Word Stemming for Arabic: The Case for Simple Light Stemming

Suleiman H. Mustafa
Dept. of Comp. Info. Systems,
Faculty of Information Technology,
Yarmouk University, Irbid-Jordan,
e-mail: smustafa@yu.edu.jo

Abstract
Although a number of attempts have been made to develop a stemming formalism for the Arabic language, most of these attempts have focused merely on the lexical structure of words as modeled by the Arabic grammatical and morphological lexical rules. This paper discusses the merits of light stemming for Arabic data and presents a simple light stemming strategy that has been developed on the basis of an analysis of actual occurrence of suffixes and prefixes in real texts. The performance of this stemming strategy has been compared with that of a heavier stemming strategy that takes into consideration most grammatical prefixes and suffixes. The results indicate that only a few of the prefixes and suffixes have an impact on the correctness of stems generated. Light stemming has exhibited superior performance than heavy stemming in terms of over-stemming and under-stemming measures. It has been shown that the two stemming strategies are significantly different in retrieval performance.

Keywords
Word Stemming. Light Stemming. Heavy Stemming. Arabic. Information Retrieval. Morphological Analysis.

Introduction

Stemming for information retrieval (IR) is a computational process by which we remove potential suffixes and prefixes from a textual word to extract its basic form. The basic form produced does not have to be the actual word itself. Instead, the stem is said to be the least common denominator for the morphological variants (Carlberger, Dalianis, Hassel, & Knutsson, 2001). This process should not be confused with the process of “morphological analysis” (or word “lematization”, as called by linguists) which aims at reducing morphological variants to a linguistically correct root morpheme from which they were derived.

In IR, the notion of “correct stem” is not of direct relevance. The aim of computational stemming is to ensure that any two morphologically related words, which refer to the same concept, should be reduced to the same form – however “unnatural” that might be (Paice, 1996). Hence, IR-oriented stemmers are not usually judged on the basis of linguistic correctness, though the stems they produce are usually very similar to root morphemes (Frakes, 1992).

The importance of word stemming for information retrieval and computational linguistics was recognized a long time ago. As pointed out by (Lennon et al., 1981), the notion is thought to be useful for two reasons. Firstly, it reduces the total number of distinct terms present with a consequent reduction in dictionary size and updating problems. Secondly, similar words generally have similar meanings and thus retrieval effectiveness may be increased. From an application perspective, stemming has been seen useful in two ways (Khoja & Garside, 1999). In the first, roots extracted can be used in text compression, text searching, spell checking, dictionary lookup, and text analysis. In the second, affixes recognized can be used in determining the grammatical structure of the word, which is important to linguists.

The effect of term stemming on the performance effectiveness of information retrieval has been the subject of several investigations. Most notably of these investigations are those reported by (Lennon et al., 1981; Fuller & Zobel, 1998; Paice, 1994, 1996; Hull, 1996). The general indication coming out of most studies is that stemming can improve retrieval performance, but by a small factor. And it has also been considered to improve recall more than precision (Kraaij & Pohlmann, 1996).

However, it should be noted that inconsistent results were reported in some cases. Either stemming did not show any consistent average performance improvement (Harman, 1991) or the performance increased by a factor which ranged between 15% and 35% (Krovertz, 1993). This should be compared to the average absolute improvement reported by (Hull, 1996) which ranged from 1-3%. This inconsistency could be attributed to variations in the length of documents used in the retrieval experiments. It seems that the smaller the size of documents the greater the improvement realized in performance due to stemming.

Variation in the results of stemming effectiveness also exists across languages. Popovic & Wilett (1992) showed that stemming on Slavic document abstracts increased precision in information retrieval with 40%. They concluded that stemming should be particularly effective for languages with more complex morphology. This conclusion was re-emphasized later by Pirkola (2001) and Larkey et al. (2002).

Working on the assumption that Arabic is a complex inflectional language, Larkey et al. (2002) have demonstrated that stemming has a large effect on Arabic information retrieval due (at least in part) to the inflected nature of the language. Their results indicated an average improvement in precision performance of about 100% due to stemming. For thesaurus-based cross-lingual retrieval, the results showed even larger effect on retrieval. This seems to be inconsistent with the results reported by Xu et al. (2002) who used the same corpus (i.e., the TREC 2001 data) and found that stemming had little impact on cross-lingual retrieval.

A number of research studies (Al-Khrashi,1994; Abu-Salem & Al-Omari, 1995; Hmeidi, 1995; Al-Tayyar & Bechkoum, 1998) have focused on the impact of the level of word stemming on Arabic information retrieval. Basically, they have examined three different levels including word-based retrieval, stem-based retrieval, and root-based retrieval. But, no underlying stemming algorithms have been reported due to the fact that many of these studies have used manual stemming techniques to create index terms. The results of all these studies indicate that root-based retrieval provides the highest level of performance, followed by stem-based retrieval and finally word-based retrieval.

Hence, it comes no coincidence that much of the efforts at developing stemming techniques, such as those reported by Al-Fedaghi & Al-Anzi (1989), Beesley (1996), Al-shalabi (1998), Khoja (1999), Mustafa & Masoud (2000), and Roeck & Alfares (2000), have been root-driven. Typically, in root-based stemming algorithms, root candidates are checked against a root lexicon. If no match is found, affixes and patterns are readjusted and the new candidate is checked. The process is repeated until a root is found (De Roeck & Al-Fares, 2000).

This three-tier view of Arabic IR method has emerged from the classical morphological and grammatical rules of how Arabic words can be formed within lexical and textual contexts. However, as we will see later in this paper, this view suffers from a number of drawbacks. In the present study, an attempt is made to present the case for using light stems and propose a simple light stemming technique which has been based on the characteristics of Arabic prefixes and suffixes as they occur in real texts. Some of these affixes are heavily used while many others are rarely encountered in any type of text.

Related Work

Light stemming refers to a process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognize patterns and find roots (Larkey, Ballesteros, & Connell, 2002). Other terms, such as "elementary" stemming (Harman, 1991) or "shallow" stemming (Monz & Rijke, 1991), are used sometimes to convey the same meaning. The notion of light stemming was used early in what was described by Harman (1991) as an “S” stemming algorithm, in which only a few common word endings were removed: “ies”, “es”, and “s” (with certain exceptions).

As the word “light” suggests, the term is used to indicate the opposite of heavy stemming in which the whole set of possible prefixes and suffixes are removed. Each of these two strategies has its own strengths and weaknesses. A light stemmer plays safe in order to avoid over-stemming errors, but consequently leaves many under-stemming errors. A heavy stemmer, on the other hand, boldly removes all sorts of endings, some of which are decidedly unsafe, and therefore commits many over-stemming errors (Paice, 1994).

Algorithmic light stemmers which remove Arabic affixes (prefixes, infixes, and suffixes), at various levels of stemming, have been reported by a number of authors. But, in some of these studies (Aljlayl et al., 2001, DeRoeck & Al-Fares, 2000), we find no indication of the type of algorithms or heuristics being applied or the affixes being removed. In the other studies (Larkey, Ballesteros, & Connell, 2002; Darwish, 2003), where lists of affixes are explicitly given, the affixes being stripped off seem to have been selected on the basis of authors' intuition and knowledge of Arabic.

De Roeck & Al-Fares (2000) found empirically that light stemming gave better results than heavy stemming. They pointed out that heavy stemming brought the risk of root consonant loss. The word “*t'amyn*” (insurance), for instance, which comes from the ground root “*amn*” (secured) is stemmed by a heavy stemmer into: “*t'am*” [1]. According to the authors, the same word will be treated by light stemming as “*t'amn*”, after removing the vowel “Yaa”.

(Larkey, Ballesteros, & Connell, 2002) developed several light stemmers for Arabic which remove a small number of prefixes and suffixes and a co-occurrence based statistical stemmer which creates large stem classes by vowel removal and then refines these classes using co-occurrence. The set of affixes removed included six prefixes, four vowels and eleven suffixes[2]. Besides, some normalization was carried out which involved unifying the letters: “alif” to “alif without hamzah”, “alif-maqsurah” to “yaa”, and “taa-marbootah” to “haa”. The best stemmer was a light one that removed stop words, definite articles, and the letter “waw” from the beginning of words, and a small number of suffixes from the ends of words. The authors pointed out that, although light stemming improves performance, it fails to conflate a number of forms that should go together.

Improvement in performance, due to light stemming, was also reported by Darwish (2003). He built a light stemmer (called Al-Stem) in which only a small list of prefixes and suffixes were considered[3] based on some probability threshold and personal judgment. Using mean interpolated average precision as a measure of retrieval effectiveness, index terms based on lightly stemmed words statistically significantly outperformed those based on words and roots.

A Rationale for Light Stemming

Word stemming has been based on a general assumption of semantic equivalence. In most cases, morphological variants have similar semantic interpretations and can be treated as equivalent for information retrieval applications As (Hull, 1996). However, it may be objected that natural words do not fall into entirely clear-cut semantic classes. In the first place, pairs of etymologically related words sometimes differ sharply in meaning. In the second place, some affixes may alter the meaning of a word so greatly that to remove them would be to discard vital information (Paice, 1994).

---

[1] The full word is (تأمين) and the stem is (تأم) after striping the letters (ين) as a potential suffix.

[2] The list included the prefixes (ال، وال، بال، كال، فال، و), the suffixes (ه، ها، ـة، ة، ات، ان، ون، ين، يه، ية، ي) and the infixes (ا، و، ي، ء).

[3] The final list included the prefixes (وال، متـ، وتـ، ستـ، نتـ، بمـ، لمـ، ومـ، كمـ، فمـ، الـ، للـ، ويـ، ليـ، فيـ، وا، فا، لا، با، لتـ، يتـ، بتـ، بالـ، فالـ) and the suffixes (ـات، وا، وه، ـان، تي، ته، تم، كم، هم، هن، ها، ية، تك، نا، ين، يه، ـة، ـه، ي، ا).

Speaking of Arabic, the semantic equivalence issue is further complicated by the fact that words follow the model represented in Figure 1, in which words are formed according to a three-level morphological structure: ground roots, morphological stems, and full textual words. We can view a word as derived by first adding morphological affixes, which conform to a given pattern, to a ground root to generate a stem and then attaching grammatical prefixes and suffixes to the stem to generate the full textual word[1].

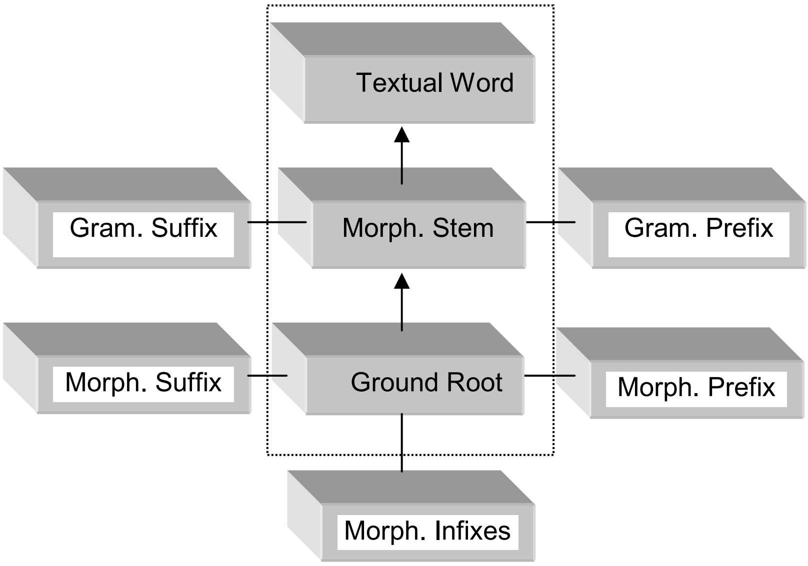

Figure (1): The morphological structure of Arabic textual words (Note that the diagram should be viewed from right to left and Gram. stands for Grammatical and Morph. for Morphological)

Given this structure and the associated lexical and syntactic rules of forming textual words, a given word can take a huge number of morphological variants in textual contexts. In some cases, this might get close to the theoretical maximum length in words such as “wabil-istiqlal-ieh” (with independence)[2], which is composed of thirteen letters. However, this is not the usual case. In reality, none of the Arabic derived words can assume the theoretical maximum length of textual words.

The average length of Arabic words in a normal text does not usually exceed six letters (Mustafa, in press). This comes as a consequence of the fact that, a large number of words appearing in a natural Arabic text do not involve any grammatical prefixes or suffixes. Table 1 shows the distribution of such affixes in two samples of text. The first represents a set of document titles, while the other comes from a narrative text.

Table (1): Prefixed and suffixed words in two samples (figures refer to distinct words)

	Sample 1		Sample 2	
	Num	%	Num	%
Prefixed only	3820	58.9	544	40.0
Suffixed only	341	05.3	157	11.6
With prfx+sufx	298	04.6	95	07.0
None	2022	31.2	563	41.4
Total	**6481**	**100.0**	**1359**	**100.0**

[1] In both types, the number of affixes added can be zero.

[2] The word ““وبالاستقلالية is composed of three grammatical prefixes (4 letters), a morphological prefix (3 letters), an infix (1 letter) , two grammatical suffixes (2 letters), and a ground root (3 letters).

Further analysis of the figures presented in this table shows that only a small number of the grammatical prefixes and suffixes are frequently used. As Figure 2 shows, out of the total number of prefixed words in each of the two samples, about 60% or more of these words start with either the prefix "*al*" (the definite article) or the prefix "*waw*" (a conjunction)[1]. This suggests that the general practice of removing all candidate prefixes and suffixes does not seem to be based on a reasonable rationale. Once prefixes and suffixes with high probability of occurrence in normal texts are removed, no more significant overall improvement is expected to be realized. This provides a strong argument in favor of light stemming.

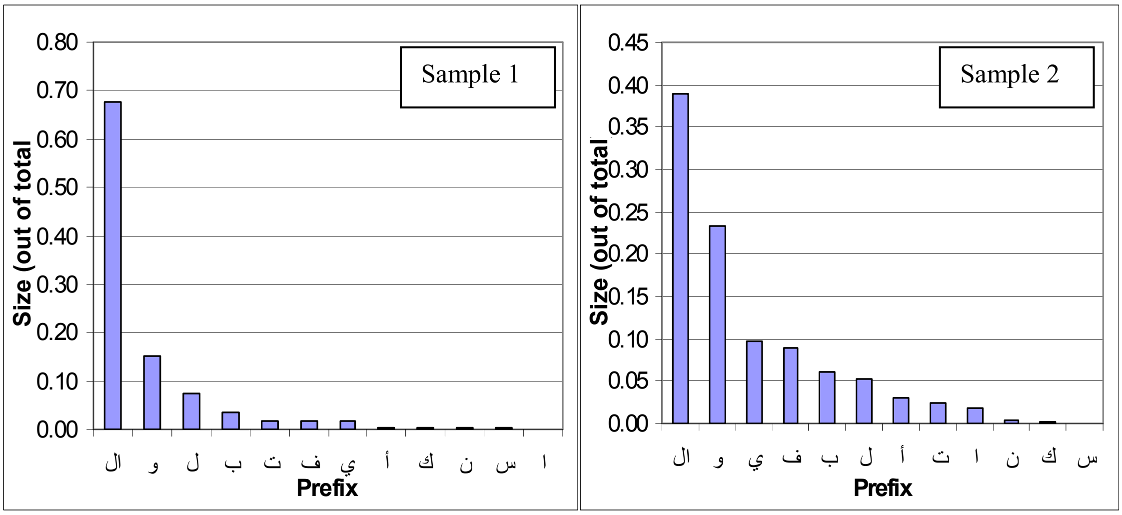

Figure (2): Distribution of prefixed words according to the first prefix (Sample1: Total = 18550 prefixed words, with all occurrences, Sample2: Total = 1020 prefixed words, with all occurrences)

More support for the case of light stemming for Arabic can also be found in the distribution of textual words according to the average number of compound prefixes and compound suffixes. As Figure 3 points out, only a small fraction of words usually involve compound prefixes (2 or 3 prefixes). More than 80% of the words included in each of the two samples either involve no prefix at all or have one single grammatical prefix. The case is more evident when we consider the occurrence of suffixes as shown in Figure 4. Only a small percentage of words involve a single suffix, while it is almost negligible in the case of compound suffixes (i.e., two or three suffixes combined)[2].

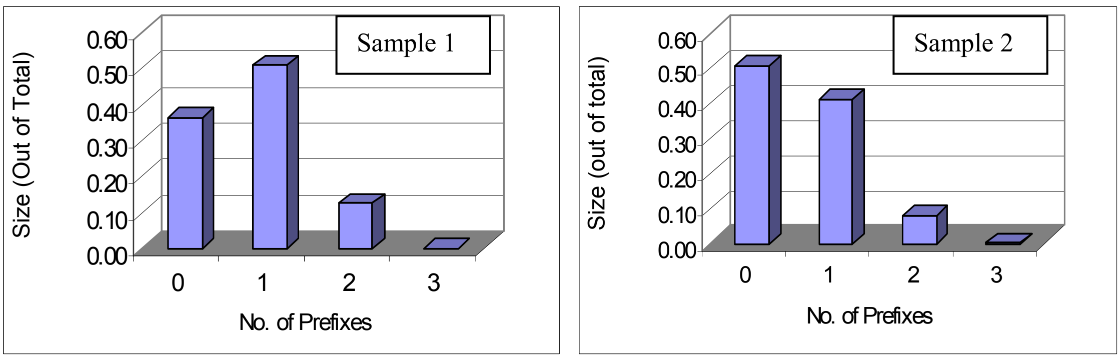

Figure (3): Distribution of distinct words according to the number of prefixes, where Total (sample 1) = 6481, (sample 2) = 1359

[1] The list of grammatical prefixes includes 93 combinations. Details of the rate of occurrence of prefixes are given in Table A1 and Table A2 in the appendix.

[2] The list of suffixes includes 50 combinations. For semantic reasons, some of the suffixes (including "‘‘ات, "‘‘ة, "‘‘ى, and "‘‘ية) were not considered. Details of the rate of occurrence of suffixes are given in Table A3 in the appendix

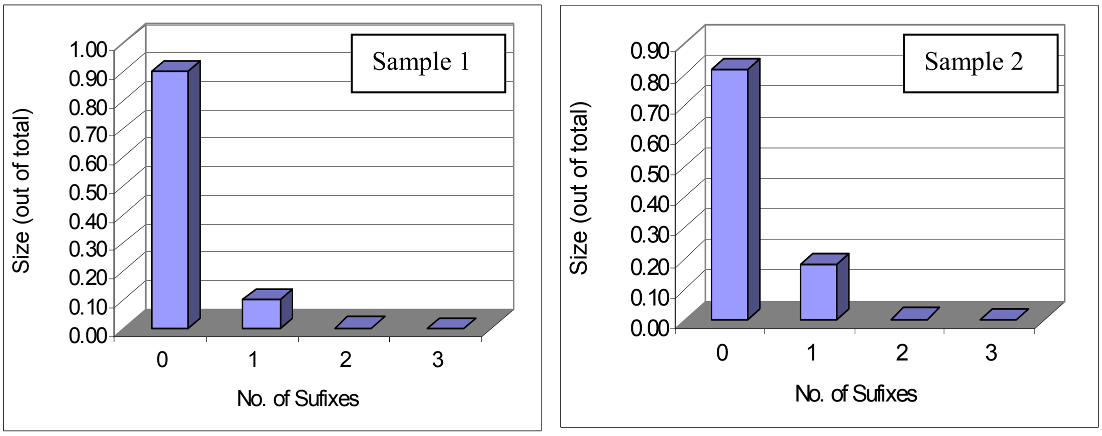

Figure (4): Distribution of distinct words according to the number of suffixes, where Total (sample 1) = 6481, (sample 2) = 1359

Given this lexical reality and the support it provides for light stemming, further support is also evident in the semantic reality. The semantic equivalence of terms must be viewed according to the information content to be conveyed by conflated terms. Most of the work in word stemming for Arabic has relied on the assumption that words sharing a root are semantically related (Hmeidi, Kanaan & Evens, 1997). This is justified on the grounds that Arabic is a derivative language (Ali, 1988; Al-Fedaghi & Al-Anzi, 1989).

A typical Arabic word contains a trilateral or quadrilateral root which involves the basic essence. The role of affixes added to it is to qualify this essence by modifying its lexical and/or syntactic role to represent various inflection aspects such as case, gender, number, tense, person, mood, or voice. The purpose of stemming is to make it possible for a user to retrieve morphologically related terms which may have a semantic relationship (Al-Tayyar, 1998).

However, it may be objected that the root of the word provides the best strategy for Arabic information retrieval. It is true that, recall performance is improved, as we move from the textual-word level down to the root level, but this is accompanied by a corresponding decrease in the precision performance. Searching based on full textual words offers the highest level of precision, since it relies on exact matching. As we start removing letters from a given word, some information is being lost from the semantic content of the word. By the time we arrive at the root, we have reached the lowest level of semantic content.

How much of the basic essence provided by a given root is carried to the various words derived from it is also subject to question. It can be easily argued that words sharing the same root do not necessarily convey the same semantic content. A typical example is when a root-based stemming procedure conflates all words derived the ground root "JM3"[1] under one basic form (which is the root in this case).

When this basic form is used in the searching process for retrieving information items related to any word derived from "JM3", many of the items retrieved will have very little, if any, semantic equivalence. Table 2 lists some of these words and the different meanings they can convey. Consider, for instance the word "jami3ah" (university). It might be said that the information conveyed by this word cannot be considered equivalent to the information conveyed by other words in the table such as "jam3iah" (association) or "jami3" (mosque).

[1] The trilateral root (جمع : put together), pronounced as "jama3a".

Table (2): words derived from the ground root (جمع “JM3”)

Word	meaning	word	meaning
جمع	crowd	جمعية	association
جماعة	group	جامعة	university
جماع	mating	مجتمع	society
مجموع	sum	اجتماع	meeting
جامع	mosque	جمعة	Friday

A Light Stemming Procedure

Given the lexical and semantic realities pointed above, a simple light stemming procedure was developed. The procedure considers only a small subset of the grammatical prefixes and suffixes, which have been found to occur in normal texts more frequently than others. The list of prefixes and suffixes includes the following:

Prefixes: (أ، ال، بال، ت، ست، فت، في، ل، لل، و، وال، وبال، ولل، ون، وي، ي).
Suffixes: (ا، ت، كم، نا، ه، ها، هم، هما، وا).

Since infixes are integral parts of the morphological forms (known in Arabic as “Awzan”) by which stems are formulated, they are treated as such and no attempt has been made to remove any of them in the procedure.

The light stemming procedure accepts a single Arabic word *W* which is tokenized from a normal text *T*. It works by first checking if W starts with any of the prefixes listed above. It does so by examining the first letter of W as follows:

If *W*[1] in [أ، ا، ب، ت، س، ف، ل، و، ي] then find_prefix(W)

If the result is true, the procedure continues looking for the rest of letters making up a given prefix. For efficiency reasons, the procedure uses binary search for accessing the list of prefixes. The presence of a suffix in W is also determined by the same technique, except that the checking is performed backward. The procedure starts by examining the last letter (with *n* denoting its position) as follows:

If *W*[*n*] in [ا، ت، ك، م، ه] then find_suffix(W)

Once a prefix or a suffix (if any) is determined, it is removed from the tokenized word W and the resulting stem is reported. A stem is considered valid if its length is greater than two letters, otherwise W is treated as the stem. If the last letter in the stem is hamzated-waw “ؤ” or leaned hamzah “ئـ” , the letter is converted into single-hamzah form “ء”.

Testing the Light Stemmer

There are several criteria for judging stemmers: correctness, retrieval effectiveness, and compression performance (Frakes, 1992). Of these three criteria the first has been chosen to test the proposed light stemmer. Correctness has been measured using two commonly known parameters: over-stemming and under-stemming. Each provides an indication of some erroneous stemming judgment. When too much of a word is removed, it is likely that the stemmer will conflate unrelated terms, thus leading to retrieving non-relevant information items. When, on the other hand, too little of a word is removed, it is likely that the stemmer will fail to conflate related forms that should be grouped together, thus preventing related items of information from being retrieved.

Using these two parameters, the performance of the proposed light stemming procedure was compared to the performance of a heavy stemming strategy, whereby almost all grammatical prefixes and suffixes were removed. The testing was carried out using a set of Arabic textual data containing a total of 29988 words, distributed over 6481 distinct textual words. Of these words, about 31.2% did not involve any prefixes or suffixes.

To provide a basis for empirical analysis and assessment, all words were stemmed and analyzed manually. A distinction was made between four categories of words: prefixed only, suffixed only, prefixed and suffixed, and non-affixed words.

Each of the two stemming strategies was run twice on the given set of data: once with removing stop words and the other without handling stop words. The set of stop words was not intended to be exhaustive. It consisted of only 342 various forms of particles, pronouns, and adverbs. Figure (5) shows the size distribution of stems generated by the two stemming strategies.

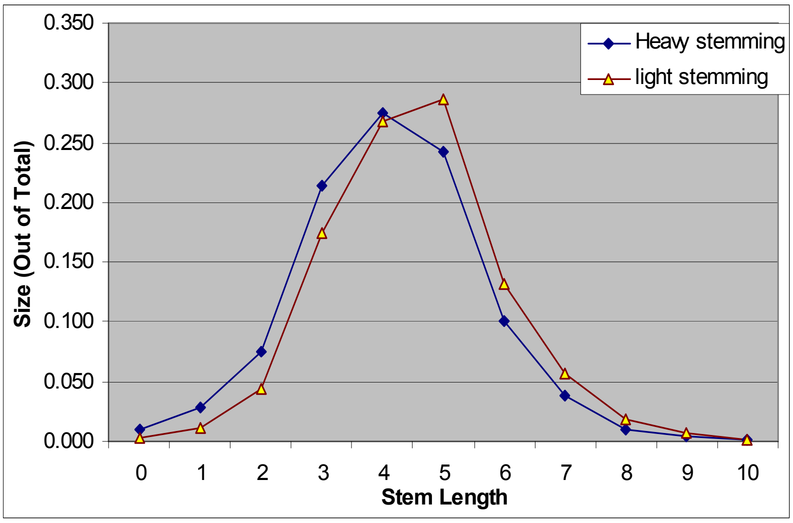

Figure (5): Distribution of stem lengths using two stemming techniques: light stemming and heavy stemming (Size is based on the total number of unique words = 6481)

To test the significance of difference between light stemming and heavy stemming, a set of randomly selected retrieval queries consisting fifty terms was matched against a corpus of about twenty-eight thousand document titles. The test of significance used was the Sign Test (a test of difference in location for two dependent groups), with level of significance being (α = 0.5) and the formula for calculating χ^2 being:

$$\frac{(|f_{o+} - f_{e+}| - .5)^2}{f_{e+}} + \frac{(|f_{o-} - f_{e-}| - .5)^2}{f_{e-}}$$

Where,

f_{o+} : obtained positive frequencies f_{e+} : expected positive frequencies

f_{o-} : obtained negative frequencies f_{e-} : expected negative frequencies

With *df* = 1, *Chi-square* (as determined by the χ^2 *Distribution*) must reach or exceed 3.84 to be significant at the 5% level.

Results and Discussion

Table (3) presents the results of heavy stemming and light stemming strategies against the actual figures of stems as determined by manual stemming for the four types of words contained in the sample. The difference in performance between the two computational strategies is shown in Figure (6). The bars under the zero-axis provide an indication of over-stemming while the corresponding bars with positive values provide an indication of under-stemming.

As we examine these results, the following observations can be made:
1. Heavy stemming failed to recognize prefixes in about nine percent of the actual number of prefixed words. It also erroneously treated about nineteen percent as having prefixes and suffixes when they actually do not. In comparison, light stemming failed to recognize only a small fraction of prefixes and gave erroneous results for about eleven percent.
2. Heavy stemming treated about four percent of the total number of words as having suffixes, and about twenty-four percent as containing prefixes and suffixes, when they actually do not. In comparison, light stemming gave about three percent erroneous results, in the case of suffixed words, and about nine and half percent erroneous results, in the case of words containing prefixes and suffixes.

Table (3): Performance of two word stemming strategies against actual number of stems as determined by manual stemming for each group

Strategy	Manual Stemming		Heavy Stemming		Light Stemming	
Words	Num	%	Num	%	Num	%
Prefixed only	3820	**58.9**	3240	**50.0**	3776	**58.3**
Suffixed only	341	**05.3**	597	**09.2**	519	**08.0**
Suf and Pref	298	**04.6**	1841	**28.4**	910	**14.0**
No Suf/Pref	2022	**31.2**	803	**12.4**	1276	**19.7**
Total	**6481**	**100.0**	**6481**	**100.0**	**6481**	**100.0**

A more accurate view of the erroneous stemming judgments can be obtained by analyzing the actual figures of over-stemmed and under-stemmed words. As Table 4 indicates, the majority of incorrect results came in the form of over-stemming and only a small percentage of words were under-stemmed. In either case, light stemming outperformed heavy stemming. About eighteen percent (18%) of the total number of distinct words were over-stemmed by the heavy stemmer with respect to the removal of prefixes, compared to about ten percent in the case of light stemming.

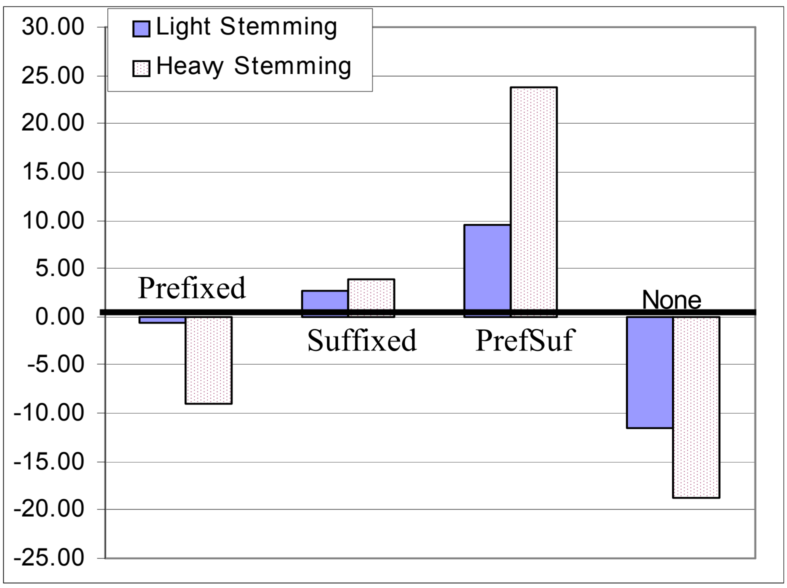

Figure (6): Viewing the results of light stemming and heavy stemming in terms of over-stemming and under-stemming percentages.

The highest percentage of erroneous judgments is encountered in the case of handling suffixes and non-affixed words. While the sample involves only a small percentage of suffixed words (i.e., about 10%), almost about thirty percent were over-stemmed by the heavy stemmer against about thirteen percent in the case of light stemming.

Further analysis of the results based on the type of affixes, as presented in Figure (7), shows that the two stemming strategies treated many instances of non-affixed words as having prefixes or suffixes which increased the number of words being considered as having prefixes or suffixes. The fact that some prefixes and suffixes are one-letter affixes increases the likelihood of mistaking original final or initial letters for affixes. The suffixes “taa” (ت), “noon” (ن), and “yaa” (ي) contributed about sixty percent of the total number of incorrect results made by the heavy stemming strategy under the “suffixed-words” category in Table 4.

Table (4): Over-stemmed and under-stemmed words involving prefixes and suffixes

Strategy	Heavy Stemming	Light Stemming
Prefixed Words		
Over-Stemming	18.24%	09.81%
Under-Stemming	03.38%	01.11%
Suffixed Words		
Over-Stemming	29.95%	12.87%
Under-Stemming	00.83%	00.68%

As pointed out earlier, an attempt was also made to examine the impact of stop words (such as separate pronouns, prepositions, and conjunctions) on the performance of the two stemming strategies. Based on the results shown in Figure (7) and Figure (8), the removal of stop shows considerable improvement, especially with respect to suffixed words. The improvement was more apparent in the results provided by light stemming than heavy stemming.

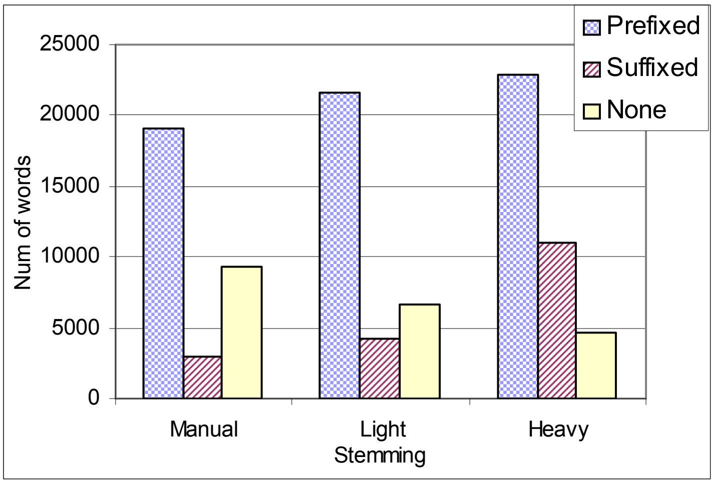

Figure (7): Performance of “heavy” and “light” stemming strategies against manually determined number of prefixed, suffixed, and non-affixed words (stop words were no removed).

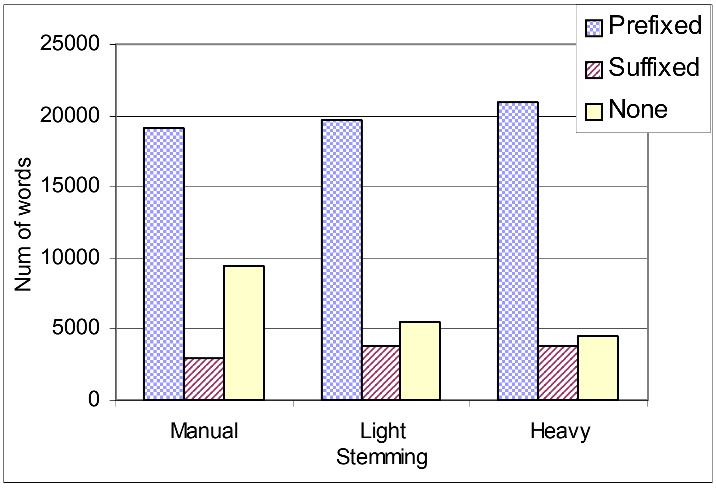

Figure (8): Performance of “heavy” and “light” stemming strategies after removing a set of stop words.

Further evidence for the superiority of light stemming over heavy stemming comes from the results of the retrieval experiment conducted over a set of fifty query items as outlined above. With *Chi-square* (χ^2) = *5.6* (i.e., exceeding 3.84 to be significant at the 5% level), the test of significance has shown that light stemming performs significantly better than heavy stemming. However, it has been observed that performance of the two strategies gets closer (and becomes similar in some cases), as the level of stemming needed goes down. A case in point is a word such as (استثمار) “*istithmar* / investment”, for which zero stemming is performed by both strategies. Hence, the two strategies will exhibit similar performance.

Conclusion

The fact that Arabic prefixes and suffixes do not occur real texts in the same rate of frequency gave the underlying rational for conducting the study presented in this paper. It has been noted that a high percentage of word affixes are caused by only a small number of suffix and affix combinations. It has been demonstrated that the definite article “Al” and the connected conjunction “Waw”, for instance, have the highest rate of frequency among all prefixes, while some other prefixes are rarely encountered in real texts. It has been assumed, accordingly, that a light stemmer, in which only the highly occurring prefixes and suffixes are removed will exhibit better stemming performance than a heavy stemming strategy in which most of the prefixes and suffixes are removed.

It has been shown in the present study that light stemming significantly outperforms heavy stemming. This conclusion confirms the findings reported by some of the researchers in the field, specifically those reported recently by Larkey et.al (2002) and Darwish (2003). However, a few remarks have to be made about the results of this study. The first of which is that, even though light stemming seems to perform better than heavy stemming, it fails in many instances to conflate related terms as a result of ignoring infixes in some instances and as a result over-stemming or under-stemming in others.

The other remark relates to the level of stemming required for a given term. If the term to be handled has no prefixes or suffixes to be removed, the two stemming strategies are expected to exhibit similar performance. It has been observed in this study that, as the level of stemming required for certain words (especially words that start and end with letters which are not confused with prefixes or suffixes) decreases, the likelihood increases of having the two strategies getting closer in performance.

The final remark that should be made here relates to the fact that some Arabic words go through a set of transformations due to the existence of weak letters. No matter how well a stemming technique is, the fact remains that all the techniques that have been tried so far to do not offer an efficient way to handle this type of words. In some cases, even if you may have the right stem for the item to be searched for, you may not find the corresponding right match in the text due to the lexical or grammatical transformation. Could the solution come from a corpus-based stemming, whereby the appropriate stem of a given word is looked up from, or checked against the text of document(s) rather than just relying on rules of prefixing and suffixing? The answer to this question should come from further research.

REFERENCES

Abu-Salem, H., Al-Omari, M., and Evens, M. (1999). Stemming Methodologies over Individual Query Words for an Arabic Information Retrieval System. JASIS, 50(6), 524-529.

Al-Fedaghi, S.S. and Al-Anzi, F.S. (1989). A New Algorithm to Generate Arabic Root-Pattern Forms. Proceedings of the 11th National Computer Conference and Exhibition (Dhahran, Saudi Arabia, 4-7 March 1989), pp. 391 - 400.

Al-Kharashi, I.A. and Evens, M.W. (1994). Comparing Words, Stems and Roots as Index Terms in an Arabic Information Retrieval System. Journal of the American Society for Information Science, 45, 1994, 548-560.

Al-Shalabi, R. (1996). Design and implementation of an Arabic morphological system to support natural language processing. Ph.D. thesis, Computer Science, Illinois Institute of Technology, Chicago.

Al-Shalabi, R. & Evens, M. (1998). A computational morphology system for Arabic. Workshop on Semitic Language Processing. COLINGS-ACL'98, University of Montreal, August 16, 1998, pp. 66-72.

Al-Tayyar, M.S. and Bechkoum, K. (1998). The effectiveness of the morphological analysis for text retrieval in Arabic. Proceedings of the 6th International Conference and Exhibition on Multilingual Computing, Cambridge, 17-18 April 1998, pp. 2.4.1-2.4.13.

Beesley, K.R. (1996). Arabic finite-state morphological analysis and generation. Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), Vol.1, pp.89-94.

Buckwalter, T. Arabic Lexicography. [Available at: http://members.aol.com/ArabicLexicons/]

Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O. (2001). Improving Precision in Information Retrieval for Swedish Using Stemming. In: Proceedings of NODALIDA'01 – 13th Nordic Conference on Computational Linguistics, Uppsala.
[Available at: http://stp.ling.uu.se/nodalida/pdf/carlberger.pdf]

Croft, W.B. and Xu, J. (1995). Corpus-Specific Stemming Using Word Form Co-occurrence. 4th Annual Symposium on Document Analysis and Information Retrieval. Las Vegas: Univ. of Nevada , 147-59.

Darwish, K. (2003). Probabilistic methods for searching OCR-degraded Arabic text. Unpublished Ph.D. Thesis. University of Maryland (USA).

De Roeck, A.N., and Al-Fares, W. (2000). A Morphologically sensitive clustering algorithm for identifying Arabic roots. Proceedings of 38th Annual Meeting of the ACL, Hong Kong. Retrieved June 2003, from http:// citsseer.nj.nec.com /deroeck00morphologically.html.

Frakes, W.B. (1992). Stemming Algorithms. In: Frakes, W.B. & Baeza-Yates, R. (eds.) Information Retrieval: Data Structures and Algorithms. Englewood Cliffs: Prentice-Hall, 131-160.

Freund, G.E.and Willett, P. (1992). Online Identification of Word Variants and Arbitrary Truncation Searching Using a Similarity Measure. Information Technology: Research and Development, 1: 177-87.

Harman, D. (1991). How Effective is Suffixing? Journal of the American Society for Information Science, 42, 1991, 7-15.

Hmeidi, I., Kanaan, G., and 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 algorithm – a case study for detailed evaluation. JASIS, 47(1), 70-84.
Khoja, S. and Garside, R. (1999). Stemming Arabic Text. Computing Department, Lancaster University, Lancaster. Internet Home Page, 1-7. [Available at: http://www.comp.lancs.ac.uk/computing/users/khoja/stemer.ps]
Kraaij, W. and Pohlmann, R. (1996). Viewing stemming as recall enhancement. In: Proceedings of ACM SIGIR96, pp. 40-48.
Krovetz, R. (1993). Viewing morphology as an inference process. Proceedings of the 16th ACM/SIGIR Conference . New York, Association for Computing Machinery, pp. 191-202.
Larkey, L.S., Ballesteros, L., and Connell, M.E. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02), Tampere, Finland, 275-282. [Available at: http://ciir.cs.umass.edu/pubfiles/ir-249.pdf]
Lennon, M. , Perce, D.S., Tarry, B.D., and Willett, P. (1981). An evaluation of some conflation algorithms for information retrieval. Journal of Information Science, 3, 177-183.
McNamee, P., Piatko, C., and Mayfield, J. (2002). JHU/APL at TREC 2002: Experiments in filtering and Arabic retrieval. Proceedings of the 11th Text Retrieval Conference (TREC '02). [Available at: http://ciirs.cs.umass.edu/pubfiles/ir-278.pdf]
Monz, C. & de Rijke, M. (2001). Shallow morphological analysis in monolingual information retrieval for German and Italian. In: Proceedings of the CLEF 2001 Workshop: Cross-Language Information Retrieval and Evaluation, C. Peters, Ed., Springer Verlag.
Mustafa, S.H. and Masoud, F. (2000) A Backward algorithm for lexical analysis of textual Arabic words. Abhath Al-Yarmouk: Basic Science and Engineering Series, 9 (1), 91-125.
Paice, C.D. (1994). An evaluation method for stemming algorithms. Proceedings of the 7th Annual International ACM-SIGIR Conference on Research and Devlopment in Information Retrieval (UK), edited by W.B. Croft and C. Van Rijsbergen, London, Springer-Verlag, pp. 42-50.
Paice, C.D. (1996). Method for evaluation of stemming algorithms based on error counting. Journal of the American Society for Information Science, 47(8), pp.632-649.
Pirkola, A. (2001). Morphological typology of languages for IR. Journal of Documentation, 57(3), 330-348.
Popovic, M. and Willet, P. (1992). The effectiveness of stemming for natural language access to Slovene textual data. *Journal of the American Society for Information Science,* 43(5), 384-390.
Xu, J., Fraser, A., and Weischedel, R. (2002). Empirical Studies in Strategies for Arabic Retrieval. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*, Tampere, Finland 269-274. [Available at: http://*www.isi.edu/ ~fraser/pubs/sigir2002.pdf*]

APPENDIX

Table A1: Distribution of -words based on the first prefix (Sample 1)

Prefix	Distinct	Ratio	Freq.	Ratio
ال	2093	0.508	12522	0.68
و	1075	0.261	2788	0.15
ل	439	0.107	1355	0.07
ب	188	0.046	642	0.03
ت	110	0.027	344	0.02
ف	9	0.002	323	0.02
ي	121	0.029	315	0.02
أ	24	0.006	78	0.00
ك	19	0.005	59	0.00
ن	20	0.005	59	0.00
س	10	0.002	39	0.00
ا	10	0.002	26	0.00
Total	**4118**	1.00	**18550**	1.00

Table A2: Distribution of words based on prefixes

Sample1 (6481 distinct words**)**			**Sample2 (**1359 distinct words**)**		
Prefix	**Count**	**Ratio**	**Prefix**	**Count**	**Ratio**
أيـ، فالـ، وبالـ، ول، وا	5	0.000	أتـ، لتـ، ليـ، وبـ، وبالـ	5	0.005
سنـ	2	0.000	كالـ	2	0.001
فيـ	2	0.000	فالـ	3	0.002
كالـ	2	0.000	نـ	3	0.002
وللـ	2	0.000	ونـ	3	0.002
لنـ	3	0.000	وأ	4	0.003
ست	4	0.001	فيـ	5	0.004
سيـ	4	0.001	فتـ	10	0.007
فـ	6	0.001	بالـ	11	0.008
ويـ	9	0.001	فأ	11	0.008
ا	10	0.002	إ	14	0.010
كـ	17	0.003	لـ	14	0.010
نـ	20	0.003	ويـ	14	0.010
أ	23	0.004	للـ	16	0.012
بالـ	81	0.012	فـ	20	0.015
بـ	107	0.017	أ	22	0.016
تـ	110	0.017	تـ	22	0.016
يـ	121	0.019	والـ	27	0.020
ل	215	0.033	بـ	30	0.022
للـ	221	0.034	يـ	76	0.056
والـ	481	0.074	و	116	0.085
و	580	0.089	الـ	242	0.178
الـ	2093	0.323			
Total	**4118**	**1.00**	**Total**	**670**	**1.00**

Table A3: Distribution of words based on suffixes

Sample1(6481 distinct words**)**			**Sample 2(**1359 distinct words**)**		
Suffix	**Count**	**Ratio**	**Suffix**	**Count**	**Ratio**
ان	1	0.000	ن	1	0.001
تا	1	0.000	هن	1	0.001
وها	1	0.000	ون	1	0.001
وه	2	0.000	وه	1	0.001
ته	3	0.000	تها	2	0.001
ك	4	0.001	هما	2	0.001
كم	5	0.001	وها	3	0.002
هما	8	0.001	هم	4	0.003
ي	10	0.002	نا	9	0.007
ون	26	0.004	كم	11	0.008
وا	28	0.004	ي	12	0.009
ت	31	0.005	وا	14	0.010
ا	32	0.005	ك	18	0.013
هم	41	0.006	ها	25	0.018
نا	63	0.010	ت	26	0.019
ه	155	0.024	ا	39	0.029
ها	228	0.035	ه	83	0.061
Total	**639**	**1.00**	**Total**	**252**	**1.00**