Research Article

International Journal of Advanced Computer Research, Vol 8(35)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2017.733023

Arabic root extraction using a hybrid technique

Hayel Khafajeh[1*], Nidal Yousef[2] and Mahmoud Abdeldeen[3]
Faculty of Information Technology, Zarqa University, Zarqa, Jordan[1]
Faculty of Information Technology, Al-Isra University, Amman, Jordan[2]
Faculty of Arabic Language and Culture, National Chengchi University, Taipei, Taiwan[3]

Received: 25-August-2017; Revised: 24-February-2018; Accepted: 27-February-2018

Abstract
Root extraction is one of the main text operations conducted by converting the conflation into its root. This process aims to overcome the morphological richness problem of the Arabic language. Root extraction gives a valuable support to many natural language processing applications such as information retrieval, machine translation, and text-summarizing applications. In this research, a hybrid technique to extract Arabic word roots has been developed. The proposed technique depends on optimization function, which is the enhancing process performed by playing a set of non-morphological rules to enhance the n-gram technique. The proposed technique is tested using a dataset containing more than 6000 distinguished words belonging to 141 different roots. The results show a marked improvement after using the hybrid method, the proposed technique extracts correctly about 99% of tripartite strong roots and about 86% of tripartite vowels roots.

Keywords
Arabic root extraction, Natural language processing, Hybrid technique, Similarity.

1.Introduction
The Arabic language is one of the major languages in the world. The language is spoken by nearly 400 million people and ranks fifth in the world's languages [1]. It is also the language of the holy Quran used by more than a billion and a half Muslims in their prayers. The Arabic alphabet consists of 28 letters written from right to left using cursive letters. Arabic words are derived from their roots by adding postfixes, infixes, and suffixes or by amending the center of the word. Many applications in the Arabic language computerization field utilize the conversion of Arabic words into their roots to use their roots instead of the word. The main examples of these applications are information retrieval systems, document classification systems, text summarizing, automatic translation systems, and optical character systems (e.g., optical character recognition (OCR)) [2].

Arabic roots can be classified according by containing vowels into two types [3].

*Author for correspondence

The research is financed by Deanship of Research and Graduate Studies at Zarqa University, Jordan.

The first type, which is called the vowel root, is the root that contains at least one vowel. The second type, which is called the strong root, is a root that does not contain a vowel. We can classify Arabic roots into the four following types according to the number of letters forming the root: trio, which forms most of the words in the Arabic language [4], quartet, quintet, and hexagon.

Many techniques are employed to extract the roots of Arabic words. However, no agreement has been reached as to which one method should be used because of the morphological richness of the Arabic language [5] and the presence of a large number of different conflations for each word. Researchers have presented several approaches of extracting the roots of Arabic words [6, 7], especially the strong trilateral roots. Many of the researchers rely on the morphological rules and analysis to extract the word roots [1], which makes the root extraction process difficult and complex because of the multiple formulas of the morphological forms for each word. The researchers herein present a hybrid method (i.e., statistical approach plus some non-morphological rules) that reduces the complexity of the extraction process of the word roots.

The Arabic language is described as a derivative language, which is a prominent feature of Semitic languages. The Arabic language always retains this feature, which gives it flexibility, vitality, and verbal wealth that amounts to over 70,000 of pure derivatives leading to the description of the Arabic language as a propagation language [8]. Many researchers consider that derivation is "the process of creating". The creativity in language is an important tributary that supplies all that is needed from the vocabulary and formulas. It is also a factor of language growth and evolution as well as a means of enriching the vocabulary.

1.1Derivation

Derivation is defined as the formation of a word from another word with the fittings between the two words in the pronunciation and meaning [9].

We can form from the root (ف هـ م) using the following words: (يفهم) understand (افهمْ)(understand) and (فاهِم)(realizable) and (مفهوم)(the concept) and (فَهْم)(understanding), (فهّم)(understand), and (أفهمَ)(understand), and (فهّام)(to make anybody understand), and (استفهم)(inquired), and (يستفهم)(ask), and (يستفهم)(question), and (استفهام)(questioning), and (مستفهِم)(questioning), (مستفهَم)(inquired). Each derived formula has a new significance that includes the original meaning of the root (ف هـ م).

The derived root must include its original letters and meaning. Therefore, the semantic of the original root is conserved. However, this remains an ongoing indication of the root with words derived from it. Language scientists have called derivation in this form as the morphological derivation, which relies on certain standard formulas for the derivation of words from each other. Morphology is defined as the science of studying formulas and how to derive and change them, which leads to the development of new formula meanings.

The morphological derivation has undoubtedly continued as a method of enriching the Arabic language to keep up with the linguistic needs of the language users, which resulted from the development of life in different areas. Deriving new words, according to the standard rules, which theoretically means the ability to derive many words from any root according to the standard format, is necessary. However, the need for this derivation is decided by the presence or absence of a derived word. Ibrahim Anis said that "*many of those formulas that may be derived do not exist actually in true Arabic texts as there is a big difference between what we can derive from formats, and what actually is derived and used in the narrated Arabic language*" [10].

Derivatives grow when needed because the word derivations are not derived at the same time, but derived according to new needs. In other words, derivation is a standard method for word Symantec expansion, as required by the necessities and the development of life. Derivation is also an important tributary to produce vocabulary.

1.2Linguistic root

Linguistic root is defined as the origin, which generates a word and consists of a silent without any diacritics. The linguistic root specializes in providing indications on a particular meaning, which remains inseparable from the derived words [11]. The great majority of the roots of words in the Arabic language are the trilogy, which forms up to 85% of all Arabic language roots. Meanwhile, the quartets' roots are few [12].

The Arabic language has about 10,000 roots and 900 morphological weights [13]. The most important things to notice when examining the roots are the following:

1. Most of the Arabic roots consist of correct consonants. These correct consonants are correct safety (درس), Mahmozh (سأل), or intensity (شدّ).
2. Some roots consist of vowel consonants. This letter is held in the first letter of the root (وقف), middle (قام), or end (قضي). Some possibility of meeting more than one silent in one root exists.

From the roots, we can derive many conflations that contain the original meaning of the root. These conflations are formed by adding affixes to the root. These affixes are collected in the Arabic word "سألتمونيها".

2.Previous studies

Many researchers have presented different methods of extracting the roots of Arabic words. Some of these methods rely on morphological analysis. Others are based on statistical methods. Boudlal et al. [14] presented a method of finding the roots of Arabic words through a series of operations on the text. These operations include the removal of diacritics, removal of stop words, removal of the extra "waw", and the removal of prefixes and suffixes. After these operations, the resulted word is compared to a list of roots. Hmeidi et al. [15] used the hidden Markov model (HMM) to extract the consonant roots. The

researchers relied on context to determine the root. The system consists of two parts. In the first part, the word is morphologically analyzed without relying on context to identify the potential roots by dividing the word into the prefix, stem, and suffix. In the second part, the context is used to identify the correct root among the potential ones.

Al-Kabi et al. [16] developed and evaluated a new method of extracting the tripartite Arabic roots. The proposed method is based on light and heavy approaches (on the root basis) and uses three stages of processing to generate the word root. The first phase is responsible for the removal of prefixes and suffixes, while the second phase compares the output of the first phase of the sources of the word or the standard forms. The third phase is responsible for correcting the extracted root.

Hmeidi et al. [15] proposed a new algorithm based on the bigram to extract the roots of Arabic words. They used the similar Manhattan standards and DICE. The researchers tested the proposed algorithm using the holy Quran. Their algorithm extracted three-letter, quartet, quintet, and hexagon roots from the summaries of 242 texts in addition to the Arabic roots of seven. Abuata and Al-Omari [17] suggested the root analyzer for various Arabic accents. They described an algorithm based on the new rules of extracting the roots of the tone of the text of the Arabian Gulf accent. Boubas et al. [18] showed a new algorithm to strip Arabic words using genetic algorithms and verb pattern matching. This algorithm was mainly based on the automatic learning system and the rules and patterns of the Arabic word formation. They adopted an Arabic language analyzer capable of generating Arabic roots for any (stream) of Arabic words.

3.The proposed algorithm

The hybrid algorithm developed herein to extract the roots was reviewed in this section. We developed the proposed algorithm to extract the roots of Arabic words using a statistical technique known as the bigram. The bigram was originally proposed to remove prefixes and suffixes. We relied on *Equation (1)* [19] as follows to calculate the degree of similarity between the word and its suggested roots from the list:

$$S = 2C / (A + B) \quad (1)$$

Where

C: parts uniquely shared between words A and B
A: the number of unique bigrams in word A
B: bigrams in the number of unique word B

In the word root extraction process, we found that the root of the word we want to extract must be available with the words in addition to a group of potential roots for this word to benefit from *Equation (1)* and calculate the value of the similarity in the same equation. We used for this a list of roots consisting of 3500 roots. We extracted a word's root using the proposed algorithm by first dividing the word into bilateral parts. The value of (s) was then calculated as shown in *Table 1*. For example, the values of A, B, C, and S were obtained as shown in *Table 2* if we had the word "مستعرض", whose root is "عرض".

Table 1 Calculating equation (1) values

Ngram	مستعرض (مس ست تع عر رض) عرض (عر رض)
A	5
B	2
C	2
S	S = (2 * 2) / (5 + 2)

Yousef et al. [20] suggested the following algorithm to extract roots:

1. Normalization by deleting the word diacritics, Alhmza (ء), and converting (ة) to (ه)
2. Cutting the word to its bigram parts
3. Calculating the similarity S between the word and a list of roots starting with the roots matching the first character in the word
4. Repeating the previous step for all characters that form the word
5. The root with the highest similarity value S will be chosen as the word root

3.1Optimization methods

This research uses the objective function optimization to solve the problems that researchers in [21] face, which yield to root redundancy, prevent and find the exact roots, and find the near optimal solution (root). Extracting the root by calculating the similarity degree between the word and its roots might be a trap at redundancy. In some cases, several words have the same similarity with two or more roots. This research proposes three constraints, which must be satisfied as follows:

C1: The extracted root letters should have the same order with any given word.
C2: The extracted root letters should not count more than the word letters.
C3: All extracted root letters should exist in the word.

The penalty weight for each constraint violation is determined by human experts. According to the

preliminary experiments, ten points are provided for all constraints.

1. Initialization: extract the roots using the similarity(s), as proposed in [20]
2. Fitness: evaluate the fitness f(x) of each root as follows:

For example: if the word is (نضال):

3	2	1	0
ل	ا	ض	ن

Pre-objective	3	2	1	0	Root
0.4			ر	ض	Root 1
0.4		ل	ض	ن	Root 4
					Root 5

Pre-objective: the objective function, which came from the similarity in the first phase of this work. The new root objective function is calculated as follows:

$$f(x) = f(x)' + H1 + H2 + H3$$

Root 1:

H1	H2	H3
The letter ر does not exist in the original word (نضال). The penalty is calculated by subtracting (0.4/2) from the current objective function.	The letter order is not correct. The penalty is calculated by subtracting (0.4/2) from the current objective function.	The penalty is 0, while Root 1 letters do not count greater than the original word.

$$f(x) = 0.4 + (-0.4/2) + (-0.4/2) + 0 = 0$$

Root 4:

H1	H2	H3
All the letters exist in the original word. The penalty is calculated by adding (0.4/2) to the current objective function.	The letter order is correct. The penalty is calculated by adding (0.4/2) to the current objective function.	The penalty is 0, while Root 1 letters do not count greater than the original word.

$$f(x) = 0.4 + (0.4/2) + (0.4/2) + 0 = 0.8$$

The new objective function calculation states that the greater objective presents the best solution. Root 4 is the best quality solution.

3. Test: Stop and return the best solution (Root) if the end condition is satisfied.

4. Experiments

We design a corpus consisting of 141 roots to examine the proposed algorithm. The corpus contains 6308 morphological forms derived from 141 roots. The corpus also includes 1318 morphological forms belonging to 21 vowel roots.

Table 2 demonstrates the morphological forms used in the experiments for each root.

Table 2 Example of morphological forms for each root for the verb "كتب"

الرقم	المشتقة	الوزن	الدلالة
1	كتب	فعَل	الأفعال الماضية
2	كتبت		
3	كتبا		
4	كتبتا		
5	كتبوا		
6	كتبن		
7	كتبتم		
8	كتبتن		
9	كتبنا		
10	يكتب	يفعل	الأفعال المضارعة
11	تكتب		
12	يكتبان		
13	تكتبان		
14	يكتبون		
15	يكتبن		
16	تكتبوا		
17	تكتبون		
18	اكتب	افعل	أفعال الأمر
19	اكتبي		
20	اكتبا		
21	اكتبوا		
22	اكتبن		
23	كتابة	فعالة	الحدث مجرد من الزمان والمكان
24	كتابات		
25	كاتب	اسم فاعل	الشخص الذي قام بالفعل
26	كاتبان		
27	كاتبات		
28	كاتبون		
29	مكتوب	اسم مفعول	المادة التي وقع عليها الفعل
30	كَتْبَة	مصدر مرة	عدد مرات صول الفعل
31	كِتبة	مصدر هيئة	هيئة حصول الفعل
32	مكتب	اسم مكان	مكان حصول الفعل
33	مكتبان		
34	مكاتب		
35	مكتبات		
36	مكتبة		
37	كُتّاب	فعّال	صيغة مبالغة للفاعل الواحد
38	كُتّاب	فُعّال	صيغة مبالغة للفاعل المجموع
39	اكتَب	أفعَل	اسم تفضيل
40	كتيّب	فعيعِل	صيغة تصغير
41	مكاتبة	مفاعلة	
42	كتاب		
43	كتابان		
44	كتب		

4.1Results without optimization

The results obtained after running the proposed algorithm on the designed corpus without optimization are discussed below.

Tripartite strong roots

Table 3 shows the results obtained when the morphological forms of strong triple roots that do not contain a vowel are examined.

Table 3 Tripartite strong root results

Results	Root's number	Ratio
Correct root	3971	0.79659
Wrong or no root found	1194	0.239519

Tripartite roots with vowels

Table 4 presents the results obtained when the morphological forms for the tripartite roots with vowels are examined. The results are similar to the strong triple roots because of the reliance on statistical methods without considering the vowels.

Table 4 Tripartite roots with vowel results

Results	Root's number	Ratio
Correct root	987	0.747161
Wrong or no root found	334	0.252839

All roots

Table 5 shows the results obtained when all morphological forms for the designed corpus are examined.

Table 5 All root results

Results	Root's number	Ratio
Correct root	4958	0.603
Wrong or no root found	1348	0.3971

4.2Results after optimization

The results obtained are as follows after running the hybrid algorithm on the designed corpus:

Tripartite strong roots

Table 6 shows the results obtained when the morphological forms of strong triple roots that do not contain a vowel are examined.

Table 6 Tripartite strong root results

Results	Root's number	Ratio
Correct root	4930	0.990358
Wrong or no root found	48	0.009642

Tripartite roots with vowels

Table 7 shows the results when the morphological forms for the tripartite roots with vowels are examined.

Table 7 Tripartite roots with vowel results

Results	Root's number	Ratio
Correct root	1143	0.865909
Wrong or no root found	177	0.1340909

Table 8 shows the results obtained when all morphological forms for the designed corpus are examined.

Table 8 All root results

Results	Root's number	Ratio
Correct root	6073	0.964274
Wrong or no root found	225	0.035726

Table 9 illustrates the comparison between the algorithm proposed by Yousef et al. [20] and the hybrid algorithm proposed in this research.

The standard measures such as precision, and recall are used in order to evaluate the effectiveness of the proposed hybrid technique that used to extract Arabic word roots. Therefore, the following formulas are used to compute each of the above three measures [22]:

$$\text{Precision} = \frac{\text{Correct}}{\text{Correct} + \text{Incorrect}} \quad (2)$$

$$\text{Recall} = \frac{\text{Correct}}{\text{Correct} + \text{UnAnalyzed}} \quad (3)$$

After applying the *Equation (2) and (3)* we have the following results:
Precision = 0.987
Recall = 0.963

Table 9 Comparison between two algorithms

Root \ System		N-Gram	Hybrid
Strong	Number	3971	4930
	Ratio	0.79659	0.990358
Vowels	Number	987	1143
	Ratio	0.747161	0.865909
All	Number	4958	6073
	Ratio	0.603	0.964274

Figure 1 shows that the proposed algorithm extracts 99% of strong roots while the n-gram algorithm extract about 80% of strong root, in the case of words with vowel letters the proposed algorithm extract 87% of these roots.

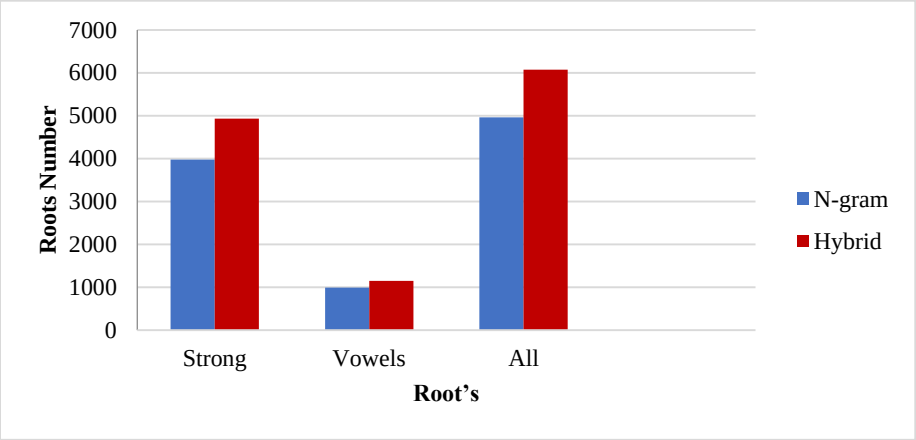

Figure 1 Comparison between N-gram method and hybrid method

5. Conclusions

In this work a hybrid multi-objective function with a statistical algorithm for extracting Arabic roots has been proposed. A multi-objective function is used to avoid getting trapped in similarity roots by finding the best quality solution calculated using new proposed constraints. The aim is to guide the search to other promising regions probably different from the current local optimum. The multi-objective function aims to enhance the ability of extracting the root by escaping from the local optimum and diverting the search to another promising region when the searches are trapped in the local optimum. The computational results show that the new hybrid methods, improve performance and outperform other static methods for extracting the Arabic roots. Furthermore, the performance of the new hybrid method tested on the given Arabic words and compared to statistical algorithm results.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Abu-Errub A, Odeh A, Shambour Q, Hassan OA. Arabic roots extraction using morphological analysis. International Journal of Computer Science Issues. 2014; 11(2):128-34.

[2] Elazhary H, Alharthi A, Balkhi E, Aljahdali G, Zagzoog D, Alkhammsh A. Automated tutoring of Arabic word root extraction. International Journal of Scientific & Engineering Research. 2015; 6(7):687-91.

[3] Wightwick J. Arabic verbs & essentials of grammar. McGraw Hill Professional; 2017.

[4] Yousef N, Al-Bidewi I, Fayoumi M. Evaluation of different query expansion techniques and using different similarity measures in Arabic documents. International Journal of Computer Science Issues. 2010; 43:156-66.

[5] Al-Fedaghi S, Al-Anzi F. A new algorithm to generate Arabic root-pattern forms. In proceedings of the national computer conference and exhibition 1989 (pp. 391-400).

[6] Alsaad A, Abbod M. Arabic text root extraction via morphological analysis and linguistic constraints. In international conference on computer modelling and simulation 2014 (pp. 125-30). IEEE.

[7] Hawas FA. Towards a new Approach for Arabic root extraction: exploit relations between the word letters and their placement in the word for Arabic root extraction. Computer Science. 2013; 14(2):327-41.

[8] Mustafa SH. A relational approach to the design of an Arabic lexical database. Journal of King Saud University-Computer and Information Sciences. 2002; 14:1-23.

[9] Jurjani AI, Al-Sharif AS. Kitab al-Ta'rifat. Al-Hakawati; 2014.

[10] Anis I. Min Asrär Al-Lugah. Among Language Secrets.1975.

[11] R. M. Baalbaki, Comparative philology of the Arabic language. Beirut House of Science for Millions, 1999.

[12] De Roeck AN, Al-Fares W. A morphologically sensitive clustering algorithm for identifying Arabic roots. In proceedings of the annual meeting on association for computational linguistics 2000 (pp. 199-206). Association for Computational Linguistics.

[13] Al Ameed H, Al Ketbi S, Al Kaabi A, Al Shebli K, Al Shamsi N, Al Nuaimi N, et al. Arabic light stemmer: a new enhanced approach. In the second international conference on innovations in information technology. 2005 (pp. 1-9).

[14] Boudlal A, Bebah MO, Lakhouaja A, Mazroui A, Meziane A. A markovian approach for Arabic root

extraction. The International Arab Journal of Information Technology. 2011; 8(1):91-8.
[15] Hmeidi II, Al-Shalabi RF, Al-Taani AT, Najadat H, Al-Hazaimeh SA. A novel approach to the extraction of roots from Arabic words using bigrams. Journal of the Association for Information Science and Technology. 2010; 61(3):583-91.
[16] Al-Kabi MN, Kazakzeh SA, Ata BM, Al-Rababah SA, Alsmadi IM. A novel root based Arabic stemmer. Journal of King Saud University-Computer and Information Sciences. 2015; 27(2):94-103.
[17] Abuata B, Al-Omari A. A rule-based stemmer for Arabic Gulf dialect. Journal of King Saud University-Computer and Information Sciences. 2015; 27(2):104-12.
[18] Boubas A, Lulu L, Belkhouche B, Harous S. GENESTEM: a novel approach for an Arabic stemmer using genetic algorithms. In international conference on innovations in information technology 2011 (pp. 77-82). IEEE.
[19] Frakes WB, Baeza-Yates R. Information retrieval: data structures and algorithms. Englewood Cliffs, New Jersey: Prentice Hall; 1992.
[20] Yousef N, Abu-Errub A, Odeh A, Khafajeh H. An improved Arabic word's roots extraction method using n-gram technique. Journal of Computer Science. 2014; 10(4):716-9.
[21] Ababneh M, Al-Shalabi R, Kanaan G, Al-Nobani A. Building an effective rule-based light stemmer for Arabic language to improve search effectiveness. International Arab Journal of Information Technology. 2012; 9(4):368-72.
[22] Al-Kabi M, Al-Mustafa R. Arabic root based stemmer. In proceedings of the international Arab conference on information technology, Jordan 2006 (pp. 1-7).

Hayel Khafajeh obtained his Ph.D. in Computer Information Systems in 2008 in Jordan. He joined Zarqa University, Jordan in 2009. In 2010, he served as Head of the CIS Department for two years. Since academic year 2014/2015, he has served as the Vice Dean of the IT College at Zarqa University. Professor Hayel Khafajeh has worked for 23 years in the educational field as programmer, teacher, head of IT division, and manager of ICDL center. He has published many educational computer books for the Ministry of Education in Jordan. In 2013, he published his Java Programming book for university students. His research interests include information retrieval and E-learning. He is the author of several publications on these topics. His teaching interests focus on information retrieval, C++ programming, Java programming, and DBMS (ORACLE & MS Access). In 2015-2017 he served as Chairman of the Committee for the Supervision of Computer Books for the seventh, eighth, ninth, tenth, eleventh and twelfth grades.
Email: hayelkh@zu.edu.jo

Nidal Yousef received his B.S. degree in 1997. In 2000, he earned his Master degree in Computer Information System. A Ph.D. received in 2008 in Computer Information System. He joined King Abdulaziz University, in KSA in 2009, as an Assistant Professor and in 2010 he moved to Al-Esra University in Jordan as Assistant Professor in the college of computing and Information Technology. In 2015 he obtained a scientific degree as an Associate Professor. He has published 9 research papers in International Journals and Conferences.
Email: nidal.yousef@iu.edu.jo

Mahmoud Abdeldeen received his B.S. degree in 1991. In 2006, he earned his Master degree in The Syntax of Arabic language. A Ph.D. received in 2011 in Syntax and Morphology Issues in Old Arabic Proverbs. He worked for one semester at the University of Philadelphia in Jordan, as a part time lecturer, and then he joined the National Chengchi University in Taiwan in 2015, as Assistant Professor. He has published 2 research papers in International Journals.
Email: mahmoudtalab@yahoo.com