Malay Interrogative Knowledge Corpus

Fatimah Sidi[1], Marzanah A. Jabar [2], Mohd Hasan Selamat[2], Abdul Azim Abdul Ghani[2], Md Nasir Sulaiman[1], Salmi Baharom[2]

[1]Department of Information System
[2] Department of Computer Science
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia (UPM),
43400 UPM Serdang,
Selangor, MALAYSIA
{fatimah, marzanah, hasan, azim, nasir, salmi}@fsktm.upm.edu.my

ABSTRACT

The growth in the number of documents written in Malay language is enormously available on the web and intranets. There is a need to identify the information in the Malay documents that contain knowledge. This triggers the need to investigate the availability of knowledge in them. This study uses interrogative theory to identify knowledge from documents or texts. The results are expected to lead towards establishment of new set of interrogative rules for Malay corpus. This study contributes the interrogative knowledge identification thru the development of Malay Interrogative Knowledge Corpus (MalayIK-Corpus). It facilitates to explicitly capture and make available Malay knowledge representation in a knowledge-base system.

Keywords
Interrogative Theory, Knowledge Identification, Document

1.0 INTRODUCTION

The development of the Malay Interrogative Knowledge Corpus (Malay*IK*-Corpus) is due to unavailable public domain utilities or tools for Malay language to codify computational grammar and collect morphological rules, semantic or syntactic templates. Even, there is no public domain parser to analyze Malay texts and general computational lexicon for Malay words. Ahmad (1995) reports that the use of dictionary for Malay words is inevitable as far as Malay documents are concerned. Unfortunately, there is no Malay corpus that has been published yet except a dictionary of root words which contain 22,433 entries (Ahmad, 1995; Abdullah, 2006).

Therefore, the development of Malay*IK*-Corpus has to manually modify the dictionaries into a Malay*IK*-Corpus. Firstly this paper presents the development of the corpus. Then, it highlights stop words and the development of stop words list in texts processing and follow by results and discussion. Finally is the conclusion.

2.0 DEVELOPMENT OF THE CORPUS

The Malay*IK*-Corpus is a Malay language corpus where the Malay dictionary of *Kamus Dewan* (Dewan Bahasa Perpustakaan, 2002; 2005) and the dictionary of root words act as important secondary controls of the lexicon entries. It is derived from 6,000 word entries (about 4,000 root words and 2,000 derivations). It also refers to the dictionary of *Kamus Imbuhan Bahasa Melayu* (Ali *et al.*, 1993), *Kamus Dwibahasa Oxford Fajar* (Hawkins, 2001), and *Kamus Komprehensif Bahasa Melayu* (Othman, 2005). Besides, books on Malay language are also used in preparing the grammatical information entries (Ahmad, 2001; Asraf, 2002; Latif & Rashid, 2003).

It looks upon the interrogative theory of knowledge identification and representation as the background theory for the foundation of the Malay*IK*-Corpus development. The interrogative-based approach is described as the "who, when, what, where, how and why" analysis (Quigley & Debons, 1999). It makes distinctions between data, information, and knowledge. The Malay*IK*-Corpus used grammatical information of lexicon to answers the interrogative-based question. The "when/where/who/what" identifies the information. The "how/why" identifies the knowledge. While the grammatical information of lexicon that answers no question identifies data. Hence the most important attribute is the grammatical information of lexicon entry to answer the question of the lexicon grammatical information interrogatively besides the root word.

2.1 Attributes of the Interrogative Knowledge Corpus

For the purpose of this development, Microsoft Access is used as a database for the Malay*IK*-Corpus. It is easier to maintain and develop because the lexicon capacity is not huge. The task is merely done to create and update information of lexicons for the corpus. Some other available databases or tools that can also be used according to the needs of the task are Oracle, SQL Server, XML and others. The lexicons

entries are manually inserted in the database using standard Data Manipulation Language (DML) of the related database. Each entry of the Malay*IK*-Corpus contains attributes of:

i. root word *(kata dasar)*);
ii. lexicon *(perkataan)*;
iii. grammatical information of lexicon entry *(kata masuk)*;
iv. interrogative element *(elemen interogatif)* - may consists of either what (*apa)*, when *(bila)* (when), who (*siapa)*, where *(di mana)*, why (*mengapa)* or how *(bagaimana)* which answers the grammatical information of the word entry; and
v. status - indicates status of the lexicon for processing purposes which includes stop words. Status 1 indicates noun (*kata nama am)* or adjective (*adjektif)* while status 2 indicates stop word.

In order to create a general purpose corpus for Malay, the Ahmad's and Abdullah's stop words (Ahmad, 1995; Abdullah, 2006) are included which indicate pronoun, auxiliary verb, adverb, predicate, preposition, negative, conjunction, relative and determinant.

2.2 Classification of the Word Entry

Table 1 presents examples of words entry extracted from Malay*IK*-Corpus in a table format (by columns and rows). The header row of Table 1 refers the attributes of corpus by columns. The rest of the rows are examples of words entries for *'rumah'* (house), *'sejak'* (since), *'penyelidik'* (researcher), *'di'* (at), *'kerana'* (because) and *'dengan'* (with). It answers the question of interrogative of *'apa'* (what), *'bila'* (when), *'siapa'* (who), *'di mana'* (where), *'mengapa'* (why), and *'bagaimana'* (how) respectively.

Table 1: Entries of MalayIK-Corpus.

Root Word	Lexicon	Grammatical Information	Interrogative Element	Status
rumah (house)	*rumah* (house)	*kata nama am benda* (noun)	*apa* (what)	*1*
sejak (since)	*sejak* (since)	*kata sendi nama masa* (preposition)	*bila* (when)	*2*
selidik (research)	*penyelidik* (researcher)	*kata nama am orang* (noun)	*siapa* (who)	*1*
di (at)	*di* (at)	*kata sendi nama tempat dan arah* (preposition)	*di mana* (where)	*2*
kerana (because)	*kerana* (because)	*kata hubung pancangan* (conjunction)	*mengapa* (why)	*2*
dengan (with)	*dengan* (with)	*kata sendi nama bersama-sama* (preposition)	*bagaimana* (how)	*2*

Basically, the grammatical information of *'rumah'* and *'penyelidik'* is noun *('kata nama am')* but classified as different category. The word *'rumah'* (house) falls under categorization of 'Things' which answers the interrogative question of 'what'. While *'penyelidik'* (researcher) falls under categorization of 'People' which answers the interrogative question of 'who'. However, in Malay language, *'sejak'* and *'kerana'* which answer the interrogative question of *'bila'* (when) and *'mengapa'* (why) are conjunctions. The word entry of *'di'* and *'dengan'* are prepositions which answer the interrogative question of *'di mana'* (where) and *'bagaimana'* (how) respectively. Those words of when, why, where and how are listed as stop words. Since, there is no computational grammatical information available in public domain, the interrogative element of Malay*IK*-Corpus has to define all of them primarily in the corpus. The following are steps taken in building up the Malay*IK*-Corpus:

i. create attributes for corpus;
ii. extract lexicons from the document collection;
iii. verify the lexicons entries with Malay language expert;
iv. insert lexicons entries in the database; and
v. extend words encountered which are ambiguous or unclear in its context of answering the interrogative question, then the opinion of the Malay language expert will be referred.

3.0 STOP WORD LIST

Stop words, or stopwords, is a name given to words which are filtered out prior to, or after, processing of text. A stop word list (stoplist) is a set of or list of stop words which is typically language specific, although it may contain words (and other character sequences like numbers and punctuations). A search engine or other natural language processing system may contain a variety of stop lists, one per language, or it may contain a single stop list that is multilingual. These stop words are poor discriminators and cannot possibly be used by them to give any hint value and identify document content. Hence, they are eliminated from the set of index terms (van Rijsbergen, 1979) in

search engine or document retrieval system. Salton and McGill (1983) report that such words comprise about 40% to 50% of a collection of documents text words. There is no definite list of stop words, which all natural language processing tools incorporate. Not all NLP tools use a stop list. Some tools specifically avoid the use of a stop list in order to support phrase searching.

3.1 Development of a Stop Word List

A list of stop words is included in the development of the Malay*IK*-Corpus, in order to eliminate words which have no values. The development of a stop words list in Malay*IK*-Corpus adopts approaches used by van Rijsbergen (1979). The purpose is for identification of such stop words list having the same aim to find those of no values. The approach used is the combination of manual selection method and statistical counting of high frequent words. The statistical method of occurrences is to find words of high and very low number of occurrences that are taken as stop words. The total numbers of 6,479 words are extracted from the test collection of Malay unstructured documents collection. The extracted words are ranked by frequency of occurrence in decreasing order.

3.2 Foundation of the Stop Word List

Table 2 presents a list of the 35 most frequently occurring words in the test collection documents.

Table 2: A list of the 35 most frequently occurring words.

Rank	Lexicon	Frequency	%	Malay*IK*-Corpus Status
1	*dan*	190	2.9	stop word
2	*yang*	170	2.6	stop word
3	*di*	131	2.0	stop word
4	*ini*	76	1.2	stop word
5	*dengan*	70	1.1	stop word
6	*itu*	58	0.9	stop word
7	*tidak*	53	0.8	stop word
8	*kita*	51	0.8	stop word
9	*dalam*	50	0.8	stop word
10	*dari*	43	0.7	stop word
11	*untuk*	43	0.7	stop word
12	*halal*	41	0.6	adjective
13	*kepada*	38	0.6	stop word
14	*mereka*	38	0.6	stop word
15	*juga*	37	0.6	stop word
16	*pada*	37	0.6	stop word
17	*bagi*	34	0.5	stop word
18	*pertanian*	33	0.5	noun
19	*akan*	31	0.5	stop word
20	*umat*	29	0.4	noun
21	*telah*	28	0.4	stop word
22	*tetapi*	28	0.4	stop word
23	*seperti*	27	0.4	stop word
24	*makanan*	26	0.4	noun
25	*negara*	26	0.4	noun
26	*oleh*	26	0.4	stop word
27	*rakyat*	26	0.4	noun
28	*ada*	25	0.4	stop word
29	*dunia*	24	0.4	noun
30	*berkata*	23	0.4	verb
31	*ke*	23	0.4	stop word
32	*daripada*	22	0.3	stop word
33	*beliau*	21	0.3	stop word
34	*bukan*	21	0.3	stop word
35	*boleh*	20	0.3	stop word
Total number of words		6,479		

Table 2 shows that the most frequent lexicons in the test collection documents are conjunction of *'dan'* (and), relative of *'yang'* (which), and preposition of *'di'* (at). These words are created by Ahmad (1995) and Abdullah (2006) as stop words. This shows that these words are function words and commonly appeared in any text documents. Abdullah (2006) reports that inclusion of these words in the list of stop words comply with the fact that these words will not contribute to the content of the collection. The reason being, these words will mark the whole collection as relevant document in a query. With that, it complies with the fact that these words need to be eliminated in order to build up knowledge representation. However, in constructing phrases and identifying interrogative

elements of when, where, why and how, the stop words list is being avoided for its usage.

The stop words list that is created by Ahmad (1995) contains 314 entries and 20 entries from Abdullah (2006). This makes a total of 334 entries of Malay stop words originated from Quranic documents. It is interesting to note that content-bearing words, i.e., *'pertanian'* (agriculture), *'halal'* (lawful), and *'makanan'* (food), also appear in Table 2. Their high positions derive from the fact that the lengthiest documents in the test collection documents is from newspaper which reports on the main domain of agriculture.

4.0 RESULTS & DISSCUSION

Interrogative Knowledge Identification Framework is used to address the need for the mechanism to identify knowledge from unstructured document (Sidi *et al.*, 2009). They used lexicon interrogative analysis to identify and extract knowledge in each of the complete sentences written in the document. It is also used to extract interrogative lexical constructs from the individual unstructured document. Each of the lexicons is analyzed with lexicon interrogative analysis matching rules of Malay*IK*-Corpus using the standard DML. The DML is used to analyze, check and insert the lexicon into interrogative annotation as interrogative lexical construct if it exists. Any new lexicon analyzed and existed is inserted and defined primarily in Malay*IK*-Corpus.

The sample used in this experiment, 15% of 42,733 words from Malay*IK*-Corpus are sufficient and justified to produce better results in extracting identified knowledge. It is more than the suggested by Gay and Airasian (2003, page 113) for sample of more than 5,000 units, a sample size of 400 (8%) should be adequate. The results obtained are measured in terms of percentage of quantitative retrieval performance recall and precision metrics (Baeza-Yates & Ribeiro-Neto, 1999) coupling with research methods and concept in information system research (Jabar, 2009). The accuracy of the knowledge extracted is measured by precision (fraction of the retrieved knowledge which has been relevant), and recall (fraction of the relevant knowledge which has been retrieved). Comparison of results is done with an expert evaluation. The Malay documents collection is given to the expert to identify the knowledge that resides in the collection interrogatively.

Results of the experiments in the form of precision and recall tables were explained in detail in Sidi (2008). The interrogative element of why has shown a significant accuracy in identifying knowledge. Unfortunately, it is not true for the interrogative element of how. Both these interrogative elements are used to identify knowledge within the text in unstructured document. Moreover, the analysis of results has also confirmed significant accuracy in identifying and extracting information for the interrogative elements of what and who. Unfortunately, the accuracy differences are not significant for the interrogative elements of where and when. The reasons for the performances differences are possibly caused by the quality of various formats and styles of writing the Malay documents collection used.

5.0 CONCLUSION

The paper presents a development of Malay*IK*-Corpus to identify knowledge in documents. It facilitates to identify and explicitly capture and make available Malay knowledge representation in a knowledge-base system. This leads to potential increase sharable and reusable of the knowledge in documents among the community. However, the Malay*IK*-Corpus is lacking of ease for navigation in its system interface. It is not fully automated on the creation of the Malay corpus.

REFERENCES

Abdullah, M. T. (2006). *Monolingual and Cross-Language Information Retrieval Approaches for Malay And English Language Documents.* PhD Thesis. Universiti Putra Malaysia.

Ahmad, F. D. (1995). *A Malay Language Document Retrieval System: An Experimental Approach and Analysis.* PhD Thesis. Universiti Kebangsaan Malaysia.

Ahmad, S. (2001). *UPSR Bahasa Malaysia.* Petaling Jaya, Malaysia: Sasbadi Sdn. Bhd.

Ali, H. M., Shariff, M. N. M., & Dewa, W. M. W. (1993). *Kamus Imbuhan Bahasa Melayu Edisi Kedua.* Kuala Lumpur, Malaysia: Penerbit Fajar Bakti Sdn. Bhd.

Asraf. (2002). *UPSR Tatabahasa.* Petaling Jaya, Malaysia: Sasbadi Sdn. Bhd.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* New York: Addison Wesley.

Dewan Bahasa Perpustakaan. (2002). Kamus Dewan Edisi Ketiga. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Dewan Bahasa Perpustakaan. (2005). Kamus dewan edisi keempat. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Gay, L. R., & Airasian, P. (2003). *Educational Research: Competencies for Analysis and Application* (7th edition) .Merrill, New Jersey: Upper Saddle River.

Hawkins, J. M. (2001). *Kamus Dwibahasa Oxford Fajar* Edisi Ketiga. Kuala Lumpur, Malaysia: Penerbit Fajar Bakti Sdn. Bhd.

Jabar, M. A., Sidi, F., Selamat, M. H., Ghani, A. A. A., & Ibrahim, H. (2009). An Investigation into Methods and Concepts of Qualitative Research in Information System Research. *Computer and Information Science, 2*(4), 47-54.

Latif, M., & Rashid, A. (2003). *Tatabahasa Melayu.* Shah Alam, Malaysia: SNP Panpac (M) Sdn. Bhd.

Othman, A. (2005). *Kamus Komprehensif Bahasa Melayu*. Selangor Darul Ehsan, Malaysia: Penerbit Fajar Bakti Sdn. Bhd., a subsidiary of Oxford University Press.

Quigley, E. J., & Debons, A. (1999). Interrogative Theory of Information and Knowledge. *In the Proceeding of the 1999 ACM SIGCPR Conference on Computer Personnel Research.*New Orleans, Louisiana, United States. pp. 4-10.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sidi, F. (2008). *Transformation of Extracted Knowledge in Malay Unstructured Documents into an Interrogative Structured Form.* PhD Thesis, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia.

Sidi, F., Jabar, M. A., Selamat, M. H., Ghani, A. A. A., & Sulaiman, M. N. (2009). Framework for Interrogative Knowledge Identification. *Computer and Information Science, 2*(4), 109-115.

van Rijsbergen, C. J. (1979). *Information Retrieval. 2nd Edition.* London: Butterworths.