International Journal of Trend in Research and Development, Volume 3(6), ISSN: 2394-9333
www.ijtrd.com

An Automatic Scoring System for Essay by Using Methods Combination of Term Frequency and n-Gram

[1]Munir, [2]Lala Septem Riza and [3]Asep Mulyadi,
[1,2,3]Program Studi Ilmu Komputer, Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam,
Universitas Pendidikan, Indonesia

***Abstract*—**Based on the type, examination is divided into two types, i.e. multiple choices and essay. Multiple choices are often used in the national examination, since the judgement will not be difficult and time consuming. Yet, this kind of exam has weaknesses such as difficult to measure students' understanding, enables speculation, no analysis, etc especially for certain materials. Essay demands a better understanding as well as used to find out students' ability so that the depth level of understanding can be measured. The same as multiple choices, essay has weaknesses also, for example the difficulty of assessing the answers and time consuming. Hence, the research aims to make an automatic scoring system for essay by using methods combination of Term Frequency and n-Gram, in which wording becomes its highlight. In the process of extracting input, the research will use techniques of case folding, stopword, and stemming. The result of the process in the terms of unigram, bigram and unigram+bigram will be tested, in order to obtain numbers in scoring students' answers. To measure the scoring, the research uses Cosine and Jaccard Similarity. The testing accuracy is conducted to calculate mean absolute error and person correlation coefficient results. The result shows that using Jaccard Similarity for Indonesian language is more suitable instead of Cosine Similarity.

Keywords*—Automatic Scoring Sytem for Essay; n-Gram; Cosine Similarity; Jaccard Distance*

I. INTRODUCTION

Evaluation is a process of taking decision by using information from the measurement of learning outcomes, either using test or non-test instrument [13]. There are 2 types of test, i.e subjective and objective test. Subjective test is generally in the form of essay. The test in the form of essay is a kind of test of the advancement learning which needs explanation or description as the answer. The characteristics of question are preceded by words such as describe, explain, why, how, compare, conclude, and so on. While the objective test is a test that has true or false answers and so can be marked objectively. Objective test includes true/false answers, multiple choice, multiple-response and matching questions as well as completion tests. Essay is commonly used to find out student's ability since he/she is able to express and develop answers based on his/her understanding.

There are several weaknesses of essay, namely difficult in assessing the answers and time consuming, since they are various. Besides, subjectivity will be probably occurred during scoring. To obtain an excellent educational process, an excellent scoring system will be needed indeed in order the subject absorption by student measured definitely.

Along with the technology development, many researchers have been developed a system to score evaluation automatically. It aims to increase effectiveness in scoring so that subjectivity can be reduced. However, the automatic scoring system is available for multiple choices only; for essay it is still developed and cannot be applied yet.

Many claimed that the subjective characteristics of an essay created differences in value given by different people, which considered as injustice by students. Based on computing technology development, the problem could be resolved by using automatic scoring for essay so that manpower could be reduced and scoring objectivity increased.

In 2008, a research toward scoring system for essay was conducted by using K-Nearest Neighbour (KKN) algorithm with the accuracy of 76% [1]. While in 2010, a research toward Chinese essays used several combinations of Vector Space Model (VSM), namely Latent Semantic-based Vector Space Model (LS-VSM), Sequence Latent Semantic-based Vector Space Model (SLS-VSM), Word-based Vector Space Model (W-VSM), and Weight Adapted Word-based Vector Space Model (WAW-VSM)[2]. VSM is a model uses to measure similarity of a document with a query.

In 2011, Probabilistic Latent Semantic Analysis (PLSA) method and Cosine Similarity implemented to count similarity on text of an essay. This PLSA method is an improvement of LSA [3]. Either PLSA or LSA has not been able to notice wording. Then in 2014, Latent Semantic Analysis (LSA) and Support Vector Machine (SVM) methods were implemented to count similarity on text of an essay [4]. LSA is a method to determine a relationship between a set of documents and the terms they contain by analyzing the larger text corpus. The research obtained accuracy of 88.8%.

Still in the same year (2014), another method namely Regularized Latent Semantic Indexing (RLSI) was used toward Chinese language and obtained accuracy of 89% [5]. Besides the methods used in the previous research, an implementation of Winnowing Algorithm and Jaccard Similarity was conducted in 2014 [6]. Winnowing Algorithm is a matching string algorithm. The accuracy obtained was 75-80%.

The research continues. In 2015, a research using n-Gram method and Cosine Similarity have been applied to English essays [7]. n-Gram is a contiguous sequence of n-items from a given sequence of text or speech. A result was obtained that unigram+bigram was better than unigram and/or bigram, with 7.5872 as mean absolute error and 0.2843 as person correlation coefficient, which showed low positive correlation. The numbers of mean absolute error show the length between similarity and lecturer's value. Thus, the smaller the value of mean absolute error, the closer the similarity toward lecturer's value is. Person correlation coefficient shows a relationship status between similarity and lecturer's value.

Based on the above literature study, the researcher will implement methods combination of Term Frequency and n-Gram by applying several innovations into Indonesian language. Term Frequency (TF) is the number of times a term (t) occurs in a document (d) and used to calculate words appearance in a sentence. Besides both methods, the researcher tried to compare Cosine Similarity and Jaccard Similarity. Cosine Similarity is often used to give a useful measure of how

similar two documents are likely to be in terms of their subject; while Jaccard Similarity is a statistic used for comparing the similarity and diversity of sample sets. Through the research a comparison between unigram, bigram and unigram+bigram are created.

II. THEORETICAL FOUNDATION

A. Information Retrieval (IR)

Information Retrieval (IR) is the activity of obtaining materials (document), which are unstructured (text), acts as information in a large scale and oftenly stored in a computer [8]. How the Information Retrieval (IR) works explained below:

1. Tokenization or word token is the process of breaking a set of words in a sentence, a paragraph into termed word, removing characters in punctuation as well as modifying termed word into lower case, for example "I learn Information Retrieval", which will result: "I", "learn", "information", and "retrieval".
2. Stopword removal or filtration is words which are filtered out in a document. It is used to describe the content, and distinguish document content from another one. Stopword removal is also used to sort the terms by collection frequency (the total number of times each term appears in the document selection), for example: words such as and, or, not, etc which are frequently appeared.
3. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form, which usually exists in the similar document or word (synonym); for example: if the word "find" input, then the query will recommend "discover", "detect", "notice", "encounter" and so on.
4. Term weighting is a procedure that takes place during the text indexing process in order to assess the value of each term to the document. Term weighting is divided into local, global and normalization.

B. Pre-processing

Pre-processing is an early stage in processing input data before entering the primary stage. Pre-processing aims to equate and facilitate reading. The processes are:

a. Case Folding

Case folding is a process of converting all characters in a document into the same case, either all upper case or lower case, in order to speed up comparisons during an indexing process.

b. Stopword

Stopword is the process of filtering words that frequently appeared; yet do not give a significant information [9]. Thus, stopword reduces words volume. Stopword can be in the form of preposition, conjunction, and substitue. Yet, stopword depends on the type of classification and data collected [10]. Words such as "and", "from", "then", "which", "to", "in/at/on", "I", "you", "he/she", etc are eliminated since they are time consuming and confusing to the document search by user.

c. Stemming

Stemming is a process of reversing derived words into their word stem, base or root form, for example: "finish off" into "finish", "wash up" into "wash", etc. Yet, the algorithm stemming for Indonesian language is different than any other language. English has a different morphology than Indonesian. On English text, the process conducted only eliminating suffix, while on Indonesian text, suffix, prefix and confix should also be eliminated. The research uses the stemming algorithm of Nazief and Andriani. Augusta stated that the accuracy by using Nazief and Andriani Algorithm is higher than Porter Algorithm [11]. During this process, the dictionary used really affects the stemming result. The complete the dictionary used, the more accurate the stemming result is. Hence, the stemming performance is also various depends on language domain used.

C. Methodology

The research uses methods combination namely Term Frequency and n-Gram.

a. Term Frequency

Term Frequency (TF) is a heuristic weighting algorithm which determines a document weighting based on term appearance. The more a term appears, the higher the document weighting is and vice versa.. There are several formulations can be used in TF such as Binery TF, Raw TF, Logarithmic TF and Normalization TF.

b. n-Gram

n-Gram is a sequence of n-word, such as 2-gram (bigram) for example: "please turn", "turn your", or "your homework", and 3-gram (trigram) for example: "please turn your" or "turn your homework" [12]. Unigram is the simplest model of n-Gram, consists of 1 word. The sentence "kedaulatan rakyat adalah suatu kekuasaan pemerintahan yang sepenuhnya ada di tangan rakyat", then the unigram will be "kedaulatan", "rakyat", "adalah", "suatu", "kekuasaan", "pemerintahan", "yang", "sepenuhnya", "ada", "di", "tangan", and "rakyat". An n-Gram of size 2 is reffered to bigram, for example: "kedaulatan rakyat", "rakyat adalah", "adalah suatu", "suatu kekuasaan", "kekuasaan pemerintahan", "pemerintahan yang", "yang sepenuhnya", "sepenuhnya ada", "ada di", "di tangan", and "tangan rakyat". To analyze the effectiveness of n-Gram in an automatic scoring system for essay, then unigram and bigram are combined into unigram+bigram, for example: "kedaulatan rakyat adalah", "rakyat adalah suatu", "adalah suatu kekuasaan", "suatu kekuasaan pemerintahan", "kekuasaan pemerintahan yang", "pemerintahan yang sepenuhnya", "yang sepenuhnya ada", "sepenuhnya ada di", and "di tangan rakyat".

D. Similarity Calculation

Cosine Similarity and Jaccard Similarity are two very common measurements while comparing item similarities. The results obtained from the methods usage are the answer key vector and the student answer vector. The next stage will be calculating the similarity of both vectors by using Cosine Similarity and Jaccard Similarity.

a. Cosine Similarity

Cosine Similarity is a measure of similarity between two non zero vectors of an inner product space that measures the cosine of the angle between them.

$$\text{CosSim}(d,q) = \frac{\sum_{i=1}^{t}(d_i * q_i)}{\sqrt{\sum_{i=1}^{t} d_i^2 * \sum_{i=1}^{t}}} \quad (1)$$

Description:
d = the student answer vector
q = the answer key vector
t = the component numbers of vector d (the same as the component numbers of vector$q$)

d_i = the component of vector d to i (student answer)
q_i = the component of vector q to i (the answer key)

b. Jaccard Similarity

Jaccard Similarity is a statistic used for comparing the similarity and diversity of sample sets, represented by X and Y. X represents the answer key, while Y represents the student answer. The equation will be:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (2)$$

The equation 2 can be simplified into,

$$J(X,\Psi) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \quad (3)$$

Description:
X = the answer key vector
Y = the student answer

III. THE RESEARCH RESULT AND DISCUSSION

A. Model Development

Models developed in this scoring system include pre-processing, n-Gram calculation, and similarity. During the scoring process, user should input question and answer in the form of csv file with the structure provided by including the numbers of quenstions and students. Pre-processing will then be conducted to be futher processed by n-Gram. The output of n-Gram process are the vector of answer key and student answer which will be calculated by using Similarity namely Cosine Similarity and Jaccard Similarity. The last but not least, the data resulted from Similarity calculation will be put in the form xls file.

B. Experiment Design

Testing toward the system is conducted through 2 scenarios, namely:

1. Scenario 1, conducted by using Cosine Similarity

2. Scenario 2, conducted by using Jaccard Distance

Each scenario conducts running system 3 times (see TABLE I).

Table 1: Experiment Design

No	n-Gram	Similarity
1.	Unigram	Cosine Similarity
2.	Bigram	Cosine Similarity
3.	Unigram+Bigram	Cosine Similarity
4.	Unigram	Jaccard Similarity
5.	Bigram	Jaccard Similarity
6.	Unigram+Bigram	Jaccard Similarity

From the result, the accuracy between data using Cosine Similarity and Jaccard Similarity can be analysed.

C. The Experiment Result

a. Scenario 1 Testing

On scenario 1 testing, Cosine Similarity was conducted on input data as much as 3 times running system, by using unigram, bigram and unigram+bigram (see TABLE II).

Table 2: Testing Result Of Scenario 1

No	NIS	Cosine Similarity			
		Lecture rValue	Unigram	Bigram	Unigram+ Bigram
1	02	20	16,95	5,94	2,58
2	02	20	17,22	2,67	0,00
3	02	20	18,53	12,56	1,71
4	02	15	14,33	2,36	0,56
5	02	20	17,65	4,78	0,00
..	…	…	...	…	…
215	46	20	11,79	2,67	0,00

b. Scenario 2 Testing

Jaccard Similarity was conducted on scenario 2 testing. It was conducted to Class B by also running system 3 times as scenario 1 did. The result shown by TABLE III.

Table 3: Scenario Testing Result

Problem Code	NIS	Jaccard Similarity			
		Lecturer Value	Unigram	Bigram	Unigram+ Bigram
1	02	20	15,00	13,33	20,00
2	02	20	12,00	10,00	13,33
3	02	20	8,57	10,00	15,00
4	02	15	16,00	15,00	10,00
5	02	20	6,67	20,00	20,00
..	…	…	...	…	…
215	46	20	20,00	20,00	20,00

D. Analysis

The experiment was already conducted and the data was obtained. The next step will be analysing mean absolute error and person correlation coefficient result. The result of mean absolute error is the average difference between value from lecturer and system.

$$MAE(\bar{x}) = \frac{\sum_{i=1}^{n} |Xi - Yi|}{n} \quad (4)$$

Description:
$\bar{x}$ = average value from lecturer and system
Xi = value from lecturer
Yi = value from system
n = numbers of data

Person Correlation Coefficient Results is conducted to find out realationship degrees between dependent and independent variable. The dependent variable in this case is the value given by lecturer and denoted as X, while independent variable is the value given by system and denoted as Y.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{\sum X^2}{N} \sum Y^2 - \frac{\sum Y^2}{N}}} \quad (5)$$

From the result, TABLE IV shows the explanation of r value.

Tabel 4: Person Correlation Coefficient Results

Range	Description
0.5 s.d 1	high positive correlation
0 s.d 0.49	low positive correlation
-0.5 s.d 0	weak negative correlation
-0.49 s.d -1	strong negative correlation

Based on TABLE IV, the r value obtained ranges from 0 to 1 are categorized as good though range from 0 to 0.49 is still classified as low. If the r value ranges from 0 to -1, the result is

calssified as less good and even will be classified as poor if the value is between -0.49 to -1.

Based on the equation 4 and 5, for scenario 1 the value of mean absolute error and person correlation coefficient results are obtained (see Picture 1).

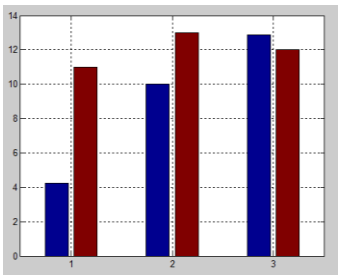

Figure 1: Mean Absolute Error and Person Correlation Results of Cosine Similarity

Description:

1 = Unigram

2 = Bigram

3 = Unigram+Bigram

= Person Correlation Coefficient Results

= Mean Absolute Error

Based on Picture 1, mean absolute error shows the result that unigram is better than bigram and unigram+bigram with the error value of 4.23. Yet, there is an increasing error from unigram to unigram+bigram. While person correlation coefficient results show that either unigram, bigram or unigram+bigram has low positive correlation with the sequence value 0.11/11, 0.13/13 and 0.12/12. In scenario 2, mean absolute error and person correlation coefficient results can be seen in Picture 2.

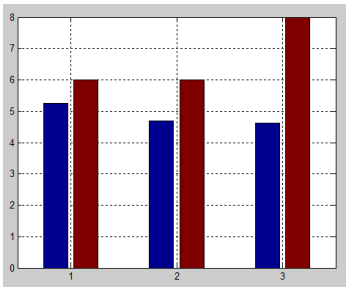

Figure 2: Mean Absolute Error and Person Correlation Results of Jaccard Similarity

Decsription:

1 = Unigram

2 = Bigram

3 = Unigram+Bigram

= Person Correlation Coefficient Results

= Mean Absolute Error

Picture 2 presents mean absolute error of 4.23 only and shows unigram+bigram is better than unigram or bigram. Besides, there is a decreasing eror from unigram into unigram+bigram. Person correlation coefficient results shows that either unigram, bigram or unigram+bigram has low positive correlation with the sequence value 0.06/6, 0.06/6, 0.08/8.

E. Analysis and Disscusion

The analysis has been conducted, therefore a table of comparison to facilitate conclusion withdrawing. The result comparison between scenario 1 and 2 is presented in TABLE V.

Table 5: The Comparison Ofmean Absolute Error And Person Correlation Coefficient Results

	Unigram		Bigram		Unigram+ Bigram	
	Cos.	Jac.	Cos.	Jac.	Cos.	Jac.
Mean Absolute Error	4,23	5,26	9,99	4,70	12,89	4,63
Person Correlation Coefficient Result	0,11	0,06	0,13	0,06	0,12	0,08

Seen from the TABLE V, a difference in results of cosine similarity and jaccard similairity obtained. According to Cosine Similarity, unigram is better since the value of mean absolute error is the lowest and classified as low positive correlation; while according to Jaccard Similarity, unigram+bigram is better since the value of mean absolute error is the lowest and classified as low positive correlation. There are several things which affect the difference such as:

1. The answer data is too short, so that will weaken the theory of bigram and unigram+bigram. If the pre-processing resulted 1 word, then it does not applied to either bigram or unigram+bigram which need 2 and 3 words. If the pre-processing resulted 2 words then it does not applied into unigram+bigram which need 3 words, yet it is applicable into unigram.
2. There is an answer in the list form, so that will need wording. Thus, it weakens the theory of bigram and unigram+bigram since list does not concern wording.
3. The similarity usage can affect the result.

CONCLUSIONS

The research obtained several results, which are:

1. The methods combination of Term Frequency dan n-Gram have been implemented. Yet, there are several factors that affect the result, namely stopwords, stemming, similarity, and data used.
2. The accuracy level of the system is calculated by using mean absolute error and person correlation coefficient results. By using Cosine Similarity, Mean Absolute Error is obtained and the sequence of unigram, bigram, and unigram+bigram are 4.23, 9.99, dan 12.89. As for Person Correlation Coefficient Results are 0.11, 0.13, dan 0.12. By using Jaccard Similarity, the Mean Absolute Error for unigram, bigram, and unigram+bigram are obtained i.e 5.26, 4.70, dan 4.63; while the Person Correlation Coefficient Results are 0.06, 0.06, dan 0.08.
3. The usage of similarity turns out affecting the scoring result; proven by Unigram of Cosine Similarity which is better than Unigram+Bigram of Jaccard Similarity.

References

[1] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, “Automated essay scoring using the KNN algorithm,” Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008, vol. 1, pp. 735–738, 2008.

[2] X. Peng, D. Ke, Z. Chen, and B. Xu, "Automated Chinese Essay Scoring Using Vector Space Models," 2010.

[3] Y. Wihardi, "Sistem Penilaian Jawaban Esai Secara Otomatis Menggunakan Metode Probabilistic Latent Semantic Analysis," 2011.

[4] M. Zhang, S. Hao, Y. Xu, D. Ke, and H. Peng, "Automated essay scoring using incremental latent semantic analysis," J. Softw., vol. 9, no. 2, pp. 429–436, 2014.

[5] S. Hao, Y. Xu, H. Peng, K. Su, and D. Ke, "Automated chinese essay scoring from topic perspective using regularized latent semantic indexing," Proc. - Int. Conf. Pattern Recognit., pp. 3092–3097, 2014.

[6] S. Astutik, A. D. Cahyani, and M. K. Sophan, "Sistem Penilaian Esai Otomatis Pada E-Learning Dengan Algoritma Winnowing," vol. 12, no. 2, pp. 47–52, 2014.

[7] O. E. Oduntan, I. A. Adeyanju, S. O. Olabiyisi, and E. O. Omidora, "Evaluation of N-Gram Text Representations for Automated Essay-Type Grading Systems," Int. J. Appl. Inf. Syst. – ISSN, vol. 9, no. 4, pp. 25–31, 2015.

[8] C. D. Manning, P. Raghavan, and H. Schutze, "An Introduction to Information Retrieval," Inf. Retr. Boston., pp. 1–18, 2009.

[9] R. Puri, R. P. S. Bedi, and V. Goyal, "Automated Stopwords Identification in Punjabi Documents," vol. 8, no. June 2013, pp. 119–125, 2013.

[10] E. Rasywir and A. Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin," J. Cybermatika, vol. 3, no. 2, pp. 1–8, 2015.

[11] L. Agusta, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia," Konf. Nas. Sist. dan Inform. 2009, no. KNS&I09–036, pp. 196–201, 2009.

[12] D. Jurafsky and J. H. Martin, Speech And Language Processing: An Introduction to Natural Language Processing , Computational Linguistics, and Speech Recognition. 2007.

[13] Zainul and Nasution, Penilaian Hasil Belajar, Jakarta: Dirjen Dikti, 2001.