ISSN: 0067-2904

New Arabic Stemming based on Arabic Patterns

Rafea Mohammed*
College of Islamic Science, The Iraqi University, Baghdad, Iraq

Abstract

Algorithms for Arabic stemming available in two main types which are root-based approach and stem-based approach. Both types have problems which have been solved in the proposed stemmer which combined rules of both main types and based on Arabic patterns (Tafealat[1]) to find the added letters. The proposed stemmer achieved root exploration ratio (99.08) and fault ratio (0.9).

Keywords: Arabic stemming, Arabic pattern, root-based, stem-based.

ايجاد جذور الكلمات العربية بالاعتماد على التفعيلات

رافع محمد*
كلية العلوم الاسلامية، الجامعة العراقية، بغداد، العراق

الخلاصة

الخوارزميات العربية لايجاد الجذور متوفرة بنوعين رئيسيين هما النهج قائم على الجذر و النهج القائم على الجذع. كلا النوعين لهما مشاكل و التي حُلت في محلل الجذوع المقترح الذي يجمع بين قواعد كلا النوعين واعتمادا على تفعيلات اللغة العربية التي تساعد في ايجاد الحروف المضافة على الكلمة. حقق محلل الجذوع المقترح نسبة اكتشاف جذر (99.08) ونسبة خطأ (0.9).

Introduction

Arabic stemming is an approach which goes after finding the origin (root) of words in natural Arabic language by getting rid of any additions (affixes) in words, because Arabic words may have more complicated forms than any other language with such additions. Changes in the forms of words morphology have the same significance like changes in the meaning of words and they can be considered equivalent to what is intended in retrieving information systems.

Arabic stemming algorithms can be classified, according to the desirable level of analysis: root-based approach [1] and stem-based approach [2]. Root-Based approach uses morphological analysis to find the root of a given Arabic word. Many algorithms have been proposed for this approach [3-5]. The aim of the Stem-Based approach is to eliminate the most frequent prefixes and suffixes [6-9]. Most of the endeavours in this area were a number of rules to attain the least number of additions before or after the words (prefixes, suffixes), also there is no certain list of strippable of these additions (affixes) [10].A Light stemming is one of the most superior in morphological analysis, it passed the other algorithms in terms of performance (precision and recall) [11].

Many stemming researches were introduced to reduce words to its original root, like Khoja's stemmer [1], Buckwalter's morphological analyzer [12] and the Tri-literal root extraction algorithm [13]. The Khoja stemmer gets the highest accuracy, next is the tri-literal (triple) root extraction algorithm and the last one is the meaning analyzer Buckwalter. In [14] a light and heavy Arabic stemmer was introduced. Results showed that accuracy of the stemmer is slightly better than the accuracy yielded by each one of those two well-known Arabic stemmers used for evaluation and

*Email: rafea_sweetest@yahoo.com

1 "Arabic patterns" and "Tafealat" are used interchangeably in this research.

comparison ((i.e. Khoja and Garside (1999), Ghwanmeh et al. (2009)). Evaluation tests on our novel stemmer yield 75.03% accuracy, while the other two Arabic stemmers yield slightly lower accuracy. In [15] the algorithm first preprocessed the document to be stemmed, then it matched the resulted words against Arabic patterns to get the stems of the words. In this research, the proposed light stemming algorithm for Arabic Languages showed better results.

Arabic Language Characteristics

Arabic language is the mother language for over than 300 million people in comparison to the English language, the Arabic language has its own special characteristics that are diversified [16].Arabic is a very rich and complex language. Arabic has 28 characters and is written from right to left. Arabic language differs from English and European languages and the morphological representation of Arabic is rather complex because of the morphological variation and the agglutination phenomenon [17]. Arabic language is described as algebraic language which makes its morphological analysis process very difficult and hard.

Arabic Roots and Patterns

Arabic language is based on set of roots [18]. 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. An Arabic root is an ordered sequence of three (**فعل**) or four letters (**فعلل**) from alphabet [21]. The root has a general, basic meaning which forms the basis of many related meanings [22]. 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 derivations from them).
- 7198: Three character roots.
- 3739: Four character roots.
- 295: Five character roots.

These roots join with various vowel patterns to form simple nouns and verbs to which affixes can be attached for more complicated derivations [19]. Arabic patterns are part of the Arabic grammar. They are formed based on the Arabic root [20]. Patterns play an important role in Arabic lexicography and morphology [19]. They are generated from the process of vocalization and affixation [23]. Each root can canonically combine with orthographically distinct patterns to form another words, for example, the root"لعب"is consisting of three characters root, the root "لعب" corresponds to the pattern "فعل"and the pattern preserves "ف", "ع", "ل" in the same order, where other letters can be added to form another pattern. For example, several patterns are derived from the base pattern "فعل" of the morpheme "لعب". The pattern "مفعل" form the word "ملعب" by adding the letter "م" to the morpheme "لعب" [19]. Table-1 shows a sample of the Arabic Patterns (Three-Consonant root).

Table 1- Arabic patterns sample (Trilateral roots)

Arabic Patterns									
fa'ala ↓	فعل	mustafa'ael	مستفعل	mafa'ael	مفاعيل	tafa'alon	تفعلون	fea'altan	فعالتان
maf'al	مفعل	mustafa'aelat	مستفعلات	afa'al	افعل	tafa'aln	تفعلن	fa'a:la	فعالى
mafa'aloon	مفعلون	mustafa'alon	مستفعلون	afa'aela	افعلاء	tafa'aelat	تفعيلات	fa'ala	فعلى
mafa'aleen	مفعلين	mutafa'ael	متفاعل	ifti'a:l	افتعل	Fa:'il	فاعل	fea'ali	فعلي
mafa'alan	مفعلان	mutafa'alat	متفاعلات	eftea'al	افتعال	faa'aelan	فاعلان	fa'aol	فعول
mafa'aool	مفعول	mufa'ael	مفاعل	tafa'aul	تفاعل	fa'ael	فعائل	yafa'al	يفعل
maf'ala	مفعلة	mufa'aeloon	مفاعلون	tafa'aulan	تفاعلان	fi'a:al	فعال	yafa'alan	يفعلان
mefa'aal	مفعال	mufa'aelat	مفاعلات	tafa'alen	تفعلين	fea'ala	فعالة	yafa'alon	يفعلون

Variations of the root and patterns determine the actual meaning of the word. For example, the root (ktb كتب) with the addition of the letters (i ,a) gives the word (kita:b كتاب), which means book, ↓ The combination (a'a) represents the letter ' ع' of the Arabic alphabet.

While the root pattern combination of (ka:tib كاتب) means "one who writes" or "clerk". There are also some prefixes and suffixes which determine whether a word is a subject marker, pronoun,

preposition, or a definite article. Table-2 illustrates set of derivatives patterns, its corresponding English word, the position in the language and its Arabic patterns from the Arabic trilateral verbal root 'k t b' [24].

Table 2- Derivatives of the Arabic trilateral root 'k t b'Arabic English POS Pattern Arabic English POS Pattern

Arabic	English	POS	Pattern	Arabic	English	POS	Pattern
Ktb	Write	V	Fa'ala	Maktab	Office	N	Maf'al
Kita:b	Book	N	Fi'a:l	Maktabah	Library	N	Maf'ala
Kita:bah	Writing	N	Fi'a:lah	Muka:tabah	Correspondence	N	Mufa:'alah
Ka:tib	Writer	N	Fa:'il	Iktita:b	Subscription	N	Ifti'a:l
Ka:taba	correspond	V	Fa:'ala	Kita:bi	Clerical	Adj	Fi'a:li

Arabic Language Affixes

Arabic language, unlike English, both prefixes and suffixes are removed for efficient result, but Arabic provides the additional difficulty of infixes [26]. The difficulty arises because Arabic has two genders, feminine and masculine; three numbers, singular, dual, and plural; and three grammatical cases, nominative, genitive, and accusative. A noun has the nominative case when it is a subject; accusative when it is the object of a verb; and genitive when it is the object of a preposition. The form of an Arabic noun is determined by its gender, number, and grammatical case [25].

Stemming

Stemming is a very essential technique for processing strong morphological languages such as Arabic [16]. Word stemming in Arabic is the process of removing all of a word's prefixes and suffixes to produce the stem or root. Simply, it is a conversion of plurals to singulars, or derivation of a verb from the gerund form. There are also other possibilities such as deriving the root from the pattern words.

The importance of the stemming process is in the classification and index builders/searchers because it makes the operations less dependent on particular forms of words and reduces the potential size of vocabularies, which might otherwise have to contain all possible forms.

Since Arabic is complexly morphological language, so it requires a further effort of morphological analysis, where absolutely morphological techniques are required that eliminate suffixes from words according to their internal structure [24].

Root Extraction Stemmer

Arabic words are formed from abstract forms named roots, the root is the 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. A root based stemmer main goal is to extract the basic form for any given word by performing morphological analysis for the word, Table-3 shows an example root "لعب" and a set (not all) derivations can be obtained from this root [27]:

Table 3- Some derivations of the root لعب

يلعب	ملعب	لاعب	ملعوب	لعبة
Play	Playground	Player	Played	Game

Khoja [1] stemmer basically attempts to find roots for Arabic words which are far more abstract than stems. It first removes prefixes and suffixes, then attempts to find the root for the stripped form [28].So ending up with the fact that root extraction stemmers increase word ambiguities and that inflected and derived words can have a vigorous impact on the retrieval effectiveness of any information retrieval system and a good stemmer should recognize the different forms of a word [29].

Arabic Light Stemmers

There are several stemming approaches that are applied for Arabic language; one of them is light stemmer algorithm. It is not an aggressive practice as the root-based algorithm. The aim of this approach is not to produce the linguistic root of a given Arabic surface form; rather, it is to remove the most frequent suffixes and prefixes. In Arabic, unlike English, both prefixes and suffixes are removed for efficient results, but Arabic provides the additional difficulty of infixes [24].

The Proposed Arabic Stemmer

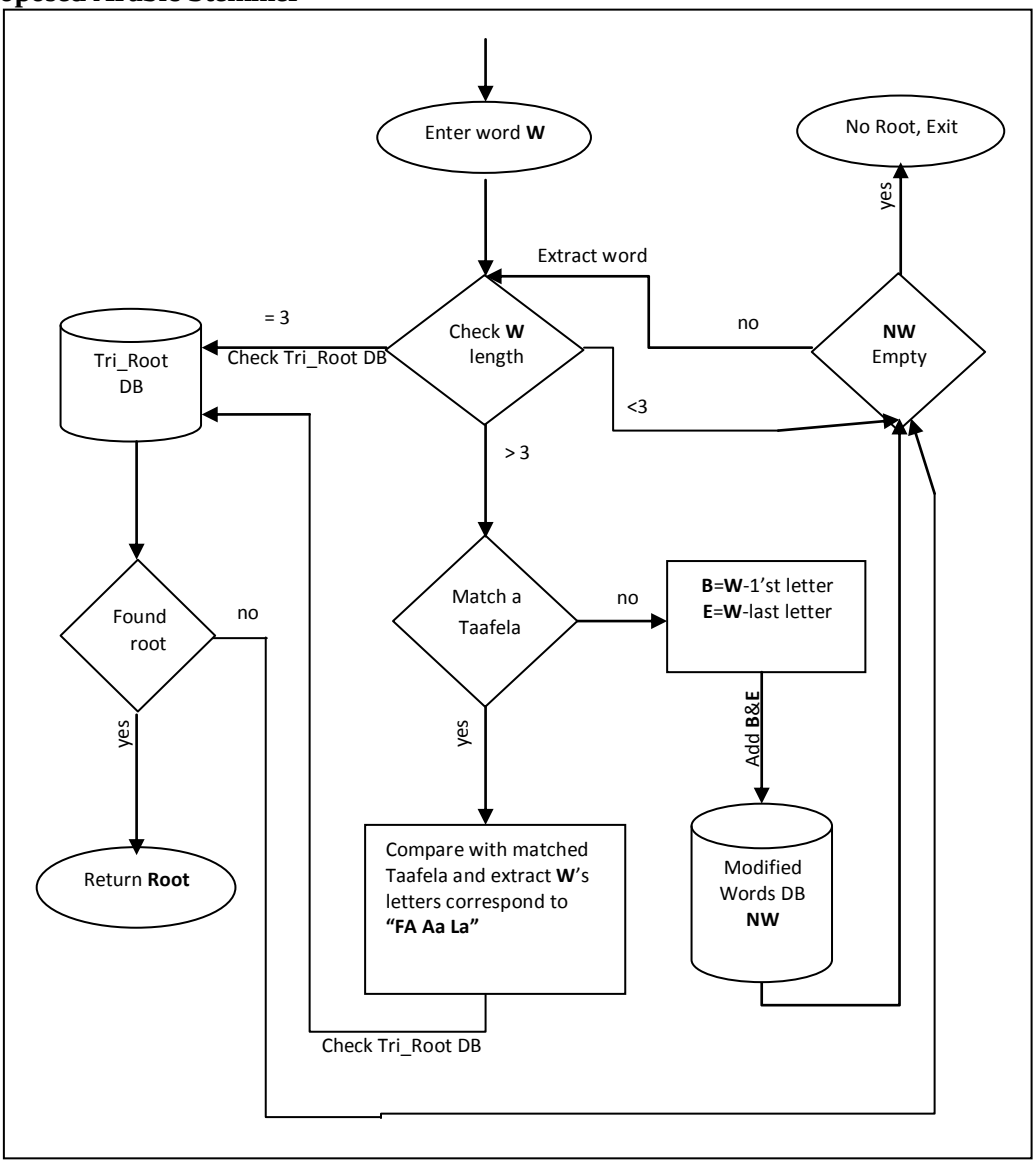

Figure 1- Stages of the Proposed Arabic Stemmer

This research proposed an Arabic stemmer that combined the rules of *root-based* stemmer and *light-based* stemmer to overcome the mistakes, which they are: affix removing before matching with a Tafeala and not dealing with word of three letters length, resulted from these two stemmers as shown in Figure-1 as follows:

Checking word[1]'s length due to number of its letters to examine three cases:

* *First case*: the word composed of three letters, it will be matched against *tri_rootDB* (this is not done by the previous stemmers). If a match found with one of these roots the matching is stopped and return that root, else if no match is found:-

-in situation of the word length is three letters stop matching because its roots is not found.

- Other situation the word is three letters length and it is not the original word (i.e. modified word) resulted from either Tafeala or removing a letter from beginning or end, then a check on New Word Database will be done as it happens in next steps and follow the same procedures.

**Second case*: word's length is more than three letters, it first will be match with Tafealat (which are quad, penta and hexaletters) shown in Table-1. This is not done by the previous stemmers.

[1] The term "word" may refer either to "original word" or "modified word"

- if a match with one Tafela is found, matching is stopped and the word's letters correspond to "Fa Aa La" will be extracted to be match with *tri_rootDB*, and if the root found the matching is stopped , else if no match is found then repeat steps of the first case when do not matching any roots .
-If no match exists with any Tafela , one letter will be removed :-
a) Once from word's beginning if it match one of these prefixes ('ل', 'ف', 'ت', 'م', 'ن', 'س', 'ي', 'ب', 'و', 'ا') and append the modified word to *New Word Database*.
b) and another from word's end if it match one of these suffixes ('ه', 'ة', 'ك', 'ن', 'ت', 'ا', 'ي', 'م','و') and append the modified word to *New Word Database*.
If *New Word Database* is empty so this mean that word's root not found and the matching is stopped, else pull a word from database and repeat steps that have been done on the original word.
**Third case*: word's length is less than three letters, it will be discarded and repeat the previous step.

Experimental Results

The proposed stemmer has been tested on 1634 words with these Tafealat (يفعل تفعل تفعلين يفعلان تفعلان يفعلون تفعلون يفعلن فعله فعلها فعلهم فاعل فاعلان مفعول) and the number of correctly stemmed words is 1619 words with percentage of correctness equal to 99.08 % and error ratio equal to 0.9 %. In Table-4 a sample of words stemmed with the proposed Arabic stemmer.

Table 4- Sample of words stemmed with the proposed Arabic stemmer

Word	Stemmed root	Correct root		Word	Stemmed root	Correct root
يلوم	لوم	لوم		عاكس	عكس	عكس
يلهو	لهو	لهو		مطبوخ	طبخ	طبخ
تندمين	ندم	ندم		تطحن	طحن	طحن
تفرحون	فرح	فرح		طبعهن	طبع	طبع
عمله	عمل	عمل		يلهين	لهي	لهو

Conclusion

These conclusions reached from the execution of the proposed stemmer and the obtained results.

- Matching a word against Tafealat before removing any affixes to avoid deleting a genuine letter of a word unlike other stemmer researches.
- The proposed stemmer has been solved the situation of three letters words while most of stemming researches like [1, 15] did not treat it. Table-5 illustrates a comparison with the stemming results of these two stemmers.
- Proposed stemmer has been tested with the same words that used for testing stemmers in [1,15]. Stemming results of [1,15] have been obtained from their papers.
- It was reached experimentally that dealing with affixes of one letter length is best than that of two or three letters as it is obvious from the obtained exploration results.
- Reducing Tafealat number to as little as possible in order to include Tafealat without suffixes.

Table 5- Stemmers' Results Comparison

Stemmer	Stemming Result
Proposed Stemmer	99.8 %
Khoja Stemmer	85.7 %
[15] Stemmer	85 %

References

1. Khoja S. and Garside R. **1999**. Stemming Arabic text. Technical report, Computing Department, Lancaster University, Lancaster.
2. Larkey S., Ballesteros L., and Margaret E. Connell. **2002**. Improving Stemming for Arabic Information Retrieval: Light Stemming and Occurrence Analysis. In Proc. of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR'02), Tampere, Finland, pp: 275–282.
3. Al-Fedaghi S. and Al-Anzi F. **1989**. A new algorithm to generate Arabic root-pattern forms. In proceedings of the 11th national Computer Conference and Exhibition. pp: 391-400.
4. Al-Shalabi R. and M. Evens. **1998**. A computational morphology system for Arabic. In Workshop on Computational Approaches to Semitic Languages, COLING-ACL98.

5. Freund, G. and Willett P. **1982**. Online Identification of word variants and arbitrary truncation searching using a string similarity measure. *Information Technology: Research and Development 1*. pp: 177-187.
6. Aljlayl M. and Frieder O. **2002**. On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. Proceedings of the eleventh international conference on Information and knowledge management. pp: 340-347.
7. Chen A. and F. Gey. **2002**. Building an Arabic Stemmer for Information Retrieval. In Proceedings of the 11th Text Retrieval Conference (TREC 2002), National Institute of Standards and Technology.
8. Larkey L. and M. E. Connell. **2001**. Arabic information retrieval at UMass in TREC-10. Proceedings of TREC 2001, Gaithersburg: NIST.
9. Larkey S., Ballesteros L., and Margaret E. Connell. **2002**. Improving Stemming for Arabic Information Retrieval: Light Stemming and Occurrence Analysis. In Proc. of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR'02), Tampere, Finland.pp.275–282.
10. Syiam M., Fayed Z. and Habib M. **2006**. An Intelligent System for Arabic Text Categorization. *IJICIS*.vol.6, no. 1.
11. Fouzi H., Aboubekeur H., and Abdulmalik S. **2010**. *Comparative study of topic segmentation Algorithms based on lexical cohesion: Experimental results on Arabic language. The Arabian Journal for Science and Engineering*. vol. 35, no. 2C.
12. Buckwalter T.**2004**. Buckwalter Arabic Morphological Analyzer Version 2.0. *Linguistic Data Consortium (LDC) catalogue number LDC2004L02*. ISBN 1-58563-324-0.
13. Al-Shalabi R., Kanaan G., & Al-Serhan H. **2003**. New approach for extracting Arabic roots. In proceedings of The International Arab Conference on Information Technology.
14. Al-KabiM., KazakzehS. and et al **2015**. A novel root based Arabic stemmer. *Journal of King Saud University – Computer and Information Sciences*. vol.27, 94–103.
15. Sameer R.**2016**. Modified Light Stemming Algorithm for Arabic Language. *Iraqi Journal of Science*. vol. 57, no.1B, pp: 507-513.
16. Otair M. **2013**. Comparative Analysis of Arabic Stemming Algorithms. *International Journal of Managing Information Technology (IJMIT)*. vol.5, No.2.
17. Abdusalam N., Seyed T., and Falk S.**2005**. Stemming Arabic Conjunctions and Prepositions. In Proceedings of the 12th international conference on String Processing and Information Retrieval, Heidelberg, pp. 206-217.
18. Marwan B. **2004**. Arabic Language Processing in Information Systems. *Springer*.
19. Aitao C.**2003**. Building an Arabic Stemmer for Information Retrieval. In Proceedings of the Eleventh Text Retrieval Conference, Berkeley.pp. 631-639.
20. www.mesiti.it **2005**. Arabic Grammer. Available at: http://www.mesiti.it /arabic/grammar /lessons/ lesson2/roots.html.
21. David Crystal. **1997**. *A Dictionary of Linguistics and Phonetics*. (Language Library). Fourth Edition. Blackwell Publishers.
22. Jonathan Owens. **1988**. *The foundations of grammar: an introduction to medieval Arabic grammaticall theory*. John Benjamins Publishing Company.
23. Joseph Dichy and Ali A. Farghaly. **2003**. Roots & Patterns vs. Stems plus Grammar-lexisSpecifications: On What Basis Should a Multilingual Database Centered on Arabic be Built?. MTSummit IX -- workshop: Machine Translation for Semitic Languages, New Orleans, USA.
24. Al Ameed H., Al Ketbi S., and et al.**2005**. Arabic Light Stemmer: Anew Enhanced Approach. The Second International Conference on Innovations in Information Technology (IIT'05).
25. Aitao Chen, and Fredric C. Gey. **2003**. Building an Arabic Stemmer for Information Retrieval. Available at: http://metadata.sims.berkeley.edu/papers/trec.
26. Monica Rogati, Scott McCarley, and Yiming Yang. **2003**. Unsupervised Learning of Arabic Stemming Using a Parallel Corpus. Available at: http://acl.ldc.upenn.edu /acl2003/main/pdfs/ Rogati.pdf.

27. Ababneh M., Al-Shalabi R., and et al. **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,
28. Kazem T., Rania E., and Jerey C. **2005**. Arabic Stemming Without A Root Dictionary. *Information Science Research Institute,* USA.
29. Aitao C.**2003**. Building an Arabic Stemmer for Information Retrieval. In Proceedings of the Eleventh Text Retrieval Conference, Berkeley. pp:631-639.