Libri, 2001, vol. 51, pp. 63–74
Printed in Germany · All rights reserved

Libri
ISSN 0024-2667

Information Retrieval from Full-Text Arabic Databases: Can Search Engines Designed for English Do the Job?

HAIDAR MOUKDAD AND ANDREW LARGE
School of Library and Information Studies, Dalhousie University, Halifax, Canada
Graduate School of Library and Information Studies, McGill University, Montreal, Canada

The amount of electronic information in Arabic and other non-English languages available, especially on the World Wide Web, is increasing. Searches for such information can be undertaken on engines developed with the English language in mind, but will these engines work as effectively in other languages? This article investigates the impact on retrieval of prefixes in Arabic, which are far more common than in English. Typically search engines such as AltaVista designed implicitly for English include right hand (suffix) but not left hand (prefix) truncation. A test collection of 271 Arabic HTML records was created and indexed using the personal version of AltaVista. A series of searches was conducted on this collection, again using AltaVista. The results showed that searches on nouns stripped of prefixes reduced recall, in some cases dramatically, and that total recall of nouns can only be guaranteed by repeating searches that include the various prefixed versions of the nouns. The research questions the assumption that search engines designed with English in mind will work as well with different language structures.

Introduction

Despite the dominance of the English language in electronic information resources (Large and Moukdad 2000), a growing amount of electronic information is being generated in other languages. The global outreach of the World Wide Web in particular has accelerated this process. Many search engines developed explicitly or implicitly for the English language and the Roman script can also cope with other languages and other scripts; they will accept a search query and retrieve information in that language. The question that motivates this paper, however, is whether such engines will function as effectively as they do for English-language databases, or whether recall and / or precision ratios will be reduced because of the different linguistic structures involved. In a multilingual environment this is an important issue to address, because it is expensive either to modify existing search engines with new algorithms that take account of different morphological and syntactic structures or to design new search engines for each language.

There is a growing interest in cross-language information retrieval (CLIR) where a query can be input in one language but matched against a data set in another language (see, for example, Grefenstette 1998). Again, this interest has been stimulated by the multilingual environment offered on the Web. Many of the problems associated with CLIR relate to translation, but clearly difficulties will be exacerbated if the search engine itself functions less optimally with some languages than others.

As a contribution to the discussion on information retrieval in a multilingual environment, this paper deals specifically with the contrast between English and just one other language – Arabic. Does Arabic exhibit certain linguistic features that not only differentiate it from English but also

Haidar Moukdad, School of Library and Information Studies, Faculty of Management, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada, E-mail: hmoukdad@is.dal.ca
Andrew Large, Graduate School of Library and Information Studies, McGill University, Montreal, Quebec H3A 1Y1, Canada, E-mail: large@gslis.lan.mcgill.ca

might complicate information retrieval using search engines designed with English in mind?

Arabic

Arabic is one of the oldest languages in the world. It belongs to the Semitic family of languages and its "relatives" are Akkadian, Amharic, Aramaic and Hebrew. It is significantly different from English and other Indo-European languages in a number of important respects:

a) it is written from right to left;

b) it is mainly a consonantal language in its written forms, i.e. it excludes vowels;

c) its two main parts of speech are the verb and the noun in that word order, and these consist, for the main part, of triliteral roots (three consonants forming the basis of noun forms that derive from them); and

d) it is a morphologically complex language in that it provides flexibility in word formation: complex rules govern the creation of morphological variations, making it possible to form hundreds of words from one root (al-Fedaghi and al-Sadoun 1990).

The Arabic alphabet, adopted by several other languages of which Persian and Urdu are the major ones, consists of 29 letters including the glottal 'hamza' (it sounds like the 'a' in 'answer'), which appears as a separate character in written language, but is rarely used alone. In its complete written form, Arabic consists of two parts: the consonants (letters) and the vowel signs (over and underscores used with letters to indicate proper pronunciation). The consonants are the more important part of the written language, for they convey the basic meanings of the words (Ghani 1987). In fact, the common practice in written Arabic is to omit vowel signs in all published works, except poetry and some religious texts.

The following is a list of other characteristics that could create potential problems for Arabic IR that do not apply to English IR:

1) Some Arabic words, particularly the definite article 'al' and a number of conjunctions and prepositions, are not separated from their following word by a space. This results in a large number of entries being clustered together alphabetically in index files.

2) The peculiar morphology of Arabic might render methods used for English text retrieval inappropriate. As an example, the English phrase "and she wrote it" comprising four words would be written in Arabic as one word "wakatabathu" (wa=and, kataba=wrote, t=she, hu=it). In this case, "waw" (meaning "and") has been slightly transformed and linked and with the following word; this would create problems were it decided to treat "and" as a stopword (not uncommon in English-language retrieval systems).

3) It is common to find many Arabic words that have different pronunciations and meanings but share the same written form (homonyms), making finding the appropriate semantic occurrence of a given word a problem. English also has many homonyms, of course, but the problem is aggravated in Arabic by the absence of vowels in the written form, which then produces many identical consonant groupings.

4) In order to look up an Arabic word in a dictionary or index, it must be reduced to its root; unlike English, truncating the beginning or end of an Arabic word does not lead invariably to its root.

5) Every Arabic letter is pronounced as a word and cannot be used to represent one character like in English; therefore, in Arabic, acronyms and abbreviations are not found.

6) Arabic plurals are formed more irregularly than in English; depending on the root and the singular form of the word, the plural form might be produced by the addition of suffixes, prefixes or infixes, or by a complete reformulation of the word.

7) Most Arabic letters can be connected to other letters on both the left and right sides. However, there are some letters like the (waw) and the long vowel (alif) that cannot be connected on the left side to other letters. This results in the appearance of one space or more separating the letters forming one word.

8) A double letter (a silent letter followed by a vocalised one) in Arabic is denoted with a pronunciation mark (shadda) and is never spelled out.

The research problem

This paper deals with one specific but crucial issue in Arabic IR – does the common absence of left-hand truncation in search engines designed for English constitute a major problem for IR in Arabic, given the prevalence of prefixes in Arabic. [1] Although some engines designed for use with English-language databases do offer left-hand truncation (typically those operating in specialized domains such as chemical nomenclature), this is not a common feature (unlike right-hand truncation) because prefixes in English are much less likely than in Arabic to modify (rather than totally change) the meaning of a word. Prefixes exist in English, of course, but typically the resulting new word has little semantic relationship with the root term. For example, the addition of prefixes to the word "position" produces new

words such as preposition, supposition, deposition, imposition and so on, that can be ignored when trying to retrieve records about "position".

Research in Arabic IR

Morphological differences between Arabic and English seem to have inspired the few experiments that have been conducted so far on Arabic IR. Interestingly, research on Arabic IR has focused on using word roots and stems as index terms in collections of bibliographic records. This is based upon the assumption that the affix-rich morphology of the Arabic language will make any other indexing method ineffective. The first experiment that heralded interest in Arabic IR was conducted by al-Kharashi (1991), who explored the problems of storing and displaying Arabic bibliographic data, selection of index terms, ranking of retrieved Arabic records, and stemming algorithms for Arabic index terms. The main goal of his research was to try to find the best way to solve the problem of stemming for documents in Arabic. To test the proposed indexing methods, a microcomputer system for Arabic Information Retrieval (Micro-AIRS), developed by al-Kharashi, was used with a set of bibliographic records extracted from the databank at King Abdulaziz City for Science and Technology in Saudi Arabia. He performed a series of experiments using three indexing methods: the word itself, the stem, and the root. The root is defined as a bare verb form that can be triliteral, quadriliteral, or pentaliteral. (For example, "run" in English would be considered the triliteral root of runner, running and ran; talk the quadriliteral root of talker, talking and talked; and train the pentaliteral root of trainer, training and trainee). Most Arabic nouns and all Arabic verbs are morphologically derived from a short list of generative roots, about 1200 according to Hegazi and Elsharkawi (1985). The stem is a combination of a root and derivational morphemes to which one or more affixes can be added (to give an example using English, for the word "talkers", talker (talk= root, er=derivational morpheme) would be the stem and "s" would be the affix. In order to assess the effectiveness of the three indexing methods, 29 queries were entered against the database of 355 bibliographic records covering computer and information science. The results of implementing recall and precision measures demonstrated the superiority of root/stem-retrieval methods over word-retrieval methods, and underlined the contrast with IR methods in English.

Abu-Salem (1992) constructed an experimental Arabic IR system with 120 abstracts, applying the same indexing methods as al-Kharashi (1991) and repeating his experiments. He confirmed the results of al-Kharashi, ranking roots as the best indexing terms in Arabic, followed by stems and then words. He also concluded that the presence of abstracts improves retrieval regardless of the indexing method, and that the interactive use of a relational thesaurus, linking morphologically related words, gives the same good results as using roots as index terms.

Wien (n.d.) wanted to find out whether records in the Arabic script could be merged in an OPAC along with records in the Latin script and then searched without need for modifications to the engine (that had been designed for languages using the Latin script). Using a set of Arabic and English bibliographic records provided by RLIN (the Research Libraries Information Network), she describes a methodology to investigate the effect of Arabic prefixes, infixes and suffixes on retrieval.

Continuing on from the experiments of Abu-Salem (1992) and al-Kharashi (1991), Hmeidi, Kanaan and Evens (1997) built a database of 242 abstracts to determine the usefulness of automatically indexing Arabic words and to investigate the use of roots, stems and full words as index terms. The Arabic text was automatically indexed by every word according to specific rules and guidelines. Traditional measures of recall and precision were applied to the results of searches using both manual and automatic indexes, and the superiority of the latter was proved. One reason given for the feasibility of automatic Arabic indexing is that Arabic words typically appear less often than English ones in any given text. This has to do with the pattern and root rules mentioned above and with the morphological structure of Arabic. Because one root can produce a large number of words, and many words are created by adding affixes and connecting the definite article 'al', a large proportion of Arabic roots will appear only once, making the frequency of index terms (roots) low. This research also concluded that Arabic documents were best indexed by word roots. Root indexing increased recall and bypassed complex problems created by Arabic

morphology: a root index term would retrieve all variations of this root and eliminate the need to enter complex search queries taking account of all the words derived from that root. Likewise, the authors argued that roots made better index terms than words or stems, at least when phrases were not involved. When phrase searching is involved, however, searching by root is simply not feasible: it would produce root derivations that change the meaning of the phrase and therefore produce unexpected (and most probably undesirable) results. For example, searching for the Arabic equivalent of the phrase "international organization" would produce "organized nation", "organizations of nations", and even "international member".

While these various studies suggest that an ideal IR system working with Arabic text should have the capability of root and stem indexing, it is not clear if these types of indexing would be practical in a large bibliographic system, not to mention full-text searching environments. In fact, anyone who has tried looking up Arabic words in a dictionary will have noticed that words derived from one single root can cover several pages. It seems reasonable to assume, for example, that searching by roots in a full-text database will produce high recall but low poor precision. Other likely problem areas are phrase searching and searching for words that are not derived from known roots.

Web searching

The Web is fast becoming a huge reservoir of multilingual information. As such, it is creating a demand for IR using queries and data sets in languages other than English. Yet the focus of IR research largely has been on English-language documents and queries, using search engines designed with English in mind (see, for example, Taube 1995 and Wildstrom 1995). A number of evaluation studies have appeared, but typically they are based on the personal experience of the author with the search engines (for example, Chu and Rosenthal 1996). Such research has mostly led to recommendations about which engines to use in specific situations and which ones to avoid (Notess 1995; Courtois, Baer and Star 1995). Nicholson (1997) conducted a systematic analysis of Web searching, examining the indexing and abstracting methods of six popular search engines (Lycos, AltaVista, Excite, Open Text, Yahoo and Magellan). The evaluation was undertaken from the viewpoint of an indexer / abstracter, with emphasis on three IR aspects: collection methods, indexing, and abstracting.

Experimental procedure

The research described here had two objectives in mind:

- first, to investigate how Arabic words are handled by one Web search engine that has been developed primarily to handle the English language but is also capable of handling Arabic (and Arabic script), and to obtain a fairly credible estimate of the distribution of prefixed and non-prefixed nouns in documents;
- and second, to investigate the significance of statistical differences in the distribution of documents retrieved by searches with prefixes and without them.

In other words, the experiment was intended to reveal the extent to which the absence of a left-hand truncation capability in the search engine adversely affected retrieval from a language in which prefixes play a much greater role than in English. The overall goal was to explore the likely benefits of developing specific search algorithms for Arabic-language databases, and of adding indexing features to accommodate Arabic prefixes.

Database selection

In order to achieve these objectives, it was first necessary to create a test database of Arabic records extracted from the Web, and a set of query terms to match against that database. Full-text documents, as opposed to bibliographic records, have both a larger and richer vocabulary and therefore offer more investigation points in terms of the number of word occurrences and of word variations (and in any case, are more common on the Web). Possible Arabic documents types on the web include newspaper articles, technical documents, religious and literary texts, electronic books, business and country information, and government documents (consult URL: http://www.ayna.com [viewed May 18, 2001] for a comprehensive list). The documents selected had to meet two main criteria. First, they had to be reasonably homogeneous in subject content and therefore vocabulary to ensure that a limited set of queries could meaningfully be matched against

Table 1. *AltaVista* Search Engine Features

Features	Possible advantage	Possible problem
Every string of characters denoted by spaces and / or punctuation is indexed	N / A	The definite article, prepositions, conjunctions, etc. are connected to the beginnings of Arabic words and therefore will be indexed as part of the word rather than as separate words.
Wild card searching: Right-hand truncation	It should be helpful by stripping possessive suffixes, regular plural endings, and allowing spelling variations of the last letter (common in Arabic)	This will not handle irregular plurals (common in Arabic).
Wild card searching: Middle truncation	Spelling variations in the middle of Arabic words are rare. It might be useful in retrieving a group of words having the same root.	N / A
Character mapping (diacritical marks)	N / A	Latin characters with accents can be mapped to the original form. Diacritical marks in Arabic represent vowels and might be present or not present (problems of homonyms). Does character mapping work?
Phrase searching	N / A	In principle, this should work the same way it does in English. But Arabic words in a phrase may be arranged in many different ways while conveying the same meaning.
One-character words	N / A	Arabic documents might contain one-character abbreviations. Some letters might not be connected to others in a word and therefore be indexed (and retrieved) as separate words.
Lower and upper case	N / A	There is no upper or lower case in the Arabic alphabet, but letters can have up to four forms depending on their position in the word. This might create a problem if they are wrongly identified as being different letters.
Identifying important words in phrases	N / A	The software analyses queries and identifies meaningful English words for searching, discarding others. How does this feature behave in Arabic environment?
Ranking	N / A	One ranking criterion is the rarity of occurrence of a word in the index. Arabic words tend to occur less often than English words: This might affect the ranking process or render it useless.

them (see below). Second, they had to be capable of division into discrete records that could be searched. Random sampling of Arabic documents on the Web revealed that newspaper articles and electronic books would both meet the second criterion: they could be broken down into a series of records by article (newspapers) or chapter (book). Newspapers did not meet the first criterion, however, as their subject coverage inevitably is diverse.

One Web site was identified as meeting these criteria. It is dedicated to the publications of an Egyptian religious scholar, Yusuf al-Qaradawi, (URL: http://www.qaradawi.net [viewed May 18, 2001]) and includes electronic versions of a number of his published books in their original Arabic. While the subject matter of all these books deals with religious issues, the subject content and organization of one book, *the Lawful and Unlawful in Islam*, was especially appropriate. It deals with al-Qaradawi's interpretation of Islam's governance of the daily and social life of its adherents and is divided into four chapters containing 271 individual rulings on a wide range of topics. Each ruling has its own unique title and therefore can be treated as a separate information record.

The entire text of this book was downloaded from the Web and saved as 271 individual records, each representing one ruling. Each ruling

Table 2. Test Nouns and English Equivalents

	Arabic	English
N1	رسول	Prophet
N2	أولاد	Children
N3	قرآن	Quran
N4	طلاق	Divorce
N5	زوج	Husband
N6	كذب	Lying
N7	دين	Religion
N8	تجارة	Trade
N9	خمر	Alcohol
N10	شعر	Hair
N11	لباس	Clothes
N12	تحريم	Prohibition
N13	زواج	Marriage
N14	أسرة	Family
N15	حق	Right
N16	حب	Love
N17	أرواح	Souls
N18	حكم	Ruling
N19	مسلم	Muslim
N20	شرع	Law
N21	حلال	Lawful
N22	حرام	Unlawful
N23	لحم	Meat
N24	دم	Blood
N25	فائدة	Interest
N26	طاعة	Obedience
N27	طعام	Food
N28	إيمان	Faith
N29	موت	Death
N30	شرب	Drinking
N31	بنت	Daughter
N32	نفقة	Support
N33	أرض	Land
N34	أمانة	Honesty
N35	رشوة	Bribe

Table 3. Prefixes and Prefix Combinations

Arabic	Possible English Meaning(s)
A1: بـ (P1)	in, inside, by
A2: وبـ (P5+P1)	and in, and inside, and by
A3: كـ (P2)	as, like
A4: فـ (P3)	so, then, and
A5: الـ (P4)	the
A6: والـ (P5+P4)	and the
A7: فالـ (P3+P4)	so the, then the, and the
A8: كالـ (P2+P4)	as the, like the
A9: بالـ (P1+P4)	in the, inside the, by the
A10: وبالـ (P5+P1+P4)	and in the, and inside the, and by the
A11: و (P5)	and
A12: لـ (P6)	to, for
A13: ولـ (P5+P6)	and to, and for
A14: للـ (P6+P4)	to the, for the
A15: وللـ (P5+P6+P4)	and to the, and for the

was copied and pasted into a new HTML file that was saved under the title of that ruling (the resulting directory occupied 955,492 bytes of storage space). The HTML files (records) ranged in length from one paragraph to three pages. The indexing and searching software was configured to index every word in each file, producing an index of 69,209 words.

Search engine selection

Most search engines on the Web accept extended character sets and can therefore search and browse languages written in scripts other than Roman. The extended character sets may be available within the "standard" browser, or alternatively it may be necessary to install browsers specifically designed to handle individual scripts. A quick examination of the popular Web search engines using Sindbad, an Arabic add-on to the Netscape browser developed by Sakhr (URL: http://www.sakhr.com [viewed May 18, 2001]) revealed that most do accept Arabic characters (exceptions are Infoseek, Webcrawler and HotBot).

One of the most popular search engines on the Web, AltaVista, was selected for this study because it meets three critical requirements:

1) it indexes all words in a document (an important consideration in this research);
2) it was the only search engine at the time to offer a free version that can be installed on a local PC; and
3) it includes a variety of search features that can be tested in future experiments to measure their effects on Arabic-language searching (these features are briefly summarized in Table 1).

Search term selection

A set of search terms relevant to *The Lawful and Unlawful in Islam* was needed to match against the test database. Fortunately, a book called *fatawa moasera* (Current Rulings) can also be found on the Web (URL: http://www.qaradawi.net/arabic/books/fatawa-moasera/index-all.htm [viewed May 18, 2001]) that includes 43 questions posed to al-Qaradawi by a group of his followers on topics covered by the test database.

All nouns found in these 43 questions were identified (nouns are the most common search terms chosen by users to query an IR system) and

Table 4. Prefixed and Non-Prefixed (Naked) Noun Searching: Number of Retrieved Words

	Naked	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
N1	210	1	0	0	0	78	2	0	0	0	0	1	2	0	1	0
N2	2	0	0	0	0	7	2	0	0	0	0	0	0	0	0	0
N3	0	0	0	0	0	99	4	3	0	0	0	0	0	0	1	0
N4	9	0	0	0	0	71	1	1	0	4	0	0	0	0	0	0
N5	11	1	0	0	0	26	1	0	0	0	0	0	1	0	7	0
N6	0	0	0	0	1	8	2	0	0	0	0	0	0	0	0	0
N7	30	0	0	0	0	56	1	0	0	5	0	2	0	1	0	0
N8	12	2	0	0	0	29	3	1	0	0	1	0	1	0	3	0
N9	8	0	0	0	0	70	2	0	2	5	0	0	0	0	3	0
N10	6	1	0	0	0	9	4	0	0	1	0	0	0	0	0	0
N11	7	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
N12	70	0	0	0	2	35	8	0	0	3	0	8	1	0	0	0
N13	21	1	0	1	0	40	3	0	0	2	0	0	0	0	2	0
N14	3	0	0	0	0	11	2	0	0	0	0	0	0	0	0	0
N15	47	4	1	0	1	43	2	0	0	5	0	1	1	0	3	0
N16	6	0	0	0	0	3	1	0	0	1	0	1	1	0	0	0
N17	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
N18	20	3	0	1	0	20	2	0	0	0	0	1	0	0	0	0
N19	41	0	0	0	0	180	4	0	0	3	0	1	18	0	47	1
N20	9	1	0	0	1	4	1	0	0	1	0	3	1	0	0	0
N21	29	0	0	0	1	60	1	0	0	2	0	0	0	0	0	0
N22	112	3	0	0	2	70	28	0	0	1	0	0	0	0	0	0
N23	20	0	0	0	0	1	0	0	0	1	0	5	0	0	0	0
N24	9	0	0	0	0	19	8	0	0	0	0	0	0	0	0	0
N25	4	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0
N26	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
N27	7	1	0	0	0	15	0	0	0	0	0	7	0	0	0	0
N28	0	0	0	0	0	15	2	0	0	2	0	0	0	0	0	0
N29	2	1	0	0	0	13	0	0	0	0	0	0	0	0	0	0
N30	8	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0
N31	10	0	2	0	0	2	1	0	0	0	0	2	0	0	0	0
N32	0	0	0	0	0	9	3	0	0	1	0	0	1	0	0	0
N33	6	0	0	0	0	97	10	1	0	0	0	0	0	0	0	0
N34	2	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
N35	1	0	0	0	0	12	0	0	0	2	0	0	0	0	1	0

35 of these nouns were randomly selected to use as test search terms. These 35 nouns (with their English translations) are listed in Table 2.

Only single-word terms (rather than phrases) were required for this research as the objective was to establish how effectively AltaVista can cope with the morphological structure of individual Arabic words.

Unlike in English, one Arabic word can be formed from a combination of up to five linguistic entities: (conjunction + preposition + definite article + noun + pronoun). In addition to the 35 nouns used in their naked form, six prefixes were systematically attached to these nouns. To further complicate matters (and increase the actual number of search terms used), these prefixes can be combined in nine possible ways, as up to three different prefixes can be attached to a word at the same time (see Table 3).

The searches

Each of the 35 nouns was searched in 16 combinations: one search was undertaken using the noun's naked form without prefixes (as its English equivalent would be used); and 15 more searches were undertaken using the six prefixes and their nine possible combinations. In total, then, 560 searches were conducted.

Data analysis

Words retrieved in naked (non-prefixed) and prefixed noun searches.

The numbers of words retrieved by the 560 searches are shown in Table 4. *AltaVista* provides such a word count at the end of the results of a search. The 35 rows represent the 35 individual nouns

Table 5. Prefixed and Non-Prefixed Noun Searching: Number of Retrieved Documents

	Naked	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
N1	97	1	0	0	0	58	2	0	0	0	0	1	1	0	1	0
N2	2	0	0	0	0	5	1	0	0	0	0	0	0	0	0	0
N3	0	0	0	0	0	69	4	3	0	0	0	0	0	0	1	0
N4	6	0	0	0	0	17	1	1	0	2	0	0	0	0	0	0
N5	8	1	0	0	0	19	1	0	0	0	0	0	1	0	7	0
N6	0	0	0	0	1	8	2	0	0	0	0	0	0	0	0	0
N7	24	0	0	0	0	31	1	0	0	5	0	2	0	1	0	0
N8	8	2	0	0	0	9	3	1	0	0	1	0	1	0	2	0
N9	4	0	0	0	0	22	2	0	2	3	0	0	0	0	3	0
N10	3	1	0	0	0	3	2	0	0	1	0	0	0	0	0	0
N11	5	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
N12	38	0	0	0	2	27	4	0	0	3	0	7	1	0	0	0
N13	11	1	0	1	0	24	3	0	0	2	0	0	0	0	2	0
N14	3	0	0	0	0	9	2	0	0	0	0	0	0	0	0	0
N15	34	4	1	0	1	24	2	0	0	5	0	1	1	0	3	0
N16	6	0	0	0	0	3	1	0	0	1	0	1	1	0	0	0
N17	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
N18	17	3	0	1	0	16	2	0	0	0	0	1	0	0	0	0
N19	30	0	0	0	0	65	4	0	0	3	0	1	14	0	39	1
N20	7	1	0	0	1	4	1	0	0	1	0	2	1	0	0	0
N21	18	0	0	0	1	25	1	0	0	2	0	0	0	0	0	0
N22	46	3	0	0	2	31	7	0	0	1	0	0	0	0	0	0
N23	11	0	0	0	0	1	0	0	0	1	0	4	0	0	0	0
N24	7	0	0	0	0	8	7	0	0	0	0	0	0	0	0	0
N25	4	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0
N26	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
N27	5	1	0	0	0	12	0	0	0	0	0	5	0	0	0	0
N28	0	0	0	0	0	13	2	0	0	2	0	0	0	0	0	0
N29	1	1	0	0	0	7	0	0	0	0	0	0	0	0	0	0
N30	8	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0
N31	9	0	1	0	0	2	1	0	0	0	0	1	0	0	0	0
N32	0	0	0	0	0	5	3	0	0	1	0	0	1	0	0	0
N33	5	0	0	0	0	33	7	1	0	0	0	0	0	0	0	0
N34	2	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
N35	1	0	0	0	0	3	0	0	0	1	0	0	0	0	1	0

(identified here as N1 to N35), and the 16 columns represent the 16 different combinations (one naked and 15 prefixed) of the noun that were searched (identified as Naked, and A1 to A15). The individual cells reveal the number of words found within the 271 records by each noun in each of its possible combinations. So, for example, Noun 35 was found once in its naked form, 12 times when used in combination with prefix A5, twice with prefix A9, and once again with prefix A14.

Documents in naked and prefixed noun searches

The number of discrete documents (in contrast to words) retrieved by the 560 searches are shown in Table 5. As in Table 4, the rows represent the 35 nouns and the columns the one naked plus the 15 prefix combinations for these nouns. In this case, for example, Noun 35 retrieved one document in its naked form, three documents when used with prefix A5, one with prefix A9 and once again with prefix A14.

Discussion

Number of words in prefixed and non-prefixed searching.

The first column of Table 4 indicates that some Arabic nouns may not occur in a given set of records in their naked form, that is, without a prefix of some kind. For example, N3 is a very common noun, but its occurrences are limited to instances when it is prefixed. In total, five of the 35 nouns did not occur in this particular case in their naked form. More generally, there are wide differences between the frequency of occurrences of individual nouns. In their naked forms, the 35 nouns

Table 6. Total of Retrieved Words

	Naked	Prefixed
N1	210	85
N2	2	9
N3	0	107
N4	9	77
N5	11	36
N6	0	11
N7	30	65
N8	12	27
N9	8	82
N10	6	15
N11	7	2
N12	70	57
N13	21	49
N14	3	13
N15	47	61
N16	6	7
N17	0	3
N18	20	27
N19	41	254
N20	9	12
N21	29	64
N22	112	104
N23	20	7
N24	9	27
N25	4	5
N26	4	1
N27	7	23
N28	0	19
N29	2	14
N30	8	3
N31	10	7
N32	0	14
N33	6	108
N34	2	3
N35	1	15
Total	726	1413

Table 7. Total of Retrieved Documents

	Naked	Prefixed
N1	97	64
N2	2	6
N3	0	77
N4	6	21
N5	8	29
N6	0	7
N7	24	40
N8	8	19
N9	4	32
N10	3	7
N11	5	2
N12	38	44
N13	11	33
N14	3	11
N15	34	42
N16	6	7
N17	0	3
N18	17	23
N19	30	127
N20	7	11
N21	18	29
N22	46	44
N23	11	6
N24	7	15
N25	4	3
N26	3	1
N27	5	18
N28	0	17
N29	1	8
N30	8	3
N31	9	5
N32	0	12
N33	5	41
N34	2	3
N35	1	5
Total	423	815

occur in the database a total of 726 times, that is, an average of 20.75 occurrences per noun. Noun N1 was by far the most frequently occurring (210 times), almost twice as often as the next ranked, N22. Only eight nouns (N1, N7, N12, N13, N15, N19, N21, N22) have a close to or above-average frequency of occurrence, and they alone account for more than two thirds of word occurrences (560 out of 726). One prefix (A5) is particularly effective in retrieval, whereas others are of little effect (A10 and A15) although each of the 15 prefixes / prefix combinations increases word recall compared with the naked form by at least one.

Table 6 makes the difference in word recall clearer between the naked form and any of the prefixed forms. It compares the number of words retrieved without prefixes (column 1) with the number of words retrieved using one or more of the 15 prefix combinations (column 2). A total of 1413 words were retrieved using prefixed nouns in searches, almost double the number of words retrieved without prefixes. Of the 35 nouns, only eight retrieved more words when used without prefixes (N1, N11, N12, N22, N23, N26 N30, N31). This suggests that Arabic words occur more often with prefixes than without them, but how does this affect the number of retrieved documents?

Number of documents in prefixed and non-prefixed searching

Table 5 reveals that the 35 searches on the naked noun form retrieved 423 records, or an average of 12 records per noun. As with word retrieval,

Table 8. Ranking of Prefixes and Prefix Combinations

	A5	A6	A14	A9	A11	A12	A1	A4	A7	A2	A3	A8	A10	A13	A15
Words	1110	104	68	39	32	27	21	8	7	3	2	2	1	1	1
Documents	556	71	59	34	26	22	21	8	7	2	2	2	1	1	1

Noun N1 retrieved the largest number of documents (97), or more than twice as many as the next ranked (N22 with 46 hits). Only 10 naked nouns retrieved an above-average number of records (N1, N7, N12, N13, N15, N18, N19, N21, N22, N23), and these nouns account for more than three quarters of the retrieved documents (326 out of 423). Every prefix increased by at least one document the recall compared with use of the naked form alone (N26 was the least effective). Again, therefore, excluding prefixes from searches significantly decreases the number of retrieved documents.

Table 7, generated from Table 5, compares the numbers of documents retrieved with prefixes and without them. A total of 815 documents were retrieved using prefixes, almost double the number of documents retrieved when using the naked noun forms alone. In some cases (for example, N3 and N19) the increased recall by using prefixes is dramatic. Individually, only eight nouns (N1, N11, N22, N23, N25, N26, N30, N31) retrieved more documents when used in their naked form without prefixes.

Occurrence of prefixes

When non-prefixed nouns were used in searches, then, the number both of retrieved words and documents was substantially decreased. However, the data in Tables 4 and 5 suggest that not all prefixes and prefix combinations have the same effect on retrieval. Since prefixes are common in Arabic, a close look at their distribution in documents may be helpful in determining the importance of including them in search terms and in identifying the most common ones among them. Of the 15 possible prefix combinations, nine significantly increase the recall level. Generated from the numbers in Tables 4 and 5, Table 8 ranks the 15 prefixes by the number of times they retrieved words and documents. It reveals that some prefixes are much more effective than others in terms of retrieving both words and documents.

Conclusions and future research

An experimental information retrieval system was built using AltaVista indexing and search software and an Arabic document collection taken from the Web. A series of systematic searches using prefixed and non-prefixed Arabic nouns was carried out. The following conclusions can be drawn:

1. Arabic words are more likely to occur with prefixes.
2. Prefixes are a potential problem in Arabic IR.
3. Using a prefixed noun greatly increases the number of retrieved documents.
4. There is a possibility that some Arabic prefixes and combinations of prefixes can be excluded from search terms.
5. A search engine designed for English would not work as effectively with Arabic data.
6. Handling prefixes in Arabic words necessitates the development of new information retrieval algorithms for this language.

While the search engine handled Arabic records, its failure to undertake left-hand truncation was a clear drawback. To retrieve all possible forms of a given Arabic noun and make up for the absence of left-hand truncation, prefixes would have to be entered manually by the searcher. This is a time-consuming process, especially if every possible form of the Arabic word, using all possible prefixes, must be tried. In practice, it may be possible to exclude some of the prefix forms with any particular word as being unlikely to retrieve documents, but this pre-supposes that the searcher is well versed in the linguistic properties of Arabic in order to know which prefixes to use and which to safely ignore. Root indexing and searching rather than word indexing and searching might seem a solution for this problem, but this awaits further research; it is likely to improve recall but at the expense of precision.

Although the research described in this paper was limited to experimental conditions with a relatively small database, the high occurrence rate

of prefixes in Arabic words suggests the need to develop new search algorithms to accommodate Arabic data and improve search effectiveness, at least in terms of recall. The next stage will be to apply the methodology of the present research to a larger and more comprehensive test collection and to expand its focus to include additional aspects of linguistic structures that might affect information retrieval in Arabic.

The Web offers a wide variety of materials not only in terms of content and quality, but also in terms of language. Although English is the dominant language, considerable and growing holdings are available in many other languages. This fact has been accepted by the developers of most of the major Web search engines that offer users various specialised services: searching can be confined to records in a specific language, and search words can be entered in several non-Roman scripts providing access to data collections in these scripts. However, the search engines themselves operate in the same way regardless of the language being used in searching. These engines were designed to work with the linguistic structures of just one language – English. Can we assume that a search engine developed with English in mind will work just as effectively with other languages and therefore other linguistic structures? If not, is it necessary to develop new language-dependent search engines with information retrieval algorithms targeted at the linguistic structures of other languages? Or, will it be sufficient to simply add features to an existing search engine in order to accommodate these languages? It is interesting to note that Popovic & Willett (1992) concluded that the morphological complexity of any given language determined the extent to which stemming algorithms improved retrieval effectiveness (in this case they were reporting on an experiment dealing with right-hand stemming of Slovene words).

This research suggests that in the case of Arabic, at any rate, recall will be adversely affected if a typical English-language search engine is employed, and that modification (in this case by the addition of left-hand truncation) is necessary. Further research is needed to explore other possible problems with such search engines and Arabic data in order to clarify whether modification is possible, or rather whether specialised search engines targeted specifically at individual languages will be needed to optimize search efficiency.

Ever since the development of information retrieval as a discipline and a research area, the effectiveness of a given system or algorithm has been measured by the twin instruments of recall and precision. Not surprisingly, all the research that has been conducted on Arabic information retrieval has employed theses measures. It is not our contention to question this approach or its validity in typical situations, but we veered in our research away from these evaluation methods and went directly to what we see as the fundamental problem of information retrieval in any language: words and their structures. The crux of our approach is that measures of precision and recall (and the necessary judgements on relevance) cannot be employed before studying the linguistic problems that hinder the retrieval of documents: a document has to be found before it can be judged.

Arabic bibliographic records, like similar records in any other language, do not represent the richness of the language and, consequently, lack the linguistic variations that are present in full-text documents. Measuring the frequency of occurrences of prefixes, infixes and suffixes in such linguistically limited texts risks drawing the wrong conclusions about the significance of affixes on retrieval. Yet the previous research on Arabic IR has focused on bibliographic rather than full-text records. Today's data sets no longer are predominantly bibliographic, and it is important that research recognises the importance of experiments using full-text data.

While work such as that produced by al-Kharashi (1991) and Abu Salem (1997) present interesting findings, they fall short of addressing the fundamental problem we describe above. Our methodology has taken the issue of prefixes and treated it as the single most important morphological feature, with the potential to hinder effective retrieval of Arabic words and documents. By analysing the frequency of prefix occurrences and identifying the most common prefixes, we hope to draw attention to potential solutions to Arabic information retrieval problems. It is our belief that these problems are being greatly accentuated by the presence of Arabic on the Web, and this is where research should be focused. The ultimate goal is to find ways of accommodating

linguistic differences within existing search engines. A research endeavour such as ours not only contributes to the developments of new approaches to information retrieval in languages other than English, it, perhaps more importantly, promotes the idea of treating the linguistic problem at its root.

Note

1. Although Arabic words are written from right to left, in fact left-hand truncation here does refer to truncating at the beginning of a word, just as it would with English words. This is used to avoid confusion and preserve the standard terminology used in IR research.

References

Abu Salem, H. 1992. *A microcomputer based Arabic bibliographic information retrieval system with relational thesauri*. Unpublished doctoral dissertation, Computer Science department, Illinois Institute of Technology, Chicago, IL, USA.

Chu, H. and M. Rosenthal. 1996. Search engines for the World Wide Web: A comparative study and evaluation methodology. In *Global complexity: Information, chaos and control: ASIS annual meeting 1996*. Washington, D.C.: American Society for Information Science.

Courtois, M., W. Baer, and M. Stark. 1995. Cool tools for searching the Web: A performance evaluation. *Online* 19(6): 14–32.

Al-Fedaghi, S and H. Al-Sadoun. 1990. Morphological compression of Arabic text. *Information Processing & Management* 26(2): 303–16.

Ghani, A. 1987. Arabic literature: Uniterm indexing system for storage and retrieval. *International Library Review* 19(4), 321–333.

Grefenstette, G., ed. 1998. *Cross-Language Information Retrieval*. Boston: Kluwer

Hegazi. M and A. Elsharkawi. 1985. An approach to a computerized lexical analyzer of natural Arabic. *Computer processing of the Arabic Language, Workshop Papers*, 1.

Hmeidi, I., G. Kanaan and M. Evens. 1997. Design and implementation of automatic indexing for information retrieval with Arabic documents. *Journal of the American Society for Information Science* 48(10): 867–81.

Al-Kharashi, I. 1991. *Micro-Airs: Microcomputer based Arabic information retrieval system, comparing words, stem, roots as index terms.* Unpublished doctoral dissertation, Computer Science department, Illinois Institute of Technology, Chicago, IL, USA.

Large, A. and H. Moukdad. 2000. Multilingual access to Web resources: an overview. *Program* 34(1): 43–58

Nicholson, S. 1997. Indexing and abstracting on the World Wide Web: an examination of six Web databases. *Information Technology and Libraries* 16(2): 73–81.

Notess, G. 1995. Searching the World-Wide Web: Lycos, WebCrawler and more. *Online* 19(4): 48–53.

Popovic, M. and P. Willett. 1992. The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science* 43(5): 384–90.

Taubes, G. 1995. Indexing the Internet. *Science* 269: 1354–1356.

Wien, C. (n.d.). The tale of the unhappy Arabic books, or on the retrieval effectiveness of Arabic. URL: http://www.lib.umich.edu/libhome/Area.Programs/Near.East/cwien.htm [Viewed May 15, 2001.]

Wildstrom, S. 1995. Feeling your web around the Web. *Business Week* September 11: 22.