American Journal of Economics and Business Administration 3 (1): 171-176, 2011
ISSN 1945-5488

Malay Interrogative Knowledge Corpus

[1]Fatimah Sidi, [2]Marzanah A. Jabar, [2]Mohd Hasan Selamat,
[2]Abdul Azim Abdul Ghani, [1]Md Nasir Sulaiman and [2]Salmi Baharom
[1]Department of Information System,
[2]Department of Computer Science,
Faculty of Computer Science and Information Technology,
University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Abstract: Problem statement: 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. **Approach:** This study uses interrogative theory to identify knowledge from documents or texts. **Results:** The results are expected to lead towards establishment of new set of interrogative rules for Malay corpus. **Conclusions/Recommendations:** 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.

Key words: Interrogative theory, knowledge identification, knowledge corpus, Data Manipulation Language (DML), Malay documents, interrogative knowledge, knowledge-base system

INTRODUCTION

The development of the Malay Interrogative Knowledge Corpus (MalayIK-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 MalayIK-Corpus has to manually modify the dictionaries into a MalayIK-Corpus. This has in turn motivated by studies done on corpus-based analysis (Mustapha *et al.*, 2010; Saif and Aziz, 2011); semantic relations (Thangamani and Thangaraj, 2010); term extraction (Syafrullah and Salim, 2010); fuzzy-based Decision Support System (Hartati and Sitanggang, 2010) to develop and experiment in Malay Corpus (Tan and Sh-Hussain, 2009).

Firstly this study 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.

Development of the corpus: The MalayIK-Corpus is a Malay language corpus where the Malay dictionary of Dewan (1996; 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. Besides, books on Malay language are also used in preparing the grammatical information entries (Masri, 1997).

It looks upon the interrogative theory of knowledge identification and representation as the background theory for the foundation of the MalayIK-Corpus development. The interrogative-based approach is described as the "who, when, what, where, how and why" analysis (Quigley and Debons, 1999). It makes

Corresponding Author: Fatimah Sidi, Department of Information System, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

distinctions between data, information and knowledge. Knowledge is easy to codify and document which is known as explicit knowledge. However tacit knowledge is difficult to capture and store (Jabar *et al.*, 2010).

Hence the MalayIK-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.

Attributes of the interrogative knowledge corpus: For the purpose of this development, Microsoft Access is used as a database for the MalayIK-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 MalayIK-Corpus contains attributes of:

- Root word (kata dasar))
- Lexicon (perkataan)
- Grammatical information of lexicon entry (kata masuk)
- 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
- 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 (1995) and Abdullah (2006) stop words are included which indicate pronoun, auxiliary verb, adverb, predicate, preposition, negative, conjunction, relative and determinant.

Classification of the word entry: Table 1 presents examples of words entry extracted from MalayIK-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.

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 MalayIK-Corpus has to define all of them primarily in the corpus. The following are steps taken in building up the MalayIK-Corpus:

- create attributes for corpus
- extract lexicons from the document collection
- verify the lexicons entries with Malay language expert
- insert lexicons entries in the database and
- 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

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.

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

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

Rank	Lexicon	Frequency	(%)	MalayIK-Corpus status
1	Dan	190.000	2.9	Stop word
2	Yang	170.000	2.6	Stop word
3	Di	131.000	2.0	Stop word
4	Ini	76.000	1.2	Stop word
5	Dengan	70.000	1.1	Stop word
6	Itu	58.000	0.9	Stop word
7	Tidak	53.000	0.8	Stop word
8	Kita	51.000	0.8	Stop word
9	Dalam	50.000	0.8	Stop word
10	Dari	43.000	0.7	Stop word
11	Untuk	43.000	0.7	Stop word
12	Halal	41.000	0.6	Adjective
13	Kepada	38.000	0.6	Stop word
14	Mereka	38.000	0.6	Stop word
15	Juga	37.000	0.6	Stop word
16	Pada	37.000	0.6	Stop word
17	Bagi	34.000	0.5	Stop word
18	Pertanian	33.000	0.5	Noun
19	Akan	31.000	0.5	Stop word
20	Umat	29.000	0.4	Noun
21	Telah	28.000	0.4	Stop word
22	Tetapi	28.000	0.4	Stop word
23	Seperti	27.000	0.4	Stop word
24	Makanan	26.000	0.4	Noun
25	Negara	26.000	0.4	Noun
26	Oleh	26.000	0.4	Stop word
27	Rakyat	26.000	0.4	Noun
28	Ada	25.000	0.4	Stop word
29	Dunia	24.000	0.4	Noun
30	Berkata	23.000	0.4	Verb
31	Ke	23.000	0.4	Stop word
32	Daripada	22.000	0.3	Stop word
33	Beliau	21.000	0.3	Stop word
34	Bukan	21.000	0.3	Stop word
35	Boleh	20.000	0.3	Stop word
Total number of words		6.479		

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.

Development of a stop word list: A list of stop words is included in the development of the MalayIK-Corpus, in order to eliminate words which have no values. The development of a stop words list in MalayIK-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.

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 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.

MATERIALS AND METHODS

The research setting is an experimental approach using 15% of 42,733 words from MalayIK-Corpus are sufficient and justified to produce better results in extracting identified knowledge. It is more than the suggested by Gay *et al*. (2008) for sample of more than 5,000 units, a sample size of 400 (8%) should be adequate.

The sample is drawn from different topics such as main news, technology, editorial columns, sports, letters and e-mails, while texts from children story books, articles and magazines are drawn from Internet or retyped from the printed materials.

Each document drawn is assigned with a serial number and number of words. The documents drawn are grouped according to the source of documents and range of number of words. The points of the range are defined at positions of 50-150, 151-300, 301-500 and the final range is more than 500.

For each range, five unstructured documents are selected and sorted in ascending order by total number of words. This makes the Malay unstructured documents collection consists of 23 unstructured documents which comprises of 6,479 words.

Interrogative Knowledge Identification Framework is used to address the need for the mechanism to identify knowledge from unstructured document (Sidi *et al*., 2009; Sidi, 2007). The development of the system based on architecture of framework. The system consists of four major processes:

i. Prepare the unstructured documents to be processed and converted it into extension of plain text file
ii. Invoke lexicon identifier that uses lexicon interrogative analysis matching rules. It is used to identify and extract knowledge in each of the complete sentences written in the unstructured document. It is also used to extract interrogative lexical constructs from the individual unstructured document
iii. Invoke object recognizer that uses matching rules of object interrogative analysis to extract ontological constructs from the interrogative lexical constructs
iv. Transform ontological constructs to populate database scheme by connecting ontology model with conceptual modeling of object-relationship model. This is used to structure the extracted knowledge into interrogative structured form.

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 MalayIK-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 MalayIK-Corpus.

RESULTS AND DISSCUSION

The results obtained are measured in terms of percentage of quantitative retrieval performance recall and precision metrics (Baeza-Yates and Ribeiro-Neto, 2000) coupling with research methods and concept in information system research (Jabar *et al*., 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. Table 3 shows the results of the experiments in the form of precision and recall.

The results show a decrease in the precision of 98% for the interrogative element of what. This is due to the reason of the inability of the system to identify and extract the special characters of proper nouns in the document. This indicates limitation of the lexicon identifier in the system process. The 100% recall and precision determine that the identification and extraction of the interrogative lexical constructs from the corpus can achieve a great fidelity and certainty. Should the lexicon interrogative analysis matching rules of the Malay*IK*-Corpus are incorrectly defined and lexicons do not exist in the corpus, fidelity and certainty cannot be achieved which are shown from the results.

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

Table 3: Results of Interrogative Elements Recall and Precision

Interrogative Elements	Recall	Precision
where	87%	97%
when	71%	82%
how	83%	87%
why	96%	91%
who	95%	94%
what	95%	89%

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.

CONCLUSION

The study presents a development of MalayIK-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 MalayIK-Corpus is lacking of ease for navigation in its system interface. It is not fully automated on the creation of the Malay corpus. In the future, the system development will be based on the concept of rapid application, information retrieval and neuro-identifier (Choo and Lee, 2008; Alallayah *et al.*, 2010; Bouramoul *et al.*, 2010).

REFERENCES

Abdullah, M.T., 2006. Monolingual and Cross-Language Information Retrieval Approaches for Malay and English Language Document. PhD thesis, University Putra Malaysia. http://psasir.upm.edu.my/5869/

Ahmad, F., 1995. A Malay Language Document Retrieval System: An Experimental Approach and Analysis. PhD thesis, Universiti Kebangsaan Malaysia. http://www.waset.org/journals/waset/v10/v10-18.pdf

Alallayah, K.M., W.F.A. El-Wahed, M. Amin and A.H. Alhamami, 2010. Attack of against simplified data encryption standard cipher system using neural networks. J. Comput. Sci., 6: 29-35. DOI: 10.3844/jcssp.2010.29.35

Ali, H.M., M.N.M. Shariff and W.M.W. Dewa, 1993. Kamus Imbuhan Bahasa Melayu Edisi Kedua. 1st Edn., Fajar Bakti, Kuala Lumpur, Malaysia, ISBN: 9676562505, pp: 394.

Baeza-Yates, R. and B. Ribeiro-Neto, 2000. Modern Information Retrieval. 3rd Edn., Addison Wesley, New York, ISBN: 0-201-39829-X, pp: 513.

Bouramoul, A., M.K. Kholladi and B.L. Doan, 2010. PRESY: A context based query reformulation tool for information retrieval on the web. J. Comput. Sci., 6: 470-477. DOI: 10.3844/jcssp.2010.470.477

Choo, C.H. and S.P. Lee, 2008. Towards persistence framework-based rapid application development toolkit for web application development. J. Comput. Sci., 4: 290-297. DOI: 10.3844/jcssp.2008.290.297

Dewan, B.P., 1996. Kamus Dewan. 1st Edn., Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia, ISBN: 9836244565, pp: 1566.

Dewan, B.P., 2005. Kamus dewan. 1st Edn., Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia, ISBN: 9836283390, pp: 1817.

Gay, L.R., P. Airasian, G.E. Mills and P.W. Airasian 2008. Educational Research: Competencies for Analysis and Application. 9th Edn., Prentice Hall, USA., ISBN: 0135035015, pp: 648.

Hartati, S. and I.S.M. Sitanggang, 2010. A fuzzy based decision support system for evaluating land suitability and selecting crops. J. Comput. Sci., 6: 417-424. DOI: 10.3844/jcssp.2010.417.424

Hawkins, J.M., 2001. Kamus Dwibahasa Oxford Fajar Edisi Ketiga. Kuala Lumpur. Penerbit Fajar Bakti Sdn. Bhd, Malaysia, ISBN: 9676562548.

Jabar, M.A., F. Sidi and M.H. Selamat, 2010. Tacit knowledge codification. J. Comput. Sci., 6: 1170-1176. DOI: 10.3844/jcssp.2010.1170.1176

Jabar, M.A., F. Sidi, M.H. Selamat, A.A.A. Ghani and H. Ibrahim, 2009. An investigation into methods and concepts of qualitative research in information system research. Comput. Inform. Sci., 2: 47-54;

Masri, S., 1997. Tatabahasa Melayu. 1st Edn., the University of Michigan, USA., ISBN: 9676541125, pp: 197.

Mustapha, A., M.N. Sulaiman, R. Mahmod and M.H. Selamat, 2010. Corpus-based analysis on cross-domain experiments in classification-and-ranking generation. J. Comput. Sci., 6: 1326-1333. DOI: 10.3844/jcssp.2010.1326.1333

Quigley, E.J. and A. Debons, 1999. Interrogative theory of information and knowledge. Proceeding of the 1999 ACM SIGCPR Conference on Computer Personnel Research. New Orleans, Louisiana, United States, pp: 4-10; DOI: 10.1145/299513.299602

Saif, A.M. and M.J.A. Aziz, 2011. An automatic collocation extraction from Arabic corpus. J. Comput. Sci., 7: 6-11. DOI: 10.3844/jcssp.2011.6.11

Salton, G. and M.J. McGill, 1983. Introduction to Modern Information Retrieval. 2nd Edn., McGraw-Hill, New York, ISBN: 0070544840, 448.

Sidi, F., 2007. Transformation of Extracted Knowledge in Malay Unstructured Documents into an Interrogative Structured Form. PhD thesis, University Putra Malaysia. http://psasir.upm.edu.my/5887/

Sidi, F., M.A. Jabar, M.H. Selamat, A.A.A. Ghani and M.N. Sulaiman, 2009. Framework for interrogative knowledge identification. Comput. Inform. Sci., 2: 109-115.

Syafrullah, M. and N. Salim, 2010. Improving term extraction using particle swarm optimization techniques. J. Comput. Sci., 6: 323-329. DOI: 10.3844/jcssp.2010.323.329

Tan, T. and Sh-Hussain, 2009. Corpus design for malay corpus-based speech synthesis system. Am. J. Applied Sci., 6: 696-702. DOI: 10.3844/ajassp.2009.696.702

Thangamani, M. and P. Thangaraj, 2010. Integrated clustering and feature selection scheme for text documents. J. Comput. Sci., 6: 536-541. DOI: 10.3844/jcssp.2010.536.541

Van Rijsbergen, C.J., 1979. Information Retrieval. 2nd Edn., Butterworths, London, ISBN: 0408709294, pp: 208.