Semantic Similarity Measures for Malay-English Ambiguous Words

Nurul Husna Mahadzir[1], Mohd Faizal Omar[1] and Mohd Nasrun Mohd Nawi[2]
[1]*School of Quantitative Sciences, University Utara Malaysia, 06100 Sintok, Kedah.*
[2]*School of Technology Management and Logistic, University Utara Malaysia, 06100 Sintok, Kedah.*
faizal_omar@uum.edu.my

***Abstract*—The measurement of semantic similarity between words or concepts has been widely applied in various applications such as Artificial Intelligence, Information Processing and Natural Language Processing (NLP) field. In this paper, we are going to demonstrate how several semantic similarity measures can be adapted to disambiguate words that belong to two different languages. The measures include two path-based and three information content (IC)-based measures. These five measures are evaluated against a test bed of 40 word pairs with eight Malay-English ambiguous words. The experimental results on a common benchmark, created by human judgments show that Wu and Palmer's measures have given the best performance as compared to the other four semantic similarity measures.**

***Index Terms*—Ambiguous Words; Information Content Based; Path-Based; Semantic Similarity Measures.**

I. INTRODUCTION

The arisen of social media usages such as Facebook and Twitter has become a new challenge for NLP. In addition, it is common to see that social media users tend to mix up languages in one sentence in expressing their emotions or opinions [1]. In Malaysia for instance, as multilingual countries, people use Malay and English language in their daily online or offline conversation [2]. Unlike the highly edited genres that conventional NLP tools have been developed for, the online conversational text contains many non-standard and unstructured data which makes it hard for the same tool to be applied in this kind of data.

One of the prominent issues in regard to the use of mixed language is ambiguity where one single word contains different meanings depending on which language it belongs [3]. Generally, there are many words which belong to two different languages such as English and Malay where they share similar spelling but have a different meaning. This scenario is called as 'ambiguous word' as it could belong to more than one language. For instance, the word "fail" exists in English which means unsuccessful and in Malay which give a meaning of a folder. Individually, there is no way of knowing whether the word is written in the English "fail" or the Malay "fail" since they are spelt similarly. However, since we deal with sentences, we take into consideration the language and the meaning of the surrounding words in to get a sense of the context of the sentence.

Automatically assigning the exact language of ambiguous words is becoming the essential task due to the growing amount of information available through the Internet. This paper is proposing to adopt semantic similarity approach in order to automatically assign the correct language of the ambiguous word that belongs to either Malay or English language.

Semantic similarity measures used in this research utilize WordNet as knowledge sources in order to obtain the score on the relatedness between the ambiguous word and its surrounding words [4]. Although it has been widely applied in many applications, measuring the semantic similarity for words in different languages is considered new and has not been explored before.

The rest of the paper is organized as follows. We review the existing application of semantic similarity measurement in Section 2. We describe in detail the semantic similarity measures in Section 3. We present the experimental evaluation of various semantic similarity measures in Section 4 and discuss the result achieved by each measure in Section 5. Lastly, we conclude this paper and discuss the future studies in Section 6.

II. RELATED WORKS

The use of mixed language in daily conversation or in social media platform has given a new challenge in NLP task as some words are ambiguous. If this ambiguousness is remained unresolved, it will lead to misinterpretation and will reduce the efficiency in obtaining the correct information [5]. In the literature, this problem has been addressed by proposing few approaches. The nearest neighbour approach [18] has been used as one of the ways to perform disambiguation where the corresponding language of the ambiguous word is determined by making an assumption that the language of an ambiguous word is similar to the word that appears before and after that word in the same sentence. However, this approach is considered inaccurate as the certain ambiguous word may appear 'stand-alone' in a mixed language sentence and it does not belong to the same language as its surrounding words. Another approach is the disambiguation task has been manually performed by human annotators [6, 7]. This paper has investigated one approach to automate the disambiguation process called semantic similarity measures.

Semantic similarity measurement is a field of research where two words are assigned a quantitative score based on the likeness of their meaning [8]. Automatic measurement of semantic similarity is considered as one of the principals for various computer-related fields since a wide variety of techniques rely on deciding the meaning of data they work with. The idea of measuring semantic similarity between words is to find a model that can simulate similar thinking process of human [9]. The righteous solution for a human to compare between two objects is to find the similarity

between those two objects. It will not be a problem for a human to compare between two words and they can easily recognize if one word is similar to a given word than another.

Obtaining semantic similarity between words is necessary for many applications in text analytics. There are a lot of efforts on measuring similarity in general and on word similarity in particular. The following section will discuss the available measurement which will be applied to the experiment to disambiguate Malay-English words in this research.

III. Semantic Similarity Measures

Computing semantic similarity between words has been used in various applications and many measurement calculations have been proposed. It expresses the degree of likeness of the meaning between two related words. Measures of semantic similarity are often based on information regarding *is-a* relations found in a set of words hierarchy and it utilizes a lexical ontology such as WordNet. WordNet is utilized because it contains rich information about senses of a word and their relations [19]. In WordNet, one specific word and its synonyms are grouped together as synsets.

Let be w_1 and w_2 are two words that belong to two different nodes n_1 and n_2 in a given ontology, the distance between the nodes (n_1 and n_2) determines the similarity between these two words w_1 and w_2. Both w_1 and $w_2$can be considered as an ontology (also called concept nodes) that contains a set of terms synonymous and consequently. Two terms are synonymous if they are in the same node and their semantic similarity is maximized.

Existing semantic similarity measures can be categorized into two groups; path-based and information content (IC) based [10]. For the experiment in this article, we adapted two path-based measures and three IC-based measures as illustrated in Figure 1.

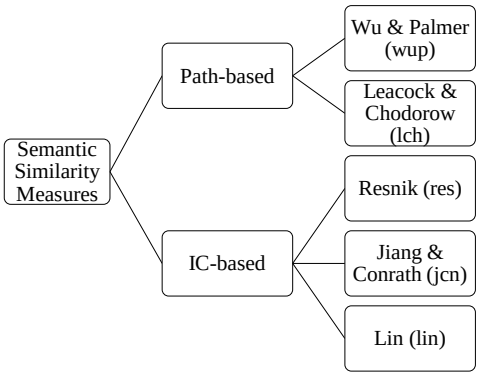

Figure 1: Semantic Similarity Measures

A. Path-Based

Path-based measures [11] rely on the shortest path information whereas IC based measures incorporate the probability of the set of words or concepts occurring in a corpus of text.

Wu and Palmer [12] present a measure of similarity for general English that relies on finding the most general concept that subsumes both of the words being measured. Mathematically, the similarity of two words w_1 and w_2 using the Wu and Palmer (wup) measure is computed as:

$$sim_{wup}\ (w_1, w_2) = 1/p \quad (1)$$

where p is a number of nodes on the shortest path between the two words in WordNet. Similarly, Leacock and Chodorow (lch) [13] define a similarity measure that is based on finding the shortest path between two concepts:

$$sim_{lch}\ (w_1, w_2) = -\log(\frac{p}{2.depth}) \quad (2)$$

where depth is the maximum depth of the hierarchy.

B. Information Content-Based

Information Content (IC)-based measures [14] are estimated by counting the frequency of that concept in a large corpus of text. A concept with high information content is very specific while lower information content values are associated with more general concepts.

Resnik (res) [15] defined a measure of similarity between two words w_1 and w_2 as:

$$sim_{res}\ (w_1, w_2) = IC(lcs(w_1, w_2)) \quad (3)$$

Jiang and Conrath [16] and Lin [17] developed measures that scale the information content of the subsuming concept by the information content of the individual concepts. Lin does this via a ratio and Jiang and Conrath with a difference. The Jiang and Conrath (jcn) measure computes the semantic distance (inverse of similarity) of words w_1 and w_2 as:

$$dist_{jcn}\ (w_1, w_2) = IC(w_1) + IC(w_2) - 2.IC(lcs(w_1, w_2)) \quad (4)$$

and the Lin measure (lin) computes semantic similarity of words w_1 and w_2 as:

$$sim_{lin}\ (w_1, w_2) = \frac{2.IC(lcs(w_1, w_2))}{IC(w_1)+IC(w_2)} \quad (5)$$

where $lcs(w_1, w_2)$ is the lowest common subsumer of w_1 and w_2 and IC returns the information content of the word.

IV. Experimental Details

A. Dataset

Currently, there is no existing set of Malay-English ambiguous words that could be used as a direct means of evaluation. In this research, we created a test bed of word pair that was assessed by human experts according to their relatedness. We have assigned three annotators how are proficient in both Malay and English to annotate 50 word pair in the test set. To derive a more reliable test set, we extracted only those pairs whose agreement was high. This resulted in a set of 40 word pairs with 8 Malay-English ambiguous words has been generated and the human experts were requested to classify each word pair either Malay (My) or English (En) based on the relatedness between the pair. Table 1 shows the sample of the annotation.

We implemented five measures of semantic similarity based on *is-a* relations as found in WordNet [20]. We have used Python package called Sematch [21] and its module WordNet::Similarity in order to obtain the score. The advantage of using this package is we do not have to translate the Malay word into English in order to utilize the English WordNet because the cross-lingual word similarity is available in the same package.

Table 1
Sample Dataset

Word pair	Expected Result
(fail, computer)	My
(main, football)	My
(main, road)	En
(air, plane)	En
(tan, water)	My
(cat, canvas)	My
(cat, pet)	En
(liar, tiger)	My
(beg, mercy)	En
(beg, grocery)	My
(beg, forgiveness)	En
(jam, traffic)	En
(jam, wrist)	My

Below is the example of the disambiguation process for the word pair *main* and *football*. Consider the following sentence containing the ambiguous word *main* which means principal or most important in English and plays in Malay.

In this example shown in Figure 2, we took football as the word pair for this ambiguous word. Then, the ambiguous word *main* is divided into two categories; Malay (My) and English (En). Then, the similarity score obtained from WordNet::Similarity will be assigned to each word pair from both categories. The corresponding language with the highest similarity score will be assigned to the ambiguous word; in this case Malay (My).

Football?? Last aku *main* two years ago!
(Football?? the last time I played was two years ago!)

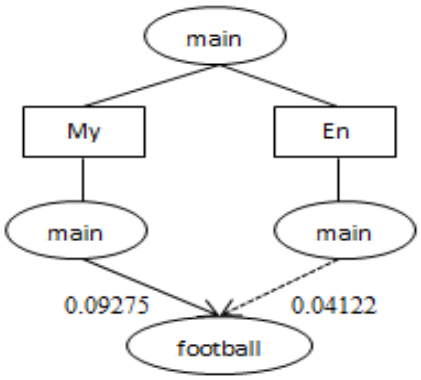

Figure 2: Similarity Score

Table 2
Semantic Similarity Score

Category	w_1	w_2	lch	wup	jcn	lin	res
My	fail	computer	**0.29627**	**0.63157**	**0.07605**	**0.34906**	**3.25768**
En	fail	computer	0	0	0	0	0
My	fail	drama	0.20557	0.46153	0.07088	0.28558	2.61964
En	fail	drama	0	0	0	0	0
My	main	road	0.29627	0.625	0.15224	0.2817	2.33223
En	main	road	**0.2997**	**0.66666**	0.07755	**0.67002**	**5.6535**
My	main	ball	0.24257	**0.63636**	**0.13946**	**0.64417**	**5.58543**
En	main	ball	0.25109	0.55555	0.08231	0.41581	3.96759
My	tan	color	0.08877	0.30769	0.05855	0.06904	0.59622
En	tan	color	**0.4471**	**0.75**	**0.15314**	**0.7058**	**6.63318**
My	tan	water	0.0595	0.26666	0.05663	0.06328	0.59622
En	tan	water	0.16831	0.42857	0.07214	0.32075	3.03657
My	tan	leather	0.03265	0.22222	0.05316	0.06275	0.59622
En	tan	leather	**0.16831**	**0.42857**	**0.07214**	**0.32075**	**3.03657**
My	tan	weight	**0.44199**	**0.71428**	**0.09219**	**0.46531**	**4.28482**
En	tan	weight	0.16831	0.42857	0.07214	0.32075	3.03657
My	tan	skin	0.0595	0.26666	0.04904	0.05794	0.59622
En	tan	skin	**0.23347**	**0.53333**	**0.0641**	**0.2989**	**3.1399**
My	air	drink	**0.77517**	**0.88888**	**0.31582**	**0.85609**	**6.69284**
En	air	drink	0.36187	0.53333	0.10816	0.48522	4.08877
My	air	tap	0.36187	0.66666	0.06095	0.23037	**4.07648**
En	air	tap	0.36187	0.6666	0.0655	0.2464	2.33223

B. Results

Table 2 presents the results obtained using five semantic similarity measures. The similarity score with the red colour indicated the contradict result as compared to human judgments while the score in the bold show similar result as annotated by human experts. Based on this table, few ambiguous words under English category (e.g *fail*) always give zero results. One of the reasons for this result is due to the absence of either the ambiguous word itself or its word pair in WordNet. The inconsistency of the result can be seen for the word *tan* which shares the same spelling as English once it is translated.

As presented in Table 3, it shows an error where both categories (My and En) shares the same similarity score for word pair (jam, wrist) and (air, tap).

Table 3
Same score for lch and res

Category	w_1	w_2	lch	res
My	Jam	wrist	**0.0267**	0.8018
En	Jam	wrist	0.0219	0.8018
My	Air	tap	0.3619	**4.0765**
En	Air	tap	0.3619	2.3322

Table 4 shows the accuracy percentage that has been obtained by measuring the correlation between human judgments and the highest score for both path-based and IC-based measures. The most accurate result is achieved by Wu & Palmer measure with 67.5% and the worse result are given by Jiang & Conrath and Lin measures. However, the result is considered as satisfying since the accuracy for all measures has exceeded 50%. While Jiang & Conrath and Lin measures share the same accuracy which is 57.5%.

Table 4
Summary of the experimental results

Measures	Accuracy (%)
Wu & Palmer	**67.5**
Leacock & Chodorow	60
Resnik	62.5
Jiang & Conrath	57.5
Lin	57.5

V. DISCUSSION

The main focuses of this research are to find an automatic tool to disambiguate Malay-English words. The result tells us that semantic similarity measures can be very useful when supporting the disambiguation task that has been manually done currently. The achieved accuracies as compared to human judgments show that semantic similarity measurement is a suitable technique that can be applied to disambiguation Malay and English words. However, there are still some major problems when certain words do not exist in the available source which is WordNet where it leads to an inaccurate score. Another problem is regarding the selection of surrounding words. An improvement can be made by identifying a proper technique to select the surrounding words in order to improve the result.

For the errors report in Table 3, it proved that Resnik (res) and Leacock & Chodorow (lch) were not able to disambiguate few words as both measures have given a similar score. An important avenue to resolve this error is to perform more experimentation with various word pair with different knowledge sources.

VI. CONCLUSION AND FUTURE WORKS

Analyzing the unstructured text in social media platform has given a new challenge in NLP field. The growing number of multilingual speakers in a multilingual country such as Malaysia has increased the use of mixed language in their daily communication. Due to this growth, it is necessary for research in text analytics area to deal with ambiguity where more than one language shares the same words.

Therefore, the aim of this research is to resolve the ambiguity issue in mixed language content. We have adapted semantic similarity approach in performing disambiguation of Malay-English ambiguous words. Different semantic similarity measures have different characteristic. Path-based measures take the path length linking between the two words or concepts while IC based measure the relatedness based on the assumption that the more common information two words share, the more similar the words are.

In this article, we have conducted an experiment on 40 word pairs to show the efficacy of adopting semantic similarity measures using WordNet to disambiguate Malay and English word. Three annotators who are proficient in both Malay and English have been assigned to classify each word pair. As a result, Wu & Palmer has given the most accurate compared to the other four measurements.

In the near future, we plan to extend the experimentation with more data sets and to find the better technique in selecting the word pair for the ambiguous words in order to improve the accuracy of the disambiguation. A new measurement calculation will be explored as well in order to deal with word pair with a similar score. We also would like to explore another available knowledge sources such as SentiWordNet or SentiStrength in order to overcome the limited number of words in WordNet.

REFERENCES

[1] J.J. Gumperz. 1982. Discourse strategies, volume 1. Cam- bridge Univ Pr.

[2] Chuah, K. (2013). Aplikasi Media Sosial Dalam Pembelajaran Bahasa Inggeris : Persepsi Pelajar Universiti. *Issues In Language Studies*, *2*(1), 56–63.

[3] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A., & Zhou, Q. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Cognitive Computation*, *8*(4), 757–771

[4] Patwardhan, S., & Pedersen, T. 2006. Using WordNet- based Context Vectors to Estimate the Semantic Relatedness of Concepts. In EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together (pp. 1-8). Trento, Italy.

[5] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, *28*(2), 15–21.

[6] Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2016). A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. *Knowledge-Based Systems*, *105*, 236-247.

[7] Zhang, Q., Chen, H., & Huang, X. (2014, February). Chinese-English mixed text normalization. In *Proceedings of the 7th ACM international conference on Web search and data mining* (pp. 433-442). ACM.

[8] Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003, October). Semantic similarity measures as tools for exploring the gene ontology. In *Pacific Symposium on biocomputing* (Vol. 8, No. 04, pp. 601-612).

[9] Hao D, Zuo W, Peng T, He F (2011) An Approach for Calculating Semantic Similarity between Words using WordNnet. In: Second International Conference on Digital Manufacturing and Automation, 177-180

[10] McInnes B, Pedersen T, Liu Y, Pakhomov S, Melton G. Using second-order vectors in a knowledge-based method for acronym disambiguation. In: Proceedings of the conference on computational natural language learning. Portland, 2011

[11] R. Rada, H. Mili, E. Bicknell and M. Blettner, "Development and Application of a Metric on Semantic Nets", IEEE Transactions on Systems, Man and Cybernetics, vol. 19, Issue 1, (1989) January-February, pp. 17 - 30.

[12] Z.Wu and M.Palmer.Verb semantic and Lexical Selection.In Proceedings of 32 Annual Meeting of the association of computer Linguistics(ACL 994),Las Cruces,New Mexico,1994.

[13] Claudia Leacock and Martin Chodorow.Combining Local Context and WordNet Similarity for word Sense identification. In [10], 1998.

[14] Pirró, G., 2009. A semantic similarity metric combining features and intrinsic information content. Data Knowledge Engineering 68(11)

[15] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, (1995) August 20-25, Canada.

[16] J.J.Jiang and D.W Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, Procs.ROCKLING X, 1997.

[17] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, (1998) July 24-27; Madison, Wisconsin, USA.

[18] Sharma, S., Srinivas, P. Y. K. L., & Balabantaray, R. C. (2015, August). Text normalization of code mix and sentiment analysis. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on* (pp. 1468-1473). IEEE.

[19] Wan S, Angryk RA (2007) Measuring Semantic Similarity using WordNet-based Context Vectors. In: Proceedings of the IEEE International Conference on systems, man and cybernetics, Canada

[20] Budnitsky, A., Hirst, G., 2006. Evaluating wordnet-based measures of lexical semantic relatedness. Computer Linguistic, 32(1)

[21] Zhu, G., & Iglesias, C. A. (2017). Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 72-85.