Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi

Vol. 8, No. 4, Desember 2016

ISSN 0216 – 0544
e-ISSN 2301– 6914

PRE-PROCESSED LATENT SEMANTIC ANALYSIS FOR AUTOMATIC ESSAY GRADING

[a]Ruth Ema Febrita, [b]Wayan Firdaus Mahmudy
[a,b] Magister of Computer Science Brawijaya University, Veteran Street, Malang 65145
E-mail: ruthemaf@gmail.com

Abstract

In education, essay is considered as the best tool to evaluate student's high order thinking and understanding. In the other hand, manual processing and grading essay answers by a teacher need much time and tending to subjectivity grading. Meanwhile automatic essay grading in e-learning system find the difficulties in comparing model or key answer to student's answer because student's can answer the question with so various way. That means a right answer also can be so various, for they have same semantic meaning. This paper proposed automatic essay grading using Latent Semantic Analysis. But before the texts being scored, they will be pre-processed using stop words removal and synonyms checking. Calibration process implemented for dealing with the various possible right answer and help to simplify the term matrix. Implementation of this approach using Java Programming Language and WordNet as lexical database for searching the synonyms of every given words. The accuracy obtained by this method is 54.9289%.

Keywords: Automatic Grading, Essay Grading, Latent Semantic Analysis, preprocessed LSA

INTRODUCTION

Evaluation in education is a series of activities for measuring and evaluating students' abilities and understanding toward the learning process [1]. There are few techniques for measuring student's ability, such as multiple-choice question, essay question, short-answer question, and project-based examination. Since 2014, many senior high schools in Indonesia left Paper Based Test (PBT) and began to implement Computer Based Test (CBT). CBT system currently provides a series of multiple-choice questions for measuring the student's abilities and comprehension. Multiple-choice questions also widely applied to e-learning system for the ease of implementation and robust assessment grading [2].

Several studies in the education field has reviewed the comparison of several assessment techniques, especially the comparison between multiple-choice assessment and essay assessment. Scouller [3] found that the essay-based assessment techniques can measure the student understanding and ability in high order thinking compared to multiple-choice questions [4]. The obstacles encountered in implementing essay in CBT system is the difficulties in grading process, for essay examination allows students to express their answer in very open answer, use their own language, diction and their creativity. Meanwhile grading essays manually by teachers require a lot of time, as well as allowing their subjective grading. Therefore, automatic essay grader is needed.

Some previous studies has used many approaches in grading essay automatically. Zen [5] used Latent Semantic Analysis (LSA) for grading computer programming assignment and found that LSA can grade essay consistently and faster than manual grading by human, but LSA is lack in detecting the order of commands and symbols in the program. Meanwhile, He et.al. [6] used the combination of LSA and n-gram technique for grading summary assessment. N-gram is used to cover the lack of LSA in detecting the order of words. The combination of LSA and n-gram can achieve better accuracy.

In this research, LSA with pre-processing method is proposed to build automatic essay grading. Pre-processing method is removing the stop words and checking the synonym of each significant words. We want to know how far synonym checking can affect the accuracy of the essay grading. Synonym checking is used to increase the system's flexibility in assessing students' answers which have the same meaning but use different vocabulary. The novelty of this research is the use of score calibration in determining the standard of answer similarity allowed.

The rest of this paper is organized as follows: in Section 2, we'll present some previous researches in automatic essay grading; in Section 3, we'll explain more about the research methodology, the framework to follow and also the score and accuracy calculation; in Section 4, we'll present the research result and discuss about the performance of the proposed method; in Section 5, we'll present the conclusion of this research.

LITERATURE REVIEW

There are several methodologies had been applied in the automatic essay grading researches, which are elaborated below.

Latent Semantic Analysis (LSA)

LSA is a basic method that analyzes the texts to extract its semantic meaning by using support vector machine (SVM) and checked their similarities between the two texts using cosine similarity [5][7][8]. In LSA approach every term found in the text, sentences or documents will be mapped into the term matrix. The meaning of a text will be measured in statistic way based on the relationship among terms in the matrix [9]. LSA does not extract the meaning of the text from the sequence of the words, so that LSA cannot be applied to extract information from text, which sequence and order are important, like programming code and grammatical essay [5][9].

LSA matrix record every term appeared in every document with the appearance frequency. The row in the matrix shows the list of terms found in documents. Meanwhile, the column shows the list of documents compared and the cell will record the appearance frequency of terms in a document. That is why the dimension of the matrix depends on the number of terms and documents.

The use of LSA as the single method achieving low accuracy due to the limited checking of the text order and have a very high-

dimensional matrix. Hybridization can be applied to improve the accuracy of LSA. Genetic algorithms can be applied to reduce the dimensions of the matrix [10]. LSA method is used as the basic method of building automatic graders because it is able to compare the meaning of student answers and key answers. This is important because there is no guarantee that student will answer exactly same answer as the key answer, so syntactic text analysis is not possible.

N-gram

N-gram is a method to analyze text by splitting the text into a set of single words (unigram), two words (bigram), three words (trigrams), and so on. Research conducted by Tripathy [11] showed that the combina-tion of unigram, bigram and trigram will achieve better accuracy in the user sentiment classification. The combination between LSA and n-gram machines also shows better accuracy [6].

Natural Language Processing (NLP)

NLP approach can also be used for grading essay automatically, by extracting the information contained in the answer model text and student's answer text, by labeling each word as predicates, nouns, etc [12]. The similarity between the two texts can be calculated based on the similarity of the structure of parse tree [13].

METHODOLOGY

This study proposes pre-processed LSA as semantic-based techniques, which process the texts to extract the semantic meaning of the text. Figure 1 will show the methodology used in this paper. As the general steps in LSA [5][7], the text will go through stop words removal, building the term matrix and calculating the similarity score between two texts using cosine similarity.

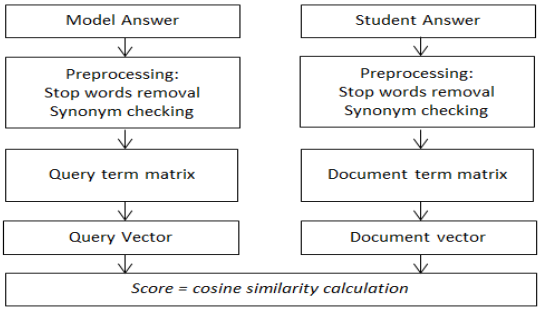

Figure 1. Proposed methodology diagram

Preprocessing

Preprocessing will be done in two steps. First is stop-words removal. Stop words means every word that doesn't have important meaning, like "the", "is", "are", "there", etc. NLTK feature will be adopted in this study to remove the stop-words.

The second step is synonym adding to provide flexibility to the vocabulary used by students in their answer. The synonyms for every term which is not in the stop word list will be added in the document text. WordNet Lexical Database [14] is used for checking the synonyms of a given word.

Term Matrix Building

In this stage, any terms contained in each text answers will be mapped into the matrix. Each row in the matrix shows the terms found, while the columns describe each document/ text answers to be tested. The document (d) refers to each student answer, while the query (q) refers to the model answer that will be used as a comparison standard in similarity calculation. Figure 2 will show the example of document term matrix.

Terms	D1	D2
transaction	1	0
buyer	1	0
seller	1	0
product	1	0
service	1	1
activity	0	1
buy	0	1
sell	0	1
good	0	1
generate	0	1
profit	0	1

Figure 2. Document term matrix

Building Vector Space using TF-IDF

A weighting value is given to each term to measure how important the term in the document and relationship with another document. The weighting value is calculated with Term Frequency Inverse Document Matrix like shown in Equation (1) below:

$$w_{t,d} = tf_{t,d} \cdot \log \frac{N}{d_{tf}} \quad (1)$$

Where $tf_{t,d}$ refers to the frequency of term t appears in document d. $\boldsymbol{N}$ refers to a number of the document processed, dtf shows the number of documents which contain terms t. Table 1 shows the TF-IDF calculation. The more often a

word appears in many documents indicates that the word has no such important role in context.

Building Vector Space using TF-IDF

Cosine similarity method will be used to measure the similarity between text, which is shown in Equation (2).

$$\text{cossim}(\text{dj}, \text{q}) = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \quad (2)$$

where,

$\sum_{i=1}^{N} w_{i,j}\, w_{i,q}$ is a vector multiplication between document d vector to query vector.

Measuring the Accuracy

The accuracy means how close the score that generated by the system to manual teacher's scoring. RMSE like shown in Equation (3) is used to calculate the accuracy.

Table 1. TF-IDF Calculation

Term	d1	d2	DF	IDF	ID
Transaction	1	0	1	log(2/1)	0,3
Buyer	1	0	1	log(2/1)	0,3
seller	1	0	1	log(2/1)	0,3
Product	1	0	1	log(2/1)	0,3
Service	1	1	2	log(2/2)	0
Activity	0	1	1	log(2/1)	0,3
Buy	0	1	1	log(2/1)	0,3
Sell	0	1	1	log(2/1)	0,3
Good	0	1	1	log(2/1)	0,3
Generate	0	1	1	log(2/1)	0,3
Profit	0	1	1	log(2/1)	0,3

$$\text{RMSE} = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} e_i^2 \quad (3)$$

The greater RMSE value indicates lower accuracy. So the accuracy can be calculated using Equation (4).

$$akurasi = 100 - RMSE \quad (4)$$

RESULT AND DISCUSSION

Answer Key

LSA method is implemented using Java programming language. To connect with WordNet for searching the synonyms we used RiTa API.

The program was tested on 10 students' essay with the topic is about economic activity. To test the performance of similarity checking, we need an essay answer topic that has an explanatory answer, not an implementation of formulas (such as Mathematics or Physics). Therefore in this experiment, we choose Economic subjects that explain about the understanding of the economic activity. There are nine key answers provided by researchers which are obtained from several sources of books and internet. Each key answer is then assessed by an economist and an Economics teacher and found that A1 got the highest score because it was a complete answer and it can cover the other answer keys. The following is a list of the key answers that have been provided by the researcher:

A1: Activity to obtain good and service to meet the need, includes production, distribution, and consumption
A2: Activity interchange good or service to gain profit
A3: Produce good or service to provide human need, sell and buy them
A4: All activities to provide human needed, such as good and service
A5: Trading good or service to obtain human need and benefit
A6: Human activity to get human need
A7: Produce good or service with added value to reach human need
A8: Trading good or service
A9: Using good or service to fulfill human need

Because A1 got the highest score, so A1 is chosen as the key answer for direct comparison to student's answer.

Score Calibration

The experts have chosen A1 as the complete answer key, so A1 will be compared with student's answer to calculate the score. The selection of just one key answer to be compared directly with student's answer will make the term matrix much simpler so the score processing becomes faster. To accommodate students' answers which match or are similar to the other answer keys which are also considered as the correct answer, then calibration process is needed.

Table 3. Score calibration

Key Code	Score
A2	58,08
A3	19,15
A4	3,4
A5	11,95
A6	5,6
A7	0
A8	0
A9	3,34
Average	**12,69**

The calibration process is done by finding the average of similarity score between the selected answer key (A1) with the other answer keys. The calibration process is presented in Table 3.

The student who gets a score greater than calibration value can be considered as a perfect answer and receive the maximum score (100). Equation (5) show the score calculation for students' answer.

$$score\ (LSA)\begin{cases}100, & LSA \geq 12{,}69\\ \frac{LSA}{12{,}69}\ X\ 100, & LSA < 12{,}69\end{cases} \quad (5)$$

Essay Grading

In this study, we used 10 distinct student answers that were sufficient to illustrate the diversity of student answers related to the questions given. Below is a list of 10 distinct student answers.

S1 : Transaction between seller and buyer for product or service
S2 : An exchange of service or good for reach mutual benefit between the two sides
S3 : An activity that focus on gaining advantages, specifically money through simple actions
S4 : An activity involving economic sector, such as selling product or buy product
S5 : A series of activity to increase interest in economic area, such as profit, sales, marketing, etc.
S6 : Activity of buy and sell good and service that generate profit (money)
S7 : An activity or process that results in a transaction between two person
S8 : Activity sell or buy goods made by two people
S9 : Activity to obtain good and service to meet the need, includes production, distribution, and consumption
S10 : Activity done by two or more people, with the goal of mutual benefit

Each student' answers compared with A1 and assessed using Equation (5) to obtain results as described there Table 4. Based on Table 4, RMSE can be calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{10}}\ \text{x}\ 20314{,}06$$
$$= 45{,}0711$$

So the accuracy of the LSA system developed in this study amounted to 54.9289%.

Table 4. Comparison between expert score and system score

	Expert score	LSA score	Student	error	error^2
S1	80	7,345	57,88	-22,12	489,2846
S2	70	2,55	20,09	-49,91	2490,553
S3	70	5,64	44,44	-25,56	653,0864
S4	70	2,35	18,52	-51,48	2650,343
S5	70	0	0	-70	4900
S6	70	6,2	48,86	-21,14	447,0109
S7	70	0	0	-70	4900
S8	70	6,18	48,7	-21,30	453,7001
S9	100	100	100	0	0
S10	70	1,56	12,29	-57,71	3330,081
				Total	**20314,06**

Discussion

Building the term matrix is a very important process in LSA because it will be the reference of calculating the term frequency and inverse document frequency to get the similarity score. The problems are often encountered in the LSA is the high-dimension of the term matrix, which causes the process of analysis and similarity score calculation to be long enough. In this study, the selection of one key answer to be compared directly with student key answers is quite effective in simplifying the term matrix so that the process of score calculation becomes faster. Moreover, all the synonyms found for a word will be added to the term matrix so the size of the term matrix becomes larger. In this study, it takes less than one second to process and score all students answers.

The problem encountered is the difficulty of finding the proper synonym of two words which have the same basic word but spelled in different forms. For example, the words "selling" and "sell". Both words only show the different uses of tenses, but they have two different meanings when searching for the synonyms. The word "selling" has several synonyms, are commerce, commer¬cialism, and mercan¬tilism. While the word "sell" has several synonyms namely exchange, change, and interchange. This causes less accuracy. Therefore another pre-processing is needed to get the basic form of a word before checking for the synonyms.

CONCLUSION AND CONTRIBUTIONS

This automatic essay grader has developed to assess student answer in open-mind text answer. This grader is limited to grade English text only due to the synonym checking provided by WordNet Lexical Database. In this research, we used score calibration to get the minimum score considered as the right answer according to several answer keys that have been provided. The calibration process contributes in giving ideas to simplify the term matrix in the LSA method so the scoring process becomes faster.

The results indicate that the score generated by the proposed method gives a value of 54.9289% of the expected score in average. This means that if the student should get the score is 100, the score generated by the system is 54.93. For further research, the addition of the lemmatization process to get the basic form of the word can be done to improve the accuracy because by getting the basic form of the words will result the same synonyms, so the similarity score will be higher. In addition, the algorithm for choosing the most appropriate synonyms also can be applied so the term matrix can be more effective.

Acknowledgement

We would like to appreciate Karunia Tiara Vani, an Economic subject teacher in Charis National Academy and Chris Wijayanti Puspita, a graduate student of Malang State University for contribution as subject experts who check the students' answers. We also would like to say thank to all graduate students in Intelligent System Research Group for all the support and open-mind discussion during the research.

REFERENCES

[1] D. Wiliam, "What is assessment for learning?," *Stud. Educ. Eval.*, vol. 37, no. 1, pp. 3–14, Mar. 2011.

[2] M. Nakayama, H. Yamamoto, and R. Santiago, "The Role of Essay Tests Assessment in e-Learning: A Japanese Case Study.," *Electron. J. E-Learn.*, vol. 8, no. 2, pp. 173–178, 2010.

[3] K. Scouller, "The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay," *High. Educ.*, vol. 35, no. 4, pp. 453–472, Jun. 1998.

[4] S. Brown and P. Knight, *Assessing Learners in Higher Education*. Psychology Press, 1994.

[5] K. Zen, D. N. F. A. Iskandar, and O. Linang, "Using Latent Semantic Analysis for automated grading programming assignments," in *2011 International Conference on Semantic Technology and Information Retrieval*, 2011, pp. 82–88.

[6] A. Abdi, N. Idris, R. M. Alguliev, and R. M. Aliguliyev, "Automatic summarization assessment through a combination of semantic and syntactic information for intelligent educational systems," *Inf. Process. Manag.*, vol. 51, no. 4, pp. 340–358, Jul. 2015.

[7] "GLSA based online essay grading system," *IEEE Conf. Publ.*, 2013.

[8] G. R. Perera, D. N. Perera, and A. R. Weerasinghe, "A dynamic semantic space modelling approach for short essay grading," in *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2015, pp. 43–49.

[9] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, *Handbook of Latent Semantic Analysis*, vol. 1, 2011.

[10] W. Song and S. C. Park, "Genetic algorithm for text clustering based on latent semantic indexing," *Comput. Math. with Appl.*, vol. 57, no. 11–12, pp. 1901–1907, 2009.

[11] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016.

[12] A. O. Ade-Ibijola, I. Wakama, and J. C. Amadi, "An expert system for automated essay scoring (AES) in computing using shallow NLP techniques for inferencing," *Int. J. Comput. Appl.*, vol. 51, no. 10, 2012.

[13] B. Galitsky, "Machine Learning of Syntactic Parse Trees for Search and Classification of Text," *Eng Appl Artif Intell*, vol. 26, no. 3, pp. 1072–1091, Mar. 2013.

[14] Princeton University, "About WordNet," *Wordnet*, 2010.