ASSESSING WRITING FLUENCY OF NON-ENGLISH-SPEAKING STUDENT FOR AUTOMATED ESSAY SCORING

How to Automatically Evaluate the Fluency in English Essay

Min-Chul Yang, Min-Jeong Kim, Hyoung-Gyu Lee and Hae-Chang Rim

Department of Computer and Radio Communications Engineering, Korea University, 136-713, Seoul, South Korea

Keywords: CBA, Computer-based Assessment, AES, Automated Essay Scoring, Writing, Fluency, Machine Learning.

Abstract: Automated essay scoring (AES) systems have provided a computer-based writing assessment comparable to expert raters. However, the existing systems are inadequate to assess the writing fluency of non-English-speaking students, while they detect grammatical errors relatively well. The writing fluency is one of the important criteria in essay scoring, because most of non-English-speaking students have much difficulty in expressing their thoughts in English. In this paper, we propose an automated essay scoring system focusing on assessing the writing fluency by considering the quantitative factors such as vocabulary, perplexity in a sentence, diversity of sentence structures and grammatical relations. Experimental results show that the proposed method improves the performance in automated essay scoring.

1 INTRODUCTION

In the era of globalization, the English language has become the most common language. For this reason, many people in non-English-speaking countries have had a great interest in learning English language. English certification tests such as TOEIC, TOEFL and GRE are being used to admit the applicants of many universities or companies. Among the assessment criteria of English, writing has become an important part due to recent upsurge of interest in practical English, but assessing the writing ability is the most expensive and time-consuming activity. Therefore, automated essay scoring (AES), which is one of the educational applications of natural language processing, is standing out as a replacement of expert rater.

Automated essay scoring system is a real-time system which can automatically evaluate an essay and give a score and a feedback to the writer of the essay without any human efforts. E-rater(Attali and Burstein, 2006) and IEA(Foltz et al., 1999) have been used by many students as commercial automated essay scoring systems. In particular, e-rater has been actually utilized in GMAT test. Unfortunately, the core technology of these commercial systems is not published.

The advantages of automated essay scoring system are as follows. 1) It can evaluate an essay in real time. Students can immediately receive ratings for their essay. 2) It requires low cost. 3) It is easily accessible over the web, while an expert rater is required for the manual assessment.

Many studies have found that many automated essay scoring systems are able to provide an automatically measured score that is comparable to human score(Warschauer, 2006; Wang and Brown, 2007). However, they are suitable for assessing essays of English-speaking students, but not non-English-speaking students. The main difference between English-speaking and non-English-speaking students is the ability to express their thoughts in English. Actually, English-speaking students may get a high score of writing if they improve their reasoning skill. However, non-English-speaking students should learn not only the reasoning skill but also the ability to express their thoughts in English. Therefore, for assessing the writing of non-English-speaking students, the evaluation of English expression ability is the essential criteria. In order to automatically assess it, an automated essay scoring system should evaluate an essay by considering the fluency of a sentence or a whole essay, as well as detecting the grammatical error of a sentence. Nevertheless, the existing systems have not fully considered the writing fluency. In other words, an additional technology for measuring the fluency is required to precisely evaluate an essay of a non-English-speaking student.

In this paper, we define the writing fluency as fol-

Yang M., Kim M., Lee H. and Rim H..
ASSESSING WRITING FLUENCY OF NON-ENGLISH-SPEAKING STUDENT FOR AUTOMATED ESSAY SCORING - How to Automatically Evaluate the Fluency in English Essay.
DOI: 10.5220/0003898200830087
In *Proceedings of the 4th International Conference on Computer Supported Education* (CSEDU-2012), pages 83-87
ISBN: 978-989-8565-07-5

Topic: Neighbors are the people who live near us. In your opinion, what are the qualities of a good neighbor?

Good Essay [Grade 5]
... to interrupt other's privacy ... to interfere in almost cases. ... bothers me ... not meddling in other people's concern ...

Poor Essay [Grade 2]
... I will describe that what are the quality of good neighbors ... I would think that good neighbors quality are ... I always think that ...

Figure 1: The examples of good and poor essay.

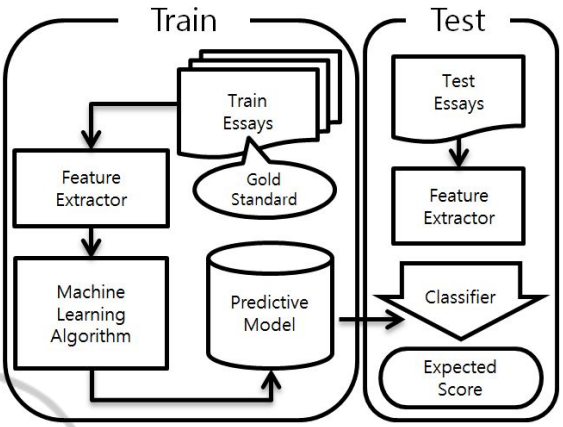

Figure 2: System architecture.

lows: The writing has an easy flow and rhythm when reading aloud(Quinlan et al., 2009) and has similar expressions to native English speakers. According to the definition, we classify the essays into the good and the poor as shown the example in Figure 1.

In the example, a well written essay has synonyms and seldom has repeated words, while a poorly written essay has multiple sentences with same structure (e.g., subject + auxiliary verb + verb + that) and also has same word sequences, which are repetitively used in the essay. The repetition of the same sentence structure and vocabulary usage makes the reader feel uncomfortable. The poorly written essay also has a grammatical error of using number disagreement. Accordingly, in this paper, we propose an automatic method of evaluating the writing fluency by analyzing these common errors of non-English-speaking students.

The rest of the paper is organized as follows. In Section 2, we examine previous works related to this paper. In Section 3, we describe how to measure the writing fluency based on a machine learning technique. In Section 4, we show the experimental results and analysis. Finally, Section 5 concludes the paper.

2 RELATED WORK

Early study was to apply a document classification method to automated essay scoring(Larkey, 1998). It used only simple features such as the number of type and token of words, sentences and words longer than a certain length. Unfortunately, it could just evaluate surface part of the essay.

Other method for assessing writing style of essay(Burstein et al., 2003) tried to include proportional occurrence of the word and previous occurrence distance in addition to proposed features in the previous study. It focused on utilizing styles for evaluating fluency of the essay. However, implied fluency of the essay was not considered.

The fluency is an important factor besides automated essay scoring. The research on detecting a grammatical error in a sentence(Sun et al., 2007) measured perplexity of lexical and part-of-speech n-gram. Nonetheless, the specific measurement of naturalness of a sentence using sentence perplexity is required in order to assess the fluency of the essay.

Other evaluation method about spoken English fluency(Deshmukh et al., 2009) utilized similar n-grams as an additional feature. A method for finding similar n-grams was used to measure similarity by considering edit-distance such as insertion, deletion, and substitution. However, it only considered surface lexical sequences, but not sentence structures of part-of-speeches.

3 PROPOSED METHOD

In this study, we have implemented an automated essay scoring system based on a machine learning framework. The system architecture is shown in Figure 2. Preprocessing techniques are performed by using natural language processing techniques such as sentence segmentation, morphological analysis, part-of-speech tagging[1], and dependency parsing[2]. Then, discriminative features are extracted. The machine learner makes a predictive model with the features in the training step, and the classifier uses the model and extracted features to score the essays.

In this paper, we propose the following features for assessing fluency.

[1]http://nlp.stanford.edu/software/tagger.shtml
[2]http://nlp.stanford.edu/software/lex-parser.shtml

3.1 Vocabulary

Firstly, we measure the author's paraphrasing ability. In order to paraphrase an expression, the author needs to have rich vocabulary knowledge. Therefore, we measure the vocabulary usage as follows. 1) We pairwise the words which have same part-of-speeches, especially noun, verb, adverb, and adjective. 2) We check if they are synonyms based on WordNet[3](Miller, 1995).

3.2 Diversity of Sentence Structure

Secondly, the diversity of sentence structure is measured based on the monotony of the sentence form. For simplicity, we assume that the sentence structure is represented by n-gram(contiguous sequence of n items). We count the repetitions of each n-gram in an essay. Lexical bigram and part-of-speech trigram are used because of data sparseness problem.

3.3 Perplexity in a Sentence

Thirdly, we compute the perplexity sentences of essay by using external documents according to our intuition; if an essay consists of expressions which are frequently used by skilled authors, the essay must be fluent. The perplexity is a common way of evaluating a language model that is a probability distribution over sentences. As our language model unit, lexical trigram and part-of-speech 5-gram are used. The lexical language model can identify grammatical errors such as passive, agreement and prepositions. The part-of-speech language model can detect errors of sentence structure.

3.4 Grammatical Relation

Finally, we examine relationships between words like subject-predicate. Even if the sentence is complex, proper use of advanced grammatical relations, such as direct object and conjunction, can make the essay natural and clear. We use a syntactic parser(Marneffe et al., 2006) to extract these relations.

3.5 Feature Set

The whole features implemented in our experiments are shown in Table 1. [F_1, F_2, F_3] are the feature sets used in previous studies and [F_4] is the proposed feature set.

[F_1] is a surface feature set which does not require natural language analysis. [F_2] is style feature set, which includes various frequency features representing the author's writing style. Among these, the advanced word features are counted according to high-level dictionary, and the density feature, known as one of the most useful features measuring quality of a document, is calculated as '# of word types / # of word tokens'. [F_3] is the fluency feature set presented in the previous study(Burstein et al., 2003), and [F_4] is the proposed feature set which is separately grouped for the comparative evaluation.

Table 1: Feature description.

Category	Description
Surface (F_1)	Number of [word tokens / characters] Number of sentences
Style (F_2)	Number of word types Number of each POS Number of word longer than N characters Length of [words / sentences] Use of [advanced words / word phrases] Word density ($= \#types/\#tokens$)
Fluency (F_3)	Previous occurrence distance Proportional occurrence of word
Proposed Fluency (F_4)	Number of synonym word pairs Number of [lexical 2-gram / POS 3-gram] Number of each grammatical relation Perplexity in a sentence

The actually used feature values in the training step and the testing step are determined according to the characteristics of each feature. For example, in perplexity, the average as a representative value is selected and inappropriate sentences are determined by counting sentences more than a threshold. In addition, the maximum, minimum and variance of perplexity can be used for identifying unknown properties. We also use normalized values to produce the normal form of the entire corpus.

4 EXPERIMENTS

We have evaluated the performance of our automated essay scoring system according to correlation and accuracy. To analyze the results in more detail, we have used two kinds of scores for training and evaluation: total score and style score.

4.1 Experimental Setup

We have collected 2,675 essays written by Korean stu-

[3]WordNet 3.0: http://wordnet.princeton.edu/wordnet/

dents. The corpus consists of 10 topics and each essay is rated based on the total score and the style score by two humans who are expert raters. Both scores have 6-point score scale; the total score is based on the rubric of TOEFL Writing scoring and the style score is graded according to fluency, readability, simplicity, and word usage. The language model, for the perplexity feature described in Section 3.3, is trained from news corpus consisting of 33 million sentences. To build the language model and measure perplexity, SRILM(Stolcke, 2002) toolkit[4] is used. The results are reported using 10-fold cross-validation.

4.2 Evaluation Metrics

We report the performance of our system in terms of correlation and accuracy. Eqs. (1)-(3) define each of these metrics.

$$r = Correlation(X,Y) = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y} \quad (1)$$

Where X and Y are random variables, and E is the expected value operator. μ means expected value, and σ means standard deviation. Since the Pearson correlation refers a statistical relationship with scores of the human and the system, it is the most commonly used evaluation metric in automated essay scoring area.

$$Exact\,Accuracy = \frac{D_0}{N} \quad (2)$$

$$Ad\,jacent\,Accuracy = \frac{D_{-1} + D_0 + D_1}{N} \quad (3)$$

Where N is the number of total documents, and D_n is the number of documents that have n points of difference between the human and the system score. Likewise, *Ad jacent Accuracy* means that the human and the system score are within 1 point of difference.

4.3 Results and Analysis

The result of correlation between scores of the human and the system is shown in Table 2 and 3. $H_{1,2}$ indicates the human scores of essays. G_{total} and G_{style} indicate gold standard which represents the average of two humans' score in each part. The notations related to features are described in Table 1. The combinations of these feature sets indicate the baseline systems described in Section 2 and the proposed system described in Section 3. The correlation scores of the baseline systems are shown on columns 3 and 4 of Tables 2 and 3. In the same manner, the correlation scores of the proposed system are shown on the column 5 of the Tables 2 and 3.

Table 2: Correlation of scores on total part.

Total Part				
X	H_1	G_{total}	G_{total}	G_{total}
Y	H_2	F_1+F_2	$F_1+F_2+F_3$	*All*
r	0.578	0.497	0.501	**0.525**

Table 3: Correlation of scores on style part.

Style Part				
X	H_1	G_{style}	G_{style}	G_{style}
Y	H_2	F_2+F_3	F_2+F_4	$F_2+F_3+F_4$
r	0.465	0.342	**0.368**	**0.365**

Overall, the performance of our method is similar to humans in terms of both total score and style score. Experimental results show that the performance of our method is better than the performance of all baseline systems with regard to correlation. In particular, F_4, which is the proposed feature set, improved about 5% in total part and 15% in style part. Therefore, our system is proved to be able to measure the fluency of the essay better than others.

Table 4: Accuracy of system scores on total part.

	F_1	F_1+F_2	$F_1+F_2+F_3$	*All*
Exact	43.55%	57.12%	57.87%	**58.32%**
Ad jacent	90.50%	95.18%	95.25%	**95.93%**

Table 4 reports the accuracy of the proposed system. It also shows that the proposed system outperforms the baseline systems.

5 CONCLUSIONS

In this paper, we have proposed an automatic measurement for scoring non-English-speaking student's essay. The proposed method used a combination of novel semantic and implicit features in machine learning to compute the overall score of fluency.

In order to verify the proposed method, we have evaluated essays of Korean as non-English-speaking students. Experimental results showed that our system outperforms all baseline systems. It also had a small difference with expert raters in terms of correlation.

For future works, we are going to build a language model based on large corpus which can reflect the characteristics of essays. Furthermore, we will develop a system which can give feedback to students who want to learn English.

[4] http://www.speech.sri.com/projects/srilm/

REFERENCES

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment (JTLA)*, 4(3).

Burstein, J., Wolska, M., and Saarlandes, U. D. (2003). Toward evaluation of writing style: Finding overly repetitive word use in student essays. In *In proceedings of EACL-03*.

Deshmukh, O., Kandhway, K., Verma, A., and Audhkhasi, K. (2009). Automatic evaluation of spoken english fluency. In *ICASSP'09*, pages 4829–4832.

Foltz, P. W., Laham, D., and Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Education Journal of Computer enhanced learning On-line journal*, 1(2).

Larkey, L. (1998). Automatic essay grading using text categorization techniques. In *In Proceedings of SIGIR-98*, pages 90–95.

Marneffe, M. D., Maccartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *In LREC 2006*.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Quinlan, T., Higgins, D., and Wolff, S. (2009). Evaluating the construct-converage of the e-rater scoring engine.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *ICSLP*, pages 901–904.

Sun, G., Liu, X., Cong, G., Zhou, M., Xiong, Z., Lee, J., and yew Lin, C. (2007). Detecting erroneous sentences using automatically mined sequential patterns. In *In Proceedings of ACL-07*.

Wang, J. and Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *The Journal of Technology, Learning and Assessment (JTLA)*, 6(2).

Warschauer, M. Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *LANGUAGE TEACHING RESEARCH*, 10(2).