Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Text mining: A survey of Arabic root extraction algorithms

Manar Ahmed Mohammed Hamza [1, 2], Tarig Mohamed Ahmed [3, 4], Anwer Mustafa Mohamedsalih Hilal [1, 2, *]

[1]Department of Computer and Self Development, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
[2]Faculty of Computer Science and Information Technology, Omdurman Islamic University, Omdurman, Sudan
[3]Department of Information Technology, Faculty of Computing and Information Technology, King Abdul-Aziz University, Jeddah, Saudi Arabia
[4]Department of Computer Sciences, University of Khartoum, Khartoum, Sudan

ARTICLE INFO

Article history:
Received 6 May 2020
Received in revised form
22 July 2020
Accepted 17 August 2020

Keywords:
Accuracy
Arabic word root
Stemming algorithm
Text mining

ABSTRACT

In all Arab countries, the Arabic language is the official language spoken and written and is one of the oldest known languages. This paper aims to explain and discuss the work done on extracting the root of the Arabic word and Stemming algorithms. Text mining has become of interest to scientists, researchers, and users because of the existence of big data and deep learning algorithms that can analyze giant sets of unstructured data. The basic algorithms are used to extract and classify texts, information retrieval systems, and indexes. Algorithms are used to extract the root of a word from different natural languages. This paper will present a brief background and comprehensive presentation of a number of algorithms that handle the Arabic text to extract the word root in its light, heavy, hybrid, leading, and Markovian form. There are a number of papers, articles, and research papers that deal with extracting the Arabic root from the word. This paper will present a brief background for a number of stemming algorithms on how to extracting the root and stem of the Arabic word, then make a comparison and discussion of a number of selected algorithms in terms of accuracy, data set, method of stemming regarding of strengths and weakness.

1. Introduction

The Arabic language defines as a complex language based on root-pattern format. Application of the natural language processing includes multiple applications such as text processing, speech tagging information retrieval, machine translation, and the most important topic is the context is root extraction and stemming.

We can use stemming algorithms in text mining, text classification, information retrieval systems, and indexers. A lot of stemming algorithms are built in different natural languages. We will introduce an overview of multiple algorithms work in this field, such as the light, heavy, hybrid, novel, and Markovian Arabic stemmer. We found that there are multiple algorithms that made a comparison between their own stemmer and Khoja stemmer (Khoja and Garside, 1999), which is the standard Arabic stemmer.

When developing Information Retrieval (IR) systems, Arabic light stemmers have been developing by many approaches to use in multiple applications and projects. Researchers select which stemmer to use in their project after evaluate and compare between stemmers.

2. Related works

In this section, we will present and compare multiple papers and researches worked in data mining, text mining, Arabic morphological analysis, and Arabic root extraction and stemming.

A literature survey allows researchers and readers to simply get to investigate a specific subject by selecting high-quality articles or considers that are related, significant, important, and substantial and summarizing them into one complete report. Moreover, it gives a great beginning point for analysts starting to investigate in an unused region by constraining them to summarize, assess and compare unique investigation in that particular zone, and it makes sure of that don't duplicate work that has as of now been done.

* Corresponding Author.
Email Address: a.hilal@psau.edu.sa (A. M. M. Hilal)
https://doi.org/10.21833/ijaas.2021.01.002
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-4658-8941

It can give clues as to where future inquire about is heading or prescribe regions on which to center. It gives a supportive examination of the ways and approaches of other researchers. In the following section, there are related works we will discuss them.

2.1. Root extraction algorithm

Boudlal et al. (2011) presented an Arabic morphological examination framework that appoints, for each expression of a vowel Arabic sentence, an individual root contingent upon the condition. The proposed framework is made out of two modules. The first comprises of an investigation. It separates each expression of the sentence into its rudimentary morphological units to distinguish its potential roots. For that, they embrace the division of word into three little portions (prefix, stem, and suffix). In the subsequent module, they utilize the condition to recognize the right root among all the potential underlying foundations of the word. For this reason, they utilize a Concealed Markov Models approach, where the supervisions are the words and the potential roots speak to the shrouded states. They approve the methodology utilizing the NEMLAR Arabic composition corpus comprising of 500,000 words.

The framework finds the percentage of the right root in the preparation set is over 98% and in the testing set of words, practically is 94%. It can generally have excellent outcomes in picking the right root of the word.

After the effects of tests did on the two pieces of the framework are empowering. They can be improved by further investigation of hamzated words in the examination out of condition and by utilizing a bigger corpus in the Markovian methodology. They broaden crafted by the principal module so as to make different labels of the words (thing, action word, molecule, (a word that portrays a thing), (a word that depicts an action word or a descriptor), conceivable vowelizations).

They utilize a supportive difference in the Markovian way to deal with recognize the best vowelization of the word in the setting.

Al-kabi et al. (2015) presented another light and heavy Arabic stemmer and contrasted with two surely understood Arabic stemmers. The result demonstrated the accuracy of the proposed stemmer is somewhat low to those two stemmers. The tests on novel stemmers show accuracy 75.03%, while low accuracy show by the two Arabic stemmers.

They proposed, created, and evaluated another Arabic stemmer. Three principles handling root starting exploratory outcomes demonstrated a satisfactory for roots foreseeing. They contrasted their stemmer, and two Arabic stemmers, where the equivalent dataset is staged, were applying to make Arabic roots from words. Stage 1 is in charge of expelling prefixes and suffixes, Stage 2 is in charge of contrasting yield with standard word sources or shapes, and stage 3 is in charge of revising the developing utilized.

Results demonstrated that their calculation is better regarding accuracy much of the time of various word lengths in examination with the other two Arabic stemmers.

Alkhatib et al. (2017) proposed another algorithm, which is called a novel methodology for building up Al-Hadith Al-Shareef WordNet linguistic to fills its needs for various tasks of Arabic natural language processing. Particularly, they build up semantic associations between words in order to achieve a decent comprehension of the implications of the words in Al-Hadith. Their procedure is to use the ontology of Al-Hadith and Traditional Arabic lexicons.

This algorithm capacity was demonstrated in a classification that they created for procedure estimation. The classifier has been applied on around 8500 synsets that incorporate 6126 titulars, 310 adjectives, 1990 verbal and 71 adverbial expressions.

Taghva et al. (2005) executed an Arabic stemmer for root extraction, which is the same as the stemmer used by Khoja, but it does not use the root dictionary. It is a light stemmer and applied to the Arabic Trec-2001 collection.

The result of testing demonstrated that in an Arabic stemmer, there is no need for stem lists. Larkey et al. (2002) found that in general, the performance of a light stemmer such as the proposed one that removes affixes without pattern more complex stemmers and the Khoja stemmer, which uses stem lists and pattern checking.

Kreaa et al. (2014) proposed a new stemming system (AKK stemmer) for Arabic words, which consolidates Light Stemmer and Looks in tables strategies to take care of the issue of the broken plural, which is the irregular nouns in the Arabic language.

AKK stemmer furnishes exact outcomes in examinations with a different algorithm. They made a correlation between the algorithms; they utilize general techniques are two. Either like Khoja stemmer, which extracting the root of the word, or like Light stemmer that truncation of affixes. The primary technique has numerous issues. Mainly, the root word dictionary expects support to ensure.

Ababneh et al. (2012) presented another light stemming system. It is a rule-based light stemmer. They presented another arrangement of PC directions that uses a lot of principles to choose if a specific succession of characters is a piece of the first word or not all that this can help understanding some confounding problems. Additionally, they presented a route for dealing with most broken plural structures and diminishing them to their single to gathering expressions of a similar significance in a typical structure.

Al-Omari et al. (2013) introduced an algorithm, which finds the Arabic word root. It uses a set of mathematical rules and relations between letters in Arabic light stemming. It was called the Arabic Rule-

Based Light stemmer (ARBLS). It is tested and compared with a Khoja stemmer. The tests and their results had shown a major margin of difference in favor of ARBLS. It needs to be improved its abilities to extract Arabic roots of a large number of Arabic words correctly.

Otair (2013) explained that the definitions concerning hybrid stemming approaches where analyzed then summarize the main characteristics of the Arabic language.

This paper expects to look at the greater part of the usually utilized light stemmers in terms of affixes lists, algorithms, main ideas, and information retrieval performance. The outcomes demonstrate that the light 10 stemmer outperformed the other stemmers.

Al-Lahham et al. (2018) illustrated Light 10 stemmer; it is the best one among a grouped of light stemmers. It defines a table with a list of suffixes and prefixes with better retrieval and high performance. Light 10 has no confinements on the affixes, so it is possible to have two distinct terms having the same symbol while they have different meanings. Light 10 stemmer proposes adding to the table more affixes and power a few conditions on removing these affixes. The accomplishment and testing of the proposed strategy show high quality than the Light 10 stemmer.

The proposed stemmer recommends removing affixes if they fulfill one or more of a set of proposed conditions. The utilization of the proposed light stemmer demonstrates that adding a few conditions to light stemmers improves the retrieval at the lower recall levels.

El-Defrawy et al. (2015) estimated different Arabic stemmers by doing a progression of comparisons utilizing a series of comparisons using a manually annotated dataset, which shows the performance of Arabic stemmers, and points out potential enhancements to existing stemmers. They also present improved root extractors by using light stemmers as a preprocessing stage.

The study and the results did show that there is a relationship between linguistic accuracy and other measures. If linguistic accuracy increases, the other related measures will increase. The more Arabic stemmers exist, it will make the stemming analysis job richness. Any stemmers have their own strengths and weaknesses, where the weaknesses could be reduced by combining many stemmers in effective ways.

Yaseen and Hmeidi (2014) presented a new algorithm called WSS. It does not remove any affixes. It creates a set of all substrings of an Arabic word and employments a set of rules to extract root from substring Arabic, roots file, and Arabic patterns file. The accuracy of the proposed algorithm is 83.9%. The algorithm utilized the Holy Quran for testing. This stemmer considers as competitive, and the accuracy can be moved up to 9.9% after doing multiple tests. In most cases, two candidates for the proper root were retrieved by the WSS algorithm.

Almusaddar (2014) focused on improving Arabic information recovery by making strides light stemming and preprocessing arranges and includes to the open-source community, moreover, construct a rule for Arabic alteration and stop-word evacuation. To achieve these objectives, he makes a GUI toolkit that performs reprocessing arrange that's essential for information retrieval. One of these steps is alteration, which we made strides and presented a set of rules to do and advance by other researchers.

The following reprocessing step they made strides is stop-word removal, the presented two diverse stop-word lists, the primary one is an intensive stop-word list for decreasing the measure of the file and befuddling words, and the other is a light stop-word list for better results with recall in information retrieval applications. He presents the utilize of Arabized words, 100 words manually collected, these words should not follow the stemming rules since they came to the Arabic language from other languages, and show how these progress results compared to two well-known stemming algorithms like Khoja and Larkey stemmers. The proposed toolkit was combined with a prevalent IR platform known as the Terrier IR platform. He utilized the TF-IDF scoring model from the Terrier IR platform and tested the results utilizing OSAC datasets. He utilizes an existing open-source application that already bolsters other languages, at that point, including the Arabic dialect back to it. He progressed the preprocessing step, which affects the results of any IR framework. The proposed GUI toolkit has numerous alternatives, including reading and writing dataset files, show output in tables, and create statistics around preprocessing steps. This toolkit might be accepted as the first step through a standard and may well be altered broadly within the Arabic language preprocessing and Arabic IR frameworks. Using UTF-8 is imperative and extraordinary alternatives, especially communicating with other schemes. The presented light-stop-word list that contains 119 words and seriously stops word list manually collected and merged from three other stop-words lists and contains 13957 stop-words. These alternatives permit more alternatives for researchers to test and move forward the impact of stop-words evacuation on a different application like TC or IR. They declare the utilize of Arabized words and clear how these words must not comply with any Arabic stemming rules since these words are not Arabic words, they had collected 100 Arabized words and incorporate them within the progressed light stemming calculation to progressing the effect of any stemming algorithm.

They made a comparison of a dataset that contains Arabized words with two well-known Arabic stemming Larkey that failed with 32 Arabized words and Khoja stemmers that failed with 5 Arabized words. The proposed toolkit increments the preprocessing for IR frameworks and permits simple creating and (a combination of distinctive and other groups. Terrier IR now bolsters Arabic

language utilizing the proposed toolkit and offers wide alternatives for preprocessing information before indexing it.

Larkey et al. (2002) presented numerous light stemmers based on co-occurrence for Arabic recovery. The recovery adequacy of proposed stemmers and a morphological compared to the TREC-2001 information. The finest light stemmer was more viable for cross-language retrieval than a morphological stemmer, which attempted to discover the root for each word. A repartitioning prepares to comprise of vowel expulsion taken after by clustering utilizing co-occurrence examination created stem classes which were way better than no stemming or exceptionally light stemming, but still the second rate to good light stemming or morphological analysis. It appears advancements of around 100% in average high quality due to stemming and related forms, and an indeed bigger impact for dictionary-based cross-language recovery. An online dictionary offers assistance to extricate words, so it contained far fewer distinctive. Without stemming, the dictionary interpretations of question terms were improbable to coordinate the shapes found in documents.

Boudchiche and Mazroui (2018) clarified an Arabic root extraction framework that gives the root of each word of a given sentence. It is a critical instrument for a part of natural language preparing applications such as search engines, text classification, and information recovery. The strategy of extraction utilized in this work runs in two steps. The primary one consists in trying to find of all the conceivable roots of each word carefully examined with the morphological analyzer Alkhalil Morpho Sys 2. Then, in the second step, a declaration approach is based on nonstop quadratic splines to select from these roots the one that matches the word context. They got encouraging results with an accuracy of 96%.

Sameer (2016) proposed a Modified light stemming algorithm for Arabic Languages. It is relay on the understanding of Arabic morphology. It is dependable and precise for words that have diverse length and distinctive affixes agreeing to the test but erroneously stems appropriate names and foreign words.

Boudad et al. (2018) surveyed the major works that had been managed to Sentiment Analysis in Arabic. This audit appeared that Arabic Sentiment Analysis had become one of the research ranges that have been drawn the consideration of many researchers. Examination of these works appeared that three sorts of approaches, to be specific administered, unsupervised, and hybrid, were utilized to handle an assortment of Estimation Investigation task.

Alhanini and Aziz (2011) discovered how to improve finding the Arabic words stem; the used stemmer is a light stemmer and dictionary. The improved stemmer incorporates the dealing with named entity recognition and word expressions. They have utilized an Arabic corpus that comprises ten records in arrange to figure out the improved stemmer. They detailed the improved stemmer accuracy values, light stemmer, and word reference-based stemmer in each document. The average of accuracy in an improved stemmer is 96.29%. The test shows that the improved stemmer fulfills the most elevated accuracy values, and it is superior to the dictionary-based and light stemmer.

Albogamy and Ramsay (2016) presented a light Arabic stemmer for Arabic tweets. The results increment the fulfillment of some well-known stemmers for Arabic. A new stemmer does not depend on any root dictionary, which is extremely imperative for stemming Arabic tweets, since they have a very open vocabulary. It has two stages: stage 1 is committed to creating a list of all conceivable stems by utilizing the grammar, and stage 2 is to select the shortest stem as the right stem. They compared the new stemmer with three Arabic stemmers, where one of them uses almost the same approach to the new stemmer. Results appeared that the accuracy is better to compare with the other three Arabic stemmers.

Momani and Faraj (2007) proposed extracting Arabic trilateral roots using a novel algorithm. The words that have no root where filter then remove the suffixes and prefixes and remove any repeated letters find in the Arabic word “sāltmwnyhā” after sorting term letters. Letter removal was conducted until three letters remain. Lastly, according to the order in the original word, the remaining letters will be arranged. Two Arabic text documents were chosen to make a performance test. After testing 1500 words, it produces the proper root, and the accuracy is 73%.

Al-Kabi (2013) displayed a standard Arabic stemmer called Khoja stemmer and explain its deformity. He makes a comparison between various studies and this one. Al-Kabi (2013) found that the Khoja stemmer is better than other ones assess in his study. These stemmer and Khoja stemmer based on Patterns, Shapes, weight. Adding more Patterns increase 5% of the accuracy.

Alshalabi (2005) presented a strategy for extricating the triliteral Arabic root for an unvocalized Arabic corpus. It gives a productive way to expel suffixes and prefixes from the curved words. Then it matches the coming about the word with the accessible designs to discover the appropriate one and, after that, extricates the three letters of the root by expelling all infixes in that design. This procedure does not utilize any lexicon to check the coming about the stem. A few rules had been describing that offer assistance to choose if the letters have a place to the root or not. This algorithm has been tested on a corpus of 72 abstracts (10582 words) from the Saudi Arabian National Computer Conference; the algorithm accuracy is approximately 92%.

Khafajeh et al. (2018) explained a crossover strategy to extricate Arabic word roots had been creating. The proposed method depends on optimization work, which is the improving operation

performed by playing a set of non-morphological rules to improve the n-gram method. The new method tried employing a dataset containing more than 6000 recognized words having a place to 141 distinctive roots. The results appear a stamped change after utilizing the crossover strategy; the proposed method extricates accurately approximately 99% of three-part solid roots and almost 86% of tripartite vowels roots.

The proposed strategy utilized multi-objective work with a measurable algorithm for finding Arabic roots, a multi-objective work utilized to avoid getting caught within the same roots by finding the best quality solution calculated utilizing modern proposed confinements.

Hawas (2013) presented how to assign an individual root without depending on a dataset of word roots, a list of all the prefixes and the suffixes of the Arabic words, or a list of word patterns. It tries to portray a possible case of the root-letters positions one by one based on a few rules and relations among the word letters and their situation within the word. It centers on two parts of the approach. The proposed approach had been assessed utilizing the Holy Quran words. The assessed results appear as a favorable extraction algorithm.

This approach comprises of two-stage. The first stage shows the ability to discover relations between the word letter and its situation within the word. The assessed result of this organization display a favorable root extraction calculation. The second stage represents an assessment of the classification of the Arabic letter. This stage tests the classification of Arabic word letters. A comparison is made between the roots letters produced each word and the ones put away into the roots file taking into account that the system is in its first stage. If the entire coordinate or sub coordinate is found, then the examination of a root is considered correct. On the other hand, in the case at least one letter created by analyzing the tried word is wrong, the root analysis is considered inaccurate.

AbuSafiya (2017) clarified a new strategy based on two phases: The first stage is the generation stage, which makes an introductory set of candidate roots. The second stage is a filtering stage where the root set that was made at the first stage filtered to remove the wrong roots. The primary power of this approach is that it treats distinctive mistakes of Arabic language morphology by making all conceivable roots, including those with remove, flipped, or repeated letters. The other advantage is that it is easy to put into utilize and can extend to effectively deal with new words of new derivation forms. This may be done by fixing the generator to include new roots in case the proper root isn't made within the generation stage and including modern channels to leave out wrong filters. This made the advancement of the system simple and can be improved.

Elazhary and Khodeir (2017) proposed a new approach called Art (Arabic word Root Extraction Tutor. It is a cognitive instructor implied to educate students generation rules required for Arabic word root extraction. It works in two-mode active and passive mode and combines numerous ways of doing things for progressed instructing. It gives a positive result for a correct answer and a negative one something else. Art could be a cognitive instructor implied to educate students generation rules required for Arabic word root extraction.

Hajjar and Zreik (2010) displayed a new method that assesses the implementation of some Arabic root extraction algorithm. The utilized strategies in this framework chosen agreeing to a previous ranking, where these strategies are classified into five groups. They have chosen a strategy for each group. These strategies are Arabic Stemming without a root dictionary, Light Stemmer, N-gram based on contrast coefficient, MT-based Arabic Stemmer N-gram based on the likeness coefficient. This estimation was conducted on the same terms in a corpus of two thousand words and their roots. These words are taken from the Arabic lexicon “Lesan Al-Arab.” This framework works in two ways: Typical and automatic.

The view of this method is to apply these strategies in the waterfall on the same entry to investigate the viability of each combination of methods and compare the execution of this new strategy with a combination of the strategies as of now created. In expansion, this framework can be distributed as a web location to permit all people groups to select their own corpus and to create a joining assessment of these methods.

Al-Kabi et al. (2011) presented an assessment of four heavy Arabic root based stemmers. The assessment showed the strength and accuracy of these stemmers.

Jaafar et al. (2017) discovered a new Arabic stemmer gives solutions to a few bad results. Also, it estimates and compares Arabic stemmers that take into consideration measurements related to the accuracy of results as well as the execution time of stemmers. The results appear that the stemmer finishes the highest rate of accuracy with 33.7% and occurs in the second position in terms of the Gs-Score metric with 0.1.

Nehar et al. (2016) presented a new Arabic root extraction approach dealing with Text Classification, a new approach utilizing transducers and rational kernels. It displays the premise to utilize Arabic Pattern-Based Stemmer designs. Transducers utilized to show these designs and root extraction were done on three-word collections without utilizing a dictionary. The accuracy yields 75.6%. Classification tests were done on the Saudi Press Agency dataset, and N-gram kernels, are tried with different values of N. accuracy and F1 report 90.79% and particularly 62.93%. These results mean that this approach is more accuracy and F1 than other approaches.

2.2. Related papers in text mining and data mining

Bharati and Ramageri (2010) examined a few of the data mining techniques, algorithms, and a few of the organizations which have adjusted data mining technology to improve their businesses and found great results. They clarify a few strategies of algorithms and techniques like Artificial Intelligence, Clustering, Regression, Classification, Neural Networks, Association Rules, Genetic Algorithm, Decision Trees, the Nearest Neighbor strategy utilized for knowledge discovery from databases. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be utilized to discover patterns and connections that would otherwise be difficult to discover. This innovation is well known for numerous businesses since it allows them to learn more about their customers and make smart marketing decisions. FBTO Dutch Protections Company, Provident Financials Domestic Credit Division, joined together Kingdom and Standard Life Common Monetary Services Companies are a few companies presented by this paper using data mining technology to found solutions for business problems.

Gridach and Chenfour (2011) described an approach for the Arabic morphological examination. It is called the Arabic Morphological Automaton (AMAUT). They have evaluated the presented approach utilizing Xerox Arabic Morphological Analyzer and Arabic Morphological Analyzer by Otakar Smrz because they are considered as the most referenced approaches for Arabic morphological investigation, and they are accessible for research and evaluation. There are a few preferences by utilizing the Arabic morphological automaton; it makes the framework portable and reusable since it's created utilizing Java dialect and XML innovation. Another advantage makes the morphological analyzer productive and exceptionally quick. Concerning the improvement of the lexicon, they have utilized XMODEL language for speaking to, planning, and actualizing the lexical resource.

Alsaad and Abbod (2014) showed an improved root extraction algorithm for Arabic words, which is based on morphological analysis and linguistic constraints. The algorithm removes prefixes, suffixes then checking the word against a predefined list of patterns. In expansion, a few issues of extricating the roots have been handled by recognizing phonetic based rules to re-place, eliminate or duplicate certain letters where required. The algorithm collects Arabic words from an online Arabic corpus then done an experiment and testing on it. The assessment of the results and accuracy of the algorithm was connecting by human judgment.

Saad and Ashour (2010) clarified and surveyed the existing common Arabic stemming, light stemming algorithms. They perform and combine Arabic morphological examination tools into the driving open-source machine learning and data mining tools, Weka, and Rapid Miner.

Salloum et al. (2018) presented a wide study of several considerations related to the Arabic text mining with more concentrate on the Holy Quran, estimation analysis, and web documents and their implementation. The study discusses the later development within the field of intelligent computing, and it gives a total rundown of the existing text mining methods, which can be utilized for the extraction of logical patterns from the grammatically incorrect and unstructured textual data.

3. Discussion

In linguistic morphology and information retrieval, stemming is the method of lessening bent words to their word stem, base, or root. Many algorithms for finding the stem or root from the Arabic word mentioned above, and we can classify some techniques used to extracting root as follow:

1. Root extraction using light stemmers.
2. Root extraction using light and heavy stemmers.
3. Root extraction-using rules and based on a dictionary.
4. Root extraction-using rules without a root dictionary.
5. A Markovian Approach for Arabic Root Extraction.
6. Pattern-based Stemmer for finding Arabic Roots.
7. Root extraction without removing affixes.
8. Root extraction using Rational Kernels and text classification.

We notice that many stemming algorithms are built in different techniques, and some of them depend on the Khoja stemmer, which was widely used and known. Table 1 shows the accuracy of the stemmer algorithm.

It explains the comparison between the accuracy, technique used, and type of the data set of eleven algorithms to extract Arabic root, which is mention in Table 1. In section two, a short description of multiple algorithms had been discussed. The algorithms were tested to utilize Holley Quran words, Arabic newspaper, website, Arabic Trec-2001, Corpus, and NEMLAR Arabic Writing Corpus as a dataset.

The results obtained from Alsaad and Abbod (2014) proposed root extraction algorithm is worth being connected in different Arabic language handling programs, and it is promising. The WSS approach occupies the third position through the algorithms. For this reason, it has competitive accuracy. Ghawanmeh's stemmer (Ghawanmeh et al., 2009) accuracy is 95%. It involves the first position, whereas Taghva's et al. (2005) stemmer accuracy is 38%, which is the slightest accuracy. Boudlal et al. (2011) stemmer accomplish in testing set 93.81% in preparation and in training set 98%.

Khoja stemmer begins by remove diacritics, punctuation, and non-characters of the input word.

Then predefine a set of paths, which is based on word length to let words follow these paths. Then define prefixes and suffixes that had been removed. Apply a set of linguistic rules. Finally, validate the extracted root against a set roots dictionary, then stop if the root is correct. If the extracted root is incorrect, the stemmer continues searching for other root possibilities. An exhaustive search is done if the stemmer doesn't find the root then marked as an un-stemmed word.

Table 1: The accuracy of the stemmer algorithm

No.	Algorithm	Technique	Type of Data Set	Accuracy
1	Khoja and Garside (1999)	Pattern and form	Arabic corpora (Arabic newspaper)	72.1%
2	Alshalabi (2005)	Pattern-based	Corpus	92%
3	Taghva et al. (2005)	Light stemmer and patterns	Arabic Trec-2001	38.0%
4	Boudlal et al. (2011)	A Markovian Approach	NEMLAR Arabic writing Corpus	93.81% In the testing set and 98% in the training set
5	Al-Kabi (2013)	Light Stemmer	Holy Quran	60.0%
6	Yaseen and Hmeidi (2014)	Arabic patterns file and set of rules	Holy Quran	83.9%
7	Alsaad and Abbod (2014)	Remove affixes, apply rules	online Arabic corpus	70.5%
8	Al-Kabi et al. (2015)	Improve Khoja Rule-base stemmer	Arabic newspaper and website	75.89%
9	Khafajeh et al. (2018)	Statistical technique	distinguished words	99% of strong ternary roots, 86% of ternary vowels roots.

Khoja stemmer does not explore all linguistic possibilities, and there are some missing Patterns not used, which is occurred in its accuracy. The Al-Kabi (2013), Taghva et al. (2005), and Yaseen and Hmeidi (2014) algorithms normalized words by removing affixes; it considers as an extra process deal to a major drawback to producing the wrong root because the stemmer cannot differentiate between extra letters (non-root) and root, another drawback is the overhead produced when extracting the root. The WSS-based algorithm reduces the time of extracting root without removing affixes. Al-Sarhan's stemmer (Al-Sarhan, 2003) assigning words weights (real numbers between 0-5) and ranks (order of a letter in a word) to the letters of words. Weights figure out by some tests on an Arabic text then multiplies the rank of a letter by its weight. Then, the three letters with the least product value were chosen as the three-letter root. This stemmer depends on the correctness of the weights of letters and the formula it uses, which could be tested by running some evaluations on a good dataset. The algorithm of Alshalabi (2005) shows an accuracy of about 92%. It normalizes the corpus by remove stop words, determiner, prefixes, and suffices, then reducing the inflected word. Some rules are added for removing suffixes and defined some constraints. These rules and constraints cannot be for all words, Boudlal et al. (2011) n A Markovian approach system deal with unvoweled words only. It gives correct root in training set more than 98% and in the testing set about 94%.

Many algorithms can be improved by adding extra rules and by combining the strength of two or more strong stemmers to deal with special cases that didn't follow the rules and to find the correct roots which are extracted wrong by some weak stemmers.

4. Conclusion

The Arabic internet content in the last years raised up the need for effective stemming techniques for the Arabic language. Arabic stemming algorithms can be classified into three categories, root-based approach (ex. Khoja), stem-based approach (ex. Larkey), and statistical approach (ex. N-Garm).

In this paper, we had displayed and discussed many papers and articles work on extracting Arabic word root, data, and text mining and explain the strength and weakness points in them.

The advantages of related works serve many purposes, several of which relate directly to reviewing, the person handling the submission will use the referenced papers to identify good reviewers, reviewers will look at the references to confirm that the submission cites the appropriate work, everyone will use the section to understand the paper's contributions given the state of existing research and future researchers will look to the Related Work section to identify other papers they should read.

Acknowledgment

This publication was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ababneh M, Al-Shalabi R, Kanaan G, and Al-Nobani A (2012). Building an effective rule-based light stemmer for Arabic language to improve search effectiveness. The International Arab Journal of Information Technology, 9(4): 368-372.

AbuSafiya M (2017). Arabic root extraction through generation and filtering. In the International Conference on Mathematics

and Information Technology, IEEE, Adrar, Algeria: 191-195. https://doi.org/10.1109/MATHIT.2017.8259715

Albogamy F and Ramsay A (2016). Unsupervised stemmer for Arabic tweets. In the 2nd Workshop on Noisy User-generated Text, Osaka, Japan: 78-84.

Alhanini Y and Aziz AMJ (2011). The enhancement of Arabic stemming by using light stemming and dictionary-based stemming. Journal of Software Engineering and Applications, 4(9): 522-526. https://doi.org/10.4236/jsea.2011.49060

Al-Kabi MN (2013). Towards improving Khoja rule-based Arabic stemmer. In the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, IEEE, Amman, Jordan: 1-6. https://doi.org/10.1109/AEECT.2013.6716437

Al-Kabi MN, Al-Radaideh QA, and Akkawi KW (2011). Benchmarking and assessing the performance of Arabic stemmers. Journal of Information Science, 37(2): 111-119. https://doi.org/10.1177/0165551510392305

Al-Kabi MN, Kazakzeh SA, Ata BMA, Al-Rababah SA, and Alsmadi IM (2015). A novel root based Arabic stemmer. Journal of King Saud University-Computer and Information Sciences, 27(2): 94-103. https://doi.org/10.1016/j.jksuci.2014.04.001

Alkhatib M, Monem AA, and Shaalan K (2017). A rich Arabic word net resource for Al-Hadith Al-Shareef. Procedia Computer Science, 117: 101-110. https://doi.org/10.1016/j.procs.2017.10.098

Al-Lahham YA, Matarneh K, and Hasan M (2018). Conditional Arabic light stemmer: Condlight. The International Arab Journal of Information Technology, 15(3A): 559-564.

Almusaddar MY (2014). Improving Arabic light stemming in information retrieval systems. M.Sc. Thesis, Islamic University of Gaza, Gaza, Palestine.

Al-Omari A, Abuata B, and Al-Kabi M (2013). Building and benchmarking new heavy/light Arabic stemmer. In the 4th International Conference on Information and Communication Systems (ICICS 2013).

Alsaad A and Abbod M (2014). Arabic text root extraction via morphological analysis and linguistic constraints. In the UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, IEEE, Cambridge, UK: 125-130. https://doi.org/10.1109/UKSim.2014.43

Al-Sarhan H, Al-Shalabi R, and Kanaan G (2003). New approach for extracting Arabic roots. In the 2003 Arab Conference on Information Technology (ACIT 2003), Egypt: 42–59.

Alshalabi R (2005). Pattern-based stemmer for finding Arabic roots. Information Technology Journal, 4(1): 38-43. https://doi.org/10.3923/itj.2005.38.43

Bharati M and Ramageri M (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering, 1: 301-305.

Boudad N, Faizi R, Thami ROH, and Chiheb R (2018). Sentiment analysis in Arabic: A review of the literature. Ain Shams Engineering Journal, 9(4): 2479-2490. https://doi.org/10.1016/j.asej.2017.04.007

Boudchiche M and Mazroui A (2018). Improving the Arabic root extraction by using the quadratic splines. In the International Conference on Intelligent Systems and Computer Vision, IEEE, Fez, Morocco: 1-5. https://doi.org/10.1109/ISACV.2018.8354062

Boudlal A, Bebah MOAO, Lakhouaja A, Mazroui A, and Meziane A (2011). A Markovian approach for Arabic root extraction. The International Arab Journal of Information Technology, 8(1): 91-98.

Elazhary HH and Khodeir N (2017). A cognitive tutor of Arabic word root extraction using artificial word generation, scaffolding and self-explanation. International Journal of Emerging Technologies in Learning (iJET), 12(05): 36-49. https://doi.org/10.3991/ijet.v12i05.6651

El-Defrawy M, El-Sonbaty Y, and Belal N (2015). Enhancing root extractors using light stemmers. In the 29th Pacific Asia Conference on Language, Information and Computation: Posters, Shanghai, China: 157-166.

Ghawanmeh S, Al-Shalabi R, Kanaan G, Khanfar K, and Rabab'ah S (2009). Enhanced algorithm for extracting the root of Arabic words. In the 2009 Sixth International Conference on Computer Graphics, Imaging and Visualization, China: 388–391. https://doi.org/10.1109/CGIV.2009.10

Gridach M and Chenfour N (2011). Developing a new approach for Arabic morphological analysis and generation. arXiv:1101.5494. https://doi.org/10.1155/2011/629305

Hajjar M and Zreik K (2010). A system for evaluation of Arabic root extraction methods. In the 5th International Conference on Internet and Web Applications and Services, IEEE, Barcelona, Spain: 506-512. https://doi.org/10.1109/ICIW.2010.98

Hawas FA (2013). 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, 14(2): 327-341. https://doi.org/10.7494/csci.2012.14.2.327

Jaafar Y, Namly D, Bouzoubaa K, and Yousfi A (2017). Enhancing Arabic stemming process using resources and benchmarking tools. Journal of King Saud University-Computer and Information Sciences, 29(2): 164-170. https://doi.org/10.1016/j.jksuci.2016.11.010

Khafajeh H, Yousef N, and Abdeldeen M (2018). Arabic root extraction using a hybrid technique. International Journal of Advanced Computer Research, 8(35): 90-96. https://doi.org/10.19101/IJACR.2017.733023

Khoja S and Garside R (1999). Stemming Arabic text. Lancaster University, Lancaster, UK.

Kreaa AH, Ahmad AS, and Kabalan K (2014). Arabic words stemming approach using Arabic WordNet. International Journal of Data Mining and Knowledge Management Process, 4(6): 1. https://doi.org/10.5121/ijdkp.2014.4601

Larkey LS, Ballesteros L, and Connell ME (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, Tampere, Finland: 275-282. https://doi.org/10.1145/564376.564425

Momani M and Faraj J (2007). A novel algorithm to extract tri-literal Arabic roots. In the 2007 IEEE/ACS International Conference on Computer Systems and Applications, IEEE, Amman, Jordan: 309-315. https://doi.org/10.1109/AICCSA.2007.370899

Nehar A, Ziadi D, and Cherroun H (2016). Rational kernels for Arabic root extraction and text classification. Journal of King Saud University-Computer and Information Sciences, 28(2): 157-169. https://doi.org/10.1016/j.jksuci.2015.11.004

Otair MA (2013). Comparative analysis of Arabic stemming algorithms. International Journal of Managing Information Technology, 5(2): 1-13. https://doi.org/10.5121/ijmit.2013.5201

Saad MK and Ashour WM (2010). Arabic morphological tools for text mining. In the 6th International Conference on Electrical and Computer Systems, Lefke, North Cyprus: 1-6.

Salloum SA, AlHamad AQ, Al-Emran M, and Shaalan K (2018). A survey of Arabic text mining. In: Shaalan K, Hassanien A, and Tolba F (Eds.), Intelligent natural language processing: Trends and applications: 417-431. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-67056-0_20

Sameer RA (2016). Modified light stemming algorithm for Arabic language. Iraqi Journal of Science, 57(1B): 507-513.

Taghva K, Elkhoury R, and Coombs J (2005). Arabic stemming without a root dictionary. In the International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II, IEEE, Las Vegas, USA, 1: 152-157. https://doi.org/10.1109/ITCC.2005.90

Yaseen Q and Hmeidi I (2014). Extracting the roots of Arabic words without removing affixes. Journal of Information Science, 40(3): 376-385. https://doi.org/10.1177/0165551514526348