Classification of Arabic Information Extraction methods

Abd El Salam AL HAJJAR

Institute University of Technology
Lebanese University
Lebanon
Paragraph Laboratory
University of Paris 8- Vincennes- Saint-Denis
France
abdsalamhajjar@hotmail.com

Mohammad HAJJAR

Institute University of Technology
Lebanese University
Lebanon
m_hajjar@ul.edu.lb

Khaldoun ZREIK

Paragraph Laboratory
University of Paris 8- Vincennes- Saint-Denis
France
zreik@univ-paris8.fr

Abstract

The performance of information retrieval in arabic language is very problematic due to the specific morphological and structural changes in the language. To extract information from an arabic document, the involved methods must answer the following question: "How can we find the root of the word we search". To find a word in an arabic dictionary, first you must extract the root of this word and then find this root in the dictionary. This is because the vocabulary of the arabic language is essentially built from the roots derivation. The roots are words composed of three to five consonants letters. To address these problems, several methods have been proposed. The aim of this paper is to propose a preliminary classification of arabic information extraction methods. These methods can be classified into two main categories. The first one is called "Stemmer". This category includes the following subcategories: Stemmer based on affixes, Stemmer based on translation and Stemmer based on pattern and affixes. The second is called "N-gram". This category regroups the subcategories: N-gram based on Dice's similarity coefficient and N-gram based on "Manhattan distance" dissimilarity coefficient. However, we find methods which implement the two approaches "Stemmer" and "N-gram". This work contributes to decide on the more appropriate arabic information extraction method.

Introduction

Arabic language is used by more than 330 million arabic speakers that are spread over 22 countries (Ghosn, 2003; Censure of the Internet in the arab countries, 2006). However, the performance of information retrieval in arabic language is very problematic due to the specific morphological and structural changes in the language: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short (diacritics) and long vowels, most of the arabic words contain affixes (Table 1, 2). To address these problems, several methods have been proposed.

The aim of this paper is to propose a preliminary classification of arabic information extraction methods. These methods can be classified into two main categories. The first one is called "Stemmer" which requires specific knowledge about the language (Al Ameed et al., 2005; Larkey et al., 2002; Larkey, 2005; Darwish, 2002; Chen & Gey, 2002; Kanaan et al., 2004; Thabet, 2004; Kadri & Nie, 2006; Al-Shalabi & Evens, 1998; Taghva et al., 2005, Khoja & Garside, 1999; Hammo et al., 2002). The second is called "N-gram". It based on statistical approaches to retrieve the information independently of the language complexity (Adamson George & Boreham, 1974; Suleiman Mustafa, 2004; Ahmed & Nürnberger, 2007; M. Sinane et al., 2008; Khreisat, 2006). However, we find methods which implement the two approaches "Stemmer" and "N-gram" (De Roeck & Al-Fares, 2000).

Problematic

To find a word in an arabic dictionary, first we must extract the root of this word and then find this root in the dictionary (Ibn Manzour, 2008). This is because the vocabulary of the arabic language is essentially built from the roots derivation. The roots are words composed of three to five consonants letters. The arabic language has about ten thousand roots, 85% of them are trilateral. The derivation of words is done by adding affixes (prefix, infix, or suffix) to the root according to several patterns that are around 120 (Al Kharashi, 1999). For example, let us take the root (كتب); the words (كاتب, كاتبة, مكتوب) (Table

3) are respectively derived from this root according to the patterns (فاعل, فاعلة, مفعول) (Table 4).
To extract information from an arabic document, the involved methods must answer the following question: "How can we find the root of the word we search". To answer this question, we must perform a morphological analysis. In the arabic language, this consists to identify the morphemes of a word (Stem): the affixes (prefix, infix, and suffix) and the root. A stem can be a noun, verb or particle. It can be composed of: One part (a root, for example: (ب ك ت)); Two parts (a root + a pattern, for example: (ك ت ب): root (ت ب ك) + a pattern (CuCiC where C is the consonants of the root (the radicals)); Three parts (a root + a pattern + affixes, for example: (ال ك ا ت ب و ن): root (ك ت ب) + a pattern CaCiC + affixes (prefix (al) (ال), infix (a) (ا) and the suffix (a) (ون))) (Table 2, 3).

Results

The study that we realized permits to identify several methods which address the problems of information extraction from arabic documents. We have found that these methods can be classified into three categories: "Stemmer", "N-gram", and "Stemmer and N-gram". The first category requires specific knowledge about the language. The second is based on statistical approaches to retrieve the information independently of the language complexity. In the last category we find methods which implement the two approaches "Stemmer" and "N-gram".
The diacritics and the variation of the letter forms according to its positions take an important role in the arabic reading and writing complexity and reduce the Arabic Information Extraction methods performance. To resolve these problems, the normalization phase is applied before applying these methods, the text normalization takes a character string as input and tries to remove or replace some characters under the predefined rules to converts it into a string of letters (Figure 1). Every method has the specific rules, in general a text is normalized by removing (the tatweel character "-", the diacritics and the shedda " ّ ", the punctuations, the non letters, the stop words, the specials characters, and the numbers) and replacing ("أ", "إ" and "آ" by alif bar "ا", " ى" by " ي" at the end of the words, " ة" by " ه" at the end of the words, " يء " by "ئ", ؤ by و ء , and ….) (Chen, A. & Gey, 2002; Kadri & Nie, 2006; Khreisat, 2006; Larkey et al., 2002; Larkey, 2005; Douzidia & Lapalme, 2005; Darwish, 2002).

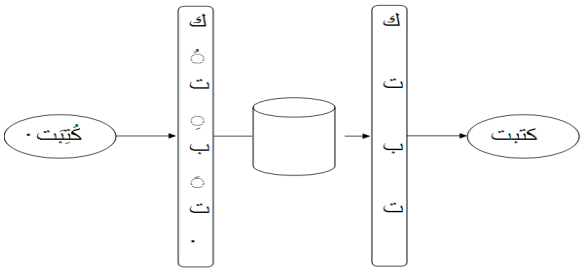

Figure 1: Normalization Process

Stemmer Category

Stemmer is an automatic process used to reduce the different morphological forms of words into common root (Stem) to improve the performance of the extraction system. This category includes the following subcategories: Stemmer based on affixes, Stemmer based on translation and Stemmer based on pattern and affixes.

Stemmer based on affixes
Several Stemmer algorithms use the predefined rules to remove the affixes (prefix, infix, suffix ...) from the word to extract the root. This category allows remarkably good information retrieval without providing correct morphological analysis.

Several algorithms have been developed (Al Ameed et al., 2005; Larkey et al., 2002; Larkey, 2005; Chen & Gey, 2002; Kanaan et al., 2004; Thabet, 2004; Kadri & Nie, 2006).
The normalization phase is applied before applying these algorithms (Replacing " أ", " إ" and " آ" by alif bar "ا", Replacing " ى" by " ي" at the end of the words, Replacing "ة" by " ه" at the end of the words, Replacing the sequence " يء " by "ئ", etc.) (Table 1).
Al Ameed, H. et al. (2005) and Larkey, L. et al.(2005) developed a light stemmer which is based on the suppression of "و" if it is initial at the beginning of the word, of the prefixes (ال, وال, بال, كال, فال, لل), and of the suffixes (ها, ان, ات, ون, ين, يه, ية, ه, ة, ي). A. Chen and F. Gey (2002) identified other sets of prefixes and suffixes.
To remove the prefixes and suffixes in the pre-defined sets, each algorithm proposes their own rules. For example, A. Chen and F. Gey (2002) apply the following rules: If the word is at least five-character long, remove the prefixes of three characters: ال, بال, فال, كال, ولل, مال, ااا, سال, لال. If the word is at least four-character long, remove the first two characters: وا, ال, سي, وس, با, لل, وم, وت, وب, لا, وي, ول, كا, فا. If the word is at least four-character long and begins with و remove the initial letter و. If the word is at least four-character long and begins with either ب or ل remove ب or ل (Table 2).
Kanaan et al. (2004) presented a new stemming algorithm to extract quadrilateral arabic roots. The algorithm starts by excluding the prefixes, and then checks the word characters starting from the last letter backward to the first one. A temporary matrix is used to store the suffix letters of the arabic word, and another matrix is used to store the roots. Algorithm checks the letters of any word, also checking whether the tested letter is included within the general standard arabic word.
The large-scale use of diacritics (ْ, ّ, ٌ, ٍ, ً, َ, ُ, ِ) (Table 1) representing short vowels are prevalent in the Qur'an. Every word, even every letter is marked with a diacritic. (For example: مُلك "reign", مَلِك "king" ...). N. Thabet (2004) proposes a new stemming approach based on a light stemming technique that uses a transliterated version of the Qur'an in western script (Table 3).
Y. Kadri and J. Nie (2006) defined that the arabic words are usually formed as a sequence of antefixes are generally prepositions joined to words at the beginning (وبال, وال, بال, فال, كال,...), prefix are usually represented by only one letter and indicate the conjugation person of verbs in the present tense (ا, ن, ي, ت...) (Table 1), core, suffixes are the conjugation terminations of verbs and they are the dual/plural/female marks for the nouns (تما, يون, تين, تان, ات, ان...), and postfixes represent pronouns attached to the end of the words(كما, هما, كن...)) (Table 2).

Stemmer based on translation

The algorithms of english Stemmer have better performance than arabic. For that, several methods use the translation technique to allow any languages (like arabic) use the Stemmer of another language (like english) to extract the root of a word. A. Chen and F. Gey (2002) have built a MT-based arabic stemmer from the arabic words found in the arabic documents, and their english translations are using the online "Ajeeb" machine translation system. They divided the arabic words into clusters based on the english translations of the arabic words. The arabic words whose english translations, after removing english stop words, are conflated to the same english stem that made from one cluster. All the arabic words in the same cluster are conflated to the same arabic word, which is the shortest arabic word in the cluster. For example: أطفالنا (our children), remove "our" is a word parasite, أطفالنا is apparent that in relation to "child". So أطفالنا is related to طفل (Table 3).

Stemmer based on pattern and affixes

Several Stemmer algorithms based on the patterns and affixes have been developed, to find roots with three letters, four and five letters, starting from verbal forms, nouns and adjectives derived from verbs (Khoja & Garside, 1999; Hammo et al., 2002; Al-Shalabi & Evens, 1998; Taghva et al., 2005; Darwish, 2002).

S. Khoja and R. Garside (1999) have proposed a method that involves removing diacritics representing vowelization, the stop words, the punctuation, the numbers, the definite article (ال), the inseparable conjunction (و), and the longest prefix and suffix. Then, the result is compared to a list of patterns. If a match is found, the characters representing the root in the pattern are extracted (Table 1, 2).

The QARAB system is developed by B. Hammo et al. (2002) is based on the Khoja Stemmer (Khoja & Garside, 1999).

R. Al-Shalabi and Mr. Evens (1998) proposed an approach of Stemmer where the first step is to remove the longest possible prefix. The three letters of the root must lie somewhere in the first four or five characters of the remainder. They checked all the possible trigrams within the first five letters of the remainder. Then, they check the following six possible trigrams: First, second, and third letters. First, second, and fourth. First, second, and fifth, etc.

K. Taghva et al. (2005) implemented a root-extraction stemmer for arabic which has a performance equivalent to the Khoja stemmer (Khoja & Garside, 1999). To implement this algorithm, they have defined several sets of the affixes (D diacritic: ... سْ, سُ, سِ, سَ, س P3 prefix of length 3:وال, وللـ . P2 prefix of length 2: لل, الـ. P1 prefix of length 1: ف, ب, ل. S3 suffix of length 3: تان, همل, تمل. S2 suffix of length 2: ان, ات, ون. S1 suffix of length 1: ي)(Table 1, 2) and several sets of models (PR4 model of length 4: فعول, فاعل, فعلة. PR53 model of length five and a root of length 3: أفعال, افتعل, تفاعل . PR54 model of length five and a root of length four: مفعلل, افعلل, تفعلل. PR63 model of length 6 and a root of length 3: استفعل مفعالة ...افتعال. PR64 model of length 6 and a root of length 4: ...افعلال, متفعلل) (Table 4). The extraction of the root in this algorithm is based on the length of the normalized stem.

N-gram Category

The second category is based on statistical approaches: N-gram. The N-gram can find if two words are semantically similar or dissimilar from the structures of characters of these words. Two words are considered similar if they have in common several substring of N characters, this is done by calculating a coefficient on these two words. The advantages of N-gram are that it does not require a preliminary knowledge of the language, does not require predefined rules, and does not require the construction of a database of vocabulary. This class gives a good result in multiple languages, even in arabic (3-gram and 4-gram). This category regroups the subcategories: N-gram based on Dice's similarity coefficient and N-gram based on "Manhattan distance" dissimilarity coefficient. In this category you can find the following classes:

N-gram based on the calculation of similarity coefficient

W. Adamson George and J. Boreham (1974) has been developed the first automatic classification technique based on the character structure of words. Dice's Similarity Coefficient is computed from the number of matching bigrams (2-gram) in pairs of character strings, and used to cluster sets of character strings (Figure 2).

H. Suleiman Mustafa (2004) assesses the performance of two N-gram matching techniques for arabic root-driven string searching: contiguous N-grams and hybrid N-grams, combining contiguous and non-contiguous.

F. Ahmed and A. Nürnberger (2007) have been presented the n-gram model which can be used to compute the similarity between two strings by counting the number of similar n-grams they share. The more similar n-grams they found between the two strings exist the more similar they are. Based on this idea the similarity coefficient can be derived. The similarity coefficient δ is defined by the following equation:

$$\delta_n(a,b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|}$$

Where α and β are the n-gram sets.

For Example: shows an example of two arabic words: الإستمرارية and استمرار (Table 3):

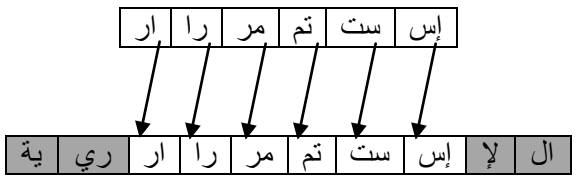

Figure 2: Bigram similarity measure between two words الإستمرارية and استمرار

M. Sinane et al. (2008) presented an approach that uses N-gram based on the word and characters. Four basic types have been explored, sometimes separately and sometimes in combination: Word, lexical root, root, and N-gram. In general, N-grams based on the stems are better than those based on words, because the N-grams based on words could have prefixes and suffixes which make more mistakes in the Similarity between the document and query.

N-Gram based on the Frequency Statistics technique

L. Khreisat (2006) presented the N-Gram Frequency Statistics technique for classifying arabic text documents. The technique employs a dissimilarity measure called the "Manhattan Distance", and "Dice's measure". A corpus of arabic text documents was collected from online arabic newspapers, 40% of the corpus was used as training classes and the remaining 60% of the corpus was used for classification. All documents, whether training documents or documents to be classified went through a preprocessing normalization phase that remove the punctuation marks, the stop words, the diacritics, and the non letters. For the training documents, the N-gram (N=3) (the trigrams of the word المودعين are: الم, مود, ودع, دعي, عين) (Table 2) was generated for each document and saved in text files. Then for each document to be classified, the N-gram frequency profile was generated and compared against the N-gram frequency profiles of all the training classes. This comparison is done by calculating Manhattan distance and Dice's measure.

Category: Stemmer and N-gram

Each of the two approaches has advantages and disadvantages, as long as the Stemmer approach depends on the language, and its morphological complexity, and always does not give the best performance... and the statistical approach N-gram is independent of language but has drawbacks in terms of synonyms. To do this, there are authors trying to merge the two approaches, in order to have a good method.

A. N. De Roeck and W. Al-Fares (2000) presented a method for arabic words sharing the same root. To implement this method "Clustering Algorithm", depends on two stages which is called "Two-Stage". In first step, they applied the Light Stemming to remove affixes, the second step is based on the Adamson algorithm with some modifications. Each bi-gram assigned a weight (0.25 for bi-gram containing low letter, 0.5 for bi-gram containing the non-low letter, 1 for all other bi-gram).

Conclusion

In this article we have proposed a first classification of arabic information extraction methods into three categories: Stemmer, N-gram, and "Stemmer" and "N-gram". In the stemmer category we find the following subcategories: Stemmer based on affixes, Stemmer based on translation, and Stemmer based on pattern and affixes. In the N-gram category we find the following subcategories: N-gram based on Dice's similarity coefficient and N-gram based on "Manhattan distance" dissimilarity coefficient. However, we find a method which implements the two approaches "Stemmer" and "N-gram". The next step will be the making of a detailed comparative study of the early described categories. This study will cover mainly the following topics: performances, stabilities, usability, advantages, and disadvantages. Another possible extension of the present work is to test these categories in similar conditions. These studies and tests will permit to designate the more appropriate arabic information extraction method or to propose a new one.

Letter	Transcription	Writing			Letter	Transcription	Writing		
		At Begin	In Middle	At End			At Begin	In Middle	At End
ً	Tanween Fatha				ر	Raa	ر	ـر	ـر
ٌ	Tanween Dama				ز	Thal	ز	ـز	ـز
ٍ	Tanween Kasra				س	Seen	سـ	ـسـ	ـس
َ	Fatha				ش	Sheen	شـ	ـشـ	ـش
ُ	Dama				ص	Saad	صـ	ـصـ	ـص
ِ	Kasra				ض	Daad	ضـ	ـضـ	ـض
ْ	Sokon				ط	T'aa	طـ	ـطـ	ـط
ّ	Shedda				ظ	Zha	ظـ	ـظـ	ـظ
~	Maada				ع	Ain	عـ	ـعـ	ـع
ء	Hamza				غ	Jain	غـ	ـغـ	ـغ
ا	Alef				ف	Faa	فـ	ـفـ	ـف
إ	Alef with Hamza on bottom				ق	Qaf	قـ	ـقـ	ـق
أ	Alef with Hamza on top				ك	Kaf	كـ	ـكـ	ـك
آ	Alef with Maada				ل	Lam	لـ	ـلـ	ـل
ب	Baa	بـ	ـبـ	ـب	م	Meem	مـ	ـمـ	ـم
ة	Taa Marbouta	X	X	X	ن	Noon	نـ	ـنـ	ـن
ت	Taa	تـ	ـتـ	ـت	ه	Haa	هـ	ـهـ	ـه
ث	Tha	ثـ	ـثـ	ـث	و	Waw	و	ـو	ـو
ج	Jeem	جـ	ـجـ	ـج	ؤ	Hamza on waw	X	ـؤ	ـؤ
ح	H'a	حـ	ـحـ	ـح	ى	Alif Makzora	X	X	
خ	Khaa	خـ	ـخـ	ـخ	ي	Yaa	يـ	ـيـ	ـي
د	Dal	د	ـد	ـد	ئ	Hamza on yaa	ئـ	ـئـ	ـئ
ذ	Zain	ذ	ـذ	ـذ					

Table1: Arabic diacritics and letters transcription. Empty case means no writing change in the corresponding letter and position. X-case means no existing of the corresponding letter

Affix	Transcription		Affix	Transcription
اال	Alef Alef Lam		يون	Yaa Waw Noon
ات	Alef Taa		ات	Alef Taa
ال	Alef Lam		الم	Alef Lam Meem
ان	Alef Noon		ان	Alef Noon
بال	Baa Alef Lam		با	Baa Alef
تم	Taa Meem		بال	Baa Alef Lam
دعي	Dal Ain Yaa		تان	Taa Alef Noon
سْ	Saa Sokon		تما	Taa Meem Alef
سال	Saa Alef Meem		تين	Taa Yaa Noon
عين	Ain Yaa Noon		سَ	Seen with Fatha
فال	Faa Alef Lam		سي	Seen Yaa
كال	Kaf Alef Lam		فا	Faa Alef
كن	Kaf Noon		كال	Kaf Alef Lam
لا	Lam Alef		كما	Kaf Meem Alef
لال	Lam Alef Lam		وال	Waw Alef Lam
لإ	Lam Alef with Hamza on bottom		وبال	Waw Baa Alef Lam
لل	Lam Lam		بية	Yaa Yaa Taa Marbouta
مال	Meem Alef Lam		ار	Alef Raa
مود	Meem Waw Dal		إس	Alef with Hamza on bottom Seen
ها	Haa Alef		تم	Taa Meem
هما	Haa Meem Alef		تمل	Taa Meem Lam
همل	Haa Meem Lam		را	Raa Alef
وال	Waw Alef Lam		ري	Raa Yaa
وب	Waw Baa		سُ	Seen with Dama
وت	Waw Taa		ست	Seem Taa
ودع	Waw Dal Ain		كا	Kaf Alef Lam
وس	Waw Seen		مر	Meem Raa
ولل	Waw Lam Lam		وا	Waw Alef
وم	Waw Meem		ول	Waw Lam
ون	Waw Noon		ون	Waw Noon
وي	Waw Yaa		ية	Yaa Taa Marbouta
ين	Yaa Noon		يه	Yaa Haa

Table2: Arabic Affix Transcription cited in this paper and their transcription

Word	Transcription	Translation
كاتب	Kateb	Writer
العدوان	Aleidwan	Attack
الحرب	Alharb	War
المعركة	Almaaraka	Battle
كتب	Kataba	Write
استمرار	Estemrar	Continuity
طفل	Tofol	Child
كاتبة	Kateba	Writer
مَلِك	Malek	King
مكتوب	Maktob	Written
الإستمرارية	Estemrareya	Continuities
المودعين	Al modeoon	Depositors
أطفالنا	Atfalona	Our children
مُلك	Muluk	Had
الكاتبون	Alkateboun	Writers

Table3: Arabic words cited in this paper, their transcription, and their translation

Pattern	Transcription
افتعل	Eftaala
افعلل	Afaalal
أفعال	Afaal
فاعلة	Faaela
فعول	Faaol
مفعول	Mafool
افتعال	Efteaal
متفعلل	Moftaeel
مستفعل	Mostafeel
مفعالة	Mafaala
مفعلل	Mafaalal
استفعل	Iestafaal
افعلال	Afaalal
تفاعل	Tafaool
تفعلل	Tafaalal
فاعل	Faeel
فعلة	Faaela

Table4: A sample of arabic pattern cited in this paper and their transcription

Acknowledgements

This work has been done as a part of the project “Arabic Web Intelligence” supported by the Lebanese National Centre of Scientific Research (CNRSL).

Bibliographical References

Adamson George, W. & Boreham, J. (1974). "The use of an association measure based on character structure to identify semantically related pairs of words and document titles", Information Storage and Retrieval, Vol. 10, pp 253-260, 1974.

Ahmed, F. & Nürnberger, A. (2007). “N-grams Conflation Approach for Arabic”, ACM SIGIR Conference, Amsterdam, 27 July.

Al Ameed, H. & Al Ketbi, S. & Al Kaabi, A. & Al Shebli, K. & Al Shamsi, N. & Al Nuaimi, N. & Al Muhairi, S. (2005). "Arabic Light Stemmer: A new Enhanced Approach", The Second International Conference on Innovations in Information Technology (IIT’05).

Al Kharashi, I. (1999). "A Web Search Engine for Indexing, Searching and Publishing Arabic Bibliographic Databases".

Al-Shalabi, R. & Evens, M. (1998). "A Computational Morphology System for Arabic", Proceedings of COLING-ACL, New Brunswick, NJ.

Censure of the Internet in the Arab countries (2006). Human Rights Tribune - Geneva 2006 - www.humanrights-geneva.info.

Chen, A. & Gey, F. (2002). "Building an Arabic stemmer for information retrieval".TREC-11 conference.

Darwish, K. (2002). "Building a Shallow Arabic Morphological Analyzer in One Day". In The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, USA.

De Roeck, A. N. & Al-Fares, W. (2000). "A morphologically sensitive clustering algorithm for identifying Arabic roots". In Proceedings ACL-2000. Hong Kong.

Douzidia, F. & Lapalme, G. (2005). "Un système de résumé de textes en arabe", 2ème Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la langue, Alger.

Ghosn, Z. (2003). government Web sites in the Middle East in 2003, The Arab Advisors Group.

Hammo, B. & Abu-Salem, H. & Lytinen, S. & Evens, M. (2002). "A Question Answering System to Support the Arabic Language". Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages Philadelphia, Pennsylvania Pages: 1 – 11.

Ibn Manzour (2008). Lisan Al-Arab. www.muhaddith.org.

Kadri, Y. & Nie, J. (2006). "Effective Stemming for Arabic Information Retrieval" in proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni.

Kanaan, G. & Al-Shalabi, R. & Jaarn, J. & Al-Kabi, M. & Hasnah, A. (2004). "A New Stemming Algorithm to Extract Quadri-Literal Arabic Roots".

Khoja, S. & Garside, R. (1999). "Stemming Arabic text", Computing Department, Lancaster University, Lancaster, ww.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, 1999.

Khreisat, L. (2006). "Arabic Text Classification Using N-gram Frequency Statistics a Comparative Study". The 2006 International conference on Data Mining Part of the 2006 World Congress in Computer Sciences DMIN 2006: 78-82.

Larkey, L. S. & Ballesteros, L. & Connel, M. E. (2002). "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis", in Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275 – 282.

Larkey, L. & Ballesteros, L. & Connell, M. (2005). "Light Stemming for Arabic IR" Arabic Computational Morphology: Knowledge-based and Empirical Methods, A.Soudi, A. van en Bosch, and Neumann, G., Editors. Kluwer/Springer's series on Text, Speech, and Language Technology.

Sinane, M. & Rammal, M. & Zreik, K. (2008). "Arabic documents classification using N-gram", Conference ICHSL6, Toulouse.

Suleiman Mustafa, H. (2004). "Character contiguity in N-gram based word matching: the case for Arabic text searching". Information Processing and Management.41 (4), 819-827.

Taghva, K. & Elkoury, R. & Coombs, J. (2005). "Arabic Stemming without a root dictionary". International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I pp. 152-157.

Thabet, N. (2004). "Stemming the Qur'an" WORKSHOP ON Computational Approaches to Arabic Script-based Languages, University of Geneva, Geneva, Switzerland, August 28.