

ENHANCING THE ACCURACY OF SONBOL'S ARABIC ROOT EXTRACTION ALGORITHM

Nisrean Thalji¹, Nik Adilah Hanin¹, Ziad Thalji² and Sohair Al-Hakeem³

(Received: 11-Sep.-2018, Revised: 5-Oct.-2018, Accepted: 25-Oct.-2018)

ABSTRACT

Root extraction is an important primary process in most Arabic applications, such as information retrieval systems, text mining, text classifiers, question answering systems, data compression, indexes, spelling checkers, text summarization and machine translation. Any weaknesses of root extraction will affect negatively the performance of these applications. Sonbol's Arabic root extraction algorithm achieves high accuracy of performance and gives new classification for Arabic's letters which minimizes the affix ambiguity. The comparison and testing of the existing Arabic root extraction algorithms on unify datasets shows that they still need some enhancements. Arabic root extraction is mainly based on using patterns, where as much as the algorithm has patterns as much as the accuracy is better. In this study, we improve Sonbol's Arabic root extraction algorithm, by enhancing its rules and increasing its patterns. We use 4320 patterns to extract the roots, which is the largest patterns' list extracted by Thalji's corpus. We test the new algorithm on Thalji's corpus that contains 720,000 word-root pairs. This corpus is mainly built to test and compare Arabic root extraction algorithms. The new algorithm is compared with Sonbol's Arabic root extraction algorithm. The algorithm of Sonbol et al. achieves an accuracy of 68%, whereas the new algorithm achieves an accuracy of 92%.

KEYWORDS

Arabic root extraction algorithm, Stemming, Arabic language processing.

1. INTRODUCTION

Arabic language is one of the most used Semitic languages. Semitic languages are spoken in a number of regions that were common in distant times in many regions of Africa and Asia over many decades. Some of these languages are not used now, such as Akkadian, Assyrian and Babylonian and some languages are still used nowadays, such as Arabic, Hebrew and Syriac. Semitic languages are a branch of the Afro-Asiatic language family originating in the Middle East (Bennett, 1998) [1].

In Arabic, vowels are used to ensure the exact meanings of words. If the word is non-vocalized, in many cases it is an ambiguous word and then we need to read the sentence and sometimes the whole text to understand the exact meaning. These vowels are written above or under the letter. Table 1 shows vowels in Arabic and corresponding letter/s in English. Some words in non-vocalized texts may have more than one meaning (ambiguous words). So, they have different roots. For example, the non-vocalized Arabic word "والدين/WALDN" has three possible words "والدين/WALEDAYN", "والدين/WA ADDAYN" and "والدين/WA ADDEEN". And then, the possible roots are "والد/WALAD" (son), "دين/DAYN" (debt) and "دين/DEEN" (religion). Another example is the non-vocalized Arabic word "كتب/KTB" which has three possible interpretations: "كتب/KATABA" (he wrote), "كتب/KUTIBA" (has been written), and "كتب/KUTUBUN" (books) [2]. We converted the Arabic letters and words into Latin characters (uppercase), so that the reader who is not familiar with Arabic texts can read it with ease. This way, the reader will be able to read words as the way they sound phonetically.

Root extraction is an important primary process in most Arabic applications, such as information retrieval systems, text mining, text classifiers, question answering systems, data compression, indexes, spelling checkers, text summarization and machine translation.

-
1. N. Thalji and N. Hanin are with Department of Computer Engineering, School of Computer and Communication Engineering, University Malaysia Perlis, Malaysia. Emails: nthalji1980@gmail.com, adilahanin@unimap.edu.my
 2. Z. Thalji is with Department of Management Information System, Imam Abdulrahman Bin Faisal University, Kingdom of Saudi Arabia. Email: zythalji@iau.edu.sa
 3. S. Al-Hakeem is with Department of Computer Science, Ajloun National University, Jordan. Email: drsohair@gmail.com

Table 1. Vowels in Arabic and corresponding letter/s in English.

No.	Vowels in Arabic	Corresponding letter/s in English
1	◌َ	E
2	◌ِ	O
3	◌ُ	A
4	◌ْ	No letter
5	◌َ◌ِ	En
6	◌ِ◌ِ	Un
7	◌ِ◌ُ	An
8	◌َ◌ِ◌ُ	Duplicate the letters

Therefore, many Arabic root extraction algorithms are presented in many different studies that tried to algorithm of Sonbol, Ghneim and Desouki [3]. The algorithm of Sonbol et al. comes after the algorithm of Khoja and Garside [4], which is well-known in extracting Arabic roots. Khoja and Garside algorithm's accuracy amounted to 95% when they tested their algorithm with their own dataset. Sonbol et al. tried to improve Arabic root extraction algorithms in order to increase the percentage of accuracy. The algorithm of Sonbol et al. accuracy amounted to 98% when they tested their algorithm using their own dataset.

Al-Shawakfa, Al-Badarneh, Shatnawi, Al-Rabab'ah and Bani-Ismael [5] made a comparison study of existing Arabic root extraction algorithms. This comparison included the algorithm of Sonbol et al., Khoja and Garside algorithm and other algorithms. This comparison was conducted in a unified dataset, in order to evaluate these algorithms fairly. Khoja and Garside algorithm's accuracy was 34%, whereas Sonbol algorithm's accuracy was 24%. Variance in the accuracy values is due to the differences of datasets that were used in the testing process. The study by Al-Shawakfa et al. revealed that existing Arabic root extraction algorithms still need more improvement. It also presented some weaknesses of Khoja and Garside's algorithm and the algorithm of Sonbol et al. In this study, we continue completing Sonbol's work by improving their algorithm.

There are three different approaches to extract the roots of the word; rule-based approach, lookup table approach and statistics-based approach. Recently, most of the root extraction algorithms are rule-based approach [5]. This approach mainly has two parts; the lists of affixes (roots and patterns) and the rules. Each Arabic root extraction algorithm tried to enhance the lists and/or the rules.

However, most algorithms suffer from the following problems:

- There are neither standard dataset nor complete lists of Arabic affixes, patterns and roots. Each work collects or generates their own dataset or lists. Most of the lists which contains the affixes, patterns, roots and lists that they used in each work are not publicly available. They just wrote samples of these lists. As a result, every time a new root extraction method is proposed, the researchers need to collect their own data or generate their own list independently. This will cause overlapped works, where each work is trying to solve the same problem instead of improving each other's work, which resulted a waste of time and resources. In addition, the lists used might have significant difference in terms of number of words, which will make it difficult for researchers to fairly compare the performance of existing works. Therefore, in recent works, the researchers tried to extend these lists by adding new contents [6].
- Arabic has a complex structure, which makes it difficult to extract the roots [7]. All Arabic root extraction algorithms suffer from affixes' ambiguity, so that it is difficult to distinguish between affix letters and root letters.

This work focus on solving these problems. The structure of this paper is organized as follows; in Section 2, different related previous studies and their drawbacks are discussed. Section 03 describes the proposed methodology which includes the details of each process. Section 0 explains the experimental implementation of our algorithm and its evaluation process. Section 5 concludes the main points of the paper and gives some future directions.

2. PREVIOUS STUDIES

In this section, we give a brief overview of previous rule-based Arabic root extraction algorithms. Khoja and Garside and Garside algorithm [4] is a very popular rule-based Arabic root extraction algorithm.

Khoja and Garside and Garside algorithm reported 96% accuracy of their stemmer using newspaper text. Al-Shalabi [8] presented an Arabic root extraction algorithm, which is a rule-based algorithm that is used to extract trilateral roots of Arabic words. This algorithm has been tested on a corpus of 72 abstracts 10582 words from the Saudi Arabian National Computer Conference, where its accuracy was about 92%.

Al-Kabi and Al-Mustafa algorithm [9] is based on affix removal. They tested their algorithm on small data sets containing 1,827 words. The system failed to analyze 55 words, since their patterns are unknown. This failure was mostly due to foreign (Arabized) words. The system enables to analyze the rest 1,772 words and achieved 91% of accuracy.

Sonbol, Ghneim and Desouki [3] algorithm is a rule-based root-extraction algorithm, the principal idea of which is based on encoding Arabic letters with a new code that preserves morphologically useful information and simplifies its capturing toward retrieving the root. They conducted their experiments using two different corpuses. The first corpus consisted of lists of word-root pairs 167,162 pairs. The second corpus was a collection of 585 Arabic articles from different categories (policy, economy, culture, science and technology and sport). This corpus consisted of 377,793 words. In general, the accuracy was about 96%-98%.

Another work is Ghwanmeh, Al-Shalabi, Kanaan, Khanfar and Rabab'ah algorithm [10], which is a rule-based algorithm used to find trilateral Arabic roots. According to Ghwanmeh et al., their algorithm has only failed to analyze words that are normally foreign, irregular or do not have trilateral roots. A corpus of 242 abstracts from the Proceedings of Saudi Arabian National Computer conference in machine-readable form was used in the testing procedure. The set of abstracts was chosen randomly from the corpus for analysis. The results obtained showed that the algorithm extracts the correct roots with an accuracy rate up to 95%. Many algorithms have been conducted under this type, like the Kchaou and Kanoun algorithm [11], El-Defrawy, El-Sonbaty and Belal algorithm [12] and Ayedh and Guanzheng algorithm [13].

Also, many morphological analyzers have been conducted to properly provide maximum morphological information on Arabic words, such as the proclitic, the prefix, the lemma, the suffix, the stem, the root, the enclitic, the tag and the pattern. One of them is MADAMIRA [14], a morphological analyzer that provides many information on Arabic words. MADAMIRA combines two morphological analysis systems; MADA [15], [16] and [17] and AMIRA [18].

In addition, Al-Khalil Morphological System 2 [19] is a recent morphological analyzer that provides many information on Arabic words. It deals with vocalized and non-vocalized Arabic words. It overcomes many errors of the previous system Al-Khalil Morphological System 1.

In general, all rule-based Arabic roots extraction algorithms share seven processes, which are: normalization, removing prefixes and suffixes, matching the word against patterns, extracting the roots from the patterns, comparing the roots with the roots' list, returning the extracted roots and finally making enhancement to extract the correct roots, as shown in Figure 1.

The difference between one algorithm and the others is in the details of each process. Also, every algorithm has different lists of prefixes, suffixes, roots and patterns. The main problem is in process two and process four. In process two, in many cases, the algorithms remove the matched prefixes and suffix letters, but these letters are part of the root. So, the result is a wrong root. Our proposed solution to this problem is done by not removing prefixes or suffixes and collecting more rules to reduce affix ambiguity. In process four, the algorithms match the word against patterns to extract the root. The main problem is the limited number of patterns that are collected till now. The extraction accuracy will improve if the algorithm can test as many as existing word patterns. Our proposed solution to this problem is using longer pattern lists. The proposed algorithm uses Thalji's pattern [6], which is the most recent list. This list is automatically generated from most of the Arabic dictionaries and contains (4320) patterns, which is the longest list discovered until now. These patterns contribute to enhancing the accuracy of Arabic root extraction algorithms.

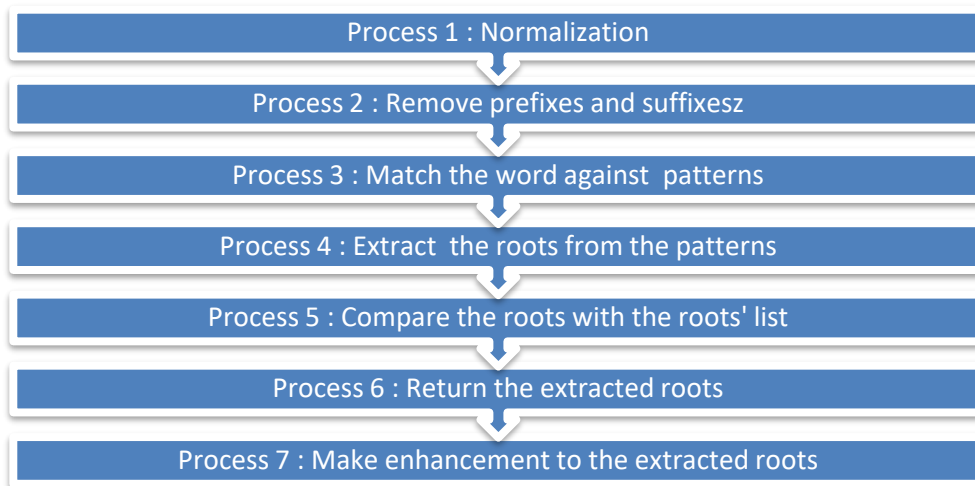


Figure 1. Main processes in rule-based Arabic root extraction algorithms.

3. METHOD

The root is the base form of the word that gives the main meaning of the word. In this section, the methodology for the proposed Arabic root extraction algorithm is explained. The proposed algorithm is an enhancement of the algorithm of Sonbol et al. by increasing the rules and the lists to find all possible roots of the word.

3.1 Normalization

The normalization steps for the algorithm of Sonbol et al. are as follows:

- Removing the kasheeda symbol (" _").
- Removing the diacritics.
- Replacing the Hamza's forms (ء, ؤ, آ, إ, ؤ, ؤ) with the letter (أ).

In this section, we extend Sonbol's normalization process by the following steps:

- Removing the punctuations.
- Removing the non-letters.
- Duplicating any letter that has the Shaddah: "ّ" symbol.

3.2 Encoding

In this step, Sonbol coded the Arabic letters based on six symbols {O, P, S, PS, U, A}, representing six groups of letters each of which shares certain characteristics:

O: Original letters. These letters are surely part of the root. They are: ("ث", "THAA"), ("ج", "JEEM"), ("ح", "HAA"), ("خ", "KHAA"), ("د", "DAL"), ("ذ", "THAL"), ("ز", "RAA"), ("ز", "ZAY"), ("ش", "SHEEN"), ("ص", "SAD"), ("ض", "THAD"), ("ظ", "DAA"), ("ط", "TAA"), ("ع", "AYEN"), ("ق", "GAF"), ("غ", "GAYEN"). This means that if the word contains one or more of these letters, these letters should be part of root's letters.

P: Prefix letters "ب, ف, ل, س" (BAA, FAA, SEEN, LAM). These letters should be treated as part of the root word if they appear in a different part of the word other than the prefix part. Otherwise, these letters are considered ambiguous letters (can be part of the root word or added letters to the root word). If these letters are ambiguous letters, the algorithm initially considers them as prefix letters. There is a possibility for these letters to become (O) letters (root's letters) in the next steps.

S: Suffix letter ("ه", Haa). This letter should be treated as part of the root word if it appears in a different part of the word other than the suffix part. Otherwise, this letter is considered as an ambiguous letter. If this letter is an ambiguous letter, the algorithm initially considers it as suffix letter. There is a possibility for this letter to become (O) letter (root's letter) in the next steps.

PS: Prefix-Suffix letters "ك, م, ن" (KAF, MEEM, NOON). These letters can appear only on both sides of the word; i.e., in the suffix part or in the prefix part. These letters should be treated as part of the root

word if they appear in a different part of the word other than the prefix-suffix part. Otherwise, these letters are considered ambiguous (can be part of the root word or added letters to the root word). If these letters are ambiguous letters, the algorithm initially considers them as prefix-suffix letters. There is a possibility for these letters to become (O) letters (root's letters), (P) prefix letters or (S) suffix letters in the next steps.

U: Uncertain letters ("ا, و, ي, ت") (TAA, YAA, WAW, ALEF). These letters can appear anywhere in the word. It is not possible to verify whether these letters are part of the root word letters. Several cases are associated with these letters as they may change, omit or convert from one letter to another during the derivation process following well known Arabic rules "EBDAL and EALAL". For example, the Arabic word "قيل/KEEL" (It was said) and its root "قول", the letter "و" is converted into "ي" during the derivation process.

A: Added letter ("ة" only (TAA MARBUTA)). This letter is always considered an additional letter. This letter is always deleted.

3.3 Some Improvements to the Last Coding

In this section, Sonbol et al. added some improvements to the last coding, by applying the following conditions:

- The letter "ب" (BAA) is a prefix letter if it is situated in the first three letters; otherwise, it is an original letter (part of the root word letter). This rule means that if the letter "ب" (BAA) is situated among the first three letters, it is an ambiguous letter. On the other hand, it will be part of the root word if it appears in place other than the first three letters of the word. For example, with the words "بأمرك, وبالصعيد" (BEAMRK, WBSA"EED), the letter "ب" (BAA) is situated among the first two letters. So, it's an ambiguous letter. Initially, the algorithm considers it as a prefix letter. Then in the next steps, it may change to (O) letter (root's letter). Another example is the word "أفبالبطل/AFBLBATL"; this word has the letter "ب" (BAA) appearing twice in the word. The first "ب" (BAA) letter is an ambiguous letter, as it appears as the third letter in the word. Initially, the algorithm considers it as a prefix letter. Then, in the next steps, it may change to (O) letter (root's letter). The second "ب" (BAA) will be considered part of the root word, as it appears as the sixth letter of the word.
- The letter "ف" (FAA) is a prefix letter if it appears among the first two letters; otherwise, it is an original letter. This means that if the letter "ف" (FAA) appears among the first two letters of the word, it is an ambiguous letter. If the letter "ف" (FAA) appears in a place other than the first two letters, it is part of the root word. Initially, the algorithm considers it as prefix letter. For example, in the words "فهد, فاستأجرتها, افغير" (FHD/FASTAJRTHA, AFGYR), the letter "ف" (FAA) appears among the first two letters of the word. So, it's an ambiguous letter. Initially, the algorithm considers it as a prefix letter. It may change to (O) letter (root's letter) in the next steps of the algorithm following certain rules. Another example; the words "متفائلين/MTFAELEEN", "تسافهت/TSAFHT", "وأنتفكت/WETFKT"; the letter "ف" (FAA) is considered to be part of the root word as it appears in places other than the first two letters of the word.
- The letter "س" (SEEN) is a prefix letter if it is followed by one of the letters "أ, ن, ي, ت" (HAMZA, NOON, YAA AND TAA); otherwise, it is part of the root word. For example, the words "سنتواجد, سينجلي, ساتغير" (SNTWAJD, SYNJLE, SATGYR); the letter "س" (SEEN) is an ambiguous letter, as it is preceding one of the letters "أ, ن, ي, ت" (HAMZA, NOON, YAA AND TAA). Initially, the algorithm considers it as a prefix letter. It may change to (O) letter (root's letter) in the next steps of the algorithm. Another example is with the words "كالأسد, مأسور" (KALASD, MASOR); the letter "س" (SEEN) is considered part of the root word, as it is not preceding one of the letters "أ, ن, ي, ت" (HAMZA, NOON, YAA AND TAA).
- The letter "ل" (LAM) is considered a prefix letter if it appears among the first five letters of the word; otherwise, it is part of the root word.
- The letter "ه" (HAA) is considered a suffix letter if it appears among the last three letters of the word; otherwise, it is part of the root word.
- The letter "ك" (Kaf) is considered a prefix letter if it appears among the first three letters of the word; otherwise, it is a suffix letter.

If the algorithm of Sonbol et al. finds three O-Letters (or more) in the encoded word, these letters are considered root letters and the algorithm is terminated. However, in this work, the enhancement of the algorithm of Sonbol et al. is to continue searching for other possible roots and for longer roots (more than three letter root word).

3.4 Applying Transformation Rules

In this section, the algorithm of Sonbol et al. applies transformation rules between groups to obtain a maximum number of original letters. Transformation rules are mentioned below:

- R1) Change each (P) after (O) to (O).

For example, with the word "ضيوف/ DYUF", "ض" (DAA) is an (O) letter, "ف" (FAA) is a (P) letter; in this case, (P) comes after (O). So, "ف" (FAA) is changed to (O) letter, which means that it should be one of the root's letters. Until now "ض, ف" (FAA, DAA) are part of the root's letters.

- R2) Change each (S) before (O) to (O).

For example, with the word "الهداية/ ALHDAYH", "د" (DAA) is an (O) letter, "هـ" (HAA) is a (S) letter; in this case, (S) comes before (O). So, "هـ" (HAA) is changed to (O) letter, which means that it should be one of the root's letters. Until now "د, هـ" (DAA, HAA) are part of the root's letters.

- R3) Change each (PS) before (P) to (P).

For example, with the word "كالسويف/ KALSOYUF", "ك" (KAA) is a (PS) letter, "س" (SEEN) is a (P) letter; in this case, (PS) comes before (P). So, "ك" (KAA) is changed to (P) letter, which means that it should be one of the root's letters or prefix letters, but not a suffix letter.

- R4) Change each (PS) before (O) to (P).

For example, with the word "منتقمون/ MNTKMON", "ن" (NOON) is a (PS) letter, "ق" (KAA) is an (O) letter; in this case, (PS) comes before (O). So, "ن" (NOON) is changed to (P) letter, which means that it should be one of the root's letters or prefix letters, but surely not a suffix letter.

- R5) Change each (PS) after (S) to (S).

For example, with the word "منتك/ MNTHK", "ك" (KAF) is a (PS) letter, "هـ" (HAA) is a (S) letter; in this case, (PS) comes after (S). So, "ك" (KAF) is changed to (S) letter, which means that it should be one of the root's letters or suffix letters, but surely not a prefix letter.

- R6) Change each (PS) after (O) to (S).

For example, with the word "علمك/ ELMK", "ك" (KAAF) is a (PS) letter, "ع" (A'A) is an (O) letter; in this case, (PS) comes after (O). So, "ك" (KAAF) is changed to (S) letter, which means that it should be one of the root's letters or suffix letters, but surely not a prefix letter.

- R7) Change each (P) after (S) to (O).

For example, with the word "التهبت/ ELTHBT", "ب" (BAA) is a (P) letter, "هـ" (HAA) is an (S) letter; in this case, (P) comes after (S). So, "ب" (BAA) is changed to (O) letter, which means that it should be one of the root's letters. Until now, "ب, هـ" (BAA, HAA) are root's letters.

- R8) Change each (S) before (P) to (O).

For example, with the word "البهتان/ ALBHTAN", "هـ" (HAA) is an (S) letter, "ب" (BAA) is a (P) letter; in this case, (S) comes before (P). So, "هـ" (HAA) is changed to (O) letter, which means that it should be one of root's letters. Until now, "ب, هـ" (BAA, HAA) are the root's letters.

As in the previous step, if the algorithm of Sonbol et al. has three O-letters in the encoded word, these letters are considered root letters and the process will terminate here. However, the enhancement of the algorithm of Sonbol et al. is to continue searching for other possible roots and for longer roots, with more than three root length.

3.5 Extracting All Possible Patterns of the Word

The algorithm of Sonbol et al. uses the idea of traditional algorithms, but with the aid of the encoded

word. Traditional algorithms store lists of Arabic prefixes, suffixes and patterns. These algorithms delete prefixes and suffixes, then use the pattern to extract the root from the remainder. The enhancement of this process is done by using larger lists' content and not removing the prefixes or suffixes, but applying the patterns. It's worth to mention that the presented algorithm assumes that patterns are composed of three elements: prefix, stem and suffix. One of the problems that were experienced by most of the previous algorithms is that they delete clitics before comparing with patterns. And in many cases, these clitics are parts of roots and not clitics. For example, with the words "كالحون, التقى" (Altka, kalehon), removing "كال, ال" (AL, KAL) will give these roots "حون, تقى" (TKA, HON), ignoring other possible roots "كالح, لقي" (LKA, KLH). In this section, we use the pattern's list of Thalji's corpus that was automatically extracted [6]. Up until now, this corpus contains the largest list of 4,320 patterns, which is the most appropriate list to be used in this work. Previous algorithms have used short lists that were manually collected. In addition, they did not publicly publish all the lists' contents. Thalji's patterns are listed in appendix A, so that future researchers can benefit from them.

In this step, we compare the word with the Thalji's list of patterns and return all matched patterns. For example, for a word like "فهد/ FHD", the algorithm found two original letters, "د, هـ" (DAA, HAA). Next, the word is compared to the list of patterns and all matched patterns were returned. The word "فهد/ FHD" matches the pattern "فعل/ FA'L". The word "فهد/ FHD" is the first possible root. Another example is the word "البحر/ ALBAHER", where the algorithm just finds two original letters, which are "ر, ح" (RAA, HAA). The word is compared to the list of patterns and all matched patterns were returned. The word "البحر/ ALBHR" matches the pattern "الفعال/ ALFA'L". The word "بحر/ BHR" is the first possible root. Also, this word matches the patterns "افعل, فعلل" (FA'LL/ EFA'L), then "البحر, البحر" (ALBHR/ ABHR) are also possible roots.

3.6 Extracting All Possible Roots for the Word

All possible roots are found by matching the words against the list patterns. All the possible roots that match the patterns are extracted after finding all possible patterns.

3.7 Solving the Problems with Ealal Rules and Ebdal Rules

When we have a weak letter (ALEF, YAA AND WAW), we replace this letter with the two other letters and check if the result is a valid root. If so, we add this root to the possible roots. For example, in the word "قال/ KAL", the algorithm replaces "ا/ ALEF" with "ي/ YAA" and "و/ WAW". So, "قول, قيل" (KEEL, KAWL) are possible roots.

3.8 Minimizing Possible Roots by Comparing them with Roots' List

In this section, we use the roots' list of Thalji's corpus that was automatically extracted from most well-known Arabic dictionaries. It is the largest roots' list found till now with 12,000 roots. This list is longer than the list that is used in Ababneh, Al-Shalabi, Kanaan and Al-Nobani stemmer [20]. It has about 11,347 roots. The distribution of roots for these two different lists is shown in Table 2.

Table 2. The distribution of the roots for two different lists.

Roots	List of Thalji's corpus	List of Ababneh, Al-Shalabi, Kanaan and Al-Nobani stemmer
Two-letter roots	500	115
Three-letter roots	7912	7198
Four-letter roots	3180	3739
Five-letter roots	360	295
Six-letter roots	48	0

The presented algorithm uses Thalji's list to minimize the possible roots, whereas the algorithm of Sonbol et al. used a short list of roots. For example, in the word "البحر/ ALBHR", the possible roots are "بحر, البحر, البحر" (BHR, ALBHR, ABHR), while the roots "البحر, البحر" (ALBHR, ABHR) are excluded, because they are not found in the roots' list.

4. EXPERIMENT AND EVALUATION

In this section, the presented algorithm is compared with other algorithms with the same approach, which is the rule-based approach. These algorithms are Khoja and Garside's Arabic root extraction algorithm and Sonbol's Arabic root extraction algorithm. In addition, the presented algorithm is also compared with one of the most recent morphological analyzer systems, which is Al-Khalil Morphological System 2. Al-Khalil Morphological System 2 gives maximum morphological information of Arabic words, such as the proclitic, the prefix, the lemma, the suffix, the stem, the root, the enclitic, the tag and the pattern. A complete comparison was conducted between the algorithms on Thalji's corpus in terms of accuracy. Thalji's corpus is an automatic corpus that is built from ten old Arabic dictionaries; this corpus is mainly built to test and fairly compare Arabic root extraction algorithms. This corpus contains 720,000 word-root pairs, which helps to avoid the interference of a human expert normally needed to verify the correct roots of each word used in the testing or comparison process. Moreover, this corpus has more than 4,320 types of words derived from 12,000 roots. Therefore, the list used in this experiment is more comprehensive compared to previous works.

The result of testing shows that the accuracy of Khoja and Garside's algorithm was 63%, the accuracy of the algorithm of Sonbol et al. was 68%, the accuracy of Al-Khalil Morphological System 2 was 75%, whereas the accuracy of the presented algorithm was 92%. Figure 2 shows the performance accuracy of all compared algorithms.

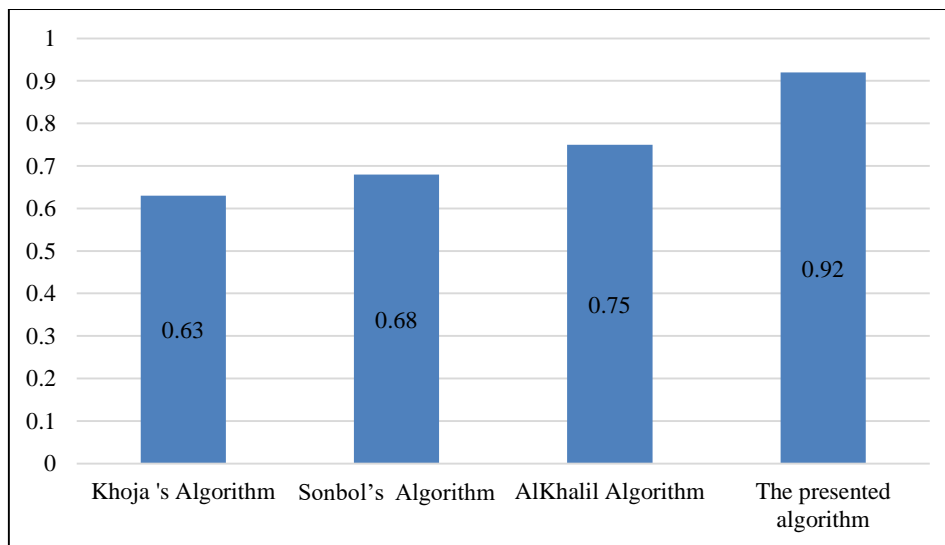


Figure 2. Accuracy of the algorithm of Khoja and Garside, the algorithm of Sonbol et al., the algorithm of Al-Khalil and the presented algorithm.

The main problems of Khoja and Garside's algorithm are that it does not consider many roots, prefixes, suffixes and patterns. It suffered from affix ambiguity problems. In addition, it returned just one solution for non-vocalized words, ignoring other possible solutions. Besides, it replaced a vowel letter with the letter "و" that sometimes returns a root that is not related to the derivation word. Finally, it produced wrong roots being unsuccessful to extract roots for derivation words that contain the "ابدال/ EBDAL" rule.

The main problems of the algorithm of Sonbol et al. arise if the root does not contain any constant letter, if the root does not start with a constant letter or if the root contains only one constant letter. Also, it does not consider many roots, prefixes, suffixes and patterns. In addition, it returned just one solution for non-vocalized words, ignoring other possible solutions.

The main problems of Al-Khalil Morphological System 2 are that it failed to analyze some words, which were about 25% of the input words. Table 3 shows a sample of these words. For example, the words

“ارصدت, مقاصر, مدارج, جارودة, شجير” (ARSDT, MKASR, MDARJH, JARODH, SHJER) are straightforward to find the roots, because they contain three original letters, but the algorithm fails to analyze them. And in some cases, it returns non-acceptable roots. For example, with the word “التقى/ALTKA”, the generated roots are “وقى, لقي”(WKA,LKE), where the root “وقى/WKE” is not an acceptable root, because “ل/LAM” letter cannot be an infix letter. Also, the algorithm matches the word to the wrong pattern, which is “فاة/FA”AH”. In addition, the algorithm fails to find all possible roots of non-vocalized words, like the word “التقى/ALTKA”, where it doesn’t return the possible root “تقى/TKA”.

The main problems of the proposed algorithm are that it fails to extract the root of derivation words with one letter length. In Arabic language, there are some few derivation words with one letter length, like “ق, ر, ع”(KE, RE, A’E). These derivation words are derived from a weak root with a length of three letters and these weak letters are deleted during the derivation process.

Table 3. A sample of unanalyzed words by Al-Khalil Morphological System 2.

سخرىا	البضيع	دواعب	ارصدت	مقاصر	مضمار	مدارج	قضم	عواطف	عفير
الجريم	الجبابة	الجارودة	الشجير	التواجد	التماجيد	التقدمية	التساخين	البهيم	البنود
الرجيع	الرجوليه	الدبسه	الخليف	الخبيص	الحوارد	الحكومات	الحريصه	الجورب	الجفلى
الطوارف	الضراغه	الصناديد	الشريم	السميد	السبله	الزهرة	الزغاليل	الرعاديد	الرحيق
زهيره	رصيد	دهريا	جوشا	بالمناقيش	المغيرة	المراشد	اللطم	العنزي	العريكة

Another case in which the proposed algorithm still fails is to find the root of derivation words as in the word “درهم/DERHAM”. The algorithm produces these roots “در, درأ, دري, دور, ودر”(WDR, DWR, DRE, DRA, DRR). In this derivation word’s matter, the algorithm finds two constant letters in the derivation word and tries to find the third constant letter in order to produce trilateral roots. However, the proposed algorithm is stopped to continue looking for the fourth one.

The proposed algorithm and Al-Khalil Morphological System 2 produce more than one possible root of the derivation words. In contrast, Khoja and Garside’s algorithm and the algorithm of Sonbol et al. produce just one root. In this section, the proposed algorithm and Al-Khalil Morphological System 2 are evaluated in terms of the average of possible roots per word and the number of processed words per second. The result is summarized in Table 4.

Table 4. Comparison between the proposed algorithm and Al-Khalil Morphological System 2.

The algorithm	The average of possible roots per word	The number of processed words per second
The proposed algorithm	3	101
Al-Khalil Morphological System 2	5	105

5. FUTURE WORK

The presented algorithm particularly contributes to enhancing the algorithm of Sonbol et al. by increasing its rules and extending its lists’ contents by using Thalji’s lists. The presented Arabic root extraction algorithm is compared with Khoja and Garside’s Arabic root extraction algorithm, Sonbol’s Arabic root extraction algorithm and Al-Khalil Morphological System 2. The testing and comparing processes are conducted on Thalji’s corpus, where the result of testing shows that the accuracy of Khoja and Garside’s algorithm was 63%, whereas the accuracy of the algorithm of Sonbol et al. was 68%, the accuracy of Al-Khalil Morphological System 2 was 75%. The presented algorithm achieved an accuracy of 92%.

In future, we plan to enhance the accuracy of the presented algorithm, overcome some weakness points and enhance the result to return just the exact root word. In order to implement this, the system must have the ability to understand the whole sentence or sometimes the whole paragraph.

REFERENCES

- [1] W. Abo Thuaaib, *History of Sematic Languages*, Lebanon: Darul Kalam for Pub. and Printing, 2016.
- [2] A. Al-Taani and S. A. Al-Rub, "A Rule-based Approach for Tagging Non-vocalized Arabic Words," *The International Arab Journal of Information Technology*, vol. 6, no. 3, pp. 320-328, 2009.
- [3] R. Sonbol, N. Ghneim and M. S. Desouki, "Arabic Morphological Analysis : A New Approach," *Information and Communication Technologies: From Theory to Applications*, Proc. of the IEEE 3rd International Conference, pp. 1-6, 2008.
- [4] S. Khoja and R. Garside, "Stemming Arabic Text," Computing Department, Lancaster Univ., UK, 1999.
- [5] E. Al-Shawakfa, A. Al-Badarneh, S. Shatnawi, K. Al-Rabab'ah and B. Bani-Ismail, "A Comparison Study of Some Arabic Root Findings," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 5, pp. 1015-1024, 2010.
- [6] N. Thalji, N. A. Hanin, Y. Yacob and S. Al-Hakeem, "Corpus for Test, Compare and Enhance Arabic Root Extraction Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, pp. 229-236, 2017.
- [7] M. Sawalha and E. Atwell, "Comparative Evaluation of Arabic Language Morphological Analyzers and Stemmers," *Proc. of COLING 22nd Inter. Conference on Comptational Linguistics*, pp. 107-110, 2008.
- [8] R. Alshalabi, "Pattern-based Stemmer for Finding Arabic Roots," *Information Technology Journal*, pp. 38-43, 2005.
- [9] M. N. Al-Kabi and R. Al-Mustafa, "Arabic Root-based Stemmer," *Proceedings of the International Arab Conference on Information Technology*, 2006.
- [10] S. Ghwanmeh, S. Rabab'Ah, R. Al-Shalabi and G. Kanaan, "Enhanced Algorithm for Extracting the Root of Arabic Words," *Proc. of the 6th International Conference on Computer Graphics, Imaging and Visualization*, pp. 388-391, 2009.
- [11] Z. Kchaou and S. Kanoun, "Arabic Stemming with Two Dictionaries," *IEEE International Conference on Innovations in Information Technology*, pp. 688-691, 2008.
- [12] M. El-Defrawy, Y. El-Sonbaty and N. Belal, "A Rule-based Subject-correlated Arabic Stemmer," *Arabian Journal for Science and Engineering*, vol. 41, no. 8, pp. 2883-2891, 2016.
- [13] A. Ayedh and T. Guanzheng, "Building and Benchmarking Novel Arabic Stemmer for Document Classification," *Journal of Computational and Theoretical Nanoscience*, vol. 13, no. 3, pp. 1527-1535, 2016.
- [14] A. Pasha, M. Al-Badrashiny, M. Diaby, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R. Roth, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Japan, 2014.
- [15] N. Habash and O. Rambow, "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop," *Proceedings of the 43rd Annual Meeting of Association for Computational Linguistics*, pp. 573-580, Association for Computational Linguistics, Michigan, 2005.
- [16] N. Habash, O. Rambow and R. Roth, "MADA+ TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization," *Proc. of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Egypt, 2009.
- [17] N. Habash, R. Roth, O. Rambow, R. Eskander and N. Tomeh, "Morphological Analysis and Disambiguation for Dialectal Arabic," *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, 2013.
- [18] M. Diab, K. Hacioglu and D. Jurafsky, "Automated Methods for Processing Arabic Text: from Tokenization to Base Phrase Chunking," *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, Kluwer/Springer, 2007.
- [19] M. Boudchiche, A. Mazroui, M. Bebah, A. Lakhouaja and A. Boudlal, "Al-Khalil Morphological System 2: A Robust Arabic Morpho-syntactic Analyzer," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 141-146, 2017.
- [20] M. Ababneh, R. Al-Shalabi, G. Kanaan and A. Al-Nobani, "Building an Effective Rule-based Light Stemmer for Arabic Language to Improve Search Effectiveness," *Int. Arab Jour. of IT* vol. 9, no. 4, 2012.

ملخص البحث:

إنّ استخراج الجذور عملية أساسية مهمة في غالبية تطبيقات اللغة العربية، مثل: أنظمة استرجاع المعلومات، وتحليل النصوص، وتصنيف النصوص، وأنظمة الإجابة عن الأسئلة، وضغط البيانات، والفهرسة، والتدقيق الإملائي، وتلخيص النصوص، والترجمة الآلية. وإن أي ضعف في استخراج جذور الكلمات من شأنه أن يؤثر سلباً على الأداء في تلك التطبيقات.

تُحقق خوارزمية سُنْبُل لاستخراج الجذور دقة عالية في الأداء وتقدّم تصنيفاً جديداً للأحرف العربية من شأنه أن يقلل كثيراً من الغموض المتعلق بالبادئات واللاحقات. وتُبين المقارنة والفحص للخوارزميات المتاحة لاستخراج الجذور في اللغة العربية أن تلك الخوارزميات لا تزال في حاجة إلى بعض التحسينات.

يعتمد استخراج الجذور في اللغة العربية على استخدام الأوزان، إذ تزداد دقة العملية بزيادة عدد الأوزان. في هذه الدراسة، يجري تحسين خوارزمية سُنْبُل لاستخراج الجذور من خلال تحسين قواعدها وزيادة أوزانها. وتم استخدام 4320 وزن لاستخراج الجذور، وتعد هذه القائمة أطول قائمة أوزان تم استخراجها من مجموعة قواعد بيانات "تلجي". كما تم فحص الخوارزمية الجديدة باستخدام تلك المجموعة من قواعد البيانات التي تحتوي على 720000 زوج من الجذور والكلمات. والجدير بالذكر أن مجموعة قواعد البيانات المذكورة إتّمت بُنْيَت لفحص خوارزميات استخراج الجذور ومقارنتها.

لقد تمت مقارنة الخوارزمية الجديدة بخوارزمية سُنْبُل لاستخراج الجذور، وكانت النتيجة أنّ خوارزمية سُنْبُل حققت دقة بلغت 68%، في حين بلغت دقة الخوارزمية المحسّنة الجديدة 92%.



This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).