Applied Science University Journal (ASUJ)

http://dx.doi.org/10.18576/jasu/060206

A Priority Based Approach for Stemming Arabic Word

Riyad Al-Shalabi[*]

Applied Science University-Bahrain

E-mail: Riyad.alshalabi@asu.edu.bh

Revised: 1 May. 2022 *Accepted: 30 May 2022* *Published: 1 Jun. 2022*

Abstract: Recently Natural Language Processing of Arabic text has gained a lot of attention. We need to understand many problems in Arabic linguistics in order to carry out the Natural Language Processing projects we want to attack.

This paper describes an algorithm for stemming Arabic words. The algorithm calculates the root based on removing the suggested prefixes and suffixes and then finds out all the possibilities in which the root could be one of them, finally it starts minimizing these possibilities in order to find out the correct root based on the calculation of the number of the original letters of each possibility formed. If the word does not contain any original letters the algorithm tries to find the root letters according to some priorities assigned to letters as described below. We have tested our algorithm on a different set of Arabic abstracts taken from the proceedings of the Saudi Arabian National Computer Conferences, the performance of the algorithm is very high where in some abstracts we reached to 95%.

Keywords: Natural Language Processing, Stemming, Arabic Morphology.

1 Introduction

Arabic is a complex and wonderful language at the same time. The Arabic language expresses a Semitic language and is spoken by nearly 380 million people as the first official language, while it is known as a second language by about 250 million people. Most of the peoples of the Arab countries use the Arabic language in their education, daily life, official dealings, and in social media and television channels. There are several forms of the Arabic language, such as Modern Standard Arabic and Standard Arabic, which is the language of the Holy Qur'an and is also used in literary texts and poems. For more than fourteen centuries, the people of the Arabian Peninsula have spoken the Arabic language. There is another form that differs according to the place you live in, which is the colloquial dialects. This is why the Arabic language faces many challenges. The number of letters in the Arabic language is 28, in addition to the hamza, which is sometimes considered a letter [1]. The way to write in the Arabic language is to start from the right of the paper and end to the left, unlike the English language. The shapes of the letters in this language differ according to their position in the world, for example, the shape of the letter differs if it is at the beginning of the word than if it is in the middle of it. The different shapes and diacritics in the Arabic language make analyzing a difficult task. In addition to the fact that more than 80 percent of Arabic words can be traced back to a root, most Arabic words have their roots. by good fortune, the Arabic language has a built-in filtering process that allows different phrases to be linked to their source. It is very important to represent words for their source and it will undoubtedly reduce the number of words [2].

Stemming is the process of truncating different morphologies of a word to result in a single morphem or base. As a result, text dimensionality is decreased and helps efficient computation. There is a difference between root or heavy stemming and light stemming. The former returns the words to their roots besides affix removal while the last only concerns removing suffixes, prefixes and stop words.

[*] Corresponding author E-mail Riyad.alshalabi@asu.edu.bh

Consequently, root stemming becomes more difficult to process and construct specially with Arabic words since there are so many words that do not depend on specific regulations [3].

2 Arabic Morphology

The morphology science is the one that is specialized in the Arabic language words and their reconstruction which is changing the word form from one skeleton to another or to other different structure that by the end would give certain intended meaning. The morphology science explores the well formed names which are through that are not unchangeable as pronouns. The words in Arabic language are divided into three known types: *verbs* which indicates an action and related to time, *nouns* which indicates meaning not related to time, and *letters* which indicates meaning when connected together.

In a recent study executed by linguistics and statisticians founded that 4814 verbs out of 5629 verbs that form about 85.5% if Arabic verbs were triples, and so the morphologists preferred using the triple scale since the addition is easier than deletion when changing the triple verbs into quadruple verb and penta-letters verbs.

Most ancient language scientists agree on that in Arabic a verb can’t be made of less than three letters and suggest that in most words the roots are consist of three letters. Roots that are less than three letters are have three letters in origin but some letter(s) was/were deleted.

From the root verb we can find many different forms of the same word by adding some prefixes, suffixes or infixes, these additions to the root comes in two forms either from the original letters by repeating some letter(s) or from extra letters addition: which can be exemplified in the mnemonic “ " سألتمونيهاthese extra letters do not exceed 10 letters and no other letters can be used, so the original triple noun can get one, two, three or four letters and so will end up after addition to be of 5 or 6 letters, which most of the time the maximum a triple root can be enlarged to be.

3 Algorithm Descriptions

- Define the following letters (أ- ا- إ - ؤ- و- ئـ - ي - هـ - م - ن- ت - ك) as not original Letters in the Arabic words.
- Define the following Priorities of the following letters from right to left as follows:
  (هـ - ك – ن - م – و- أ – ي – ت- ؤ - ئـ - ا- ء- إ)
- Vowel letters are: **(**و- أ - ي **).**

* remove the following prefixes from the word if any: ال - بال - فال.
* If the word length is five or more letters and have the prefixes تت – سأ - لل then
we remove them as in the words "للذهاب “ or "تتوقع “ or “سأذهب " .
* If the word length is six or more letters and have the prefixes است – أست – مست – إست
then we remove them as in the words “ استشفى“or “مستحسن ".
* If the word length is five letters and has two original letters and ends with the
suffix (ني)then we remove it as in the word “قادني ”.

2. Remove the letters (ة ، هـ) if they occurs at the end of the word as in the words "مدرسة “ or "فكره “.

- ■ If the word length is six or more letters and have one of the following suffixes ات – ني – تي – هما – هم – هن – ها – كن – كم - كما then remove them.
- ■ If the word length is five or more and contains a vowel letter between the second and the one before the last, then this vowel will be removed; else don’t do anything, as in the word "ملعون “ remove the letter (و) but no letters are removed for example in the word "طاولة “ .

- If the word length is five or more letters and has one of the following suffixes: كن – كم – كما – ها هن – هم – هما – ني – تي - ات –then remove them.
- If the word length is four letters or more and ends with the letter (ت) or begins with the letter (ت)then remove it.

6. If the word contains three original letters then the root will be the first three original letters; Else
7. Find all possible words of length three.
8. If the word contains two original letters then
8-1. If there is letter(s) between them then
8-1-1. The priority is for the possibility which contains the letter (و).
8-1-2. If there are more than one possibility, then the priority is given to the possible word that begins with an original letter.
8-1-3. If there are more than one possibility, then the priority would be according to the priority given to the letters as shown above.
8-2. If the word does not contain letter(s) between the originals, then
8-2-1. If the second original letter is in the middle of the word, then we take the letter that comes after the second original one if it is not a vowel,but if it is a vowel then we take the letter that comes after it, and if there is no letters following the vowel, then we take the letter that precedes the first original letter unless if it is "ت or ا" . If it is "ت or ا " then take the letter which precedes them both.
8-2-2. If the second original letter is at the end of the word, then take the preceding letter of the first original unless if it is "ت or ا" If it is "ت or ا “then take the letter that precedes them both.
9. If the word contains one original letter, then
9-1. If the word contained the letter (ك) in the first half of the word (the smallest integer close to n/2) then the priority will be given for the possibility containing this letter; else we examine this condition on the letter (ن) and the priority will be given for the possibility containing this letter; else we examine this condition on the letter (هـ) and the priority will be given for the possibility containing this letter (here, we are trying to find a letter as a second original letter).
repeat steps (8-1) to (8-2) again if the word contains the previous letters (ك، ن،هـ) in the first half of the word.
9-2. else, if we didn’t find a letter, then choose the letter according to its priority and do step (8) again, as in the word “وجوهها”.
10. If the word does not contain any original letters, then get two originals letters according to the priorities given above, and find out the root as if it has two original letters**.**

4 Conclusions

An algorithm was designed and implemented for analyzing Arabic words to find their roots. The system was implemented using C++. The algorithm was tested on different set of Arabic abstracts taken from the proceedings of the Saudi Arabian National Computer Conferences. The performance of the algorithm is very high where in some abstracts we reached to 95%. But as we know that Arabic language has many special cases where some words contain vowel letter, the algorithm suffer from handling some of these cases, and I hope that in the near future I can overcome these cases and have an algorithm that will handle a high percentage of irregular Arabic words.

References

[1] P. Shinde, P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. In 2018 Fourth International Conference on Computing Communication Control and Automation, 1-6.

[2] Kanan, T., Odai, S., Almhirat, A., & Kanan, E. (2019). Arabic light stemming: A comparative study between p-stemmer, khoja stemmer, and light10 stemmer. 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 511-515). IEEE.

[3] M. Ababneh, R. Al-shalabi, G. Kanaan and A. Al-nobani, building an Effective Rule-BasedLight temmer for Arabic Language to Improve Search Effectiveness," International Arab Journal of Information Technology, vol. 9, no. 4, p. 5,2012.

[4] Aldabbas O., Al-Shalabi R, Kanaan G., Shehab M., Albadarneh M., and Mahyoub N. (2016). Arabic Light stemmer Based on regular Expression technique. International Journal of Advenced Studies in Computer Science and Engineering. Vol. 5, issue 11.

[5] Al-Shalabi, R and Evens, M. A computational morphology system for Arabic. In Workshop on computational Approaches to Semitic Languages, COLING-ACL98, August 1998.

[6] Mustfa, S. and Masoud, F. A Backward Algorithm for Lexical Analysis of Textual Arabic Words. Abhath Al-Yarmouk: Basic and Engineering Sciences Series. Vol 9, No. 1, 2000, pp. 91-122.

[7] Hmeidi, I., Al-Shalabi, R., Al-Taani, A., Najadat, H., and Al-Hazaimeh, Sh. (2010), A novel approach to the extraction of roots from Arabic words using bigrams. JASIST: 583~591

[8] Kanan, T., Sadaqa, O., Aldajeh, A., Alshwabka, H., AL-dolime, W., AlZu'bi, S., .A. Alia, M. (2019). A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media. Jordan International Joint Conference on Electrical Engineering and Information Technology, 622-628.

[9] Yahya, A. H. Morphology and Syntax of the Arabic Language. Computers and the Arabic Language, ed. by Pierre A. MacKay. Hemisphere Publishing Corporation, New York, NY, 1990, pp.201-207.