تقنية جديدة لاشتقاق جذور اللغة العربية

د . خليفة عبدالرؤف نصرات ـ كلية التربية الزاوية ـ جامعة الزاوية

ملخص الورقة البحثية

اللغة العربية تتوسع في العالم يوما بعد يوم. نما وجود اللغة العربية على الإنترنت بحوالي 6.091٪ في الخمسة عشر عامًا الماضية، وهو أعلى نمو لأهم عشر لغات على الإنترنت، ويزداد عدد المستندات العربية بسرعة. لذلك من الضروري تحسين إحدى الأدوات المهمة (الاشتقاق) لتقنيات استرجاع المعلومات العربية. يتفق العديد من الباحثين على فوائد الاشتقاق لتعزيز الكفاءة في نظام استرجاع المعلومات. نقدم في هذا البحث نهج اشتقاق جديد (قائم على الجذر) للغة العربية، تعتمد هذه التقنية على إزالة البادئة واللاحقة ومطابقتها مع الأنماط العربية. يُظهر تنفيذ وتقييم هذا الجذع التحسن في الدقة بالنسبة للخوارزميات الأخرى.

New root-based stemming approach for Arabic language

Dr . Khalifa Abdullrauof Nusrat
k.nusrat@zu.edu.ly

Abstract

The Arabic language is expanding in the world day after day. The presence of the Arabic language on the internet grew around 6.091% in the last fifteen years, it is the highest growth of the ten top online languages, and the number of Arabic documents increases rapidly. Therefor it necessity to improve the one of important tools (stemming) for the Arabic Information Retrieval (IR) techniques. Many researchers agree on the benefits of stemming to enhance the efficiency in information retrieval system. In this paper we presents a new (root-based) stemming approach for Arabic language this technique is based on prefix and suffix removal and matching with Arabic patterns. The

implementation and evaluation of this stemmer shows the improvement in the accuracy relative to other algorithms.

Keywords: information retrieval, Arabic language, Stemming, Prefix, Suffix, NLP.

1. Introduction

Arabic language is the fifth most widely spoken language in the world. It belongs to the Semitic family, so it differs from the Indo-European languages morphologically, semantically, and syntactically. The Arabic alphabet contains twenty-eight letters, always written from right to left in cursive form. Diacritical marks (harakat) (tashkiil) appear either above or below the letters, and play an essential role in many cases in distinguishing semantically and phonetically between two identical words with the same characters, but with different diacritics[1]. Recently, the big growth of the Arabic internet content has raised up the need to an effective stemming techniques for Arabic language. All Arabic words belong to three main categories: noun, verb or particle. Around 64% of Arabic words are derived from trilateral verbs (three consonants), but there are also bilateral verbs (two consonants), quadrilateral verbs (four consonants), and pentaliteral verbs (five consonants). Naturally these verbs represent the roots for which stemming algorithms typically search. This stemming process excludes words derived from nouns and particles [2].

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 [3]**.** The

main goal behind building any stemmer is to improve the search effectiveness so an IR system can match user's queries with relevant documents. Users form their query terms in many different formats. However, they are looking for the same thing [10]. Now an IR system should be able to translate all these forms that have the same meaning to a standard form. Thus grouping all these different formats in a singular or standard format on both users' queries and index terms sides.

In Arabic language has two complex paradigms derivational and inflectional morphology, which are based on roots and patterns. The interaction between roots and patterns has intrigued lexicographers and morphologists for centuries. IR tools are challenged by these characteristics of morphological structure of Arabic. Unfortunately, several stemmers remove blindly the most frequent affixes from Arabic words [5], which result high stemming error ratio and produce incorrect Arabic words [6].

Arabic language is based on set of roots, a root is the base form of a word, which cannot be further analyzed without the loss of the word's identity, or it is that part of the word left when all the affixes are removed. additional the root is basic form of word from which many derivations can be obtained by attaching certain affixes so we produce many nouns and verbs and adjectives from the same root .Table -1 shows an example root "كتب" and a set derivations can be obtained from this root [7].

Table -1. Some derivations of the root "كتب"

Writes	Library	writer	Written	Write
يكتب	مكتبة	كاتب	مكتوب	كتابة

All nouns and verbs are generated from a set of roots which is about 11,347 root distributed as follow [18].

- 115 : Two character roots (and these roots have no derivation from them).
- 7198: Three character roots.
- 3739 : Four character roots.
- 295: Five character roots

Arabic stemming algorithms developed to reduce surface words to their base (stems or roots) and it can be ranked, according to three category, as root-based approach (Khoja [7]); stem-based approach (Larkey [8]); and statistical approach (N-Garm [4]). Although many stemming methods have been developed for Arabic language, they suffer from many problems [12]. In this paper, we introduce a new stemming technique for Arabic words. The rest of this paper is organized as follow. A brief review on related work in stemming Arabic word is presented in the next section. Section 3 describes the proposed Stemming technique for the Arabic Language. Section 4 presents the experimental result .finally section 5 conclusion this work.

2. Related work

Arabic stemming is a technique that aims to find the stem or lexical root for words in Arabic natural language, by eliminating affixes stuck to its root, because an Arabic word can have a more complicated form than any other language with those affixes. The stem is simply defined as a word without a prefix or/and suffix. For example, stem of the Arabic word (" المعلمون ", the teachers) is (" معلم " ,teacher). Arabic stemming algorithms are classified under three categories Root-Based Approach (Khoja Stemmer).

Stem-Based Approach (Larkey Light Stemmer).Statistical Approach (often involves N-Gram) [11]**.**

2.1 Khoja Stemmer

Root-based stemmers: the main goal of this type of stemmers is to separate the root of a specific surface word. The prefixes and suffixes are removed and they followed by the extraction of root. The residual stem is then compared with the similar patterns and length to extirpate the root as depicted in Table -2 by projecting the matching related letters [14].

The weaknesses of root-based stemmers are Increases word ambiguity; all possible patterns are not involved

Table -2. Extracting root from stem by comparing patterns

Pattern	ف ع ا ل ة
Surface word	د ر ا س ة
Root	درس

In 1999 Khoja and Garside produced an effective root-extracting stemmer. (Khoja

Stemmer) [9] . The Khoja stemmer follows this procedure:

1. Remove diacritics representing vowelization.
2. Remove stopwords, punctuation, and numbers.
3. Remove definite article " ال"
4. Remove inseparable conjunction " و"
5. Remove suffixes, prefixes.
6. Match result against a list of patterns. If a match is found, then extract the characters in the pattern representing the root.

7. Match the extracted root against a list of known "valid" roots.
8. Replace weak letters " و , ا, ي" with "و"
9. Replace all occurrences of hamza " ء , ى , ؤ" with "أ"
10. Two letter roots are checked to see if they should contain a double character. If so, the character is added to the root.

2.2 light stemmer

Light stemming refers to the process of stripping off a small set of prefixes and/or suffixes without trying to deal with infixes this causing a serious issue in Arabic documents stemming since it is hard to differentiate between root characters letters and affix letters or recognize patterns and find roots. One of best stem-based stemmer is Larkey stemmer [11].

In 2002 Larkey et al proposed the Light stemmer. The purpose of the Stem-Based Approach or Light Stemmer is to remove the most frequent suffixes and prefixes. However, this algorithm changes the form of the words in some cases [8].The light stemmer, light10, strips off initial "و " (and), definite articles:

(" لل" "وال" "ال " "كال " "فال" "بال")

and suffixes:

("ي" , "ة" , "ه" , "ية" , "يه" , "ين" , "ون" , "ات", "ان" , "ها")

("ها" "ان" "ون" "ات" "ين" " يه" "ية" " ة" "ه" "ي")

It was designed to strip off strings that were frequently found as prefixes or suffixes at the beginning or ending of stems. It was not intended to be exhaustive.

There are several versions of light stemming; all of them follow the same steps [8]:

1. Remove " و " from lgiht2, light3, and light8 and light10 if the remainder of the word is 3 or more characters long
2. Remove any definite article if this leaves 2 or more characters.

2.3 Statistical Approach

Statistical Approach related words are grouped based on various string similarities measures. Such approaches often involve N-Gram [4]. The main idea of N-Gram approach is that the similar words will have a proportion high of N-Grams. And extract the root after the comparing the similar characters. Specific values for n-Gram are bigrams or trigrams Note that, among the best known Arabic Stemming algorithms for each approach, we can select the Khoja Stemmer as Root-Based [13]; Larkey as Stem-Based [8] and the N-Gram as Statistical Based [4]. There are some weakness in this algorithm where the N-Gram algorithm returns many of documents that are not necessarily relevant, and the production of an index files size exorbitant.

3. Our approach technique

To conduct this study, a system (stemmer) is built to find Arabic roots using Python programming language version 3.2 . This stemmer of the word started by eliminate the " فال الا كال بال والا لبال وبال فبال والا لبال وبال فبال " , while Khoja stemmer initially removes affixes. Next if no matching one of set of the patterns it repeatedly eliminates the affixes in order to extract the root match Fa Aa La.

- **T(i)** be any term
- **Let LenT(i)** be the length of each term
- **Let chr (i)** be the character position within a term
- **Let Len P(j)** be the length of the pattern
- **Affixes** means prefixes and suffixes

- T-string Corresponding affixes

1. Remove (" فال الا كال بال ")
2. Remove (" فبال وبال لبال والا ")
3. Remove (" ال لل ")
4. Normalization

3.1 Remove diacritics depending on a list of diacritics characters

3.2 Remove tatweel symbol("_")

5. if Len T(i) = 3 then

It will be matched against (rootDB) if found( ) with one of these roots

Return that Root (T (i)) = T(i)

else

Stop matching because root not found ( )

endif

Endif

6. If *Len (T(I))* >= 4 then

For j ← 1 to number of patterns of length = Len T(i) do

If T(i) match pattern then

Extracted the word 's letter correspond "Fa Aa La " "فعل" by

Remove the T-string (affixes) character from T(i) and match

with (rootDB) repeat step No.5 to Return Root (T(i)) or not.

Endif

Next j

Table -3 Trace how to extract the correct root.

Original	Normalization	T-string	Root
التعليمات	تعليمات	تيات	علم
البدور	بدور	و	بدر
الوطنية	وطنية	ية	وطن
المكتبة	مكتبة	مه	كتب

4. Experimental Results

The proposed stemmer has been tested on 1450 words with most of Tafealat (يفعل تفعل تفعلين يفعلان تفعلان مفعول فاعلان فاعل فعلهم فعلها فعله يفعلن تفعلون يفعلون فاعلات فعول (مفاعل فعلين فعلان مفعولات مفعولان مفعلة مفاعيل تفعيلات مستفعل متفاعل مفاعلون مفعل مفعلة and the number of words correctly is 1398 words with percentage equal 0.96 % .The simple of words stemmed with our proposal stemmer in Table - 4

Table 4- sample of words stemmed with proposed stemmer

Word	Stemmed root	Correct root	Word	Stemmed root	Correct root
يلعب	لعب	لعب	مطحون	طحن	طحن
عامل	عمل	عمل	تطبخ	طبخ	طبخ
تدرسين	درس	درس	يمزحن	مزح	مزح
تذهبون	ذهب	ذهب	مسالمون	سلم	سلم
كسره	كسر	كسر	مقبولان	قبل	قبل
ضربه	ضرب	ضرب	مشاهير	شهر	شهر
متقاعس	قعس	قعس	تدريبات	درب	درب
رسمين	رسم	رسم	منتصران	نصر	نصر
مقبولات	قبل	قبل	مستعطف	عطف	عطف

$$\text{Precision} = \frac{\text{correct}}{\text{correct} + \text{incorrect}}$$

5. Conclusion

Arabic content on the internet has been increased rapidly over the last years. Thus, Arabic stemmer it can be used to enhance the efficiency a number of systems such as Information Retrieval systems, Text mining systems, Text analysis systems, etc. Arabic as a highly inflected language requires a good stemming process to make information retrieval effective. We have get the results after execution of the proposal stemmer. Started by deleting prefixes and then matching a word against Tafealat and removing affixes to extract the Root matching Fa Aa La. The proposal stemmer has been tested about 1450 words and the Stemmer' result comparison show in Table-5

Table-5 stemmer' Results

Stemmer	Stemmer Result
Proposal stemmer	96.1 %
Khoja Stemmer	85.7 %
Light Stemmer	85 %

Finally, the Arabic stemmer is always subject to further studies and the hope of building effective Arabic stemmer in the future.

References

[1] Kanaan, G.; Al-Shalabi, R.; AL-Kabi, M.N.; Jaam, J.M.; Hasnah, A.; . 2004. "New
Approach for Extracting Quadriliteral/Quadrilateral Arabic Roots ", In proceedings
of 1st International Conference on Information & Communication Technologies:
from Theory to Applications, ICTTA'04, (Damascus, Syria, April 2004). IEEE-
France.

[2] Khoja S., Research Interests, Pacific University, 2043 College Way, Forest Grove,
Oregon 97116 http://zeus.cs.pacificu.edu/shereen/research.h tm, July 8, 2006.

[3] AL-OMARI A. and ABUATA B. **2014**. ARABIC LIGHT STEMMER (ARS). *Journal of
Engineering Science and Technology*. vol. 9, pp. 702 – 716.

[4] Khreisat, L. "Arabic text classification using N-Gram frequency statistics a
comparative study". Proceedings of the 2006 International Conference on Data
Mining (pp. 78–82). Las Vegas, NV: USCCM.

[5] L. S Larkey , L. Ballesteros and M. E. Connell, "Light stemming for Arabic
Information retrieval ," in Arabic computational morphology , Springer, 2007, pp.
22-243.

[6] E. T. Al-shammari, lemmatizing, stemming, and query expansion method and
system . Google Patents, 2013.

[7] Khoja S.. "Stemming Arabic Text". Lancaster, U.K., Computing Department,
Lancaster University. 1999.

[8] Larkey L., L. Ballesteros, and M. E. Connell. "Improving Stemming for Arabic

Information Retrieval: Light Stemming and Co-occurrence Analysis". Proceedings of SIGIR'02. PP 275–282.2002.

[9] Purwitasari D, Najibullah A, Zainal Arifin A "Modification of Khoja Stemmer for Searching Arabic Text "Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Indonesia 2005.

[10] Abdusalam N., Seyed T., and Falk S., "Stemming Arabic Conjunctions and Prepositions," in Proceedings of the 12th international conference on String Processing and Information Retrieval.

[11] 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.

[12] Aitao C., "Building an Arabic Stemmer for Information Retrieval," in Proceedings of the Eleventh Text Retrieval Conference, Berkeley, pp. 631-639, 2003.

[13] Rafea Mohammed " New Arabic Stemming based on Arabic Patterns" *Iraqi Journal of Science, 2016, Vol. 57, No.3C, pp:2324-2330* **ISSN: 0067-2904**

[14] Mohammed A. Otair "Comparative Analysis of Arabic Stemming Algorithms" Article in International Journal of Managing Information Technology · May 2013 DOI: 10.5121/ijmit.2013.5201.