Semantic Similarity Measures for Malay Sentences

Shahrul Azman Noah, Amru Yusrin Amruddin, and Nazlia Omar

Faculty of Information Science & Technology
Universiti Kebangaan Malaysia
Bangi Selangor
samn@ftsm.ukm.my, amruyusrin@yahoo.com, no@ftsm.ukm.my

Abstract. The concept of semantic similarity is an important element in many applications such as information extraction, information retrieval, document clustering and ontology learning. Most of the previous works regarding semantic similarity measures have been traditionally defined between words or concepts (i.e. word-to-word similarity), thus ignoring the text or sentence that the concepts participate. Semantic text similarity was made possible with the availability of resources in the form of semantic lexicon such as the WordNet for English and GermaNet for German. However, for languages such as Malay, text similarity proved to be difficult due to the unavailability of similar resources. This paper, however, describe our approach for text similarity in Malay language. We used a preprocessed Malay dictionary and the overlap edge counting based method to first calculate the word-to-word semantic similarity. The word-to-word semantic similarity measure is then used to identify the semantic sentence similarity using a modified approach for English language. Results of the experiments are very encouraging, and indicate the potential of semantic similarity measure for Malay sentences.

Keywords: Sentence similarity, semantic similarity measures, information retrieval.

1 Introduction

Most of the previous work in information retrieval regarding similarity has been focused primarily on text similarity whereby input query is compared with collection of documents and some ranking results will be obtained. The vector space model is perhaps the most popular approach still employed in text similarity [1]. Text similarity has also been used in relevance feedback [2], document clustering [3], information extraction [4] and ontology learning [5]. Semantic text similarity on the other hand is a concept whereby a set of sentences or terms within term lists are assigned a metric based on the likeness of their meaning content [6].

Measures of semantic similarity have been traditionally defined between words or concepts, and much less between text segments of two or more words. The emphasis on word-to-word similarity metrics is probably due to availability of resources that explicitly specify the relations among words such as the WordNet [7]. Although the method to measure the similarity between pair of texts can be done by measuring similarity of co-occurring words, the chances to get good measures are very slim and therefore few other aspects need to be considered such as word ordering and semantic word meanings.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 117–126, 2007.

In the case of other non-dominant languages such as the Malay language, the measures for text similarity proved to be difficult due to the non-availability of a lexical database similar to the Wordnet. In this paper, we describe the sentence similarity measure for Malay language. We based our approach from the work of Li et al. [8] with some modifications particularly on measuring the word-to-word similarity. The next section provides a brief review on related work in this area particularly in Malay language. Section 3 describes the proposed approach and section 4 discuss our initial experiments findings. Section 5 concludes our work and provides future work directions.

2 Background

Works relating to measuring similarity between sentences and documents in English are extensive [8, 9, 10], but there have been very little or any work which relate to semantic sentence similarity for Malay language. The nearest to our knowledge would be the work of [11] which exploit word-to-word semantic similarity to enhance Malay documents retrieval. In their work, the similarity between words is defined by direct translation of English WordNet.

Most of the sentence similarity measures mainly concern with 'calculating' the availability or non-availability of words in the compared sentences [9, 10]. Therefore, the word overlap measures, TF-IDF measures, relative frequency measures and probabilistic models have been the popular method for evaluating similarity.

In semantic sentence similarity measure, the first task is to get the word-to-word semantic measures of the participating sentences. There is a relatively large number of word-to-word similarity measures previously proposed in the literature, which according to [7] can be clustered into two groups: corpus based measures and knowledge based measures. Corpus-based measures of word semantic similarity seek to identify the similarity between words using information derived from large corpora [12, 13]. Turney [12] proposed Pointwise Mutual Information measures which was based from the term co-occurrence method using counts over large corpora. Another popular approach is the Latent Semantic Analysis (LSA) whereby the term co-occurrences are captured by means of dimensionality reduction operated by a singular value decomposition (SVD).

Knowledge-based measures on the other hand identify the semantic similarity between words by calculating the degree of relatedness among words using information from dictionary or thesaurus [14, 15]. For example the Leacock and Chodorow method [14] count the number of nodes of the shortest path between two concepts in WordNet. The work by Resnik [15] and Li et al. [8] also use the Wordnet to calculate the semantic measures. The Lesk method [16] defined semantic similarity between two words based on overlap measures between the corresponding definitions as provided by a dictionary.

As can be seen work focusing on Malay semantic sentence similarity is little or none. Most of the established and proven works were in English. In this work we experiment the use of semantic sentence similarity for Malay. The proposed method is typically comparable with other methods for English sentences [7, 8] of which pair of texts is compared base on the derived syntactic and semantic information. In this method, we follow the approach proposed in [8] with some modifications.

3 The Proposed Text Similarity Method

Fig. 1 illustrates the procedure for measuring the sentence similarity between two Malay sentences. In adopting the method proposed by [8], a joint distinct word set is formed for the two sentences. For each submitted sentence, a raw semantic vector is obtained by exploiting an open Malay dictionary database. Unlike other English text similarity methods which rely on the WordNet in calculating the word semantic similarity measures, our approach uses the overlap edge counting-based method which was originally proposed by Lesk [16]. In this case the semantic similarity between words is based on the counting of overlaps between dictionary definitions of the compare words.

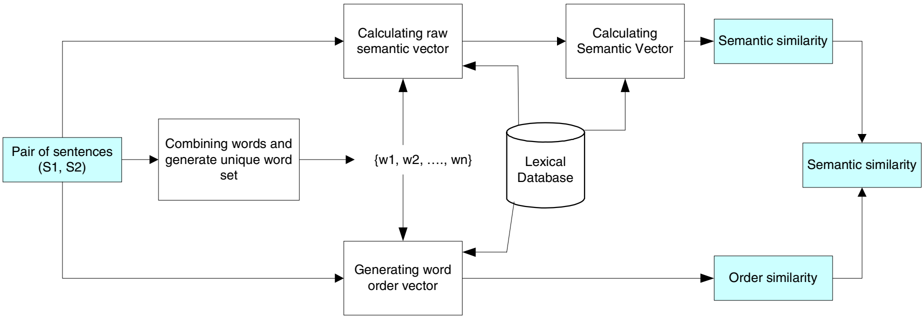
Calculating raw semantic vector
Calculating Semantic Vector
Semantic similarity
Pair of sentences (S1, S2)
Combining words and generate unique word set
{w1, w2, …., wn}
Lexical Database
Semantic similarity
Generating word order vector
Order similarity

Fig. 1. The process for semantic similarity between sentences

A word order vector is formed for each sentence, again using information from the open dictionary. Since each word in a sentence contributes differently to the meaning of the whole sentence, the significance of a word is weighted by using information content derived from the open dictionary. This is another limitation of Malay language which does not yet contain any document corpus such as the Brown corpus for English. By combining the values of raw semantic vector with information content from the open dictionary, a semantic vector is obtained for each of the two sentences. Semantic similarity is then computed based on the two semantic vectors, whereas an order similarity is calculated using the two order vector. Finally, the sentence similarity is derived by combining semantic similarity and order similarity.

We describe in detail the aforementioned procedure in the following sections.

3.1 Semantic Similarity of Words

As previously mentioned, the semantic similarity measures between words can be grouped into corpus-based measures and knowledge measures. We chose to focus on the knowledge-based measure as a large corpus for Malay language related sources is not currently in existence. Furthermore as lexical database for Malay language similar to WordNet is not yet available, we chose to use an open dictionary. The open dictionary contains 69,344 rows of data with 48,177 Malay words which is based from the Kamus Dewan 3rd Edition. The dictionary, however, is still not yet in a Machine Readable Dictionary (MRD) format (i.e the dictionary is still in a human readable format), therefore a few pre processing are required. The dictionary was parsed by filtering and eliminating symbols, short form words, verbs, and other words not found in the dictionary.

After investigating a number of methods for knowledge-based measures, the only suitable method to use is the Lesk's method [16]. This is due to the nature of the generated MRD dictionary which only contains meanings of words and not the hierarchical structure of words that models the human common sense knowledge of general language usage similar to WordNet [17].

Therefore, we proposed that the similarity *sim(*w_1*,* w_2*)* between words w_1 and w_2 is the multiplication of ratios for the meanings of words w_1 and w_2 as follows:

$$sim(w_1, w_2) = r(C, M_{w_1}) \cdot r(C, M_{w_2}) \tag{1}$$

where C is the set of unique overlap words found in the meanings of w_1 and w_2 and M refers to the meanings of the respective words. Therefore, $r(C, M_{w1})$ refers to the ratio between the counts of meanings that contains any of the words in C with all the meaning associated with w_1.

3.2 Semantic Similarity Between Sentences

Sentences are aggregation of words, therefore, it is common to use words in the sentences to represent the sentences. Using the method proposed by [8], the semantic vector of words is dynamically formed solely based on the compared sentences. This approach is slightly different with the conventional vector space model which requires the comparison of all words existed on the document corpus.

So, assuming we are comparing between sentences, S_1 and S_2, a joint distinct word set S, is formed between S_1 and S_2:

$$\begin{aligned} S &= S_1 \cup S_2 \\ &= \{w_1, w_2, \ldots\ldots., w_n); \; w_i \text{ are distincts} \end{aligned}$$

We don't consider morphological variants among words. Therefore, the words *makan* (eat), *makanan* (food) and *pemakanan* (nutrition) are all considered as three distinct words and forms part of the set *S*. For example if we have the sentences: S_1: *Saya berjalan ke sekolah* (I walked to school); S_2: *Dia berkereta ke bandar* (He drived to town), then we will have *S* = {*Saya berjalan ke sekolah Dia berkereta bandar*}. The joint word set *S*, is viewed as the semantic information for the compared sentences. In other words the semantic information for sentences S_1 and S_2 are derived from the joint set *S*. To derive the semantic information content of S_1 and S_2, a term-term matrix is constructed as follows:

$$\begin{array}{rc|ccccccc|}
 & S & =w_1 & w_2 & \cdots & \cdots & \cdots & \cdots & w_n \\
 & q_1 & x_{1,1} & x_{1,2} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & x_{1,n} \\
 & q_2 & x_{2,1} & x_{2,2} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & x_{2,n} \\
S_i = \Bigg\{ & \cdot & \cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot \\
 & \cdot & \cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot \\
 & q_m & x_{m,1} & x_{m,2} & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & x_{m,n}
\end{array}$$

whereby $x_{i,j}$ represents the similarity measure between the i-th word in the compared sentence and the j-th word of the joint set. The value of $x_{i,j} = 1$, if q_i and w_i are the same words, whereas if $q_i \neq w_i$, the similarity measure is computed using the word-to-word semantic similarity method previously described. In the case of $q_i \neq w_{i,}$ the similarity value of $x_{i,j}$ is only considered if $x_{i,j} > \xi$, whereby ξ is a specified threshold value. Anything less than ξ, is assumed not semantically similar.

From our experiment of 200 pairs of antonyms, the $\xi = 0.18$ has been selected due to its dominance as shown in Fig. 2.

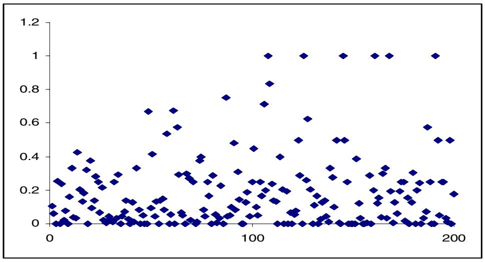

Fig. 2. Distribution of the word-to-word similarity measures for antonyms

The raw semantic vector of S_i (i = 1,2), i.e. $\check{s}$; can then be computed, whereby $\check{s}$ = {*(max* $(x_{1,1}...x_{m,1})),......,$ *(max*$(x_{1,n},...x_{m,n}))$}. For example if we compared between, the joint set S = {*negara, Malaysia, aman, sentosa, jepun, maju*} with the compared sentence S_1 = {*negara, Malaysia, aman, sentosa*}, we will get the following term-term matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.327 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

therefore, the raw semantic vector $\check{s}$ for S_1 will be {1 1 1 1 0 0.327}.

For the calculation of the semantic vector s_i, the following formula is used:

$$s_i = \check{s} \cdot I(w_i) \cdot I(\tilde{w_i}) \tag{2}$$

where w_i is a word in the joint word set S, and w_i is its associated word in the sentence. The value of $I(w)$ is calculated by referring to the MRD dictionary, using the following formula:

$$I(w) = 1 - \frac{\log(n+1)}{\log(N+1)} \tag{3}$$

where n is the number of rows of meaning containing the word w and N is the total number rows (meaning) in the dictionary. Then, the semantic similarity between the two compared sentences is simply the cosine coefficient between the two semantic vectors.

$$S_s = \cos(s_1, s_2) = \frac{s_1 \cdot s_2}{|s_1| \times |s_2|} \tag{4}$$

3.3 Word Order Similarity Between Sentences

Measuring the word similarity is rather a straightforward process and used the similar joint word set as discussed in the previous section. Assuming that we have a pair of sentences, L_1 and L_2 of which:

L_1: *Negara Malaysia aman sentosa*
L_2: *Jepun negara maju*

therefore, we will have a join set $L = \{$*Negara, Malaysia, aman, sentosa, Jepun, maju*$\}$. Similarly with the semantic similarity, the vector of the word order is derived from the joint set *L*. A term-term matrix is constructed and the word-to-word similarity measure is calculated using the method discuss in section 3.1. The resulting matrix for the sentence L_1 and the joint set *L* is similar to the one presented in section 3.2.

The word order vector for L_1, i.e. u_1 is constructed based on the existence or the highest word-to-word similarity between the joint set *L* and L_1. Therefore we will have u_1 = (1 2 3 4 0 2), the last value of u_1 is equal to 2 because the word *maju* in *L* is strongly similar with the word Malaysia, which is the second position in L_1. Similarly we will get u_2 = (2 2 3 3 1 3), derived from the following matrix.

L L_2	negara	Malaysia	aman	sentosa	Jepun	Maju
Jepun	0	0	0	0	1	0
negara	1	0.58	0	0	0	0
maju	0	0	0.282	0.163	0	1
u_2	(2	2	3	3	1	3)

Using the word order similarity as follows:

$$S_r = 1 - \frac{|u_1 - u_2|}{|u_1 + u_2|} \tag{5}$$

we will get $S_r(L_1, L_2)$ = 0.828. The word order similarity in (5) is determined by the normalized difference of word. Li et al. [8] has demonstrated that the formula is an efficient metric for measuring word order similarity.

3.4 Combine Sentences Similarity

The combine sentences similarity represents the overall sentences similarity, which is the summed of the semantic similarity and the word order similarity as follows:

$$Sim(S_1, S_2) = \delta S_s + (1 - \delta)S_r \tag{6}$$

whereby δ is a damping factor, which decides the contribution of the involved similarity measures (i.e. S_s and S_r). Li et al. [8] suggested that δ should be greater than 0.5 due to the importance of lexical elements presented in semantic similarity [18].

4 Initial Experimental Testing

We have conducted an initial testing in order to evaluate of the proposed modified approach. The result of the testing is as illustrated in Table 1. Due to brevity, only portion of the result is shown.

Table 1 separates the result into semantic similarity, order similarity and sentence similarity for $\partial = 0.5$. The testing as illustrated in Table 1 compares the first sentence of the list with the remaining six sentences. To assist discussion, we called the first sentence of the list as the 'target sentence' and the remaining six sentences as the 'compared sentences'. Each of the six compared sentence is being weight against the target sentence. Human ranking of similarity is just our (human) opinion about the relevancy ranking of the compared sentences with the target sentence.

Result in Table 1 shows consistent outcome between our ranking of similarity and the approach sentence similarity measures, with very minimal differences.

Table 1. Initial testing result*

Sentence Tested	Sentences Compared	Human Ranking of Similarity	Semantic Similarity	Order Similarity	Sentence Similarity
Target sentence 1					
Saya pergi ke sekolah.					
	Saya pergi ke sekolah	1	1	1	1
	Saya berjalan ke sekolah	2	0.95	1	0.98
	Saya pergi ke madrasah	3	0.90	1	0.95
	Saya pergi ke kedai	4	0.61	0.89	0.74
	Dia pergi ke kedai	5	0.59	0.89	0.74
	Saya makan nasi di kedai	6	0.46	0.67	0.57
Target sentence 2					
Saya membaca buku sambil minum air kopi.					
	Saya membaca buku sambil minum air kopi.	1	1	1	1
	Saya membaca buku sambil minum air teh.	2	0.87	0.62	0.74
	Saya membelek majalah sambil minum air teh.	3	0.54	0.60	0.57
	Saya menonton televisyen sambil minum air teh	4	0.62	0.60	0.61
	Ahmad menonton televisyen sambil minum air teh	5	0.55	0.64	0.60
	Saya menonton televisyen sambil baring.	6	0.44	0.61	0.53

* For brevity, we don't provide the English translation of the tested sentences presented in Table 1

As mentioned earlier we do not consider morphological variant among words. However, further testing and analysis of the approach, found that morphological variants do play a significant role. For example, consider the following compared sentences and their respective similarity measures. The underlined words are the morphological variants in Malay although they referred to the same words when translated into English. In this case '*kahwin*' (married) is the root word for '*berkahwin'* (got married).

S_1 = Saya suka lelaki bujang itu. => I like that bachelor man
S_2 = Saya suka lelaki belum berkahwin itu. => I like that *unmarried* man
S_3 = Saya suka lelaki belum kahwin itu. => I like that *unmarried* man

$S(S_1, S_2) = 0.579$; $S(S_1, S_3) = 0.902$

As we can see, words that were stemmed to their root word give higher similarity measures. Therefore, aspects of morphological variants should be considered. The machine processing however might be the drawback if the morphological variants are to be considered.

We have also conducted some random testing by selecting pair of sentences of which the relevancies are known. Table 2 shows portion of the testing result. If the paired sentences are assumed to be relevant if the similarity measures > 0.5; then we have consistent result except for pair number 3. For pair number 3, the high similarity measures value is due to the high word-to-word similarity between '*tidur*' (sleep) and '*katil*' (bed).

Table 2. Similarities between selected sentences

	Sentences Pair	Sentences Pair (English Translation)	Rel.	Similarity Measure
1.	*Saya hendak tidur* *Saya sangat mengantuk*	I want to go to sleep I am very sleepy	Y	0.772
2.	*Ali sangat lapar* *Ali hendak makan*	Ali is very hungry Ali wants to eat	Y	0.643
3.	*Saya hendak tidur* *Saya bermain atas katil*	I want to go to sleep I am playing on the bed	N	0.667
4.	*Ahmad ke kuliah* *Ali belajar di kelas*	Ahmad went to a lecture Ali study in class	Y	0. 547
5.	*Tayar kereta pancit* *Tayar motosikal pancit*	Punctured car tyre Punctured motorcycle tyre	Y	0.939
6.	*Saya bermain di padang* *Ibu memasak kari ikan*	I am playing at the field Mother is cooking fish curry	N	0.133
7.	*Saya ada tukul* *Beri limau itu*	I have a hammer Give that lemon	N	0.314

5 Conclusions and Near Future Works

This semantic sentence similarity is important in many applications such as information retrieval, information extraction and ontology learning. While research in this area has been dominated for English language, little or no work has been focus for Malay language. This paper has presented an approach based on the work of [8] to provide a semantic measure for Malay sentences. This approach compares pair of sentences by first finding the similarity measures among words. The word-to-word similarity measure is derived from an on-online Malay dictionary using the overlap edge counting based method. The obtained word-to-word similarity measures are then used to construct the semantic vector and the word order vector. Lastly the sentence similarity is derived from the sum of the aforementioned vectors.

Initial experiments have shown consistent and encouraging results which indicate the potential use of this modified approach to practical applications previously mentioned. However more testing and evaluation works need to be conducted particularly involving real test data and human experts. Therefore further testing has been our main near future work.

As previously mentioned, morphological variants do provide significant similarity measures. Few stemming algorithms for Malay words are currently in existence such as [19, 20]. However, we need to further investigate on how significant morphological variants have in terms of sentence similarity. This is quite crucial due to the limitation of stemming algorithm relating to understemming and overstemming [21].

Our other future works include applying the sentence similarity measures to information retrieval activities of Malay documents. Apart from that the evaluation of word-to-word similarity should be extended to other method such as the term co-occurrence corpus base method and the semantic network which requires the construction of a linguistic ontology similar to WordNet.

Acknowledgments. The authors wish to thank the Ministry of Higher Education for the fund provided for this project and also the anonymous referees for their helpful and constructive comments of this paper.

References

1. Salton, G., Lesk.: Computer evaluation of indexing and text processing. Prentice Hall, Englewood Cliffs (1971)
2. Smucker, M.D., Allan, J.: Find-similar: similarity browsing as a search tool. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 461–468. ACM Press, New York (2006)
3. Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., Ma, J.: Learning to cluster web search results. In: SIGIR 2004 (2004)
4. Mooney, R.J., Bunescu, R.: Mining Knowledge from Text Using Information Extraction. SIGKDD Explorations 7(1), 3–10 (2005)
5. Buitelaar, P., Cimiano, P.: Bernardo Magnini Ontology Learning from Text: An Overview. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Evaluation and Applications Frontiers in Artificial Intelligence and Applications Series, vol. 123, IOS Press, Amsterdam, Trento, Italy (2005)
6. Cilibrasi, R., Vitanyi, P.M.B.: Similarity of objects and the meaning of words. In: Cai, J.-Y., Cooper, S.B., Li, A. (eds.) TAMC 2006. LNCS, vol. 3959, Springer, Heidelberg (2006)
7. Mihalcea, R., Corley, C., Strapparave, C.: Corpus based and knowledge based measures of text semantic similarity. In: Proceedings of the American Association for Artificial Intelligence (AAAI 2006) (2006)
8. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering 18(8), 1138–1150 (2006)
9. Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., Zobel, J.: Similarity Measures for Tracking Information Flow. In: Proceedings of the CIKM 2005, pp. 571–524 (2005)
10. Tatu, M., Moldovan, D.: A semantic approach to recognizing textual entailment. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 371–378 (2005)
11. Hamzah, M.P., Sembok, T.M.: Enhance retrieval of Malay documents by exploiting implicit semantic relationship between words. Enformatika 10, 89–94 (2005)
12. Turney, P.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the 12th European Conference on Machine Learning (2001)
13. Karov, Edement.: Similarity-based Word Sense Disambiguation. Computational Linguitics 24(1), 41–59 (1998)

14. Leacock, C., Chodorow, M.: Combining local context and WordNet sense similarity for word sense identification. WordNet, An Electronic Lexical Database. The MIT Press, Cambridge (1998)
15. Resnik, P.: Using information content to evaluate the semantic similarity. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (1995)
16. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone (1986)
17. Miller, G.A.: WordNet: a lexical database for English. Communication of the ACM 38(11), 39–41 (1995)
18. Wiemer-Hastings, P.: Adding syntactic information to LSA. In: Proceedings of the 2nd Annual Conference on Cognitive Science, pp. 989–993 (2000)
19. Ahmad, F., Yusoff, M., Sembok, T.M.T.: Experiments with a Stemming Algorithm for Malay Words. JASIS 47(12), 909–918 (1996)
20. Othman, A.: Pengakar perkataan melayu untuk sistem capaian dokumen. MSc Thesis. National University of Malaysia (1993)
21. Xu, J., Croft, W.B.: Corpus-based stemming using coocurrence of word variants. ACM Transactions on Information Systems 16(1), 61–81 (1998)