Webology, Volume 3, Number 1, March, 2006

Stemming and root-based approaches to the retrieval of Arabic documents on the Web

Haidar Moukdad
Assistant Professor, School of Information Management, Dalhousie University, Halifax, Nova Scotia B3H 3J5 Canada Phone: 1-902-494-2462, E-mail: haidar.moukdad (at) dal.ca

Received December 29, 2005; Accepted March 25, 2006

Abstract

Using information retrieval systems to gain access to documents in languages other than English is becoming an increasingly significant problem. Rules, theories, algorithms, and retrieval methods designed and developed for English and other morphologically similar languages may or may not apply in the linguistic environments of other languages. The problem is particularly acute in languages that differ radically from English on account of morphological rules. This paper compares the effects stemming and root retrieval on information retrieval in Arabic through an exploratory study of the handling of Arabic words by an English-language search engine (ELSE). Search experiments, using 2000 Arabic documents and 40 Arabic search terms (nouns), were conducted in a Web search engine developed for English (AltaVista) and in an Arabic search engine (al-Idrisi) to compare the performances of stemming and root retrieval and to investigate the possibility of adapting AltaVista for use with Arabic text. The results of the experiments show that more effective retrieval can be accomplished through stemming, and that it is possible to adapt an ELSE for use with Arabic without the need to develop root-retrieval features.

Keywords

World Wide Web, Search engines, Arabic language

Introduction

With the continuing explosive growth of the Internet and the proliferation of textual information in a multitude of languages other than English on the Web, retrieval of documents in these languages is becoming an increasingly significant problem. Rules, theories, algorithms, and retrieval methods designed and developed for English and other morphologically similar languages may or may not apply in different linguistic environments. Nowhere could the problem be more acute than in languages that differ radically from English in morphology and word-formation rules. Words, being the gist of written and spoken information queries, are by far the most fundamental elements of expression, and they form the basic components of meaningful information exchanges.

Information retrieval (IR) is a communication process that relies on language to perform its functions. The content of documents and information records are represented by language elements, and the information problems of users are also expressed in terms of

language (Harter, 1986). All human languages have vocabularies, corpora of words whose elements constitute the building blocks from which meaningful communication constructions can be formed. Words form phrases, phrases form sentences, sentences form paragraphs, and paragraphs form documents. If we think of a document as a collection of words, then it is easy to contemplate the role played by the structure of a language in providing access to information within this document. Words are formed according to specific rules and guidelines that differ among languages, creating IR problems and potential solutions that need to be investigated with the language involved in mind. In the early days of IR systems and for a few decades after that, this issue was not as crucial as it is today. In fact, the systems of those days were developed in English, for English, and with English in mind. And, since most of the available electronic databases were in English, search and retrieval software, indexing methods, and user interfaces were designed specifically for this language. As this is no longer the case, IR systems have been developed for languages other than English, and search engines have increasingly been modified to handle these languages. This paper explores the handling of Arabic words in English search and retrieval environment represented by AltaVista, and it presents specific approaches to assessing stemming and root-based retrieval methods to accommodate the peculiarities of Arabic word-formation rules within the framework of this environment.

Arabic is one language that is likely to present challenges in a traditional IR environment (Khurshid, 1997) and in popular search engines, because its morphology and word-formation rules are radically different from those of English. These rules are based on a root-and-pattern system that has been long thought to be a major factor in hindering IR operations (Beesley, 1996). Finding all possible words that are derived from an Arabic root might not necessarily lead to better IR performance. This, coupled with the lack of an empirical evidence to support the need for major changes that necessitate the development of dedicated Arabic IR systems, highlights the need for studies focused on identifying alternatives. The cost of developing new systems or radically modifying existing ones can be prohibitive, and without a clear need for such undertakings they will be no more than expensively futile exercises. While researchers on Arabic IR (al-Kharashi, 1991; Abu Salem, 1992; al-Kharashi & Evens, 1994; Hmeidi, Kanaan & Evens, 1997; Abu Salem, al-Omari, & Evens, 1999) advocate the use of advanced word stemming and root extraction algorithms, the limited scope of their research leave many questions unanswered. Careful examinations of the effects of specific aspects of Arabic word-formation rules on IR (Moukdad, 1999) could be indispensable to isolate manageable and necessary areas of improvement. With the notion of taking advantage of existing English systems in mind, the objectives of this paper are to

1. Present an overview of the Arabic language and compare Arabic and English word-formation rules with focus on nouns;
2. Develop a methodology for studying the possibility of effective Arabic IR in a non-dedicated search engine;
3. Explore the handling of Arabic words and document retrieval in an English IR environment;
4. Propose techniques to handle Arabic words in an ELSE; and
5. Conduct a preliminary examination of the practicality of root-based retrieval.

The Arabic Language

1. Arabic and the Root-and-Pattern System

Arabic is written from right to left and belongs to the Semitic family of languages. Although different spoken Arabic dialects exist throughout the Arab world, there is only one form of the written language found in printed works, and it is known as فصحه or

Standard Arabic (henceforth referred to as Arabic). Semitic languages differ in structure and grammar, but they share one characteristic that facilitated transition from one to another. In most cases, lexical forms (words) in these languages are derived from basic building blocks with tri-consonantal roots at their bases. The word building process starts with the three letters of a root and follows a regular set of word patterns. All traditional Semitic-language dictionaries and most modern ones are arranged by root. Instead of listing alphabetic entries, these dictionaries arrange words under entries of the roots that produce them. To look up a specific word, the user has to have enough knowledge to isolate the root then locate its entry. It is as though words like ascribe, describe, subscribe, circumscribe, proscribe, prescribe, inscribe were listed in an English dictionary under the Latin root "scribere" that describes the basic idea of writing/drawing (DeYoung, 1999). The difference is that the words grouped under an Arabic root can be analyzed down to the letters of a root and the predefined morphological patterns that created them.

One of the standard Arabic lexicons, (لسان العرب or the Language of the Arabs), lists 6,350 triliteral roots and 2,500 quadriliteral ones. Out of these, only about 1200 are still used in modern Arabic vocabulary (Hegazi & Elsharkawi, 1985), and the great majority of words can be analyzed down into triliteral roots consisting of three consonants or radicals (Ziadeh & Winder, 1957). Although the description given here focuses on triliteral roots and the patterns that apply to them, it should be sufficient to give a general idea about how the system works.

Words constructed from the same root constitute what is traditionally called a morpho-semantic field, where semantic attributes are assigned through patterns governed by morphological rules. The meaning that is inherent in the root is shared by all words in this field. However, the patterns that produce these words make them semantically distinguished (Rafea & Shaalan, 1993). A similar process can be noticed in English if we look at "necessitate", "necessary", "unnecessary" and "necessarily". While all four words share the basic meaning that is inherent in necess (need), they convey different semantic messages: necessitate (to produce the need), necessary (needed), unnecessary (not needed), and necessarily (in need condition/mode). We could say that adding "itate" to the root created the verb necessitate, "ary" the adjective necessary, and so on.

In general, each pattern is associated with a meaning which, when combined with the meaning conveyed by the root, gives a final meaning to the derived word (Moutaouakil, 1987). Using patterns to create different morphological variations from a root is a fairly regular process. It is similar to a mathematical formula, where the original letters are constant variables, and changing variables are letters added in the beginning, middle or end of the root. Patterns may also be indicated by vowel changes only; in these cases no letters are added to the root and, for present purposes, the structure of the word is considered unchanged. Traditionally, Arab grammarians have used the letters ف , ع and ل as generic letters to represent the root and the patterns. Patterns are based on these three letters, and, in derived words, the order of these letters is always the same: ف is first, ع second, and ل last. Table 1 shows a selection of noun-derivation patterns and illustrates examples of their usage.

Table 1. A Sample of Noun-Derivation Patterns

Pattern	Sample Roots	Derived Nouns
فاعول	جرر (to pull), حسب (to count)	جارور (drawer), حاسوب (computer)
فعال	حرم (to deny), دوم (to last)	حرام (unlawful), دوام (work shift)
فعاله	زرع (to plant), صنع (to make)	زراعه (agriculture), صناعه (industry)
فعيل	كبر (to grow), غسل (to wash)	كبير (big), غسيل (laundry)
فعلان	زعل (to grieve), كسل (to neglect)	زعلان (sad), كسلان (lazy)

فعله	حرب (to battle), دفع (to pay)	حربه (spear), دفعه (installment)
فعول	خجل (to hesitate), شكر (to thank)	خجول (shy), شكور (thankful)

Table 1 shows only a small fraction of Arabic patterns. There are hundreds more that convey all kinds of meanings. It is important to keep in mind that these patterns are not arbitrary and should not be used as so. Learners of Arabic have traditionally relied on the root-and-pattern system to practice correct use of words and to enhance their vocabulary knowledge. This system is also used to derive different forms of a base noun as explained below.

2. Arabic Word Formation

In linguistics, word formation is a function of morphology. Morphological analysis of human languages is largely based on the following linguistic elements: root, stem, affixes (prefixes, infixes and suffixes), and morphemes De Guzman & O'Grady, 1987). These elements are used in the IR field; therefore, a clear definition of their roles in word structures is essential. The function of the Arabic root has already been explained, but a general explanation of the term root as used in IR and in general linguistics is given here.

Arabic roots are forms of the verb; while in English and many other languages a root can be an adjective, a noun or a verb. A global definition of the root is that it is word that can stand on its own without the need for additional morphological elements. At the same time, this word cannot be broken down to smaller elements. However, a root can accept the addition of elements to create new words (Crystal, 1985). Run for example is a root: it is a complete word with a meaningful semantic representation. This word cannot be broken down to generate new words like ru or un. However, we can add s to run to obtain runs, ing to obtain running, and er to obtain runner. When we add s, ing and er to run it is also called a stem. The linguistic elements s, ing, and er are suffixes because they are added at the end of the stem and they cannot exist in isolation from the word. That said, a morpheme is the smallest meaning-bearing unit of the composition of a word. For example, run has one morpheme (run), runs has two (run and s). Table 2 uses examples from English and Arabic to illustrate the relationship between root and stem, to show the differences between prefixes, infixes and suffixes, and to explain the concept of morphemes.

Table 2. Roots, Stems, Affixes, and Morphemes in English and Arabic Words

Word	Root	Stem (s)	Prefix	Infix	Suffix	Morphemes
attract	attract	none	none	none	none	attract
attractive	attract	attract	none	none	ive	attract, ive
attractively	attract	attract, attractive	none	none	ive, ly	attract, ive, ly
unattractive	attract	attract, attractive	un	none	ive	un, attract, ive
قتل	قتل	none	none	none	none	قتل
قتيل	قتل	قتل	none	ي	none	قتل, ي
مقتول	قتل	قتل, قتول	م	و	none	م, قتل, و
مقتولون	قتل	قتل, قتول, مقتول	م	و	ون	م, قتل, و, ون

In Table 2, an adjective is created from the verb attract by adding the suffix ive; similarly, an adverb is created from the adjective attractive by adding the suffix ly. The suffixes ive and ly are derivational suffixes, and the generation process is called derivational morphology, because new grammatical categories of the word (parts of speech) are derived: verb –> adjective and adjective –> adverb. Conversely, the process of attaching a suffix like s to a noun (car –> cars) or to a verb (eat –> eats) is called inflectional

morphology, because it does not create a new grammatical category from the word (word class is not affected); inflections typically encode person, number, gender features (Matthews, 1974). In this case, car and cars are both nouns and eat and eats are verbs. The inflectional suffix s inflects the noun to indicate the number (singular and plural) and the verb to indicate the person of the subject (first and third).

Arab linguists identify only three parts of speech: 1) the verb, 2) the noun, and 3) the particle (Mehdi, 1986). This is a broad categorization by which nouns (as defined in English), adjectives, and pronouns are all classified as nouns. As opposed to English, Arabic adjectives are not treated separately from nouns. In fact what is considered an adjective in English can be an adjective or noun in Arabic. Let us look at the English phrase: "the big boy of the class". The Arabic equivalent reads something like "كبير اولاد الصف", which translates roughly as "the big of the boys of the class". The English adjective "big" translates as كبير but in the Arabic phrase كبير is a noun. However, we could say "ولد كبير" (a big boy); كبير, in this case, is an adjective. For present purposes, Arabic nouns and adjectives are simply referred to as nouns. There would be no distinction made between the two, and the treatment of word formation disregards any discrepancies with English terminology.

Particles, the third part of Arabic speech, include prepositions, conjunctions, interjections, question particles and answer particles. Only particles that attach to nouns will be treated here because they are considered affixes and they affect IR procedures. In general, these are prepositions and conjunctions like ل (to) and و (and) in للمدرسه (to the school) and وكره (and a ball). Particles that cannot be attached to nouns are usually connected to pronouns or they occur alone like في in فيه (in it) or عن in عني (about me).

Based on the categories of Arabic speech, the concept "word formation" is used for present purposes to describe the use of inflectional affixes to generate new forms (sub-classes) from the base form of an Arabic noun, e.g., singular to plural or masculine to feminine. It does not, however, cover proper nouns (such as people, place, day and month names, etc.); these types of nouns usually do not have variants and are not affected by word-formation rules. Prefixes and suffixes in the form of particles and pronouns that do not create sub-classes from the base noun are treated separately below. The base form of the noun is the masculine singular or the feminine singular form if a masculine one does not exist. For example, the masculine noun كتاب (book) is a base form for the plural كتب (books) and for the feminine singular كتابه (writing). By the same token, the feminine noun طاوله (table) is the base form for the plural طاولات (tables) since it does not have a masculine form.

There is no neutral gender in Arabic; nouns are divided between masculine and feminine. This division is grammatical not natural, because nouns do not necessarily have to be male or female (Cowan, 1958). In general, the feminine is formed from the masculine by adding the suffix h. For example, طالب is a male student طالبه is a female student. In other instances, the masculine and feminine forms of a noun do not share a common root as in رجل (man) and امراه (woman).

Dual indicates a number of two and is form by adding the suffix ان to singular masculine and feminine nouns in the nominative case. The dual form of the masculine قلم (pen) is قلمان (two pens); مدرستان (two schools) is the dual form of the feminine مدرسة (school). The suffix ان is changed to ين to indicate accusative or genitive cases: الولد اكل تفاحتين (the boy ate two apples).

The plural form indicates any number higher than two; it is of three types. The first type, the sound masculine plural, is formed by adding the suffix ون to the base masculine noun in the nominative case: معلمون (teachers) is the plural of معلم. In the accusative and genitive cases the ون is changed to ين as in معلمين (teachers in the accusative). The second type, the

sound feminine plural, is constructed by dropping the suffix ه from feminine nouns in the nominative cases and adding ات at in its place. For example, ورقات (papers) is the plural of the feminine noun ورقه. The suffix ات does not change in the accusative and genitive cases. Constructing the third type, the broken plural, is more complex than the sound ones. Broken plural forms of masculine and feminine nouns are derived through the use of a pattern system similar to the one mentioned above, and case is indicated through the use of the vowels. Murtonen (1964) lists 82 of the most common patterns in addition to many rarely used ones. Table 3 shows a sample of ten of these patterns and their usage. The patterns are used with masculine and feminine nouns where it is not possible to construct a sound plural form. Applying these patterns might involve the addition or omission of prefixes, infixes, suffixes, or a combination of two or three of these affixes. The pattern فعال, for example, is applied to create the broken plural رجال of the masculine singular رجل (man). The pattern فواعل produces the plural عواصف of the feminine singular عاصفه (storm), and افاعل produces the plural اغاني of اغنيه (song).

Table 3. A Sample of Broken Plural Patterns

Pattern	Singular noun	Plural Noun
افاعيل	حديث (conversation)	احاديث (conversations)
افعال	حزب (political party)	احزاب (political parties)
فعل	كتاب (book)	كتب (books)
فعلان	قطيع (flock)	قطعان (flocks)
فعول	اسد (lion)	اسود (lions)
مفاعيل	مقعد (seat)	مقاعد (seats)

3. Particles and Pronouns

Particles and pronouns affect the construction of Arabic words because, as opposed to their English counterparts, they are usually attached to verbs and nouns (Haywood, 1960). Possessive pronouns and particles (including the definite article) are attached to nouns in the form of non-inflectional prefixes or suffixes (see Tables 4 and 5). For instance, possessive pronouns are always attached as suffixes (the ي in بيتي (my house)), while the definite article ال is attached as a prefix (البيت (the house)). This phenomenon is so widespread in the language that the number of occurrences of nouns with these prefixes and suffixes is much higher than without them (Yahya, 1989). For example, virtually every Arabic noun accepts the prefix ال (the definite article), and the conjunction و (and) is always attached to the word that follows it. The prefix ك (The equivalent of the English word like in "sweet like honey") does not occur in isolation from the noun. Instead, the Arabic equivalent of "sweet like honey" is "حلو كالعسل", where العسل is honey.

Table 4. The Non-Inflectional Suffixes (possessive pronouns)

Suffix	Person/gender/number	Example
ي (my)	First/both/singular	بلدي (my country)
ك (your)	Second/both/singular	بلدك (your country)
كما (your)	Second/both/dual	بلدكما (your country)
كم (your)	Second/masculine/plural	بلدكم (your country)
ه (his)	Third/masculine/singular	بلده (his country)
ها (her)	Third/feminine/singular	بلدها (her country)
هما (their)	Third/both/dual	بلدهما (their country)
هن (their)	Third/feminine/plural	بلدهن (their country)
هم (their)	Third/masculine/plural	بلدهم (their country)

Table 5. The Most Common Prefix Particles

Prefix particle	Meaning	Example
ال	the	الشارع (the street)
ب	in, with	بمجالك (in your field)
ف	and, therefore	فرئيس (and president)
ك	like, as	كجامعه (like university)
ل	for, to	لمدينه (to city)
و	and	وجرس (and bell)

Particles are far more common than possessive pronouns and they can occur alone or in combination in the beginning of a noun. Up to three of them can be attached to a noun. For example, the definite article can be preceded by any one of the other five prefixes. Table 6 shows some of the most common combinations and gives examples of their use.

Table 6. Prefix Particle Combinations

Combination	Meaning	Example
بال	in the	بالشارع (in the street)
فال	and the, therefore the	فالمدينه (therefore the city)
كال	like the	كالرئيس (like the president)
لال	for the, to the	للمجال (to the field)
وال	and the	والجامعه (and the university)
فبال	therefore in the	فبالحق (therefore in the right)
وبال	and in the	وبالوسط (and in the center)
وكال	and like the	وكالشمس (and like the sun)
ولال	and for the	ولليسار (and for the left)
فب	and in, therefore in	فبنوم (therefore in sleep)
وب	and in	وبحركه (and in movement)
فل	and for, therefore to	فلمعركه (and for battle)
ول	and for, and to	ولزمان (and to time)

4. Arabic Nouns in IR

The most salient problem in an IR system is to improve recall rates while retaining a high level of precision (van Rijsbergen, 1979). Retrieving morphological variants of a word is a technique that is meant to enhance recall. Because of the dominance of the root system, and the large number of derivation possibilities, morphological variants of a word are not always semantically related. Under the root قصد, for example, we can find قصد (intention) and قصيده (poem). It is safe to assume that a user searching for قصيده would not be interested in قصد. Instead, this user would be interested in قصيدتان (two poems), قصائد (poems), and in all occurrences of these words with possessive pronouns and prefixes. For present purposes, morphological variants of an Arabic noun are divided into three groups: root based (nouns grouped under one root), inflected (feminine, dual, plural, etc.), and affixed (attached to particles and possessive pronouns).

In theory, looking up a word in an IR system with root searching capabilities is a concept related to using a traditional Arabic lexicon. However, instead of figuring out the root of the word and then looking it up in the lexicon, the IR system analyzes the word down to its root and retrieves documents that contain any morphological variation derived from that root (al-Kharashi & Evens, 1994). The IR system should also retrieve inflected and affixed

variants of a noun, which are not usually listed in a lexicon. With this ultimate variant retrieval, potential problems might arise. As explained above, save for the root, the search noun might not have much in common with many of the retrieved nouns. For instance, searching for علم (flag) will retrieve any document that contains words such as علامه (scholar), تعليم (teaching), and عليم (expert). It will also retrieve all affixed and inflected variants of these words in addition to all possible forms of the verb علم (to know).

The problem of retrieving inflected and affixed variants of Arabic nouns has to be approached from two directions: one dealing with suffixes (feminine, dual, sound plural, personal pronouns, etc.), and another dealing with prefixes and infixes (particles and broken plurals). In a traditional IR system, suffixes can be handled through stemming (at the indexing stage) or end truncation (using a wild card character, like * or ?, to replace a string of characters at the end of the word at the search stage); this will reduce the search term to a stem and allow the retrieval of documents containing its variants. Searching for the English truncated term run*, for example, will retrieve runner, runners, and running. Arabic Suffixes can be handled in the same way. To retrieve variants of a noun, it is sufficient to truncate the search term: *مكتوب (letter) will retrieve مكتوبي (my letter), مكتوبان (two letters), مكتوبها (her letter), etc.

A traditional IR system will handle infixed variants, but the user has to be well versed in Arabic to use middle truncation. In the simplest forms of broken plurals, this will involve correct insertion of the wild card character in the middle of the word. The term در*س will retrieve the singular form of درس (lesson) and its broken plural دروس (lessons). Other more complex broken plural forms (Table 3) have more than one infix, or a prefix and an infix added to the base form, making truncation a challenging task. The plural of مسجون (prisoner) is مساجين; middle truncation involves inserting the wild card betweens س and ج, and between ج and ن in the singular form. In the case of the singular مرض (disease), the plural امراض is formed by adding ا as a prefix and an infix. This poses a new problem, because middle truncation is not enough: the beginning of the word has to be truncated too. This type of truncation is also needed to strip nouns of any particles (Tables 5 and 6) that might be attached to them. In this case, an IR system should have beginning truncation capabilities or be able to identify and isolate these particles at the indexing stage. A search using the truncated term بريد* (mail) would retrieve documents that contain بريد or any of its variants like البريد (the mail), وبريد (and mail), and كالبريد (like the mail). By the same token, if the indexing mechanism can isolate the particles, وبريد, البريد, and كالبريد would be stripped of ال ,و, and كال and indexed under بريد.

Most of the noun-formation rules that may hinder retrieval in Arabic either do not exist in English (infixes) or have minimal effect on retrieval (prefixes). Any attempt to adapt an ELSE to Arabic will have to take into account the similarities and differences between the noun-formation rules of these two languages. Although some morphological rules are shared among languages, attention in an IR environment should be focused on the differences between languages and on ways to accommodated them. Arabic rules differ radically from those of English, and this degree of difference is likely to adversely affect the processing of Arabic nouns in an ELSE. The problem of retrieving English noun variations is not as difficult as that of Arabic nouns. While Arabic nouns can be present with all kinds of affixes, the process of isolating the basic form of a suffixed English noun is relatively straightforward. A simple stemming procedure, for example, is all that is needed to reduce many plural forms to their singular forms. A similar procedure will also isolate the basic form of a feminine noun or a genitive form (through the elimination of -ess, the apostrophe, and -'s).

Theoretically, at least, most of the IR problems in English environments are not related to morphological variations of search terms (they are related to the particularly rich

vocabulary of English, drawn from several linguistic sources, that has produced large numbers of synonyms and homonyms). Stemming has been traditionally implemented to handle word variants, although effectiveness has been debated (Harman, 1991). The negligible effect of prefixes on the retrieval of English nouns, coupled with the absence of infixes in English morphology, have made stemming and truncation stable features of IR systems designed for this language, and they are virtually the only features needed to handle its morphology. The morphology of the language is simple enough to eliminate the need to undertake complex morphological analyses that might be necessary for other languages, a fact that has been illustrated in research on English IR and on IR in other languages.

Background Work

Interest in Arabic IR did not materialize until the 1990s. Before that, specialists in Arabic computing focused their efforts on presenting the language in a computer environment and finding solutions for display and coding problems. In the early 1990s, this changed, and research started to appear on the automation of Arabic online library catalogs and on IR issues. The literature on Arabic in electronic environments includes works ranging from descriptive articles to works relevant to the topic of this paper, such as experimental research on IR systems and on indexing methods.

Hegazi, Ali and Abed (1987) tackled the measurement of redundancy caused by the morphological nature of the Arabic language (compared to English, redundancy in Arabic was assumed to be higher, because Arabic words are derived from roots according to certain patterns, depending on fixed rules, in addition to suffixes, prefixes and infixes). Their study measured the information content per letter and per letter complexes. This kind of measurement can be helpful in many areas, such as information retrieval or text compression. In order to reveal the true characteristics of the Arabic language, full-text documents were used, i.e., full words as they appear in any text with their morphological extensions and not merely their roots. The n-gram technique was applied (the n-gram is defined as a string of n letters occurring frequently in a text, justifying their consideration as symbols by themselves in addition to the symbols that comprise the text). Examples of the full-text documents that were used in the study are books, newspapers, and social magazines. Systematically, studies of the dependencies of characters on each other were done, as well as a study on the average distribution of word lengths. This identified the most and the least frequent characters in any Arabic text. By comparing the results with those from research on English, Arabic was found to have a greater redundancy, and the average word length for Arabic is greater than for English, making Arabic potentially more compressible than English.

Bachir and Baxton (1991) tried to provide a partial answer to the question of whether Arabic periodical article titles can be relied on as a basis for keyword indexing techniques. Another aim of their research was to compare the characteristics of Arabic titles with those of English titles, which according to previous studies have been found sufficiently informative to be used for indexing. They examined the information content of Arabic titles in 16 scientific and non-scientific fields by counting their number of substantive words and comparing the results with those for English periodical articles in the same subject areas. Although significant differences were found between the two samples in some subjects, such as agriculture, philosophy, linguistics, law, and library and information science, Arabic titles generally appear to be as informative as English titles. Where there is a difference, the main problem is that Arabic titles tend to be longer, and contain words that are not indicative of the subject matter. Some practical problems are found in using Arabic titles for indexing, for example, the need to strip prefixes from keywords, and the presence of some words in Roman rather than Arabic script.

The problem of handling Arabic text compression was tackled by al-Fedaghi and al-Sadoun (1990). Their research is concerned with finding a method to reduce the storage space necessary to contain Arabic text in a computer system, in order to decrease the cost of data storage. The morphological compression of Arabic text was thought to be the most effective compression method, replacing some words in the original text by their roots and morphological patterns. In order to examine its effectiveness and measure its reduction ratio, a new combinational method was developed and tested utilizing different texts. The morphological compression was performed in two steps. First, a triliteral root for a compressible word and a morphological pattern were extracted; and second, the compressible words were stored in a three-byte format while the uncompressible words were stored at one character per byte. Large sample data were used to test experimentally this morphological compression scheme. The reduction effect of the morphological property of the language was between 25% and 31.2%, but if the method is used in conjunction with other compression techniques (space elimination from the original text), it is not difficult to achieve reduction ratios of above 40%.

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 Arabic records, and stemming algorithms for Arabic index terms. This work was supplemented by that of al-Kharashi & Evens (1994). The basic goal of the two works was to find the best way to solve the problem of stemming for documents in Arabic. To test the proposed indexing methods, the Micro-AIRS System, a microcomputer system for Arabic information retrieval developed by al-Kharashi, was used. A series of experiments was performed 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. The stem is a combination of a root and derivational morphemes to which one or more affixes can be added. The bibliographic records were extracted from the databank at King Abdulaziz City for Science and Technology in Saudi Arabia. A small word-stem-root dictionary was created and used during the indexing and retrieval process to identify the stem or the root of a given word and also to identify stop words. In order to assess the effectiveness of the three indexing methods, 29 queries were performed against a database of 355 Arabic bibliographic records, covering computer and information science. The results demonstrated the superiority of root/stem-retrieval methods over word-retrieval methods, and underlined the contrast with IR methods in English. Moreover, the root performs as well as or better than the stem at low recall levels and definitely better at high recall levels. This experiment was limited in scope, however, because the collection had short records without abstracts, and the title field alone could be used for information retrieval.

Abu Salem (1992) constructed an experimental Arabic IR system with 120 records (fewer than al-Kharashi) but this time including abstracts. He used the same indexing methods as al-Kharashi (1991) and repeated the latter's experiments. He confirmed the results of al-Kharashi, rating roots as the best indexing terms in Arabic, followed by stems and 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.

Building on the experiments of Abu Salem (1992) and al-Kharashi (1991), Hmeidi, Kanaan and Evens (1997) built a database comprising 242 records, all with abstracts, with the intention of determining the usefulness of automatically indexing Arabic words and investigating the use of roots, stems and full words as index terms. The authors defined automatic indexing as a task performed by a program that would take Arabic text and index every word according to specific rules and guidelines. Traditional measures of recall and precision were applied to searches using 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. 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 ال, a large proportion of Arabic roots will appear only once, making the frequency of index terms (roots) low. As for index terms, this research found that Arabic documents were best indexed by word roots, because 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. As for the effectiveness of searching, the authors argued that roots made better index terms than words or stems, at least when phrases were not involved.

In their work on approaches to improving Arabic IR, al-Jlayl and Frieder (2002) presented two stemming algorithms for Arabic IR systems and investigated the effectiveness of surface-based retrieval (an IR approach based on stemming mechanisms similar to those used for English). They concluded that this approach negatively affected the precision of IR operations because of the high inflection rate in Arabic words. Consequently, they proposed a root-based retrieval algorithm, which performed better than surface-based retrieval but produced extraneous terms that led to the retrieval of document containing word unrelated to query terms. Finally, they employed a light-stemming algorithm that was not as aggressive as the root-based algorithm, and concluded that it significantly outperformed the root-based algorithm.

Most of the research on Arabic language processing and IR has focused on the script, on the linguistic properties of the language in general, or on its morphological structure in particular. Crucial to IR is the treatment of Arabic morphology for indexing and retrieval purposes. Stemming and root indexing have been adopted by researchers as necessary tools for effective IR. Complex linguistic analyses have been conducted to prove this point, but the feasibility of implementing Arabic IR tools in an ELSE has not been discussed. In the experimental Arabic IR systems that have been developed so far, stemming and root indexing have been employed to find word variants, and their effectiveness in IR environments has been measured using recall and precision. This paper, as detailed below, introduces a method to compare stemming and root-based approaches to Arabic IR and to investigate the feasibility of using existing ELSEs for Arabic IR.

Method

1. Introduction

Adapting an ELSE to use with other languages can be a challenging task, not to be lightly undertaken. The morphological properties of a language are the single most important issue that must be tackled in indexing, searching and retrieval. The developers of Arabic IR systems have identified stemming and root indexing as two methods that must be implemented in any system to effectively handle the language (Abu Salem, al-Omari, & Evens, 1999). Stemming is a universal IR technique that is used with different degrees of success to enhance retrieval in any language, while root indexing is a language-specific technique that has been developed for Arabic.

This paper is based on the premise that the process of adapting an ELSE to use with Arabic texts must start at the word level, and specifically with the morphological variants of nouns. This involves providing the system with indexing and searching features that enable users to retrieve the variants of a noun by simply entering that noun or any of its variants as a search term. In English, this is accomplished through stemming. Stemming also has been used along with root indexing in experimental Arabic IR systems to ensure effective

IR (Abu Salem, al-Omari, & Evens, 1999) and by Arabic search engines on the Web. Root indexing requires morphological analysis to identify the Arabic root of words and then group all words that are derived from one root under one index term: the root itself (al-Fedaghi and al-Sadoun, 1990). Logistically, implementing a mechanism to handle these analyses and performing them within a search engine system will likely be more time- and resource-consuming than stemming, which is a common feature of most popular search engines. On the other hand, implementing a stemming mechanism may not be enough to ensure retrieval of word variants in an IR environment. How does root indexing compare with stemming, and which technique is a better choice for Arabic nouns?

2. Search Engines

Two search engines were selected for this research: an Arabic-language engine (al-Idrisi) that employs stemming and root-indexing and whose advanced search features were publicly available at the time of the research (2002), and an ELSE (AltaVista) that employs stemming only. Al-Idrisi is still available through the search field on the Web site of its developer (Sakhr Software), and AltaVista has been supported by Yahoo since 2004. The two engines were used to answer the following questions:

1. How do AltaVista's stemming search features compare with root searching in al-Idrisi?
2. How might the performance of AltaVista be improved, and how can the engine be modified to handle Arabic documents?
3. Does root searching actually outperform stemming?

3. Test Database and Queries

In experimental IR studies, the most common methodological approach involves creating an experimental text collection with known relevant documents, and computing evaluation measures to validate the effectiveness of the strategy (Hull, 1996). The experimental collection of documents (document database) used in this research comprises Web pages (documents) retrieved by initial searches using al-Idrisi to locate nouns extracted from real Web searches.

In traditional IR experiments, queries expressing information needs are selected and matched against documents to measure recall and precision. This research, however, deals with the issue of matching nouns to documents that contain not only those nouns but also other nouns belonging to the same noun blocks. Therefore, the queries were selected to include only one noun and/or its variants: the command issued to the IR system can be conceptualized as follows: find documents containing this noun or any of its variants.

4. Searches

Searches were designed to compare the performance of AltaVista with al-Idrisi's, and to evaluate stemming as an alternative to root-retrieval. The document database was created using the results of root-retrieval by al-Idrisi. The only way to determine if a retrieved document is relevant to a query is to display the document and then check the highlighted terms to see if any belong to that query's noun block. It is important to stress that because the initial objective was to compare the performance of AltaVista, with its stemming capabilities, against that of al-Idrisi and its root capabilities, the AltaVista searches were performed against the document set retrieved by al-Idrisi; at this stage, an assumption was made that all the documents retrieved by al-Idrisi were relevant.

One purpose of the experiment is to explore how well AltaVista performs in terms of document retrieval using its current search features. A second purpose is to establish to

what extent manual manipulation of AltaVista's search features can increase the number of retrieved documents, thereby suggesting ways in which it might be improved for Arabic-language searching. The final purpose is to identify those documents that still have not been retrieved by AltaVista in order to determine whether in fact they should have been retrieved (that is, they contain nouns belonging to the block of the noun in the query) or not (they do not contain nouns belonging to the block of the noun representing the query). Such an examination of the documents not retrieved by AltaVista after all stemming manipulations have been implemented will reveal two critical pieces of information: how many of the documents initially retrieved by al-Idrisi using its root indexing technique have wrongly been missed by AltaVista (that is, the shortcomings of stem in comparison with root searching); and how many of those documents were wrongly retrieved by al-Idrisi in the first place (that is, the shortcomings of root in comparison with stem searching).

5. Recall, Precision and Relevance

Experiments were conducted to investigate how closely AltaVista can approach the performance level attained by al-Idrisi, and to suggest ways of improving the former to make it get even closer. The experiments called for noun queries searched using AltaVista to be matched against documents retrieved earlier by the al-Idrisi. In theory, investigating how close AltaVista can get to al-Idrisi involves counting how many documents it retrieves out of the ones retrieved by al-Idrisi using the root of a specific noun. It was not at all clear at the outset of this research whether the search-by-root feature of al-Idrisi always retrieves relevant documents, but earlier research by others (al-Kharashi, 1991; Abu Salem, 1992; al-Kharashi & Evens, 1994; Hmeidi, Kanaan & Evens, 1997; Abu Salem, al-Omari, & Evens, 1999) had strongly suggested that this indeed was the case. The initial assumption, therefore, was that all documents retrieved by root searching would be relevant to the query containing the noun that retrieves them. The question of whether in fact this is the case was left to the final stages of the methodology.

Relevance is one of the most critical concepts in information. Previous research on Arabic IR (al-Kharashi, 1991; Abu Salem, 1992; al-Kharashi & Evens, 1994; Hmeidi, Kanaan & Evens, 1997; Abu Salem, al-Omari, & Evens, 1999) concluded that root retrieval extracts the highest number of Arabic noun variations and, therefore, produces the maximum possible number of retrieved documents. An ELSE can only match this recall performance if it provides indexing and search capabilities that facilitate the retrieval of an equal number of documents. For example, if a search-by-root query in al-Idrisi retrieves 35 documents, AltaVista also should be able to retrieve those same 35 documents. This assumes, however, that all 35 documents are relevant to the subject encapsulated in the search statement. If this is not the case, then it does not follow that any shortfall by AltaVista represents in fact a criticism of or a failing in the search engine. In practice, then, the question of relevance cannot be ignored; criteria must be in place to judge relevance and to compare the performance of the two systems using this as a measure.

In traditional evaluation studies of IR systems, recall and precision measures are based on the relevance of retrieved documents to the information needs expressed in queries. Determining relevance is not a morphological/linguistic process and should be considered a language-independent exercise: it does not involve looking at the success of a system in retrieving variants of words included in a query; rather, it assesses the extent to which a retrieved document matches the information needs represented by those words. This standard definition of relevance does not apply in this paper, as it is looking at the performance of a system only at the word level, where a retrieved document is examined in order to check if it contains a variant of a query word or not. More specifically, a document is relevant to a query if it contains the noun included in the query or any variant

of that noun. The variants of a noun form a block of nouns; the occurrence of any noun in this block within a document makes it relevant to a query that contains a noun belonging to the same block. A block can include the masculine (m.) noun, the feminine (f.) noun, their dual (d.) and plural (p.) forms, and any forms of these nouns attached to the definite article, to particles, or to possessive pronouns. Given the large number of variants, in addition to the fact that this number can vary from noun to noun, only a very partial listing is shown in Table 7, using the noun معلم (teacher).

Table 7. A partial list of variants in an Arabic noun block

معلم (m.)	معلمه (f.)	معلمان (m. d.)	معلمون (m. p.)
معلمتان (f. d.)	معلمات (f. p.)	معلمي (my teacher m.)	المعلم (the m. teacher)
المعلمه (the f. teacher)	معلمتها (her f. teacher)	ومعلم (and a teacher)	المعلمون (the m. teachers)
المعلمان (the m. d. teachers)	ومعلمات (and f. teachers)	معلمهم (their m. teacher)	وللمعلم (and for the m. teacher)

Recall used in this restricted sense is a measure of the extent to which the IR system retrieves all documents in the database containing a noun or nouns that belong to the noun block of a noun present in a query. Precision is a measure of the extent to which the system only retrieves those documents that contain block nouns, and rejects all others.

6. Arabic Noun Selection

The first step in conducting the experiments was to construct a set of Arabic nouns. This was done by selecting and translating into Arabic 40 nouns from a collection of 907 English nouns entered by real users in real searches conducted on the Web. The nouns were originally obtained in English using SearchSpy, a service provided by Webcrawler. A total of 1891 search queries (predominantly in English) were captured, and 4236 individual English search terms (strings of characters separated by spaces) were extracted from these queries. Of the 4236 terms, 2109 terms were nouns. If a term was a homograph, it was considered a noun ('show', for example, was considered the noun 'show' not the verb 'to show'). These nouns were entered into a Microsoft Access database file. Using the sorting facilities available in Access, 808 duplicate nouns were eliminated, leaving 1301 unique nouns in the set. Further examination of the noun set revealed the presence of 394 proper nouns (countries, cities, people, etc.). Proper nouns were excluded because they do not usually have dual, plural, or feminine forms in Arabic, and typically do not generate morphological variations. Loan-words from other languages were also excluded because they are untypical in Arabic, usually not having an identifiable Arabic root, a fact that makes them fall outside the scope of this research.

After these exclusions, 907 nouns remained in the set. Each noun was numbered for identification purposes. Applying a random selection process, 40 numbers were generated, and the corresponding nouns were selected from the list of 907 nouns. These 40 nouns were translated into Arabic to form the noun data set (Table 8). It comprises the basic forms of Arabic nouns: singular masculine or singular feminine nouns that are not attached to any prefixes or suffixes.

Table 8. Noun data set

Noun	English eq.	Root	Noun	English eq.	Root
وكاله	agency	وكل	بيت	house	بيت
حيوان	animal	حيو	صناعه	industry	صنع
فنان	artist	فنن	معلومه	information	علم

ولاده	birth	ولد	ساكن	inhabitant	سكن
ولد	boy	ولد	معهد	institute	عهد
شركه	company	شرك	بريد	mail	برد
وصل	connection	وصل	وجبه	meal	وجب
متسابق	contestant	سبق	مكتب	office	كتب
تحكم	control	حكم	خيار	option	خير
خلق	creation	خلق	قصيده	poem	قصد
وكيل	dealer	وكل	حمل	pregnancy	حمل
دفاع	defense	دفع	ثمن	price	ثمن
قسم	department	قسم	قراءه	reading	قرء
تنزيل	download	نزل	وصفه	recipe	وصف
بيئه	environment	بوء	نتيجه	result	نتج
نار	fire	نور	خدمه	service	خدم
صديق	friend	صدق	تسوق	shopping	سوق
لعبه	game	لعب	عرض	show	عرض
دليل	guide	دلل	جهه	side	وجه
تاريخ	history	ارخ	جامعه	university	جمع

7. Document Data Set Creation

The next step involved the creation of a document data set that could form a test database for searches using AltaVista. Each of the 40 nouns was entered in al-Idrisi as a single search term using the search-by-root option. The searches on these terms produced hits ranging from 85 to 1046 documents (Table 9). From the results of each of the searches, 50 documents were selected randomly and displayed using the "highlight feature" implemented by al-Idrisi to distinguish words that cause a document to be retrieved. The randomness was achieved using a process similar to the one used to select the 40 nouns. For every search, each retrieved document was numbered for identification purposes. Applying a random selection process, 50 numbers were generated through the random-number-generator function of a calculator, and the corresponding documents were selected from the list of documents retrieved by the search. Because root searching was used, every occurrence of a word that is derived from the root of the noun used as a search term was highlighted. The 50 selected documents from each search were saved in a separate folder on a local computer. For example, the search for خلق (creation) produced 113 documents; 50 documents were randomly selected out of 113 and saved in a folder named "creation" on the local hard drive. As a result of this process, 40 folders (one for each search) were created, each containing the 50 randomly selected documents resulting from the corresponding search. This procedure resulted in 2000 HTML documents that formed the test database.

Table 9. al-Idrisi's search results (number of hits)

Noun	Hits	Noun	Hits	Noun	Hits	Noun	Hits
وكاله	119	وكيل	119	بيت	173	حمل	252
حيوان	176	دفاع	167	صناعه	418	ثمن	146
فنان	169	قسم	230	معلومه	1046	قراءه	382
ولاده	400	تنزيل	163	ساكن	113	وصفه	137
ولد	400	بيئه	120	معهد	190	نتيجه	409
شركه	643	نار	85	بريد	347	خدمه	715
وصل	605	صديق	132	وجبه	128	تسوق	332

متسابق	271	لعبه	112	مكتب	807	عرض	489
تحكم	407	دليل	315	خيار	344	جهه	473
خلق	113	تاريخ	250	قصيده	169	جامعه	829

8. Document Indexing

Similar to Google Desktop, a personal version of AltaVista has been available (although no longer supported) for local use. This version was used in this research, and, for the sake of simplicity, it is referred to hereafter simply as AltaVista. It is a fully functioning version of the main engine with similar indexing and searching features; it was selected because it could be installed and controlled locally. AltaVista was installed on the same personal computer as the test database, and was used to index the 2000 HTML documents contained in the 40 folders. A separate index was built for each folder to allow searching against individual folders (as explained below).

9. AltaVista Searches

In Stage 1 of the searches, each of the 40 Arabic nouns was matched against its corresponding folder in AltaVista's index. First, the nouns were entered in their complete form, exactly as they had been entered earlier using al-Idrisi. In this way AltaVista searched for an exact match of the noun (the column labeled SS (simple searches) in Table 10).

The next stage (Stage 2) was to use the truncation feature available on AltaVista in order to ignore the endings of Arabic nouns that are not part of the root (a manual stemming of the nouns). Each noun was truncated after the occurrence of the third and last letter of the root. For example the noun خيار (from the root خير) was truncated after the letter "ر", the last letter of the root; and the noun وجبه (from the root وجب) was truncated after the letter "ب". When a noun had only three letters, it was truncated after these letters, because this is the minimum number of pre-truncation characters allowed by AltaVista. The column labeled AS (advanced searches) in Table 10 shows the truncated forms of sample nouns as they were entered in AltaVista.

Table 10. Samples of simple and advanced searches in AltaVista

Noun	Root	SS	AS
شركه	شرك	شركه	شرك*
متسابق	سبق	متسابق	متسابق*
صديق	صدق	صديق	صديق*
لعبه	لعب	لعبه	لعب*
دليل	دلل	دليل	دليل*
تاريخ	ارخ	تاريخ	تاريخ*
صناعه	صنع	صناعه	صناع*
معلومه	علم	معلومه	معلوم*
جامعه	جمع	جامعه	جامع*

At the next stage (Stage 3) it was necessary to use AltaVista to retrieve documents using the 40 nouns and after specific prefixes and prefix combinations had been added to these nouns. Because AltaVista does not offer beginning truncation, this stage involved manual modification of the search nouns. The seven most common prefixes/prefix combinations (hereafter referred to as prefixes) were one by one added to the noun. To ensure the retrieval of the noun in its basic form as well as attached to any of these prefixes, each noun was entered in eight forms: in its exact form (as described above), and in the other

seven forms with the seven prefixes attached to it. The column labeled MMS (manually modified searches) in Table 11 shows samples of how these searches were entered. For example, the query of the noun بيئه (environment) contains eight nouns: بيئه, البيئه, والبيئه, للبيئه, بالبيئه, وبيئه, لبيئه, and ببيئه. (Note that when no Boolean operators are used between search terms, AltaVista defaults to OR, and a document is retrieved when it contains any one of the terms). The first noun is the basic form, with no attached prefixes, the remaining seven nouns are forms of the basic noun attached respectively to the prefixes: ال (the), وال (and the), لل (for the), بال (in the), و (and), ل (for), and ب (in).

The fourth and last stage of the searches utilized queries that produced the maximum possible number of documents (the highest recall level) for each of the 40 nouns. These queries were designed to retrieve all documents retrieved by the first three stages, in addition to documents that were retrieved by modifications made to the noun forms used in queries in Stage 3, the stage of manually modified searches (MMS). The noun forms used in the searches in Stage 3 were truncated after the last letter of the root. That meant a query would retrieve documents containing the basic truncated noun or any of its prefixed forms. The column labeled AMMS (advanced manually-modified searches) in Table 11 shows samples of how these searches were entered.

Table 11. Samples of manually modified and advanced manually-modified searches

Noun	MMS	AMMS
وصل	وصل الوصل والوصل للوصل بالوصل ووصل لوصل بوصل	وصل*الوصل*والوصل*للوصل*بالوصل*ووصل*لوصل*بوصل*
وكيل	وكيل الوكيل والوكيل للوكيل بالوكيل ووكيل لوكيل بوكيل	وكيل*الوكيل*والوكيل*للوكيل*بالوكيل*ووكيل*لوكيل*بوكيل*
بيئه	بيئه البيئه والبيئه للبيئه بالبيئه وبيئه لبيئه ببيئه	بيئه*البيئه*والبيئه*للبيئه*بالبيئه*وبيئه*لبيئه*ببيئه*

Upon completion of the four stages of searches, each document that had not been retrieved by AltaVista (a missed document (MD)) was displayed to identify the words that had caused its initial retrieval by al-Idrisi. This was easily accomplished because, as explained above, the documents were saved in "highlighted" formats, where the words that caused their retrieval were highlighted in red. After the highlighted terms were extracted, a database file was created in Access to organize the terms and link them to their respective documents, and consequently to the noun. Let us suppose that after performing all four stages of the searches using the noun نار (fire), 15 MDs were identified. Each one of these MDs is displayed and the highlighted words in it are extracted and entered in a record containing pointers to the document that contains them and to the noun نار. Later, this type of information can be consulted to analyze the causes of retrieval failure and to determine if a document should have been retrieved by AltaVista or, alternatively, if it should not have been retrieved by al-Idrisi in the first place. For example, if an MD has one highlighted term نور (light), which does not belong to the noun block of نار, it is judged irrelevant: it should not have been retrieved by al-Idrisi. By contrast, if an MD contains the highlighted term نيران (fires), which belong to the noun block of نار, it is judged relevant: it ideally should have been retrieved by AltaVista.

Following similar procedures, each document that was retrieved by AltaVista at the last stage of searches was displayed to verify that it was relevant to the noun. Each retrieved document was checked to confirm that it contained at least one highlighted term that belonged to the noun block of the noun that retrieved it, and all retrieved documents were judged relevant to their respective queries.

Results and Analysis

1. The Searches

The search experiments in AltaVista were conducted with the following objectives in mind:

1. To compare the recall of AltaVista with that of al-Idrisi--in document retrieval; that is, to determine how many of the documents originally retrieved by al-Idrisi could also be retrieved by AltaVista.
2. To explore ways of improving the recall achieved by AltaVista in order to identify ways of adapting AltaVista for use with Arabic text.
3. To isolate documents retrieved by al-Idrisi that were not retrieved by AltaVista in order to analyze these documents to see if they are relevant to the nouns used in the searches; that is to determine whether al-Idrisi is retrieving documents that are unrelated to the search noun.

Objective (1) was achieved through the first two stages of the searches (SS and AS) using AltaVista's existing search algorithms. Objective (2) was achieved through the last two stages of the searches (MMS and AMMS) using a manually enhanced AltaVista that involved adding prefixes to the search nouns. This procedure simulated an AltaVista that in effect has beginning-truncation capabilities or automated prefix attachment to Arabic nouns. AMMS by itself provided the highest maximum recall level (closest to that achieved by al-Idrisi through its root searching capability). Therefore, the documents that were not retrieved after this stage were assumed to be missed documents (MDs) pending the analysis of AltaVista's failure to retrieve them. Once the missed documents had been identified (after the four stages of the searches), each was examined, and the word/words (keywords) that caused its retrieval in al-Idrisi were analysed to determine if they belong to a block of the noun used in the search, that is, if they are relevant.

An overview of AltaVista searches is provided in Table 12. For each of the 40 Arabic nouns it shows the results for the four stages, plus the number of missed documents and the failure rate. The AMMS column shows the highest possible number of document that could be retrieved in any of the four stages of searching and, for present purposes, it is assumed to represent the optimal performance of AltaVista. Subtracting the number in this column from the original number of 50 documents in the document set that was retrieved by al-Idrisi produces the number in the last column (MD). For example, if the number in the MD column is 11, this means that the fourth stage of searching retrieved 39 documents out of 50 and failed to retrieve 11 (as in the case of صناعه (industry)). The documents referenced in the MD column are analyzed later to determine why they were not retrieved and if the keywords that retrieved them in al-Idrisi belong to the noun block and are, therefore, relevant.

Table 12 shows that there were many MDs. In total 1120 documents were not retrieved by AltaVista out of the 2000 documents retrieved by al-Idrisi. For some nouns, the number of MDs was almost 100%: وجبه (meal), 49 MDs, and متسابق (contestant), 46 MDs. The lowest MDs are for بيئه (environment) (2 MDs). The rest of the nouns have MDs ranging from 3 to 45, with the largest concentration of numbers in the 20s and 30s. Consequently, the failure rates varied considerably from noun to noun. Of the 40 nouns, 24 experienced a failure rate of 50% or more, with four reaching rates of 90% or higher. Eight of the nouns experienced a failure rate between 40% and 48%, five between 16% and 38%, and only three nouns experienced rates of 10% or less.

Table 12. Number of documents retrieved in the four search stages in AltaVista

Noun	SS	AS	MMS	AMMS	MD	Failure rate
وجبه	0	1	0	1	49	98%
متسابق	0	2	0	4	46	92%

تنزيل	2	3	5	5	45	90%
ثمن	3	3	5	5	45	90%
ولد	5	7	6	7	43	86%
فنان	1	2	3	7	43	86%
قصيده	1	2	3	8	42	84%
تحكم	3	5	6	8	42	84%
دفاع	0	1	8	11	39	78%
وكيل	3	3	3	12	38	76%
معهد	5	5	5	13	37	74%
وصل	8	13	10	15	35	70%
نار	1	2	14	15	35	70%
خيار	7	12	11	16	34	68%
صديق	4	11	7	16	34	68%
حمل	5	13	6	17	33	66%
لعبه	6	10	8	17	33	66%
جامعه	11	14	12	19	31	62%
حيوان	0	9	2	19	31	62%
جهه	7	7	19	19	31	62%
معلومه	8	9	20	21	29	58%
خلق	16	18	22	23	27	54%
وصفه	0	15	0	24	26	52%
خدمه	7	10	16	25	25	50%
نتيجه	20	20	26	26	24	48%
قسم	18	18	24	26	24	48%
ولاده	1	19	21	27	23	46%
مكتب	3	16	16	28	22	44%
تسوق	3	11	4	28	22	44%
وكاله	9	25	11	29	21	42%
قراءه	10	13	18	29	21	42%
عرض	19	23	27	30	20	40%
دليل	15	15	31	31	19	38%
بيت	23	33	29	37	13	26%
صناعه	10	18	19	39	11	22%
بريد	21	24	38	41	9	18%
شركه	32	35	37	42	8	16%
ساكن	17	25	25	45	5	10%
تاريخ	28	32	41	47	3	6%
بيئه	22	30	40	48	2	4%

2. Recall Rates in the Search Stages

The simple searches (SS) conducted in the first stage used the exact form of the noun (the basic noun without prefixes or suffixes). Entering exact nouns in search queries does not seem to be a viable option for effective IR from Arabic databases. Using non-affixed nouns in searching substantially reduces the number of retrieved nouns and therefore adversely affects the number of retrieved documents. In Table 13, the SS column indicates the recall

rate (the percentage found of the original 50 documents retrieved by al-Idrisi). Only two exact nouns retrieved more than 50% of the documents. Four nouns retrieved documents accounting for more than 40% and less than 50% of the documents. Eight nouns retrieved numbers of documents ranging between 20% and 38%, while the remaining 28 nouns retrieved less than 20% of the documents, including five nouns that retrieved no documents at all. Using the simple form of the Arabic noun in an ELSE risks, then, missing many of the noun variants and reduces recall levels.

Truncating an Arabic noun after the third letter of the root slightly improves recall levels. The AS column in Table 13 indicates the AS recall rate. Eleven truncated nouns retrieved between 30% and 48% of their respective documents. The remaining 24 nouns retrieved less than 30% of their documents, with nine of them retrieving less than 10%. The slight improvement is related to the morphological nature of Arabic words. While truncation helps in solving the problem of suffixes, the prefix-rich forms of Arabic nouns cannot be retrieved with end truncation. This works well for English words because of the importance of suffixes compared with prefixes, but for Arabic it is not enough because it does not solve the important problem created by prefixes (which requires beginning truncation).

Table 13. Recall rates

Noun	SS	AS	MMS	AMMS	Noun	SS	AS	MMS	AMMS
بيئه	44%	60%	80%	96%	جامعه	22%	28%	24%	38%
تاريخ	56%	64%	82%	94%	حيوان	0%	18%	4%	38%
ساكن	34%	50%	50%	90%	جهه	14%	14%	38%	38%
شركه	64%	70%	74%	84%	حمل	10%	26%	12%	34%
بريد	42%	48%	76%	82%	لعبه	12%	20%	16%	34%
صناعه	20%	36%	38%	78%	خيار	14%	24%	22%	32%
بيت	46%	66%	58%	74%	صديق	8%	22%	14%	32%
دليل	30%	30%	62%	62%	وصل	16%	26%	20%	30%
عرض	38%	46%	54%	60%	نار	2%	4%	28%	30%
وكاله	18%	50%	22%	58%	معهد	10%	10%	10%	26%
قراءه	20%	26%	36%	58%	وكيل	6%	6%	6%	24%
مكتب	6%	32%	32%	56%	دفاع	0%	2%	16%	22%
تسوق	6%	22%	8%	56%	قصيده	2%	4%	6%	16%
ولاده	2%	38%	42%	54%	تحكم	6%	10%	12%	16%
نتيجه	40%	40%	52%	52%	ولد	10%	14%	12%	14%
قسم	36%	36%	48%	52%	فنان	2%	4%	6%	14%
خدمه	14%	20%	32%	50%	تنزيل	4%	6%	10%	10%
وصفه	0%	30%	0%	48%	ثمن	6%	6%	10%	10%
خلق	32%	36%	44%	46%	متسابق	0%	4%	0%	8%
معلومه	16%	18%	40%	42%	وجبه	0%	2%	0%	2%

The third stage of searching in AltaVista (MMS) involved manually modifying the nouns to include prefixes as part of the search term. This produced eight search terms for each original noun query. The recall levels (Table 13, MMS column) achieved for each of the 40 nouns by manually attaching prefixes to these nouns but not truncating their ends. Attaching the seven prefixes/prefix combinations to two of the nouns improved the recall rate to 82% and 80% respectively. The recall rate of seven queries ranges from 50% to 76%, while a range of 20% to 48% represents the recall rates of 14 queries. The remaining

17 queries retrieved less than 20% of their documents, including three that did not retrieve any documents.

The terms used in the third stage of AltaVista searching were truncated after the occurrence of the last letter of the root. The AMMS technique in the fourth stage produced the highest number of retrieved documents: 880 documents (44% of the 2000 documents found by al-Idrisi). Three truncated and prefixed nouns respectively retrieved 96%, 94% and 90% of the documents (Table 13). Fourteen other nouns retrieved numbers of documents ranging from 50% to 84% of the total. A range of 22% to 46% of documents was retrieved by 15 queries, and the remaining eight queries retrieved less than 20% of their documents. Adding prefixes to the truncated basic noun in the AMMS stage produced the highest rates of recall among the four stages of the searches, because this technique produced the combined results of the previous three stages.

3. The Root Factor

Queries entered in the fourth stage of AltaVista's searches produced the highest number of documents. Any document that was not retrieved by any of the queries on the 40 nouns has been considered a missed document (MD) that should have been retrieved in an Arabic IR environment, using a search-by-root feature. Since al-Idrisi retrieved all the documents that AltaVista failed to retrieve, each one of these documents must be related in one way or another to the noun by the Arabic root. The search-by-root option used in al-Idrisi produced these documents, and so far all 50 documents retrieved by each one of the nouns have been treated as relevant documents based on their containing a derivation of the root of the noun. What has not been considered yet is the validity of the assumption that all these documents should have been retrieved in the first place and, consequently, if the keywords that retrieved these documents are actually related to the original noun.

To investigate the morphological reasons behind the MDs for each noun, the keywords that retrieved each of these documents were extracted and analyzed. The analysis included an assessment of the root factor in the success of the search (how the keyword is related to the original noun). MDs that were retrieved because of the occurrence of a keyword that is not related to the original noun were judged as false hits. Table 14 summarizes the results of the analyses, and tabulates the numbers of MDs and false hits (FHs). The fourth column (AVDs) stands for AltaVista documents and tabulates the number of documents that were missed by AltaVista but ideally should have been retrieved. The values in this column are obtained by subtracting the value of FHs from the value of MDs. For example, there are 19 MDs for دليل (guide), of which 15 are FHs, leaving the number of AVDs that were missed but nonetheless are morphologically related to the search noun as four. The table clearly shows the high number of false hits and, therefore, the adverse effect of root retrieval on precision. In 26 cases, all MDs are false hits, meaning that there are no documents that ideally should have been retrieved by AltaVista. The remaining 14 nouns share among them 74 documents that should have been retrieved by AltaVista, for an average of less than six documents per noun, ranging from a low of one document to a high of 12 documents.

Table 14. AltaVista's performance record

Noun	MDs	FHs	AVDs	Noun	MDs	FHs	AVDs
وكيل	38	26	12	معلومه	29	29	0
لعبه	33	25	8	شركه	8	8	0
بيت	13	5	8	صناعه	11	11	0
معهد	37	30	7	دفاع	39	39	0
قسم	24	18	6	بريد	9	9	0

نتيجه	24	18	6	وجبه	49	49	0
صديق	34	28	6	ولاده	23	23	0
ولد	43	38	5	خيار	34	34	0
نار	35	30	5	قصيده	42	42	0
دليل	19	15	4	ساكن	5	5	0
خلق	27	23	4	حمل	33	33	0
مكتب	22	20	2	ثمن	45	45	0
جهه	31	30	1	قراءه	21	21	0
فنان	43	42	1	وصفه	26	26	0
تنزيل	45	45	0	وكاله	21	21	0
بيئه	2	2	0	خدمه	25	25	0
حيوان	31	31	0	تسوق	22	22	0
تحكم	42	42	0	عرض	20	20	0
متسابق	46	46	0	تاريخ	3	3	0
وصل	35	35	0	جامعه	31	31	0

The cause of failure in AltaVista, and therefore the presence of AVDs in Table 14 are mostly related to keywords that represent the irregular plural forms of nouns. These are usually formed through the addition of infixes and cannot be retrieved through truncation or through manual attachment of prefixes (but can be retrieved by a root search). In addition, a character called كشيده presented by a long underscore "__" prevented the retrieval of some documents. The use of this character is a peculiar aspect of presenting Arabic words in electronic format; it is used between two characters for the sole purpose of lengthening the distance between them, making the word more visually appealing. In two cases, the cause of failure is the presence of a prefix or a prefix combination that was not included in the prefixes/prefix combinations that were added to the nouns in the third and fourth stages of the searches. The prefix ك and the prefix combination كال occur in two documents that were not retrieved by AltaVista.

Table 15 shows a breakdown of the numbers of AVDs among the 14 nouns that produced them (as shown in Table 14), and it relates them to the three causes of AltaVista's failure mentioned above. Irregular forms of the plural of nouns caused by far the highest number of failures, accounting for a total of 60 from 12 of the 14 nouns. Second came the كشيده, which caused 13 failures in four nouns. Prefixes affected only two nouns, with a total number of two failures.

Table 15. Causes of failure in AltaVista

Noun	Irregular plural	كشيده	Prefix/prefix combination	Total
وكيل	11		1	12
لعبه	8			8
بيت	4	4		8
معهد	7			7
قسم	6			6
نتيجه	6			6
صديق	5	1		6
ولد	5			5
نار	1	4		5
دليل	4			4

خلق		4		4
مكتب	2			2
جهه	1			1
فنان			1	1

Conclusion and future research

The first two stages of searching in AltaVista - using only the original noun, and then the noun plus end truncation - are easy to implement on a typical ELSE, but showed how the engine as a consequence produced low recall levels, missing a high number of documents. The performance of the engine in these two stages was affected by the absence of beginning truncation that allows truncation at the beginning of an Arabic noun, and therefore can take account of the presence of prefixes in these nouns. Once the prefixes were added to the search terms in the last two stages of the searches, the recall levels increased dramatically. In this experimental environment, it can be concluded that the biggest obstacle facing effective retrieval in Arabic is the occurrence of prefixes. These are very commonly used with Arabic nouns, and it is extremely important that an ELSE should be able to accommodate them. Unfortunately, the manual addition of such prefixes is both very time consuming and prone to spelling mistakes at the query input stage; it also requires a very good knowledge of Arabic morphology.

While the manual additions of prefixes enabled AltaVista to handle the problem of prefixes, other problems arose because of the presence of morphological variants in the Arabic nouns. Many of the plural forms of Arabic nouns are irregular, which are usually formed by the addition of infixes to the stem forming the basic noun. Theoretically, this can be handled with middle truncation, but the user has to be well versed in the language to know where to place the truncation symbol.

The last retrieval problem identified in the search experiments is the occurrence of the special character (كشيده), which is indexed by AltaVista as a separate character. If this character is present in an Arabic noun, that noun cannot be retrieved unless the character is entered. The user must know the position of this character in the noun and enter it accordingly.

AltaVista's lack of beginning truncation constituted a major drawback in using it as a search engine for Arabic retrieval. One solution to the prefix problem is offered in existing Arabic IR systems, including al-Idrisi. Advanced stemming is applied to the words to strip them of prefixes and suffixes. This is accomplished through the implementation of algorithms that isolate the prefixes and suffixes and allow the entry of index terms under the stemmed noun (the noun stripped of prefixes and suffixes). In an ELSE, it might be difficult to implement such algorithms if the system does not recognize the language being indexed. The ELSE has to have a mechanism by which it identifies the Arabic words at the indexing stage and applies the prefix and suffix-stripping algorithms to them. Otherwise, this system will not know when to apply the algorithms and when to ignore them.

An alternative to automatic stripping at the indexing stage is automatic inclusion of prefixes at the search stage. In these search experiments, seven prefixes/prefix combinations were used to improve recall levels. The ELSE can be modified to automate what was done here manually, but it must have a mechanism to identify the word being entered in a search as an Arabic word. Once the word is identified, the system must then include the word in its original form in addition to the other seven forms, thereby generating a query that would retrieve documents containing any of the eight noun forms.

The irregular plurals of Arabic nouns presented the most challenging problem for retrieval using AltaVista. Most ELSEs provide automatic stemming of regular English plural forms. Irregular plural nouns are not as common in English as they are in Arabic. Retrieval by the root of the noun can solve this problem in an Arabic IR system because the singular and plural forms share the same root: both forms are retrieved when either one of them is entered as a search term. In an ELSE, a possible solution for this problem could be the inclusion of a list of irregular plural forms along with their singular forms in the indexing algorithms. At the indexing stage, whenever a document containing either form is countered, it is indexed in a way that allows its retrieval no matter which one of the forms is entered in a query. This is analogous in English to including the singular noun "tooth" plus its corresponding irregular plural form "teeth" as a linked pair in the index. A document including the noun "teeth" would then be retrieved whenever a query contains either the nouns "tooth" or "teeth".

For obvious reasons, AltaVista failed to retrieve nouns that contained the كشيده between their characters; these nouns could have been retrieved only if the exact position of this character had been known. But there is no way for the user to know this; the use of this character is arbitrary, and even if it exists in a noun in one document, it may not exist in the same noun in another document. The best way for an ELSE to deal with the كشيده is to ignore it altogether at the indexing stage. AltaVista and other search engines do ignore certain special characters in indexing; modifications can be made to ignore the كشيده as well.

Previous research on Arabic IR mainly has compared root retrieval with stemming as indexing methods for effective IR. Advocating the use of root retrieval, this research was conducted on experimental IR systems designed specifically for Arabic and including stemming and root indexing capabilities, but it did not discuss the feasibility of modifying existing ELSEs for use with Arabic. The results presented in this paper strongly suggest that in order to adapt an ELSE to operate effectively with Arabic, stemming in fact is a better approach than root retrieval. When performed together, truncating the ends of words and adding prefixes to them have an effect equivalent to the advanced stemming of an Arabic noun (stripping it of both prefixes and suffixes). When the documents that had not been retrieved by AltaVista after both suffix and prefix stemming had been applied were judged for morphological relevance, a majority were found to be irrelevant. In other words, the root retrieval capabilities available on al-Idrisi were generating many irrelevant hits. This finding strongly supports the case for stemming rather than root retrieval as an effective means to retrieve Arabic documents.

Future research plans related to the findings of this paper include an investigation of their implications when applied to actual queries that express actual information needs of real users. The investigation will be based on a study of genuine user queries in a system that implements the improved search/indexing techniques suggested by the findings, employing traditional measures of recall and precision to evaluate the effectiveness of the system. In another future study, it is planned to evaluate the performance of Morfix, a publicly available search engine that combines the features of a typical ELSE and those of a specialized Arabic one, in addition to cross-language IR capabilities. The evaluation will contribute to an understanding of the emerging problems associated with IR in different languages, and will advance ideas on the investigation of language-dependent aspects of IR.

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.

- Abu Salem, H., al-Omari, M., & Evens, M. (1999). Stemming methodologies over individual query words for an Arabic information retrieval system. *Journal of the American Society for Information Science*, 50(6), 524-529.
- Bachir, I. & Buxton, A. (1993). The use of topic sentences for evaluating the representativeness of Arabic article titles. *Journal of Information Science*, 19(6), 455 65.
- Beesley, K. (1996). Arabic finite-state morphological analysis and generation. In *COLING-96 Proceedings*, Volume 1. Copenhagen: Center for Sprogteknologi, the 16th International Conference on Computational Linguistics, 89-94.
- Cowan, D. (1958). *An introduction to modern literary Arabic*. Cambridge: Cambridge University Press.
- Crystal, D. (1985). *A dictionary of linguistics and phonetics* (2nd edition updated and enlarged). Oxford: Blackwell/AndrÃ© Deutsch.
- De Guzman, V. & O'Grady, W. (1987). Morphology: the study of word structure. In O'Grady, W. and M. Dobrovolsky (eds.) *Contemporary linguistic analysis*. Toronto: Copp Clark Pitman Ltd, 127-155.
- DeYoung, T. (1999). Arabic language history. Retrieved March 28, 2005, from http://www.indiana.edu/~arabic/arabic_history.htm
- al-Fedaghi, S. & al-Sadoun, H. (1990). Morphological compression of Arabic text. *Information Processing & Management*, 26(2), 303 316.
- Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 7-15.
- Harter, S. (1986). *Online information retrieval: concepts, principles, and techniques*. Orlando: Academic press, Inc.
- Haywood, J. (1960). *Arabic lexicography*. Leiden: E. J. Brill.
- Hegazi, M. & Elsharkawi, A. (1985). An approach to a computerized lexical analyzer of natural Arabic. *Computer processing of the Arabic Language, Workshop Papers*, Vol. 1. Kuwait: Kuwait Institute for Scientific Research (KISR).
- Hegazi, N., Ali, N., & Abed, E. (1987). Information content in textual data: Revisited for Arabic text. *Journal of the American Society for Information Science*, 38(2), 133 137.
- Hmeidi, I., Kanaan, G., & Evens, M. (1997). Design and implementation of automatic indexing for information retrieval with Arabic documents. *Journal of the American Society for Information Science*, 48(10), 867 881.
- Hull, D. (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 70 84.
- al-Jlayl, M., & Frieder, O. (2002). On Arabic search: Improving the retrieval effectiveness via light stemming approach. *Proceedings of the 11th ACM International Conference on Information and Knowledge Management, Illinois Institute of Technology*. New York: ACM Press, 340-347.
- al-Kharashi, I. (1991). *Micro-Airs: Microcomputer based Arabic information retrieval system, comparing words, stem, and roots as index terms*. Unpublished doctoral dissertation, Computer Science department, Illinois Institute of Technology, Chicago.
- al-Kharashi, I. & Evens, M. (1994). Comparing words, stems, and roots as index terms in an Arabic information retrieval system. *Journal of the American Society for Information Science*, 45(8), 548 560.
- Khurshid, Z. (1997). Arabic online catalog. *Information Technology and Libraries*, 11(3), 244 51.
- Matthews, P. (1974). *Morphology*. London: Cambridge University Press.
- Mehdi, S. (1986). Arabic language parser. *International Journal of Man-Machine Studies*, 25, 593-611.

- Moukdad, H. (1999). An investigation of the necessity of information retrieval algorithms for full-text Arabic databases. *Information Science: Where has it Been, Where is it Going*? *Proceedings of the 27th Annual Conference of the Canadian Association for Information Science*, UniversitÃ© de Sherbrooke, June 1999. [Toronto]: CAIS, 207-227.
- Moutaouakil, A. (1987). Lexical derivation in Arabic: roots and patterns. In Descout, R. (ed.) *Applied Arabic linguistics and signal & information processing*. Washington: Hemisphere Pub. Corp., 93-97.
- Murtonen, A. (1964). *Broken plurals: origin and development of the system*. Leiden: E. J. Brill.
- Rafea, A. & Shaalan, F. (1993). Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network. *Software-Practice and Experience*, 23(6), 567-588.
- van Rijsbergen, C. (1979). *Information Retrieval, Second Edition*. London: Butter Worths.
- Yahya, A. (1989). On the complexity of the initial stage of Arabic text processing. Paper presented at the *First Great Lakes Computer Conference*, Kalamazoo, MI.
- Ziadeh, F. & Winder, R. (1957). *An introduction to modern Arabic*. Princeton: Princeton University Press.

Bibliographic information of this paper for citing:

Moukdad, H. (2006). "Stemming and root-based approaches to the retrieval of Arabic documents on the Web." *Webology*, **3**(1), Article 22. Available at: http://www.webology.org/2006/v3n1/a22.html

---

This article has been cited by other articles.
