The effect of using a thesaurus in Arabic information retrieval system

Mohammad Wedyan , Basim Alhadidi and Adnan Alrabea

Computer Science Department, Al-Balqa Applied University, Al-Salt, Jordan

Abstract

Automatic query expansion methods for English and other languages text retrieval have been studied for a long time. In this research we study the retrieval effectiveness, achieved when we apply a successful automatic query expansion method in Arabic text retrieval based on an automatic thesaurus. Our experiments show that the automatic query expansion method resulted in a notable improvement in Arabic text retrieval using a sample of abstracts of Arabic documents. The study showed that the use of a thesaurus has improved information retrieval system by 10% -20%. The study also shows that the greater the number of documents in the building thesaurus, Thesaurus was more accurate.

Keywords: *Arabic retrieval, thesaurus, stop words, indexing, information retrieval system.*

1. Introduction

Arabic is a language that holds the miracle of holy Quran, and that accomplished all the requirements of Arabic and Islamic civilization in its peak flourishing. Arab books in Medicine and Science had been the main reference books for the west and in most of its important universities. [1]

Internationally, it gained full acceptance and recognition and become a credited language in UN institutions along side with the other five languages previously used. [1]

Arabic has many Properties, first, Arabic language consists of 28 letters, 16 of them have one dot, two or three dots. Second, Writing is from right to left. Third, varying ways of writing. For example completely mashkool (all signs of tashkeel are used) or partially mashkool or Not mashkool Fourth, Letters change their shape according to the place of occur in. fifth, Dual language formal and informal.sixth, Grammatical flexibility, words may be arranged in many different ways. [2]

Experimental results show that spelling normalization and stemming can significantly improve Arabic monolingual retrieval. Character tri-grams from stems improved retrieval modestly on the test corpus, but the improvement is not statistically significant. [3]

Therefore this study will statement effect of using a thesaurus on the information retrieval system (IRS), and compared the improvement after using automatic thesaurus from the traditional system.

2. Evaluating information retrieval systems.

Any retrieval system is usually evaluated according to its efficiency and effectiveness. There are two aspects of efficiency, they are time and space. Time is the speed of matching the in-use queries with the document descriptions. Space is the space needed in a disk that the system needs. Efficiency is determined according to the ability of the system to return documents relevant to the user query. The perfect status of the system is referring all the files that are relevant to the process of query and never referring any irrelevant files. The difficulty lies in the determination of relevance because the process of determining relevance of documents is a subjective one. [4]

The decision of the person depends much on many factors; experience, for example. Any professional in a certain field may see the general information retrieved from a system as irrelevant while any amateur (beginner) sees it as fully relevant. This may lead to increasing in the determination of relevance. In research, researchers usually consider the process of determination of relevance as an objective process. [4] We suggest here that evaluation process is objective and previously agreed on.

Criteria used in the process of evaluating the performance of a system are precision and recall. Precision means the ability of the system to return documents that have relevance to the query. [4]

The most commonly used measurements of retrieval performance are precision and recall. Precision measures the ability of the system to retrieve only the documents that are relevant to a query [4]

$$\text{Precision} = \frac{\text{A mount of relevant documents retrieved}}{\text{A mount of documents retrieved}}$$

Recall measures the ability of the system to retrieve all documents that are relevant to a query [4]:

$$\text{Recall} = \frac{\text{A mount of relevant documents retrieved}}{\text{A mount of relevant documents in the collection}}$$

3-Idexding

Indexing is defined as the process of choosing a term or a number of terms that can represent what the document contains. These terms are called (Index terms). [3]

Indexing can be performed either manually (Manual Indexing) or through using computers software and programs (automatic Indexing) [4].

Manual indexing has some weaknesses that mentioned. The person who performs indexing must have the complete knowledge of what the document contains, and what the document talks about. The result may vary due to different experiences of indexers. This leads to increasing cost.[5]

This research uses automatic indexing, so it will be our focus.

3-2 Automatic Indexing

The first step in indexing is the Lexical Analysis. The process of changing the text into a group of separate words, each word is called (token), a token is a group of letters. Lexical analysis is also the first step in queries analysis [6].
The process of lexical analysis may present idioms that can be used as (Index Terms), in order to assign the suitable index term to reach the suitable document.[6]

Then comes the process of separating unnecessary words, they are called 'Stop Words" as (قد) and (هذا), they are repeated in all documents and texts. The importance of this step is discussed later in this study.

3-3 Eliminating Stop words:-

Stop words are those words that are repeated in every document, so they are considered as weak to be distinguished, we cannot distinguish the content of a text depending on them.[5]
There are other benefits from eliminating them as "shortening indexing structure"[7]and are useful in making the process faster and doesn't have information Retrieval and the degree of the efficiency of recalling system. [6]

It doesn't also burden the system with unnecessary information [8]

It is not clear which words can be considered on stop words and which cannot. Traditional methods consider that words that are repeated many times are stop words, but there are some words that are repeated in a certain document and considered as important words "indexing terms". But when the subjects are more specialized, as to say a subject specialized in data base. Then the use of repeated words, even if simply, as "index terms" as computer language engines" are useless to be "index terms". [6]

The other way is to save stop words in a list, then we search for each token separately. That result from lexical analysis and comparing it with the list, if it is in the list, it will be ignored and not processed later. [6]

Arabic is very rich in lexical tokens, that means stop words are available in big quantities. [8]

Swaine said several characteristics of stop words in his book. First, they have no meaning if they are used separately. Second, appear many times in a text. Third, necessary for the construction of the language. Fourth, mostly adjectives. Fifth, general words and not particularly used in a certain field. Sixth, any researcher doesn't ask about such words. Seventh, never form a full sentence when used alone.

4. Thesaurus

Thesaurus is an efficient tool in IRS specially in the modern systems, in indexing or in searching which helps in extending queries through using more suitable tokens. [4]

Constructing thesauruses has a great benefit in IRS, it strengthens precision and control of idioms in order to serve and increasing format in the process of documents. Indexing and retrieval and in using the best idioms and helps the user to reformulate his queries if necessary [6].

Simply the thesaurus consists of a list of the important words, a certain subject, each word is connected with other words in the list. [7]. Most thesauruses we use have been built manually depending on experts in certain fields or on the experts in the field of document description. Building thesauruses manually is a waste of time and money, the result may also be subjective, because the person who builds it may use his own choices which may affect the construction of the thesauruses, so we are in need of an automatic construction of thesauruses which will save time, effort and cost and make the results more objective easy to be modified in the future [4]
Taking into consideration what is mentioned previously, we will build Automatic Thesauruses which have many benefits over the manual one [7].
It supports standard vocabulary in indexing or in searching it helps the user in putting down the suitable expressions in queries. It supports different hierarchies as it allows broadening or narrowing the query according to the user needs.

4.1 Automatic Thesaurus Construction.

in vector space models documents are represented by vectors as bellow:

$$\vec{D_j}=(W_{1,j}, W_{2,j}, W_{3,j}, \ldots\ldots\ldots, W_{t,j})$$

W: weight

$\vec{D_j}$: Vector for doc j

We can compute the weight by these equations:

$$W_{i,j} = f_{i,j} * \log N/n_i \text{ ----------[7]}$$

ni: the number of documents that term i appear in it.

Fi.j: Normalized Frequency and compute by

$f_{i,j} = freq_{i,j} / MAX_L\ freq_{L,j}$ -----[7]

Freqi,j: the nuber of times the term i appeared in the text of the document j.

$MAX_L\ freq_{L,j}$: the maximum is compute over all terms which are mentioned in the text of the documents dj.

These vectors of a group of documents van be represented as follows:

	T_1	T_2	T_3	…	T_n
D_1	W_{11}	W_{12}	W_{13}	…	W_{1n}
D_2	W_{21}	W_{22}	W_{23}	…	W_{2n}
D_3	W_{31}	W_{32}	W_{33}	…	W_{3n}
:	…	…	…	…	:
:	…	…	…	…	:
D_m	W_{m1}	W_{m2}	W_{m3}	…	W_{mn}

Figure (1): Documents Vectors

Then comes the step of calculating similarity between index terms using any of the equations of similarity calculations as in the following table:

	T_1	T_2	T_3	…	n
D_1	W_{11}	W_{12}	W_{13}	…	W_{1n}
D_2	W_{21}	W_{22}	W_{23}	…	W_{2n}
D_3	W_{31}	W_{32}	W_{33}	…	W_{3n}
:	…	…	…	…	:
:	…	…	…	…	:
D_m	W_{m1}	W_{m2}	W_{m3}	…	Wm_n

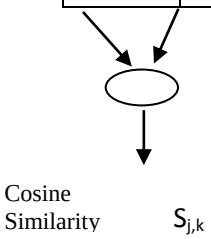

Figure (2): compute the term-term similarity

$$\text{Cosine similarity } S_{j,k} = \frac{\sum_{i=1}^{n}(W_{i,j} * W_{i,k})}{\sqrt{\sum_{i=1}^{n} W_{i,j}^2 * \sum_{i=1}^{n} W_{i,k}^2}}$$

These equations to calculate similarity between each index term brings out a matrix as the following:

	T_1	T_2	T_3	…	T_n
T_1	S_{11}	S_{12}	S_{13}	…	S_{1n}
T_2	S_{21}	S_{22}	S_{23}	…	S_{2n}
T_3	S_{31}	S_{32}	S_{33}	…	S_{3n}
:	…	…	…	…	:
:	…	…	…	…	:
T_m	S_{m1}	S_{m2}	S_{m3}	…	S_{mn}

Figure (3): The term-term similarity

Sm,n: resembles the similarity between the term (N) and the term (M).

We have now similarity matrix; because the similarity between (Tx) and (Ty) equals the similarity between (Tx) and (Ty).

5. RELATED WORK

Despite the very little Arabic efforts in developing thesauruses, the theoretical efforts supported and opened new paths for building Arabic thesaurus, even though very limited, the first trials in this field were translation of foreign thesauruses, example of this is the list of Arabic Idioms prepared by Industrial Development Center for the Arab World in 1970, and the Islamic thesaurus which was built manually[9].

Some studies in IRS and in building thesauruses. Abu salem (1992) for example, studies the IR in Arabic Language. His study was based on 120 documents he received from the Saudi Arabian National Computer Conference and on 32 queries. in his research, he studied indexing by using full words and by using the roots only. He found that using the roots is superior to other ways. He also built a manual thesaurus using the relation between expressions to test the possibility of supporting an IRS through this thesaurus. He found that the thesaurus makes IR much better.

The General Thesaurus presented by UN Aid Program.The Program of Authorization in the Arabic World (2003). This one uses initially synonyms that help the researcher to choose his expressions that he has to look for. This thesaurus includes also the relations of origin and branches and those of contextualization between expressions. This helps in boarding the search, if the search has no

matches when using a certain expression, the researcher can use either broad terms or narrower ones. Synonyms are the first step in this thesaurus

Kanaan and wedyan (2006). Their study was based on 242 documents they received from the Saudi Arabian National Computer Conference and on 24 queries. In their research, they studied indexing by using full words and by using the roots. They found that using the roots is superior to other ways. They also built a Automatic thesaurus using the relation between expressions to test the possibility of supporting an IRS through thesaurus. They found that the thesaurus makes IR much better between 1% and 10%.

6. Conclusions

This study aims at reinforcing IRS depending on Arabic. The results after applying 35 queries, this study was based on 500 documents those were given to a group of students who have certain links with those subjects to determine the relevant document to each query. According to the determination of those students, work on these results began and results were analyzed using the criteria of Precision and Recalling and by using smoothing Algorithm that was used by Abu Salem (1992) and by Kanaan (1997). Average Recall Precision was calculated.

Average Recall Precision			
Recall	without using thesaurus	with use thesaurus	**Improvement (%)**
0.1	0.706	0.884	17.8
0.2	0.7	0.872	17.2
0.3	0.63	0.81	18
0.4	0.45	0.607	15.7
0.5	0.352	0.498	14.6
0.6	0.25	0.38	13
0.7	0.18	0.305	12.5
0.8	0.09	0.198	10.8
0.9	0.05	0.151	10.1
1	0.021	0.121	10

Table (1): The above Table Showing how better were the results when using with the thesaurus.

Figure A comparison between the values of average Recall Precision when full words were used with and without the thesaurus.

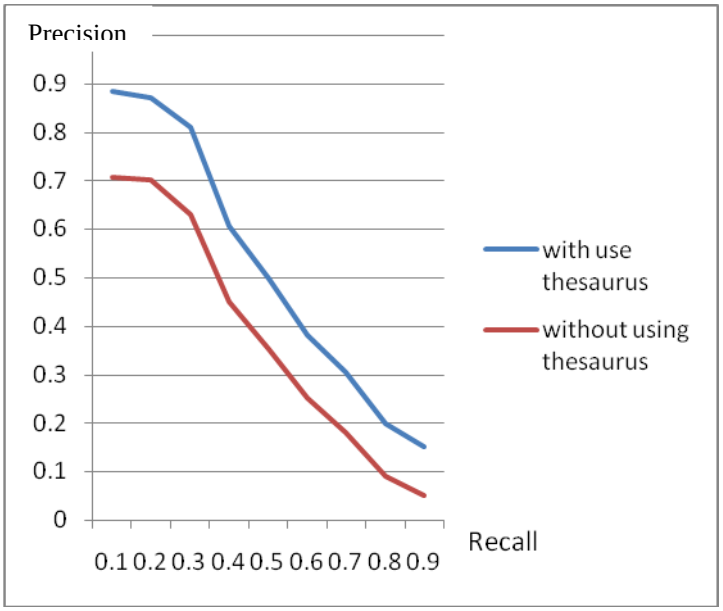

Figure (4): Showing how better were the results when using full words with the thesaurus.

The previous chart shows the effect of using the thesaurus on making the system efficiency that depends on whole words better by applying the criterion of average recall precision. When the thesaurus was used, the results were better. This goes well with what Hani Abu Salem(1992) and Kannan(2006) calculated when he aid that the use of thesaurus in Arabic will make the efficiency of the Arabic IRS better when full words were used. And when we increase number of documents that used to build thesaurus the result will be better. Kannan and wedyan (2006) used 242 documents to build their thesaurus and in this study we use same equations to build our thesaurus but we used 500 documents

This study may be applied on other equations as Jaccard and Dice or be applied on huge number of documents. The user can be utilized in feeding the system in order to have a high precision thesaurus.

References:

[1] Khatib, Ahmed Shafiq,1997," terminological specifications and applications in the Arabic language", cultural fifteenth season of the Arabic Language Academy of Jordan, Amman, Jordan, pp. 177-213.(Arabic)
[2]Ali, Nabil, 1988,"Arabic and computer, localization", Cairo. (Arabic)
[3] J. Xu, A. Fraser, and R. Weischedel, 2002, "Empirical studies in strategies for Arabic retrieval," Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland: ACM, pp. 269-274.
[4] Lassi, M., 2002," Automatic Thesaurus Construction", university collage of boras,
[5] Salton, G., and McGill, M., 1983," Introduction to Modern Information Retrieval", McGraw-Hill, New-York.
[6] Frakes, W., and Baeza-yates, R.,1992," Information Retrieval Data Structures & Algorithms", P T R Prentice Hall, New Jersey.

[7] Baeza-yates, R.,and Rierio-neto, B.,1999," Modern Information Retrieval" , Addison-Wesley,New-York.
[8] Soana, Ali Suleiman,1994," information retrieval in the Arabic language", King Fahd National Library.(Arabic).
[9] Abdul-Jabbar,Abdul Rahman,1993," The use of a system consultant in building thesauruses", scientific record of the Symposium on the use of Arabic in Information Technology organized by the King Abdul Aziz Library public, Riyadh, Saudi Arabia.(Arabic).
[10] Abu Salem, H.,1992," A Microcomputer BasedArabic Bibliographic Information Retrieval system With Relational Thesau"ri, Ph.D. Thesis, University of Illinois,Chicago,USA.
[11] Kanaan, G.,1997, Comparing Automatic Statistical and Syntactic Phrase Indexing for Arabic Information Retrieval,1997, Ph.D. Thesis, University of Illinois, Chicago, USA.
[12] Kanaan, G., M, Wedyan.,2006," Constructing an Automatic Thesaurus to Enhance Arabic Information Retrieval System", The 2nd Jordanian International Conference on Computer Science and engineering, JICCSE , Salt, Jordan. 89-97