INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING & COMPUTER SYSTEMS (IJSECS)
ISSN: 2289-8522 e-ISSN: 2180-0650
VOL. 6, ISSUE 1, 25 – 40
DOI: https://doi.org/10.15282/ijsecs.6.1.2020.4.0067

ORIGINAL ARTICLE

Semantic similarity measurement for Malay words using WordNet Bahasa and Wikipedia Bahasa Melayu: issues and proposed solutions

T.N. Tuan Zakaria[1,*], M.J. Ab Aziz[1], M.R. Mokhtar[1] and S. Darus[2]

[1]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Selangor, Malaysia.
[2]Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, 43600 Selangor, Malaysia.

ABSTRACT – Semantic similarity between words is a very important task and widely practiced in the field of natural language processing. Knowledge-based lexical resources like WordNet and Wikipedia are useful for this task. WordNet Bahasa (WB) and Wikipedia Bahasa Melayu (WikiBM) are the example of lexical resources for Malay language. However, these lexical resources are still ongoing and limited semantic information. This paper aims to discuss some issues regarding semantic similarity for Malay language, propose a framework using WB and WikiBM, and evaluate the performance of both. An experiment was done using 150 Malay translated words (75 word-pairs). The result showed that the WB and WikiBM are capable to be adapted to literature techniques. For WB, we tested the coverage of WB based on three word-levels (stem, root and mix level) to find the most applicable word level as our dataset. The test indicated that the mix level (86.7%) outperformed the stem (78.7%) and root level (68.0%). For WikiBM, we evaluated the coverage of three main features in its article (gloss definitions, hyperlinks and categories) where these features are important in some previous techniques. The results of the experiment revealed that the gloss definition gave full coverage (100%) for our 75 word-pairs input compared to hyperlinks and categories (88.0%).

ARTICLE HISTORY
Received: 6 Feb 2020
Accepted: 27 May 2020

KEYWORDS
Malay, semantic similarity, WordNet Bahasa, Wikipedia Bahasa Melayu, language processing, explicit semantic analysis

INTRODUCTION

Human language is very complicated and has rich information that is often misleading. Some information can be presented with different words and structures. A word may have different meanings, yet these different meaning can be presented in different ways. Hence, the ambiguity and diversity of the nature of human language are often a barrier that separates human understanding with computer comprehension. This problem needs to be resolved before the language processing tasks can be run. There are various sources such as language tools that have been developed to overcome this problem. However, the source of language tools for Malay language processing is still limited and it has become a key constraint in Malay language processing studies.

Knowledge based lexical source such as WordNet are very useful in a variety of language processing applications such as semantic search (Ma et al., 2016; Liu et al., 2017) and word sense disambiguation (Chen et al., 2014; Moro et al., 2018). WordNet has features such as semantic relationships (hypernyms, hyponyms, synonyms and so on) and definition (gloss) that can be used for semantic relatedness studies. Semantic information derived from WordNet can form a taxonomy, which can be used in semantic measurements between words (Wu & Palmer, 1994; Zhou et al., 2008; Sanchez et al., 2012). The top of WordNet taxonomy is a root called 'entity'. While the gloss feature can be used to measure the semantic relatedness by looking at the overlapping of definition of two words, such as the studies conducted by (Lesk, 1986; Banarjee & Pedersen, 2003; Ponzetto et al., 2010).

Despite of its popularity, WordNet lacks useful information for language processing (Fernando et al., 2012). For example, WordNet cannot connect related words based on topics like 'drivers' and 'car'. In addition, WordNet is only maintained by a small group of experts and makes it less useful for large-scale studies (Agirre & Edmonds, 2006; Meyer & Gurevych, 2010).

This problem leads to the use of other lexical sources like Wikipedia. Wikipedia has the added advantage of the latest information that is frequently updated by millions of volunteers Just like WordNet, Wikipedia also has been developed in several languages and is available online for free. Wikipedia contains information such as semantic information, intentions of words and extraction of information. There are three main features in Wikipedia that are often used in research: hyperlinks, categories and definitions (gloss). The definitions in Wikipedia is similar to the gloss in WordNet while the categories can form taxonomy such as hypernym relations on WordNet. The top of taxonomy in Wikipedia is a root called 'content'. Several studies have used Wikipedia as the lexical source such as (Chen et al., 2017; McCrae, 2018).

*CORRESPONDING AUTHOR | T.N. Tuan Zakaria | ✉ tn_hafizah@yahoo.com

© UMP Press

Extensive developments in the use of WordNet and Wikipedia have motivated many researchers to study these two lexical sources. They conduct research to map between WordNet's synset with articles in Wikipedia. Although both of these lexical sources perform well, the ambiguity problem still exists. This is because human language has many different meanings and structures. In WordNet, a word may have different senses. For example, the word "web" has 7 meanings: (i) "an intricate network suggesting something that was formed by weaving or interweaving"; (ii) "an intricate trap that entangles or ensnares its victim"; (iii) "the flattened web like part of a feather consisting of a series of barbs on either side of the shaft". Words with many of these meanings need to be matched first with the proper meaning before language processing applications can be developed. Wikipedia also contains the same problem of simplicity in which there are words that give a few articles output. For example, the word "web" also provides two article output options in Wikipedia: (i). "Spider web: a silken structure created by the animal" and (ii) "World Wide Web or the Web: an Internet-based hypertext system". Before the language task is resumed, this ambiguity needs to be addressed with certain techniques to match the input word to the correct article.

Due to its excellent performance, WordNet has grown and developed into other languages such as Arabic, Chinese, Malay and so on. WordNet Bahasa (WB) is a lexical source of innovation from English WordNet used for Malay and Bahasa Indonesia, while the Wikipedia Bahasa Melayu (WikiBM) is an innovation developed from English Wikipedia. Good performance has been shown by semantic relatedness study between words for other languages such as Arabic (Almaayah et al., 2014; Mahyoub et al., 2014; Alkhatib et al., 2017; Saif et al., 2018), Thai (Na Chai et al., 2017; Netisopakul & Thong-Iad, 2019) and Chinese (Zhang, 2013; Lee & Hsieh, 2015). However, in Malay, studies on the semantic relatedness between words using WordNet Bahasa and Wikipedia Malay are still lacking. This is due to the lack of resources such as tools for processing Malay language. The Malay lexical resources such as WB are still ongoing and limited of semantic information. The coverage of WB is limited especially for compound words such as 'hamba abdi' and 'tanah hutan'. As for WikiBM, the articles were written by inexpert authors. The texts are plain, long and unstructured. These affects the performance of similarity measurements. Some methods need to be found to reduce the information to increase measurement.

This paper aims to discuss some issues on Malay semantic similarity measurements, propose a framework to solve these issues and evaluate the performance of two Malay lexical resources, namely WordNet Bahasa (WB) and Wikipedia Bahasa Melayu (WikiBM). This paper highlights the related works on semantic similarity measurements, introduction to Malay lexical resources (WB and WikiBM), issues on semantic similarity measurement for Malay words and the proposed solutions for Malay semantic similarity. In order to test the Malay lexical resources used in this research, a brief experiment was done, and the analysis of the experiments were presented in this paper. Finally, the conclusion and future work will be discussed in the end of this paper.

RELATED WORK

Previous researchers have suggested several measurement techniques such as path-based, intrinsic information content and gloss-based technique to measure semantic similarity between two words. Path-based measure is a similarity method to measure the distance between the nodes of two concepts in the semantic taxonomy. Examples of measurements using this method are Sim_{Rada} (Rada et al., 1989), Sim_{WP} (Wu & Palmer, 1994) and Sim_{LCH} (Leacock et al., 1998).

Rada et al. calculate the length of the nearest path between the nodes of the two concepts, c1 and c2. Wu and Palmer improve the path-based measurement by adding depth feature in their formula. They also introduce 'least common superconcept' (LCS) in their measurement (the meeting point between concept 1 and concept 2 in semantic taxonomy). Meanwhile, Leacock et al. take into account the minimum number of edges that separate two nodes and the taxonomic depth between two words in their measure.

Information content is an important dimension of knowledge when measuring the relation between two terms or the meaning of the word (sense). The conventional method for measuring the content of this information is to combine the knowledge of hierarchical structure of each word on ontology such as WordNet. There are several studies using Intrinsic Information Content (IIC) such as (Seco et al., 2004; Zhou et al., 2008; Sanchez et al., 2011). Measurement based on the gloss in WordNet has been done by researchers such as (Lesk, 1986; Banarjee & Pedersen, 2003; Strube & Ponzetta, 2006). Lesk measures the similarity between two words by calculating the length of concept definition 1 and concept definition 2. Banarjee and Pederson measured the similarity by calculating the overlap of definitions of two concepts (concept A and concept B). Whereas, the measure by Strube and Ponzetta considered the overlaps of the definition and length of definition of both concepts.

There are several techniques for mapping WordNet synsets to Wikipedia articles. The Explicit Semantic Analysis (ESA) method, proposed by (Gabrilovich & Markovitch, 2009) is a vector space model to interpret the semantic words based on knowledge encoded in Wikipedia articles. The meaning of the word can be expressed by the explicit concepts in the language resources. The tf-idf function (word frequency with the inverse document frequency) is used in this method. This method uses the gloss definition in Wikipedia articles and it can also construct the concept vector of the text fragments (sentence or paragraph). The experimental results from previous researches indicated that ESA outperforms the state-of-the-art methods in assessing semantic relatedness of words and texts.

Hassan and Mihalcea (2009) was proposed salient semantic analysis (SSA). This method represented the meaning of a word by using the hypertexts over Wikipedia articles. Unlike the ESA model that depends on the text content of articles, SSA vector is created according to salient concepts (hyperlinks) that are found in the same contexts of the given word.

Each element in the representation vector is calculated using Point-wise Mutual Information of the associated salient concept and the word. Category Semantic Depiction (CSD) method was proposed by Taieb et al. (2011). They proposed CSD to represent the semantic of words over Wikipedia categories instead of the articles. Each category in Wikipedia article is represented as the vector containing the terms. In this method, each article is represented as the set of extracting terms with their frequencies in this article after the stemming and stop words removal tasks. The new similarity measure was proposed to estimate the similarity between two categories from their representation vectors of concept 1 and concept 2.

INTRODUCTION TO WORDNET BAHASA AND WIKIPEDIA BAHASA MELAYU

WordNet Bahasa (http://wn-msa.sourceforge.net/) built in 2011, is a source of lexical knowledge built up in Malay and Indonesian language. This electronic dictionary is inspired by Princeton WordNet, an English cyber dictionary and the Global WordNet Grid. WordNet Bahasa contains 49,668 synsets, 145,696 senses and 64,431 different words.

WordNet Bahasa has several types of lexical relationships such as WordNet English (WN). Among the lexical relationships found in WordNet Language are hypernym relations, hyponyms, domains and so on. This synset relationship is connected with synset relationships in English WordNet.

Semantic information provided by WordNet Bahasa can be presented in semantic taxonomic form. Hypernym relations (IS-A) are used to generate semantic taxonomy. Each taxonomy ends with the root of 'entity'. The path between the concepts in the taxonomy is known as a node. Figure 1 shows the taxonomy for the example of the word 'kereta'.

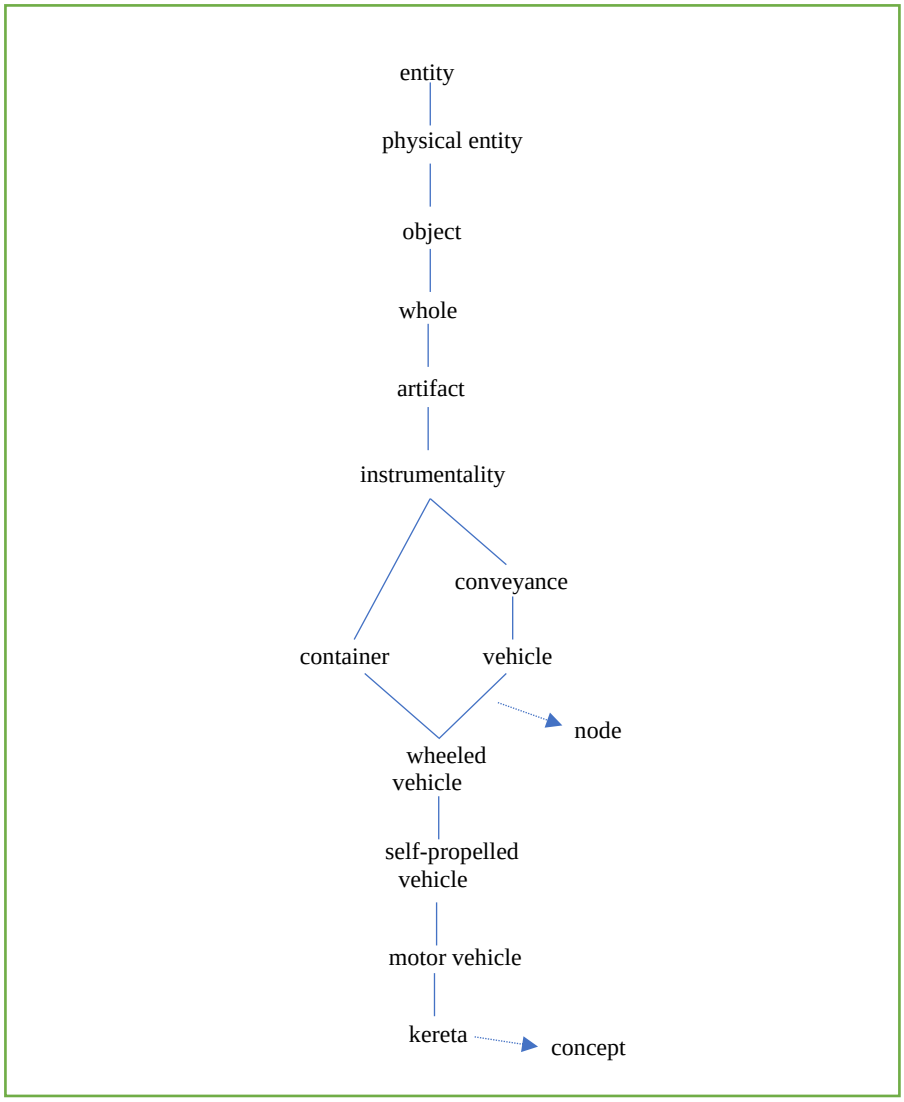

Figure 1. Semantic taxonomy of word ‘kereta’ (car)

In a semantic taxonomy, a concept is linked to another concept by a node. For example, a concept 'kereta' (car) is linked to another concept 'instrumentality' by two paths: (path 1: instrumentality – container – wheeled vehicle – self-propelled vehicle – motor vehicle – car) and (path 2: instrumentality – conveyance – vehicle – wheeled vehicle – self-propelled vehicle – motor vehicle – car). These paths are important in measuring the similarity value of two concepts especially using path-based measures.

From this taxonomy, there are three important features that can be used in path-based measures; depth, distance and LCS (least common subsume). [Rada et al., 1989] calculate the semantic similarity by using the distance between the nodes in semantic taxonomy. They calculated the shortest path that connected two concepts. For example, the distance of path 1 for concept 'kereta' (car) to 'instrumentality' is 5 while the distance of path 2 is 6. Hence, path 1 was chosen for this measure.

Wu and Palmer (1994) added depth in their proposed measurement. They indicated that the more deeper a concept located in a taxonomy, the concept are more specific. For example, the depth of the concept 'instrumentality' is 6. Least common subsume (LCS) is the most specific common ancestor of two concepts found in a given semantic taxonomy. It represents the commonality shared between two concepts. For example, the LCS of concept 'container' and 'vehicle' in Figure 1 is 'instrumentality' with depth value is 6. The depth of LCS is calculated to be used in semantic similarity measure such as measure by Wu & Palmer (1994).

Wikipedia Bahasa Melayu (https://ms.wikipedia.org/wiki/Laman_Utama) is an online encyclopedia for Malay version. Currently (2019) there are 326,477 articles in the Wikipedia Bahasa Melayu.

ISSUES ON SEMANTIC SIMILARITY MEASUREMENT FOR MALAY WORDS

The similarity of words is widely used in word processing applications such as text classification and machine translation. The semantic similarity based on semantic taxonomy has been widely used by previous researchers especially in English semantics. This semantic taxonomy is derived from a popular lexical database called WordNet. Good performance by WordNet has made other countries develop version of WordNet in different languages such as Arabic WordNet (Arabic language) and WordNet Bahasa (Malay and Indonesian language). The arising of similarity measurement studies leads to the use of lexical resources other than WordNet, which is Wikipedia. Wikipedia is used because the information obtained is larger than WordNet. However, the similarity issues are different between languages as every language has its own structures, including Malay language. The semantic similarity is often discussed based on three main issues: semantic representation, semantic measurements, and mapping a synonym set from WordNet to Wikipedia article.

1. Semantic representation

The problem of semantic representation arises when how can the lexical information of a word be delegated to a language that a computer can understand. In Wikipedia, gloss is often used. This text consists of various words with different types of words. In addition, the Wikipedia text is written by voluntary writers who make texts in Wikipedia unstructured. There are many unimportant words in gloss. The problem is how to reduce the gloss information from Wikipedia to a more concise but more valuable vectors. Various studies have been conducted and various techniques have been proposed such as grouping of texts, bag of words and so on (Zhang et al., 2014; Saif et al., 2016; Wu et al., 2017).

2. Semantic Similarity Measurement

Various techniques have been developed to measure the semantic similarity between two words using WordNet and Wikipedia. Three main methods previously used were path-based (Rada et al., 1989; Wu & Palmer, 1994; Leacock et al., 1998), information-based (Seco et al., 2004; Zhou et al., 2008; Sanchez et al., 2011). and gloss-based (Lesk, 1986; Banarjee & Pedersen, 2003; Strube & Ponzetta, 2006). However, fewer studies have been conducted by using other language lexical sources especially Malay (WordNet Bahasa and Wikipedia Bahasa Melayu). WordNet Bahasa and Wikipedia Bahasa Melayu are inspired by English WordNet and English Wikipedia, but the information sources are still limited.

WordNet was developed to represent its core unit, which is a synonym set (synset). This synset is divided into four words classes namely nouns, verbs, adjectives and adverbs. WordNet works by organizing words (inputs) into synsets, then providing a brief definition and usage example. It also records some relations between related synsets. Each synset is linked to another set through several relations such as hyponym, hypernym, meronym and holonym. WordNet Bahasa is inspired by English WordNet but with some different features. In English WordNet, there is a feature called 'inherited hypernyms' that list all hypernyms contained in the set. This feature is very helpful in generating a semantic taxonomy for a synset. However, this feature is not available in WordNet Bahasa. The list of hypernyms can be obtained by clicking on the 'hypernym' feature one by one until arrive at the 'entity' level, which is the root of the taxonomy in WordNet.

Malay is a language rich in morphology (word formation process) in which a root word could be transformed to other word through several morphological process such as affixation, compounding and duplication. In this study, we use English to Malay translated dataset because there is no baseline Malay dataset that can be used for semantic similarity study. Problem arises when the word of the English dataset is translated into Malay need to be matched with the lemma in the WordNet Bahasa. There are some compound words that cannot be processed by WordNet Bahasa. For example,

the word ‘woodland’ when translated into Malay will become ‘tanah hutan’ which is known as a compound word. WordNet Bahasa could not provide the proper output for this compound word as it does not match the lemma in WordNet Bahasa. A lemmatization task is required to solve this problem.

The semantic information could also be represented by using the definition (called as ‘gloss’) in WordNet. However, the gloss in WordNet Bahasa is written in English. In order to use the gloss as our data for our research, we need to translate the gloss to Malay language. This limitation is time consuming process.

Therefore, research has to be done to adapt and evaluate the effectiveness of previous techniques and measures for Malay semantic information using Malay lexical sources such as WordNet Bahasa and Wikipedia Bahasa Melayu. The best technique for Malay words needs to be determined.

3. Mapping concept in WordNet to the correct article in Wikipedia

Wikipedia works by matching the term to one or more related articles. Wikipedia uses a search engine and a search box. The input in the search box is called a search string. This string will go through a stem matching process in which the input will be parsed into several strings called the ‘original string’ and ‘stemmed string’. For example, ‘cloud’ input will be parsed to some strings such as ‘clouds’, ‘clouded’ and ‘clouding’. Then, the title matching technique will be performed where there are two possibilities: title match and non-match title. If there is a title match, the term will be matched to two possibilities: direct match or redirect match. Redirect match is usually matched to a title that has the same meaning as the string’s input. For example, direct math will be matching the term ‘grave’ into article entitled ‘grave’ while redirect match will match the term ‘graveyard’ to the article entitled ‘cemetery’. If there is no title that matches with the input term, Wikipedia will display the ‘search results’ page. These results provide a number of possible articles (called candidate articles). To match the term with the right article in Wikipedia, a technique needs to be used.

Several article matching techniques have been proposed by previous researchers such as Explicit Semantic Analysis (ESA), Salient Semantic Analysis (SSA) and Category Semantic Depiction (CSD). ESA uses the text features in Wikipedia’s article to match the terms, while SSA uses hyperlinks and CSD uses category features. However, because Wikipedia Bahasa Melayu is still developing and lack semantic information, there are some articles that do not have the category and hyperlinks features. As a result, ESA is a technique that can be used for Wikipedia Bahasa Melayu. In ESA, semantic representation depends on the existence of words in the text collection. ESA uses the tf-idf function where it compares the terms to the exact words in the text. The text in Wikipedia article is also long and unstructured. This makes the ESA high dimensional.

Malay morphological processes also affect the tf-idf function in the matching of term-article. This is because tf-idf cannot match terms with words in the text that have the same meaning although it has gone through the morphological process. For example:

Term: ‘perkuburan’ (graveyard)

Text in the article entitled ‘kubur’ (cemetery): Kubur adalah tempat di mana jasad, biasanya manusia ditanam. Selain jasad, kubur biasanya mengandungi objek yang menjadi petunjuk penting bagi menentukan corak kehidupan dan kebudayaan bangsa yang dikuburkan itu. Kubur merupakan sumber maklumat kebanyakan kebudayaan silam yang penting disebabkan bukti-bukti lain telah lenyap ditelan zaman.

Figure 2. Example term and article retrieved from Wikipedia Bahasa Melayu

From the example term and article text in Figure 2 above, tf-idf cannot match ‘perkuburan’ with ‘kubur’ although these two words are of the same root word ‘kubur’. This gives the article a low tf-idf value. Therefore, one method needs to be studied to reduce the dimensions of ESA technique. One of the potential methods that can be used is pre-processing method such as stemming.

The next important step in the semantic similarity using WordNet and Wikipedia is the alignment of concepts from WordNet to articles in Wikipedia. Previous studies have shown that similarities between a set of synonyms with one article can be calculated using two main techniques: overlap-based or WordNet-based (through semantic relations in taxonomy). Overlapping techniques depend on the number of words shared in the text representation between one set of synonyms and one article. However, Wikipedia’s text is still long, redundant and semantically insignificant. In order to reduce the length of the text to a more useful vector as a semantic representation, one method needs to be identified. Among these is the text clustering method.

PROPOSED SOLUTIONS FOR MALAY SEMANTIC SIMILARITY

The issues on semantic similarity measurements for Malay words have been discussed earlier in section “Issues On Semantic Similarity Measurement for Malay Words”. Based on these issues, we will propose a framework for our

semantic similarity measurement research. Our proposed framework consists of four main stage: problem identification, data collection, design and implementation and evaluation.

The first stage in our framework is problem identification. The main issues have been discussed in previous section in this article. Our main focus is on three main issues: representation of semantic information, adaptation of previous similarity technique and mapping synset from WordNet Bahasa to article in Wikipedia Bahasa Melayu. The main problem in our research is the lengthy text of gloss definition especially definition collected from WikiBM. It contains redundant and repeated words. our aim is to reduce the lengthy texts to more valuable semantic information. Our propose method for these three issues is discussed in Step 1, Step 2 and Step 3.

The second stage of our proposed framework is data collection. Since there is no available baseline Malay dataset for semantic similarity, English dataset is used. The data was translated into Malay language. The semantic information is collected from WordNet Bahasa and Wikipedia Bahasa Melayu. For WordNet Bahasa, the semantic information includes hypernyms, gloss and synsets. The hypernyms are collected to build semantic taxonomy for each input concept. Then, this taxonomy will be used to calculate the semantic similarity between two Malay words, together woth the gloss information. While the synsets will be used in Stage 3, in the process of mapping between WB synset to WikiBM article.

Next, our method that suits Malay language behaviors and requirements is designed and then implemented to Malay data. The design and implementation are based on our three main focuses which are the representation, adaptation and mapping. For semantic representation, we applied text clustering algorithm to reduce the long, plain gloss definition from Wikipedia to more valuable vectors. For adaptation, we adapted the semantic similarity techniques based on the semantic taxonomy such as path-based measurements, information contents and gloss-based. Firstly, we adapted semantic similarity measurement to WordNet Bahasa by generating a semantic taxonomy using hypernyms (is-a relation). Then, we adapted semantic similarity measurements to Wikipedia Bahasa by collecting the ‘category’ feature and built a semantic taxonomy based on the category of the article. For mapping, we used ESA (Explicit Semantic Analysis) technique to map a synset from WordNet Bahasa to a correct article in Wikipedia Bahasa Melayu. ESA technique will match the synset to article using term while the gloss similarity technique uses the overlaps of the text in gloss definition. The gloss similarity can improve mapping by calculating the similarity shared between gloss in WordNet Bahasa and gloss in the Wikipedia Bahasa Melayu.

Finally, several experiments are conducted to evaluate our methods. Our evaluation stage focuses on three main issues: the performance of reduced semantic representation, the correlation of semantic similarity for Malay words and the performance of our mapping method. For semantic similarity, we will use three type of measures; path-based, intrinsic information content and gloss-based. The correlation are compared to human score. For semantic representation, we evaluated the effect of text clustering to text similarity value between original texts and clustered text based on Malay part-of-speech (POS). while, for mapping process, we will evaluate the performance of the mapping process using ESA (tf-idf function) with original text, ESA with reduced semantic representation (RSR) and the combination of ESA and gloss similarity technique towards original text and RSR. Figure 3 presents our proposed framework. Step 1, 2 and 3 present the details of the steps in solving our three main issues.

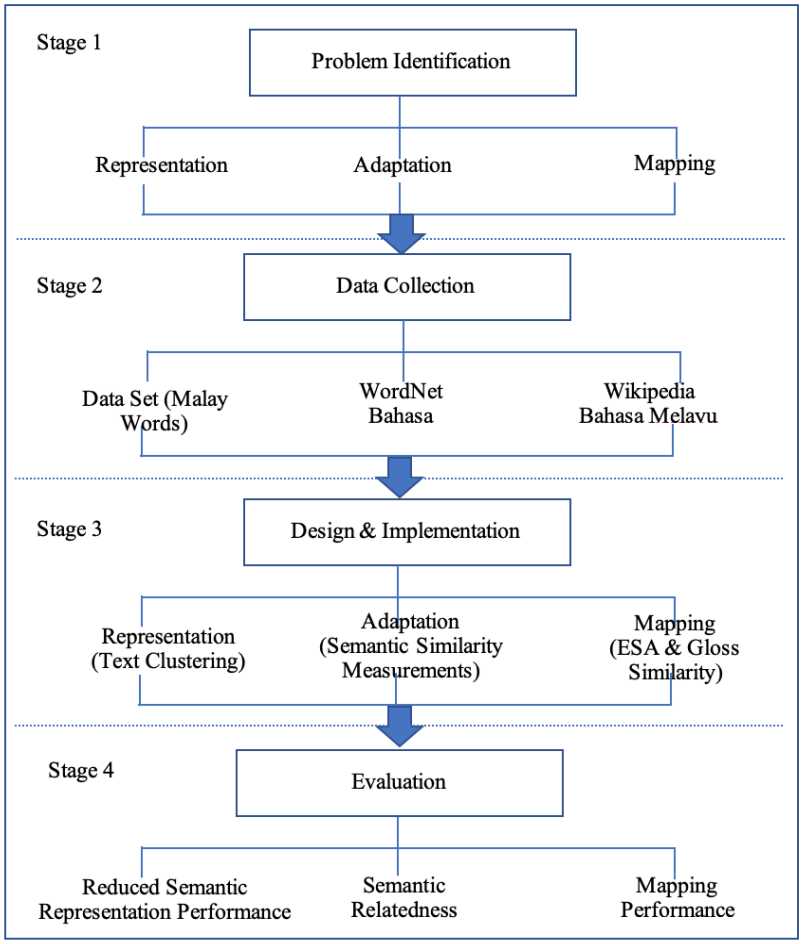

Figure 3. Proposed framework for Malay semantic similarity measurement

Step 1. Reducing the semantic representation

Wikipedia gloss contains plain and long texts. Many redundant and unnecessary words exist in the text. Several techniques were developed to reduce the information into valuable vectors. In this study, we propose text clustering method based on part of speech for our Malay data.

Text clustering is the application of cluster analysis to text-based documents. This method uses machine learning and natural language processing to understand and categorize unstructured textual data. Typically, text clustering can be done based on two methods which are entity recognition and part-of-speech tag. For our research, we chose text clustering based on part-of-speech as for Malay, the 'Kata Nama' (noun) and 'Kata Kerja' (verb) are the most important type of words in Malay language. In order to do this, some preprocessing need to be done. Figure 4 shows the steps in this text clustering method.

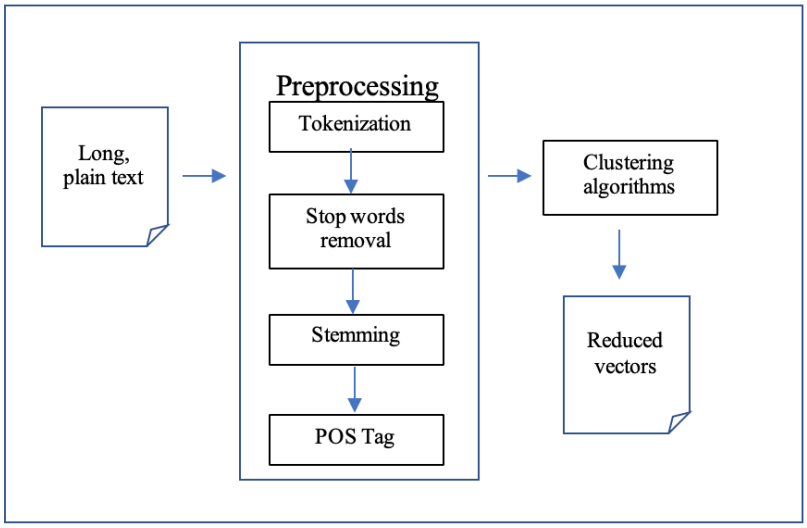

Figure 4. Steps in text clustering

The input of this step is a long and plain text. The input will go through preprocessing tasks such as tokenization, stop words removal, stemming and part-of-speech (POS) tag to reduce the semantic of the words. Then, these reduced words will be clustered using clustering algorithms. Finally, the output of this process is a reduced text which is shorter and contains more valuable semantic vectors.

Step 2. Adapting the semantic similarity measurement

In this research, we adapted the semantic measurement techniques from previous researches. The previous researches used English words as their dataset and using English WordNet and English Wikipedia as their lexical sources. However, for our research, we adapted some measurement techniques into Malay words as our dataset. For lexical sources, we used WordNet Bahasa and Wikipedia Bahasa Melayu. Figure 5 shows the semantic similarity measurements which were examined in our study.

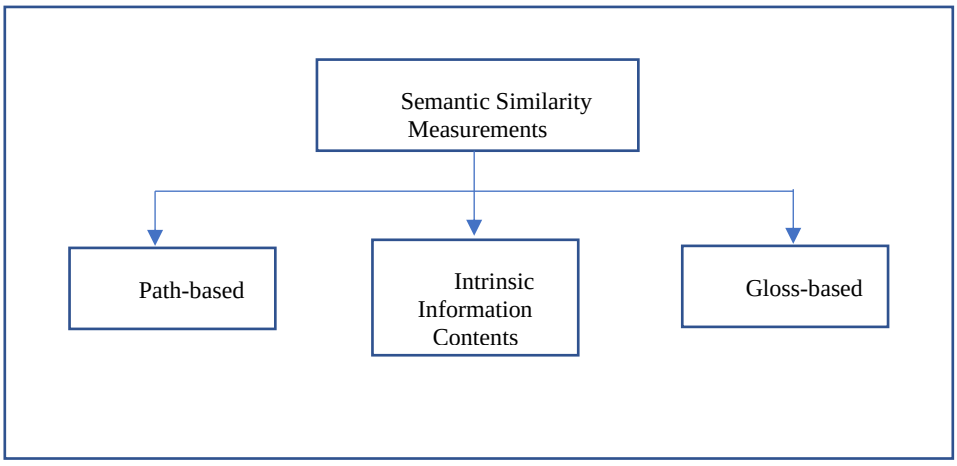

Figure 5. Semantic similarity measurements

In order to do this, the comparison between English WordNet, WordNet Bahasa, English Wikipedia and Wikipedia Bahasa Melayu were considered. For instance, we focused on path-based measurements, intrinsic information contents and gloss-based measurements. The path-based and intrinsic information contents are the measurements using semantic taxonomy while the gloss-based is using the gloss in lexical sources. These measurements need to be examined and the best measurement for Malay words will be selected.

Step 3. Mapping the WordNet synset to Wikipedia article

The final steps in this research was mapping the WordNet synset to Wikipedia article. The popular technique used is called “Explicit Semantic Analysis” or ESA and gloss similarity. According to Wikipedia, ESA is a vectoral representation of text (individual words or entire documents) that uses a document corpus as a knowledge base. A word

is represented as a column vector in tf-idf matrix of the text corpus while a document (string of words) is represented as the centroid of the vectors representing its words.

Gloss similarity is a text overlapping technique. The gloss of a word was collected in the WordNet Bahasa while the gloss definition of that word was collected from Wikipedia Bahasa Melayu. The overlapping of gloss in the WordNet Bahasa and Wikipedia Bahasa Melayu was calculated and the highest value of article candidates was selected as the best mapping article. The detailed algorithm for this technique will be described later in another article. Figure 6 shows the process in the mapping technique.

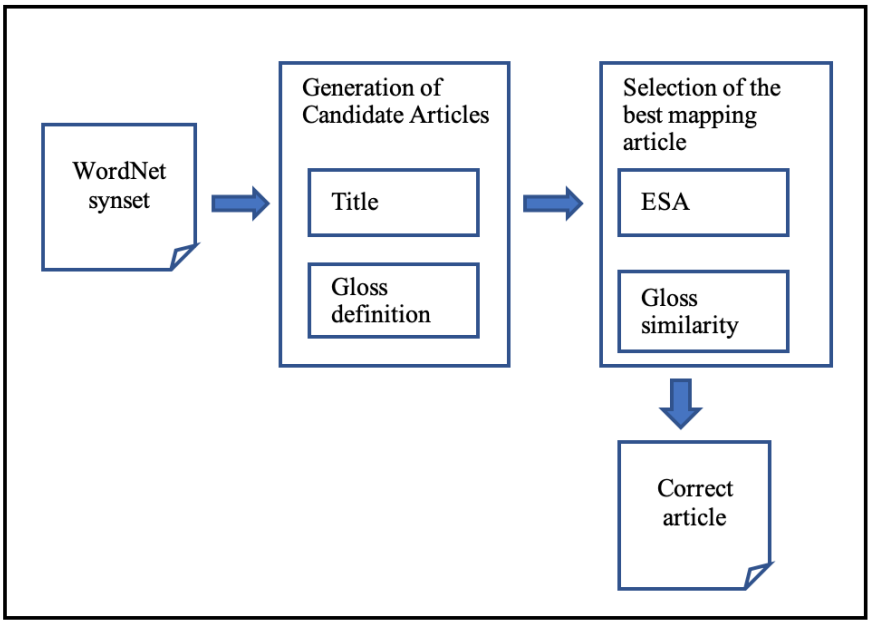

Figure 6. Mapping process

The mapping process started with the collection of WordNet synsets from WordNet Bahasa. Using the synsets, we will generate the candidate articles based on the title of the article in Wikipedia Bahasa Melayu. The title and gloss definition of each articles will be collected. These articles will be assigned as 'candidate article'. Then, selection of the best mapping article needs to be conducted. The best article will be selected using ESA and gloss similarity methods. The ESA method will be done using tf-idf function while the gloss similarity method using text similarity measurement. The output of this step is a correct article for each synset.

EXPERIMENT ON WORDNET BAHASA AND WIKIPEDIA BAHASA MELAYU

A brief experiment was done to test the coverage of WordNet Bahasa and Wikipedia Bahasa Melayu. For WordNet Bahasa, the test was done on the original translated words (stem level) and lemmatization process (root level and mix level). The purpose of this experiment was to evaluate the effect of lemmatization task to the percentage of coverage of WordNet Bahasa. Meanwhile for the Wikipedia Bahasa Melayu, the experiment was done to test the coverage of three main features in Wikipedia Bahasa Melayu (gloss definition, hyperlinks and categories). The purpose of this experiment on three features was to evaluate and determine which technique (ESA, SSA or CSD) was suitable to be used in our research based on the coverage of each feature. The percentage of each coverage was calculated and presented later.

1. Data Set

Currently, there is no existing baseline dataset which can be used in semantic similarity research. However, we can use the dataset from other language for this experiment. A total of 75 word-pairs (150 words) have been selected from three popular dataset: Rubeinstein & Goodenough (1965), Miller & Charles (1991) and Finkelstein et al. (2002). These datasets were provided with word pairs and their senses in wordnet.

2. Translating Data Set to Malay

The selected words from three datasets were translated into Malay. The translation of each word was done at three stages: stem level, root level and mix level (combination of stem and root to suits the WordNet Bahasa). This is to cover the lack of WordNet Bahasa.

a) Stem level

The Malay translated words are based on three type of words which are root words (eg: 'kereta'), affixations (eg: 'pelayaran') and compound words (eg: 'hamba abdi'). Some of these translated words especially compound words can't be retrieved from WordNet Bahasa such as 'hamba abdi' and 'tanah hutan'. However, some of the compound words can match the lemma in WordNet Bahasa such as 'makanan laut'.

b) Root level

The compound words were separated into root words to cater for the semantic information in WordNet Bahasa. The word that can match the lemma in WordNet Bahasa and have the semantic information was saved as candidate word in final data. For example, the root word 'hutan' for a compound word 'tanah hutan' was saved as the word 'hutan' as it can match the lemma in WordNet Bahasa.

c) Mix level

The meaning of mix level in this experiment was we chose one (original or the root word) as the final data. The semantic information of the word was collected from WordNet Bahasa. The sense of the final data must match with the sense of the original english words. For example, the word 'woodland' becomes 'tanah hutan' when translated to malay. There is no output for this compound word in WordNet Bahasa. However, WordNet Bahasa can retrieve the semantic information for the root word 'hutan'. Therefore, the word 'hutan' was selected as the final data for this experiment.

3. Testing the Coverage of WordNet Bahasa and Wikipedia Bahasa Melayu

The purpose of this experiment was to evaluate the effect of lemmatization task to the percentage of coverage of WordNet Bahasa. Meanwhile for the Wikipedia Bahasa Melayu, the purpose of this experiment on three features in Wikipedia Bahasa Melayu was to evaluate and determine which technique between ESA, SSA and CSD was most suitable to be used in our research based on the coverage of each feature.

Testing the coverage of WordNet Bahasa

The final of 75 word-pairs were input into WordNet Bahasa and the coverage of WordNet Bahasa based on the original sense were collected. The coverage of WordNet Bahasa was based on the sense of each word. It should be noted that there are some differences between English WordNet and WordNet Bahasa:

i) The amount of senses between English word in WordNet and Malay translated word in WordNet Bahasa are different. For example, the word 'marathon' in the WordNet has 3 senses while the word 'maraton' in WordNet Bahasa has 2 senses.

ii) Some sense in English WordNet does not exist in WordNet Bahasa.

Testing the coverage of Wikipedia Bahasa Melayu

Previous researchers proposed several techniques to map the WordNet synset to Wikipedia's article such as Explicit Semantic Analysis (ESA), Salient Semantic Analysis (SSA) and Category Semantic Depiction (CSD). Each technique uses different Wikipedia's feature. ESA uses the gloss in Wikipedia's article to match the terms. SSA uses hyperlinks while CSD uses category features. As we mentioned previously, Wikipedia Bahasa Melayu is still a developing lexical database and lack semantic information: there are some articles that do not have all these three features. This experiment was done to clarify why we chose ESA technique for our research. The total of 75 word-pairs (150 words) were used as an input term. The gloss definition, hyperlinks and categories for each word was collected.

a) **Calculating the coverage percentage of WordNet Bahasa and Wikipedia Bahasa Melayu**

The coverage percentage of WordNet Bahasa and Wikipedia Bahasa Melayu were calculated and presented in Table 1 and Table 2.

Table 1. The coverage of WordNet Bahasa based on the word level of input word

Word Level	Words	Percentage	Word Pairs	Percentage
Stem Level	133	88.7	59	78.7
Root Level	122	81.3	51	68.0
Mix Level	139	92.7	65	86.7

Table 2. The coverage of Wikipedia Bahasa Melayu based on the features

Features	Word covered		Word Pair covered	
	Total covered	Percentage	Total covered	Percentage
Gloss definition	150	100	75	100
Hyperlinks	140	93.3	66	88
Categories	142	94.7	66	88

b) Discussion

Based on the analysis on page 14, Table 1 shows the coverage percentage of WordNet Bahasa based on three word-levels (stem level, root level and mix level) while Table 2 shows the coverage percentage of Wikipedia Bahasa Melayu based on three main features in Wikipedia's article (gloss definition, hyperlinks and categories). WordNet Bahasa was developed based on English WordNet but still lack of information. Since there is no available baseline Malay dataset for semantic similarity measurements, we used English dataset. Our constraint with WordNet Bahasa was to match the translated words with the lemma in WordNet Bahasa. To resolve this problem, lemmatization task was required in order to find the best final words as our dataset. From the experiment, we can conclude that the mix level of words gave the highest percentage of WordNet Bahasa's coverage with 92.7% for words and 86.7% for word pairs.

For Wikipedia Bahasa Melayu, previous techniques in literature mentioned three different techniques using three different features of Wikipedia's article. The techniques are ESA, SSA and CSD. ESA uses gloss definition features, SSA uses hyperlinks while CSD uses categories. Due to the lack of semantic information in Wikipedia Bahasa Melayu compared to English Wikipedia, we need to evaluate the Wikipedia Bahasa Melayu based on the features that can give the most effective information. The coverage of three features of each words and word pairs in Wikipedia's article were examined. From the analysis of this experiment, the result shows that gloss definition gave full coverage for our 75 word-pairs (100%) while hyperlinks and categories gave the same percentage (88%).

CONCLUSION AND FUTURE WORK

Our main focus of this research was to build a method to measure semantic similarity of Malay words using some techniques that can suit Malay language behaviour and features. In this research, we used WordNet Bahasa and Wikipedia Bahasa Melayu as our lexical resources. In order to develop this research, firstly we need to understand the issues of semantic similarity measurements for Malay words. Then only we can propose some solutions to solve all the issues.

This article describes some issues related to the semantic similarity measurements for Malay words and proposed the solutions for each issue. There are three main issues regarding to the semantic similarity measurements for Malay words which are: i) representation; ii) semantic similarity measurements; and iii) mapping WordNet synset to the Wikipedia article. In this article, we presented the proposed frameworks and described the details of each steps.

To ensure that this research can be carried out effectively, we need to study the previous techniques in the literature in order to choose the most effective techniques for Malay words and test the lexical resources that will need to be used in this study. A brief experiment was done to examine the coverage of WordNet Bahasa and Wikipedia Bahasa to test its efficiency as our lexical resources. From our experiment, we can conclude that WordNet Bahasa and Wikipedia Bahasa Melayu are capable to be used as lexical resources for semantic similarity research with some adjustments.

In future, we will adapt previous semantic similarity techniques for our Malay dataset and finding solutions to suit Malay language behaviour. As for the semantic similarity using WordNet Bahasa, the mix-level translated words will be used as our dataset. For Wikipedia Bahasa Melayu, the gloss definition will be collected. We will use ESA technique and study efficient methods to adapt with Malay language requirements based on our proposed framework.

REFERENCES

Agirre, E., Edmonds, P. (2006). Word Sense Disambiguation: Algorithms and Applications. *Text Speech and Language Technology*.

Alkhatib, M., Monem, A. A., Shaalan, K. (2017). A Rich Arabic WordNet Resource for Al-Hadith Al-Shareef. *Procedia Computer Science*.

Almaayah, M., Sawalha, M., Abushariah, M. A. M. (2014). A Proposed Model for Quranic Arabic WordNet. Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts.

Banerjee, S., Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. IJCAI International Joint Conference on Artificial Intelligence.

Chen, D., Fisch, A., Weston, J., Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Chen, X., Liu, Z., Sun, M. (2014). A Unified Model for Word Sense Representation and Disambiguation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Fernando, S., Stevenson, M., Court, R. (2012). Mapping WordNet synsets to Wikipedia articles. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116-131.

Gabrilovich, E., Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34, 443–498.

Hassan, S., Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.

Leacock, C., Chodorow, M., Miller, G. A. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*.

Lee, C.-Y., Hsieh, S.-K. (2015). Linguistic Linked Data in Chinese: The Case of Chinese Wordnet. Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries. Proceeding of SIGDOC '86.

Liu, Z., Zheng, V. W., Zhao, Z., Zhu, F., Chang, K. C., Wu, M., Ying, J. (2017). Semantic Proximity Search on Heterogeneous Graph by Proximity Embedding. Aaai 2017.

Ma, B., Zhang, N., Liu, G., Li, L., Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing and Management*.

Mahyoub, F. H. H., Siddiqui, M. A., Dahab, M. Y. (2014). Building an Arabic Sentiment Lexicon Using Semi-supervised Learning. *Journal of King Saud University - Computer and Information Sciences*.

McCrae, J. P. (2018) Mapping WordNet instances to Wikipedia. GWC 2018 - 9th Global WordNet Conference.

Meyer, C. M., Gurevych, I. (2010). Worth its weight in gold or yet another resource - A comparative study of Wiktionary, OpenThesaurus and GermaNet. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Miller, G., Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.

Moro, A., Raganato, A., Navigli, R. (2018). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics*.

Na Chai, W., Ruangrajitpakorn, T., Buranarach, M., Supnithi, T. (2017). A framework to generate carrier path using semantic similarity of competencies in job position. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Netisopakul, P., Thong-Iad, K. (2019). Thai sentiment resource using Thai wordnet. *Advances in Intelligent Systems and Computing*.

Ponzetto, S. P., Navigli, R. (2010). Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems. Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics.

Rada, R., Mili, H., Bicknell, E., Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1), 17–30.

Rubenstein, H., Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Saif, A., Ab Aziz, M. J., Omar, N. (2016). Reducing explicit semantic representation vectors using Latent Dirichlet Allocation. *Knowledge-Based Systems*.

Saif, A., Omar, N., Zainodin, U. Z., Aziz, M. J. A. (2018). Building sense tagged corpus using wikipedia for supervised word sense disambiguation. *Procedia Computer Science*.

Sanchez, D., Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, 44(5), 749–759.

Sanchez, D., Batet, M., Isern, D., Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9), 7718–7728.

Seco, N., Veale, T., Hayes, J. (2004). An intrinsic information content metric for semantic similarity in word net, *Frontiers in Artificial Intelligence and Applications*.

Strube, M., Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. Proceedings of the National Conference on Artificial Intelligence.

Taieb, M. A., Ben Aouicha, M., Tmar, M., Ben Hamadou, A. (2011). New information content metric and nominalization relation for a new WordNet-based method to measure the semantic relatedness. Proceedings of 2011, 10th IEEE International Conference on Cybernetic Intelligent Systems, CIS 2011.

Wu, Z., Palmer, M. (1994). Verbs semantics and lexical selection. 32nd Annual Meeting of the Association for Computational Linguistics.

Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J., Chen, E., Xu, G. (2017). An efficient Wikipedia semantic matching approach to text document classification. *Information Sciences*.

Zhang, C., Zhang, L., Wang, C. J., Xie, J. Y. (2014). Text Summarization Based on Sentence Selection with Semantic Representation. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI.

Zhang, P. Y. (2013). Word Similarity Computation Based on WordNet and HowNet. *Applied Mechanics and Materials*.

Zhou, Z., Wang, Y., Gu, J. (2008). New model of semantic similarity measuring in wordnet. Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering, 256–261.

APPENDICES

Appendix A

Translation of English word pairs to Malay

Word Pairs	Stem level	Root level	Mix level
	Translated Word Pairs	Translated Word Pairs	Translated Word Pairs
car - automobile	kereta - automobil	kereta - automobil	kereta - automobil
journey - voyage	perjalanan - pelayaran	jalan - layar	perjalanan - pelayaran
forest - graveyard	hutan - kuburan	hutan - kubur	hutan - kubur
monk - slave	sami - hamba abdi	sami - hamba, abdi	sami - hamba
beach - forest	pantai - hutan	pantai - hutan	pantai - hutan
forest - woodland	hutan - tanah hutan	hutan - tanah, hutan	hutan - hutan
noon - string	tengah hari - rangkaian	tengah, hari - rangkai	tengah hari - rangkaian
seafood - food	makanan laut - makanan	makan, laut - makan	makanan laut - makanan
marathon - sprint	maraton - lari pecut	maraton - lari, pecut	maraton - lari pecut
train - car	kereta api - kereta	kereta api - kereta	kereta api - kereta

Appendix B

The coverage of WordNet Bahasa

Word Pairs	Stem level		Root level		Mix level	
	Translated Word Pairs	Covered by WordNet Bahasa?	Translated Word Pairs	Covered by WordNet Bahasa?	Translated Word Pairs	Covered by WordNet Bahasa?
car - automobile	kereta - automobil	Yes	kereta - automobil	Yes	kereta - automobil	Yes
journey - voyage	perjalanan - pelayaran	Yes	jalan - layar	Yes	perjalanan - pelayaran	Yes
forest - graveyard	hutan - kuburan	Yes	hutan - kubur	Yes	hutan - kubur	Yes
monk - slave	sami - hamba abdi	No	sami - hamba, abdi	Yes (hamba)	sami - hamba	Yes
beach - forest	pantai - hutan	Yes	pantai - hutan	Yes	pantai - hutan	Yes
forest - woodland	hutan - tanah hutan	No	hutan - tanah, hutan	Yes (hutan)	hutan - hutan	Yes
noon - string	tengah hari - rangkaian	Yes	tengah, hari - rangkai	No	tengah hari - rangkaian	Yes
seafood - food	makanan laut - makanan	Yes	makan, laut - makan	No	makanan laut - makanan	Yes
marathon - sprint	maraton - lari pecut	Yes	maraton - lari, pecut	No	maraton - lari pecut	Yes
train - car	kereta api - kereta	Yes	kereta, api - kereta	No	kereta api - kereta	Yes

Appendix C

The coverage of gloss definition in Wikipedia Bahasa Melayu

Word 1	Covered by Wikipedia BM?	Word 2	Covered by Wikipedia BM?	Word Pair covered by Wikipedia BM?
kereta	Yes	automobil	Yes	Yes
perjalanan	No	pelayaran	Yes	No
rahib	Yes	sami	Yes	Yes
perjalanan	No	kereta	Yes	No
hutan	Yes	kuburan	Yes	Yes
sami	Yes	hamba abdi	Yes	Yes
pantai	Yes	hutan	Yes	Yes
kaca	Yes	ahli silap mata	Yes	Yes
hutan	Yes	tanah hutan	Yes	Yes
tengah hari	Yes	rangkaian	No	No

Appendix D

The coverage of hyperlinks in Wikipedia Bahasa Melayu

Word 1	Covered by Wikipedia BM?	Word 2	Covered by Wikipedia BM?	Word Pair covered by Wikipedia BM?
kereta	Yes	automobil	Yes	Yes
perjalanan	No	pelayaran	Yes	No
rahib	Yes	sami	Yes	Yes
perjalanan	No	kereta	Yes	No
hutan	Yes	kuburan	No	No
sami	Yes	hamba abdi	Yes	Yes
pantai	Yes	hutan	Yes	Yes
kaca	Yes	ahli silap mata	Yes	Yes
hutan	Yes	tanah hutan	Yes	Yes
tengah hari	Yes	rangkaian	No	No

Appendix E

The coverage of categories in Wikipedia Bahasa Melayu

Word 1	Covered by Wikipedia BM?	Word 2	Covered by Wikipedia BM?	Word Pair covered by Wikipedia BM?
kereta	Yes	automobil	Yes	Yes
perjalanan	No	pelayaran	Yes	No
rahib	Yes	sami	Yes	Yes
perjalanan	No	kereta	Yes	No
hutan	Yes	kuburan	Yes	Yes
sami	Yes	hamba abdi	Yes	Yes
pantai	Yes	hutan	Yes	Yes
kaca	Yes	ahli silap mata	No	No
hutan	Yes	tanah hutan	Ya	Yes
tengah hari	Yes	rangkaian	No	No