

# Analysis of translated query in Quranic Malay and English translation documents with stemmer

*Mohd Amin Mohd Yunus*<sup>1</sup>, *Aida Mustapha*, and *Noor Azah Samsudin*

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400 Batu Pahat, Johor, Malaysia

**Abstract.** Quranic documents result has a limited query due to focusing on exact words to retrieve those relevant documents. Therefore, there is variety of results to be useful for the target users to explore Quran documents in proper manner. Thus, this paper presents analysis according to conducted empirical experiments in 12 retrieval processes. Thus a system is needed to retrieve relevant documents across language boundaries as well as monolingual. Therefore, empirical experiments are conducted with the purposes to investigate English-Malay translation approach and vice versa against monolingual searching process. Furthermore, it is also conducted to investigate the performance between keywords and querywords based on total retrieve and relevant for each retrieval process. The retrieval however, included the unnecessary documents because of the translation polysemy. This research also is being applied in retrieving Quran English and Malay translated documents with queries compared to monolingual query searching retrieval. Furthermore, in order to produce more significant result, the comparison between stemmer and monolingual results are successfully analysed to evaluate precision and recall percentages. The most important findings are the use of stemmer more beneficial to the query and documents simultaneously regardless the experiments applied translation or not. It leads more and more relevant results displayed.

## 1 Introduction

It is observed that the dictionary-based method is a suitable method to disambiguate every translation of query rather than machine translation and corpus parallel document collection. Therefore, the goal of this research is to enable users to query in the Malay language against an English collection. Cross Language Information Retrieval (CLIR) systems can be evaluated automatically in much the same fashion as monolingual IR systems, using queries with known relevance judgments. However, in the multilingual context, there is a strong desire to compare the results against the optimal performance of

---

<sup>1</sup> Corresponding author: aminy@uthm.edu.my

the system if the query is perfectly translated. This gives rise to the following approach to evaluation in CLIR. Start with queries, documents, and relevance judgments in a single language. The queries have been translated into another language by human translators. These translated queries are helpful for the performance of the CLIR system. In order to conduct the experiments, data regarding the Malay and English Quran translation are gathered accordingly as two sets of collection respectively. The Malay and English dictionary files are useful for translating the query. A Malay and English stopword files are used for removing the unnecessary words in the query.

## 2 Related Work

Monolingual and multilingual are two different to be applied on query. Monolingual is easier to be understood as there is only single language performed in searching process. While multilingual can be applied to the query is used in translation for searching other languages in getting more multilingual documents with single query. This concept is called multilingual query expansion for searching more documents in various languages. There are several studies conducted to investigate the query to be useful for retrieval. Therefore Nie [9] stated the query expansion that needs to be improved in monolingual while Dang and Croft [10] present query formulation improving the search result. Another perspective, Strohmaier and partners [11] describe the suggested words for query so that users may choose the right query for expected results. On top of that, stemmer [2,6] is very important to apply to that translated query [12, 13,14,15,16].

## 3 Experimental Approach

Two al-Qur'an documents are used in these experiments, Malay Quranic documents collection [3] and English Quranic documents collection [4]. Each collection has 114 surahs and 6236 documents. Every document has numbers beginning with q which means query and followed by numbers with first three numbers denote chapter and last three number denote verse for every filename such as q034006 but in the filename the number of surah begins with "." or dot, followed by the number of surah and the number of ayat begin with , or comma, followed by the number of ayat as for example ".34,6". All documents are as flat files in ASCII or EBCDIC text and searching process is through pattern matching [1]. Short query is to type some queries for retrieving relevant documents. In this study, the words as input are called query for any matched words in the documents. In this research, the Malay query words are taken from Fatimah's collection [2] and the English query words are translated from the Malay query words. Fatimah has obtained them by considering several guidelines put forward by Popovic [5] and Salton [8]. Each query would be separated and broken into keywords and replaced by target language. For example, if query is Malay, so it is called as source language and the target language is English. Thus, English represents the translated word to retrieve English documents and if the query is English, the target language is Malay. The dictionary lists 1500 Malay words and 1325 English words including 36 Malay query words selected. The translation refers to the same index between Malay natural query languages [2] and the translation of the natural query language in English. When the keyword is Malay, then reference is to the English word at the same index or when the keyword is English, then reference is to the Malay word at the same index. It is considering word by word in the text files.

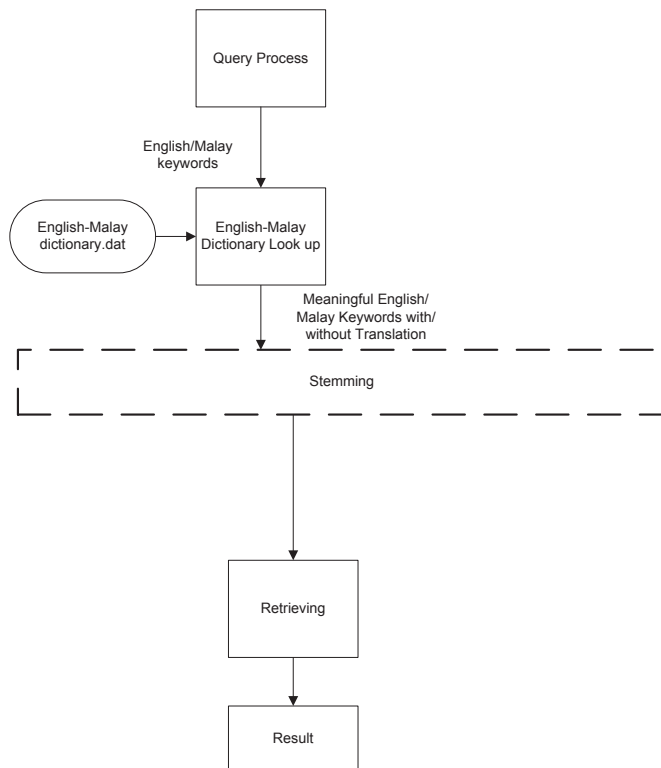
Stopwords are used for removing the meaningless words in the query during retrieval process. So, it is easier for the system for not processing meaningless words. By having Malay and English stopwords, it is easier to remove the common meaningless words and

English does the same. The Malay root words are used to make comparable to the query after stemming. Then, if stemmed words in the query are similar to root words, then retrieving process will be next to produce relevant documents and stemmed query [2]. Fatimah [2] calculated that the relevant document should be resulted from each query list with the Arab student assistance from the Universiti Kebangsaan Malaysia. The student searched the al-Quran, referring to well-known Islamic books regarding the same subject matter based on the natural language queries [2].

The flow of process is started by pre-processing, storing, retrieving, evaluation and presentation of the result. The query term can be Malay or English and also keywords or querywords according to words in the query. The query can be categorized into two which are keyword and queryword. If the query is keywords, the results retrieved according to word by word results and redundant document names existed if merged. But querywords, retrieved according to the whole words as one at all and only when no redundant or unique document names retrieved rankly. Query translation can replace the origin query in to another language of the query. This translation can be either of both languages.

When the query is conflated with stemming algorithm, this query can be translated in either Malay stemming Malay [2] or English [6] stemming algorithm according to the query keyed-in. Stemming algorithm removes the suffix, infix, and prefix of each word in the query to be root word in order to have more relevant documents in the results. Then, searching process is done according to the type of process of results required. All documents are in flat files and ASCII or EBCDIC text. For searching process, pattern matching is used in the process. Pattern matching refers to the words similarity between query and documents in retrieving process. A large amount of manual effort is needed to represent large document collection in the form of semantic relationship. File structures are stored in memory while collections or databases are usually stored on disk due to their large volumes. The query submitted to the system is also represented by a query signature that is used to search the signature file and unwanted documents are removed. Figure 1 explains the workflow of cross language information retrieval based on query translation in the retrieval process of Malay and English Quranic documents collection. Query is processed by removing the meaningless words or stopwords. Then, query can be translated into another language if needed by the dictionary. Next step is to stem the query by a stemmer if needed according to accurate language.

The query results consist of word by word result or keyword or the whole words or full phrase. This documents retrieval comes from al-Quran documents translation collection which are from Malay and English collection. For this research, when the query is Malay words, the words are translated into English, and vice versa. Then, there are two choices whether stemming words or not. For stemming process, it is going to remove suffix, prefix and infix of the query that contains word by word by using either Porter [6] for English and Fatimah [2] for Malay. Then, the stemming process is necessary for query and document in order to retrieve more relevant documents.



**Fig. 1.** Workflow of query translation information retrieval with stemmer.

## 4 Results and Discussion

The evaluation technique is used for precision and recall results [7]. Table 1 shows the formula to calculate the percentage of precision and recall. There are 12 processes of retrieving the results as shown in Figure 2. A few of processes consist of stemming approach and dictionary translation.

**Table 1.** Recall and precision formula.

$Recall (\%) = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}} \times 100$
$Precision (\%) = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}} \times 100$

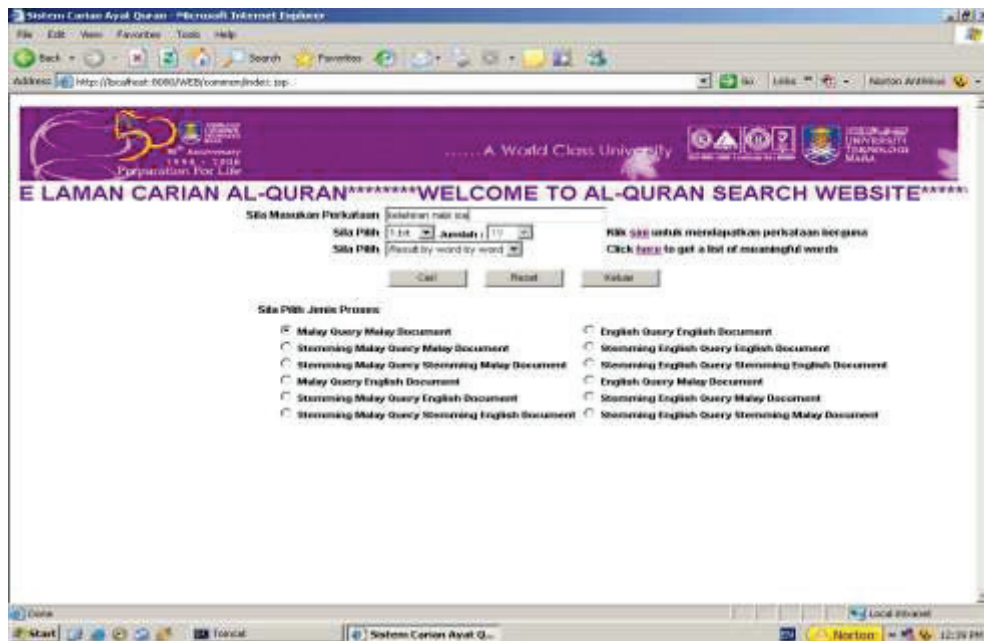


Fig. 2. Main menu of Al-Quran search website.

The results are shown for each of them according to non-stem both, stem query, stem both query and document. Their effectiveness and retrieval and relevant percentage are included as part of evaluation and analysis. Keywords (K) and querywords (Q) are different processes of retrieving results. K usually has redundant information and retrieving results word by word but not for Q. Total retrieve (TRT) is from the result and total relevant (TRE) is used from Fatimah's thesis [2]. Total retrieve and relevant is taken from comparison between TRT and TRE whereby totally matched documents. Table 2 shows that the experiments results in terms of average of recall and precision percentages. Translation of query is not affecting on recall and precision. Experiments on Malay documents show the better average Recall and Precision percentages in SMQSD and SEQSD. It means that the Fatimah [2] stemmer has more relevant documents at 44.62 for Queryword (Q). Its precision is 21.47 for and Q. Experiments on English documents show the lower performance than Malay documents because of its stemmer, Porter [6] has lower recall and precision, 16.60 and 14.69. Its performance with the stemmer shows the average recall and precision at 0.84 and 11.38 respectively. These values are lowest in the experiments. The high of recall and precision percentages depend on the efficient stemmer in order to help the queries to retrieve more relevant documents. Thus the queries can retrieve more relevant documents as the efficient stemmer taking part in the process in information retrieval. The queries explain keywords that repeat the same documents and lead to low effectiveness rather than querywords.

**Table 2.** Average percentages of recall and precision of experiments.

No	Experiments	Average of Recall		Average of Precision	
		K	Q	K	Q
1.	NSMQNSMD	34.30	30.12	23.47	23.90
2.	SMQNSMD	33.03	29.62	22.94	22.81
3.	SMQNSMD	47.34	44.62	20.56	21.47
4.	NSEQNSMD	34.30	30.12	23.47	23.90
5.	SEQNSMD	33.03	29.62	22.94	22.81
6.	SEQNSMD	47.34	44.62	20.56	21.47
7.	NSEQNSMD	20.40	16.60	15.50	14.69
8.	SEQNSMD	11.28	10.74	12.56	12.42
9.	SEQNSMD	0.84	0.84	11.38	11.38
10.	NSMQNSMD	20.40	16.60	15.50	14.69
11.	SMQNSMD	11.28	10.74	12.56	12.42
12.	SMQNSMD	0.84	0.84	11.38	11.38

## 5 Conclusions

The empirical result is presented to show the difference performance between monolingual and cross lingual. The results also show the benefits of applying keywords (K) and querywords (Q) in retrieval and relevant retrieval according to monolingual and cross lingua query. The results are concluded according to performance of query with or without translation, the benefits of K and Q, the stemming query itself or both.

The short query is quite suitable with the dictionary and retrieves better results and it helps to search more documents in other languages. These examples also include stemming words after translating the word and matching the words in every document in collection in order to retrieve relevant document required from the query given. It is called as CLIR compared to MLIR, only search specific language if given query with the same language. The relevant documents retrieval depends on combination of the efficient stemmers, dictionaries and stopwords to help the retrieval process. it has more documents retrieval when stemming both translated query and documents. While monolingual takes part of having high percentage of retrieving relevant and related documents. Then, querywords (Q) show unique documents rather instead of keywords (K) which show the redundant documents. Recall and precision percentages however, show better performance in K than Q. The queries need dictionaries to help and translate words in different languages for queries or documents. The difference languages but same meaning in queries can retrieve quite same documents as if it is monolingual search. The future study will be discussed on semantic approach related to performance of query translation. The study will reveal the effective way to use the approach to be applied on information retrieval application according to the empirical experiments.

## Acknowledgement

This work was supported in part by a grant from the Ministry of Education under the Research Acculturation Grant Scheme (R045) and in part by a grant from Research Gates IT Solution Sdn. Bhd.

## References

1. J. Elly, The Study Of Existing Malay Algorithm Performed On Words Beginning With 'D', B.Sc. Thesis, Universiti Teknologi MARA (2000)
2. A. Fatimah, *A Malay Language Document Retrieval System: An Experiment Approach And Analysis*. Tesis Ijazah Doktor Falsafah Universiti Kebangsaan Malaysia (1995)
3. H. Z. Hamidy, H. S. Fachruddin, Tafsir Quran. Translation. Klang: Klang Book Centre (1987)
4. T. A. Muhammad, M. K. Muhammad, Interpretation of the Meaning of the Noble Quran. Dar-us-Salam Publications. <http://www.amazon.com/Noble-Quran-Interpretation-Meanings-Language/dp/996074079X> (1999)
5. M. Popovic, P. Willett, The Effectiveness Of Stemming For Natural-Language Access To Slovene Textual Data. *Journal Of The American Society For Information Science*, **43**(5), 384-390 (1992)
6. M. F. Porter, An Algorithm For Suffix Stripping, *Program*, **14**(3), 130-137 (1980)
7. G. Salton, M. J. McGill, Introduction To Modern Information Retrieval. New York: Mcgraw-Hill (1983)
8. G. Salton, Experiments In Automatic Thesaurus Construction For Information Retrieval, *Proceedings Ifip Congress 1971*, Ta-2, 43-49 (1971)
9. J. Y. Nie, Query expansion and query translation as logical inference. *Journal of the American Society for Information Science*, **54**(4), pp. 335-346 (2003)
10. V. Dang, W. B. Croft, Query Reformulation Using Anchor Text. *WSDM'10*, February 4-6, 2010, New York City, New York, USA (2010)
11. M. Strohmaier, M. Kröll, Christian, Intentional Query Suggestion: Making User Goals More Explicit During Search. *WSCD'09*, Feb 9, 2009 Barcelona, Spain (2009)
12. M. A. Yunus, R. Zainuddin, N. Abdullah, Semantic Method for Query Translation. *The International Arab Journal of Information Technology (IAJIT)*, vol. **10**, no.3, pp. 253-259 (2013)
13. M. A. Yunus, R. Zainuddin, N. Abdullah, Semantic Query for Quran Documents Results. *Proceeding of IEEE Conference On Open Systems (ICOS)*, pp. 1-5, IEEE (2010)
14. M. A. Yunus, R. Zainuddin, N. Abdullah, Semantic Speech Query via Stemmer for Quran Documents Results. *Proceeding of International Conference On Electronic Devices, Systems and Applications (ICEDSA)*, pp. 17-21, IEEE (2011)
15. M. A. Yunus, R. Zainuddin, N. Abdullah, Visualizing Quran Documents Results by Stemming Semantic Speech Query. *Proceeding of International Conference On User Science and Engineering (i-USER)*, pp. 209-213, IEEE (2010)
16. M. A. Yunus, R. Zainuddin, N. Abdullah, Semantic Query with Stemmer for Quran Documents Results. *Proceeding of IEEE Conference On Open Systems (ICOS)*, pp. 40-44, IEEE (2010)

## Appendix: List of Abbreviations

<b>NSMQNSMD</b>	Non-Stem Malay Query Non-Stem Malay Document
<b>SMQNSMD</b>	Stem Malay Query Non-Stem Malay Document
<b>SMQSMD</b>	Stem Malay Query Stem Malay Document
<b>NSEQNSMD</b>	Non-Stem English Query Non-Stem Malay Document
<b>SEQNSMD</b>	Stem English Query Non-Stem Malay Document
<b>SEQSMD</b>	Stem English Query Stem Malay Document
<b>NSMQNSD</b>	Non-Stem Malay Query Non-Stem English Document
<b>SMQNSD</b>	Stem Malay Query Non-Stem English Document
<b>SMQSD</b>	Stem Malay Query Stem English Document
<b>NSEQNSD</b>	Non-Stem English Query Non-Stem English Document
<b>SEQNSD</b>	Stem English Query Non-Stem English Document
<b>SEQSD</b>	Stem English Query Stem English Document