AUTOMATIC ESSAY SCORING SYSTEM USING N-GRAM AND COSINE SIMILARITY FOR GAMIFICATION BASED E-LEARNING

M Ali Fauzi
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
moch.ali.fauzi@ub.ac.id

Djoko Cahyo Utomo
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
djoko.c.utomo@gmail.com

Budi Darma Setiawan
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
s.budidarma@ub.ac.id

Eko Sakti Pramukantoro
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
ekosakti@ub.ac.id

ABSTRACT

E-Learning is one of the great innovations in teaching methods. In the E-learning, there are several assessment methods, one of them is the essay examination. Essay assessment takes a long time if corrected manually. Therefore, researches on automatic essay scoring has been growing rapidly in recent years. The method that is usually used for automatic essay scoring is Cosine Similarity by utilizing bag of words as the feature extraction. However, the feature extraction by using bag of words did not consider to the order of words in a sentence. Meanwhile, the order of words in an essay has an important role in the assessment. In this study, an automatic essay scoring system based on n-gram and cosine similarity was proposed. N-gram was used for feature extraction and modified to split by word instead of by letter so that the word order would be considered. Based on evaluation results, this system got the best correlation of 0.66 by using unigram on questions that do not consider the order of words in the answer. For questions that consider the order of the words in the answer, bigram has the best correlation value by 0.67.

CCS Concepts

• Applied computing → Education → E-learning

• Information systems → Information retrieval → Retrieval models and ranking → Similarity measures.

Keywords

Automatic Essay Scoring; E-Learning; N-gram; Cosine Similarity

1. INTRODUCTION

By definition, Electronic learning (E-Learning) is an educational system that uses electronic applications to support teaching and learning with Internet media, computer networks, and standalone computers [1]. E-learning is the current educational innovation that widely used in teaching and learning process nowadays. Many types of e-learning have been provided to meet the users need. The most used type is media repository of learning materials with assessment and assignment features. Generally, the assessment features provided are multiple-choice and essays based examination.

In recent research, [2] developed a gamification based E-learning system. Gamification can be defined as the concept of applying game elements and gaming techniques to non-game applications, such as e-learning. It model the whole activity in the teaching-learning process as a game. There are some quests or challenges that must be solved and there are rewards for every solved quests or challenges. The challenges are assignment and assessment, while the rewards are the scores. This concept could increase students' motivation and liveliness since it is fun and engaging. After all, the features in this e-learning is not quite different compared to the common e-learning system such as the repository of learning materials, multiple-choice quizzes, and essays examination.

In its next development, essay examination feature has been a problem. Multiple-choice questions can be corrected automatically, while the essay ones is still have to be corrected manually by the lecturers. This work is very expensive and time consuming. Most of the teachers or lecturers need to spend a lot of time assessing their student or student exam answers [3]. Therefore, automatic scoring system is needed. Since the automatic essay scoring system will be applied to E-learning, there are some factors other than accuracy that have to be considered, such as the server performance. A system with low computational complexity is preferred.

Researches on Automatic Essay Scoring (AES) has grown tremendously in recent years, due to the increasing number of college students and the possibility of providing E-Learning as an asynchronous approach of education that can be accessed anywhere and everywhere [4]. There are several approaches to the automated assessment developed for English such as Project Essay Grade (PEG) [5], Intelligent Essay Assessor (IEA) [6], Electronic Essay Rater (ERater) [7], C-Rater [8], Intelligent Essay Marking System [10], SEAR [11], and Automark [12]. Another approaches using LSA conducted by [13] and [14]. Some researches using text categorization techniques [15] and Support Vector Regression [16] to tackle this problem. Assessment of essay answers in Indonesian language began to be developed since 2005 conducted by [17] used Latent Semantic Analysis (LSA) method. The next research was conducted by [18] using LSA and sicbi weighting.

One of the methods used for Automatic Essay Scoring is cosine similarity. Cosine similarity showed the best result for string similarity measure compared to Dice or Jaccard method [19]. It also has competitive result compared to LSA [20]. Moreover, cosine similarity is less complex than LSA. It has good server performance

with lower CPU usage and faster load time than LSA [20]. Generally, bag of words feature is used for cosine similarity approaches. The problem is the bag of words feature does not consider the order of words in a sentence. Therefore, two different sentences with the same words composition will be considered similar. In some cases, this condition can decline the accuracy of the system. Therefore, N-gram is proposed for automatic essay scoring system along with cosine similarity.

Generally, N-gram is a piece of N-character taken from a string [21]. In this study, N-gram used is a piece of N-word taken from a sentence. N-grams can be used to predict word order [22]. N-gram also have been proved to be better for calculating similarity in short sentences [23]. By using this N-gram, the automatic scoring system is expected to improve its accuracy.

2. N-GRAM

N-gram is a piece of N-character taken from a string [21]. We can get the full N-gram by adding blank or another character at the beginning and end of the string. For example a string "TEXT" after have been added with "_" on the beginning and ending will get N-gram as follows:

Unigram : T, E, X, T

Bigram : _T, TE, EX, XT, T_

Trigram : _TE, TEX, EXT, XT_, T_ _

Quadgram : _TEX, TEXT, EXT_, EX_ _, X_ _ _

In this study, N-gram used is a piece of N-word taken from a sentence. For example we have a sentence "Cosine similarity adalah salah satu metode yang dipelajari saat mata kuliah stki", we can get the N-gram as follows:

Unigram : Cosine, similarity, adalah, salah, satu, metode, yang, dipelajari, saat, mata, kuliah, stki.

Bigram : Cosine similarity, similarity adalah, adalah salah, salah satu, satu metode, metode yang, yang dipelajari, dipelajari saat, saat mata, mata kuliah, kuliah stki.

Trigram : Cosine similarity adalah, similarity adalah salah, adalah salah satu, salah satu metode, satu metode yang, metode yang dipelajari, yang dipelajari saat, dipelajari saat mata, saat mata kuliah, mata kuliah stki.

Combination : Cosine, similarity, adalah, salah, satu, metode, yang, dipelajari, saat, mata, kuliah, stki, Cosine similarity, similarity adalah, adalah salah, salah satu, satu metode, metode yang, yang dipelajari, dipelajari saat, saat mata, mata kuliah, kuliah stki, Cosine similarity adalah, similarity adalah salah, adalah salah satu, salah satu metode, satu metode yang, metode yang dipelajari, yang dipelajari saat, dipelajari saat mata, saat mata kuliah, mata kuliah stki.

3. COSINE SIMILARITY

Cosine similarity is a similarity measurement method between two different texts or documents by measuring the cosine of the angle between the document representation vectors. This method can be used to see the similarity score between two sentences or documents. The cosine value is between 0 and 1. The greater the cosine value, the more it similarity between the two sentences or documents. The cosine value 1 states the similarity of 100%, while if the value 0 means 100% not similar.

The Cosine similarity is measured as follows:

$$\cos(\vec{d1}, \vec{d2}) = \frac{\vec{d1} \bullet \vec{d2}}{\left|\vec{d1}\right|\left|\vec{d2}\right|}$$

where $\cos(\vec{d1}, \vec{d2})$ is the cosine similarity value between document d1 and d2.

4. PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

Pearson's product moment correlation coefficient or Pearson's *r* is used to calculate the correlation between two interval or ratio variables. Pearson's is used to know the correlation between two variables under three conditions. First, both variables must measure intervals or ratios (i.e. attitude scales, test scores). Second, the relationship between the two variables must be linear - the data points should generally fall along a straight line. The non-linear relationship between variables yields Pearson's *r* close to zero. The third condition is that both variables are normally distributed. An oblique distribution produces *r* smaller than the normal distribution [24].

Pearson's used in this study to calculate the correlation between actual test scores and the score given by the system. The correlation values are in between -1 and 1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. The higher the correlation, the better the accuracy of the system. Pearson's formula as shown in this equation.

$$r_{xy} = \frac{\sum XY - \sum X \sum Y}{\sqrt{\left[\left(\sum X^2\right) - \left(\sum X\right)^2\right]\left[\left(\sum Y^2\right) - \left(\sum Y\right)^2\right]}}$$

where

r_{xy} : Pearson's Product Moment Correlation Coefficient

X : Population data score value from automatic assessment

Y : Population data score value from manual assessment

5. RESEARCH METHOD

The dataset used in this research is essay examination of Computer Graphic course. The dataset contains the answer from lecturer, which is called ground truth, and the answers from students. In the first step, preprocessing (case folding, tokenizing, filtering, and stemming) was conducted to the answer from lecturer and the answer from student. The next step is weighting using TF-IDF and measurement of similarity between the answers from lecturer and each answers entered by students by using N-gram and Cosine Similarity. The cosine value from this process will be the test score of each students. There are 2 kinds of dataset that have been used in this study, Datasets 1 and Dataset 2. Dataset 1 contains questions with answers that do not quite consider the order of words in the answer, while Dataset 2 contains questions with answers that pay attention to it.

In the experiment, several variations of N-gram was used including unigram, bigram, trigram and combination of all three. The cosine value on each variation automatically assigned as the test score of the students. In the evaluation step, correlation measurement was conducted between the score given by system and the score given by given by the lecturer.

6. RESULT AND ANALYSIS

The dataset used in this research is essay examination of Computer Graphic course. There are 2 kinds of dataset that have been used in this study, Datasets 1 and Dataset 2. Dataset 1 contains questions with answers that do not quite consider the order of words in the answer, while Dataset 2 contains questions with answers that pay attention to it.

Table 1 Experiment Result

N-Gram	Correlation value on *dataset* 1	Correlation value on *dataset* 2
Unigram	**0.66**	0.64
Bigram	0.57	0.67
Trigram	0.41	0.59
Combination	0.62	**0.68**

The experimental results are shown in Table 1. In general, the system shows quite good results. Regarding the N-Gram variations used, unigram has the best correlation value in Dataset 1 whose answer to the questions does not pay attention to word order. The correlation value of unigram is better in Dataset 1 because the answer of the questions do not consider the order of words. Meanwhile in Dataset 2, bigram and combination have better correlation value by 0.67 and 0.68 because Dataset 2 consider the order of words in the answer. However, in some cases the n-gram and cosine similarity systems have the disadvantage of cannot detect synonyms.

7. CONCLUSION

Based on the experiment results and analysis conducted to the system performance, we can concluded that the application of the N-gram and Cosine similarity method on the automatic essay examination scoring gives a pretty good result. Unigram in the system yields the best correlation value by 0.66 for questions that do not take into account the order of words in the answer. Meanwhile the combination of n-grams yields the best correlation value by 0.67 for questions that consider the order of words in the answer.

8. REFERENCES

[1] Rossen, E. and Hartley, D., 2001. Basics of e-learning (Vol. 109). *American Society for Training and Development*.

[2] E. Sakti, "Rancang Bangun E-learning Dengan Konsep Gamification," 2014.

[3] Palmer, J., Williams, R. and Dreher, H., 2002. Automated essay grading systems applied to a first year university subject: how can we do it better?. In *IS2002 informing science and IT education conference* (pp. 1221-1229). Informing Science Institute.

[4] Valenti, S., Neri, F. and Cucchiarelli, A., 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*(1), pp.319-330.

[5] Hearst, M.A., 2000. The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, *15*(5), pp.22-37.

[6] Jerrams-Smith, J., Soh, V., & Callear D. 2001. Bridging gaps in computerized assessment of texts. *Proceedings of the International Conference on Advanced Learning Technologies*, 139-140, IEEE.

[7] Laham, D. & Foltz, P. W. 2000. The intelligent essay assessor. In T.K. Landauer (Ed.), *IEEE Intelligent Systems*, 2000.

[8] Burstein, J., Kukich, K., Wolff, S., Chi, L., & Chodorow M. (1998). Enriching automated essay scoring using discourse marking. *Proceedings of the Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada.

[9] Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essay and short answers. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

[10] Ming, P.Y., Mikhailov, A.A., & Kuan, T.L. (2000). Intelligent essay marking system. In C. Cheers (Ed.), Learners Together, Feb. 2000, NgeeANN Polytechnic, Singapore.

[11] Christie, J. R. (1999). Automated essay marking-for both style and content. In M. Danson (Ed.), *Proceedings of the Third Annual Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

[12] Mitchell, T., Russel, T., Broomhead, P., & Aldridge N. (2002). Towards robust computerized marking of free-text responses. In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughboroug University, Loughborouh, UK

[13] Abel Teklemariam Mengistu and Sebsibe Hailemariam Dadi. 2012. *Automatic Essay Scoring: Design and Implementation of Automatic Amharic Essay Scoring System Using Latent Semantic Analysis*. LAP Lambert Academic Publishing, , Germany.

[14] Kakkonen, T., Myller, N., Timonen, J. and Sutinen, E., 2005, June. Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 29-36). Association for Computational Linguistics.

[15] Leah S. Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '98). ACM, New York, NY, USA, 90-95. DOI=http://dx.doi.org/10.1145/290941.290965

[16] Yali Li and Yonghong Yan. 2012. An Effective Automated Essay Scoring System Using Support Vector Regression. In *Proceedings of the 2012 Fifth International Conference on Intelligent Computation Technology and Automation* (ICICTA '12). IEEE Computer Society, Washington, DC, USA, 65-68. DOI=http://dx.doi.org/10.1109/ICICTA.2012.23

[17] Krisnanda, B. P. 2005. Sistem penilaian essay otomatis dengan menggunakan metode LSA. Fakultas Teknik, Universitas Indonesia.

[18] Hermawandi, D. 2008. *Implementasi Skema Pembobotan Sicbi Pada Aplikasi Essay Grading Metode Latent Semantic Analisis*. Fakultas Teknik, Universitas Indonesia.

[19] Thada, V. and Jaglan, V., 2013. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, *2*(4), pp.202-205.

[20] Pramukantoro, E.S. and Fauzi, M.A., 2016, October. Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification. In *Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on* (pp. 149-155). IEEE.

[21] Cavnar, W.B. and Trenkle, J.M., 1994. N-gram-based text categorization. *Ann Arbor MI*, *48113*(2), pp.161-175.

[22] Chang, F. 2005. *Comparing different approaches for using n-grams in syntax acquisition*. NTT Communication Sciences.

[23] Nikos Malandrakis, Elias Iosif, and Alexandros Potamianos. 2012. DeepPurple: estimating sentence semantic similarity using n-gram regression models and web snippets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (SemEval '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 565-570.

[24] Yount, W.R., 2006. Research Design and Statistical Analysis in Christian Ministry. USA.