Latent Semantic Analysis Based Automatic Cross-Language Plagiarism Detector for Paragraph Written in Two Syntactically Distinct Languages

Anak Agung Putri Ratna,Universitas Indonesia, Indonesia
Emily Lomempow, Universitas Indonesia, Indonesia
Prima Dewi Purnamasari, Universitas Indonesia, Indonesia
Untung Yuwono, Universitas Indonesia, Indonesia
Boma Anantasatya Adhi, Universitas Indonesia, Indonesia

The Third Asian Conference on Society, Education & Technology 2015
Official Conference Proceedings

Abstract
The number of scientific publication in Bahasa Indonesia is now in steady rise. As a speaker of under-resourced language, Indonesian author often consult documentation in other language, especially English. The necessity for an automated cross-language plagiarism checker has now become prominent. There are several methods available for an automated cross-language plagiarism detection but, most of them only works well on syntactically similar language. Unfortunately both Bahasa Indonesia and English come from a very different language family, therefore they have completely different syntax. This paper investigates the possibility of expanding the use of Latent Semantic Analysis (LSA) for an automated cross-language plagiarism checker between two syntactically distinct languages. LSA's bag of word concept is exploited, removing the necessity to use grammatically correct automatic translator. Several modifications to the LSA algorithm are also proposed to improve its performance. The proposed a proof of concept algorithm is capable to find similarities between a paragraph and its exact translation written in different languages. The exact translation of a paragraph can be identified with 81.82% up to 90.91% accuracy in all test cases.

Keywords: latent semantic analysis; plagiarism detection

Introduction

The convenience provided by high level technology not only affect life in a positive way, but also in a negative way. Wide range of information is available on Internet which is practically accessible by everyone. Scholars are able to get the works of other people from the Internet and then by doing a few of editing, then the work can be submitted as their own. One renowned form of editing is translating the work into another language, which belongs as an act of plagiarism. While plagiarism detection can be conducted manually, it is not efficient as it takes a long time and effort to do. An automation for plagiarism detection which is also called computer assisted detection due to it uses of a certain software on a computer (El Tahir Ali, Dahwa Abdulla, & Snášel, 2011). Furthermore, the automated plagiarism detection system is required to be able to detect cross language similarity in order to detect cross language plagiarism.

The system covered in this paper was developed to be a semantic-based computer assisted plagiarism detection. The semantic-based algorithm used is Latent Semantic Analysis to find the similarity between two documents. The algorithm also have to support translation process between the two languages. Latent semantic analysis, which is a semantic-based method, describes a document as the words it contains and the frequency of occurrence of the words. Therefore, a fully accurate translation process between languages is not needed in the system. The system only has to cover translation per words and overlooks the grammar correctness.

Modified LSA

Latent Semantic Analysis is a technique used to represent words as a statistical computation based by its context in a document. It creates a term-document matrix and apply SVD followed by dimension reduction as a way to predict the relationship between words in a document.

LSA can uses words which appear in more than one document as the matrix rows (Landauer, Foltz, & Laham, 1998). While the column represents each document, so that the matrix element represents frequency of occurrence of the word represented by the row in the document represented by the column. Each element of the matrix can be transformed afterwards by assigning a weight according to the significance of each term in the document.
The next step is to apply SVD to the term-document matrix which decomposed the matrix into three component matrices. The first component matrix represents the original matrix row entities as an orthogonal vector matrix. Second component matrix is a diagonal matrix which contains a scalar value known as singular values. This matrix acts as the multiplying factor such as that if the three component matrices are multiplied, the original matrix will be reconstructed. While the third component matrix is another orthogonal vector matrix which represents the original matrix column entities.

When the singular values are reduced by setting one or more element in the second component matrix element to 0, the multiplication result of the three component matrix will be a different matrix, rather than the original. The new matrix will contain different elements from the original, which means there are changes in frequency of occurrence of the terms. As a matter of fact, a document that originally does not contain a certain term, can have a small frequency of occurrence of that term in the

new matrix. Due to dimension reduction done to the second component matrix, the reconstructing process will predict which term appears in the document based on the similarity that document has with another document. For instances, a document that contain words such as “father”, “mother”, and “son” most likely also has the word “daughter”. Therefore, by comparing it to another document that also has the words “father”, “mother”, “son”, or “daughter”, the system will be able to predict the frequency of occurrence of the certain term.

The vector space in LSA process is constructed of a large corpus of documents. Therefore, it takes relatively long time to process especially if each document contains lots of words. To overcome this problem, at the Electrical Engineering Department at the Faculty of Engineering, University of Indonesia we have developed and implemented a new LSA-based algorithm to find the similarity between two texts.

The new developed LSA algorithm uses term-sentence matrix rather than term-document matrix to separate each document into different vector spaces thus reducing the duration of the overall process. Each document is in the form of paragraph and has its own vector space which is a term-sentence matrix. The terms used in creating the matrix depend on the program function and usually is defined by the user.

To measure the similarity between two documents in the common LSA, we use the cosine between column vectors or the comparison between their lengths (Golub & Van Loan, 1996). Whilst the new developed algorithm uses frobenius normalization of the second component matrix as the vector, so that to measure the similarity between two documents we use the cosine or the comparison between the lengths of frobenius normalization of the second matrices. Equation (1) and (2) are used to normalize a matrix using frobenius normalization (Golub & Van Loan, 1996)

$$\|\boldsymbol{A}\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \quad (1)$$

$$\|\boldsymbol{A}\|_F \equiv \sqrt{Tr(\boldsymbol{A}\boldsymbol{A}^H)} \quad (2)$$

The aim of this research was to use a modified LSA-based algorithm to detect the similarity between two forms of a paragraph, one is written in English and another written in Indonesian. The similarity is measured using the cosine and the comparison between the lengths of frobenius normalization of the second matrices. The method was tested on a sample of 11 paragraphs from three different topics. The accuracy of the system was compared between four different scenarios, with the results could go into three categories. The categories of the results are able to detect the similarity between the appropriate paragraphs, able to detect the similarity of topic between the paragraphs, or unable to detect the similarity between appropriate paragraphs. The goal of this research is as a proof of concept that the new modified LSA is capable to detect the similarity between two paragraphs written in different language.

Modified LSA Implementation For Plagiarism Detection Between Indonesian And English Paragraphs

This computer assisted plagiarism detection utilized LSA, which is categorized as a semantic-based method. Latent Semantic Analysis is a vector space information retrieval model, which represents an object as the characteristics it has (Rehurek, 2007). The representation of a document in this context is the words it contains and the frequency of occurrence of each word, with the assumption that this representation

covers substantial information that the original document contains. The order of occurrence of each word is not significant in LSA process, therefore LSA could find the similarity between two documents without being affected by the grammar of each document. Thus the plagiarism detection system between Indonesian and English paragraph was developed using LSA.

LSA is a method to analyze a document uses of words, hence it can be used to detect the similarity between two documents (Landauer & Dumais, 1997). Then the similarity value will also indicate the similarity in topic and use of words between the two documents. In order to be able to compare two paragraphs written in different languages, the system was designed to translate an Indonesian paragraph into English. This is done by using a simple dictionary database to translate the paragraph per words. The translation process is conducted without being affected by grammar correctness.

This system is developed to be a proof of concept that LSA is capable to detect plagiarism between an Indonesian paragraphs that is written in reference to an existing English paragraph. A modified LSA method, which separates the test and reference vector space, is used in this system to simplify overall process.

The first step in this system development was to collect a number of paragraphs written in English about several different topics. This collection of paragraphs was saved in a database to facilitate as an input. Then each paragraph was translated into Indonesian manually, this is done to obtain the Indonesian paragraphs used to test system accuracy. The collection of Indonesian paragraphs was also saved in database for the same reason. A database consists of a collection of English and Indonesian paragraphs were obtained in this development step.

The test paragraphs, which are written in Indonesian, were translated per word into English without regardless of the grammar correctness. A dictionary database, which is obtained from a free translation website called gkamus, is used in this system. The database contained of two tables, one is for translating Indonesian to English, and another to translate English to Indonesian.

The next step was to explode the reference paragraph per word and save them into an array. This array would be used later as the keywords or terms to create term-document matrix in LSA processing.

Each test and reference paragraph was made into term-document matrix separately. This method came from the modified LSA, which is developed to overcome LSA disadvantages in duration and resource to process. This method works by separating test and reference vector space, therefore simplifying SVD process. The similarity between the two paragraphs were obtained by comparing their vector lengths or the angle formed between frobenius norm vectors of the second component matrices from SVD process.

Method

This research used 11 paragraphs from 3 different topics which are written in English as the reference. These paragraphs were translated into Indonesian and the translation results were used as the test paragraphs. The translation process was not exact because it only translated word per word and not paying attention to the grammar. Because the plagiarism detection system is prone to the exact use of words, not to the exact use of grammar. We used a non-commercial English-Indonesian dictionary database to do the translating.

Reference and test paragraphs were made into term-sentence matrices and several modifications are applied in the process to find the most proper algorithm. The accuracy of a modified algorithm was assessed by the comparison value between reference and test paragraph. If a method succeed in detecting the similarity between a reference paragraph and its translation, then the method was an effective one to detect plagiarism. However, if a test paragraph was detected to be similar to a different paragraph from different topic, then the method would be assessed to be less efficient.

The first modification was to remove stop words from LSA process. Stop words are common used words which are not significance to analysis process (Manning & Raghavan, 2009). Including stop words into LSA process will increase the possibility of disrupting the result of the process. Therefore, we developed an algorithm to automatically erase stop words from the paragraph. This process were conducted on English reference paragraph and also Indonesian test paragraph. The algorithm designed to erase stop words from each test and reference paragraph is shown in **Error! Reference source not found.**.

Whilst the next modification was the terms used to create term-sentence matrix. There were two type of terms used in in this research, the first one used words collection coming from reference paragraph as the terms, and the second one used words collection from reference and also test paragraph as the terms. The algorithm designed to use keywords from test and reference paragraphs to create term-document matrices is shown in **Error! Reference source not found.**.

Based on these two modifications, four different algorithms were designed for this system. The four algorithms are:

- Without removing the stop words and the terms used were derived from reference paragraph only.
- Without removing the stop words and the terms used were derived from test and reference paragraphs.
- By removing the stop words and the terms used were derived from reference paragraph only
- By removing the stop words and the terms used were derived from test and reference paragraphs.

Results

The test results were in form of accuracy of the program, which is the percentage of Indonesian paragraph detected to be similar with the related English version. Smaller angle value and nearer length comparison value to 100 indicates higher similarity between two paragraphs. The results of the test came into three different categories, which are:

- Test paragraph is stated to be most similar with the reference paragraph which is the English version of it. This means the test result is accurate.
- Test paragraph is stated to be most similar with a reference paragraph which is not the English version of it, yet the two paragraphs are about the same topic. This means the test result is less accurate, yet still succeed in detecting the similarity in topic between the two paragraphs.
- Test paragraph is stated to be most similar with a reference paragraph which is not the English version of it, and also the two paragraphs are not about the same topic. This means the test result is not accurate.

The test was conducted using 11 paragraphs written in English as the reference paragraphs and their Indonesian translations as the test paragraphs. The similarity values between each test paragraph and 11 reference paragraphs, so that we have the similarity values between every test paragraph and reference paragraph. The 11 paragraphs come from four different topics, which are four paragraphs about organism, three paragraphs about biology, and four paragraphs come from the biography of Albert Einstein.

The results of this tests came in as the similarity value between test and reference paragraphs. The similarity value could be in form of vector lengths comparison in the scale of 100. The data taken from this indicator is in the form of difference with 100, where smaller difference means the two paragraphs are similar. Another form of similarity value is the angle formed by the two vectors with the same length. To equalize the lengths of two vectors, we could slice the matrix which contains more element so that it has the same element with the other, therefore we would obtain the angle between matrices with least element. Or we could also pad the matrix which contains less element so that it has the same element with the other, and we would obtain the angle between matrices with most element.

The first test was conducted without removing the stop words and the terms used were derived from reference paragraph only. The results were in the range of 27.27% to 45.45% succeed in detecting the similarity between the right paragraphs. 18.18% to 36.36% of the test results only succeed in detecting the similarity in the topic covered by the paragraphs. While 27.27% to 36.36% of the test results failed in detecting the similarity between the right paragraphs, yet also failed in detecting the similarity of the topic.

Table 1. Testing Results without Removing Stop Words and Using Terms Derived from Reference Paragraph

Indicator	% Succeed in Detecting Similarity Between Paragraphs	% Succeed in Detecting Similarity of Topics	% Not Succeed in Detecting Similarity
Length Comparison	27,27	36,36	36,36
Angle Between Least Elements	45,45	18,18	36,36
Angle Between Most Elements	36,36	36,36	27,27

Table 1 shows the results of the first test. As it can be seen, the results were not quite accurate. This can be caused by many reasons. The first reason is that the dictionary database used is less accurate in translating word-to-word, so that translating a certain Indonesian word could deliver a lot of different English words. Furthermore, another case shows that there are few words not contained in the database. Because the dictionary database was not capable to detect affixed words and the words adopted from different language.

The second test was conducted without removing the stop words and the terms used were derived from reference paragraph and also the test paragraph. The results were in the range of 9.09% to 27.27% succeed in detecting the similarity between the right paragraphs. 9.09% to 27.27% of the test results only succeed in detecting the similarity in the topic covered by the paragraphs. While 45.45% to 72.73% of the test results failed in detecting the similarity between the right paragraphs, yet also failed in detecting the similarity of the topic.

Table 2. Testing Results without Removing Stop Words and Using Terms Derived from Test and Reference Paragraph

Indicator	% Succeed in Detecting Similarity Between Paragraphs	% Succeed in Detecting Similarity of Topics	% Not Succeed in Detecting Similarity
Length Comparison	27,27	27,27	45,45
Angle Between Least Elements	18,18	9,09	72,73
Angle Between Most Elements	9,09	18,18	72,73

The results of second test are shown in Table 2. The results of this test were not quite accurate. There are several causes for this. The first reason was similar to the first test. The dictionary database used in this system is less accurate in translating word-per-word. There were many Indonesian words translated into several different English words. Another reason was that the combined terms from test and reference paragraphs used to create the matrices affected the results. If the terms were dominated by test paragraph words, then the test matrix vector length could be much greater than the reference matrix vector length. Therefore, the two paragraphs could be deemed similar by LSA.

The second test was conducted by removing the stop words and the terms used were derived from reference paragraph only. The results were in the range of 81.82% to 90.91% succeed in detecting the similarity between the right paragraphs. 9.09% of the test results only succeed in detecting the similarity in the topic covered by the paragraphs. While 0% to 9.09% of the test results failed in detecting the similarity between the right paragraphs, yet also failed in detecting the similarity of the topic.

Table 3. Testing Results by Removing Stop Words and Using Terms Derived from Reference Paragraph

Indicator	% Succeed in Detecting Similarity Between Paragraphs	% Succeed in Detecting Similarity of Topics	% Not Succeed in Detecting Similarity
Length Comparison	90,91	9,09	0
Angle Between Least Elements	81,82	9,09	9,09
Angle Between Most Elements	90,91	9,09	0

The results of this test were accurate, seen from high accuracy average for the three indicators, as shown in Table 3. The anomaly in this test was caused by the same reason as the two previous tests. There were several Indonesian words which were translated into few less appropriate English words, and also several Indonesian words unable to be translated. Therefore a lot of terms used to create the matrices in LSA were not precise and the results are less accurate than expected.

The fourth scenario was a test by removing the stop words and the terms used were derived from reference paragraph and also the test paragraph. The results were in the range of 18.18% to 27.27% succeed in detecting the similarity between the right paragraphs. 18.18% to 27.27% of the test results only succeed in detecting the similarity in the topic covered by the paragraphs. While 45.45% to 63.64% of the test results failed in detecting the similarity between the right paragraphs, yet also failed in detecting the similarity of the topic.

Table 4. Testing Results by Removing Stop Words and Using Terms Derived from Test and Reference Paragraph

Indicator	% Succeed in Detecting Similarity Between Paragraphs	% Succeed in Detecting Similarity of Topics	% Not Succeed in Detecting Similarity
Length Comparison	27,27	18,18	54,55
Angle Between Least Elements	18,18	18,18	63,64
Angle Between Most Elements	27,27	27,27	45,45

Table 5. Testing Results Based on Topics

Topic	Without removing stop words and using terms derived from reference paragraph	Without removing stop words and using terms derived from test and reference paragraph	By removing stop words and using terms derived from reference paragraph	By removing stop words and using terms derived from test and reference paragraph
Organism	25%	25%	91,67%	50%
Biology	33,33%	11,11%	88,89%	11,11%
Biography of Albert Einstein	50%	16,67%	83,33%	8,33%

As shown in Table 4, the results of this test were not quite accurate. The reason of several anomalies occurred in this test was that a number of test and reference paragraphs contain many stop words, so that when the stop words were excluded from LSA process, the matrices created would be small. This caused the results to be less accurate in LSA process using terms derived from test and reference paragraphs.

The percentage shown in Table 5 is the percentage of paragraphs which were succeed to be detected as similar with its English version (reference paragraph) from total paragraphs coming from the same topic. For instances, at the first test, which is without removing stop words and using terms derived from reference paragraph, from four paragraphs about organism, vector lengths comparison as indicator succeed in detecting the similarity between the exact paragraphs one time, angle formed between least element vectors as indicator succeed in detecting the similarity between the exact paragraphs one time, and angle formed between least element vectors as indicator succeed in detecting the similarity between the exact paragraphs also one time. Therefore, from four paragraphs about organism, the average of accurate detection is 25%.

The result also shows that the test by removing stop words and using terms derived from reference paragraph only obtained the most accurate results than the other tests in three topics. While the fourth algorithm obtained the worst result in biology and biography of Albert Einstein paragraphs. This is caused by a lot of incorrect translations occurred while translating the words contained in the paragraphs about these two topics. Several Indonesia words were left untranslated and there were also few words translated into wrong English words. Furthermore, paragraphs which were tend to be short and contain many stop words gave poor results on the test by removing stop words and using terms derived from test and reference paragraph.

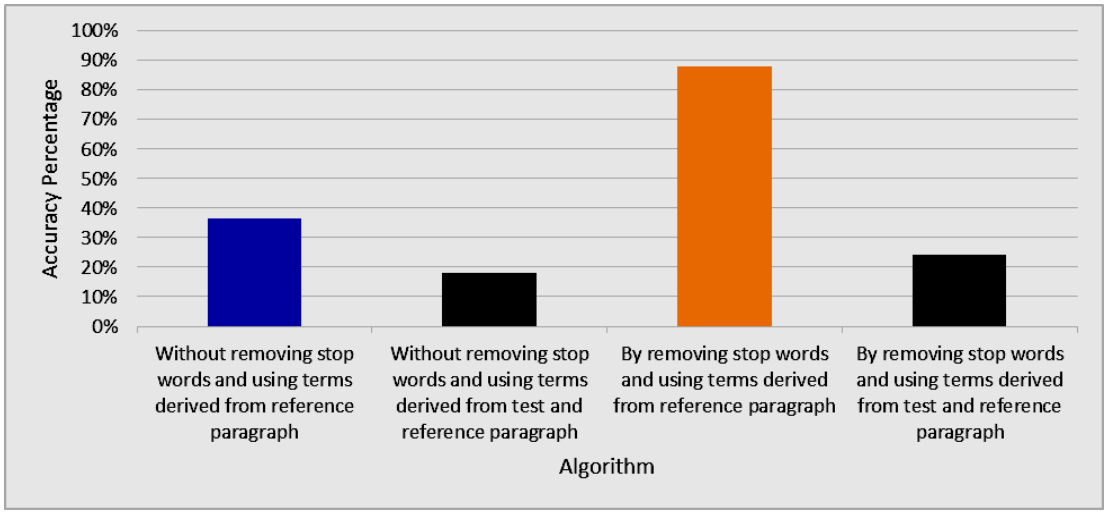

Fig. 1. Algorithm Effect to the Accuracy of Program

The testing results graph is shown in Fig. 1. Overall, the best result was obtained using the third algorithm, which is by removing stop words and using terms derived from reference paragraph to create matrices in LSA process. Whilst, the most exact indicator of similarity is the vector lengths comparison.

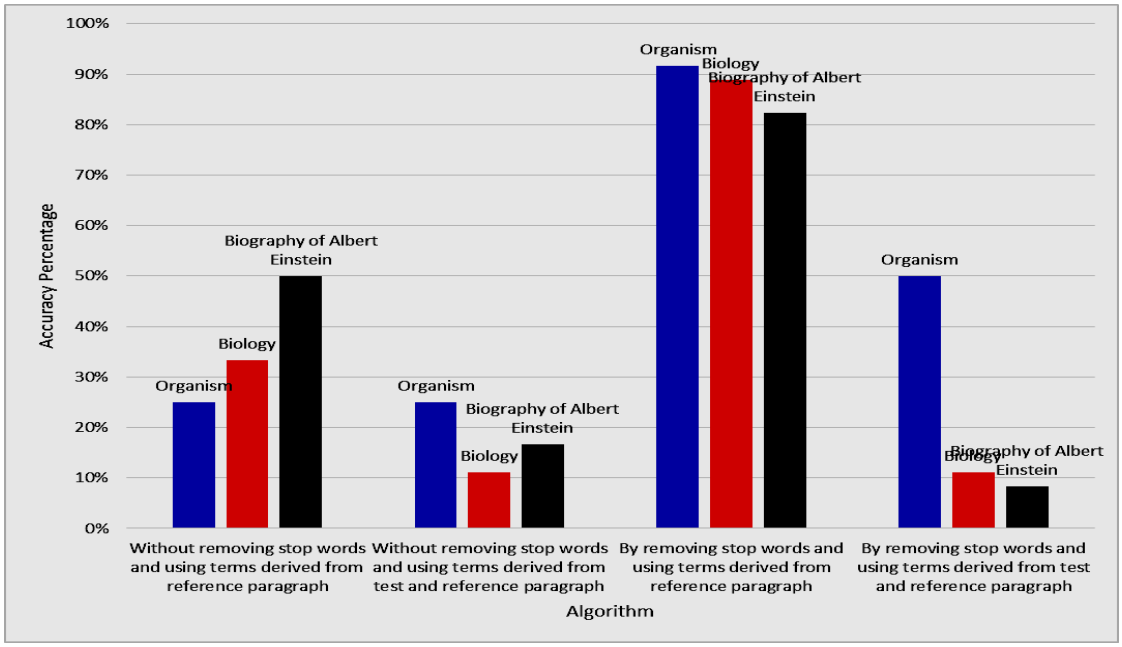

Fig. 2. Paragraph Topic Effect to the Accuracy of Program

Fig. 2 shows the effect of paragraph topic to the accuracy of plagiarism detection. As it can be seen, the test by removing stop words and using terms derived from reference paragraph gave the best results. The most accurate detection obtained from the test on paragraphs from organism topic. Whilst the least accurate detection obtained from the test by removing stop words and using terms derived from test and reference paragraph on paragraphs from the biography of Albert Einstein.

Conclusions

LSA can be used to detect the similarity between two paragraphs written in different languages, which are test paragraph written in Indonesian, while the reference paragraph written in English. Cosine and the length comparison between the Frobenius normalization of two paragraphs are able to be used as the measure of similarity. The most accurate test result acquired from scenario three, which is a test by removing the stop words and the terms used were derived from reference paragraph only. The results reaches 81.82% to 90.91% accuracy in detecting the similarity between the right paragraphs. Our research indicates that LSA are more than capable to be used to detect plagiarism between a papers written in Indonesian with another paper written in English.

References

El Tahir Ali, A. M., Dahwa Abdulla, H. M., & Snášel, V. (2011). Overview and comparison of plagiarism detection tools. *CEUR Workshop Proceedings*, *706*, 161–172.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix Computations*. *Physics Today* (Vol. 10). http://doi.org/10.1063/1.3060478

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. http://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284. http://doi.org/10.1080/01638539809545028

Manning, C. D., & Raghavan, P. (2009). *An Introduction to Information Retrieval*. *Online*. http://doi.org/10.1109/LPT.2009.2020494