ISSN: 0067-2904

New Root-based Stemmer for Arabic Language

Inas Ali
Department of Computer Science, College of Science, University of Baghdad , Baghdad, Iraq

Abstract

Importance of Arabic language stemming algorithm is not less than that of other languages stemming in Information Retrieval (IR) field. Lots of algorithms for finding the Arabic root are available and they are mainly categorized under two approaches which are light (stem)-based approach and root-based approach. The latter approach is somehow better than the first approach. A new root-based stemmer is proposed and its performance is compared with Khoja stemmer which is the most efficient root-based stemmers. The accuracy ratio of the proposed stemmer is (99.7) with a difference (1.9) with Khoja stemmer.

Keywords: Information Retrieval, Arabic Language, Stemming, Root-based, Stem-based.

محلل جديد للجذوع المعتمد على الجذر الجديد للغة العربية

ايناس علي
قسم علوم الحاسوب ، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

اهمية خوارزميات ايجاد الجذور للغة العربية ليست اقل من أهمية خوارزميات ايجاد الجذور للغات الاخرى في مجال استرجاع المعلومات. عدة خوارزميات لإيجاد الجذر العربي متوفرة وهي تصنف بشكل رئيسي تحت نهجين هما النهج المعتمد على الجذر والنهج المعتمد على الجذع. النهج الثاني هو الافضل نوعا ما من النهج الاول. تم اقتراح محلل الجذور المعتمد على الجذر الجديد وتمت مقارنة أدائه مع محلل جذور خوجة والذي يعد افضل محلل جذوع المعتمد على الجذر.نسبةالدقة للمحلل المقترحهي (99.7) مع فرق (1.9) مع محلل جذوع خوجة.

Introduction

IR is basically an issue of deciding which documents in a corpus should be regained to meet a user's information need which is represented by a query, and contains search term(s), in addition to some information such as the relatively importance. Thus, the retrieval decision is possessed by finding the similarity between the query terms with the index terms appearing in the document. It could have dual decision relevant or non-relevant or it could be appropriate to a limited extent that the document would have to query. Improving efficiency of the search is the important target behind building of any stemmer so an information retrieval system can match user's queries with relevant documents. [1]. The stemming process constitutes word morphological analysis based on the language used in order to get the words' roots or stems [2], usually by removing affixes [3], to represent the documents as well as to act as indexes to the documents for efficient and effective retrieval [2].

Email: smart_girl8120@yahoo.com

Arabic is a major international language, spoken in more than 23 countries, and the lingua franca of the Islamic world. The number of Arabic-speaking Internet users has grown over nine-fold in the Middle East [3]. Since Arabic language is a highly inflected language and has a complex morphological structure than English [4] and due to the vast growth of the Arabic internet content [1], it requires superior stemming algorithms for effective IR [4].

Arabic text is written from right to left in a cursive, consonantal script that has 28 characters [3]. Arabic Root is the base verb form which can be trilateral, which is the overwhelming bulk of Arabic words, and to a minor amount, quadrilateral, pentaliteral, or hexaliteral, each of which generates further verb forms and noun forms by the addition of derivational affixes [2]. Surface forms of Arabic words comprise two or more morphemes: a root with a semantic meaning, and a pattern with syntactic information. There are around 400 distinct patterns in Arabic. The most well-known pattern is (فعل Fa Aa La), which is often used to generically represent three-letter root words [3]. Affixes in Arabic are: prefixes that attached at beginning of the words, suffixes (or postfixes) that are added at the end, and infixes are found in the middle of the words [2].

Stemming algorithms can be used in Arabic text pre-processing to reduce words to their root/or stems [4]. Although they experience from many problems, they have been in use in many IR systems [1]. Arabic stemming algorithms can be categorized, according to the desired level of analysis, as either stem-based or root-based algorithms [4]. Light stemming, like the stemmer introduced by Leah et al [1], which is restricted to the removal of prefixes and suffixes. It does not dealing with patterns or infix [5]. Root-based stemmer like Khoja's stemmer eliminates the longest suffix and the longest prefix. It then matches the resting word with verbal and noun patterns, to extract the root [6, 7]. Stemming algorithm proposed by Khoja is one of well-known Arabic stemmers [8].

Many stemming researches were introduced to reduce words to its original root. Al-Fedaghi and Al-Anzi suggested an algorithm that tries to find the root of the word by matching the word with different patterns with all possible affixes attached to it, and does not removed any prefixes or suffixes. In this context and with other technical Al-Shalabi morphology system applied several algorithms to find the roots and patterns. This algorithm searched the root in the first five letters of the word by removing the longest prefix. As mentioned earlier, Khoja contributes with a very important algorithm [7]. The proposed new root-based stemming algorithm for Arabic language has shown better results.

Arabic Language characteristics

Arabic is a Semitic language, and a descendant of Proto-Semitic [3]. The Arabic Language is the 5th broadly used language in the world. It is pronounced by more than 422 million people as a first language and by 250 million as a second language [8]. Arabic language opposes from English and European languages and the morphological representation of Arabic is rather complicated because of the morphological variation and the agglutination phenomenon [9]. Arabic alphabet has 28 letters, see Table- 1, and may be extended up to 90 by adding vowels, marks and shapes [1]. There is no upper or lower state for Arabic letters like English letters. The letters (ا, و, ي) are vowels. The direction of writing in Arabic is from right to left [8]. The grammatical system of Arabic language is depend on a root-and-pattern structure and regarded as a root-based language with not more than 10000 roots and 900 patterns [4].

Table 1- Arabic Letters and their Pronunciation in English

Letter	أ	ب	ت	ث	ج	ح	خ
Pronunciation	alif	baa	taa	thaa	jiim	haa	kha
Letter	د	ذ	ر	ز	س	ش	ص
Pronunciation	daal	thaal	raa	zaay	siin	shiin	saad
Letter	ض	ط	ظ	ع	غ	ف	ق
Pronunciation	daad	taa	thaa	ayn	ghayn	faa	qaaf
Letter	ك	ل	م	ن	ه	و	ي
Pronunciation	kaaf	laam	miim	nuun	ha	waaw	yaa

Arabic Word Derivation –Word – Root – Pattern - Affixes

As described in previous section, Arabic language is a highly inflected language, since it uses many inflectional forms [5]. Arabic is mainly roots and templates (pattern) dependent in the formation of words [2]. Arabic *words* have many forms, and are formed by applying vowel patterns to roots that have three (7198) or four (3739) and in rare cases five letters (295) [3] (which is about 11,347 root [9]) listed in (لسان العرب LisaanulArab), one of the most respected Arabic dictionaries [3]. A *root* is the base form of a word which cannot be further analysed without the loss of the word's identity [4]. Arabic roots consonants might be altered or removed during the morphological process [2]. Root generates increased verb forms and noun forms by the addition of derivational affixes.

Arabic *patterns* are portion of the Arabic grammar. They are formed based on the Arabic root [4]. The roots join with various vowel patterns to form simple nouns and verbs to which affixes can be attached for more complicated derivations. Patterns play a significant role in Arabic lexicography and morphology. Each root can canonically combine with orthographically distinct patterns to form another words, see an example in Table-2 of a root (علمteach). The sign (*) in Table-2 refer to Feminine [9].

An Arabic *affix* is a morpheme that [4], as Arabic is written from right to left [3], can be added before (prefix) or after (suffix), or inserted inside (infix) a root or a stem to form new words or meanings [4], See an example in Table-3. Generally, ten letters are used in Arabic affixes and they are grouped in the acronym (سألتمونيها)[3]. Arabic prefixes and suffixes do not follow a systematic standard for their attachment to Arabic words [2].This affixes list is usually built based on the language morphology and statistical analysis of Arabic text [3].

Table 2 -Some of different words derived from the root (علم Teach)

Word	علم	معلم	معلمة	معلمان	معلمتان	معلمون	معلمات	يتعلم	تتعلم	تعليم
Pattern	فعل	مفعل	مفعلة	مفعلان	مفعلتان	مفعلين	مفعلات	يتفعل	تتفعل	تفعيل
Meaning	Teach	Teacher	Teacher*	2 Teacher	2 Teacher*	Teachers	Teachers*	Learn	Learn*	Teaching

Table 3- Word with Affixes

Word	**Prefix(s)**	**Infix(s)**	**Suffix(s)**	**Root**
الكتابة (Writing)	ا، ل	ا	ة	كتب (Write)

Stemming

Stemming is an essential process used in many fields of natural language processing like IR systems, Web search engines, Question Answering Systems, textual classifiers, etc.. Stemming primary task is to standardize words; which can be obtained by reducing each word [5], remove all possible affixes [4], to its base (root or stem) [5]. Stemming is used in IR systems in order to improve retrieval effectiveness [2]. Stemming algorithms for some languages have been published and applied in building of IR systems, among which for English is the popular Porter’s algorithm, for French we have Savoy’s algorithm, and for the Malay language we have Fatimah’s et al [2].

In highly inflected languages such as Arabic, stemming is considered one of the most important factors to improve retrieval effectiveness of Arabic IR systems [3]. Very little research has been carried out on Arabic text. The nature of Arabic text is inconsistent than English text, and pre-processing of Arabic text is more challenging stage in text categorization particularly and text mining generally. The effect of the preprocessing tools on Arabic text categorization is an area of research [7]. Most current stemmers eliminate affixes without checking whether the removed letters are actually affixes. This can happen in any language, but it is a major problem for Arabic [3].

Distinction in morphological properties among world's languages is high, and stemmers are language dependent as mentioned earlier. From here, it is expected to see distinct stemmers for the Arabic language that are different from the English ones. Arabic is very rough to stem [1]. Arabic stemming is in fact a process of morphological analysis utilized for the word in order to extract the correct stem [2]. Stems or roots are useful index terms for Arabic. Their benefits were clearer than for

English, as the Arabic language is a root based language [1]. For example, in Table-2, the root of the Arabic word (معلمون teachers) is (علمteach). [7].

Arabic Stemmers Approaches

Approaches for the developing of Arabic stemmers are restricted because of its complicated structure. These approaches are mainly dependent on the understanding of the Arabic morphology [2]. Arabic stemming algorithms are classified under two main approaches: Root-Based Approach; Stem-Based Approach [7].

a. *Root-Based Approach (Morphological analysis):*

To elicit the root of a given Arabic word with this approach uses morphological analysis [7]. Many morphological analyzers have been improved for Arabic; little of them receiving a standard IR evaluation. Most such morphological analyzers attain the root, or any number of possible roots for each word. A superior root-based stemmer is Khoja stemmer [5].

b. *Stem-Based Approach*

Light stemming refers to the process of stripping off a small set of prefixes and/or suffixes without trying to deal with infixes [4] (this causing a serious issue in Arabic documents stemming since it is hard to differentiate between root characters letters and affix letters [5]) or recognize patterns and find roots [4]. One of best stem-based stemmer is Larkey stemmer [7].

Khoja's Stemmer

This aggressive stemmer eliminates the longest suffix and the longest prefix. It then matches the remaining word with verbal and noun patterns, to extract the root [8]. This stemmer benefits from several linguistic data files such as a list of all diacritic characters, punctuation characters, definite articles, and 168 stop-words. The Khoja algorithm (illustrated in Figure- 1) initially removes suffixes, infixes and prefixes and uses pattern matching to extract the roots, but suffered from problems especially with nouns. [6]. Return to [10] for more detail about this stemmer.

```
Algorithm: Khoja Root-Based Stemming Algorithm
Purpose: Stemming Arabic Words
Input:
• Dataset
• Stop-word list
• Assets and patterns files
Output: Stemmed Dataset
Procedure:
1. Replace initial آ، أ، إ with ا.
2. Stop-words removal.
3. Remove punctuation, non-letters and diacritics.
4. Remove definite articles from the beginning of the word.
5. Remove the letter (و) from the beginning of the word and (ة) from the end of the word.
6. Remove prefixes and suffixes
7. Comparing the resulting word to patterns stored in the dictionary, if the resulting root is meaningless the original word is returned without changes.
```

Figure 1- Khoja Stemmer Algorithm

Proposed New Root-Based Stemmer

This research proposed a new Arabic root-based stemmer that makes use from few of linguistic data files used by Khoja Stemmer such as a list of all diacritic characters, punctuation characters, 168 stop-words, and some of its used affixes. It keeps the number of patterns to minimum by eliminating, if possible, prefixes and suffixes from patterns that having them.

The stemmer firstly matches a word against predefined set of patterns in order to extract its root, while Khoja stemmer initially removes affixes. Next if no match found, it repeatedly eliminates the suffix and the prefix separately with different lengths (not stop with one elimination of prefix and suffix) firstly to get rid of the still remaining affixes and secondly to find another roots if any available even after finding first root (for example, roots of word (يزعمون they claim) are (زعم allege – وزع distribute)), while Khoja stemmer removes affixes with the longest length with the same step one time only and satisfied with finding one root. The stemmer preserves each word after elimination of each affix while Khoja stemmer does not.

The proposed new stemmer summarized as follows:

1. Document Preprocessing.

From the document that to stem its words

a. Eliminate

i. All Stop-words.

ii. Punctuation, non-letters, other languages' letters and numbers.

iii. Diacritics of all words.

b. Replace initial آ، أ، إ with ا.

2. For each word in Document

c. Checking word[1]'s length due to number of its letters

i. If length = Two letters, then check two-length roots list

- If root found, add it to *List of roots*[2] of word
- Go to step #2.b.iii.

ii. If length = Three letters, go to step #2.b.

iii. If length > Three letters

- Remove prefix with length one letter[3], see list of proposed prefixes in Table-4, and save the modified word in *List of words*.
- Remove suffixes with length one letter, see list of proposed suffixes in Table-4, and save the modified word(s) in *List of words*.

iv. Comparing the word to patterns that equal its length, see list of the used patterns in Table-5

i. If the root not found, go to step #2.b.iii.

ii. If the root found, add it to *List of roots* of word.

iii. If *List of words* is not empty, extract a word from it and repeat steps from step #2.a.

iv. If *List of words* is empty

- If *List of roots* is empty, no root found for word in document.
- If *List of roots* is not empty, root(s) found for word in document.

v. Go to step 2.

Table 4- Used Prefixes and Suffixes

Prefixes	ا، ب، ت، س، ف، ل، م، ن، و، ي
Suffixes	ا، ت، ة، ك، م، ن، ه، و، ي

Table 5- Used Arabic Patterns

Patterns	فعل	فعول	فعال	فاعل	فعلى	افتعل	تفتعل
	افتعال	يفتعل	مفتعل	تفعيل	مفعول	فواعل	فعائل
	فعالة	فعالى	فعلان	فعلاء	مفاعيل	افعلاء	افعلال

Experimental Results

The proposed new stemmer has been examined with 1390 words with different Arabic patterns illustrated in Table-5. The stemmer's accuracy and fail ratios have been calculated as follows:

$$\text{Accuracy Ratio} = \frac{\mathit{Correctly\,Stemmed\,Words'Number}}{\mathit{Total\,Words'Number}} \qquad (1)$$

$$\text{Fail Ratio} = \frac{\mathit{Not\,Stemmed\,Words'Number}}{\mathit{Total\,Words'Number}} \qquad (2)$$

[1] The term *word* in step #2.a and step #2.b may refer either to a *word in document* or a *modified word*.

[2] There is a *List of roots* for each word in document after preprocessing.

[3] The proposed stemmer also examined with *affixes of length 2 and above*, but roots for words are mostly not found so it is not mentioned in steps of the proposed stemmer.

In Table- 6 an example that illustrates the difference between stemming steps of the proposed and Khoja stemmers. Performance of the new stemmer has been compared with Khoja Stemmer and has shown better results as illustrated in Table- 7. A sample of words that have been stemmed with both of those stemmers showed in Table- 8.

Table 6- Comparison between Steps of the Proposed and Khoja Stemmers

Proposed Stemmer		**Khoja Stemmer**	
Word	العاب	**Word**	العاب
Pattern Matching	No match	**Affixes Elimination**	لعا
Affixes Elimination	لعاب، العا	**Pattern Matching**	No match
Pattern Matching	لعاب فعال	**Root**	Not found
Root	لعب		

Table 7- Results of New Root-based Stemmer and Khoja Stemmer

Stemmer	**New Root-based Stemmer**	**Khoja Stemmer**
Accuracy Ratio	99.7	97.8
Fail Ratio	0.3	2.2

Table 8- Sample of words stemmed by the proposed and Khoja Stemmers

Word	**New Root-based Stemmer Root**	**Khoja Stemmer Root**	**Correct root**
ساجد	سجد	سجد	سجد
تكتبان	كتب	كتب	كتب
ثبته	وثب	وثب	ثبت
مستور	ستر	سور	ستر
يشربن	شرب	شربن	شرب

Conclusions

- Keeping number of patterns to minimum helps in reducing stemming errors.
- A word is first compared with patterns before any affixes elimination to avoid the loss of an original genuine letter of a word.
- Prefixes and suffixes to be used in the proposed stemmer are of length one letter. This give a better stemming results than when using affixes of two letters length and above because those long affixes may contain an original root letter so stemmer could not find a root of a word.
- Separating elimination of a prefix and a suffix from a word to preserve an original word's letter if it was not an additional letter. For example a root of word (الـوانColors) is (لـونColor) could not be find if front and end letters eliminated at same time.
- Preserve every word after each affix elimination for possibility of finding more than one root to the original word.
- Do not eliminate the prefix (وWaw) and the definite article (الـ Al) from original word, unlike Khoja stemmer, since they will be eliminated gradually through step of prefix elimination. They may be an original word's letter and so stemmer cannot find the root. For example a root of word (الـوانColors) is (لـونColor) and a root of word (وصـول Arrival) is (وصـل Arrive).

References

1. Otair M. **2013**. Comparative Analysis of Arabic Stemming Algorithms. *International Journal of Managing Information Technology (IJMIT)*. vol.5, No. 1-12.
2. SEMBOK T., ABU ATA B. and BAKAR Z. **2011**. A Rule-Based Arabic Stemming Algorithm. Proceedings of the European Computing Conference. France. Pp. 392-397.
3. Nwesri A. **2008**. Effective Retrieval Techniques for Arabic Text. Ph.D thesis. School of Computer Science and Information Technology, Science, Engineering, and Technology Portfolio, RMIT University, Melbourne, Victoria, Australia.
4. Al Ameed H., Al Ketbi S., Al Kaabi A. and Al Muhairi S. **2005**. Arabic Light Stemmer: Anew Enhanced Approach. The Second International Conference on Innovations in Information Technology (IIT'05)

5. AL-OMARI A. and ABUATA B. **2014**. ARABIC LIGHT STEMMER (ARS). *Journal of Engineering Science and Technology*. vol. 9, pp. 702 – 716.
6. Almusaddar M. **2014**. Improving Arabic Light Stemming in Information Retrieval Systems. MSC Thesis. Computer Engineering Department, Faculty of Engineering, Research and Postgraduate Affairs, Islamic University, Gaza, Palestine.
7. Hadni1 M., Ouatik S. and Lachkar A. **2013**. EFFECTIVE ARABIC STEMMER BASED HYBRID APPROACH FOR ARABIC TEXT CATEGORIZATION. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*.vol.3, no.4.
8. Brahimi B., Touahria M. and Tari A.**2016**. Data and Text Mining Techniques for Classifying Arabic Tweet Polarity. *Journal of Digital Information Management*. vol. 14, no. 1, pp.15-25.
9. Kana G., Al-Shalabi R., Ababneh M., and Al-Nobani A. 2012. Building an Effective Rule- Based Light Stemmer for Arabic Language to Improve Search Effectiveness. *The International Arab Journal of Information Technology* . vol.9, no.4, pp.368-372.
10. Khoja S. **2008**. Khoja Stemmer. Available at: http://zeus.cs.pacificu.edu/shereen/research.htm.