INTERNET ARABIC SEARCH ENGINES STUDIES

Majed SANAN
Caen university, France
Sinane80@hotmail.com

Mahmoud RAMMAL
Lebanese University
mrammal@ul.edu.lb

Khaldoun ZREIK
Paris 8 university
zreik@univ-paris8.fr

Abstract

This paper is aimed to identify and explain the limitations and the problems of Arabic texts retrieving in the general search engines, and we have made many experiences on Arabic documents from the Lebanese official journal.

In our approach, we have used three "keyword matching" Arabic search engines: Google, Yahoo, and Idrisi (in case of keyword matching), and we have calculated the recall and the precision of our search experiments, and then we have compared the results in order to realize the limitations of this method.*

Keywords: *Keyword matching, Arabic search engine, Information Retrieval, Precision, Recall.*

1 Introduction

Modern Standard Arabic (MSA) is the official language used within 22 Arab countries.

Arabic language faces certain challenges in Information Retrieval (IR) for the following reasons:

First, orthographic variations are prevalent in Arabic [1]; certain combinations of letters can be written in different ways.

For example, sometimes in glyphs combining HAMZA or MADDA with ALEF the HAMZA or MADDA is dropped, rendering it ambiguous as to whether the HAMZA or MADDA is present.

Second, Arabic has a very complicated morphology.[2]

Third, broken plurals are common.

Broken plurals often do not resemble the singular form, they do not obey normal morphological rules, and they are not handled by existing stemmers.

Fourth, Arabic words are often ambiguous due to the tri-literal root system. In Arabic, a word is usually derived from a root, which usually contains three letters. In some derivations one or more root letters may be dropped, rendering high ambiguity between Arabic words with one another.

Fifth, short vowels are omitted in written Arabic texts.[2]

* Arabic search engine from Sakhr Company, http://www.sakhr.com.

Sixth, synonyms are widespread, perhaps because the variety in expression is appreciated as part of a good writing style by Arabic speakers (Noamany, 2001).

2 Information Retrieval

2.1 Definition

Information Retrieval (IR) is a process that informs on the existence and whereabouts of information relating to a specific request. Queries supplied by users are composed of a set of words interrelated by Boolean operators; the system responds by locating the documents containing combinations of these query words. The retrieval process is influenced by the indexing process as well as by the natural language that is being indexed.[3]

2.2 Retrieval Effectiveness (Precision and Recall)

The matching results of IR are imprecise and inexact as in Database, so that we need to measure the IR effectiveness.[4]

The question of how effective the Arabic language in case of retrieval purposes, should be raised.

The retrieval effectiveness of the Arabic language has not been tested yet.

The first step taken in developing the methodology was to map the differences of the Arabic language that might affect retrieval effectiveness. These can be seen simply as three features of Arabic. The Arabic language uses:

- Prefix
  For the definite article, some particles, and some plural forms
- Infix
  For some plural forms
- Suffix
  For some pronouns

The second step was to investigate how the effect of these features could be measured. As the formulation of the problem related to the term "differences", the methodology had to incorporate a comparison. It was decided to compare the retrieval effectiveness of Arabic to the retrieval effectiveness of English. Such a comparative analysis of two languages for retrieval purposes had not been carried out before.

Ever since the 1960's information retrieval (IR) effectiveness is evaluated using the twin measures of *recall* and *precision*.[5]

a) Precision:

The precision is the proportion of retrieved documents that is relevant.

Precision = |**relevant** ∩**retrieved**| ÷ |**retrieved**| = **P(relevant | retrieved)**

Precision = $(\frac{a}{a+b}).100\%$ [4]

In the above formula, a represents the retrieved relevant documents and b the retrieved non-relevant documents.

b) Recall:

The recall is the proportion of all relevant documents in the collection included in the retrieved documents.

Recall = |**relevant** ∩**retrieved**| ÷ |**relevant**| = **P(retrieved | relevant)**

Recall = $(\frac{a}{a+c}).100\%$ [4]

In the above formula, a represents the retrieved relevant documents and c the non-retrieved relevant documents.

c) Single-number measures:

We can also use a single-number measures for the effectiveness as follows:

F1 = 2PR / (P+R) ... where F1 as a harmonic mean of precision and recall. [4]

For this study relevance has been defined conceptually as:

A document is *relevant* to a query if the document has the same *aboutness* as the query.

The *aboutness* of the document means what the document is about.

3 Challenges of Arabic Language

3.1 Arabic Information Retrieval

Research has been done to improve Arabic IR through deploying techniques and methodologies, as morphology and others, to improve the recall and the precision.[6] Text REtrieval Conferences (TREC) of 2001, 2002 and Cross-Language Evaluation Forum (CLEF) 2002 had helped to show the achievements of different research groups in the area, and allowed a concrete evaluation of the participant systems.

Table 1 summarizes the techniques and approaches used by the participants in TREC 2001.[7]

Alternative indexing terms, the query languages, and (for cross-language runs) the sources of translation knowledge has been explored by the ten participating teams. All ten participating teams adopted a "bag-of-terms" technique based on indexing statistics about the occurrence of terms in each document. A wide variety of specific techniques were used, including language models, hidden Markov models, vector space models and inference networks.

Four basic types of indexing terms were explored, sometimes separately and sometimes in combination:

- Word: a single and isolated lexeme that represent a certain meaning.
- Stem: a morpheme or a set of concatenated morphemes that can accept an affix.
- Root: a single morpheme that provides the basic meaning of a word.
- N-gram: text strings decomposed into n-grams, i.e., substrings of length n, which usually consist of the adjacent characters of the text strings. Diagrams contain two and trigrams three characters.

	Arabic Terms Indexed				Query Languages	Translation Resources Used		
Team	Word	Stem	Root	n-gram		MT	Lexicon	Corpus	Transl.
BBN		X			A,E	X	X	X	
Hummingbird		X			A				
IIT	X	X	X		A,E	X	X		
JHU-APL	X			X	A,E,F	X			
NMSU	X	X			A,E		X		
Queens	X			X	A,E	X			
UC Berkeley		X			A,E	X	X		
U Maryland	X	X	X	X	A,E	X			X
U Mass	X	X			A,E	X	X		
U Sheffield	X				A,E,F	X			

A: Arabic, E: English, F: French.

Table 1. Techniques used in TREC 2002 for Arabic IR.

3.2 Features of Arabic Language

The Arabic language is an inflectional language and not an analytic language [1]. The derivation in Arabic is based on morphological patterns and the verb plays a greater inflectional role than in other languages. Furthermore, Arabic words are built-up from roots representing lexical and semantic connecting elements.

Arabic offers the possibility of combining particles and affixed pronouns to words.

In other words, Arabic allows a great deal of freedom in the ordering of words in a sentence.

Thus, the syntax of the sentence can vary according to transformational mechanisms such extraposition, fronting and omission, or according to syntactic replacement such as an agent noun in place of a verb.

The Arabic language is distinguished by its high context sensitivity in several directions. On the writing level, the shape of the letter depends on the letter that precedes it and the one that follows it. On the syntactic level, the different synthetic coherence relations such as case-ending, matching, connecting, associating and pronominalizing represent various examples of syntactic sensitivity.

The context sensitivity feature is not only limited to letters, words, and sentences. Arabic sentences are embedded and normally connected by copulatives, exceptive and adversative particles. For this reason it is more difficult to identify the end of an Arabic sentence than is the case in other languages.

Also, the "shadda" in Arabic language represents a higher accent on the character (in other languages shadda is represented by doubling the character when writing). Then we can have two words: one with "shadda" and another the same as the first one but without "shadda"; these two words can have different signification.

3.3 Ambiguities

Ambiguity is one of the challenging issues for search engines which make it difficult to adopt non-native Arabic search engines and challenges native search engines as well. The ratio of ambiguity in Arabic found to be larger than known in other languages [1].

Implemented solution approaches were tested with some evaluation. Xu, Fraser and Weishedel (2001) tested two techniques to handle the issue. First, sure-stem: where the word will be stemmed if and only if the word has one single stem. Second, all-stem: where the word is probabilistically mapped to all possible stems and assuming that all possible stems are equally probable. If a word had n possible stems, each stem gets 1/n probability.

Solutions for the problem and enhancements for the current technologies could be found in deploying a syntax analyzer, which will produce the right part-of-speech that can determine and eliminate some of the stems that would not fit in the structure of the sentence, word sense disambiguation or through statistical measures that can be drawn from corpora analysis to weigh senses based on frequency or co-occurrence.

4 Arabic Retrieval Strategies

4.1 Introduction

The aims of IR systems are to find relevant document and provide the user with clear control mechanism and a rapid response. A number of techniques and algorithms have been implemented within search engines.

Research and Development (R&D) in the Arabic text still has long way to go. Although academia has made significant achievements, the complex morphological structure of the Arabic language provides challenges; techniques must be found to make IR efficient for the Arabic language (Abdelali, Cowie and Soliman, 2004).

Existing Arabic text retrieval systems could be classified in two groups[1]:

- Full form based IR: Most of the commercial engines used are full form retrieval system. These include Sakhr web

engine www.sakhr.com; and www.ayna.com and other Unicode multilingual engines such as www.alltheweb.com or www.google.com.

• Morphology-based IR: The efforts that have been made in the academic environment to evaluate more sophisticated systems give an idea about the next generation of the Arabic search engines. Evaluation has been performed on systems using different approaches of incorporating morphology –stem, root based, light stem [7, 8].

Generally, using stemmers improve the recall as well as the precision. (Larkey, Connell. 2002) experiments showed that the light stemmer performs better than the regular stemmer.

While each of these methods is proposed as an alternative solution for Arabic text retrieval none of them is claimed to provide the optimum solution. For example, the word and stem methods are good at providing a more focused output but may miss relevant texts.

The root method, on the other hand, is very efficient at retrieving all related text but may retrieve a great deal of irrelevant text. This is the quest for a more effective method for Arabic Information Retrieval. The next section elaborates on Arabic IR and the challenges of the field.

4.2 Spelling Normalization and Mapping

Arabic orthography is highly variable. For instance, changing the letter YEH (ي) to ALEF MAKSURA (ى) at the end of a word is very common. (Not surprisingly, the shapes of the two letters are very similar). Since variations of this kind usually result in an "invalid" word, in our experiments we detected such "errors" using a stemmer (the Buckwalter Stemmer) and restored the correct word ending.

A more problematic type of spelling variation is that certain glyphs combining HAMZA or MADDA with ALEF (e.g. إ, أ and آ) are sometimes written as a plain ALEF (ا), possibly because of their similarity in appearance. Often, both the intended word and what is actually written are valid words.

This is much like confusing "résumé" with "resume" in English. Since both the intended word and the written form are correct words, it is impossible to correct the spellings without the use of context.

We explored two techniques to address the problem.

1) With normalization technique, we replace all occurrences of the diacritical ALEFs by the plain ALEF.

2) With the mapping technique, we map a word with the plain ALEF to a set of words that can potentially be written as that word by changing diacritical ALEFs to the plain ALEF. In this absence of training data, we will assume that all the words in the set are equally probable.

Both techniques have pros and cons. The normalization technique is simple, but it increases ambiguity. The mapping technique, on the other hand, does not introduce additional ambiguity, but it is more complex.

4.3 Arabic Stemming

Arabic has a complex morphology. Most Arabic words (except some proper nouns and words borrowed from other languages) are derived from a root. A root usually consists of three letters. We can view a word as derived by first applying a pattern to a root to generate a stem and then attaching prefixes and suffixes to the stem to generate the word [10]. For this reason, an Arabic stemmer can be either root-based or stem-

based.

5 Our Approach

5.1 Corpus

Lebanon has 128-member Chamber of Deputies (Majlis al-Nuwwab), elected for a four-year term of office by universal suffrage.

The Lebanese National Assembly has large legislative control and influence. It plays a crucial role in the orientation of the public, economic, political, and social life of the country. The Assembly differs from many other countries in that there is no upper house to share in legislative processes. Because of this, we are interested in the Lebanese parliament documents in which Lebanese official journal documents form the main part.

Then our test documents will be the 2002 Lebanese official journal documents that consist of about 2667 documents.

5.2 Experiment

In our approach we have chosen three Arabic search engines:

1. Idrisi
2. Google
3. Yahoo

applied in the case of "keyword matching" and on the 2667 Lebanese official journals (for the year 2002).

5.3 Methodology

The methodology in this experiment consisted of choosing keywords by the help of juridical experts, searching these keywords by using the above three search engine in case of "keyword matching" and applying the formulas seen in paragraph 2.2. But before starting the search, with the help of juridical experts we have indicated manually what must be the relevant documents (a+c) corresponding to each query chosen by them, and we have found the following results:

1-When searching the keyword "حرب" we must find 75 relevant documents (which we have indicated their titles manually).

2- When searching the keyword "المجلس" we must find 28 relevant documents.

5.4 Results

The results obtained by choosing the following keywords:

"حرب"، "المجلس"

Then we obtained the following results:

1- Al Idrissi search engine:

Key word	a	b	c	Recall	Precision	F1
حرب	11	110	64	14.67	9.09	11.22
المجلس	11	2023	17	39.28	0.54	1.065
Mean	11	1066.5	40.5	26.975	4.815	6.1425

2- Google search engine:

Key word	a	b	c	Recall	Precision	F1
حرب	10	115	65	13.33	8	9.99
المجلس	12	2064	16	42.85	0.58	1.14
Mean	11	1147	40.5	28.09	4.29	5.565

3- Yahoo search engine:

Key word	a	b	c	Recall	Precision	F1
حرب	13	105	62	17.33	9.07	11.91
المجلس	15	1983	13	53.57	0.75	1.48

Mean	14	1044	37.5	35.45	4.91	6.695

Then we can calculate the average mean of the means of the three above search engines:

Mean	a	b	c	Recall	Precision	F1
Mean 1	11	1066.5	40.5	26.975	4.815	6.1425
Mean 2	11	1147	40.5	28.09	4.29	5.565
Mean 3	14	1044	37.5	35.45	4.91	6.695
Mean	12	1085.83	39.5	30.17	4.67	6.134

5.5 Discussion

As we have seen above that the effectiveness of using "keyword matching" Arabic search engine is not good (The final mean value F1 of all means values is 6.695%), because of the challenges of Arabic language seen in paragraph 3.

For example when searching the keyword "حرب" we have obtained the documents containing "حربتا", which is a region in Baalbeck in Lebanon, or containing the keyword "حرب" but as a name of deputies, and these documents will be retrieved but not relevant to our search.

Also, other documents containing for example the keyword "عدوان", which is a synonym of "حرب", will not be retrieved and they are relevant to our search.

6 Conclusions and Future Work

Arabic is one of the most widely used languages in the world, yet there are relatively few studies on the retrieval of Arabic documents.

The main contribution of this paper is the studies of the application of "Keyword matching" search engine within the Arabic documents.

The Arabic language problems [1] make exact keyword match inadequate for Arabic retrieval.

Two techniques, spelling normalization and stemming, are well-known techniques for IR. Previous experiments [10, 11] show that while these techniques can significantly improve retrieval, they are not adequate.

The third technique, retrieval based on character n-grams, has been used by a few studies [12]. Then we have to think about using this technique for indexing Arabic documents in order to make the Arabic information retrieval more effective.

References

[1] Ahmed Abdelali. Improving Arabic Information Retrieval Using Local variations in Modern Standard Arabic, New Mexico Institute of Mining and Technology, 2004.

[2] Xu, A. Fraser, and R. Weischedel. 2002. Empirical studies in strategies for Arabic information retrieval. In SIGIR 2002, Tampere, Finland.

[3] Haidar M. Harmanani, Walid T. Keirouz, and Saeed Raheel. A Rule-Based Extensible Stemmer for Information Retrieval with Application to Arabic. *The International Arab Journal of Information Technology, Vol.3, No.3, July 2006*

[4] Victor Lavrenko. Center for intelligent Information Retrieval University of Massachusetts Amherst. *Hopkings IR workshop, 2005*

[5] G. Salton. "Another Look at Automatic Text Retrieval systems," *CACM*, 9 (7), pp. 648-656, 1986.

[6] Leah Larkey, Lisa Ballesteros, and Margaret Connell. Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In *SIGIR 2002*, pages 269to 274, 2002.

[7] Dumais, S. T. (1994) Latent Semantic Indexing (LSI) and TREC-2. In: D. Harman (Ed.),The Second Text REtrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215 , pp. 105-116

[8] Leah S. Larkey and Margaret E. Connell. Arabic information retrieval at UMass in TREC-10. In *Proceedings of TREC 10*, 2002.

[9] Shereen Khoja and Roger Garside. Stemming Arabic text. Computer Science Department, Lancaster University, Lancaster,UK, http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps,1999.

[10] James Mayfield, Paul McNamee, Cash Costello, Christine Piatko, and Amit Banerjee, JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video, and Web Retrieval. In E. Voorhees and D. Harman (eds.), *Proceedings of the Tenth TextREtrieval Conference (TREC 2001)*, Gaithersburg, Maryland, July 2002.

[11] P. McNamee, "Knowledge -light Asian Language Text Retrieval at the NTCIR-3 Workshop," *Working Notes of the 3rd NTCIR Workshop*, 2002.

[12] Darwish et al, 2001; Mayfield et al, 2001; Kwok et al, 2001.