AN IMPROVED ALGORITHM FOR THE EXTRACTION OF TRILITERAL ARABIC ROOTS

Dr. Raed Kanaan, Assistant Prof.
Management Information Systems Amman Arab University,
Faculty of Computer Sciences and Informatics, Amman, Jordan
Dr. Ghassan Kanaan, Full Prof.
Computer Science Amman Arab University,
Faculty of Computer Sciences and Informatics Amman, Jordan

Abstract

Stemming in the Arabic language is extracting the root form of the verb, removing inflectional affixes and derivational morphemes. Stemming is a share form of language processing in the systems of information retrieval. It is similar to the morphological processing used in natural language processing, but to some extent has different aims. Stemming is used to reduce word forms to common words. Stemming is the process of removing all affixes from a word to extract its root. This paper describes a stemming algorithm that has been developed for the Arabic language. The algorithm utilizes an important morphological aspect of the Arabic language. The algorithm examines the word and extracts its root. It examines the word letter by letter starting from the end of the word, i.e., from the last letter of the word to the first. The algorithm correctly stems most Arabic words that are derived from roots, and achieves high rate of accuracy. The algorithm has been tested on a corpus of 242 abstracts of Arabic documents from the Proceedings of the Saudi Arabian National Conference.

Keywords: Arabic Language; Extraction; Roots; Stemming; Light stemmer

1. Introduction:

1.1 Linguistic Affiliation:

Arabic is a Semitic language of the Arab-Canaanite subgroup (Ruhlen, 1987). It belongs to the Afro-Asiatic family of languages--the bulk of which are spoken in Africa which has several main sections: Semitic (such as Arabic); Berber; Chadic (such as Hausa); and Ancient Egyptian descendent of modern, Coptic, is to maintain the liturgical language. Arabic and Canaanite are distantly related to Aramaic. Other relatives are even more

distant Semitic languages of Ethiopia and Akkadian, an extinct language once spoken in Mesopotamia.

The major dialects of Arabic are Classical Arabic, Eastern Arabic, Western Arabic, and Maltese. A modern form of Classical Arabic referred to as Modern Standard Arabic is used in writing Arabic today. Eastern Arabic, sometimes called Mesopotamian Arabic, includes the Arabic dialects spoken in several countries such as Egypt, Sudan, Syria, Iraq, Arabian Peninsula, and the Arabic speaking communities in Asia (Bateson, 1967).

1.2 Orthography

Arabic language uses an alphabetic system that normally represented by symbols; consonants and long vowels. In addition, there is a close match between the written symbols and their linguistic function. Short vowels, however, are not written despite the fact that much morphological and grammatical meaning is signaled by vowels. Due to the fact that only roots and stems of an inflected word are written, therefore the reader has to infer its particular meaning from context. When vowels are represented, as in children's books or learners' manuals, super- and subscript diacritics are used. Arabic language is written from right to left (UCLA, 2013).

1.3 Linguistic Sketch

Modern Standard Arabic (MSA) has a grammatical system known as a "root and pattern system." Words are composed of roots and patterns. Roots consist of three consonants, however a few have four or five; the roots, unpronounceable of itself, are allied with a general meaning, therefore the sequence ktb has an association with the meaning "writing." Patterns are vowel sequences, which can be mediated on as templates, (sometimes as prefixes and suffixes, and sometimes with additional consonants). Patterns are then "added" to, or even within roots of the word following well-defined models. For instant, consider the root d-r-s. Despite the letters (i.e., consonants) drs will always remain the same, the following are examples to confirm that the scheme and vocalization will change depending upon usage (Semitic Languages, 2013):

darasa, "to study, learn" “he studied” (the third personal singular perfect form is the reference form for the verb).
darrasa, "to teach" “he taught”
dars, "lesson, class"
durus, "lessons"
mudaaris, "teacher (male)"/mudarrisa, "teacher (female)"
madrasa, "school"

These patterns then produce various nominal and verbal stems, which have a variety of functions; in nouns for example, they imply habitual occupations, diminutives, or colors, and in verbs, they form participles, causatives, and passives.

Nouns are inflected and morphologically marked for case (nominative, genitive, and accusative), gender (masculine and feminine), number (singular, plural, dual, and collective) and determination (definite and indefinite). Plural forms of many nouns are significant by ablauts, that is, "the vowel pattern within a root varies between singular and plural forms, similar to alternations in English like those in the verb sing, sang, and sung, or the noun mouse and mice" (Touregypt, 2013).

In verbs, which occur in two basic stems, the perfect and imperfective, person, number, mood, and aspect are marked by prefixes and suffixes. Templates for verbs consists of ten commonly, (also of four rarely) used shapes and meanings. Their meanings reveal verbs that relate intensity, repetition, causation, intention, and belief. In addition to the nominal and verbal systems there is another system of particles. Particles include such things as function words, which express syntactic relationships, for example, interrogatives, prepositions, conjunctions, and pronouns (Touregypt, 2013). Compared to the root-pattern system of other word categories these are quite simple in their formation (Al-Fedaghi, & Al-Anzi, 1989; Al-Fedaghi, and Yaseen, 1990; Ali, et.al, 1984; Beesley, 1989)

2. Related Work:

According to Al-Nashashibi, et al. (2010) he suggested a new technique for root extraction as a pre-processing step in the Arabic text mining. They claimed that the available approaches in the literature does not tackling the elimination of long vowels, geminated, and hamzated. Therefore, they proposed an algorithm to handle such issues. This algorithm improved the accuracy by 14%. However, Beesley (2001) and EI-Sadany and Hashish (1989) handled the elimination of long vowel words.

Khoja (1999) proposed a light stemmer to handle weak and geminated words; however, Al-Nashashibi, et al. (2010) further asserted that none of the available approaches handling weak, geminated, hamzated and eliminating long vowels. As a result, Al-Nashashibi, et al. (2010) benefited from AI-Ameed (2006) linguistic approach and improved the accuracy by 14%.

To support the reason behind conducting this research, Aljlayl, and Frieder (2002) light stemmer for instant remove the most frequent prefixes and suffixes from the words instead of the prefix and/or suffix list must be removed from the selected words, hence our proposed algorithm.

The work of AI-Ameed et.al.(2005) or what so called TREC-2001, enhanced the performance of Larkey's (2002) light8 stemmer in two ways. In one hand, by changing the sequence of the components of algorithm execution. On the other hand, by additionally adding new affixes to the already existing ones. He claims that the prefixes list contains 17 two characters, however the only found is 15.

Apparently, the above work can be considered as a development of Darwish and Oard (2002) stemmer. The below prefixes have been removed by his stemmer:

فال, تال, تد, يد, لد, مد, سد, وخ, لم, تم, ود, وم, مم, فم, ال, للـ, وي, لي, في, فا, وا, لا, تا).

In addition to the following suffixes: اخ, وا, ون, وه, ان, ذي, ذه, ذم, مم, هم, هه, ي,ه, ج, يه ’يه, وا, ذل, يح, ها

Below section describes the proposed algorithm.

3. Algorithm Description:

We start by examining the length of the word. If it is a one or two letter word, it is probably a particle; it is treated as a stop word and the algorithm returns it as is. If the word consists of three letters we accept it as a root and do not process it further. Every letter is checked separately to determine whether it is considered additional (احرف الزيادة) or not. These letters are:

{"أ"،"ت"،"م"،"و"،"ن"،"ي"،"ا"،"ء"،"ئ}

The rest of the letters are considered original.

The first step is to find the stem by deleting the definite article "ال" and what precedes it in addition to removing the letters "ة" "هـ" from the end. Then we examine the length of the word, i.e., if it is three letters long, then it is a root, otherwise we continue.

{ملعب = للملعب} or {أكل = ليأكل}

If the first letter is "ل" we delete it and if the first two letters are "لل" we delete them too. If the second letter is "س" preceded by one of the following letters:

{"أ"،"ت"،"ي"،"م"،"ن"} and followed by "ي" or "ت"، we delete the three letters, if the length of the word is more than five letters.

عمل =استعمل، يستعمل، نستعمل، تستعمل، أستعمل

If "س" is the first letter followed by "ي" or "ت" and the word length is more than four, these letters are deleted. If the last letter is "ي", it is changed to "ا".

{سيساهم= ساهم= سهم} {أخر =أخرا }

And we start implementation to apply the main part of the algorithm:

We examine the last letter, if we have two similar letters, with a vowel between them; we delete the vowel, and add the two letters to the root.

{جنن = مجنن = مجنن}

As for the preposition ,"ب" we always delete it except in two cases: if it is followed by the letter "ي" or "ا". We also remove the letter that precedes it, then continue till the first letter, i.e., until the length of the word becomes equal to zero. The processes that are applied on letters are as follows:

Let us take the word {المدرسة} as an example. When we delete its additional letters it becomes {مدرس} so we take the last letter {س} and find it an original letter, we examine the next letter "ر" and find it also an original letter, then we examine "د" to find it also original, but when we examine "م" we find it to be an additional or extra letter so the root is {درس}.

Note: if "ا"و is found in the text, it is changed to "و", whereas if "يا" or "أ ي" are found, they are changed into "ا"

Letter "أ"

If preceded by "س" and followed by "ل", the root is {سأل}. {سأل=سألني}

If preceded by "س"+أ they are deleted {قام = قوم=سأقوم}.

If it was the initial letter, it is added to the list of root characters; otherwise, it is not original.

Letter "ا"

If preceded by "س" and followed by "ل", the root is {سأال} as in: {سال=سالني}

If the length of the root is more than five letters, and the number of original letters is more than two and preceded by "س" they are deleted

{سير = ساسايره=سار }

If it is a word initial character or in second position, it is deleted, but if it is in the third position, it is listed.

If it is a postposition and proceeded by "هم" or "كم" we delete all three letters.

{كتبهما= كتبكما = كتب}

If the word length equals the stem word length and "ا" is preceded by one of the following letters "ت" "و" "ه" "ن", {ملك = ملكها و درس = درسوا درستا و درسنا و}

If it does not match one of these cases, it is listed.

Letter "ت"

If "س" is the second letter, preceded by one of these letters "ا"،"ت"،"ي"،"م"،"ن" and followed by "ت" we deleted all three of them since the word length exceeds five letters. {مستعمل، أستعمل، نستعمل، تستعمل، يستعمل}

If "س" is in initial position, followed by "ت", and the word is more than four letters, we delete both characters. {ستساهم= سهم}

If we have the two letters "تت"or "يت", we delete the last one and keep the first if the word length is four, but if it is more, we delete the two.

{ وفق= يفق=يتفق}

If it is the last letter, and it is preceded by ""ا, we delete both letters.

If "ت" is in initial position, and the number of original letters is greater than zero, we delete ""ت and retrieve the last letter that was listed except ""ي or "ا".

{تنمية = نمي} or {تجارة= تجر} otherwise we delete "ت" {تفرح= فرح}

If "ت" is in third position, and the initial character is "ت" or "ي" we delete both the first and the third character and keep the second { يلتقيان= لقي}.

Letters "ئ" and "ء"

If one of them occurs in the initial position, it is changed into "أ".

If we find "اء" or "ائ", we turn them into "أ", when the length of the word is four letters. { نائم=نام or سيما = سماء}

If the word consists of five letters, we turn "اء" or "ائ", into "ا". In addition, we add it to the list. Otherwise, it is original. { سوداء =سود، ابناء= ابن}.

Letter "ي"

If ""س occurs as the second letter, preceded by one of the following letters "ا"، "ت"، "ي"، "م"، "ن" followed by one by "ي", we deleted the three of them if the word is more than five letters.

If "س" is in initial position followed by "ي," and the word is more than four letters, we delete both letters {سيقود= قود}

If it occurs as the second letter, preceded by ""ا we delete "ي" and keep "ا"

{ ودع = ادع =ايداع }

If {ي} is in initial position and there are only two original letters or fewer, and the list is nonempty, we delete "ي" and retrieve the last letter added except for ""ا.

Letter {و}

If it is preceded by "ال" then it is an original letter.

If "و" is in initial position, and there are two original letters or fewer, and the list is nonempty, we delete "و" and retrieve the last letter added to the list except for ""ت, "ي", "ا". {وخاف= خاف= خوف}

If it occurs in third position, and "و" is the initial letter, we delete the third position and keep the two letters remaining, otherwise they are listed. {ورود=ورد}

If it is in initial position and there are two original letters or fewer, and the list is empty, we keep the {و} {ورد= ورد}

Letter "م"

If "م" is in initial position, and there are two original letters or fewer, and the list is non empty, we delete "م" and retrieve the last letter added to the list.

{موقد= وقد، موعد= وعد}.

If it is initial position, and there are fewer than three original letters the list is empty, then it is an original letter.

If it is in final position in the base word, preceded by -"ه" or "ك" but not followed by a letter in the original word, we delete the last two letters: "هم"، "كم"

If the number of letters in the base word is greater than or equal to five letters. {كتب = كتبكم، كتبتم}.

If it occurs initially followed by "ال", the word followed the base {مالية= مال} otherwise it is original.

Letter {ن}

If there are more than five letters in the base word, and "ن" is the final letter preceded by a vowel, we delete both "ن" and the vowel. { مدرسون= مدرس= درس}

If "ن" is initial position, and there are fewer than three original letters, and the list is nonempty, we delete "ن" and retrieve the last letter added to the list so long as it is not preceded by "و"،"ت"،"ي"،"ا".

If it occurs initially, and the original letters are three or more, then treat "ن" as original.

If it occurs initially, and there are three or fewer original letters, and the list contains "ن" only, "ن" here is original. Otherwise it is an original letter {ترانيم= رنم}

Post Algorithm

If there are fewer than three original letters

If there are two original letters, we retrieve the last letter in the list.

If there is only one original letter, we retrieve the last letter in the list

If there are no original letters, we retrieve the last three letters in the list.

Sorting

We arrange the letters in order of their appearance in the original word.

If there are four letters, we re-examine the resulting word as a new word if it is the same, we examine the first letter, so if it is one of the following letters "ب"،"و"،"ف"،"ك" we delete it, and otherwise the word has a quadruple root.

{لعب,بلعب }.

If there are two letters in the word, we examine the first letter, if it is "ت" or "ا،" we retrieve it, otherwise we report it as the root as it is.

Before Print

If "و" it found, it is turned into "أ".

If the first letter is "ا" or "ي", it is changed into "و", on condition that it is not followed by a vowel.

If the second letter is ""ا it is turned into "و", except when "ب" is initial or the third is "ل".

4. Results:

Table 1: Sample of Success Rates In Tested Data

Number	Number of words on text	Percentage result
1	113	100%
2	200	98.5%
3	550	96.6%
4	650	95.2%

The percentage of the correct words in the four texts =97.6**%**

Problems And Weaknesses

The existence of "ت" in the end of the word, either original or not.

The existence of "ب" as extra letters the beginning of the words.

5. Conclusion

In this paper an Arabic Stemming Algorithm have been designed and implemented, which has been developed for the Arabic language. The algorithm utilizes an important morphological aspect of the Arabic language. The algorithm was implemented in Vbasic. Sample output of the program is shown in Figure 1. We have tested the algorithm with sample data from the Proceedings of the Saudi Arabian National Computer Conferences with a total of 3500 words, the algorithm runs very well and achieves an accuracy rate that reached 97.6%. The words that the algorithm failed to analyze are foreign names and proper nouns.

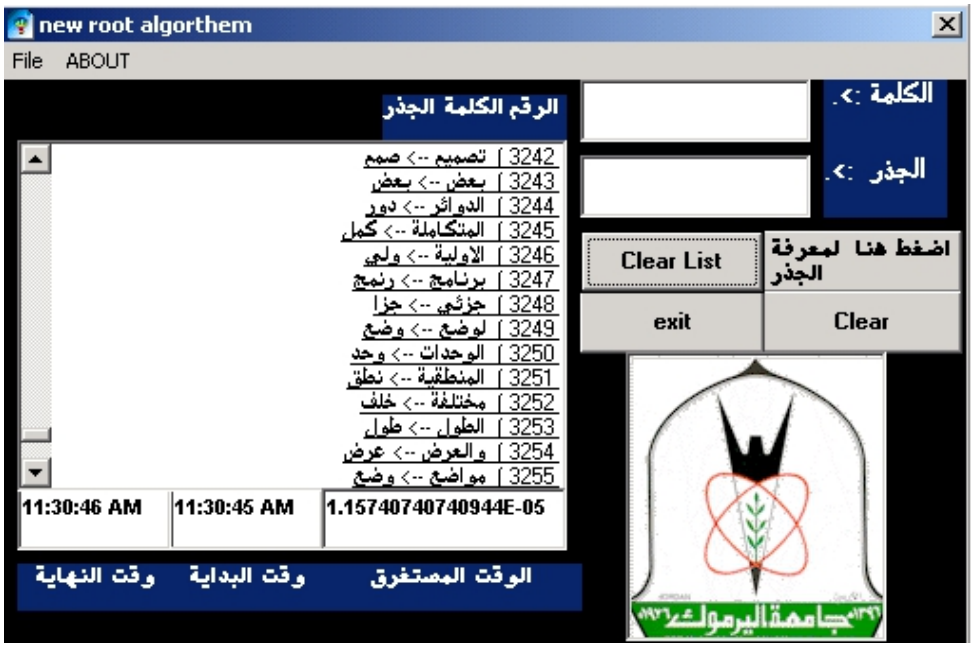

Figure 1: Sample input/output of the system

References:

AI-Ameed, KH. A proposed new model using a light stemmer for increasing the success of search in Arabic terms. PhD Thesis, Bradford, UK: University of Bradford. 2006.

AI-Ameed, KH., Al-Ketbi, O., Al-Kaabi, A., Al-Shebli, S., Al-Shamsi, F., Al-Nuaimi, H., and Al-Muhairi, S. Arabic Light Stemmer: A new Enhanced Approach, The second international conference on innovations technology (IIT'05), 2005.

Al-Fedaghi, S.S., and Al-Anzi, F.S. A New Algorithm to Generate Arabic Root-Pattern Forms, Proceedings of the 11th National Computer Conference and Exhibition, March, Dharan, Saudi Arabia, pp.391-400, 1989

Al-Fedaghi, S.S., and Yaseen, M. Theorem for Automatic Derivation in the Non-Vowelized Arabic Text. The 12th National Computer Conference, King Saud University, Riyadh, Saudi Arabia, October 21-24, Vol.II, pp.660-674, 1990.

Ali, N., Hegazi, N., and Abed, E. A Morphology-Based Data Compression Technique for Arabic Text. 1st African Conference on Computer Communications, Tunis 21-23 May, pp.241-251, 1984

Aljlayl, M., and Frieder, O. On Arabic search: Improving the retrieval effectiveness via light stemming approach. In Proceedings of the 11th ACM International Conference on Information and Knowledge Management, Illinois Institute of Technology (pp. 340–347). New York: ACM Press.2002.

Al-Nashashibi, M., Neagu, D., and Yaghi, A. An improved root extraction technique for Arabic words. 2nd International Conference on Computer Technology and Development, ICCTD 2010.

Bateson, M. C. Arabic Language Handbook. Washington: Center for Applied Linguistics. 1967.

Beesley, K. Computer Analysis of Arabic Morphology: A Two-Level Approach with Detours. Perspectives on Arabic Linguistics III, ed. by Mushira Eid and Bernard Comrie, John Benjamins, Amsterdam, Pp.155-172, 1989.

Beesley, K. Finite-State morphological analysis and generation of Arabic at Xerox Research: status and plans in 2001. In: ARABIC NLP Workshop Status and Prospects ACL-EACL, Toulouse, France 6 July, pp. 1 -8, 2001

Darwish, K., and Oard, D.W. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. Gaithersburg: NIST, pp 703-710, 2002.

EI-Sadany, T.A, and Hashish, M.A. An Arabic morphological system. IBM Systems Journal, 28(4), pp. 600-612, 1989.

Khoja, S. And Garside, R. Stemming Arabic text. Computing Department, Lancaster University, Lancaster, 1999.

Larkey, S., Ballesteros, L., and Connell, M. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval, SIGIR 2002, Tampere, 2002.

Ruhlen, M. A Guide to the World's Languages, Vol. 1: Classification. London: Edward Arnold, 1987.

Semitic Languages. Section from Chapter 5 of John Heise's Akkadian Language', about Semitic languages in general 2013. Available at http://www.sron.nl/~jheise/akkadian/semitic.html Date of Access 1 October 2013.

Touregypt. Egyptian Arabic 2013. Available at http://www.touregypt.net/featurestories/arabic.htm. Date of Access 15 September 2013.

UCLA, UCLA Language Materials Project 2013. Available at http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=52. Date of Access 10 October 2013.