Journal of Computer Science 10 (4): 716-719, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.716.719 Published Online 10 (4) 2014 (http://www.thescipub.com/jcs.toc)

AN IMPROVED ARABIC WORD'S ROOTS EXTRACTION METHOD USING N-GRAM TECHNIQUE

[1]Nidal Yousef, [2]Aymen Abu-Errub, [1]Ashraf Odeh and [3]Hayel Khafajeh

[1]Department of CIS, Faculty of Information Technology, AL Isra University, Amman, Jordan
[2]Department of CIS, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan
[3]Department of CIS, Faculty of Computing and Information Technology, Zarqa University, Zarqa, Jordan

Received 2013-08-11; Revised 2013-10-07; Accepted 2013-12-26

ABSTRACT

Arabic language is distinguished by its morphological richness, which forces the workers in the field of Arabic language Processing (i.e., information retrieval, document's classification, text summarizing) to deal with many words that seem to be different but in reality they came from an identical root word. One of the methods to overcome this problem is to return the words to their roots. This research aims to provide a new algorithm, that returns roots of Arabic words using n-gram technique without using morphological rules in order to avoid the complexity arising from the morphological richness of the language in one hand and the multiplicity of morphological rules in other hand. The proposed algorithm uses a list that contains over 4,500 identical roots words.

Keywords: Arabic Root Extraction, Natural Language Processing, N-Gram

INTRODUCTION

Arabic is one of the major languages in the world, its spoken by over 400 million people and being the language of the Holy Quran, it is also used by more than 1.5 billion Muslims all over the world, making it the largest Semitic languages Ghazzawi (1992). Arabic alphabet consists of 28 characters and written from right to left and it uses cursive letters. Each Arabic word is formed from the root word and a suffix, a prefix or an infix.

There are many Arabic language computerized applications rely on using of the roots of words, such as information retrieval systems, text classification, text summarization, auto-translation, Data mining, OCR (Ghwanmeh *et al*., 2009; Yousef *et al*., 2010) and other applications. The Arabic word's roots can be classified according to the vowels letters into two types (Wightwick and Gaafar, 2007), the first type is the strong roots which is the root that does not contain a vowel, whereas the roots that containing at least is called vocalic roots. Arabic roots can be further classified according to the number of their characters into four types: *Triliteral* (which forms most words in Arabic language) Al-Kamar (2006), *Quadriliteral*, *Quinquelitera* and *Hexaliteral*.

There are many methods to extract the roots of Arabic words, but there is no agreement on one method because of morphological affluence of the Arabic language and the large number of conflations for each word. The Arabic language researchers produce many methods to extract the roots of Arabic words during the last decade (Hajjar *et al*., 2010; Al-Nashashibi *et al*., 2010) specially strong *Triliteral* roots. Many researchers rely on morphological rules on their extraction methods, which make the process of extracting the root difficult and complex because of the multiplicity morphological formulas and multiplicity of words forms for the same root because of changing the original characters position in the word (i.e., كتب, كاتب, كتاب). In the study, the researchers will introduce an improved method to extract the word's root without using morphological rules, but using n-gram technique which simplifies the process of extracting the roots.

The study consists of four sections. The first section is a literature review in which the researchers review

Corresponding Author: Aymen Abu-Errub, Department of CIS, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan

some papers that deal with extracting Arabic word's roots. In second section the researchers introduce the proposed algorithm. The third section presents the experiment that the researches conduct in order to test their proposed algorithm and also introduces the results they obtained from these tests. In the last section the researchers conclude the research.

2. LITERATURE REVIEW

During the last two decades, many researchers proposed new approaches to extract Arabic words roots, some of these approaches using morphological analysis, whereas other approaches relied on statistical methods.

Hawas (2013) the author presents a new root-extraction approach for Arabic words that tries to assign a unique root for each Arabic word without having an Arabic roots list, a words patterns list, or the Arabic word's prefixes and suffixes list. The proposed algorithm predict the letters positions that may form the word root one by one, using rules based on the relations between the Arabic word letters and their placement in the word. The proposed algorithm consist of two parts, the first part deals with the rules that distinguish between the Arabic definite letter "ال AL, La" and the original word letters "الـ". The second part of the approach adopts the segmentation of the word into three parts and classifies its letters into groups according to their positions. The proposed approach was composed of several corporate modules. The researcher tested the proposed approach using the Holy Quran words shows a promising root extraction algorithm, the outher shows that she had the the total success ratio about 93.7% but she considered the root is correct if it has one correct letter.

Boudlal *et al*. (2011) the researchers provide a new way to find the system that assigns, for every non vowel word a unique root depending on the context of the word on the sentence. The proposed system is composed of two modules. The first one consists of analyzing the context by segmenting the words of the sentence into its elementary morphological units in order to identify its possible roots. The researchers adopt the segmentation of the word into three parts (prefix, stem and suffix). In the second module the researchers use the context to identify the correct root among all possible roots of the word. For this purpose, the researchers use a Hidden Markov Models (HMM) approach, where the observations are the words and the possible roots represent the hidden states. The researchers validate their approach using the NEMLAR Arabic writing corpus that consists of 500,000 words, in this research the proposed algorithm gives the correct root in more than 98% of the training set and in almost 94% of the words in the testing set.

Hmeidi *et al*. (2010) the researchers provide a new way to find the roots of Arabic words using bigrams. The researchers use two similarity measures; the dissimilarity measurement, or the "Manhattan distance measurement" and the "Dice's measurement". The researchers test their proposed algorithm on the Holy Quran and on a corpus of 242 abstracts from the Proceedings of the Saudi Arabian National Computer Conferences. The researcher conclude from their research that combining the n-grams with the Dice's measurement gives better results than using the Manhattan distance measurement.

The researchers in Al-Nashashibi *et al*. (2010) address linguistic approach for root extraction as a pre-processing step for Arabic text mining. The proposed approach is composed of a rule-based light stemmer and a pattern-based infix remover. They propose an algorithm to handle weak, eliminated-longvowel, hamzated and geminated words The accuracy of the extracted roots is determined by comparing them with a predefined list of 5,405 triliteral and quadriliteral roots. The linguistic approach performance was tested on an in-house text collection of eight categories, the researchers gained a success ratio about 73.74%.

Momani and Faraj (2007) the researchers proposed a novel algorithm to extract tri-literal Arabic roots. The first step of the algorithm is done by eliminating the stop words and then the prefixes and suffixes of each word are removed. In the next step the repeated word's letters that are removed until only three letters are remained. Finally, the remaining letters are arranged according to their order in the original word, which form the root of the original word. The researchers tested their algorithm on two types of Arabic text documents. The researchers claimed that the results of both runs were very promising and satisfactory showing over 73% of accuracy.

3. PROPOSED AlGORITHM

In the following section, the researches introduce the steps that have been implemented in order to reach to the new roots extraction algorithm. Because of difficulties that facing the approaches based on morphological analysis especially with words that containing vowels later, the new algorithm will use the n-gram method. N-gram is basic text analysis tool that used in natural language processing. In this technique, both the word and its assumed root are divided into pairs (called bi-gram, or digram) then the similarity between the word and the root is calculated using Equation (1) Frakes (1992). This process is repeated for each root in the roots list:

$$S=2C/(A+B) \quad (1)$$

Where:
A = Number of unique bi-grams in the word (A)
B = Number of unique bi-grams in the root (B)
C = Number of similar unique pairs between the word (A) and the root (B)

To utilize Equation (1) for extracting the word's root, we must have: the word (A) and the potential roots (B) to compare with, then the similarity measuring is conducted by computing the value of S between the word (A) and each potential roots (B). A corpus of 4500 triliteral Arabic roots was used to accomplish the similarity calculation step. Only triliteral roots were chosen because they form about 85% of the Arabic language roots.

To extract the root of the word by the proposed algorithm, both the tested word (A) and the candidate root (B) must be divided into pairs of sequence liters and then only the unique pairs will be taken to calculate the similarity (S). The root that has the highest (S) value among the roots list is considered as the root of the word.

For example, if we had the word "مستعرض", which its root is "عرض". The values of A, B, C, S are shown in the following:

The word (*A*): مستعرض (م س ت ع ر ض)
Bi-gram sequence: مس ست تع عر رض
Number of unique bigrams in the word, A = 5

The candidate root (*B*): عرض (ع ر ض)
Bi-gram sequence: عر رض
Number of unique bigrams in the root, B = 2

Number of similar unique pairs between the word and the root, C = 2

Then using Equation (1), we can calculate the similarity (S) between the word and the root:

$$S=2C/(A+B)$$
$$S=2*2/(5+2)=4/7$$

In the previous example the similarity (S) was calculated easily because we compared the word with its actual root directly, but in real situations the extraction of the root will be conducted without having the actual root, which means that we have to calculate the similarity (S) between the word and all candidate roots. For the purpose of this research, we used a list of 4500 Arabic roots. The process of extracting the actual root is shown in the following algorithm:

1. Normalization of the word: by deleting the word diacritics (Alhmza:ء) and convert the letter (ة) to the letter (ه).
2. Divide the word into bi-grams pairs.
3. Find the number of unique bi-grams in the word (A).
4. Choose a candidate root (B) from the root list and apply steps 2 and 3 to find the number of unique bi-grams in the root (B)
5. Calculate the similarity (S) between the word (A) and the candidate root (B).
6. Repeat steps 4 & 5 for the rest of roots in the roots list.
7. The root that has the highest similarity (S) among the roots in the list is chosen to be the root of the word (A).

4. EXPERIMENTAL RESULTS

To examine the proposed algorithm the researcher designed a corpus consisting of 141 roots chosen from the 4500 roots list. The corpus contains 6308 morphological forms derived from these 141 roots. Among these morphological forms there are 1318 morphological forms belonging to 21 vowel roots. **Figure 1** below demonstrates an example of the morphological forms used in experiments for the root "كتب" (i.e., write).

After running the proposed algorithm on the designed corpus the results were as follows.

4.1. Tripartite Strong Roots

When examining morphological forms of strong triliteral roots that do not containing a vowel the results were as shown in **Table 1**.

4.2. Tripartite Roots with Vowels

When examining morphological forms for the Tripartite roots with vowels the results were as shown in **Table 2**. The results are similar with the strong triliteral roots due to reliance on statistical methods, without taking into account the vowels.

4.3. All Roots

When examining all morphological forms from the designed corpus the results were as shown in **Table 3**.

كتبتم	كتبن	كتبوا	كتبتا	كتبت	كتب
يكتبان	تكتب	يكتب	كتابة	كتبنا	كتبتن
اكتب	يكتبن	تكتبوا	يكتبون	تكتبون	تكتبان
كتّبة	كتابات	اكتبن	اكتبوا	اكتبا	اكتبي
كاتبون	كاتبات	كاتبان	كاتب	كتبة	كتّبة
مكتبة	مكتبات	مكاتب	مكتبان	مكتب	مكتوب
كتابان	مكاتبة	كتيّب	اكتب	كُتّاب	كُتّاب

Fig. 1. The morphological forms of the root "كتب"

Table 1. Tripartite strong roots results

The actual root is:	Roots number	Ratio
The only root in the resultant list	1991	0.399398
In the top of resultant list With other roots	1194	0.239519
In the resultant list	786	0.157673
Not in the resultant list	1014	0.203410

Table 2. Tripartite roots with vowels results

The actual root is:	Roots number	Ratio
The only root in the resultant list	513	0.3883422
In the top of resultant list With other roots	342	0.2588948
In the resultant list	132	0.0999243
Not in the resultant list	334	0.2528388

Table 3. All roots results

The actual root is:	Roots number	Ratio
The only root in the resultant list	2504	0.3971
In the top of resultant list With other roots	1536	0.2436
In the resultant list	918	0.1456
Not in the resultant list	1348	0.2138

5. CONCLUSION

In this research an improved extraction Arabic root algorithm was proposed using bi-gram technique. The results showed that the proposed algorithm is capable of extract the most possible root for nearly 80% of the strong roots, by choosing the roots that has the highest similarity value between the desired word and the candidate roots. The proposed approach succeeded in extracting the vocalic roots in a similar ratio with the strong roots. Our future plan and works are to improve the proposed algorithm by using morphological rules and artificial intelligence techniques, to enhance the preliminary results that emerged after extracting the value of similarity.

6. REFERENCES

Al-Kamar, R., 2006. Computer and arabic language computerizing. Dar Al Kotob Al-Ilmiya, Cairo, Egypt.

Al-Nashashibi, M.Y., D. Neagu and A.A. Yaghi, 2010. An improved root extraction technique for Arabic words. Proceeding of 2nd International Conference on Computer Technology and Development (ICCTD), Nov. 2-4, IEEE Xplore Press, Cairo, pp: 264-269. DOI: 10.1109/ICCTD.2010.5645872

Boudlal, A., R. Belahbib, A. Belahbib and A. Mazroui, 2011. A markovian approach for Arabic root extraction. Int. Arab J. Inform. Technol., 8: 91-98.

Frakes, W.B., 1992. Stemming Algorithms. In: Information Retrieval: Data Structures and Algorithms, Frakes, W.B. and R. Baeza-Yates (Eds.), Prentice-Hall India, ISBN-10: 8131716929, pp: 131-160.

Ghazzawi, S., 1992. The Arabic Language in the Class Room. 2nd Edn., Georgetown University, Washington DC.

Ghwanmeh, S., G. Kanaan, R. Al-Shalabi and S. Rabab'ah, 2009. Enhanced algorithm for extracting the root of Arabic words. Proceeding of the 6th International Conference on Computer Graphics, Imaging and Visualization, Aug. 11-14, IEEE Xplore Press, Tianjin, Chain, pp: 388-391. DOI: 10.1109/CGIV.2009.10

Hajjar, A.E.S.A., M. Hajjar and K. Zreik, 2010. A system for evaluation of Arabic root extraction methods. Proceeding of 5th International Conference on Internet and Web Applications and Services (ICIW), May 9-15, IEEE Xplore Press, Barcelona, pp: 506-512. DOI: 10.1109/ICIW.2010.98

Hawas, F.A., 2013. Exploit relations between the word letters and their placement in the word for Arabic root extraction. Comput. Sci., 14: 327-431. DOI: 10.7494/csci.2013.14.2.327

Hmeidi, I.I., R.F. Al-Shalabi, A.T. Al-Taani, H. Najadat and S.A. Al-Hazaimeh, 2010. A novel approach to the extraction of roots from Arabic words using bigrams. J. Am. Soc. Inform. Sci. Technol., 61: 583-591. DOI: 10.1002/asi.21247

Momani, M. and J. Faraj, 2007. A novel algorithm to extract tri-literal Arabic roots. Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications. May 13-16, IEEE Xplore Press, Amman, pp: 309-315. DOI: 10.1109/AICCSA.2007.370899

Wightwick, J. and M. Gaafar, 2007. Arabic Verbs and Essentials of Grammar, 2E (Verbs and Essentials of Grammar Series). 2nd Edn., McGraw-Hill Companies, Inc., ISBN-10: 0071498052, pp: 160

Yousef, N., I. Al-Bidewi and M. Fayoumi, 2010. Evaluation of different query expansion techniques and using different similarity measures in arabic documents. Eur. J. Sci. Res., 43: 156-166.