Journal of Theoretical and Applied Information Technology
15th September 2021. Vol.99. No 17
© 2021 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEVELOPING A BILINGUAL MODEL OF WORD EMBEDDING FOR DETECTING INDONESIAN ENGLISH PLAGIARISM

[1]YULYANI ARIFIN, [2]SANI M. ISA, [3]LILI AYU WULANDHARI, [4]EDI ABDURACHMAN

[1,3] Bina Nusantara University, Department of Computer Science, School of Computer Science, Jakarta, Indonesia, 11480

[2]Bina Nusantara University, Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Jakarta, Indonesia, 11480

[4]Bina Nusantara University, [3]Department of Computer Science, BINUS Graduate Program-Doctor of Computer Science, Jakarta, Indonesia, 11480

E-mail: [1] yulyaniarifin@binus.ac.id,[2] sani.m.isa@binus.ac.id,
[3] lili.wulandhari@binus.ac.id,[4] edia@binus.ac.id

ABSTRACT

The Internet users that increasing can make it easier to access information even in different languages. Also, the translation application can help users to translate some idea or document without proper citation or acknowledge their idea So, plagiarism is increasing not only in the academic field but also in the industry. A lot of researchers already propose some method to detect plagiarism, but mostly in the European language. Previous research in Indonesian-English plagiarism has already proposed some methods but it is still dependent on machine translation. So, from this research, we purpose a model that can be used to detect cross-language plagiarism without depending on machine translation. The model's purpose is to use combination canonical correlation analysis with the paragraph to vector. Evaluation will be done with the monolingual task and cross-language detection plagiarism. The model evaluation has a good result in monolingual word similarity also when detecting cross-language plagiarism without depending on machine translation. After comparing with the benchmark that using Fingerprint Method with machine translation, the proposed method can detect plagiarism type with paraphrasing more accurately than the benchmark. Even the improvement compared with the benchmark not so significantly but through this proposed method can detect cross-language plagiarism in Indonesian-English language without depending on machine translation. For future work, it needs to enlarge the parallel corpus for Indonesian-English to improve the accuracy of the proposed method.

Keywords: *Word Embeddings, Plagiarism, Bilingual Model, Cross-Lingual, Canonical Correlation Analysis*

1. INTRODUCTION

Internet usage which is increasingly widespread makes it easier for many people to find the desired data. Currently, the assignments given in elementary school are already using internet search facilities. Any kind of information can be found on the internet. The types of languages in the document are English and other international languages. Therefore often the information sought is widely available in international languages compared to minority languages such as Indonesian.

.

Through translation applications, both online and offline, information in other languages can be translated into Indonesian. It easy for many people to find the information needed and it's easy to do plagiarism. According to KBBI (Kamus Besar Bahasa Indonesia), Plagiarism has acknowledged the work of others as their work (1). Meanwhile, according to Merriam-Webster plagiarize is stealing or acknowledging ideas or words from others as one's work without acknowledging other people's credit, (2). Plagiarism is an important focus, especially in the academic field. Of course, it is not easy for the teacher or the lecturer to check or detect the assignment of the student is his work or copy-paste from the work of others. Various methods of detecting plagiarism ranging from manual or automatic began to develop, such as the Turnit in an application.

The results of a literature study conducted by Alzahrani et al [3] found that the detection of plagiarism has been done by many researchers in the 1970s where the detection was carried out on plagiarism which was done in the Pascal programming language and C. Then in 1990 the detection of plagiarism was carried out on documents stored digitally on a system, for example, the library system (3). Then along with the development of research related to information retrieval, natural language processing, artificial intelligence, and even deep learning, the plagiarism detection method also experienced many changes and has a variety of forms.

Gomaa and Fahmy classify plagiarism detection methods in text into three methods namely String-Based, Corpus-Based, Knowledge-Based.(4). But other researchers classify the model plagiarism detection as syntax-based models, Dictionary-based models, Comparable Corpora-based models, and Parallel Corpora based models (5). According to (6) Plagiarism detection depends on the type of plagiarism which is divided into two groups, namely the type of plagiarism that is the same as the original document and the type of plagiarism that modifies the original document either only partially or in a translation from another language.. The type of plagiarism that translates from other languages attracts the attention of researchers to find better methods for detecting plagiarism. Cross-language detection methods that have been used include N-gram (7), Dictionaries based (8), Vocabularies Correlation (9), Machine Translation Technology (10), Latent semantic Indexing (11). Currently, the method that has been widely developed is to use word embedding, pioneered by Tomas Mikolov (12) with a method known as Word2vec. Word embedding method began to be used for detecting plagiarism between languages (13), (14), (15). But generally, the language is detected between English and European languages. There have not been many studies related to the detection of cross languages between Indonesian and other languages. This is generally due to limited available corporate resources.

The obstacle that is often faced by researchers is related to the detection of plagiarism, especially in the corpus. Corpus-based methods require large corpora. However, there are not many Indonesian-language corpora. Whereas for English corpora can be easily found. To detect plagiarism between languages requires a corporation consisting of two languages. The focus of current research is on the detection of plagiarism between Indonesian and English. The proposed solution is to build a corpus from Indonesian and English. By using the Indonesian corpus then tested on the detection method of plagiarism that has proven the results but with a different language corpus.

Comparison of cross-language plagiarism models using bilingual word embedding methods, namely Bilingual Compositional Model (BiCVM), Bilingual Skip-Gram Model (BiSkip), Bilingual Vectors from Comparable Data (BiVCD), Bilingual Correlation Based Embedding (BiCCA) conducted by Upadhyay et al.(16) The comparison of languages carried out in English in German, French, Swedish, and Chinese. The results of the study showed the BiCCA model gave better results than other methods for detecting plagiarism. Therefore this research will apply the BiCCA method in Indonesian and English using a self-developed corpus. Through this method produces that the greater the matrix produced from a document will affect the accuracy.

Then it is tested by the Cross-Language Bilingual Correlation-based embedding detection method which gives better results compared to other methods of detection between English and German languages (17).

Research that has been done related to the Indonesian language is the detection of plagiarism using the Winnowing method in parallel processing but still in the same language or monolingual (18).

Similar research has been done by Alfikri and Purwarianti [31] that proposed Fingerprint Method and CL3NG to analyzing Cross-Language Plagiarism. The Fingerprint method still needs to translate the targeted document before checking the plagiarism analysis. Ratna et al did the detection using the Latent Semantic Analysis and Learning Vector Quantization methods by using a corpus that had been made by themselves but not specifically explained also still depends on machine translation (19). So, this purpose of research is how to build a model for detect cross-language plagiarism without depending on machine translation. Based on [17] that already generate bilingual model but still in a European language, so the purpose solution is to adopt the model BICCA with combination paragraph to vector. The proposed model will be evaluated with monolingual word similarity and cross-language plagiarism detection task to measure the model performance. The last evaluation is to compare with the benchmark research that has been done by Alfikri and Purwarianti [31]. This research only limited to heuristic retrieval phase and detailed analysis phase in the Plagiarisme stages (6). It is focus on analyzing the plagiarism in level document and level sentence, without checking the similarity that be found on the document, is already using the suitable citation or not.

2. LITERATURE REVIEW

2.1 Plagiarism Type

The division of plagiarism types is divided into two namely (20) (Fig 1) :

a. Exact copy consists of large parts detected by comparison of documents and small parts can be detected by identifying chunk and style analysis
b. Modification is divided into two, namely translation from different languages can be detected by cross-language similarity analysis, and language restructuring can be detected by similarity analysis which is divided into two parts, the large part is detected by comparison of document models while the small part can be detected by fingerprint if using corpus and style references analysis if not using the corpus reference.

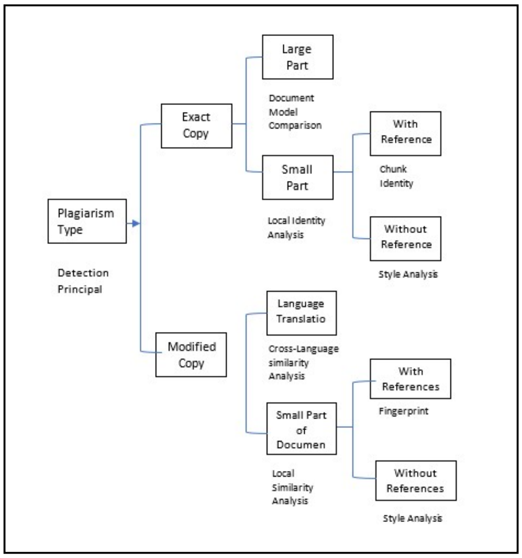


Figure 1 Plagiarism Type [20]

The Taxonomy from Literal Plagiarism that has promoted by Alzahrani et all [20]. He categorized Literal Plagiarism which plagiarism that only changing the location or position with the same words, become three categories as following :

1. The first category is Exact Copy which has a similar part for the whole document or part of the document. It means only copy-paste from the source document. This kind of category only can be done for monolingual plagiarism.
2. The second category is Near Copy which plagiarism uses insert another word, delete some words, or substituting the word with similar meaning, or splitting and joining sentences. This action still not changing the meaning from the source document.
3. The third category is Modified Copy which phrase reordering, or changing syntax example from active form become passive form.

But for advanced plagiarism, is well known as intelligent plagiarism like text manipulation, translation, idea adoption [20]. The structure can be seen in Figure 2.

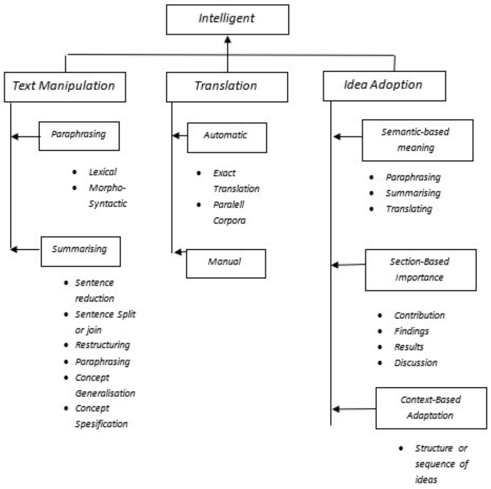


Figure 2 Intelligent Plagiarism Taxonomy [20]

The focus of this research is only on the type of plagiarism modification that results from the translation of different languages. The translation can be combined with paraphrasing.

2.2 Indonesian Language

Indonesian is a variation of Malay. Indonesian is increasingly developing with the addition of new vocabulary from assimilation with foreign and regional languages (21). Indonesian has affixes attached to the basic form (lexeme), namely Prefix, Infix, Suffix, and Circumfixes. Examples of affixes can be seen in Table 1 (22) :

Table 1 Example list of Affixes

Affixes	Example Affixes
Prefixes	pe-, ber-, di-, ter-, meN-
Infix	-er-, -em-, -el-, -in-
Suffixes	-an, -kan, -i, -or,
Circumfixes	per-an, peN-an, me-kan, ke-an, di-kan

Some examples of the use of affixes in Indonesian:

a. Added Prefix Pe on basic form will become noun example : Pe + laut (sea) = Pelaut (Sailor) , Pe + kerja (work) = Pekerja (worker)

b. Added Infix -el- combine with basic form will become Noun example -el- + tunjuk (point) = telunjuk (Index Finger) , -el- + tapak (foot print) = telapak (palm).

c. Suffix -kan added with word will become Verb example: lepas (free) + -kan = lepaskan (let it go) , tinggal (stay) + -kan- = tinggalkan (leave it).

d. Circumfixes per- an joint with word become noun example per-an + kumpul (gather) = perkumpulan (association) , per-an + lengkap (complete) = perlengkapan (equipment).

When compared with the morphology of the English language is almost the same as the Indonesian language, but there are more types of affixes in the Indonesian language compared to the English language, and the word-formation after adding an affix is different (23). This also happens to other European languages that are often used in previous cross-language plagiarism research. Because of that, it is challenging to examine the BiCCA method to be applied to low resource languages like Indonesian. (16) to be applied to low-resource languages like Indonesian.

2.3 BiCCA

A cross-language detection method that applies a canonical correlation analysis-based method on each monolingual vector from the parallel corpus to produce a correlation between the two languages (17).

An illustration of the BiCCA method can be seen in Figure 3.

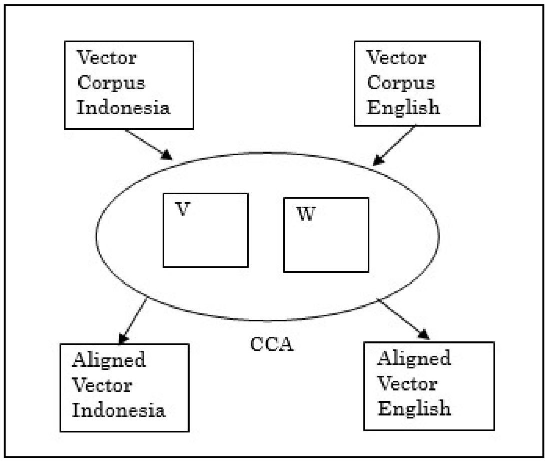


Figure 3 Bicca Illustration (Adapted From (17))

Vector space embedding of each Corpus generated, then use Canonical Correlation analysis to get the correlation that is appropriate for each vector thus producing two projection vectors. Then two projection vectors are assigned to the vector space embedding of each corpus to produce a new aligned vector that is different from the original vector.

2.4 Word Embedding

Word embedding in mono-lingual was introduced by Mikolov [12] well known as Word2Vec. It has two architecture that learns underlying word representations for each word using a shallow neural network. They are CBOW (Continuous Bags of Words) and Skip Gram. The CBOW using a distributed representation of contexts to predict the word between the context. The Skip Gram is vise versa from CBOW, it is predicted the context using a distributed representation of context. Both architectures can be seen in Figures 4-5.

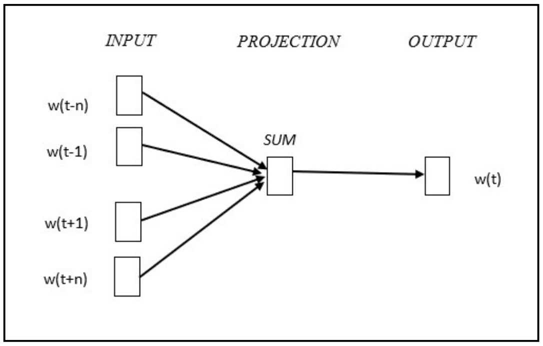


Figure 4 CBOW's Architecture[12]

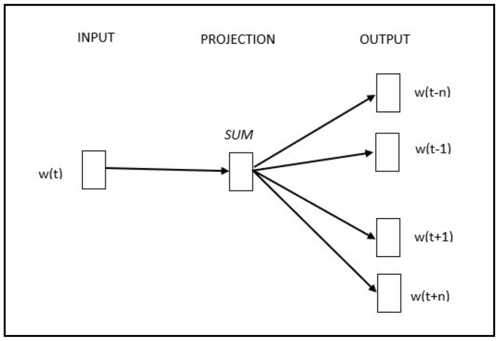


Figure 5 Skip Gram Architecture [12]

Paragraph to vector, which is well known as Doc2Vec is an enhancement from Word2vec [29]. The Doc2Vec is having vectors for words also the vector for paragraphs or sentences. It has two models, the one is Distributed Memory Model of Paragraph Vectors (PV-DM) and Distributed Bags of Words (PV-DBOW). The combination of the two models can give the best performance. The PV-DM is almost the same as the DBOW process and the PV-DBOW almost the same as the Skip Gram. Usually, Skip Gram to be used for the dataset that has a small size, but the CBOW is most used for large datasets.

2.5 Cross-Language Similar Analysis

According to Rosso et al, there is four retrieval model to detecting cross-language plagiarism such as based on syntax like model CL-CNG, based on the dictionary like model CL-VSM, based on comparable corpora like model CL-ESA, and the last one is based on parallel corpora like CL-ASA, CL-LSI and CL-KCCA [10].
The easiest way is using CL-CNG but has poor performance. The best performance is using CL-ASA but it is still machine translation.

2.6 Cross-Language Plagiarism in Indonesian English

Alfikri and Purwarianti construct the detection of cross language plagiarism in Indonesian English using Fingerprint method. But it still need to translate the suspect document first before doing the detection and also it was lacking when detecting the paraphrasing plagiarism type(24).

Another previous studies is done by Ratna et al that using Latent Semantic Analysis and Learning Vector Quantization to detect cross-language plagiarism (25) that still need to translate the suspect document before can be analyse the plagiarism.

3. THE STAGES OF MODEL DEVELOPMENT

The stages of Model development consist of several stages (Fig 6)

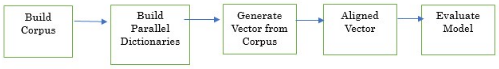


Figure 6 The stages of Bilingual Word Embedding Model

1. The stages of build Corpus

The first stage is to build Corpus. In the bilingual model of word embedding, we need a parallel corpus. The Indonesian language is the low source language. There are not so many corpora that contain parallel sentences Indonesian English sentences. The resources for the parallel corpus is dataset standard that builds by BPPT [24] also dataset new collection from [25]. They contain some news from BBC, ODB, AusAID, also SMERU. The detail of the datasets can be seen in table 2. This dataset contains news collections with various topics such as international news, economics, politics, and science. Total parallel sentences in all datasets are around 64.000 sentences.

Table 2 List Of Dataset

Dataset Name	Parallel Sentences	Number of Words EN	Number of Words ID
PAN BPPT	24.025	541.356	500.035
BBC	469	8.286	7.497
ODB	9.824	152.973	151.614
AusAID	3.113	65.555	64.024
SMERU	26.967	594.643	505.859
	64.398	1.362.813	1.229.029

One of parallel document from the dataset can be seen in Figure 7 – 8. We use the news dataset, because the standard dataset that related with academic field is not yet developed.

Tunisia membalas kekalahan dari Zambia dengan kemenangan 1-0 pada Selasa dalam pertandingan pemanasan Piala Afrika 2008.
Zambia mengejutkan negara Afrika Utara itu dengan kemenangan 2-1 pada Minggu berkat dua gol di awal pertandingan dari Felix Katongo sementara Yacine Chikhaoui mengurangi selisih angka pada babak kedua di tengah cemooh suporter tuan rumah yang gusar.

Figure 7 Example document in Indonesia Language

Tunisia turned the tables on Zambia Tuesday with a 1-0 win in a 2008 African Nations Cup warm-up match.
Zambia shocked their North Africa hosts with a 2-1 win Sunday courtesy of two early goals from Felix Katongo while Yacine Chikhaoui reduced arrears in the second half amid boos and jeers from furious local supporters.

Figure 8 Example Document In The English Language

2. . Build Parallel Dictionaries

Based on the parallel corpus then we generate a parallel dictionary. The stages to develop the parallel dictionary is started from Preprocessing process that will be done for each dataset. In the Preprocessing process, we have done tokenization for each parallel dataset. After that continue to mapping each sentence in the Indonesian Language with the English Language. After getting parallel sentences that already mapping each other, then continue to aligning process. The method to be used in generating a parallel dictionary is using the fast alignment from Chris Dryer et al [26].

The whole stages of generating the parallel dictionaries can be shown in Figure 9.

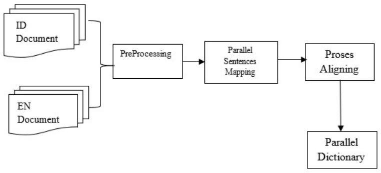
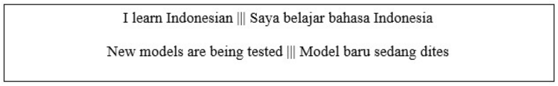


Figure 9 The Stages To Develop Parallel Dictionary

Using this alignment method from Chris Dyer can generate parallel dictionaries (See Table 3). The simulation for aligning process can be explained like this: from parallel sentences (Figure 10)


I learn Indonesian ||| Saya belajar bahasa Indonesia

New models are being tested ||| Model baru sedang dites

Figure 10 The Mapping of Parallel Sentences

Every word on each sentence will be aligning with looking at the highest probabilistic. Every word from the pairing sentence will be calculated their lexical probabilities. The simulation of Probabilities Alignment for Word 'Saya' can be seen in table 3. The number beside the word refers to the word position on the sentence. From table 3 the high probabilistic is ID word 'Saya' with EN word 'I'. So, during the alignment process, the word "Saya: will be aligned to the word 'I'. It means the translation for ID Word 'Saya' is 'I'.

Table 3 Simulation Alignment Probabilistic

ID Word	EN Word	Probabilistic
Saya (0)	I (0)	0,8
Saya (0)	Learn (1)	0,3
Saya (0)	Indonesian (2)	0,1

Every ID word will be aligned with the EN word with the highest probabilistic [Fig.11].

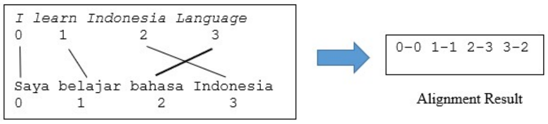


Figure 11 The Alignment Process

Using the alignment result with index mapping for each word, then all the parallel sentences will be aligned with them. The result after aligning process can be seen in Table 4.

Table 4 The Parallel Dictionary English Indonesia

Kata (EN)	Kata (ID)
Coach	Pelatih
Out	Keluar
Since	Sejak
Injuries	Cedera
Minor	Kecil
Without	Tanpa
Month	Bulan

3. Generate Vector From Corpus

The vector model will be generated from a parallel corpus using the Word2Vec method. All models will be trained with the embedding of size 200 with the Skip Gram method. So there is two model vector, let the I as the set of the vector from Indonesian Corpus and let the E as the set of a vector from English Corpus.

4. Aligned Vector

Let x as the corresponding vector from Indonesian words and y as the corresponding vector from English words. Using CCA (Canonical Correlation Analysis) maximizes the x and y correlation and will output two projection vectors A and B. During the maximizes process, CCA using the parallel word dictionary that was already generated from stages before to get the maximized correlation. The set of the vector from I and E will be projected with two projection vectors A and B and will output the new set of the vector from I* and E* (1) & (2).

$$A, B = CCA(x, y) \quad (1)$$

$$I^* = I\,A \qquad E^* = E\,B \quad (2)$$

5. Evaluate Model

After aligned vector from both corpus, then we get a bilingual model of word embeddings of English and Indonesia. To evaluate the model contains two tasks, the first task is evaluating the monolingual word similarity for English and the second task is evaluating this bilingual model to cross-lingual plagiarism. This first task is to evaluate the quality of English Embeddings in the Bilingual Model of Word Embeddings. It compares the correlation between words generate by the system with the dataset with human correlation. In this case, we compare with English Dataset Standard WS-353 and SIMLEX-999. The second task is to evaluate this model can be used to detect plagiarism in English-Indonesian documents. We evaluate this model with three plagiarism test cases, there are test cases with no plagiarism, light plagiarism, and full plagiarism.

4. RESULT AND DISCUSSION

4.1 Monolingual Word Similarity

After evaluation with Dataset Standard WS-353 and SIMLEX-999 using Spearman Rank Correlation [27], then the result can be seen in table 5

Table 5 Spearman Rank Correlation Result

Dataset	Vector Monolingual	BiCCA Model	BiCCA Model with MonoLingual Vector	Steiger Value (p < 0,1)
SIMLEX-999	0,20	0,20	0,9994	0,0037
WS-353	0,44	0,44	0,9993	0,0012

Based on the result (Table 5) if the Steiger Value (p< 0,01) [28], it means the bilingual model can improve the correlation between words in English. The Steiger value on SIMLEX-999 higher than WS-353 because the number of datasets SIMLEX-999 larger than WS-353.

4.2 Cross-Lingual Plagiarism Detection Task

The Cross-Lingual Plagiarism Detection contains three steps start from Heuristic Retrieval Step, the Detailed Analysis Step and the last one is Post Processing Step [6].

In the heuristic retrieval step, the proposed method is to generate a document to vector from the Bilingual Model of Word Embedding. The illustration from document to vector can be seen in

Figure 12. Every document will have a vector for each word in the document also it has a vector as the document identification.

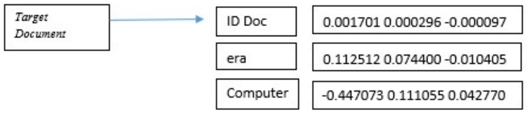


Figure 12 Paragraph To Vector For One Document

After that, using the Cosine Similarity Distance to checking the similarity between the test case document with the source document. Using this proposed method will generate a lot of list candidates who suspect have a similar part.

The second part is a detailed analysis. In this part, all the list candidates will be checking more detailed. Because of that, we must generate vectors by sentences. The target document and the list candidate will be generated vector by sentence, it is well known as Sentence2Vec.

Based on the list candidate then it will analyze more detailed, which one that has plagiarism with the test case document. The next step is to generate a paragraph to vector from all the list candidates from the Bilingual Model of Word Embedding. Using the Jaccard Similarity measure to analyzed in more detail how much the test case has plagiarism from the list candidate document. The simulation for checking the plagiarism between the candidate list and the target document can be seen in Figure 13. Each sentence in the target document will be checking with the other sentence on the candidate list. This action will be repeated until all sentences in the target document. After that using the Jaccard Similarity to calculate how many percent the similarity between the target document with the candidate list.

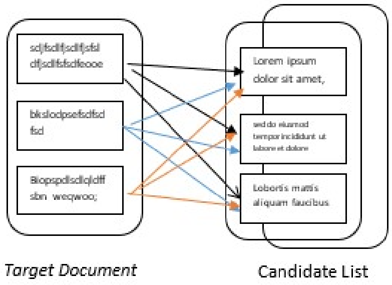


Figure 13 The Simulation Plagiarism Checking

Evaluate the model with the Cross-Lingual Plagiarism Detection Task, we must have a test case document in Indonesian text. Then, we generate a vector from each test case. Using the Cosine Similarity to measure the similarity from test case with the bilingual model of word embedding.

The parameter to be used in generating the test case based on Barron-Cedeno and Photthast [30] :

a. The length of documents. It can vary but commonly contains small documents contains three – ten sentences and large documents contains more than five sentences. For this experiment, we only use 10 target documents for each test case. There are five documents with small size and five documents with large size.
b. Plagiarism Degree. It refers to how much the plagiarism part on that suspect documents. It can be no plagiarism (0%) means the target document does not have any similar part with the source documents, light plagiarism (around 20%) means the only small part that has similarities with the source documents, half plagiarism (around 50% more) means half part from the target document similar with the source documents, and the last one is full plagiarism means all part from the target documents is translated from the source document.
c. Plagiarism Type. It is can be known as modification plagiarism due to the language in the target document is different from the source document. The target document will use Indonesian Language and the source document will be the English Language.

4.2.1 Generate Test Case

There are four scenarios to generate a test case with explaining below :

1. Test Case with Full Plagiarism

In this test case, the target document or tested document will be copied exactly with the document in the Corpus. Some parts from sources will be located randomly, so the contents are still the same but with a different location. And some of the document targets are 100% the same as the source document but with different language. In this test case, we have generated

2. Test case with half plagiarism

The target document has half the part that similar to the source document. A similar part will be a different location from the source documents.

3. Test case with light plagiarism

In this test case, only a small part of the target document is translated from the source document. It depends on how long the target document. A similar percentage will be only 20% from all lines in the target document.

4. Test case with no plagiarism

This target document will be different and there is no translation from the source documents. The context in this target document will be taken from the news that different categories with the source documents.

4.2.2. The Evaluation Measure

The evaluation measure is using the Precision, Recall, F-Measure, and Accuracy [6]. The Precision measure how many percentages of the plagiarism part will be detected by the system as a plagiarism case. The Recall measure how many plagiarism cases can be detected by the system. The F1 Measure is used to balance between the Precision score and Recall score. The accuracy score to detect the accuracy from the proposed method.

The performance result can be seen in table 6. The proposed method can detect plagiarism part as a plagiarism case with the Precision Score for test case Full plagiarism, half plagiarism, and light plagiarism. But it still needs to improve the performance for detection No Plagiarism test case. The accuracy score showed that the light plagiarism case still needs to be improved.

Table 6 The Performance From Each Test Case

Performance	Full Plagiarism	Half Plagiarism	Light Plagiarism	No Plagiarism
Precision	1	1	1	0,60
Recall	0,60	0,60	0,30	1
F1-Measure	0,75	0,75	0,46	0,75
Accuracy	0,60	0,60	0,30	0,60

4.2.3 Evaluate with the Benchmark Method

This proposed method will be evaluated with the benchmark method. The benchmark is using the method Fingerprint with CL3NGram. This method is proposed by Alfikri and Purwarianti [31]. With this fingerprint, every target document must be translated first before plagiarism checking. After translation then the document will be chunking as 3N gram. With Winnowing, the algorithm will generate a fingerprint from the target document. The analysis of plagiarism with comparing the fingerprint between the target document and the source document.

The scenario to be used will contain four test case as follows:

a. Test Case 1 is a full plagiarism test case which parallel document with target document . It means the target document is translated from the source document
b. Test Case 2 is a full plagiarism test case but with paraphrasing sentences.
c. Test Case 3 is a half plagiarism test case from source document where the similar part location is randomly.
d. Test case 4 is a no plagiarism test case .

The comparison result with the benchmark can be seen on Table 7. The test case with paraphrasing can be detected more accurately by the proposed method. But for test case that have no plagiarism , the benchmark can detect more accurate than the proposed method.

Table 7 Comparison Result With Benchmark

Test Case	Benchmark	Proposed Method
1	0,57	0,39
2	0,15	0,31
3	0,02	0,25
4	0,14	0,12

Based on table 8 that shown the performance from benchmark method and the proposed method. The benchmark can detect the plagiarism part as the plagiarism case more higher than the proposed method. But the proposed method has the Recall score higher than the benchmark. The accuracy from the proposed method more higher than the benchmark even not significant.

From this comparison, we can conclude that using Bilingual Model can detect the plagiarism in cross language without need the machine translation. The benchmark still need the machine translation to translate the target document so it can be same language with the source target. But , for test case that no plagiarism still need improve the accuracy from proposed method.

Table 8 Performance Comparison With Benchmark

Performance	Benchmark	Proposed Method
Precision	1	0,5
Recall	0,6	1
F1-Measure	0,75	0,67
Accuracy	0,60	0,62

The corpus need more bigger size to enhance the accuracy from the Bilingual Model from proposed method. As we know that the parallel sentences in the Corpus Bilingual Model of Word Embeddings only contain 64 thousands due to Indonesian Languages is a one of low resource language in the world. If the sizes of corpus can be improved , the accuracy from the proposed method can be more significant. Besides the corpus size, there is another way to improve the accuracy of the proposed method, which is to use a combination method to analyze the plagiarism part in detailed analysis.

5. CONCLUSION AND FUTURE WORK

The Bilingual model of word embedding can be used to detect cross-language plagiarism in Indonesian-English. Even, the accuracy still needs to be improved due to limited standard resources which have parallel sentences with the English Language. The accuracy from the proposed model only lights different from the benchmark. But in the deeper analysis, it can be seen that the proposed method can detect plagiarism test cases with paraphrasing but it still needs to improve the accuracy when detecting no plagiarism part.

For future work, it is still to develop Academic Parallel Corpus to be tested with this proposed method and focus on the last phase in plagiarism stage, Knowledge based post processing. Then, to improve the performance from the model using various combinations like Machine Learning or deep learning.

REFERENCES:

[1] Arti Kata Plagiat Kamus Besar Bahasa Indonesia [Internet]. Available from: https://jagokata.com/arti-kata/limbah.html

[2] Plagiarize | Definition of Plagiarize by Merriam-Webster [Internet]. Available from: https://www.merriam-webster.com/dictionary/plagiarize?show=0&t=1363947245,

[3] Brin S, Davis J, García-Molina H. Copy detection mechanisms for digital documents. ACM SIGMOD Rec. 1995;24(2):398–409.

[4] 4. Gomaa WH, A.Fahmy A. A Survey of Text Similarity Approaches. 2013;68(13):13–9.

[5] Danilova V. Cross-Language Plagiarism Detection Methods. Ranlp. 2013;(September):51–7.

[6] Photthast M, Barron-Cedeno A, Stein B, Rosso P. Cross-language plagiarism detection. 2011;45–62.

[7] Mcnamee P, Mayfield J. Character N -Gram Tokenization for European. 2004;73–97.

[8] Steinberger R, Pouliquen B, Ignat C. Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. 2006;(September 2013). Available from: http://arxiv.org/abs/cs/0609064

[9] Potthast M, Stein B, Anderka M. A Wikipedia-Based Multilingual Retrieval Model. 2008;522–30.

[10] Rosso P, Pinto D, Juan A, Barr A. On Cross-lingual Plagiarism Analysis using a Statistical Model. Proceeding Conf Proc ECAI'08 Work Uncovering Plagiarism, Authorsh Soc Softw Misuse, Patras, Greece. 2008;

[11] Landauer TK, Littman ML. Fully automatic cross-language document retrieval using latent semantic indexing. Proc sixth Annu Conf UW Cent new Oxford English Dict text Res. 1990;31–8.

[12] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013;1–12. Available from: http://arxiv.org/abs/1301.3781

[13] Ferrero J, Agnes F, Besacier L, Schwab D. Using Word Embedding for Cross-Language Plagiarism Detection. 2017; Available from: http://arxiv.org/abs/1702.03082

[14] Thompson V, Bowerman C. Detecting Cross-Lingual Plagiarism Using Simulated Word Embeddings. 2017; Available from: http://arxiv.org/abs/1712.10190

[15] Mogadala A, Rettinger A. Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-Language Text Classification. 2016;692–702.

[16] Upadhyay S, Faruqui M, Dyer C, Roth D. Cross-lingual Models of Word Embeddings: An Empirical Comparison. 2016; Available from: http://arxiv.org/abs/1604.00425

[17] Faruqui M, Dyer C. Improving Vector Space Word Representations Using Multilingual Correlation. 2015;462–71.

[18] 18. Cross P, Plagiarism L, Indonesia B, Bahasa DAN, Fuzzy M, Dan F. Yulyani Arifin. 2017;

[19] Ratna AAP, Purnamasari PD, Adhi BA, Ekadiyanto FA, Salman M, Mardiyah M, et al. Cross-language plagiarism detection system using latent semantic analysis and learning

vector quantization. Algorithms. 2017;10(2):1–14.

[20] Alzahrani SM, Salim N, Abraham A, Member S. Understanding Plagiarism Linguistic Patterns , Textual Features , and Detection Methods. 2012;42(2):133–49.

[21] Putrayasa IGNK. Sejarah Bahasa Indonesia. 2018;1–18. Available from: https://simdos.unud.ac.id/uploads/file_penelitian_1_dir/3c680101ff285bcffdcd4eb7e8862e67.pdf

[22] Suparno D. Morpholog Bahasa Indonesia. 2013;1–108. Available from: http://repository.uinjkt.ac.id/dspace/bitstream/123456789/33994/1/BUKU Morfologi Bahasa Indonesia 4 Desember 2014.pdf

[23] Pauzan. Contrastive Analysis Between English and Indonesian Prefixes and Suffixes (A Narative Text Analysis of Legends in Perspective of Morphology). J Educ Pract. 2016;7(30):1–8.

[24] Alfikri ZF, Purwarianti A. The Construction of Indonesian-English Cross Language Plagiarism Detection System Using Fingerprinting Technique. J Ilmu Komput dan Inf. 2012;5(1).

[25] Ratna AAP, Nabhastala PNY, Ibrahim I, Ekadiyanto FA, Salman M, Purnamasari PD, et al. Cross-language automatic plagiarism detector using latent semantic analysis and self-organizing map. ACM Int Conf Proceeding Ser. 2018;83–7.

[26] Hammam, Riza & Riza, Chairil & Hakim,. 2009. Resource Report: Building Parallel Text Corpora for Multi-Domain Translation System. 92-95.

[27] Desmon86 (2012) Indonesian-English-Bilingual-Corpus (BBC). Available at: https://github.com/desmond86.

[28] Dyer, C., Chahuneau, V. and Smith, N. A. ,2013, 'A simple, fast, and effective reparameterization of IBM model 2', NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, (June), pp. 644–648.

[29] Myers, J.L., Well, A. and Lorch, R.F., 2010. Research design and statistical analysis. Routledge.

[30] Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. Psychological bulletin, 87(2), p.245.

[31] Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

[32] arrón-Cedeño, A. and Photthast, M. (2012) 'Corpus and Evaluation Measures for Automatic Plagiarism Detection', pp. 1–19. doi: 10.1002/esp.590.

[33] Alfikri ZF, Purwarianti A. The construction of Indonesian-English cross language plagiarism detection system using fingerprinting technique. Jurnal Ilmu Komputer dan Informasi. 2012 Jul 28;5(1):16-23.

[34] Tung KT, Hung ND, Hanh LT. A Comparison of Algorithms used to measure the Similarity between two documents. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2015 Apr;4(4):1117-21.