International Journal of Information Technology and Language Studies
(IJITLS)

Arabic Stemmer System based on Rules of Roots

Waed Al-Abweeny [1] and **Nahed Abu Zaid** [2]
walaboweeny@iau.edu.sa; Nahedabuzaid28@gmail.com

[1] University of Imam Abdurrahman Bin Faisal, KSA
[2] University of Sattam Bin Abed Al_Azeez, KSA

Abstract. Stemmer is an automated process, which produces a base string in an attempt to represent related words, which is the main step that is used to process data in many types of applications such as text mining, information retrieval, and natural language processing. The stemmer task is to reduce words to their base. The more systems are used to analyse and understand the syntax and semantic of the documents the more accurate is the result. Arabic stemmer is not an easy task due to the morphological variants of certain words which are not always semantically related. This paper introduces an Arabic stemmer system based on Arabic rules to extract trilateral (three radicals), quadrilateral (four radicals), sometimes quintuple (five radicals) and hexagonal (six radicals) if available. In addition, it compares the Arabic stemmer with other stemmer systems, and evaluates it by four Arabic native speakers specialists where it has achieved 96.8% ratio of accuracy.

Keywords: Arabic stemmer, trilateral, quadrilateral, quintuple, hexagonal, Arabic Morphology.

1. Introduction

Arabic language is one of the most widely spoken languages in the world. Arabic is written from right to left and has 28 characters where each character has many forms and shapes. For instance, "ب" “BAA” has different forms which belong to the same letter "ـبـ" , "ـب" , "بـ" . In addition to that diacritic used in Arabic language to recognize the pronunciation of the letter like the constant "ت " / “TAA”, diacritic fatha “FATHA” is pronounced “TA” / "تَ", diacritic damma "تُ "is pronounced “ TO”, diacritic “KASRA” "تِ"is pronounced “TE”. Also, there is what is called “TANWEEN” which means two of identical diacritic like "تً " is pronounced “TAN”, "تٌ"is pronounced “TON”, "تٍ "is pronounced “ TEN “. So, building an effective stemming algorithm for Arabic language has been always an important research topic in many fields of natural language processing like information retrievals systems, web search engines, question answering systems, textual classifiers, etc (Nwesri, 2008).

1.1 Arabic Morphology

Arabic language, unlike English, has a more complicated structure and morphology. The form of nouns is determined by several criteria like gender, numbers, and grammatical cases. Moreover, Arabic nouns have a large number of variants, and some of the variants can be complex because of the prefixes, suffixes, and infixes (Chen & Gey, 2002).

To find the root of a word in Arabic language we must remove some prefixes, infixes, and affixes from it. Stem is a base form of the word. For example, the word " يكتبون" (they are writing) is pronounced “YAKTOBOON”. Its stem is "كتب" (write) “KATABA”. We remove one prefix “YA”/"ي", from the beginning of the word, and two suffixes “WAW,NOON”/"و",”ن”, from the end of the word. The word معلومة (information) is pronounced “MAALOMAH” and has one prefix “MEEM/م”, one infix “WAO/ و” , and one suffix “TAA/ـة” at the end of the word. After removing them from the word, the stem “علم” (inform) is pronounced “alema”.

1.2 Stemmer

Stemmer is an automated process that produces a base string in an attempt to represent related words (software used to produce the stem from the inflected form of words. Stemmer used as preprocessing techniques to reduce the number of tokens in textual document (AL-OMARI & AbuAta, 2014). It is used to find semantically identical terms, which are derived from the same stem/root.

Light stemming is restricted to the removal of limited number of Arabic prefixes and suffixes to output a stem. Heavy stemming is also known as root-based stemming which includes implicitly the removal of Arabic prefixes and suffixes besides extracting the appropriate root.

All nouns and verbs are generated from a set of roots, which is about 11,347, root distributed as follows (Mohammed, 2016):

115: Two character roots (and these roots have no derivations from them).

7198: Three character roots.

3739: Four character roots.

295: Five character roots.

The importance of stemming is to increase the effectiveness of information retrieval, but as we said before, Arabic Language needs more efforts because of the complexity of its morphological structure.

In this paper, we have developed a new technique of stemmer system to improve Arabic text retrieval. First, the current technique will perform preprocessing operations, and then matches the word with the suitable rule based on the number of characters per word to find and return all the roots of the word. If the word contains many roots, like those that have more than three letters, may return trilateral, quadrilateral compared with many researchers stemmer systems that just focus on returning the trilateral of the word, regardless of the fact that the word has three, four, five or six radicals. This addition improves the level of efficiency and accuracy of our system. In our system, we return the words to trilateral, quadrilateral, quintuple, hexagonal, as will be described and shown in the methodology and results.

2. Related Work

AL-OMARI and AbuAta (2014), designed and implemented a new Arabic light stemmer (ARS) which is not based on Arabic root patterns. Instead, it depends on well-defined mathematical rules and several relations between letters. ARS had shown few wrong stems when applied on a set of 6,225 so classified to 5733 as correct roots, 347 as wrong roots, and 145 as no roots. These wrong stems are grouped into three types over-stemming (stemming two words with different stem to the same root), mis-stemming (taking off what looks like an ending, but is really part of the stem) and under-stemming (the opposite). ARS also tested and compared with two similar stemmers Al-Kabi and Ghwanmeh's.

Khalid et al. (2016) presented Arabic stemming techniques depending on the root of the Arabic word. The paper presents the usage of Arabic stemming increases and accelerates the search engines output, where Google Chrome outperforms Internet Explorer and Mozilla Fire fox in terms of total number of searched pages.

Kannan et al. (2008) introduced a new method for stemming to solve many of the ambiguity problems related to light stemming. The method depends on a set of possible affixes in which they only have a prefix and suffix. In their prefixes, they combined all possible antefixes and prefixes to generate one complete list and in the suffixes, they combined all possible suffixes and postfixes.

In their rule-based light stemmer, they used a set of rules to determine if a certain sequence of characters is part of the original word or not and this helped them solve some ambiguity problems. In addition, they introduced a way for handling the majority of broken plural forms and reducing them to their singular forms. This helped grouping words of the same meaning in a common form.

Mohammed (2016) proposed an Arabic stemmer that combined the rules of root-based stemmer and light-based stemmer to overcome the mistakes. The stemmer deals with situations that have not been previously tackled by other researchers. This has solved the situation of three letters words and matching a word against Tafealat before removing any affixes to avoid deleting a genuine letter of a word.

Sembok and AbuAta (2013) developed an algorithm for Arabic stemming to increase the correctness of the stemming system compared with other systems that they mentioned in their paper. Their stemming algorithm follows several steps starting with checking the entered word with dictionary, if the word is not found then the system removes the suffixes and prefixes, and the candidates roots were generated by their system. After that, their system checked the correctness by matching the roots with Arabic template sets that sorted in a dictionary. Then they reconstructed the word from the candidate root by adding some letters to increase the correctness results and to check if the candidate root is the right output or not. Finally, their system returns the root for the word after modifies it by adding or replacing some letters to reach the correct root. If their stemming system did not find the correct stem, then the word returns without any modifications. Their system reached better results than Al-Omari system, which they mentioned in their paper from different disciplines: average recall and precisions, numbers and type of errors.

Larkey et al. (2007) developed a light stemmer for Arabic Language to try to avoid computational morphology problems because morphological analyzers make many mistakes when retrieving the information due to several factors names, many errors Tokenization, information which is got from infrared documents and inquiries may not be able to be used. For monochrome retrieval, they have shown improvements of about 100% in the average accuracy caused by the stem and related processes, and the greater impact on the dictionary-based retrieval across languages but all of the stemmer developer consistent that the output should be particularly effective for languages with more complex morphology. They found that the light stemmer is powerful in removing the specific articles, stop words, prefixes and few suffixes.

They compared the effect of their stemmer with other stemming studies and assessed their effectiveness for information retrieval. They find that stemmer for monolingual retrieval increases in average precision from the light10 stemmer and they find that the light stemmer is powerful due to many reasons such as being good and sufficient for information retrieval. It did not need complete sentences and they do not try to deal with each individual case. It is enough to retrieve information where many of the most common forms of a word are mixed. At the end of their paper, they suggested that using morphological analysis to mix words works better than light stemmer because they do not seek to represent the Arabic words with their roots, and equalize all words derived from the same root. Therefore, they recommended future work to further utilize intelligent use of morphological analysis of information retrieval.

3. Methodology

In this paper, we built an Arabic stemmer system based on rules of root, which can handle most of Arabic morphological issues. The system process is shown in Figure1.

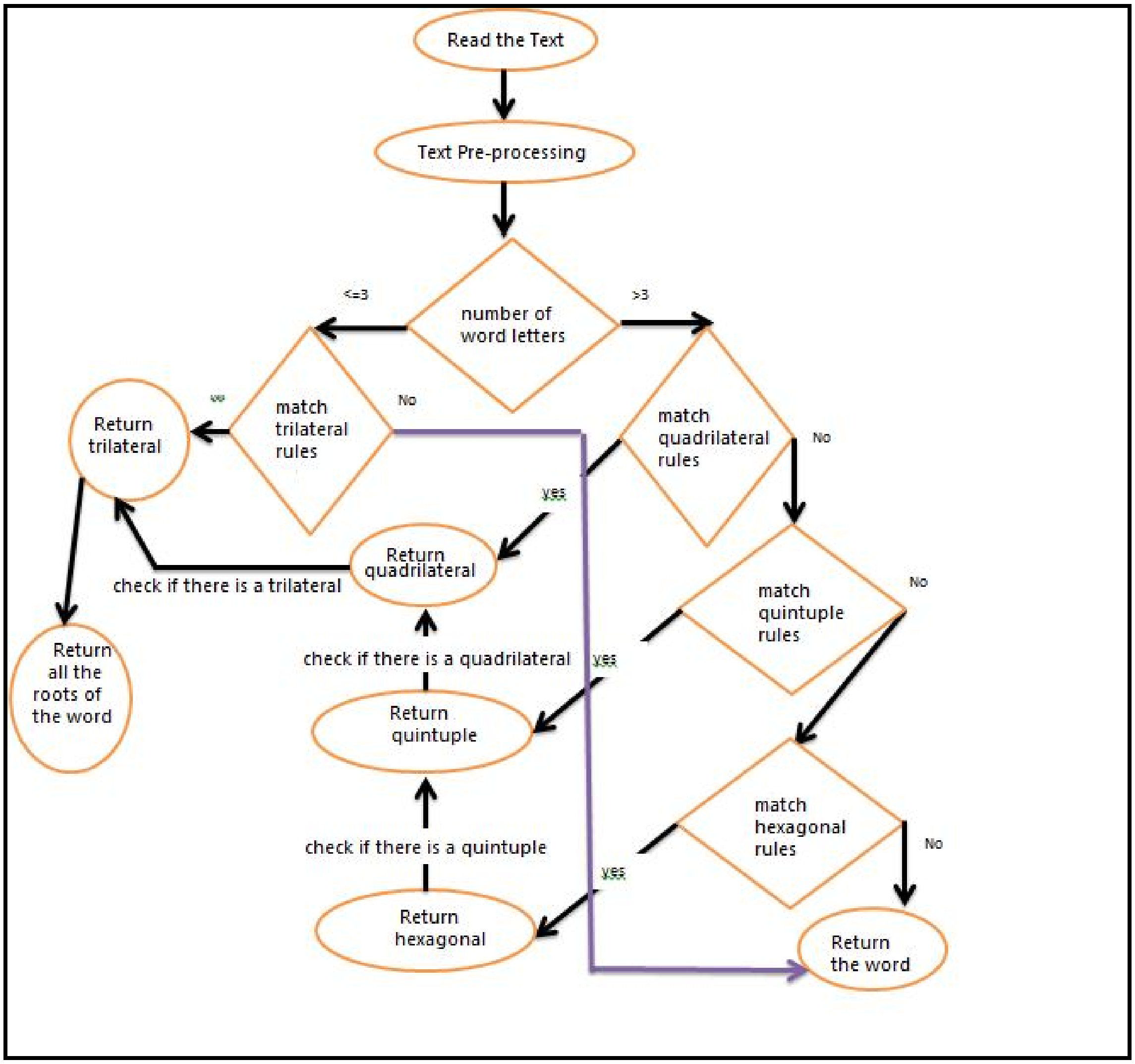

Figure 1. Stages of the Proposed Arabic Stemmer System

3.1 The process of the system: conducting Pre-processing steps before stemming:

1. Remove the diacritics as “AL-TANWEEN” / “ ٍ , ٌ ,◌ً , “AL-HRAKAT”/ “ ْ, ِ, ُ, ◌َ, but the “AL-SHADDA” / “ ◌ّ we use it in the rules because it expresses duplicating the letter.
2. Remove punctuation as “ : , . , ، ,؟ ” and others.
3. Remove the stopping word like separated pronouns, prepositions and the letters of the monument and assertion. Removing the stop words is very useful because these words do not have roots, separated pronouns such as “we/نحن”, “you/أنت”, the prepositions such as “from/من”, “on/على”, letters of the monument and assertion such as “will not/لن”, “even/حتى”.
4. Removing the” ال” (All it means the), “بال“(Bel it means with),” لل “(lel it means for). For example, take the word” للمدرسين “(for teachers) it could be “ مدرسين“after removing the prefixes from the word “لل”.
5. Check the words (الله) “لفظ الجلالة/ ALLAH”, return it as it is without any process.

3.2 Removing Suffixes, Prefixes, and Infixes from the word:

We built a system containing a huge amount of Arabic rules that are collected and extracted from (Almojaz in Arabic language grammar) " الموجز في قواعد اللغة العربية", and "الشامل في اللغة العربية " (Alshamel in Arabic language), which simulate and translate the human mind.

Checking words length due to number of their letters and then go to the suitable cases:

Case 1: If the word contains three letters, then the system returns it. , but if the word contains a letter that has" شدة "(Shada ّ) then the stemmer returns the root of the word by duplicating that letter. For example, take the word "شدّ", it will return as " شدد".

Case 2: If the word contains four letters, then the word goes through the loop to find the root of the word depending on the trilateral rules, which was extracted before. For example, take the word" يشرب "(drink), our system matches it with the suitable rule, then finds the root for the word by returning the base form "فعل", that's mean there is one prefix "ي" must be removed from the word to return the trilateral of the word and it cloud be "شرب". In our system, we almost covered the words that have trilateral as is shown in Table 1.

Entering Word	Number of Letters	Rule match	Base Rule form	Stem of the word
سهيل	4	فعيل	فعل	سهل
جلوس	4	فعول	فعل	جلس
خضرة	4	فعله	فعل	خضر
حطّب	4	فعّل	فعل	حطب
سهيل	4	فعيل	فعل	سهل

Table 1. Words Containing four Letters Having Trilateral.

Case 3: If the word contains five letters, then our system matches the word with the suitable trilateral and quadrilateral rules, and it goes into many loops until it reaches the base root of the word. For example; take the word" زراعة "it means "planting". After matching with the suitable rules, the base root is "زرع" by removing the"ا " as infix letter and "ـة" as suffix letter. However, if the word is "تجعّد" so the root of this word is "جعد" by removing "تـ" and" ّ as prefix and suffix letters. Table 2 shows some examples about five letters that have trilateral or quadrilateral.

Entering Word	Number of Letters	Rule match	Base Rule form	Stem of the word
مضروب	5	مفعول	فعل	ضرب
عطشان	5	فعلان	فعل	عطش
تقدّم	5	تفعّل	فعل	قدم
تدحرج	5	تفعلل	فعلل	دحرج

Table 2. Words Containing Five Letters Having Trilateral or Quadrilateral

Case 4: If the word contains six letters like "يتصالح", it matches the suitable rule then removes the prefixes letters as "يتـ", the root of the word is "صالح", but in this case our system continues to find other trilateral or quadrilateral of the word if that is applicable. If the root matches another rules, which means our system returns another root for the same word as "صلح". Table 3 shows some examples about six letters that have trilateral and / or quadrilateral.

Entering Word	Number of Letters	Rule match	Base Rule form	Stem1 of the word	Rule match	Base Rule form	Stem2 of the word
يتلاعب	6	يتفاعل	فاعل	لاعب	فاعل	فعل	لعب
مقاتلة	6	مفاعلة	فاعل	قاتل	فاعل	فعل	قتل
اقشعر ّ	6	افعلل ّ	فعلل	قشعر	-	-	-
زلزلته	6	فعللته	فعلل	زلزل	-	-	-

Table 1. Words Containing Six Letters Having Trilateral and /or Quadrilateral.

Case 5: If the word contains more than six letters, our system matches all the rules to extract the trilateral, quadrilateral, quintuple and hexagonal.

Case 6: If the entered word does not match with any rule then the system returns the exact word, so it may be an anomalous word.

In our system, we try to include most of the anomalous words as rules, such as the word “اعطاء” the system returns the roots “اعطى, عطى”. Table 4 shows examples about anomalous words that are handled by our system.

Entering Word	Number of Letters	Rule match	Base Rule form	Stem of the word
جرحى	4	فعلى	فعل	جرح
صيام	4	فعال	فعل	صوم
احتواء	6	انفعال	انفعل	احتوى
طالبات	6	فاعلات	فاعلة، فعل	طالبة، طلب

Table 4. Anomalous Words

4. Experimental Results

We test the efficiency of the stemming system by comparing the results of our system with other stemmer systems, and evaluating it by four Arabic native speakers’ specialists.

4.1 Comparison with other Stemmer Systems

We depend on the following equation to find the Percentage of Evolution for our system testing results:

The percentage of evolution = (no. true tested word / no. words document) * 100 (1)

From the comparison results, it is clear that each word returns the root/roots, as shown above. The system returns all the roots of the entered word not just the trilateral, like"الابداع", the roots returned (أبدع، بدع) are compared with other systems (Khalid et al., 2016). The accuracy percentage of our system is = 95.3%, and in the other systems = 92.6%. Table 5 shows the comparison.

No.	Entered word/words	Our Stemmer Results	Arabic Stemmer Results	Percentage of Evolution for their System	Percentage of Evolution for our system
1	يختلف اسنان عصر العلم و التقنية الثورة العلمية الخلاقة عن انسان العصور السابقة في عملية الإبداع و التمكن	خلف أسنن سنن عصر علم قنة نسن عصر ثور علم خلق أنس سبق عمل أبدع بدع مكن	خلف سنن عصر علم قنأ ثور علم خلق أنس عصر سبق عمل بدع مكن	0.928571	0.928571429
2	طور التعليم في المملكة العربية السعودية مقارنة بدول العالم الاسلامي	طور علم مملك ملك عرب سعود سعد قارن قرن بدل علم اسلام أسلم سلم	طور علم ملك عرب سعد قرن بدل علم سلم	1	1
3	مشاكل العلاقات الدولية و أثرها على مداخيل المواطن العربي	شاكل شكل علاقة علق دول أثر مدخل دخل واطن وطن عرب	شكل علق دول أثر دخل وطن عرب	1	1
4	طرق التحاق الطالب بالجامعات في أوروبا و أمريكا	طرق حاق طلب جامعة جمع أوروبا أمريكا	طرق لحق طلب جمع ورب وامريكا	0.833333	0.833333333
5	تحليل وتصميم أنظمة الحاسب الالي	حلل صمم أنظم نظم حسب آلي	حلل صمم نظم حسب ليي	0.8	1
6	جامعة نجران	جمع نجر	جمع نجر	1	1
Average Results				0.926984	0.953488

Table 5. Comparing Our System Results with Other Stemmer Systems Results

4.2 Comparison from Arabic Specialists

At this stage, our system is evaluated by four Arabic academic specialists, where they entered many different texts from different resources (like articles, Hadiths, poetry, etc. …) on the system to extract and return the roots of the words. After that, they evaluated the accuracy of our system and its ability to extract the correct roots. The number of words used in the evaluation equals 921, the number of corrected roots equals 892, and the number of error words that our system cannot find their roots or return the correct root equals 29. Therefore, the accuracy ratio of our system is 96.8%, which is an excellent ratio compared with other systems.

Text	Our system Result	Arabic academic specialists Result
ما أجمل الرضى.... إنه مصدر السعادة وهدوء البال يقول ابن القيم عن الرضى: هو باب الله الأعظم ومستراح العابدين وجنة الدنيا. الحمد لله الذي عافانا وأهلينا مما ابتلى به غيرنا وفضلنا على كثير من خلقه،	جمل رضى إنه صدر سعد هدء بال قول ابن قيم رضى باب الله عظم استراح راح راح عابد عبد جنة دنا حمد لله عافا عفا أهلي أهل مما بلى به غير فضل كثر خلق	جمل رضى إنه صدر سعد هدأ بال قول ابن قيم رضى باب الله عظم استراح راح عابد عبد جنة دنو حمد لله عافا عفا أهل مما بلى به غير فضل كثر خلق

Table 6. Example about Four Arabic Academic Specialists Evaluation

5. Conclusion

Arabic language has a rich and complex morphology. This paper developed Arabic stemmer system that can handle the words containing trilateral, quadrilateral, quintuple and hexagonal. The improvement of our system returns all the roots of the word not just the trilateral, and try to include most of the anomalous words as suitable rules. Arabic stemmer is also compared with and tested against other stemmer and evaluated by four Arabic specialists. The results were encouraging, showing the effectiveness of our stemmer system.

Our stemmer system needs to be enhanced and improve its functionality to increase the percentage of the correctly-extracted Arabic roots from Arabic words by handling the bilateral, some anomalous words, the connected prepositions with words like “ببابه", related pronouns “هما،نا،.....", and the related conjunctions like “فجلس”.

References

AL-OMARI, A., & AbuAta, B. (2014). ARABIC LIGHT STEMMER (ARS)”, *Journal of Engineering Science and Technology*, Vol. 9, No. 6

Chen, A., & Gey, F. (2002). Building an Arabic Stemmer for Information Retrieval”, *School of Information Management and Systems*, University of California at Berkeley,USA

Kannan, G., Shalabi, R., Ababneh, M., & Al-Nobani, A. (2008). Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness”, *International Conference on Innovations in Information Technology*, 2008

Khalid, A., Hussain, Z., & Baig, M. (2016). "Arabic Stemmer for Search Engines Information Retrieval", *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1

Larkey, L., Ballesteros, L., & Connell, M. (2007). Light Stemming for Arabic Information Retrieval, *Arabic Computational Morphology*, vol. 38, pp. 221–243.

Mohammed, R. (2016). New Arabic Stemming based on Arabic Patterns, College of Islamic Science, *The Iraqi University, Baghdad*, Iraq, Vol. 57, No.3C, pp:2324-2330

Nwesri, A. (2008). Effective Retrieval Techniques for Arabic Text, *A thesis submitted for the degree of Doctor of Philosophy*, B.Sci., M.Soft.Eng.

Sembok, T., & AbuAta, B. (2013). Arabic Word Stemming Algorithms and Retrieval Effectiveness, *Proceedings of the World Congress on Engineering*, Vol III, London, U.K.

الافغاني، سعيد (2003). "الموجز في قواعد اللغة العربية"، ص432 ، دار الفكر، بيروت، لبنان

الأزهري، مصطفى (2011). " تيسير قواعد النحو للمبتدئين “، ص 354، دار العلوم والحكم، الجيزة

النقراط، عبدالله (2002). " الشامل في اللغة العربية"، ص202، دار قتيبة

الخوص، أحمد (1999)." قصة الاعراب "، 1999، ص280، المطبعة العلمية، دمشق