International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:10, 2007

Enhancing Retrieval Effectiveness of Malay Documents by Exploiting Implicit Semantic Relationship between Words

Mohd Pouzi Hamzah, and Tengku Mohd Tengku Sembok

Abstract**—**Phrases has a long history in information retrieval, particularly in commercial systems. Implicit semantic relationship between words in a form of BaseNP have shown significant improvement in term of precision in many IR studies. Our research focuses on linguistic phrases which is language dependent. Our results show that using BaseNP can improve performance although above 62% of words formation in Malay Language based on derivational affixes and suffixes.

Keywords**—**Information Retrieval; Malay Language; Semantic Relationship; Retrieval Effectiveness; Conceptual Indexing.

I. Introduction

THE use of phrases as part of a text representation or indexing language has been investigated since the early days of information retrieval research. Cleverdon, for example, included phrase-based indexing in the Cranfield studies.

The goal of an information retrieval system is to locate relevant documents in response to a user's query. Documents are typically retrieved as a ranked list, where the ranking is based on estimations of relevance [3,8]. The retrieval model for an information retrieval system specifies how documents and queries are represented and how these representations are compared to produce relevance estimates. The performance of the system is evaluated with respect to standard test collections that provide a set of queries, a set of documents, and a set of relevance judgments that indicate which documents are relevant to each query [6,8]. These judgments are provided by the users who supply the queries and serve as a standard for evaluating performance. Information retrieval research is concerned with finding representations and methods of comparison that will accurately discriminate between relevant and non relevant documents.

Many retrieval systems represent documents and queries by the words they contain, and base the comparison on the number of words they have in common. The more words the query and document have in common, the higher the document is ranked.

Performance is improved by weighting query and document words using frequency information from the collection and individual document texts [10]. There are two problems with using words to represent the content of documents. The first problem is that words are ambiguous and this ambiguity can cause documents to be retrieved that are not relevant. The second problem is that a document can be relevant even though it does not use the same words as those that are provided in the query. The user is generally not interested in retrieving documents with exactly the same words, but with the concepts that those words represent. Retrieval systems address this problem by expanding the query words using related words from thesaurus.

The work of Fagan [7] has shown that implicit relationship between words can be exploited to enhance retrieval effectiveness. This implicit semantic relationship is in a form of phrase.

Phrases constructed with statistical approaches are usually called "statistical phrases", for some of these phrases are judged as "non-phrases" by human beings, i.e, they may not have the grammatically or semantically right structures.

Linguistic approaches usually start from studying the internal structure of phrases and the grammatical functions of the phrase components with the hope of finding rules or patterns of phrase construction. Syntactic analysis process, analyzes the whole or part of the sentence syntactic structure. Phrases are viewed as building blocks or sub-level structures of the sentential structure, therefore, phrase extraction methods developed from linguistic approaches are aimed at finding well-defined "syntactic phrases" on the basis of sentence parsing.

Yun et al. [13] propose a method of extracting Korean compounds by identifying their boundaries. A Korean compound noun consists of a set of "unit nouns" of which each is composed of 1-7 syllables. Adjacent syllables can be grouped in different ways, and hence producing different unit nouns.

The goal of phrase identification is to find out the rules of forming phrases. Unfortunately, such rules are not readily usable by machines. If machines want to use them, they should rely on many other factors such as the tagging of a word and the semantic relationships between phrase components.

Manuscript received November 21, 2005.

Mohd Pouzi Hamzah is with Department of Computer Science, University College of Science & Technology Malaysia, 21030 K.Terengganu, Malaysia (phone: 609-6683348; fax: 609-6694660; e-mail: mph@kustem.edu.my).

Tengku Mohd Tengku Sembok is with Department of Information Science, National University of Malaysia, 43600 UKM, Malaysia (e-mail: tmts@ftsm.ukm.my).

This paper will present a kind of phrase that can be used as indexes and to design an easy-to-be-implemented phrase extraction method. The rules summarized by traditional Malay grammarians are not readily usable to detect phrases. Since their purpose of compiling these rules are not for the use of automatic detection. They are used to describe the formulation of Malay phrases. Therefore, a different approach to phrase extraction will be adopted for the purpose of computer processing of languages.

Previous IR research findings suggest that language units that are used as indexes need not be of perfect grammatical structures. What is important is whether such units can capture the document content. This makes it possible to look at the to-be-analyzed text from a non-conventional angle. Since for IR purposes conceptual words have been proved to be good content descriptors while non-conceptual words can be "stopped" from entering the content representations, the idea sounds feasible of combining conceptual words into larger chunks and then using such chunks as complex indexes. *BaseNP* is such a kind of phrase that is suitable for indexing purposes.

II. BaseNP Features

The Malay BaseNP is then defined as a sequence of adjacent conceptual words that are exclusive of functional words or other units of complicated syntactic structure (e.g., a clause). The sequence should express a more specific concept than any of its individual components. BaseNPs should have the following features:

Feture 1 - Conceptual elements only

A BaseNP is composed of conceptual words only, including nouns, verbs and adjectives. These words are considered as appropriate for indexing. A BaseNP should not contain functional words like prepositions, auxiliaries, conjunctives, etc.

Feature 2 - Specific Meaning

BaseNP should express an integrated concept. If from the sequence of conceptual words, there can not be derived a specific meaning, that sequence is not a BaseNP.

Featuee 3 - Minimum of two words

The sequence must be at least 2-word long to constitute a BaseNP. The maximum length is not specified. This means that a single word can not form a BaseNP, no matter how many characters it may have.

Feature 4 - Sequential Adjacency

The sequence of conceptual words should be consecutive, i.e., appearing adjacently in a clause or sentence.

Feature 5 – Normalized Form

BaseNPs in normal form can contain the word "yang" or "untuk". For example, "Rumah untuk Ali" can be normalized to "Rumah Ali" by dropping the word "untuk" and likewise "Rumah yang besar" can be transformed to "Rumah besar".

III. BaseNP Structure

The relationship among BaseNP components is interpreted as that of "modification or specification" , very similar to Fagan's phrase definition where first component modifies the second, as in the case of "computer science" where "science" is made more specific due to the modifier "computer"[7].

Based on this interpretation, two kinds of BaseNP components can be defined: one is "modifier" and the other "head".

In a BaseNP, the component that modifies other components is called a "modifier" and the modified component is referred to as a phrase "head", hence the fundamental format of a BaseNP can be defined:

$$BaseNP = Head + modifier*$$

where the asterisk mark "*" denotes zero or more number of modifiers and

$$modifier = [N,V,A]$$
$$head = N$$

Where N is for noun, V for verb and A for adjective.

According to the definition above, a BaseNP must have at most one head and at least one modifier. The upper limit of the number of modifiers is not set, though. The first element is definitely a head while all other elements after the head are modifiers.

IV. Test Collection

For this study we use a new Malay test collection consisting of 811 documents, 39 natural language query statements and a set of expert-defined relevance judgment. Statistics of the documents and query collections are given below.

TABLE I
STATISTICS OF DOCUMENT COLLECTION

Item	Statistics
Total Number of Documents	811
Total Number of Sentences	6236
Total Number of Words	204,971
Maximum number of Sentences in a Document	69
Minimum number of Sentence(s) in a Document	1
Average sentence/document	7.689
Average word/document	252.74

TABLE II
STATISTICS OF QUERY COLLECTION

Item	Statistics
Total Number of Queries	39
Total Number of Words	253
Maximum number of Words in a Query	14
Minimum number of Words in a Query	3
Average Word/Query	6.487

V. Similarity Measure

In this study we adopt Vector Space Model as a retrieval model [9,10,11]. Similarity between query and document is based on query terms vector and document terms vector. The weight of each term is given by the following equation:

$$W_{ij} = tf_{ij} \times idf_i \quad (1)$$

$$tf_{ij} = \frac{frek_{ij}}{frek_m} \quad (2)$$

$frek_{ij}$ – no. of terms i in document j

$frek_m$ – total terms in document j

$$idf_i = \ln\left(\frac{N}{n_i}\right) \quad (3)$$

n_i – no. of documents where term i exist.

N – no of documents in a collection.

A weight for each query term is as follows [2]:

$$W_{iq} = (0.5 + 0.5 * tf_{iq}) \times idf_i \quad (4)$$

Similarity between document and query is given by cosine of the angle between the two vectors.

The Cosine of the angle between query vector and document vector is given by the following equation:

$$\cos\theta = \frac{\sum\left(W_{ij} * W_{iq}\right)}{\sqrt{\sum W_{ij}^{\ 2}} * \sqrt{\sum W_{iq}^{\ 2}}} \quad (5)$$

We use two set of vectors to represent single terms and BaseNPs. In equations 1 to 4, term refers to single terms in one vector and BaseNP in another vector. Hence similarity between queries and documents is given by equation (6):

$$sim(q,d) = c_1 * Cos\,\theta_{qd(s)} + c_2 * Cos\,\theta_{qd(f)} \quad (6)$$

In our present study no special treatment is given to BaseNP, so c1 and c2 are set to 1.

VI. Normalization of Words

To increase recall and precision depend on finding a way for non-identical terms to match. The traditional approach is through *normalization*, replacing several forms with a single canonical form. Stemming is one of the normalization based on morphology, for example:

tulis (write), tulisan (writing), penulis (writer) → tulis(write).

The formation of words in Malay differs from those languages like English and French where new word forms are created using a *root* with the addition of *derivational affixes*, and not using a stem with the addition of derivational suffixes like the two languages above [4,14,15]. Malay is characterized by a wider range of derivational affixes than is English.

The Malay stemmer has been implemented in order to carry out this experiment. It is a rule-based stemmer with 420 basic rules and 42 special rules to cater morphological variations. According to Discrimination Model [9,11], high frequency words especially function words are not good document descriptors and need to be removed from index. In Malay language, there are altogether 314 function words considered as stop words. The result of the stemming process after removal of stop words is as shown in Table III.

TABLE III
TERMS DISTRIBUTION BEFORE AND AFTER STEMMING

	No. of Terms
Number of unique terms before stemming	6907
Number of unique terms after stemming	2575

As can be seen from Table III above, a compression of 62.72% of unique terms can be achieved when stemming is applied.

In our implementation, every word in documents and queries will be normalized to its root word.

VII. Evaluation

There are many ways to evaluate document retrieval systems [12]. In our experiment we use precision at standard recall points and R-precision to compare effectiveness of the systems[1].

$$\text{Recall(R)} = \frac{\text{number of documents retrieved and relevant}}{\text{total relevant documents from collection}}$$

$$\text{Precision(P)} = \frac{\text{number of documents retrieved and relevant}}{\text{total documents retrieved from collection}}$$

To further compare effectiveness of the systems, we use R-precision that is the precision at the R-th position in the ranking of results for a query that has R relevant documents [2].

VIII. Result

The experimental results in Table IV show standard recall (R), precision (P) for retrievals using BaseNP+Stemmed Word and Stemmed word alone.

Method 1 – Stemmed word
Method 2 – Stemmed word + BaseNP

To visualize the difference, Fig. 1 shows method 1 yields lower precision at every recall point. Method 2 produces better result. If we take average precision to compare overall performance of the similarity measures, method 2 is superior to method 1 +7.17%.

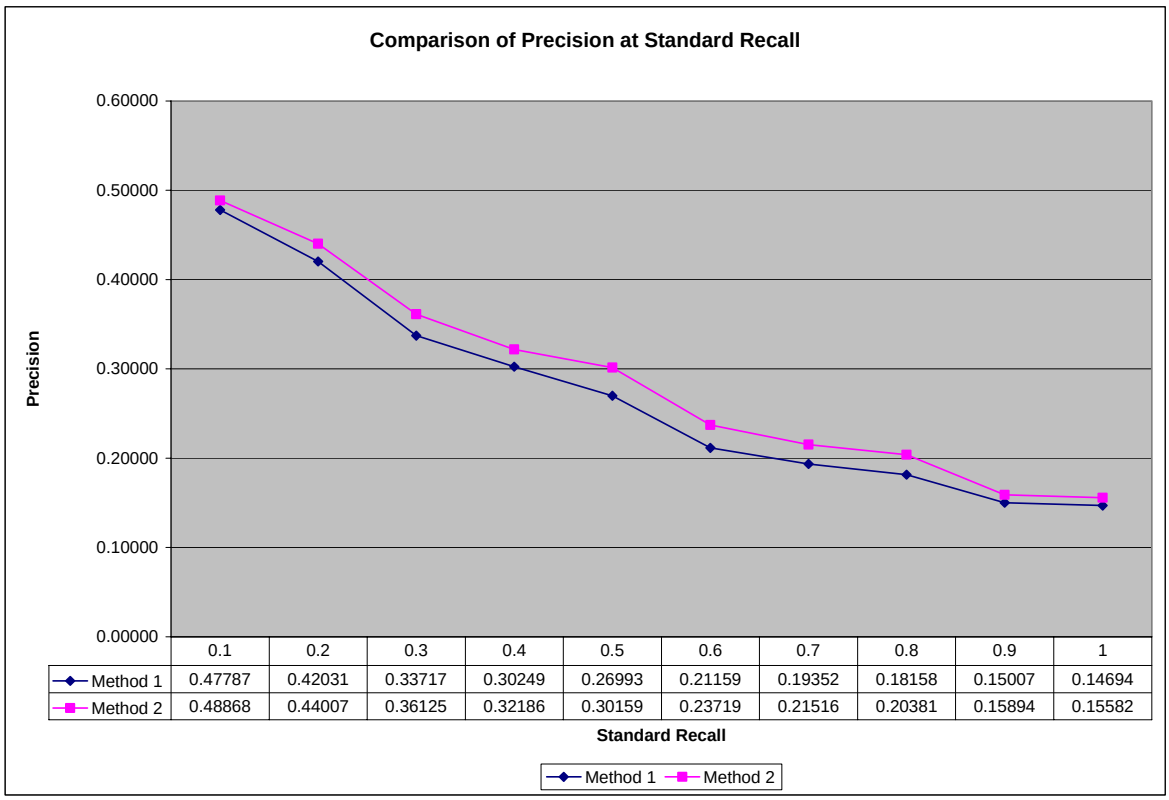
Table V presents the results obtained after calculating R-precision for each method. On average R-precision, Stemmed word+BaseNP is +0.24032% superior to Stemmed Word alone. If we look at individual query, 4 queries yield positive result and only one query negatively affected by BaseNP. The percentage of improvement is 80%.

TABLE IV
PRECISION AT STANDARD RECALL

Recall	Precision		% Increment
	Method 1	Method 2	
0.1	0.47787	0.48868	2.26124
0.2	0.42031	0.44007	4.70255
0.3	0.33717	0.36125	7.14203
0.4	0.30249	0.32186	6.40560
0.5	0.26993	0.30159	11.72901
0.6	0.21159	0.23719	12.09855
0.7	0.19352	0.21516	11.18019
0.8	0.18158	0.20381	12.24107
0.9	0.15007	0.15894	5.90978
1.0	0.14694	0.15582	6.04067
Average	0.26915	0.28844	7.16693

TABLE V
R-PRECISION FOR EACH QUERY

Query	Method 1	Method 2	% Increment
1	0.14286	0.14286	0.00000
2	0.40000	0.50000	25.00000
3	0.22727	0.22727	0.00000
4	0.16667	0.16667	0.00000
5	0.00000	0.00000	0.00000
6	0.33333	0.42857	28.57229
7	0.50000	0.50000	0.00000
8	0.85714	0.85714	0.00000
9	0.30769	0.33333	8.33306
10	0.12500	0.12500	0.00000
11	0.33333	0.33333	0.00000
12	0.00000	0.00000	0.00000
13	0.00000	0.00000	0.00000
14	0.25000	0.25000	0.00000
15	0.25000	0.25000	0.00000
16	0.28571	0.28571	0.00000
17	0.00000	0.00000	0.00000
18	0.41667	0.05202	-87.51530
19	0.40000	0.40000	0.00000
20	0.50000	0.50000	0.00000
21	0.50000	0.50000	0.00000
22	0.00000	0.00000	0.00000
23	0.20000	0.20000	0.00000
24	1.00000	1.00000	0.00000
25	0.00000	0.00000	0.00000
26	0.00000	0.00000	0.00000
27	0.00000	0.00000	0.00000
28	0.33333	0.33333	0.00000
29	0.00000	0.16667	16.66700
30	0.00000	0.00000	0.00000
31	1.00000	1.00000	0.00000
32	1.00000	1.00000	0.00000
33	0.00000	0.00000	0.00000
34	0.00000	0.00000	0.00000
35	0.00000	0.00000	0.00000
36	0.00000	0.00000	0.00000
37	0.00000	0.00000	0.00000
38	0.00000	0.00000	0.00000
39	0.00000	0.00000	0.00000
Average	0.24433	0.24492	0.24032



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Method 1	0.47787	0.42031	0.33717	0.30249	0.26993	0.21159	0.19352	0.18158	0.15007	0.14694
Method 2	0.48868	0.44007	0.36125	0.32186	0.30159	0.23719	0.21516	0.20381	0.15894	0.15582

Fig. 1 Comparison of Precision at Standard Recall Points

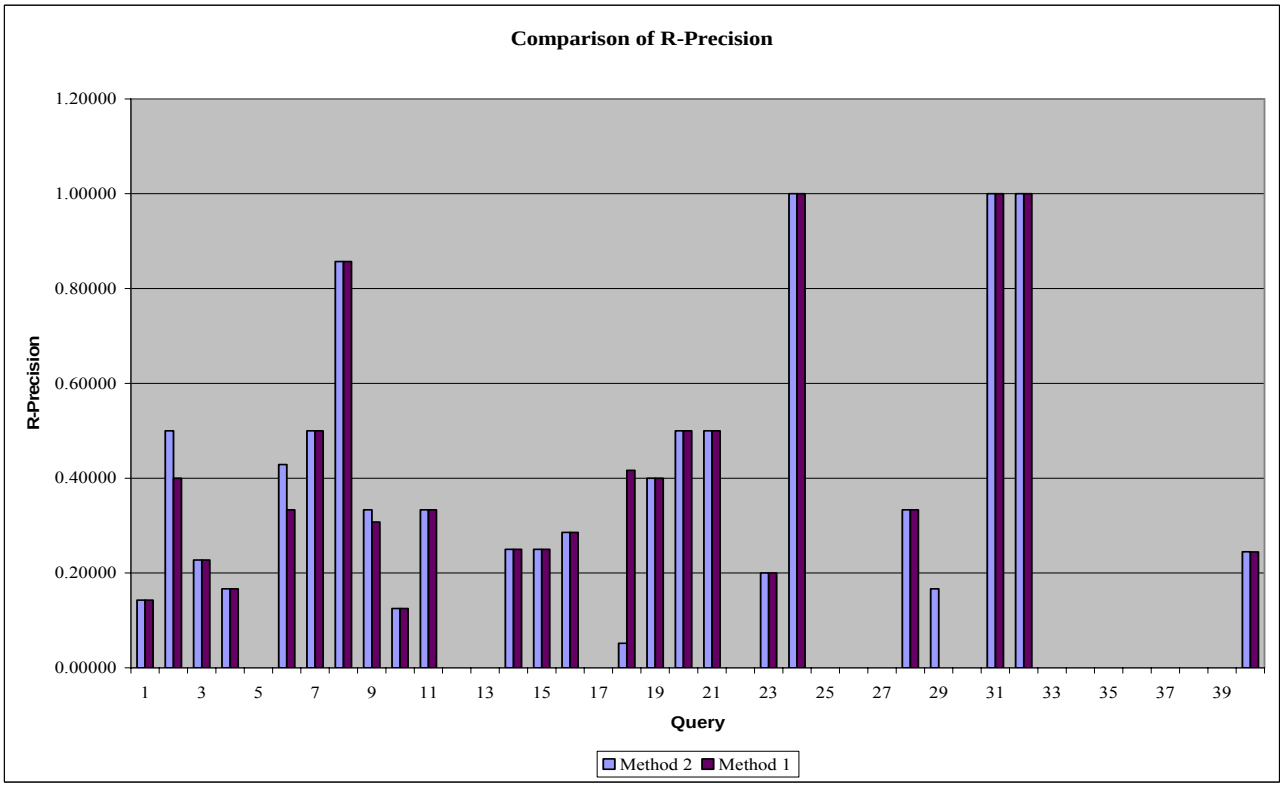


Fig. 2 Comparison of R-Precision

IX. Conclusion

In this paper we have described the use of BaseNP in information retrieval research on Malay documents. Based on the results in section VIII, we can accept the hypothesis that phrases in a form of BaseNP improve retrieval effectiveness significantly. The formation of BaseNP is done using the definition that we presented and comply to the set of features that we described.

References

[1] Atlam, E.S., Fuketa, M., Morita, K., & Aoe, J., Documents Similarity Measurement Using Field Association Terms, *Information Processing and Management Journal*, 39, 2003, pp. 809-824.

[2] Baeza-Yates, R & Ribeiro-Neto, B., *Modern Information Retrieval,* Addison-Wesley, New York, 1999.

[3] Croft, W. B., User-specified Domain Knowledge for Document Retrieval, *Proceedings Of The ACM Conference On Research And Development In Information Retrieval*, 1986, pp. 201-206.

[4] Fatimah A., *A Malay Language Document Retrieval System: An Experimental Approach And Analysis*, Ph.D Thesis, Universiti Kebangsaan Malaysia, 1995

[5] Fagan, J. L, *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*, Ph.D. Thesis, Department of Computing Science, Cornell University, Ithica, New York, 1987.

[6] Lewis, D.D. and Jones, K.S., Natural Language Processing for Information Retrieval, *Communication of the ACM*, Vol 39 No. 1 , 1996, pp. 92-100.

[7] Sanderson, M. ,Word Sense Disambiguation and Information Retrieval, *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* 1994, pp. 142-151, Springer-Verlag.

[8] Salton, G., A Blueprint For Automatic Indexing, *ACM SIGIR Forum 16*, 2 (Fall 1981), 1981, pp. 22-38.

[9] Salton, C.. and Lesk., M.E. Computer Evaluation Of Indexing And Text Processing, *Communication of the ACM*, Vol 15 No. 1 , 1986, pp. 6-36.

[10] Salton, G., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[11] Salton, G., Another Look At Automatic Text Retrieval Systems, *Communications of the ACM*, Vol 29 No. 7, 1986, pp. 648-656.

[12] Van Rijsbergen, C.J. *Information Retrieval*, 2nd edition, Butterworth.,1979.

[13] Yun, B. H., H. S. Lim and H.C. Rim, Analysis of Korean Compound Nouns using Statistical Information, *Proc. of the 22nd Korea Information Science Society Spring Conference*, 1994, pp 925-928.

[14] Zainab Abu Bakar, *Evaluation Of Retrieval Effectiveness Of Conflation Methods On Malay Documents*, Ph.D Thesis, Universiti Kebangsaan Malaysia, 1999.

[15] Zainab Abu Bakar & Nurazzah Abdul Rahman, Evaluating The Effectiveness Of Thesaurus And Stemming Methods In Retrieving Malay Translated Al-Quran Documents, *Proceeding Of 6th International Conference On Asian Digital Libraries*, 2003, pp. 653-662. Springer-verlag.