Latent Semantic Analysis and Winnowing Algorithm Based Automatic Japanese Short Essay Answer Grading System Comparative Performance

Anak Agung Putri Ratna
Department of Electrical Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
ratna@eng.ui.ac.id

Lea Santiar
Japanese Literature
Faculty of Humanities, Universitas Indonesia
Depok, Indonesia
lealas@ui.ac.id@ui.ac.id

Ihsan Ibrahim
Department of Electrical Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
ihsan.ibrahim15@ui.ac.id

Prima Dewi Purnamasari
Department of Electrical Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
prima.dp@ui.ac.id

Dyah Lalita Luhurkinanti
Department of Electrical Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
dyahluhurkinanti@gmail.com

Adisa Larasati
Department of Electrical Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
adisalarasati@gmail.com

***Abstract*——In this paper, advanced of research on e-learning application for short essay grading system had been conducted. This system was developed based on the needs of Japanese Language study program for short essay examination that required time and focus for finishing those tasks. Human abilities are limited by their energy, so that cognitive assessment objectivity could decrease in line with the elapsed time. Latent Semantic Analysis (LSA) and Winnowing Algorithm are two methods used in developing the automatic short essay answer grading system called SIMPLE-O by Department of Electrical Engineering, Universitas Indonesia. These two algorithms are chosen based on its ability to do semantic analytic without the needs of understanding about the characteristic of its languages. LSA used Singular Value Decomposition (SVD) as its main method, besides Winnowing algorithm is based on fingerprinting. These algorithms are applied into the automatic system to assess the Japanese language exam with close results between them with average accuracy of Winnowing algorithm is only 1.06% lower than LSA that could gain 87.78%. These two algorithms should be suitable for grading short essay answer in Japanese language.**

Keywords—essay grading, e-learning, japanese, latent semantic analysis, winnowing

I. INTRODUCTION

Examination is an important part of education, mainly to assess the student's understanding regarding the topics. There are type of questions that are commonly used in examination, such as multiple choice, short answer, and analysis or argument essay. In short answer, the student is required to provide the right answer in few sentences. The purpose of this type of question is mostly to assess the student's basic knowledge regarding the topic. Compared to short answer, essay answer is longer, usually consists of several paragraphs. Not just the basic understanding, but the student needs to develop analysis or argument, convey it through text, and provide examples that can support the argument.

Automatic grading system has been implemented in many examinations through Computerized Adaptive Test (CAT) with varied purpose. It is also used for e-learning process, though the question type that can be graded automatically is objective type like multiple choice. The pioneer of essay grading systems is Project Essay Grader (PEG) by Ellis Page, introduced in 1966 [1]. Since then, the research in exam grading continue to bring more development in essay grading. Other than PEG, C-rater, E-rater, and Latent Semantic Analysis (LSA) are some other applications developed for automatic essay assessment [2][3]. Automatic essay grading can also be combined with human raters like in Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT) [4].

Although short answer is not as extensive as analysis or argument essay, it is still graded manually by human rater. Giving score to each student's answer will surely takes time. By automating the grading for short answer, the student can get immediate assessment of the test. Since the score is given by system, the result that the student get should be more objective than scored by human. Moreover, the chance of human error at grading process to happen will also decrease as it will be handled by system. With those factors as consideration and to help the process of exam grading to be more efficient., implementing automatic short answer grading is needed.

Based on that problem, Department of Electrical Engineering, Universitas Indonesia had developed the automatic essay grading system to ease the examination assessment process called SIMPLE-O. This system has been developed since 2007 which originally designed for Indonesian language and based on Latent Semantic Analysis [5]. With its role benefitted in the learning environment, Japanese Language study program was interested with this application and also want to implement with its Japanese language examination. On the previous occasions, this system had been developed with many algorithms to perfect it so it could be implemented in real academic environment. This research is developed as the continuation of previous experiments that implement Winnowing algorithm and trying to be implemented in Japanese

language since it also has a unique characteristic if it is compared to Indonesian language [6].

In this research, there are two approaches that are used. The first one is Latent Semantic Analysis (LSA), a corpus-based method which is often used for text similarity measurement [7]. Corpus-based measure similarity between texts from large collection of words called corpora [8]. It does not pay attention to the word sequences or the construction of sentences itself. Another method used for this grading system in Winnowing Algorithm which is based on fingerprint, a string-based method. This research will compare the results of the those two algorithms in real examination to find the most suitable algorithms for Japanese language short essay exam implementation.

This paper is organized into several sections. After this first section, second section described about the concept of LSA and Winnowing that implemented in this research. In the third section, the system's implemented design of the system for each algorithm were explained, starting from pre-processing until the measurement of similarity between texts and system's accuracy were covered there briefly. The result of the research based on the design section presented and analysed in the next section. Then, the final section for concluding the result of this research.

II. Latent Semantic Analysis (LSA) and Winnowing Algorithm on Automatic Short Essay Grading System

In this section is covered the concept of Latent Semantic Analysis (LSA) and Winnowing to understand the processes that contained in the SIMPLE-O automatic short essay grading system.

A. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a method for extracting and representing the use of the meaning of words by using statistical computation to a reference set of text [9]. In LSA, texts are represented into matrix, the representation of text in a matrix with words as lines and sentences in document (or other contexts) as columns is called Term Document Matrix (TDM), in which the line indicates the keyword and the column indicates the sentence in the document. TDM is formed by arranging a list of keywords or unique words from a document.

After forming the matrix, LSA performs a singular value decomposition (SVD) [9], [10] of the matrix. This process will decompose the matrix into 3 matrices, in which SVD is defined as in (1) [7] below.

$$A = USV^T \tag{1}$$

A is the complex matrix with m row and n column, which in this research is the term-document matrix. U is the representation of row entity of matrix A with m row and m column, S is the diagonal matrix of matrix A with m row and n column, and VT is the representation of column entity of matrix A with n row and n column.

In SVD, the diagonal matrix's dimension will be sliced, the matrix from the process will be multiplied with row and column entity of the matrix. The result will have the same dimension as the initial matrix, but with different values from the initial matrix.Slicing the diagonal matrix results in different values for each of the unique words, the value can increase or decrease from the value of the initial matrix before SVD.

B. Winnowing

Unlike LSA, winnowing is an algorithm based on fingerprint. Using fingerprint is common techniques to compare the similarity between texts or documents. The generated fingerprint will be the representation of the text [11]. In winnowing algorithm, fingerprint is a series of numeric values selected from the output of the hashing process.

The winnowing algorithm can solve the problem of fingerprint selection in fingerprint algorithms with the windowing process and the selection of hash values based on the window [11]. This algorithm useful for essay assessment because it is expected to have some characteristics like whitespace intensity, noise suppression, and position independence [11]. This algorithm has text or document as input. This will be processed to generate fingerprint that represent the text as the algorithm's output. The generated fingerprint will be compared with the other fingerprint to find the similarity.

Before processing is done by winnowing algorithms, it first must go through the data pre-processing. It is intended to prepare the input as well as processing the Japanese characters that should be changed to the romanised form, so the results are more accurate [12]. The winnowing algorithm has several processes, such as the determination of the n-gram, hashing, windowing, and fingerprinting. N-gram is a method based on character [13]. The formation of the n-gram is the first step in the implementation of winnowing algorithm. This process makes sequences of words with n characters in each sequence. The bigger value of n will result in longer character in each word. There is a possibility that the essential words will not be noticed. Meanwhile, if the n is minimal, there will be a lot of word sequences. It will be difficult to differentiate similar word [14]. For this reason, the n is appropriate, which is a minimum value that removes the same word. N-gram is needed in the winnowing algorithm as an input to get a word from the text .

After n-gram is formed, each sequence will undergo hashing process. Hashing is a character string transformation technique that represents originals value in winnowing algorithm. It represents the fingerprint of the texts. The values are calculating from ASCII that has determined from N-gram. In this process, we use Rabin-Karp Algorithm. This algorithm was used to solve the problem of string matching. Rolling hash is used to reduce iterations in the calculation of hash values so that they can run more effectively [15]. Mathematically, hashing process is defined as in (2) below [16]:

$$H_{(C_2 \dots C_{K+1})} = \left(H_{(C_1 \dots C_K)} - C_1 * b^{(k-1)}\right) * b + H_{(C_{(k+1)})} \tag{2}$$

Windowing is a process in winnowing algorithm to store the results of rolling hash that have been obtained previously. Hash values will be grouped in windows. The length of this windows is specified as w. The result of this windowing process is a series of hashing values that has been grouped.

After the windows are formed, fingerprint that represent the text will be selected [17]. The smallest value is selected from each window. If the same minimum value is found in the same window, the fingerprint will be the one in the righter window. These selected values are put together into one fingerprint. This fingerprint will be used as a reference for calculating the similarity between the two texts.

C. Similarity Measurement

The last steps to determine the degree of similarity between the two texts is the fingerprint similarity. Input from this step is fingerprint of the two texts that will be compared. There are several methods that can be used to check the similarity between two texts or documents. In this research the methods used are cosine similarity for the winnowing algorithm-based system and Frobenius norm for the LSA algorithm-based system. The mathematic formula for cosine similarity is as shown in (3) below [14]:

$$CS = \frac{|X \cap Y|}{|X|^{1/2}.|Y|^{1/2}} \times 100\% \quad (3)$$

X and Y are the fingerprint of the compared documents. The output value will be ranged from 0 to 1. It will be multiplied by 100% in order to get the similarity percentage. While Frobenius norm [13] is used to form a vector of the diagonal matrix S of the result of decomposition. The mathematic formula for Frobenius norm is as shown in (4) and (5) below:

$$\|A\|_F \equiv \sqrt{\Sigma_{i=1}^{m}\Sigma_{j=1}^{n}\left|a_{ij}\right|^2} \quad (4)$$

$$\|A\|_F \equiv \sqrt{Tr(AA^H)} \quad (5)$$

III. Automatic Japanese Short Essay Grading System Design with Latent Semantic Analysis and Winnowing Processes

On this research, the system will be implemented using LSA and Winnowing algorithm. Each design consists of two major process: pre-processing and main process which is either LSA or Winnowing. The inputs of this system are the text which is the key answer to the problem, and the text is the answer from the students. Both these texts are in the Japanese language. The text can be in the form of katakana, hiragana, or kanji. For the inputs to be well-processed by the main algorithm, there should be some methods to prepare those inputs. This preparation process is called pre-processing.

The designed Winnowing and LSA has different way to handle pre-processing. However, there are three things that are in common in both designs: romanization and filtering. The student answers as well as the answer keys need to be converted into the Romanized form through the process called romanization . Since the students are allowed to answer in any form of Japanese characters (kanji, katakana, or hiragana), romanization will allow those various form to be read as same by the system as long as it is the right alternative form. After the inputs are changed into roman letter, filtering removes punctuation and special character, leaving only letter and number. The example is shown next column:

- Japanese text: 彼は大学生です。
- Romanization: Kare wa daigakuseidesu.
- Filtering: Kare wa daigakuseidesu

The result of pre-processing will be used as the input for main process. The purpose of the main process is to get value that represent the document. The value from student's answer and the answer key will be compared in order to get the student's score.

A. Winnowing System Design

According to Fig. 1 below, in system designed with Winnowing, the first thing that the system do is initiating the essential inputs: student's answers and answer keys. Each document will be processed with the first major process, pre-processing. The pre-processing in Winnowing consists of four processes: romanization, case folding, whitespace removal, and filtering. After the Romanized form is generated, case folding process makes sure that everything is in lower case. Space and new line from the text will be removed during whitespace removal process before it is filtered. Winnowing algorithm makes use the ASCII codes from the text in its process. Upper case and lower case have different ASCII code. Space, punctuation, and special characters also have their own code in ASCII encoding. That is why it is necessary to condition the text beforehand. While Japanese characters are usually not separated by space, whitespace removal is still done just in case in the input, spaces are included. The output from filtering process will be the input for the main process.

After pre-processing, each input will be processed with the main process, Winnowing algorithm. Winnowing algorithm also consists of four processes: n-gram, hashing, windowing, and fingerprinting. The output of this process is fingerprint from student's answer and answer key. The fingerprint from the two texts will be compared to find the similarity using the similarity measurement method. In this case, the measurement that will be used is cosine similarity. The similarity value will be used as the student's score. Fig. 2 below show the pseudocode for whole Winnowing processes.

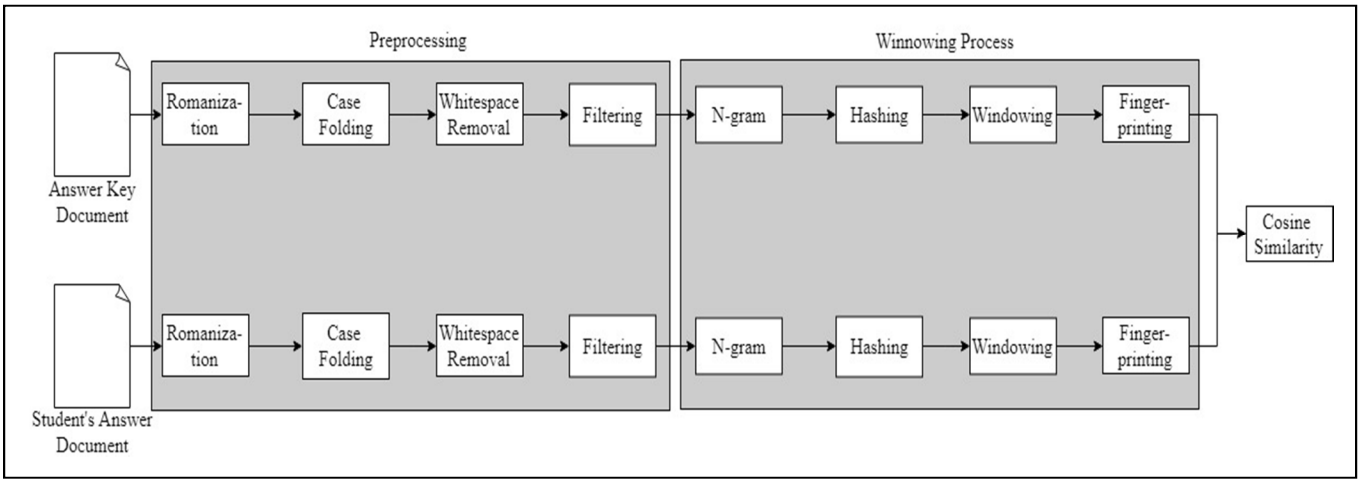

Fig. 1. Block diagram of Winnowing based automatic short essay grading system

```
1.   Input: document
2.   Define:  n_value, w_value
Ngram Process
3.    count = 0
4.    list ngram
5.    doc_length = length of document  //find the
number of characters in document
6.    for 0 <= doc_length – n_value + 1
7.     for document(count) until document(count +
n_value)
8.       append ngram  //add every ngram created to
a ngram list
9.       count = count + 1
Hashing Process
10.    for every sequence in list ngram
11.     for every character in each sequence
12.        convert to ASCII
13.     calculate rolling hash
Windowing Process
14.    count = 0
15.    list window
16.    n_length = length of list ngram
17.    for 0 <= n_length – w_value + 1
18.      for ngram(count) until ngram(count +
w_value)
19.       append window  //add every ngram created
to a window list
20.       count = count + 1
Fingerprinting Process
21.    list fingerprint
22.    for every value in list window
23.     select minimum value
24.     min_value = minimum value
25.                 if min_value window(n) =
window(n-1) AND n >= 1
26.       select min_value from the rightmost window
27.                 append fingerprint
28.
29.   Output: a series of the document's fingerprint
```

Fig. 2. Pseudocode for Winnowing algorithm in automatic short essay grading system

B. Latent Semantic Analysis System Design

According to Fig. 3 below, in system designed with LSA, the system starts by initiating the essential inputs: student's answers and answer keys. Each document will be processed with the first major process, pre-processing. The pre-processing in LSA are slightly different from the one in Winnowing. The pre-processing for answer key consists of four processes while for the student's answer there are three. The pre-processing starts with filtering to remove special characters. After that. The text will be tokenized, segmented, or separated from the sentences into one meaningful character. Then, every tokenized character will be Romanized. For answer key, keywords will be extracted after the text is converted.

LSA does not use ASCII code in its process. Rather than the text as a whole, the words that constructed the sentences are more essential. Unlike English or Indonesian, writings in Japanese are different. In Japanese, the words and sentences are not separated by a whitespace, therefore all of the characters have no clear boundaries between each other. To get the keywords, tokenization or segmentation process is needed before the romanization.

After pre-processing, each input will be processed with the main process which is LSA algorithm. It consists of two processes: the creation of Term-Document Matrix (TDM) and SVD. TDM is created from the keywords that are extracted from pre-processing. The output of LSA is the decomposed matrix from student's answer and answer key. The result from the two texts will be compared to find the similarity using the similarity measurement method. In this case, the measurement that will be used is Frobenius norm. The similarity value will be used as the student's score. Fig. 4 below show the pseudocode for LSA processes.

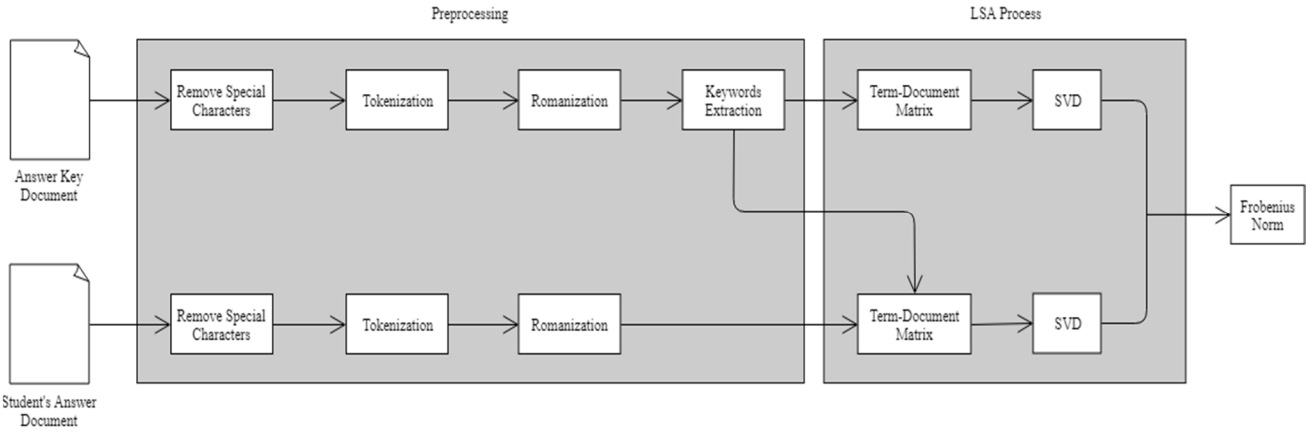

Fig. 3. Block diagram of LSA based automatic short essay grading system

```
1.   Input: document
2.   Define:  n_value, w_value
TDM Process
3.    count = 0
4.    list key_word
5.    for every keys in key_word
6.     for count <= A[keys, count] += 1
7.       append key_word
8.       count = count + 1
SVD Process
9.    singularValues = square root of eigenvalues
10.    S = diagonal of every singular values
11.    S-1 = invert S
12.    V = column matrix of eigenvectors
13.    VT = transpose V
14.    U = AVS-1
15.    A = USVT
FNorm Process
16.    fnormRef = square root of sum of square of
elements of SRef
17.    fnormTest = square root of sum of square of
elements of STest
18.    fnorm = (fnormTest/fnormRef)*100
19.   Output: a series of Frobenius norm values
```

Fig. 4. Pseudocode of LSA algorithm in automatic short essay grading system

C. Accuracy Measurement

In both Winnowing and LSA, the similarity percentage will be used for calculating the system accuracy by comparing it with the score from human rater. The formula used for calculating the accuracy is shown in (6) with a is the score obtained from system and b is the score from human rater.

$$Accuracy = 100 - abs(a - b) \tag{6}$$

IV. Experiments and Results Comparison Between Latent Semantic Analysis and Winnowing Based Systems

A. LSA Experiments

The LSA based system will be tested using 4 scenarios, where in each scenario there are 2 parameters which will be varied for each of the scenario to see the effects of the varied parameters on system accuracy. The TDM counting method will be varied between binary and non-binary, while the form of text input will be varied between Japanese characters and romaji. This experiment will be conducted on the examination answers of 43 students. There will be 5 examination questions for each of them, and for each of the questions there will be different number of answer keys. The total score for the examination of each students will be used for the analysis process.

In the first scenario, the inputs for the LSA process are sentences written in kanji, katakana, and hiragana. the values in the TDM matrix are binary numbers 0 and 1, where 0 indicates that the keyword is not in the document, where 1 indicates that the keyword exists in the document.

In the second scenario, the inputs for the LSA process are sentences written in kanji, katakana, and hiragana. Where the value in the TDM matrix is the frequency of occurrence of a keyword in a document. In the third scenario, the inputs for the LSA process are sentences written in romaji which is the pronunciation of kanji, katakana, and hiragana. the values in the TDM matrix are binary numbers 0 and 1, where 0 indicates that the keyword is not in the document, where 1 indicates that the keyword exists in the document. In the fourth scenario, In the third scenario, the inputs for the LSA process are sentences written in romaji which is the pronunciation of kanji, katakana, and hiragana. Where the value in the TDM matrix is the frequency of occurrence of a keyword in a document.

The accuracy of the result of scenario 1 is in the range of 63.83% to 99.89% with average of 86.43%, while scenario 2 is in the range of 65.20% to 99.93% with average of 87.15%, scenario 3 is in the range of 63.83% to 99.82% with average of 87.26%, and scenario 4 is in the range of 65.20% to 99.93% with average of 87.78%. Scenario 2 and scenario 4 have the same lowest and highest range but different average. This is due to different standard deviation and variation between the two scenarios caused by different text input, scenario 2 uses katakana, hiragana, and kanji, while scenario 4 uses romaji as text input for LSA process. The difference between the assessments on scenario 1, 2, 3, and 4 is shown in Table I below.

TABLE I. ASSESSMENT RESULTS FOR ALL SCENARIOS ON LSA ALGORITHMS

Scenario	Accuracy (%)			σ^2	σ
	Lowest	Highest	$\bar{x}$		
1	63.89	99.89	86.43	71%	8.41%
2	65.20	99.93	87.15	64%	7.98%
3	63.83	99.82	87.26	63%	7.95%
4	65.20	99.93	87.78	61%	7.79%

As shown in Table I, the result between the scenarios are not far different from each other, but the accuracy increases from scenario 1 to 4, while the variance and standard deviation decreases. Which shows that by using a non-binary TDM calculation and using a sentence written in romaji can increase the accuracy of the system. Because by using a sentence written in romaji, the system will still recognize two words with a same meaning but a different way of writing as the same, whether it's written in kanji, hiragana, or katakana, if the words are phonetically the same, then the system will recognize those words as the same words. And as for the TDM calculation, in theory using a binary calculation can prevent giving a high score to a student who uses a word repeatedly, but from the result using a binary calculation decreases the accuracy of the system since it strays further from the human rater score than using a non-binary calculation.

B. Winnowing Experiments

The best parameter combination for SIMPLE-O with winnowing algorithm for essay in Japanese language from the previous conducted research is n=2, p=2, and w=2 [6]. The average accuracy for the total score of the 43 students are 86.86% [6]. This result is obtained with the same or similar environment and machine specification beside the algorithms from the LSA experiments.

It is mostly high when the human rater's score is in the range of 60-80. For human rater's score above 80, the system mostly will score below the human rater and for human rater's score below 60, the system mostly gives scores above human rater. It can be known that the system's score mostly ranged from 60-80. The accuracy for the score outside that range is still lacking, even though it is not too bad. Moreover, if the accuracy is seen from each number and not total score, the accuracy still need improvement.

C. LSA and Winnowing Result Comparison

For finding the most suitable algorithm for Japanese language short answer exam, a performance comparison was conducted. For winnowing, the score used for this comparation is the score from the best parameter combination, n=2, p=2, and w=2. The graph contained the comparation of scores and accuracy from the 43 students' average score is as shown in Fig. 5.

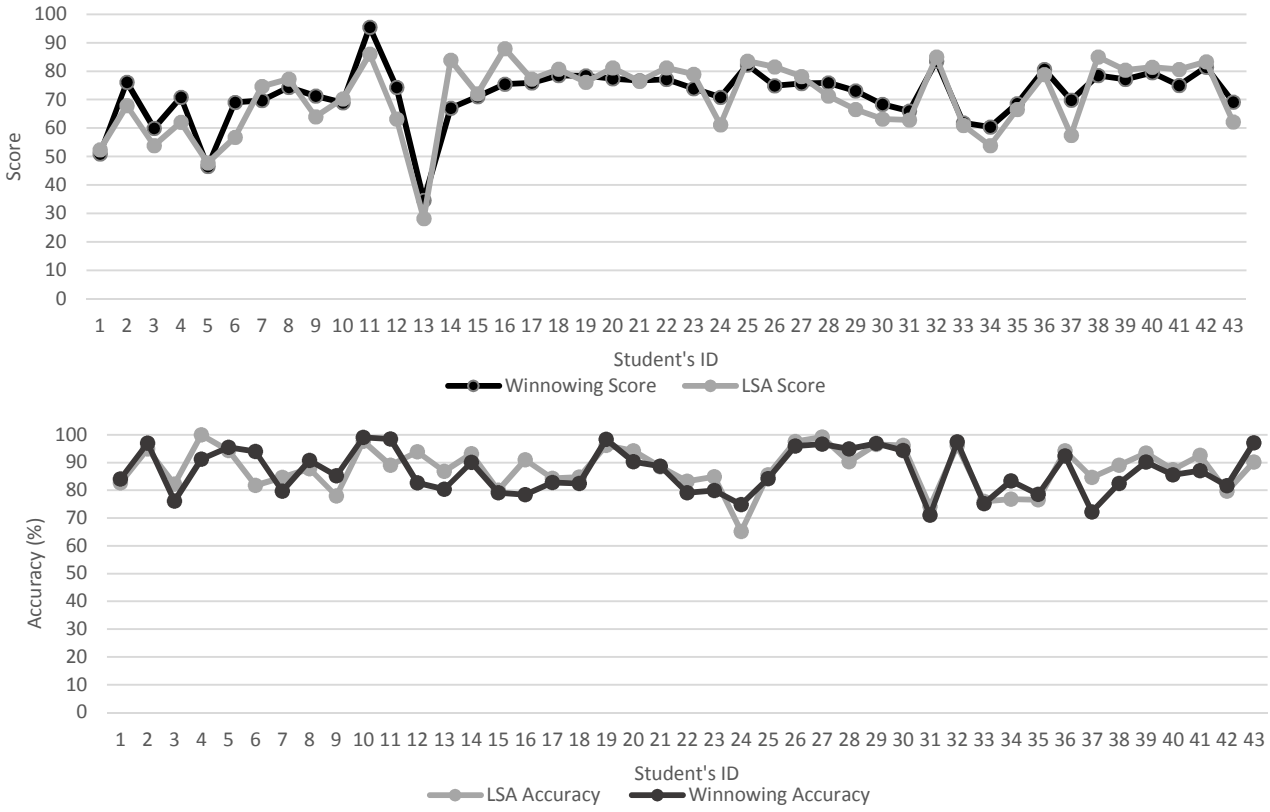

Fig. 5. LSA and Winnowing comparison results on (a) its scores (b) and accuracie

From the graph, it can be known that the system accuracy using winnowing and LSA are not too far off from each other. The lowest accuracy for winnowing is 71.06% and the highest is 99.05%. For LSA, the lowest accuracy is 65.20% and the highest is 99.93%. The average accuracy for winnowing is 86.86% which is close to LSA that have accuracy of 87.78%. These two results are close in accuracy and it could be said that Winnowing is also suitable for exam evaluation.

V. CONCLUSIONS

According to the experiments and analysis conducted for Japanese language short answer grading, there are several conclusions that can be obtained. For system with LSA, non-binary term-document matrix calculation results in higher system accuracy than binary calculation. Processed sentence in romaji results in higher system accuracy than processed sentence in kanji, katakana, and hiragana. By adjusting the parameters, the system can reach an average accuracy of 87.78%, with non-binary term-document calculation and processed sentence in romaji. While LSA and winnowing have closely similar accuracy result, the system developed in winnowing is 1.06% lower than LSA even with use of the best parameters. With their accuracy, it can be said that both LSA and winnowing are suitable to be used as method for grading short answer in Japanese language.

ACKNOWLEDGMENT

This research is fully supported and funded by Universitas Indonesia and Ministry of Research and Technology and Higher Education of Republic of Indonesia under the grant of PITTA (Publikasi Terindeks Internasional untuk Tugas Akhir Mahasiswa UI) B with contract number NKB-0702/UN2.R3.1/HKP.05.00/2019.

REFERENCES

[1] E. B. Page, "The Imminence of... Grading Essays by Computer," *Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966.

[2] L. Rudner and P. Gagne, "An Overview of Three Approaches to Scoring Written Essays by Computer." [Online]. Available: https://pareonline.net/htm/v7n26.htm. [Accessed: 03-Aug-2019].

[3] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 975–8887, 2013.

[4] S. Dikli, *An* Overview *of Automated Scoring of Essays*, vol. 5, no. 1. Technology and Assessment Study Collaborative, Caroline A. and Peter S. Lynch School of Education, Boston, College, 2006.

[5] A. A. P. Ratna and B. Budiardjo, "Simple: Automatic Essay Grading System for Exam Assessment in Indonesian Language (Simple: sistim penilai esei otomatis untuk menilai ujian dalam bahasa indonesia)," *Makara, Teknol.*, vol. 11, no. 1, pp. 5–11, 2007.

[6] A. A. P. Ratna, D. L. Luhurkinanti, I. Ibrahim, D. Husna, and P. D. Purnamasari, "Automatic Essay Grading System for Japanese Language Examination Using Winnowing Algorithm," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 565–569.

[7] T. K. Landauer, P. W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," in *Discourse Processes*, 1998, vol. 25, pp. 259–284.

[8] A. Islam and D. Inkpen, "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity," *TKDD*, vol. 2, Jul. 2008.

[9] A. K. Cline and I. S. Dhillon, "Computation of the Singular Value Decomposition," in Handbook *of Linear Algebra*, Second Edition., Leslie Hogben, Ed. Boca Raton: Chapman & Hall/CRC, 2006, pp. 45.1-45.13.

[10] T. K. Landauer, D. Laham, and P. Foltz, "Learning human-like knowledge by singular value decomposition: A progress report," *Adv. Neural Inf. Process. Syst.*, pp. 45–51, 1998.

[11] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting," *Proc. 2003 ACM SIGMOD Int. Conf. Manag.* data - *SIGMOD '03*, no. January 2003, pp. 76–85, 2003.

[12] W. Hadamitzky, "Romanization systems," 2003. [Online]. Available: https://www.hadamitzky.de/english/lp_romanization_sys.htm. [Accessed: 03-Aug-2018].

[13] A. Wibowo, K. W. Sudarmadi, and A. Barmawi, *Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents*. 2013.

[14] T. Khuat, N. Duc Hung, and L. Thi My Hanh, "A Comparison of Algorithms used to measure the Similarity between two documents," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 4, pp. 1117–1121, Apr. 2015.

[15] R. Sutoyo *et al.*, "Detecting documents plagiarism using winnowing algorithm and k-gram method," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 2017, pp. 67–72.

[16] D. Purwitasari, I. W. S. Priantara, P. Y. Kusmawan, U. Yuhana, and D. Siahaan, "The use of Hartigan index for initializing K-means++ in detecting similar texts of clustered documents as a plagiarism indicator," vol. 10, pp. 341–347, Jan. 2011.

[17] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting," in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 2003, pp. 76–85.