USING RUSSIAN AND ENGLISH ONTOLOGY IN EXPANDING THE ARABIC QUERY

Boumedyen Shannaq
Information System Department, University of Nizwa ,Sultanate of Oman

***Abstract*— This paper present and evaluate a conceptual disambiguation method using Russian and English thesauruses, also this paper shows how the Russian language is more relative to the Arabic language than the English language. Employ the Russian and English thesaurus to over come the disambiguate issues in the Arabic language by expanding the Arabic Query in IR system. The main objective of this research is to evaluate the retrieval effectiveness by employing thesaurus in the process of expanding the Arabic query Using Russian and English. The significance of this research is to improve retrieval effectiveness of Arabic information retrieval (IR) system by employing thesaurus in the process of searching for this problem domain. This paper present the results of embedding ARBIC, RUSSIAN, ENGLISH thesaurus in the process of searching ARBIC documents from Translated Russian and English retrieval system. The results obtained shows that the retrieval effectiveness improves by 9 percent when the Russian thesaurus is employed in the process of retrieving , and 4 percent when the English thesaurus is employed in the process of retrieving the Arabic documents .The obtained results support the strengthens relationships between the Arabic and Russian Language .**

***Keywords*— Thesaurus, Word disambiguation ,Arabic language ,Arabic query .**

I. Introduction

Various approaches are grouped in three categories: (a) approaches based on feedback information from the user; (b) approaches based on information derived from the set of documents initially retrieved (called the local set of documents); and (c) approaches based on global information derived from the document collection. In the first category, user relevance feedback methods for the vector and probabilistic models are discussed. In the second category, two approaches for local analysis (i.e., analysis based on the set of documents initially retrieved) are presented. In the third category, two approaches for global analysis are covered. Our discussion is not aimed at completely covering the area; neither does it intend to present an exhaustive survey of query operations. Instead, our discussion is based on a selected bibliography which, we believe, is broad enough to allow an overview of the main issues and tradeoffs involved in query operations. Thus, Thesauruses is covered throughout our discussion. However, there is no intention of providing a complete survey of Russian and English thesauruses algorithms for information retrieval. User queries are often ambiguous phrases. In Arabic information retrieval system the user query is very complex. Identifying the correct meaning of the user query is a very complex task. In fact, one single term in the source language often have several meaning in the target language, which could increase user query ambiguity. In this paper, we propose and evaluate a new method for query t disambiguation for Arabic language information retrieval. The basis of our method is that two terms are a translation of each other if they have a highly similar conceptual representation, which means that the two words are mainly used in the same circumstances and are associated with the same set of words. The terms (word) conceptual representation is constructed using a language thesaurus to identify term synonyms and related words. Our method can be applied to any language information retrieval systems. However, in this paper we are interested in applying and evaluating our proposed method on the Arabic, Russian and English language retrieval [1] [2] [3].

A. Arabic languages Characteristics

The Arabic language has more complex morphology than the other language. Arabic language is highly rich in synonyms, a simple word such خطوات which means steps has the following synonyms (خطوات - مراحل – درجات – طرق – مسارات – نقاط –سبل – درب –اجراء – مرحلة – صيغة -نظام – اسلوب - شكل). This fact causes a problem of mismatch between the user query terms and the dictionary entries, as a result, part of the user query is not correctly handled and translated, thus leading to unsuitable or incomplete target query. For more information about the Arabic languages Characteristics see [4].

II. Background

More details can be found in [5]. Let's Consider the modified query vector qm generated by the Rochio formula and assume that we want to evaluate its retrieval performance. A simplistic approach is to retrieve a set of documents using qm , to rank them using the vector formula, and to measure recall-precision figures relative to the set of relevant documents (provided by the experts) for the original query vector q. In general, the results show spectacular improvements. Unfortunately, a significant part of improvement results from the higher ranks assigned to the set R of documents already identified as relevant during the feedback process [6] . Since the user has seen these documents already (and pointed them as relevant), such evaluation is unrealistic. Further, it masks any real gains in retrieval performance due to documents not seen by the user yet.

A more realistic approach is to evaluate the retrieval performance of the modified query vector qm considering only the residual collection i.e., the set of all documents minus the set of feedback documents provided by the user. Because highly ranked documents are removed from the collection, the recall precision for qm tend to be lower than the original query vector q. This is not a limitation because our main purpose is to compare the performance of distinct relevance feedback strategies (and not to compare the performance before and after feedback).

A. *Thesauri*

Thesaurus has become another valuable structure in any Information Retrieval system. It is a list of terms and concepts that provide a controlled vocabulary of words to use in document indexing, clustering, searching and retrieval the main components of a thesaurus are its index terms, the relationships among the terms, and a layout design for these term relationships. The layout design for term relation can be in the form of a list or in the form of bi-dimensional display. According to [3], the main purposes of a thesaurus are basically: (a) to provide a standard vocabulary (or system of references) for indexing and searching; (b) to assist users with locating terms for proper query formulation; and (c) to provide classified hierarchies that allow the broadening and narrowing of the current query request according to the needs of the user. The motivation for building a thesaurus is based on the idea of, using a controlled vocabulary for the index and search. A controlled vocabulary presents important advantages such as normalization of indexing concepts, reduction of noise, identification of indexing terms with a clear semantic meaning, and retrieval based on concepts rather than on words [7] [8] [9] .

III. Matching functions

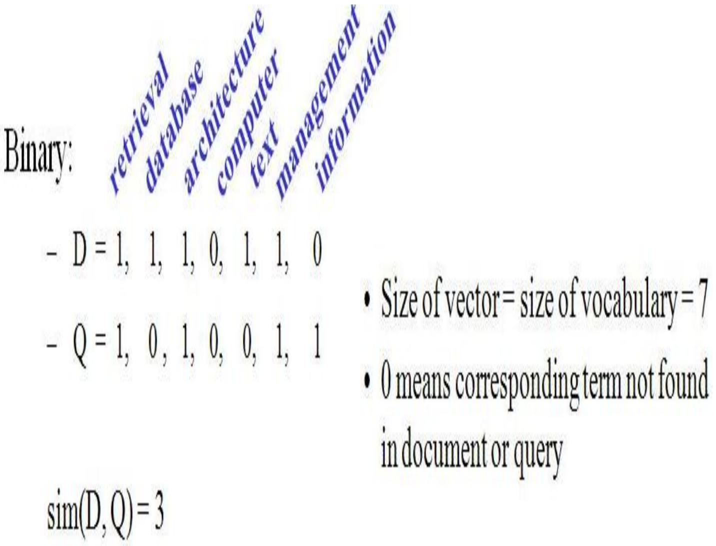
Many of the more sophisticated search strategies are implemented by means of a matching function. This is a function similar to an association measure, but differing in that a matching function measures the association between a query and a document or cluster profile, whereas an association measure is applied to objects of the same thing. Mathematically the two functions have the same properties; they only differ in their interpretations. [1] [7] there are many examples of matching functions in the literature. Perhaps the simplest is the one associated with the simple matching search strategy. If M is the matching function, D the set of keywords represents the document, and Q the set representing the query, then:

M = (2|D □ + □Q| / |D| + |Q|)

Sim (Di , Q) = (dik • qk)

where dik is the weight of Term k in document i and qk is the weight of Term k in the query .For binary vectors, the inner product is the number of matched query terms in the document For weighted term vectors, it is the sum of the products of the weights of the matched terms .

Inner Product – Examples



Binary:

retrieval database architecture computer text management information

- D = 1, 1, 1, 0, 1, 1, 0
- Q = 1, 0, 1, 0, 0, 1, 1

- Size of vector = size of vocabulary = 7
- 0 means corresponding term not found in document or query

sim(D, Q) = 3

Figure 3.1 : sample of similarity .

Weighted

D1 = 2T1 + 3T2 + 5T3 D2 = 3T1 + 7T2 + T3

Q = 0T1 + 0T2 + 2T3

sim(D1 , Q) = 2*0 + 3*0 + 5*2 = 10

sim(D2 , Q) = 3*0 + 7*0 + 1*2 = 2

IV. Experimental Details

In Thesaurus language retrieval, it is common for a single word in the source. Language has several translations, where some of them are with totally different meanings. Removing the noise terms will increase dramatically the retrieval performance. The basis of our method is that two terms are a translation to each other if they are used in the same circumstances or associated with similar set of words. In other words, have relatively similar sets of synonyms and related words. Our methodology is based on representing each term in the source (i.e., the Arabic Language) query and every possible Russian and English translation (of each Arabic term) by list of synonyms and related words using machine readable thesauruses. The identification of the set of most relevant translation in the target language (Russian, English) is based on the calculation of similarity between the Arabic term representation vector and the representation vectors of all possible Russian and English translation.

Sample of the experiment:

TABLE I

The Arabic QUERY:

خطوات الدراسة في جامعات روسيا

TABLE III

The Arabic query in Russian:

Этапы обучения в России

TABLE IIIII

The Arabic query in English:

The Way's of studying in Russia

For each keyword in the query we perform the following:

Find the similar words from Thesaurus for “خطوات"

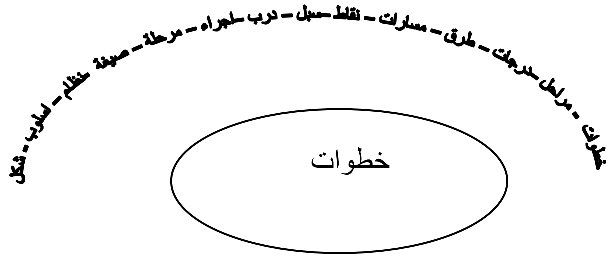

خطوات - مراحل - درجات - طرق - مسارات - نقاط - سبل - درب - اجراء - مرحلة - صيغة - نظم - اسلوب - شكل
خطوات

Figure4. 1 : Result of similar words from thesaurus for “ خطوات “

Using Russian Dictionary for translating “خطوات “into Russian

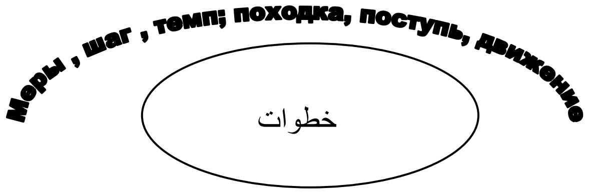


Figure4. 2 : Result of Russian translating for “ خطوات “

Find The similar words for each translating

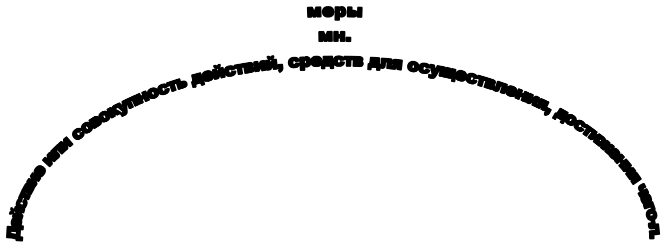

меры
мн.
Действие или совокупность действий, средств для осуществления, достижения чего-л.

Figure4. 3: Similar words for меры .

Translate all similar word of меры into Arabic , we obtained the following

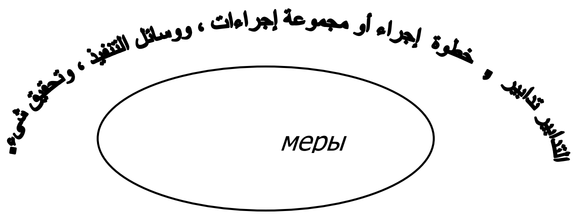

التدابير تدابير ، خطوة إجراء أو مجموعة إجراءات ، ووسائل التنفيذ ، وتحقيق شيء
меры

Figure4.4 :Arabic words related to меры

Calculate the similarities between Figure4.1 and Figure4. 3 using the following formula

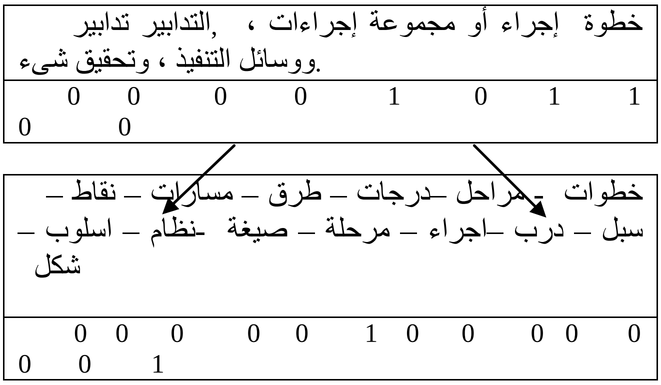
Similarity between KeywordsFig3i and KeywordsFig1Q can be computed as the inner vector product:

$$\text{sim}(Di, Q) = \sum_{k=1}^{t}(dik \bullet qk)$$

Where dik is the weight of Term k in Figure4.3 i and qk is the weight of Term k in the query in Figure4.1

For binary vectors, the inner product is the number of matched query in figure4.1 and terms in Figure4.3

Example :



خطوة إجراء أو مجموعة إجراءات ، التدابير, تدابير
ووسائل التنفيذ ، وتحقيق شيء.

0 0 0 0 1 0 1 1
0 0

خطوات ـ مراحل ـدرجات ـ طرق ـ مسارات ـ نقاط ـ
سبل ـ درب ـاجراء ـ مرحلة ـ صيغة ـنظام ـ اسلوب ـ
شكل

0 0 0 0 0 1 0 0 0 0 0
0 0 1

Sim(Fig3,Fig1)= 2

Perform the above step for English and so on .

Table 1 and table 2 shows the obtained result for Russian and English Similarity for “خطوات”

Table4. 1

Russain similarity results “خطوات ”

Word	SIMILARITY
Меры	2
шаг	3
темп	1
походка	0
поступь	2
движение	0

.

Table4. 2

English similarity results “خطوات ”

Word	Similarity
step	4
pace	1
stride	1
footstep	1
move	1
tread	1

V. Result and Discussion

Query expansion methods have been studied for a long time. While the success of expansion methods throughout the years has been debatable, more recently researchers have reached the consensus that query expansion is a useful and little explored (commercially) technique. Useful because its modern variants can be used to consistently improve the retrieval performance with general collections. Nowadays, there is controversy regarding the potential improvements to retrieval performance generated by stop words elimination, stemming, and index terms selection. In fact, there is no conclusive evidence that such text operations yield consistent improvements in retrieval performance. As a result, modern retrieval systems might not use these text operations at all. A good example of this trend is the fact that some Web search engines index all the words in the text regardless of their syntactic nature or their role in the text. Further more. It is also not clear that automatic query expansion using thesaurus-based techniques can yield improved retrieval performance. The same cannot be said of the use of a thesaurus to directly assist the user with the query formation process. In fact, the success of the `Yahoo!' Web search engine, which uses a term categorization hierarchy to show term relationships to the user, is an indication that thesaurus-based techniques might be quite useful with the highly interactive interfaces being developed for modern digital library systems . Also the results of this work support the idea of a thesaurus to directly assist the user with the query formation process , and they Can improved retrieval performance .The first experiment we carried out is an Arabic IR. Arabic queries that have been used are supplied to the system and the retrieval process is carried out on the Arabic collection of documents. This experiment is used as a base line of our disambiguation experiment. Table 6 gives the average precision of Arabic documents without running our method and three different disambiguation experiments. As we can see from Table 3, all Arabic retrieval resulted in low retrieval accuracy as compared to the Russian and English retrieval; the Every-Match disambiguation method performed the poorest while our conceptual translation method gave the best performance compared to all disambiguation and translation methods.

Table 5
Average precision and Recall value of sixty queries in different experiments

Retrieval Method	Precision	Recall
Arabic Query without using Thesauruses	70.3 %	88.5 %
Expanding the Arabic Using Russian thesaurus	79.2 %	96.5 %
Expanding the Arabic Using English thesaurus	74.4 %	89.2 %

Conclusions

Our results discover that there are a strong relationship between the Arabic and the Russian ontology, and shows that the performance of Arabic-Russian language retrieval using Russian thesaurus can be enhanced and improved to a precision of 79.2% and Recall of 96.5 of the Arabic retrieval. Also it show that the performance of Arabic-English language retrieval using English thesaurus can be enhanced and improved to a precision of 74% and Recall of 89.2% of the Arabic retrieval. Our proposed disambiguation and filtering approach overcomes some of those translation problems. It starts with considering all different translation so there is no lose of relevant translation; it then filters that different translation through a similarity comparison between the source language synonyms and related words and the different translation synonyms and related words. Translations that present a high similarity will only be considered in the retrieval process. However, if we consider the conceptual relations between the words contained in the query, we can improve translation and filtering quality. We are actually trying to define another version of query translation algorithm. In the near future, we will assess the quality of the translation, and how well it will improve document filtering.

References

[1] Boumedyen A.N. Shannaq , " Diagonal Name Search For Arabic (DNSA) " , First E-Technologies and Environment Conference (ETEC08) 15-16 April,2008 ,Sohar , Oman.

[2] P.P.Kokorin, B.Shannag, E.V.ShChelkunova , "Algorithm of normalization and ontological Clusters texts" information-measuring and operating systems Journal, http://www.radiotec.ru/catalog.php?cat=jr.(2010)

[3] Alekcandov V.V , Kuleshov S.V , Shannaq B. , " Phenomenon of identification" information-measuring and operating systems Journal, http://www.radiotec.ru/catalog.php?cat=jr.(2010)

[4] Boumedyen Shannaq, S.V. Kuleshov," Super Arabic morphological analyzer (SAMA1) "
St. Petersburg institute for Informatics and Automation of Russian RAS ,Academy of Sciences,
199178, Russia ,VAX UDC 003.9, information-measuring and operating systems Journal . http://www.radiotec.ru/catalog.php?cat=jr11.(2009)

[5] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. Journal of the American Society for Information Sciences,1976.

[6] Kokorin P. P. Kolesnikov R. A., Andreeva N. A, Frolov K. V., Boumedyen Shannaq, "The infological approach to develop edutainment systems". St. Petersburg institute for Informatics and Automation of Russian RAS ,Academy of Sciences, 199178, Russia ,VAX UDC 004.9, , information-measuring and operating systems Journal http://www.radiotec.ru/catalog.php?cat=jr11 . (2009)

[7] Boumedyen A.N. Shannaq , "Language Independent Product Name Search (LIPNS) " , First IEEE International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2008), VSB- Technical University of Ostrava, Czech Republic August 4- 6, 2008.

[8] P.P.Kokorin , Boumedyen A.N. Shannaq ,"Methods of texts normalization and ontological clustering of texts" , St. Petersburg institute for Informatics and Automation of Russian RAS ,Academy of Sciences,

199178, Russia ,VAX UDC ,information-measuring and operating systems Journal, http://www.radiotec.ru/catalog.php?cat=jr.(2010)

[9] Aksenov A.Y., Zaytseva A. A., Boumedyen Shannaq. "The rank method of text data regions localization" , information-measuring and operating systems Journal