Automated Computer-based CET4 Essay Scoring System

Yi Xi
Department of Electronic Engineering
Tsinghua University
Beijing 100084, China
xiy08mail@163.com

Weiqian Liang
Department of Electronic Engineering
Tsinghua University
Beijing 100084, China

***Abstract* –Automated Essay Scoring is a very significant research subject for the processing of machine scoring. Computer-based College English Test (National English level test) further motivates the research of AES system for Chinese English leaner. In this paper, we introduce an effective AES system based on computer-based CET4. Features belong to three domains: language quality, content and organization, are involved in our system to figure out the feature collection with high correlation coefficient. The method of improved TF*IWF*IWF in computing term weight is used to optimize content-features. Through comparison of five scoring models based on classical text classification algorithms, we proposed an algorithm based on Adaboost, Voting and KNN to improve the accuracy of machine scoring. The exact agreement attends to 60%, and the adjacent agreement is above 95%. Moreover, we also discovered that writing patterns for Chinese English learners emphasize on fluency, lexical complexity features and two features from organization.**

Key words- AES, Machine Learning, Vector Space Model

I. INTRODUCTION

Among all the English testing items, evaluating writing quality is the most challenging section. Because of large quantity of examinees, English teachers are required to score large quantity of essays in limited time, which lead to two defections of human scoring. Firstly, due to its subjectivity, teachers' personal education background, understanding for rating criteria and physical or mental status would affect the reliability of scoring results. Secondly, human resources are much expensive and time-consuming for essay assessment. The purpose of this research is designing an automated essay scoring (AES) system to facilitate the laborious and costly manual assessment in English tests.

AES is the ability of computer technology to evaluate and score written prose. Studies show that the system simulates a human rater's grading process, and performs the grading with a higher agreement with human rater.

The research of AES has a long history[1]. Since 1966, several AES systems have been already produced and even used on large-scale essay exams successfully. For example, Project Essay Grader (PEG)[1] is the first AES system, which applies regression technique to predict scores based on surface textual features. Furthermore, more advanced techniques in Natural Language Processing (NPL), Information Retrieval (IR) and Artificial Intelligent (AI) are used in AES systems [2]. Intelligent Essay Assessor (IEA)[3] analyzes and scores an essay using Latent Semantic Analysis[4]. E-rater[2] is currently used by ETS for scoring essays in GRE and TOEFL, employing statistical approaches, vector space model (VSM), and NLP. With these techniques, E-rater analyzes not just essay language, but also measures essay content and organization as well.

Even though several English AES systems have already been proposed and quite successfully developed, AES is still a difficulty, intricate and interesting area. Most of the current English AES systems[1] are based on the syntax and sentence structure to grades the essays. Moreover, there are large quantities of studies[5] that support to analyze semantic characteristics and include more features to score essays. From a theoretical point of view, the current AES systems suffered a serious restriction in native English speakers. Chinese English Learners' written English has been proved to be subject to the effect of native language and culture, and therefore has its own patterns that the current AES systems may fail to identify.

In this article, we propose an AES system to discover writing patterns for Chinese English learners for CET4. These special patterns are generated to promote the understanding of the interrelationship between different constructs of CET4 writing quality. Then we use a voting algorithm based on the initial scores and pattern features to interactively train the system to score the essays.

This paper is organized as follows: firstly, we introduced the system design and the essay quality features; secondly, several features effectiveness evaluation algorithms are implemented to analyze the writing patterns of CET4; thirdly, the experiment is presented through algorithms design and results; fourthly, we set some important conclusions from above results and implied the future work.

II. SYSTEM ARCHITECTURE

In this section, we introduce the system architecture and features. We have developed a system to give the reviewers a referenced score and a list of essay features to facilitate making comments and scoring. In the research, our AES system is more objective and reliable then humans, which is not affected by the personal emotions and knowledge background. In this system, there are three main sections: *essay Indexing, classifier learning* and *Scoring and Evaluation,* which are shown in Fig1.

Essay Indexing denotes activity of mapping an essay into compact representation of its writing quality that can be directly interpreted by a classifier. It can be subdivided into pre-processing, extracting features and feature selection, which will be discussed in section III.

Classifier Learning describes the process to building scoring model. Essays belonging to each two different categories will be tagged as ± 1 and train binary classifying

978-1-4577-0856-5/11/$26.00 ©2011 IEEE

models. Then, the result applies approaches of classifier committees to gain the final model.

Scoring and Evaluation: Get the test essay data and train models to predict the scores and evaluation that will be further discussed in Section V.

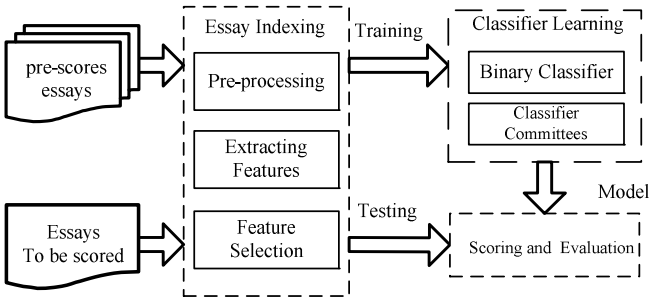

Fig. 1 System Architecture.

III. System In Detail

Our AES system contains three main kinds of features to evaluate essay quality: *language quality, content* and *organization*.

A. Language Quality

Considering how to measure the language quality, the most important evaluation criterions from the English Experts[5] are fluency, accuracy and complexity. Then the basic features we deal with are listed in Table I[5][6][13]:

TABLE I
Features in Language Quality

Fluency
Number of words in essay (Words)
Fourth root of the number of words in essay (RootWds)
Number of different words (Types)
Number of sentence (Sents)
Percentage of Preposition (Preps)
Percentage of the definite article (The)
Accuracy:
Percentage of misspelled words (SpellEr)
Number of syntax errors checked by Linkgrammar (GramEr)
Complexity:
Lexical Complexity
Average characters in a word (Wordlen)
Standard Deviation of Wordlen(WordSD)
Ratio of number of uncommon words to common word (UnC2C)
Number of nominalizations (Noms)
Type-Token Ratio(TTP=Types/Words)
Index of Guiraud(IG= Types/ $\sqrt{\text{Words}}$)
Percentage of different part-of-speech tags among words (POS)
Syntactic Complexity
Average sentence length(Sentlen=Words/Sents)
Average sentence depth (average of notes in parser tree)
Percentage of sentences contain clause (ComplexSents)
Coleman-Liau Index (5.89Wordlen-29.5/Sentlen-15.8)
Numbers of different grammatical relations between words analyzed by parser (Dependency)

B. Content

Content quality is one of the most important factors for scoring essays. Latent Semantic Analysis(LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text[4]. In our AES system, we have modified classical LSA to achieve high accuracy in two ways.

1) Term Frequency and Inverse Documentation Frequency (TFIDF)[7] is one of the key algorithms of term weighting to represent the content of the essay. Instead of TFIDF, we implement TF*IWF*IWF algorithm[8] and improved TF*IWF*IWF algorithm[9] to calculate the Term weights, which consider the skewed dataset and reduce the reliance of the weight of term on Term Frequency. The weight formula of improved TF*IWF*IWF algorithm are presented as follow

$$w(w_i,d)=\sqrt{\frac{\sqrt{\sum_j (p_{ij}-\bar{p}_i)^2}}{\sum_j p_{ij}}}\times(\log(\frac{N(w_i)}{N}))^2\times\sqrt[n]{p_{ij}} \qquad (1)$$

Here, $p_{ij}=T_{ij}/L_{ij}$, L_{ij} is the number of terms in category c_j , T_{ij} is times that term i occurs in category c_j ; $p_{id}=T_{id}/L_d$, L_d is the number of terms in essay d, T_{id} is times that term i occurs in essay d; $\bar{p}_i=\sum_j p_{ij}/m$, m is number of categories, N is the number of all essays in the data set, $n\geq 1$.

2) After SVD approach reducing synonymous noise and improving information effectiveness, we generate similarity measures between document vector and each class vector.

C. Organization

In English writing, organization is the internal structure of the writing. Proper organization may make the reader follow the writer's train of thought clearly and make reading easier. Thus, the overt use of organizational devices may help the reviewer identify the main ideas and the logic employed to develop an argument. In this sub-section, some measures [13][2] of Organization are listed as follows:

Number of paragraphs (Paragraphs)
Number of discourse conjuncts (DisConj)
Percentage of demonstratives in words (Demons)
Percentage of pronouns in words (Prons)
Number of connectives (Connects)
Number of Procedural Vocabulary Items
(PVS: Procedural vocabulary [10] plays in signalling of the rhetorical function of a specific stretch of discourse.)

IV. Feature Processing

Due to the high dimensionality (thousands of features) of the associated feature space and a relative small number of training samples, we must, therefore, guard against the potential problems of "curse of dimensionality". In our AES system, all of the three main kinds of features are used for classifying. Here, several dimensionality reduction techniques are used to generate useful features to improve classification effectiveness and computing efficiency [11][12].

1) Document Frequency (DF): the number of documents containing the feature. It assumes that the features with higher document frequency are more informative for classification.

2). Information Gain (IG): IG measures the number of bits of information obtained for category prediction by recognizing the presence or absence of a feature in a document. The information gain $G(f)$ can be present as follow:

$$G(f) = -\sum_{i=1}^{m} P(c_j)\log P(c_j) + P(f)[\sum_{j=1}^{m} P(c_j \mid f)\log P(c_j \mid f)] + P(\bar{f})[\sum_{j=1}^{m} P(c_j \mid \bar{f})\log P(c_j \mid \bar{f})] \quad (2)$$

Here, $P(c_j) = N_j / N$, N_j is the number of essays in c_j, $P(f)$ is the probability of a essay contains feature *f*, otherwise is $P(\bar{f})$; $P(c_j \mid f)$ indicates the probability that a essay contains feature *f* and also belongs to c_j; $P(c_i \mid \bar{f})$ is otherwise.

3) x^2 Statistics (CHI): The CHI measures the lack of independence between a term and the category. It is defined as

$$CHI = \frac{N_{all}(AD - CB)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)} \quad (3)$$

Here, *A, B, C* and *D* are defined in Table II.

TABLE II
COEFFICIENTS IN CHI

	contain f	not contain f
belong to c_i	A	C
not belong to c_i	B	D

5) Principal Component Analysis (PCA): PCA defined as an orthogonal linear transformation, which is used to reduce dimensionality through keeping the features with greater importance.

V. EXPERIMENTS AND RESULTS

A. Data Set

The data set in our research is the real data from computer-based CET4 of June, 2009. Comparing to the paper-based test, all essays, without the transition from hand-writing format to electronic format, would reflect the real condition of student essays. The writing section in this test requires students to write an essay (no less than 120 words) after a video. There are total 5877 essays corresponding to 8 topics in our data set.

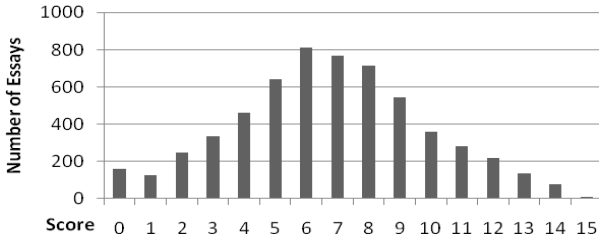

Fig. 2 Score Distribution.

In the compute-based CET4 writing, the score ranges from 0(worst) to 15(best). Due to the large scale, we rearrange the scores into 5 levels according to the essay scoring rules. For example, 13, 14 and 15 are regarded as the highest level. The scores' distribution shows in Fig.2

B. Classifier Algorithm

We have conducted novel experiments with KNN, SVM, Regression, Native Bayes, Feed forward Neural Networks, Adaboost classifiers[11][12] to explore the cross-method evaluation problem. The voting method has been implemented to improve the accuracy and solve the multiple classification problem, due to there are 5 score levels according to scoring rules in CET4. (level 1:0-3,level 2:4-6, level 3:7-9,level 4:10-12;level 5:13-15).

TABLE III
VOTING ALGORITHMS

STEP 1: Input the train data
STEP 2: Figure out the features: Fluency, Accuracy and Complexity
STEP 3: each two of them are used to train binary classifier models through features of content and selected features
STEP 4: Use the classifier committee algorithms are used to predict the final essay score through binary classifiers
STEP 5: Output the results: exact and adjacent agreement

C. Evaluation:

Exact and Adjacent agreement: The exact agreement occurs when two raters give the same score for an essay. The adjacent agreement indicates different raters who assign an essay within one point scale. Assuming the testing essays are scored from 1 to *m*, s_{ij} presents the number of essays that gain *i* point by human raters, and *j* point by scoring system.

$$Exact\ agreement \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} e_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} s_{ij}}, Where \begin{cases} s_{ij} & if\ |i-j| = 0 \\ 0 & otherwise \end{cases} \quad (4)$$

$$Adjacent\ agreement \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} s_{ij}}, Where \begin{cases} s_{ij} & if\ |i-j| \le 1 \\ 0 & otherwise \end{cases} \quad (5)$$

D. Results

1) Compartments among classifying algorithms

In our AES system, we have implemented the 6 scoring algorithms mentioned above. To examine the robustness of performance for our AES classifying algorithms, a 5-fold cross-validation testing procedure is employed to determine the average classification accuracy. From the TABLE IV, take topic 73 for example, we can see that Adaboost algorithm combined with KNN has the highest accuracy about 62%. And the accuracy of most of others is around 40%, such as the KNN, Regress, NNet and SVM, while the NB has a really low accuracy around 17%.

The Adjacent agreement rate is about twice as high as the Exact agreement rate. And all of the rates except NB are above 80% and the highest one is about 96%. It proves that our AES system is of high agreement with the human scores.

TABLE IV
ACCURACY RESULTS ON TOPIC 73

Classifier	Exact agreement rate	Adjacent agreement rate
Regress	47.33%	86 %
SVM	35.33%	80%
KNN	52%	90%
NB	17.3%	37.2%
NNet	36.67%	83.33%
Adaboost (KNN)	62%	96%

Through the cross-validation based on these algorithms, we have further study of scoring result details with AdaBoost (KNN). TABLE V shows results belong to each category for Topic 73 about 70% essays below Level 4 have been scoring exact correctly. But about 60% essays of Level 4 and 5 have

been set into wrong levels. These biased results are resulted from the skewed data. It is shown differently in adjacent agreement, above 86% essays have been arranged a score within one point scale.

TABLE V
SCORING RESULT WITH ADABOOST ALGORITHM

Data sets	1	2	3	4	5	Exact agreement rate(%)	Adjacent agreement rate(%)
1 (23 essays)	16	4	3	0	0	69.57%	86.96%
2 (32essays)	1	22	9	0	0	68.75%	100.00%
3 (50 essays)	0	9	36	5	0	72.00%	100.00%
4 (36 essays)	0	3	12	16	5	44.44%	91.67%
5 (9 essays)	0	0	0	6	3	33.33%	100.00%
Total (150 essays)	17	38	60	27	8	62%	96%
Pearson correlation						0.79876	

On another side, the KNN algorithm has the most stable performance in the features changing and we make a further study to measure how the results accuracy are sensitive to the change of the number of features.

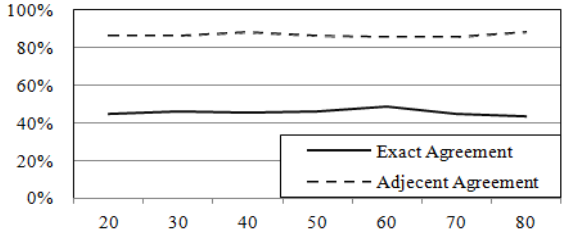

Fig. 3 Agreemnt changes with feature's numbers

During the study, the features are ranked and chosen through the methods of CHI, Information Gain or PCA. From the Fig. 3, we can found that with the increased number of features from 20 to 80, the accuracy of the system is just increasing in a small scale. That means the first 20 features play a key role in the classifying system. The latter 60 features support less information.

2) Evaluation of features effectiveness

What are the most effective features among the hundred ones after feature selection? Correlation coefficients between features and final scores are used to measure the importance for the classifiers.

TABLE VI
FEATURES DISTRIBUTION

Topic ID		1	2	3	4	5	6	7	8
Language Quality	Fluency	5	5	5	5	5	5	5	5
	Lexical Complexity	12	13	13	13	13	13	12	14
	Syntactic Complexity	1	0	1	1	1	0	0	0
Organization		2	2	1	1	1	2	3	1

From the Table VI, the features of Language Quality and Organization take up of the first 20 features. And the Lexical Complexity features support over 12 features (>50%). For the Chinese students, they have limited vocabulary and these are writing patterns for CET4. Moreover, the features of procedural vocabulary and conjuncts from organization are also very important and they are independent with topics.

VI. CONCLUSION

This paper presented a system to assess the computer-base CET4 essays by extracting features: language quality, content and organization. A novel classify algorithm is designed to reduce the negative effects from skewed data and improve the accuracy. Experiment results show the scoring from AES is of high accuracy with Human raters. Adaboost(KNN) has the best performance in our AES system with the accuracy over 62%. From the stable performance of KNN, features from 20 to 80 in the CHI ranking take little effect on the system accuracy. Furthermore, special Chinese pattern are discovered on the Lexical Complexity and Organization features, which needs to research on further.

In the next phase of our research, we will extract more features based on Organization and Content to improve the system accuracy. And we will further research on the Chinese English writing pattern to make the dimensionality reduction and design more efficient classifiers suited for the computer-based CET4.

REFERENCES

[1] S. Dikli, "An Overview of Automated Scoring of Essay", Journal of Technology, Learning, and Assessment, vol. 5, no. 1,August 2006

[2] Y. Attali and J. Burstein, "Automated Essay Scoring With e-rater V.2", The Journal of Technology, Learning and Assessment , vol. 4, no. 3, February 2006

[3] P. W. Foltz, D. Laham and T. K. Landauer, "The Intelligent Essay Assessor: Applications to Educational Technology", Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, vol. 1, 1999

[4] S. Deerwester, S. T. Dumais, G. W. Furnas and T. K. Landauer, and R. Harshman, "Indexing By Latent Semantic Analysis", Journal of the American Society for Information Science, vol. 41, pp. 391-407, 1990

[5] D. S. McNamara, S. A. Crossley and P. M. McCarthy, " Linguistic Features of Writing Quality", Written Communication, vol. 27, pp. 57-86, 2010

[6] L. Larkey, "Automatic Essay Grading Using Text Categorization Techniques", In Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, pp.90-95, 1998

[7] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34, pp. 1-47, 2002

[8] R. Basili and A. Moschitti and M. Pazienza.. "A text classifier based on linguistic processing", In Proceedings of IJCAI-99, Machine Learning for Information Filtering, 1999

[9] K. Chen, C. Zong, and X. Wang, "Analysis on Balance-Corpus and Text Categorization Based on Large-Scale Realistic Corpora",In Proceedings of JSCL'2003

[10]M. Joseâ,, L. N. Marco, and I. Castellon, "Procedural Vocabulary: Lexical Signalling of Conceptual Relations in Discourse", Applied Linguistics, vol. 20, pp. 1-21 1999

[11]H. T. Ng, W. B. Goh and K. L. Low, "Feature selection, perception learning, and a usability case study for text categorization", SIGIR '97 Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 67-73 , 1997

[12]Y. Yang, and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", Morgan Kaufmann Publishers, pp.412-420,1997

[13]E. Robert, Y. S. Schapire, "BoosTexter: A Boosting-based System for Text Categorization", Machine Learning, vol. 39, pp. 135-168, 2000

[14]M. Liang, "Constructing a model for the automated scoring of Chinese EFL learners' argumentative essays", PHD dissertation, Nanjing University,2005