International Journal on Emerging Technologies **11**(5): 346-351(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Improved Algorithm for Malay Word Sense Disambiguation

***Mohd Pouzi Hamzah*[1] *and Mohd Arizal Shamsil Mat Rifin*[2]**
[1,2]Universiti Malaysia Terengganu, Malaysia.

(Corresponding author: Mohd Pouzi Hamzah)

(Received 02 July 2020, Revised 19 August 2020, Accepted 10 September 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The vagueness of word sense is a major problem in the field of information retrieval and Natural Language Processing (NLP). This problem occurs because of computers cannot determine the correct meaning of a sentence based on the context of the sentence in a document. This is due to the nature of a word that has more than one meaning called a polysemy word. Word sense disambiguation algorithms can be categorized into three types, i.e supervised, unsupervised and knowledge-based. Our improved algorithm has been developed based on hybrid approach of unsupervised and knowledge-based. The unsupervised approach used is based on the Yarowsky algorithm which uses one sense per collocation as a determinant of the correct meaning of the ambiguous word while the knowledge-based approach used is bag of word model to map each ambiguous word to the definition in the dictionary. In addition, this algorithm also uses external sources of knowledge that are from Malay Wordnet and Google search engine. The experiment has been conducted using 10 ambiguous words and the results outperform other three algorithms namely Lesk, Yarowsky and Google Translate.

Keywords*:* Information Retrieval; Malay Text; Unsupervised; Malay Word Sense Disambiguation.

I. INTRODUCTION

Every day, we speak thousands of words, write and type millions of words to convey information that is important or unimportant, right or wrong. Without realizing that words we speak are ambiguous. Ambiguous here means that the words have more than one meaning which are called polysemy words. We as humans do not have much trouble guessing and interpreting the word even though it has more than one meaning [18]. This is because humans are endowed with a great mind-set that works so consciously that we can guess the meaning of words in a short time and we do not realize that the words are ambiguous.

However, computer as one of the man-made machines is not yet able to guess the exact definition of a word without using appropriate algorithms and through the process of word sense disambiguation. These studies are still lacking, especially in the Malay language [18]. Research on polysemy is very important because without knowing the true meaning of a word it can lead to misunderstanding about a subject. For examples in the phrase "diamenebassemak" and in the phrase "sayasukaminummadu" there are ambiguous words "semak" and "madu". The word "semak" can be translated as bush and can also be defined as check, while the word "madu" can be interpreted as honey which is the sweets produced by the bee and can also be the second wife to the husband. Clearly, certain words in Malay and other languages can contain ambiguous words which lead to ambiguity problem. Therefore a good word sense disambiguation algorithm is required to enable a paragraph or a sentence to be accurately understood by the computer.

II. RELATED WORK

The process to identify the correct sense for the ambiguous word based on contexts is known as Word Sense Disambiguation[1].There are a lot of research have been carried out by previous researchers related to word sense disambiguation. However, current approach for Natural Language Processing for Malay language is still lacking [13]. There are three categories of word sense disambiguation algorithms namely supervised approach, unsupervised approach and knowledge based approach [9, 10].

Supervised approach is an approach that use manually tagged sense-annotated data and machine learning technique. This approach gives the highest accuracy compared to two other approaches [9, 4]. However, the supervised approach is too dependent on tagged corpus as a source of knowledge and the corpus needed is huge and inadequate to cover all the ambiguous words even for English [11]. The processes of training consume too much time and cost [17]. Yamaki *et al.*, [14] (2016) proposed a method that employs sentences similarities from context word embedding for supervised word sense disambiguation.

Unsupervised approach does not use any tagged corpus to identify the exact sense of ambiguous word [9, 8, 20]. This special characteristics make it very useful to disambiguate any language without being limited by number of human tagged corpus [4]. However this approach does not use information from any dictionary or sense inventory, thus it ignores sense information in determining the true meaning of words [2].

Besides supervised and unsupervised word sense disambiguation, knowledge based approach is an approach that determine the correct sense of a word by using information retrieved from knowledge sources such as dictionary and thesauri [9]. This approach does not face bottle neck problem since no training data needed [11, 5].

Recognizing the capabilities of unsupervised and knowledge based approaches, various studies have been conducted to improve the capabilities of this approach in solving word sense disambiguation problems. One of the famous algorithm was developed by [15] named Yarowsky Algorithm. This method uses two properties of human language to disambiguate the ambiguous word which are one sense per collocation and one sense per discourse properties. This algorithm works well and nearly achieve similar performance as supervised approach.

Yarowsky algorithm is very popular and has been adopted in many research works such as in[6]. In this research the method was developed based on spectral clustering and reorders the result based on similarity value. The incidence matrix was built to identify features occurring in the document based on one sense per discourse concept. If the document contains the target several times only its first occurrence will be considered. Recently,[11] also used collocation and decision list algorithm which was introduced in [15, 16]. In this research, statistical method for collocation extraction from a big untagged corpus was used.

Issues regarding the ambiguity in the word sense not only occur in English language but also faced by the Malay language. Among the research carried out in Malay word sense disambiguation is like the one that has been done [13]. This study used sources from other languages such as AWN(Asian WordNet) and Princeton WordNet (PWN) to conduct the word sense disambiguation process. This method of recognition identifies a combination of several methods namely vector, vector pair, path equation and Lesk method. This method consists of three main modules, namely, word construction and extraction, word counting and translation, and decomposition and evaluation. In the word construction and extraction modules, segmentation tools have been used. However this method is not automatic since it requires the assistance of linguists to ensure that words are segmented well and to minimize the disambiguation time.

The next study was conducted by [12], entitled Word Prediction Algorithm in Resolving Ambiguity in Malay Text. In this study n-gram method was used to solve the problem and thus obtain the correct document. The researcher uses the hadith and the translation of Quranic verses as a source of corpus. Experiments were conducted in order to identify the better n-gram either bigram or trigram. Results of the experiment show that both bigram and trigram have their own advantages and limitations. The advantage of bigram method is that it is easier to find similar bigrams in many pre-processed documents however the disadvantage is that bigram can only look up for only one word: word before or after. However, for trigrams, the word predicted can look up the word before and after, thus this will give the better prediction.

Research conducted by [19] examined the whole process of taxonomy learning of Malay language text using unsupervised cluster approach and review the existing Malay NLP as a pre-processing tool that has the potential for the proposed ontology learning approach.

Three language tools tested in this study, two of which were developed by Translation Unit Through Computer (UTMK), Universiti Sains Malaysia (USM) which are Malay sense marker and a Malay language parser using the maximum-entropy based on open NLP package and a shallow parser based on grammar pattern developed [3]. The results of the tests are as shown in table 1.

Table 1: Result of Previous Tools.

Language Tools	Accuracy	Recall	F-Measure
Malay sense marker	0.63	0.62	0.63
Malay language parser using the maximum-entropy	0.77	0.56	0.63
Shallow parser based on grammar pattern developed [3]	0.10	0.14	0.11

III. METHODOLOGY

In this study, Malay word sense disambiguation algorithm has been developed based on a hybrid of the two approaches which are unsupervised approach and knowledge-based approach. In this algorithm, the unsupervised method used is adapted from the Yarowsky algorithm [16]. This algorithm gives high results for the English word disambiguation process. However, this algorithm is not a fully automatic algorithm and has some other disadvantage. Therefore, to make the algorithm fully functional the word sense disambiguation process is enhanced and improved with unique sequence of algorithm that will be described in this section.

There are five major steps in this algorithm which are:

1. Development of the corpus,
2. Identify the ambiguous word,
3. Identify the collocation,
4. Mapping collocation to the correct sense and
5. Mapping ambiguous words to the correct sense based on collocation.

The pseudo code of this algorithm is shown below:

```
1. Start
2. Develop the Corpus
  2.1 Collect documents according to the polysemous word
3. Identify the collocation
  3.1 Tokenize all words in the document
  3.2 Obtain token intersect for each token
  3.3 Calculate the standard deviation for each token intersect
  3.4 Select the appropriate token to be a colocation based on the intersection frequency and the standard deviation value
4. Map the colocation to the definition
  4.1 Expand collocation with the   term intersect
  4.2 Calculate the similarity with the definition of each sense in the dictionary
  4.3 Assign collocation to the definition with the highest similarity.
  4.4 If the similarity <= 0
    4.4.1 Develop collocation with synonyms from Malay Wordnet
    4.4.2 Calculate the similarity with the definition of each sense in the dictionary
    4.4.3 Assign collocation to the definition with the highest similarity.
  4.5 If the equation <= 0
    4.5.1 Expand your collage with Google search results
```

4.5.2 Calculate the similarity with the definition of each sense in the dictionary
4.5.3 Assign collocation to the definition with the highest similarity.
5. Search for meaning based on collocation
6. End

A. Development of a corpus

Development of a corpus for Malay word sense disambiguation is done by taking news from local newspapers such as BeritaHarian, Utusan Malaysia and Harian Metro. A total of 19165 newspaper clips were used as a source of corpus development. From all of the news, there are 443 documents containing the word “Madu”, 3482 documents containing the word “Semak”, 2255 containing the word “Bekas”, 552 documents containing the word “Perang”, 90 documents containing the word “Pukul”, 5 documents containing the word “Kutu”, 8 documents containing the word "Haus", 41 documents containing the word "Rendang", 445 documents containing the word "Genting" and 847 documents containing the word "Daki". These words are also used to evaluate the accuracy of the developed algorithm because all of these words are polysemous words.

Then, all of these documents are imported into the MySQL (Structured Query Language) database for subsequent processes. All of the documents are grouped into several groups according to the polysemous word that contain in the documents. For example, documents containing the word "Semak" are grouped together with other documents that have the similar polysemous word while documents containing the word "Madu" are also grouped together with documents containing the word "Madu". The same goes for other words.

B. Identify the ambiguous word

In this step, documents are analyzed to make sure they contain the targeted ambiguous word. The algorithm analyzes one ambiguous word at a time and only documents that contain the targeted ambiguous word can be in the specific database.

C. Identify the collocation

Once the corpus content is ready, the next step is to identify the collocation for each of the ambiguous word. Collocation can provide a guide for identifying the right sense of the ambiguous word. Collocation is a word that is always associated with it [1].

Processes to identify collocation consist of several steps which are:

1. Tokenization
2. Token intersection
3. Calculate standard deviation
4. Collocation list

The process to identify collocation begins with the tokenization process, which is to index all the words contained in a document in a corpus collection. The result of this process is as shown in Table 2.

Table 2: Tokenization table.

Id	Document Id	Token
6245013	11973	timbul
6235457	11945	timbul
6205349	11838	timbul
6211979	11861	timbul

After the tokenization process is done, next is to identify each token that intersect with each other. It is called token intersect. Referring to Yarowsky algorithm, terms that always occur together can be a good hint to determine the definition of the term. In other words, it can provide good hint to identify each word and also acts as an attribute of the word. The result of this step is shown in Table 3.

Table 3: Token Intersect.

Id Document	Id Token	Token	Id Token Intersect	Token Intersect
1	1	Rumah	2	dipenuhi
1	1	Rumah	3	semak
1	1	Rumah	4	samun
1	2	Dipenuhi	1	rumah
1	2	Dipenuhi	3	semak

Once information of token intersect is available, the collocation of an ambiguous word can be determined by obtaining the highest frequency of token intersect with the ambiguous word and also the standard deviation of each word [7]. The formula for calculating standard deviations is as follows:

$$\sigma = \sqrt[2]{\frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}} \qquad (1)$$

where:
n is the frequency of two words present together.
d_i is the distance between the two words to appear together with i.
μ is mean.

The next step is collocation list. In this step, the collocation is calculated separately according to the collocation distance by the targeted ambiguous word. For example, a collocation with a distance value of -1 is different from a collocation with a distance value of 1. The position of a word can also influence the definition of the word [16].

D. Mapping Collocation to Definition

Once a list of collocations has been prepared, the next step is to determine the definitions represented by each collocation. This can be done by two steps which are expanding the collocation and calculate and compare the similarity with the definition.

Collocation can be expanded with three additional information which are the tokens intersect with the collocation, synonym from Malay WordNet and result from Google search engine.

After collocation is expanded, the next step is to calculate the similarity and compare to the definition. Collocation will be mapped to the definition that gives highest similarity. Similarity is calculated using cosine similarity and “beg of word” model. Below is the similarity formula:

$$\cos\theta = \frac{def_j\,.k}{\|def_j\|\|k\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,k}^2}} \qquad (2)$$

where
$def_j\,.k$ is the intersection between definition (def_j) and colocation (k).
$\|def_j\|$ is the norm for vectors def_j.
$\|k\|$ is the norm for vectors k.

E. Search for Meaning

Once all of the above steps have been completed, the algorithm is complete and ready to determine the definition of a polysemy word by looking at the collocations that appear in the search. However additional step is added to improve the accuracy of this algorithm which is by referring to Malay language dictionary using ambiguous term, word before and word after ambiguous term as query. If the query exists in the dictionary, the definition returned by the dictionary will be selected as the definition of the ambiguous term. If not, the definition will be determined by collocation existed in the query.

IV. RESULTS AND DISCUSSIONS

Accuracy test is a series of tests that are conducted to test the accuracy of Malay word sense disambiguation. The test consists of comparing the accuracy of Malay word sense disambiguation algorithm with the real results provided by experts and then compares the accuracy of the previous algorithm.

To conduct this test, a test collection containing 500 documents of the Malay language has been marked with senses in advance. Ten ambiguous terms have been used to test the accuracy of the proposed algorithm. The selected ambiguous terms are “Semak”, “Perang”, “Madu”, “Daki”, “Pukul”, “Kutu”, “Bekas”, “Haus”, “Rendang” and “Genting”.

In order to compare with the work of Lesk and Yarowsky, Lesk algorithm and Yarowsky algorithm have been developed. All the sentences in the test collection have been marked by a linguist to identify the true meaning of the ambiguous word contained in the sentence based on the context of the sentence. The detail results of the evaluation are as shown in Figs. 1-10.

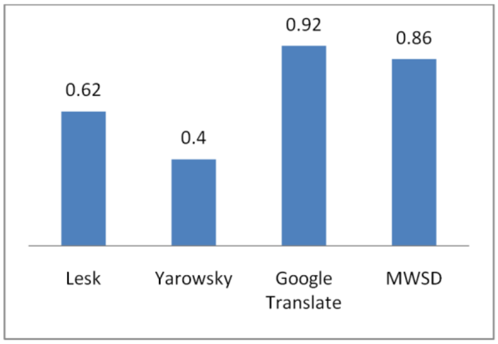

Fig. 1. Result accuracy for term “Semak”.

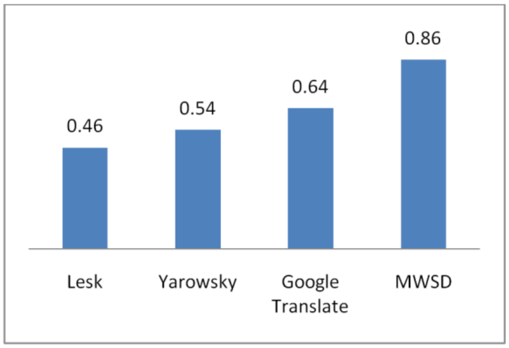

Fig. 2. Result accuracy for term “Perang”.

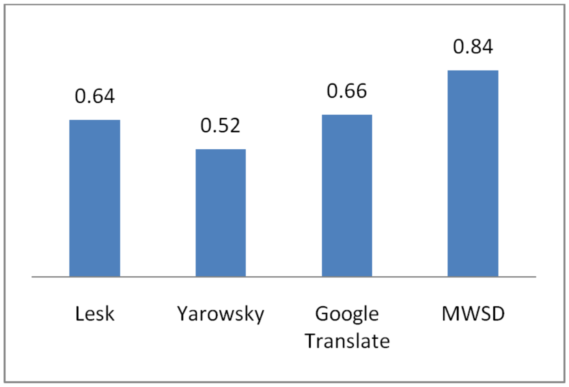

Fig. 3. Result accuracy for term “Madu”.

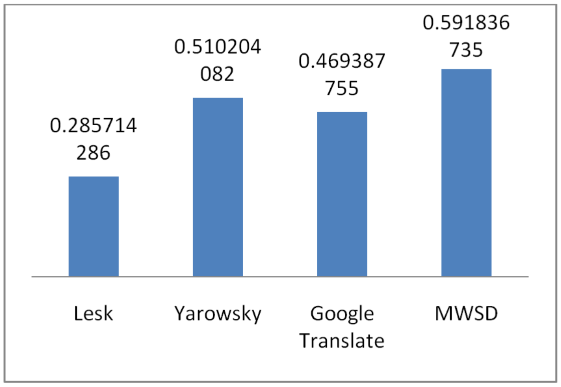

Fig. 4. Result accuracy for term “Daki”.

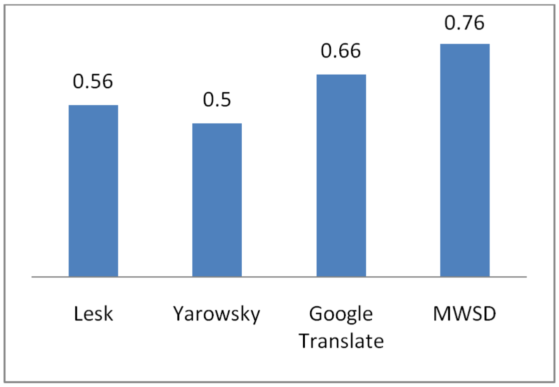

Fig. 5. Result accuracy for term “Pukul”.

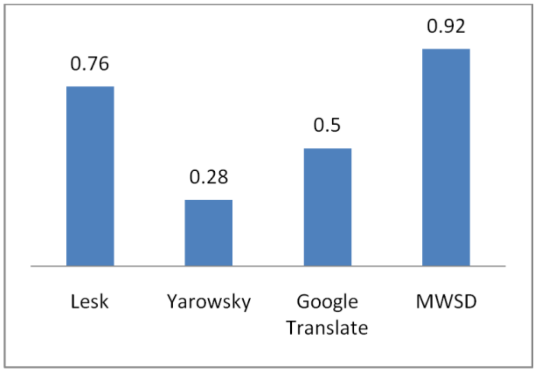

Fig. 6. Result accuracy for term “Kutu”.

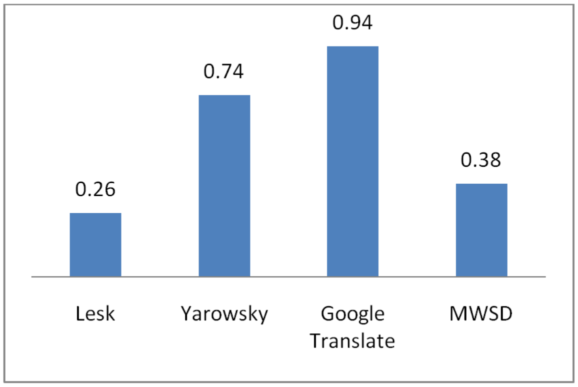

Fig. 7. Result accuracy for term “Bekas”.

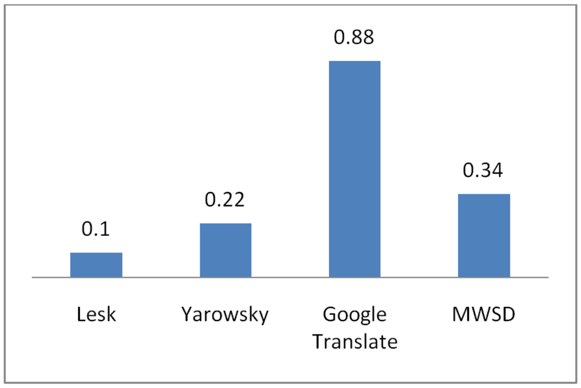

Fig. 8. Result accuracy for term “Haus”.

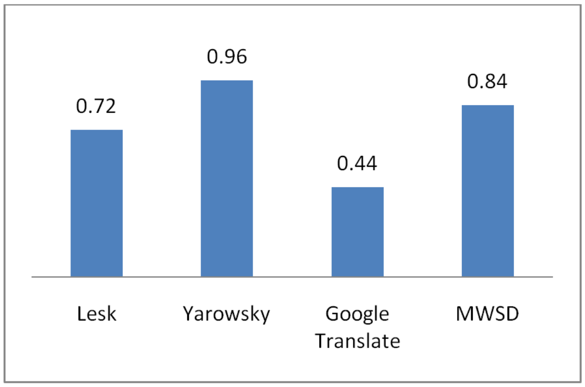

Fig. 9. Result accuracy for term “Rendang”.

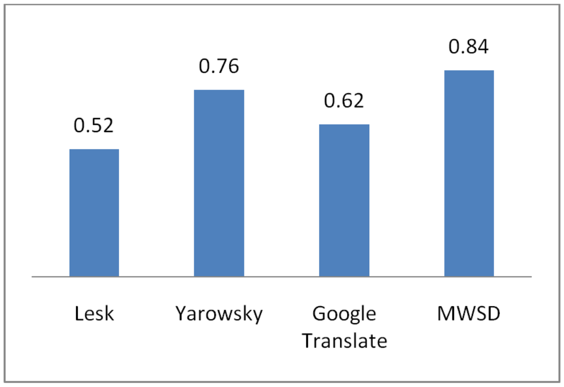

Fig. 10. Result accuracy for term “Genting”.

Six (6) out of ten (10) tests recorded the MWSD algorithm has the highest accuracy compared to the other three algorithms for words "Perang", "Madu", "Daki", "Pukul", "Kutu" and "Genting". The results of the algorithm's accuracy assessment are as shown in Fig. 11.

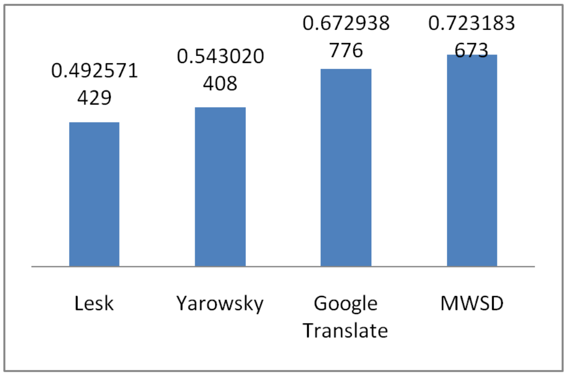

Fig. 11. Average Result Accuracy.

The results of the experiments show that the proposed algorithm outperforms other algorithms with an average 0.723183673 compared to Google Translate 0.672938776, Yarowsky 0.543020408 and the Lesk algorithm 0.492571429.

There are several factors that influence this result. Lack of corpus which serve as an unregulated source of knowledge for the MWSD algorithm is one of the factor. Due to this factor, MWSD algorithmwill poorly recognize collocation words.

In addition, there are also words that do not have a clear word collocation due to the nature of the word that can be matched with many words such as the word "Bekas", i.e"Bekastentera", "Bekasmakanan", "Bekasminuman", "Bekaspelajar", "Bekaspenagih", "Bekaspesakit" and "Bekasluka". Due to this, the MWSD algorithm that relies on word collocation as a guide to determine the meaning of a word becomes less effective in determining the true meaning of the word. However, this problem can be solved by having a larger corpus source. Probability of the correct word to be a collocation for the polysemous wordswill be higher if it has more corpus resources.

V. CONCLUSION

This study aim to compare existing algorithms of word sense disambiguation. This study also improvesthe method of identifying the exact meaning of ambiguousMalay words.A prototype has been developed to test the accuracy of this algorithm by comparing with three other algorithms namely Lesk, Yarowsky and Google Translate algorithms.

Based on the experiment that has been conducted using 10 ambiguous words and the result outperform other three algorithms namely Lesk, Yarowsky and Google Translate withaverage accuracy of 0.723183673.

VI. FUTURE SCOPE

For future undertakings, this study can be enhanced with the following steps.

a) Connect and expand each neighboring word ontology to enable the context of a sentence that can be identified more accurately and efficiently.
b) Consider the word type and grammar aspects of each word adjacent to the polysemous word.
c) Use a database software to manage words index and speed up retrieval of data.

REFERENCES

[1]. Abed, S. A., Tiun, S., & Omar, N. (2016). Word sense disambiguation in evolutionary manner. *Connection Science*, *0091*: 0–16.

[2]. AL-Saiagh, W., Tiun, S., AL-Saffar, A., Awang, S., & Al-Khaleefa, A. S. (2018). Word sense disambiguation using hybrid swarm intelligence approach. *PLoS ONE*, *13*(12).

[3]. Aziz, M. J. A., Ahmad, F., Ghani, A. A. A., & Mahmood, R. (2008). Pola Grammar For Automated Marking Malay Short Answer Essay-Type Examination, PhD thesis, Universiti Putra Malaysia.

[4]. Baskaya, O., & Jurgens, D. (2016). Semi-supervised learning with induced word senses for state of the art word sense disambiguation. *Journal of Artificial Intelligence Research*, *55*: 1025–1058.

[5]. Chaplot, D. S., & Salakhutdinov, R. (2018). Knowledge-based word sense disambiguation using topic models. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*: 5062–5069.

[6]. Chifu, A.-G., Hristea, F., Mothe, J., & Popescu, M. (2015). Word sense discrimination in information retrieval: A spectral clustering-based approach. *Information Processing And Management*, *51*(2): 16–31.

[7]. Manning, C., & Schütze, H. (1999). Collocations. In *Foundations of Statistical Natural Language Processing*:151–189.

[8]. Mihalcea, R. (2010). Word sense disambiguation. *Encyclopedia of Machine Learning*: 1027–1030.

[9]. Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, *41*, 10:1–10:69.

[10]. Ranjan Pal, A., & Saha, D. (2015). Word Sense Disambiguation: a Survey. *International Journal of Control Theory and Computer Modeling (IJCTCM)*, *5*(3): 1–16.

[11]. Riahi, N., & Sedghi, F. (2016). Improving the Collocation Extraction Method Using an Untagged Corpus for Persian Word Sense Disambiguation, *Journal of Computer and Comunications, 4*: 109–124.

[12]. Sazali, S. S., Abu Bakar, Z., & Jaafar, J. (2016). Word Prediction Algorithm in Resolving Ambiguity in Malay Text. *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*:1347–1352.

[13]. Yahaya, F., Rahman, N. A., & Bakar, Z. A. (2011). Resolving Malay Word Sense Disambiguation Utilizing Cross-Language Learning Sources Approach Conference. *Advanced Science Letters*, *4*(2): 400–407.

[14]. Yamaki, S., Shinnou, H., Komiya, K., & Sasaki, M. (2016). Supervised word sense disambiguation with sentences similarities from context word embeddings. *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 2016*, (Paclic 30): 115–121.

[15]. Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*: 189–196.

[16]. Yarowsky, D. (2007). Decision lists for lexical ambiguity resolution. *Proceedings of the 32nd annual meating on Association for Computational Linguistics (ACL '94):* 88–95.

[17]. Yuan, D., Doherty, R., Richardson, J., Evans, C., & Altendorf, E. (2016). Semi-supervised Word Sense Disambiguation with Neural Models. *Arxiv*: 1374-1385.

[18]. Zakaria, T. N. T., Aziz, M. J. A., Mokhtar, M. R., & Darus, S. (2020). Semantic similarity measurement for Malay words using WordNet Bahasa and Wikipedia Bahasa Melayu: issues and proposed solutions. *International Journal of Software Engineering and Computer Systems*, *6*(1): 25–40.

[19]. Zakree, M., Nazri, A., Shamsudin, S. M., & Bakar, A. A. (2008). An Exploratory Study of the Malay Text Processing Tools in Ontology Learning. *ISDA '08 Proceedings of the 2008 Eighth International Conference on Intelligent Systems Design and Applications*: 375–380.

[20]. Zhan, J., & Chen, Y. (2011). Research on Word Sense Disambiguation. *Advanced Materials Research*: 181–182.

How to cite this article: Hamzah, M. P. and Rifin, M. A. S. M. (2020). Improved Algorithm for Malay Word Sense Disambiguation. *International Journal on Emerging Technologies, 11*(5): 346–351.