Volume: 2 Issues: 10 [December, 2019] pp.81-90]
International Journal of Modern Trends in Social Sciences
eISSN: 2600-8777
Journal website: ijmtss.com

RETRIEVAL PERFORMANCE OF ARABIC LIGHT STEMMERS

Dr. Ouahiba Saoudi
Kulliyyah of Islamic Revealed Knowledge and Human Sciences,
International Islamic University Malaysia (IIUM), Malaysia.
(Email: ouahibasa@gmail.com)
Prof. Roslina Othman
Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia (IIUM), Malaysia.
(Email: roslina@iium.edu.my)

Received date: *11-12-2019*
Revised date: *15-12-2019*
Accepted date: *16-12-2019*
Published date: *16-12-2019*

To cite this document: Saoudi, O., & Othman, R. (2019). Retrieval Performance of Arabic Light Stemmers. *International Journal of Modern Trends in Social Sciences*, 2(10), 81-90.
DOI: 10.35631/IJMTSS.210008

Abstract: *Despite the fact that stemming greatly improves Arabic information retrieval performance, yet no standard stemmer emerges in the field of Arabic IR due to some limitations and shortcomings. Among the recurring problems is that the stemmer can reduce unrelated words to the same stem as well as fall short to reduce related words to a common stem. Many studies have suggested Arabic algorithms to address the problem associated with stemming. This paper aims to review the state of the retrieval performance of Arabic Light stemmers based on the main objectives achieved, causes for retrieval success and failure, retrieval measure, the affixes, and methodologies. The results showed that light 10 has better retrieval performance compared to other reviewed Arabic light stemmers.*

Keywords: *Arabic Information Retrieval, Arabic Light Stemming, Retrieval Performance*

Introduction

Stemming is one of information retrieval strategies that are used to retrieve and locate query search with relevant documents. In other words, stemming is a process that conflates different words together and accordingly allows the system to retrieve the documents which are not exactly matching query words. This process has positive impact on the storage and index size, as it helps in index size reduction allowing for more storage space. Furthermore, stemming can be seen as a mean to increase the number of word occurrences (Allan and Kumaran, 2003). Stemming is used in many types of language processing and text analysis systems. There are

many researches works on the application and evaluation of stemming strategies for information retrieval purposes. Stemming has been implemented in many languages and proved its effectiveness. Arabic language is among those languages. As such, in broader perspective, languages have many variations in terms of structure, morphology and function. For instance, the variation in morphology can lead to various differences at the level of the effectiveness and precision of information retrieval. More than this, in information retrieval (IR), the relationship between user information need, which will be formulated to a query and document, is mostly determined by the number and frequency of terms which they have in common. Unluckily, words have many morphological variants which will not be recognized by term-matching algorithms without some form of natural language processing. Therefore, stemming algorithms have been developed for IR in order to reduce morphological variants to their root or stem form.
There are several types of stemmers or stemming algorithms. For example, affix removal, root stemmers, which return search query into its root or stem and statistical stemmer which is based on co-occurrence.

In the context of Arabic information retrieval, many Arabic stemmers were developedby addressing the morphological variations of Arabic language in order to improve Arabic retrieval performance against information needs. However, yet no standard Arabic stemmer emerged. The most popular and successful techniques developed to extract stems of the words is light stemming technique. Light stemming is defined as a process of removal a small set of prefixes and/or suffixes without trying to deal with infixes, or recognize patterns and find roots (Larkey et al., 2002). This paper sheds some light on the retrieval performance of several introduced Arabic light stemmers. It aims at showing the strength and shortcoming of some light stemmer paving the way for a better understanding of these stemmers and the possibility of coming up with a more integrated and effective stemmer that overcomes the shortcoming of the various light stemmer which currently in use.

In specific terms, this paper aims to review the state of the retrieval performance of Arabic LIGHT stemmers based on the main objectives achieved, causes for retrieval success and failure, retrieval measure, the affixes and methodologies.The Selected Arabic light stemmers to be reviewed are as follow: AL-Stem (Darwish,2002), Berkeley, (Thabet, 2004), al-Ameed, Light1, Light2, Light3, light 8, light10, CondLight, LS, Light11, Light12 and Light13

Characteristics of Arabic Language

It is obvious that most Arabic words are morphologically derived from short list of generative roots, which constitute the bare verb form. (Abu El-Khair, p.508). For example, علم means "science" is the root base of different words with different meaning معلمات= teachers, معلومات = information, علوم = sciences, علم = flag.

The other chief feature of the Arabic language is that words are written in horizontal lines from right to left. The way of writing of letters change depending on their occurrence in the sentence whether they appear at the beginning, middle or end of a word. In Arabic language the noun can be feminine or masculine in form dual, singular or plural.

طفلان → two children, masculine.

معلمتان → two teachers, feminine.

كرسيان → Two chairs, masculine.

Another important feature of Arabic language is the existence of diacritics. The function of the eight main diacritics of Arabic language plays a vital role in the changing of meaning of words (Moukdad and Large, 2001). The same word with different diacritics makes two different meanings. These vowel diacritics appear fully in Qur'an to avoid any problem of ambiguity or mistakes. However, they are not used in modern Arabic script, except in some children stories to make it understandable.

The other leading feature of Arabic language is that it is morphologically complex in which it allows a large degree of flexibility in forming words. That is to say complex rules govern the creation of morphological variations, making it possible to form hundreds of words from one root (Fedaghi and al-Sadoum, 1990). For example, giving root words that derive many words with different meanings. In fact, all these features and others bear heavily on Arabic information retrieval function. Hence, they have to be considered in choosing appropriate indexing strategies for IR so as to avoid all sorts of complexities and problems.

Arabic Light Stemmers

Aljlayl and Frieder in 2002 have proposed root-based stemmer and examined different approaches of light stemming. Theywere the first to introduce the light stemming. Theresearch adopted Khoja's algorithm for root-based retrieval and introduced new stemming called "light stemming". The proposed light stemming (LS) did not specify the prefixes and suffixes that should be removed from Arabic word. Rather, the light stemmer algorithm that was given by the two researchers can be used for any Arabic valid affixes. The research investigated the effectiveness of light stemming against root-based and word-based for Arabic IR. The findings of the research showed light stemmer significantly outperformed the root-based approach. Khoja's stemmer was found superior over surface-based (without stemming). The differences between two approaches were statistically significant by using paired t-test and Wilcoxon sign test.

After introducing light stemming algorithm by Aljayl and his colleague, Larkeyet al., (2002) was encouraged to develop several light stemmers for Arabic retrieval light1, light2, light3, and light8. The affixes removals by those stemmers are shown in Table1. They were tested for both monolingual and cross-language retrieval. In addition, the research compared the retrieval performance of developed light stemmers and other morphological analyzers like Khoja stemmer, statistical stemming using co-occurrence analysis. Normalization is a type of stemming; it was also compared with light stemmers. The findings show that light stemming, specifically light 8 has improved the effectiveness of Arabic information retrieval for monolingual and cross-language retrieval compared to other stemmers. Measures used in the research to compare the performance of stemmers were recall and precision measures.

Al-Stem is an Arabic light stemmer developed by Darwish in 2002. The stemmer was intended for research purposes only, removed many prefixes and suffixes as illustrated in Table 1. It was claimed that the Al-Stem is more Aggressive than Light10 stemmer.

In 2003 Chen and Gey have developed two Arabic stemmers and an Arabic stop list at TREC 2001. The two researchers created a MT (Machine Translation)-based stemmer and a light stemmer. The light stemmer was called Berkeley which shares many of prefixes and suffixes that

should be removed with the light stemmers developed by Larkey et al and the one was developed by Darwish (Al-Stem). The Berkeley light stemmer was compared with two trigram indexing, MT-based stemmer, without stemming, Al-Stem which developed by Darwish for both monolingual and cross-language retrieval. The results showed that the Berkeley light stemmer performed better than automatically created MT-based stemmer.

The light stemmer followed the following sequence:

1- If the word is at least five-character long, remove the first three characters if they are one of the following: بال, فال, كال, ولل, مال, اال, لال,وال
2- If the word is at least four-character long, remove the first two characters if they are one of the following: ا كا ول وي وس سي لا وب وت وم لل با., ف, ال,وا
3- If the word is at least four-character long and begins with, remove the initial letter
4- If the word is at least four-character long and begins with either or, remove or only if, after removing the initial character, the resultant word is present in the Arabic document collection.
5- Recursively strips the following two-character suffixes in the order of presentation if the word is at least four-character long before removing a suffix: , يا, هن, كن, تم, تن, ين, ان, ات,ون يا,ني
6- Recursively strips the following one-character suffixes in the order of presentation if the character is at least three-character long before removing a suffix: ة, ه, ي,ت

Al-Ameed, et al. (2002) proposed five stemming algorithms in an attempt to enhance the TREC-2002 Arabic light stemmer presented by Kareem Darwish. In other words, the researchers used TREC-2002 Arabic light stemmer (Al-Stem) as a benchmark in order to compare their light stemmer with that presented by Kareem Darwish. The system performance evaluation was based on two testing manners; the first manner concentrated on measuring the number of acceptable produced words as an output of applying the stemmer algorithms on each test group. The second manner is grounded on measuring the frequency of removing affixation terms from the test words. The research outcomes were compared with TREC-2002 algorithm results. The results showed that proposed stemmers provide better accepted outcomes of Arabic words with up to 30-50% more than TREC stemmer outcomes. However, one cannot say that the proposed Arabic stemmers' outcomes are better than TREC-2002 stemmer results because the researchers did not use the same test collection of TREC-2002 in their research as well as standard evaluation measures. Therefore, one can conclude that one of the weaknesses of this research is that it did not follow appropriate methodological approach.

In another attempt, Naglaa Thabetin 2004 proposed a new stemming approach based on a light stemming technique that uses a transliterated version of the Qur'an in western script. These are the main procedures:

• Remove prefixes (wa, fa, la, li, lil, bi, ka, sa, s^a, al)
• After stemming, the word is inserted back into the word list.

Six groups of suffixes are identified ranging from one-letter suffixes to six-letter suffixes. The system starts stemming the words in the word lists from the longest prefixes (six-letter prefixes) to the three-letter prefixes. Stemming the one and two-letter suffixes causes some ambiguity, since some of the suffixes could sometimes be part of the word stem. To resolve this problem, the stemmer sorts the words alphabetically. In the sorted list of words, if a given sequence

displays a variety of suffixes including one and two-letter suffixes, the suffixes are removed and the stem is retained, otherwise the word is left intact. The results for seven long suras selected randomly and representing 6% of the Qur'an showed that the stemmer achieves an accuracy of 99.6% for prefix stemming and 97% for suffix stemming. The disadvantage of this study is there was no experimental evaluation provided.

In order to look at the similarities and differences of the list of prefixes and suffixes that the reviewed light stemmers stripped off Table 1 was summarized and checking if there is any failure analysis reported in the research works. Those light stemmers share many prefixes and suffixes, besides some of these light stemmer removes only few prefixes and suffixes in order to protect the same meaning of the queries like light8 and light10.

In 2005 Larkey, Ballesteros and Connell reassessed the light stemmers which were developed in the previous research at SIGIR in 2002 and developed another light stemmer called light10. The light stemmers are demonstrated in Table 1. The research covers three folds. First, compared several light stemmers with each other for Arabic retrieval and cross-language retrieval (raw, normalization, light 1, light 3, light 8 and light 10) using TREC data. The results showed that light stemming is more effective in Arabic retrieving. It was reported that each of these increments is statistically significant except light 10 vs. light 8. This means that the research run statistical analysis, which was not stated in the article and no reporting of what test was run and its full results. The second fold is comparing Khoja's stemmer with light 10, normalization, raw (means no changes made to the word) for both monolingual and cross-lingual retrieval. The results showed that light 10 stemmer is significantly more effective than the Khoja stemmer for monolingual. The same results appeared for CLR. The research also has compared light 10 stemmer with several stemmers based on morphological analysis (Khoja, Buckwalter morphological analyzer, Diab Tokenizer) with expended queries and unexpended queries. The results showed that light 10 performed effectively than other morphological analyzer based on recall and precision measures.

The other major work on AIR and particularly in light stemming is the work of Nwesri (2008) The researcher investigated several techniques to improve Arabic text retrieval. As such, many techniques that improve the performance of AIR systems were explored. Generally, the main argument was on the idea that most current stemmers remove affixes without checking whether the removed letters are actually affixes. To address this concern the researcher, propose lexicon-based improvements to light stemming that makes a distinction between the core letters from proper Arabic affixes. To this end, few rules to stem most affixes and show the effects of each individual rule on retrieval effectiveness as well as using all rules together were discussed. The researcher used the TREC 2001 test collection in order to demonstrate that applying relevance feedback with the proposed rules, that use the lexicon to differentiate core letters from actual prefixes and suffixes, produces significantly better results than light stemming.

It mainly compared effectiveness of existing AIR systems in order to prove that light stemming techniques are superior to existing systems. To achieve that, the researcher verified whether letters constitute an affix not only by checking whether the word with and without that affix exists in lexicon, but also by replacing that affix with other equivalent ones so as to examine the new instances against the lexicon.

This research has made improvement into light 10. It has developed three improved versions of the light10 stemmer namely; light11, light12, and light13. In the light11 stemmer, the researcher managed to reduce the number of suffixes (ية، ة). In the light12 stemmer, the number of stemmed suffixes was reduced to four suffixes. (ي، ه، ات، يه). In light13, he removed the same suffixes as in the case of light 12 in addition to removing the definite article (ال) and prepositions and conjunctions. The empirical result of the study has shown that light11 and light12 do not significantly improve retrieval performance, but light13 made significant improvement in recall [*t*-test, p = 0:029]. None of the algorithms led to a significant improvement in MAP or P@10. (Nwesri, 2008). As an overall result, the study has improved light stemming by introducing rules that use the lexicon to distinguish core letters from actual prefixes and suffixes, tested the effectiveness of AIR systems on a large text collection, and introduced algorithms that distinguish foreign words

One of the shortcomings of this study is the tendency of over reducing the valid prefixes and suffixes which are safe and with no negative impact on AIR like suffix (ين، ون). The latter occur frequently in Arabic documents; therefore, they should be omitted in line with other researches such as the one conducted by Larkey, Chen and Aljlayl. Furthermore, he removed the prefixes (فال، بال، كال) which are also widely occurring in Arabic document and safe to remove.

The latest Arabic light stemmer proposed by Al-Lahham et al., (2018), named conditional light stemmer (CondLight). The researchers added new prefixes and suffixes to the table of Light10 and proposed a set of conditions on removing these affixes. These conditions are derived from the morphological nature of Arabic words. The application of the proposed light stemmer showed that adding some conditions to the extended light stemmer enhances the retrieval especially at lower recall levels.

To sum up, light 10 has proved to be more effective in improving retrieval performance of AIR. As a result, light 10 was included in Lemur toolkit, for research in language modelling and information retrieval. However, it fails in addressing morphological complexity of language in Arabic retrieval. For instance, irregular plural and under-stemming errors where many other inflected words of the queries are left out not conflated with them. Besides, yet there is no commercial system available in internet which implements any kind of stemmers. This may be due to the cost of this kind of implementation or to unavailability of standard stemming algorithm to be adopted by commercial systems.

Table1. Summary of Arabic Light Stemmers Affix Removal

Author and date	**Algorithm name**	**Remove prefix**	**Remove suffix**	**Failure analysis**
Aljlayl and Frieder (2002)	LS	NA	NA	NA
Larkey, et al., (2002)	Light1	ال، وال، بال، كال، فال	none	NA
	Light2	ال، وال، بال،كال، فال، و	none	
	Light3	Same as previous	ه، ة	
	Light8	Same as previous	ها، ان، ات، ون،ين، يه، ية، ه،	

			ة، ي	
Darwish (2002)	Al-Stem	وال، فال، بال، بت، يت، لت، مت، وت، ست، نت، بم، لم، وم، كم، فم، ال، لل، وي، لي، في، وا، فا، لا، با	ات، وا، ون، وه، ان، تي، ته، تم، كم، هم، هن، ها، ية، تك، نا، ين، يه، ة، ه، ي، ا	NA
Chen and Gey (2003)	Berkeley	وال، لال، سال، اال، مال، ولل، كال، فال، بال، وا، ال، فا، كا، ول، وي، وس، سي، لا، وب، وت، وم، لل، با، و، ل، ب	ون، ات، ان، ين، تن، تم، كن، كم، هن، يا، ني، وا، ما، نا، هم، ية، ها، ت، ي، ه، ة	NA
Najwa (2004)	NA	و، ف، ل، لل، ب، ك، س، سا، ال		NA
Al-Ameed, et al. (2005)	NA	ي، ت، ن	ه، ة، ك، و، ي، ن، ا، ت	NA
		ال، لل، سي، سا، ست، سن، كا، فا، با، ب، ل، لي، لت، لن، فت، في، فن		
		وال، بال، فال، كال، ولل، وسي، وست، وسن، وسا، ولا، ولي، ولت، ولن	ان، ين، ون، ات، هم، هن، ها، كم، كن، نا، وا، تم، ني، تن، ته، يه، ما، يا، تا، تك	
		وبال		
Larkey, et al. (2005)	Light10	ال، وال، بال، كال، فال، لل، و	ها، ان، ات، ون، ين، يه، ية، ه، ة، ي	NA
Nwersi, 2009	Light11	ال، وال، بال، كال، فال، لل، و	ان، ات، ون، ين، يه، ه، ي	NA
	Light12	ال، وال، بال، كال، فال، لل، و	ات، يه،ه، ي	
	Light13	ال، وال، بال، كال، فال، لل، و	ات، يه،ه، ي	
Al-Laham, et al., 2018	CondLight	ب، ت، ل، ي، ف، س ال، وال، بال، كال، فال، لل، و	هم، هن، وا ها، ان، ات، ون، ين، يه، ية، ه، ة، ي	NA

Objectives achieved by Arabic Light Stemmers

The main objectives achieved by stemmers are that most stemmers which were developed tried to tackle the issue of morphology complexity by incorporating language morphology knowledge of the Arabic into stemming algorithm application. In addition, all stemmers were developed for AIR, whether based on affix removal or root based have achieved an important objective i.e., improving Arabic information retrieval performance against information needs.

Table 2: Summary of Achieved Objectives by Stemmers

Stemmers	Achieved objectives
Al-Jalyl	1. show that the light stemming algorithm significantly outperforms the root-based algorithm. It also shows that a significant improvement in retrieval precision can be achieved with light inflectional analysis of Arabic words. 2. In general, it appears that all stemmers significantly perform better than no stemming at all. 3. The root algorithm based on the work of Khoja, which is considered as an aggressive stemmer, has shown performance superiority over surface-based (no stemming) approach.
Larkey light1 to 8	1. The best light stemmer was more effective for cross-language retrieval than a morphological stemmer which tried to find the root for each word. *2.* A repartitioning process consisting of vowel removal followed by clustering using co-occurrence analysis produced stem classes which were better than no stemming or very light stemming, but still inferior to good light stemming or morphological analysis. *3.* Without stemming, the dictionary translations of query terms were unlikely to match the forms found in documents. In short, with sufficient parallel data, stemming may be unnecessary.
Darwish	1. The paper presents a rapid method of developing a shallow Arabic morphological analyzer. The analyzer will only be concerned with generating the possible roots of any given Arabic word. The analyzer is based on automatically derived rules and statistics. 2. For evaluation, the analyzer is compared to a commercially available Arabic Morphological Analyzer However, in this paper, it presents a quick method for performing shallow morphological analysis for use in information retrieval, which entails finding the roots of words, in one day. The method is based on collecting statistics from word-root pairs: 3. To build morphological rules for deriving roots from words, 4. To construct a list of prefixes and suffixes, and to estimate the probability that a rule will be used, or a prefix or suffix will be seen.
Larkey 2005	1. Stemming has a large effect on Arabic information retrieval, far larger than the effect found for most other languages. 2. The root detector algorithm performed much worse than the LS.

Arabic Light Stemmers Performance

Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. Therefore, precision is a measure of purity in retrieval performance. The Table6 shows the average precision for most of the Arabic light stemming reviewed in the literature. Some research works have calculated average precision with query expansion and without expansion. For example, Aljlayl and Frieder (2002) and Chen and Gey (2003) research works results showed that average precision of the light stemmers with query expansion are higher than average precision without query expansion. Larkey and Connell 2002 and 2005 research implemented for monolingual run precision at 11 recall points. However, they only provide precision measure result. Notably, light10 precision average is higher than other light stemmers that were reviewed in this research. Light Stemmers Performance Based on Precision Measure

Basically, recall is the number of retrieved relevant items for any given retrieved set. Recall is, therefore, a measure of effectiveness in retrieving performance and can be viewed as a measure of effectiveness in including relevant items in the retrieval set (Lancaster, 1979). However, most of the researches that investigated different Arabic light stemming have not stated Recall ratio results. Except for Chen and Gey (2003) research which stated the recall results for Berkerly light stemmer performance are 4952 with query expansion and 4543 when query unexpended. Al-Stem recall results showed 4864 with query expansion and 4500 queries is unexpanded. This revealed that average recall of light stemming improves when query is expanded. Even though, the research of Nagla and Al-Ameed haven't used recall and precision measures, they have shown an accuracy of 99.6% for prefix stemming and 97% for suffix stemming. Since their research works did not implement standard measures for retrieval performance. As a result, it is hard to make interpretation on their findings as well. Experimental results show that the conditional light stemmer gains about 5% enhancement of retrieval over the original Light10 stemmer. 11-recall levels precision-recall curve for Light10 and CondLight stemmers. The enhancement occurs mainly a at lower recall level which satisfies users' needs by retrieving more relevant documents at the first page.

Conclusion

Stemming is common requirement of Arabic information retrieval. It is a preliminary step in many applications involving information retrieval. Arabic language is morphologically complex which is complicated more than any other language. This paper reviewed the state of the retrieval performance of Arabic light stemmers based on the main objectives achieved, causes for retrieval success and failure, retrieval measure, the affixes and methodologies. TheSelected Arabic light stemmers to be reviewed are as follow: AL-Stem (Darwish,2002), Berkeley, (Thabet, 2004), al-Ameed, Light1, Light2, Light3, light 8, light10, CondLight, LS, Light11, Light12 and Light13.

The reviewed Arabic light stemmers showed that there was shortage of information and detailson analyzing the causes of retrieval performance success or failure in AIR especially for stemmers. The reported causes of success and failure included the stemming method and the type of affixes removal.In addition, there is no comprehensive list of Arabic prefixes and suffixes that proved to be standard and effective for an Arabic stemmer such as the case of English Porter stemmer. Therefore, more research and empirical studies are needed particularly on analysing the nature of proposed prefixes and suffixes for removal.

References

Abu El-Khair, Ibrahim Hassan., 2003. Effectiveness of document processing techniques for Arabic information retrieval. Unpublished Ph.D. Dissertation University of Pittsburgh.

Al Ameed, Hayder K. and others., 2002. Arabic Light Stemmer: a New Enhanced Approach. The Second International Conference on Innovations in Information Technology (IIT'05). Retrieved 17/12/2008, from:

Al-Fedaghi, Sabah S. and Al-Sadoun, Humoud B., 1990. Morphological compression of Arabic text. Information Processing and Management. V. 26, No. 2, pp. 303-316

Aljlayl, Mohammed and Frieder, Ophir., 2002. On Arabic search: improving the retrieval effectiveness via a light stemming approach. ACM CIKM,

Al-Lahham, Yaser, Khawlah, Matarneh, and Hassan, Mohammad. Conditional Arabic Light Stemmer: CondLight. The International Arab Journal of Information Technology, Vol. 15, No. 3, Special Issue 2018

Allan, James & Kumaran. (2003). Stemming in the language modeling framework. The 26th Annual International ACM SIGIR Conference.

Chen, Aitao and Gey, Fredric., 2002. Building an Arabic stemmer for information retrieval. Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.3777

Darwish, K. (2002). Building a shallow Arabic morphological analyser in one day. In Proceedings Of The Association For Computational Linguistics (ACL-02), 40th Anniversaty meetings, University of Pennsylvania, Philadelphia, 47-54. http://www.it-annovations.ae/lit005/proceedings/articles/G_1_IIT05_Hayder.pdf

Khoja, S., & R. Garside. (September 1999). Stemming Arabic text. Technical report, Computing Department, Lancaster University, Lancaster.

Larkey, L.S., and Connell, M.E., 2002. Arabic Information Retrieval at UMass in TREC-10. *Proceedings of the Tenth Text REtrieval Conference (TREC 2002)*, Gaithersburg, Maryland.

Larkey, Leah S., Ballesteros, Lisa, and Connell, Margaret E. (2002). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 275-282.

Larkey, Leah S., Ballesteros, Lisa, and Connell, Margaret E. (2005). Light stemming for Arabic information retrieval. Book chapter, A. Soudi, A. Van den Bosch, and Neauman, G., Editors, Kluwer/ Springer's Series on Text, Speech, and Language Technology.

Moukdad, Haidar and Large, Andrew. (2001). Information retrieval from full-text Arabic databases: can search engines designed for English do the job? Libri, 51.

Nwesri, Abdusalam F Ahmed. (2008). Effective retrieval techniques for Arabic text. Unpublished doctoral dissertation, RMT University, Melbourne, Australia.

Thabet, Naglaa. (2004, August). Stemming the Qur'an. Proceeding Semitic 04 Proceedings of the Workshop on Computational approaches to Arabic script-based languages. Retrieved February 20, 2008. http://dl.acm.org/results.cfm?h=1&cfid=42494409&cftoken=53133881.