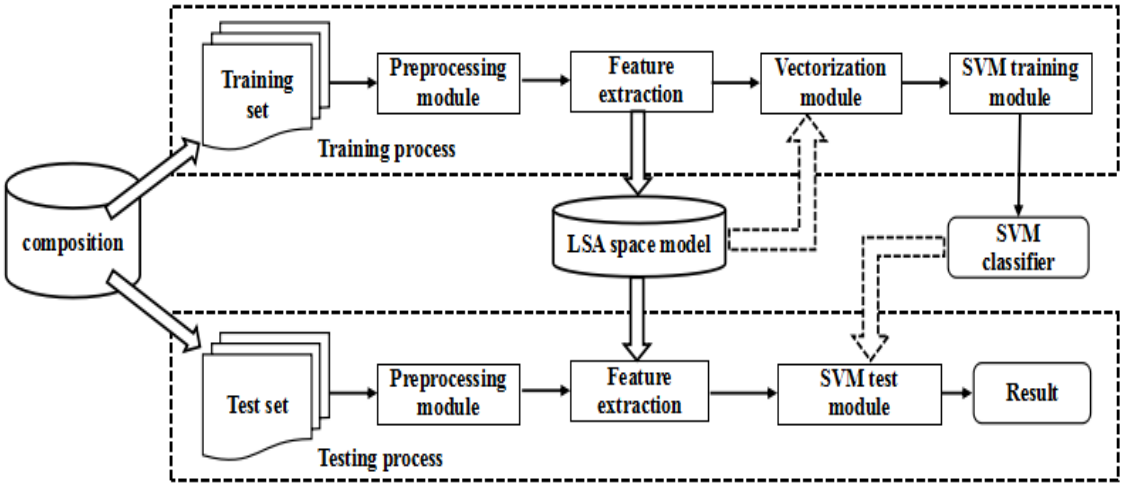


Fig. 1. SVM algorithm based on the classification process.

F-VSM algorithm process:

Input: Test set primitive feature vector *w*

Output: Classification of *grade*

Step1: Original feature vector: $\mathbf{w} \xrightarrow{\text{data normalization}} t = \{ f_1, f_2, \cdots, f_n \}$

Step2: Calculate the standard feature vector for each grade: $g_i = \{ f_{i1}, f_{i2}, \cdots, f_{in} \}$

Step3: Calculate the cosine similarity between the test set and the standard feature vector for each grade: $\mathbf{sim}_i$

Step4: According to the maximum similarity of the value of the classification results: $\mathbf{max}\ (sim_i) \rightarrow grade$

Fig. 2. F-VSM algorithm execution process.

IV. Experiment and Analysis

A. Data Collection and Processing

In the part of the experimental data, 4193 Chinese compositions from 1st to 6th grade at a primary school in Wuhan were collected, of which 333 in the first grade, 999 in the second grade, 951 in the third grade, 763 in the fourth grade, 539 in the fifth grade, 608 in the sixth grade. The experiment is programmed in Java language, and all the compositions are imported into the database and stored in grades. 45 features of each composition were extracted, and then use the "Min-Max Normalization" method to normalize the data.

B. Experiment Process and Result

The experiment process includes two parts. The first part is to classify the composition by SVM method, the second part is to classify the composition by F-VSM method. First, experiment was performed using the SVM method, each grade of the data set randomly divided into test set and training set of two parts, the ratio of 1: 3. In other words, 3/4

as a model of training data, the remaining 1/4 data (a total of 1050 compositions) as test set for testing the effect of training model. Then the data is normalized and SVM classification is performed. Finally, the results of the classification accuracy of the SVM method for each grade are shown in Table I.

TABLE I: SVM CLASSIFICATION RESULTS

Grade	Correct	Error	Total	Accuracy
grade1	41	43	84	48.81%
grade2	194	56	250	77.60%
grade3	178	60	238	74.79%
grade4	167	24	191	87.43%
grade5	91	44	135	67.41%
grade6	98	54	152	64.47%
Total	769	281	1050	73.24%

For the F-VSM method, there is no need for a model training process, so 4193 compositions are test sets. First, the average of each grade characteristic vector is obtained, and the standard feature vector of each grade is used as the test set. And then calculate the cosine similarity of each test data with the six grade standard feature vectors. Select the maximum similarity of the corresponding grade as the test classification results, and with the annotation of the grade comparison, consistent with the correct category, inconsistent is classified as the wrong category. The final number of correct classes divided by the total number is the accuracy of the classification. Finally, the F-VSM method results in the classification accuracy of each grade are shown in Table II.

TABLE II: F-VSM CLASSIFICATION RESULTS

Grade	Correct	Error	Total	Accuracy
grade1	233	100	333	69.97%
grade2	810	189	999	81.08%
grade3	638	313	951	67.09%
grade4	669	94	763	87.68%
grade5	389	150	539	72.17%
grade6	525	83	608	86.35%
Total	3264	929	4193	77.84%

From Table I and Table II, draw the trend of the classification accuracy of SVM and F-VSM in each grade section, as shown in Fig. 3, and the overall classification accuracy of SVM and F-VSM on the composition classification, as shown in Fig. 4. The ordinate of the graph indicates the classification accuracy, G1 to G6 represents six grades. It can be seen from Fig.3, the classification effect of SVM algorithm in grade 1 to 6 is more volatile, and the classification effect is poor for grade one, only 48.81%, but the classification effect of grade 2, 3, 4 is better, and all of them reach about 75%. F-VSM algorithm for grade 1 to 6 classification effect is good, and the volatility is relatively stable. It can be seen from Fig. 4, SVM algorithm and F-VSM algorithm on the composition of the classification have a good effect as a whole, and F-VSM better than SVM. In addition, Fig. 3 also shows that, F-VSM is better than SVM in classification, except for grade three. In summary, both SVM and FVSM have good effect on composition classification, but F-VSM is better than SVM.

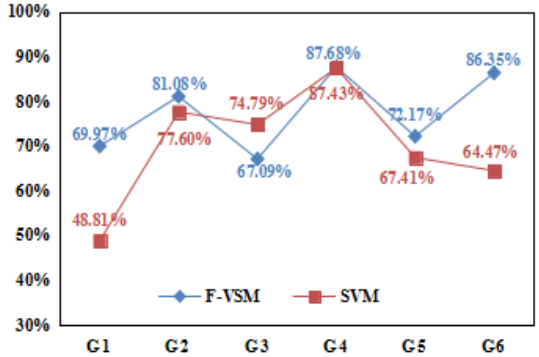


Fig. 3. SVM and F-VSM classification accuracy in each grade.

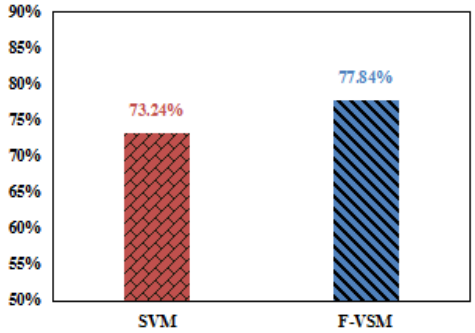


Fig. 4. SVM and F-VSM overall classification accuracy.

V. SUMMARY AND PROSPECT

This study integrates natural language processing techniques to extract features, and adopts SVM and F-VSM to classify the composition level. We investigated 45 linguistic features and divided into four aspects: text structure, syntactic complexity, word complexity and lexical diversity. Two methods were used to compare and analyze. The main contributions are as follows: 1) The feature extraction of the composition is considered from multiple aspects, which avoids the problem of low credibility due to the fact that the composition feature is too single. 2) The F-VSM method is proposed and compared with the SVM method in the composition classification effect. The results showed that the F-VSM method had better classification effect on the composition. 3) It provides some reference value for future research on automatic Chinese composition evaluation. Next, we will consider the composition of the automatic score and according to the content of the composition to give the corresponding automatic feedback.

ACKNOWLEDGMENTS

This study was supported by Technical Supporting Programs Funded by National Key Technologies R&D Program of China (NO.2015BAH33F02) and Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No.CCNU16A01022).

REFERENCES

[1] D. Semire, "An overview of automated scoring of essays," *Journal of Technology Learning & Assessment*, vol. 5, no. 1, p. 36, 2006.

[2] D. Arthur, "Computer grading of English composition," *Education Digest*, vol.55, no. 1, pp. 46-52, 1966.

[3] T. K. Landauer, "Automatic essay assessment," *Assessment in Education Principles Policy & Practice*, vol. 10, no. 3, pp. 295-308, 2003.

[4] Y. Attali and B. Jill, "Automated essay scoring with e-rater®, V.2.0," vol. 4, no. 2, p. i-21, 2006.

[5] A. Dimitrios, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," *Association for Computational Linguistics*, 2016.

[6] J.-H. Wang and M. Stallone, "Automated essay scoring versus human scoring: A comparative study," *Journal of Technology Learning & Assessment*, vol. 6, no. 2, p. 29, 2007.

[7] M. Danielle *et al*, "A hierarchical classification approach to automated essay scoring," *Assessing Writing*, vol. 23, pp. 35-59, 2015.

[8] Y. Li, "Study on automatic score of Chinese as a second language test," *Dissertations of Beijing Language and Culture University*, 2006.

[9] Y.-W. Cao and C. Yang, "The use of latent semantic analysis automated Chinese essay scoring," *Examination Research*, pp. 65-73, 2007.

[10] Z. Huang, J. Xie, and E. Xun, "Research on feature selection in HSK automated essay scoring," *Computer Engineering and Applications*, vol. 50, no. 6, pp. 118-122, 2014.

[11] G. Wang *et al.*, "Using min-max normalization to measure the differences of regional economic growth-A case study of Yulin Area, Shanxi Province," *Economy & Management*, 2016.

[12] Vapnik and N. Vladimir, "The nature of statistical learning theory," *Technometrics*, vol. 8, no. 4, p. 1564, 1996.

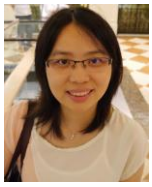
[13] H. David, M. Kearns, and R. Schapire, "Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension," *Machine Learning*, vol. 14, no. 1, pp. 83-113, 1994.

[14] W. Krzysztof *et al*., "Comparison of SVM and ontology-based text classification methods," in *Proc. International Conference on Artificial Intelligence and Soft Computing Springer, Cham*, pp. 667-680, 2016.

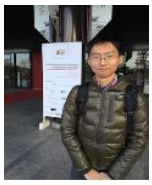
[15] C.-Y. Huang *et al.*, "An adaptation of the vector-space model for ontology-based information retrieval," *IEEE Transactions on Knowledge & Data Engineering*, vol. 19, no. 2, pp. 261-272, 2007.

Weiping Liu was born on August 14, 1993, Yichun, Jiangxi province, China. He is a master in computer application technology and studying at National Engineering Research Center for E-learning (NERCEL), Central China Normal University (CCNU). At present, his research focuses on text analysis, including the difficulty of Chinese reading text analysis and the quality of Chinese composition analysis. In addition, he is also interested in the study of NLP.

Calvin C. Y. Liao is currently an associate professor and researcher in National Engineering Research Center for e-learning (NERCEL) at Central China Normal University, China. He received his Ph.D. degree in the Institute of Network Learning Technology at National Central University in 2011. Since 2011, he was an adjunct assistant professor and a postdoctoral scholar in the Institute of Network Learning Technology at National Central University, Taiwan. His research focuses on designing Technology Enhanced Language Learning (TELL) for primary schools.



Wan-Chen Chang is assistant Professor of the Department of Human Development and Family Studies at National Taiwan Normal University in Taiwan. She received her Ph.D. in the Graduate Institute of Learning and Instruction at National Central University in 2014. Since 2014, she was an adjunct assistant professor and a postdoctoral scholar in the Institute of Network Learning Technology and the Graduate Institute of Learning and Instruction at National Central University, Taiwan. Her research focuses on emergency literacy, instructional strategies for reading comprehension, and Technology Enhanced Language Learning (TELL).



Hercy N. H. Cheng is currently an associate professor and researcher in National Engineering Research Center for e-learning (NERCEL) at Central China Normal University, China. He received his master degree in the Department of Computer Science and Information Engineering at National Central University in 2003 and the Ph.D. degree in 2009. Since 2010, he was an adjunct assistant professor and a postdoctoral scholar in the Institute of Network Learning Technology at National Central University, Taiwan. His research focuses on designing computer-based mathematical and language learning for primary schools. His current research interests include one-to-one learning environments and game-based learning, in particular, challenge design in a one-to-one classroom.

Sannyuya Liu received the B.E. and M.E. degrees in 1996 and 1999, and received the Ph.D. degree in 2003 from HUST. He devoted himself to his postdoctoral research in Xiamen University from 2003 to 2005, and worked for the field of enterprise information, business intelligence, and distributed computing. Currently. He is a professor in NERCEL, CCNU. His research interests include artificial intelligence, computer application, and educational data mining. He published at home and abroad SCI, SSCI, EI research papers more than 40 articles, edited and published 5 monographs, approved 6 national invention patents, apply for more than 30, approved software copyright more than 50. He was awarded the two prize of 1 national teaching achievement award of higher education, the first prize of teaching achievement award in Hubei higher education institutions, 2 first prize of scientific and technological progress in Hubei Province, and two 1 prize 1.