An effective automated essay scoring system using support vector regression

Yali Li
Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences
Beijing, China
e-mail: liyali@hccl.ioa.ac.cn

YonghongYan
Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences
Beijing, China
e-mail: yyan@hccl.ioa.ac.cn

***Abstract*—In this paper, we introduce an effective automated essay scoring system. To implement the system, we extract several features, including the surface features such as the number of words in the essay, number of words longer than 5, and complex features such as grammar checking, sentences, whether the essay is off-topic, the similarity to full-score essays. We get the result of 86% precision given the two scores deviation and average deviation of 0.88 compared to human score on real CET4 data.**

Keywords-component; automated essay scoring, CET4

I. INTRODUCTION

In the English learning process, writing is a major and important part. To evaluate the writing for different tests, many English teachers are required to score a large number of essays in very limited time. Human rater has many defections. Due to fatigue, teachers usually can't check the essays very carefully and different people usually give different scores on a scoring scale of 0 (worst) to 20 (best). The most important, human rater is too expensive. So how to score the essays automatically or alleviate the burdens of the scoring process become a research foci. The purpose of this research is to make the laborious and costly process of scoring the writing of CET4 easy. What about the reliability and validity of automated essay scoring [1]? Studies show that automated essay scoring technology can achieve agreement with a single human judge that is comparable to agreement between two single human judges[2], [3].

Automated essay writing has a long history. The research began in the early 1960s. The first funding to launch the inquiry of essay grade came from the College Board[4]. The approach defined a large number of objectively measurable features in the essays and the results is good

More recently, systems using more complex features received better results such as work at ETS. The mostly used applications were Project Essay Grader, e-rater and so on. Project Essay Grader is one of the earliest and longest-lived implementations of automated essay grading using surface linguistic features. E-rater is a scoring application which extracts linguistically-based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a score to the essay, typically on a scoring scale of 1 (worst) to 6 (best). And e-rater(developed at Educational Testing Service ETS) uses a corpus-based approach to model building to score the Graduate Management Admissions Test Analytical Writing Assessment. In all, there is a great deal of previous work in automated essay scoring [5-11]

Different automated essay scoring systems have different approaches. But in all, the most commonly used methods include Bayesian Text Classification, Latent Semantic Analysis(LSA), and Natural Language processing.[7,8,10,11]

Our problem is a bit different. Specifically, our task is to score the essays of CET4. Based on different writing level and different requirements, the CET4 has its own characters. For example, due to on a scoring scale of 0 (worst) to 20 (best), there will be no much difference between 17 and 18 scores. We used special measures for grammatical errors detection and the overall scoring.

This paper is organized as follows: firstly, we introduced the systems in different part since the overall score has relation with grammar, sentence topic and so on. Then we presented our experiments and results. And last, we give the conclusion and implied the future work.

II. FEATURES IN THE SYSTEM

A. *Simple features*

The basic features we deal with are text-complexity features [12].

1. The number of characters in the document(Chars)
2. The number of words in the document(words)
3. The number of different words (Diffwds)
4. The fourth root of the number of words in the document, as suggested by the Page(Rootwds)
5. The number of sentences in the document(Sents)
6. Average word length(Wordlen=Chars/Words)
7. Average sentence length(Sentlen=Words/Sents)
8. Number of words longer than five characters(BW5)
9. Number of verbs

Each feature has its own use: e.g. the number of words represented the length of the essay since the length requirement of CET4 is between 120-150. This feather can check the empty essay or essay which is so ridiculously short that it cannot be processed; the number of different words means the vocabulary of the author; average word length

978-0-7695-4637-7/12 $26.00 © 2012 IEEE
DOI 10.1109/ICICTA.2012.23

means the word complexity; average sentence length means the sentence complexity; number of words longer than five means the number of complex words and so on.

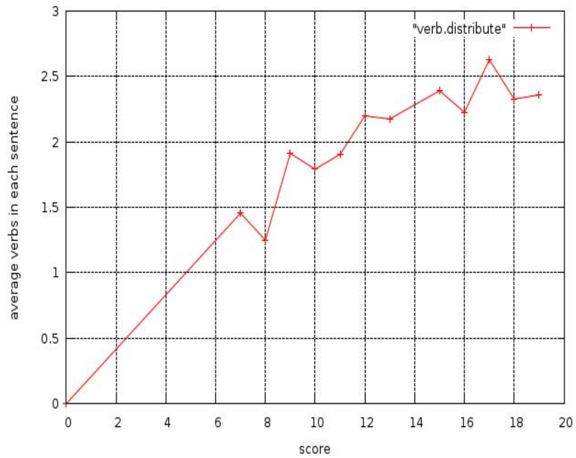

Figure 1. average verbs in each sentence according to the score

The x-label stands for the scores and y-label stands for the average verbs in each sentence. We can see that the number of verbs do affect the score of the essay although it is not the determinate one. Essays with higher scores usually have more verbs.

B. Grammar checking

Grammatical error detection systems such as ALEK(Assessment of Lexical Knowledge) [5], [6] which is a corpus-based tool to check the errors in grammar and help the students correct them. They automatically extract incorrect usage of words based on the differences between the word's context in the essay and the models of context it has derived from the well-formed sentences. They mainly using the bigram and trigram of part-of-speech tag sequence and doesn't take the words itself except the function words into account. Additionally, other than inferring one word at a time, the unit we detect is every two words.

We used a large corpus as the reference background to train models. The corpus are collected from books of 'NEW CONCEPT ENGLISH', 'NEW HORIZON COLLEGE ENGLISH', and 15 years CET4 full-score essays and so on and there are 156313 words in all. These corpus are used to train language model.

We trained both the bigram words model and part-of-speech tags model. For the part-of-speech tag, we use the toolkit of Standford. To avoid the sparseness of the words sequence, we first trained the part-of-speech tags model. We firstly test the essay using the part-of-speech tags model instead of the word model to avoid the sparseness of the word pairs. By calculating the mutual information of the every pairs in the sequence of the essay, we can exact those with low mutual information.

$$MI(a,b) = \log \frac{p(ab)}{p(a) * p(b)} . \qquad (1)$$

Here, $p(a,b)$, $p(a)$, and $p(b)$ is the frequency of ab, a, b appeared in the reference corpus.

But just using the part-of-speech tags model, we can't get the results correctly. There are cases where mutual information of the part-of-speech tags is negative while the mutual information of the words is positive.

That means the pairs is independent by the calculation of part-of-speech tags but related by the words. This accounts for 17.8 percent in the corpus.

There are cases where mutual information of the part-of-speech tags is low while the mutual information of the words is high.

This means that the pairs is wrong in grammar due to the calculation of part-of-speech tags but right by the words. "for example" is frequently used in words sequence so the mutual information is high by word, but its tags of "IN NN" is not frequently appeared and the calculation is low. So just using the tags may have erroneous judgement.

We can see the grammatical errors according to the scores below:

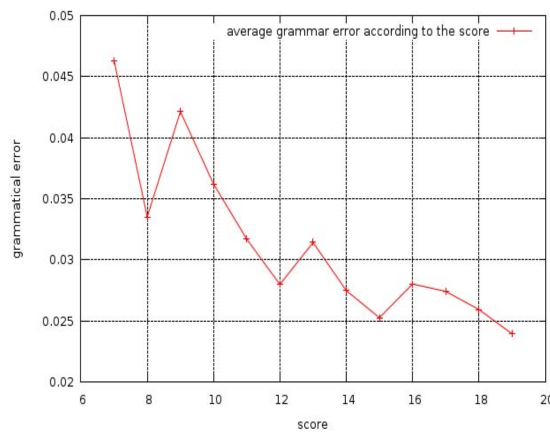

Figure 2. average grammatical error according to the score

The x-label stands for the scores and y-label stands for the average grammatical error rate. We can see that the grammar do affect the score of the essay although it is not the determinate one because the final score is effected by other features. Essays with higher scores usually have lower grammar error rate.

C. Sentence error detection

We use rule-based method to test whether the sentence is right. For example, if one sentence have two predicate verbs (VB，VBP，VBZ etc.) but it is not clause(no 'that', 'when', 'while', 'if' etc.) or has no conjunctions(no 'but', 'and' etc.), we see this sentence is wrong. This is a usual error in students essays.

We found that in CET4 test writings, there are many errors which is more than one predicate verbs in one sentence.

$$\alpha + \beta = \chi. \qquad (1) \qquad (1)$$

D. Topic

If an essay is well-formed and well written both in sentence and grammar but is not consistent with the topic or does not respond to the expected test question, it should also be given a lower score or even 0. Some students copy the essay from the comprehension part or other article firstly recited or they inadvertently cuts-and-pastes off-topic articles. Some students even play joke with the assessor: "Dear teacher, I know that you are not going to let me continue this test if I don`t write my opinion to your argument here. But please be aware that I`m a very special student...".

There are some work to evaluate the off-topic essays[13]. Higgins et.al. use no training data to identify the off-topic essays.

To assess the consistent of essays with the topic, we use two models. The first one is a simple comparing between the topic and the article, and second is the content vector analysis model.

For the simple comparison, we firstly induce the key words in the title, than test the proportion of the key words and their similar words in the article.

The content vector analysis model is suitable for us because it use no training essays based on different topic and we only have one year's corpus. It calculates the cosine value of the content vector and title vector. The process can see as follows:

- Remove the stopwords
- Put all the words except stopwords in the vector
- Calculate the $tfidf$ weight which is $\log(tf+1)*idf$ since the tf maybe zero and idf is $\log\frac{N}{df}$ where tf is the number of times the word occurs in a document and df is the number of documents containing the word and N is the total number of document.
- Calculate the CVA of essay e as

$$CVA(e)=\frac{e*\vec{t}}{|\vec{e}|*|\vec{t}|} \tag{2}$$

where t is the title..

We can see the average topic score of every all score in Fig. The x-label stands for the human scores and y-label stands for average topic scores of our rater.

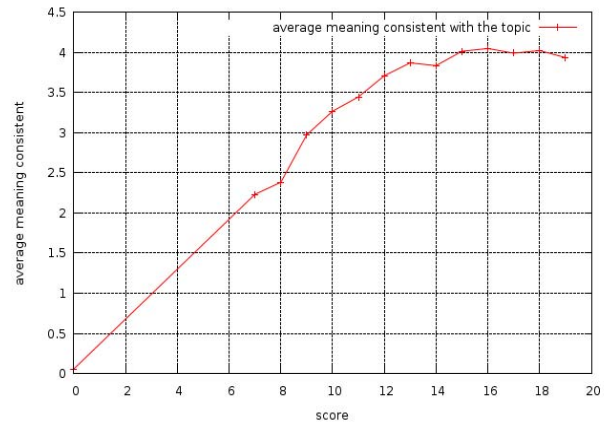

Figure 3. average meaning consistent with the topic

We can see that essays with higher score usually have more consistent with the topic.

E. Similarity to full-score essays

We choose five full-score essays in the same topic. Then we calculate the CVA values between the essay and full-score essays. For the five values, we use the largest value as the similarity to full-score feature.

This feature is important and very effective because teachers always have some references when they score essays. The higher the similarity value, the higher the score of the essay.

III. Experiments and Results

A. Data

The data we use is the real CET4 data of June, 2008. We manually transcribed the hand-writing format to electronic format which requires substantial effort. This would be solved if the test changed to use the computer instead of hand writing. Till now, there are 2041 essays in all which contains three part(training data: 621 essays; developing data: 620 essays and testing data: 800 essays) .

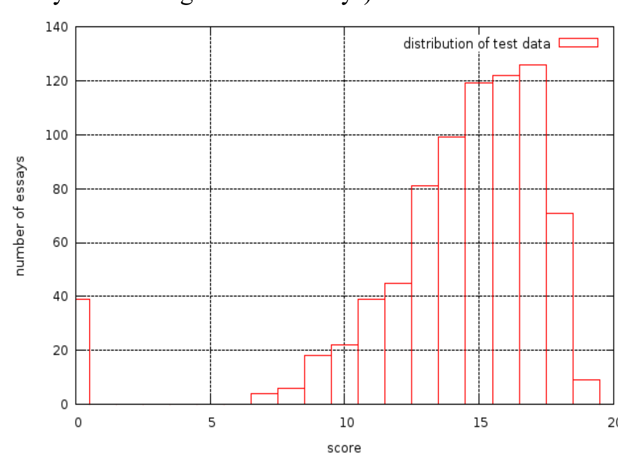

Figure 4. Distribution of the test data

From Fig.3, we can see the score distribution of the test data. The x-label stands for the score and the y-label stands for the number of the essays.

B. Merge several parts using SVR

Since we get scores of different part, how can we get a score in general become the problem. We use linear regression method.

Due to the large scale of 0 to 20, we use a 2 score window to evaluate the precision. That is to say, if the human score is 18 and the result of our automated score is 16 or 19, we think the result is right. Besides this precision, we also used the distribution of the deviation between our score and human score and the average deviation to evaluate the performance.

We use the SVM toolkit. We use regression instead of classification in the system. We first trained a model using the training data, then we test the performance on the test data.

C. Results

We test the performance on test data, and get the results below:

Score deviation	precision in single score	precision within score deviation
0	0.65	0.65
1	0.08	0.73
2	0.13	0.86
3	0.08	0.94
4	0.04	0.99
5	0.01	1.00

We can see that the precision given two deviations is 0.86 and the precision given four deviations is 0.99. We also get average minus of 0.83.

IV. CONCLUTIONS

We introduced our system for automated essay scoring. For the grammatical checking component, we take both the words and part-of-speech tag into account instead of just the part-of-speech tag. For the topic detection component, we use the simple model and CVA model, and find it can effectively detect whether an essay is off-topic especially for large number of essays. Besides this, we also use the similarity to full-score essays feature to improve the system. We get 86% precision given the two scores deviation compared to human raters.

V. . FUTURE WORK

For the future work, we wound like to further improve the performance of automated scoring. The other work we would like to do is to develop a English learning tool for Chinese students instead of just a scorer, such as article detection and discourse structure detection[14].

ACKNOWLEDGMENT

This work is partially supported by The National Science & Technology Pillar Program (2008BAI50B03), National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014).

REFERENCES

[1] J. A. Wohlpart, C. Lindsey, and C. Rademacher, “The Reliability of Computer Software to Score Essays: Innovations in a Humanities Course,” *Computers and Composition*, vol. 25, iss. 2, pp. 203–223, 2008

[2] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-harder, and M. Dee Harris, “Automated Scoring Using A Hybrid Feature Identification Technique,” in *Proc. In the Proceedings of the Annual Meeting of the Association of Computational Linguistics*, 1998, pp. 206–210

[3] P. Ellis and P. Nancy S, “The Computer Moves into Essay Grading: Updating the Ancient Test.” *Phi Delta Kappan*, 1995

[4] P. Ellis, “The Imminence of Grading Essays by Computer,” *Phi Delta Kappan 48*, pp. 238–243, 1966

[5] S. M. Phillips, “Automated essay scoring: a literature review,” TASA Institute, Society for the advancement of excellence in education, 2007

[6] M. Chodorow and C. Leacock, “An Unsupervised Method for Detecting Grammatical Errors,” in *Proc. In Proceedings of NAACL'00*, 2000, pp. 140–147

[7] L. M. Rudner, V. Garcia, and C. Welch, “An Evaluation of IntelliMetric™ Essay Scoring System,” *The Journal of Technology, Learning and Assessment*, vol. 4, iss. 4, 2006

[8] Y. Attali and J. Burstein, “Automated Essay Scoring With e-rater{\textregistered} V.2,” *The Journal of Technology, Learning and Assessment*, vol. 4, iss. 3, 2006

[9] S. Dikli, “An Overview of Automated Scoring of Essays,” *The Journal of Technology, Learning and Assessment*, vol. 5, iss. 1, 2006.

[10] M. D. Shermis and J. Burstein, Eds.*Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, 2003

[11] S. Valenti, F. Neri, and A. Cucchiarelli, "An Overview of Current Research on Automated Essay Grading." *JITE*, vol. 2, pp. 319–330, 2003.

[12] Leah S. Larkey, “Automatic essay grading using text categorization techniques”, ACM Press, New York, US, pp. 90—95, 1998

[13] D. Higgins, J. Burstein, and Y. Attali, “Identifying off-topic student essays without topic-specific training data,” *Natural Language Engineering*, vol. 12, iss. 2, pp. 145–159, 2006

[14] J. Burstein, D. Marcu, and K. Knight, “Finding the {WRITE} Stuff: Automatic Identification of Discourse Structure in Student Essays,” *IEEE Intelligent Systems*, vol. 18, iss. 1, pp. 32–39, 2003