International Conference on Engineering Research & Applications
(ICERA-17)
Istanbul (Turkey) May. 17-18, 2017

New Method for Stemming of Arabic Language Text

Amrouche Aissa, Abed Ahcène, and Boubakeur Khadidja Nesrine

Scientific and Technical Research Center for the Development of the Arabic Language, Algeria
amrouche_a@yahoo.fr; abedahcene@gmail.com; boubakeur.khadidja@gmail.com

Abstract**:** *Because of its complex morphology, the Arabic language has a very different and difficult structure than other languages. Several stemming approaches that are applied to Arabic language, but a complete stemmer for this language is not available. The existing stem-based stemmers for stemming Arabic text have a poor performance in terms of accuracy and error rates. The aim of this study is to build an effective stemmer that answer the problems of Information Retrieval (IR), and presents new way to build electronic Arabic lexicon by using the most frequency roots as the input of lexicons.*

Keywords: *Word Stemming, Arabic Language, Information Retrieval, Arabic lexicon.*

1. Introduction

Stemming algorithm for Arabic words has been an important topic in Arabic information retrieval. Many stemming methods have been developed for Arabic language in IR systems but they suffer from many problems. These stemmers are classified into two categories. The first one is root extraction stemmer like the stemmer introduced by Khoja [1]. He attempts to find roots for Arabic words by first removing prefixes and suffixes, and then tries to determine the root from the stripped words using a dictionary of root words. The second is light stemmers like the stemmer introduced, such as the algorithm developed by Larkey [2], Darwish [3] and Chen [4] select some prefixes and suffixes to be truncated from the words and produce the stems. We envisage that the approach adopted by Khoja [9] is more appropriate in determining roots or stems, since the dominant present of infixes in Arabic words. The proposed method integrates different stemming techniques, including: morphological analysis, affix-removal and patterns dictionaries.

2. Arabic language

Arabic is a Semitic language of the same family as the Syriac, Aramaic and Hebrew. Nowadays it is spoken by almost 450 million people in the world and 22 countries as well. The Arabic language is considered as difficult to master in automatic signal processing and Natural language processing because of its morphological and syntactic properties [5, 6]. The research about the automatic processing of Arabic has started in the 1970s. The first studies were primarily focused on lexicons and morphology. We will state some peculiarities of the Arabic language.

- The Arabic alphabet has 34 graphemes including 28 consonants, 3 short vowels and 3 long vowels consonants,
- Arabic is written and red from the right to the left.
- Letters take different forms depending on their position in the word: initial, median, final or isolate (Table I).

TABLE I: Example of Letter غ Ghayn Variation's

The letters change	final	medial	initial
غ	غ	ـغـ	غـ

- The Arabic language has three categories of words (part-of-speech): verbs, nouns and particles.

3. Methodology

The analysis involves the following phases (Figure 1):

- Decompose the input text into set of lexical sequences (words).
- Normalize the input text;
- Eliminate the stop words;
- Determine the morphological characteristics for each word;
- Remove prefixes and suffixes based on morphological characteristics and various dictionaries;
- Determine the possible roots for each word based on patterns dictionaries.

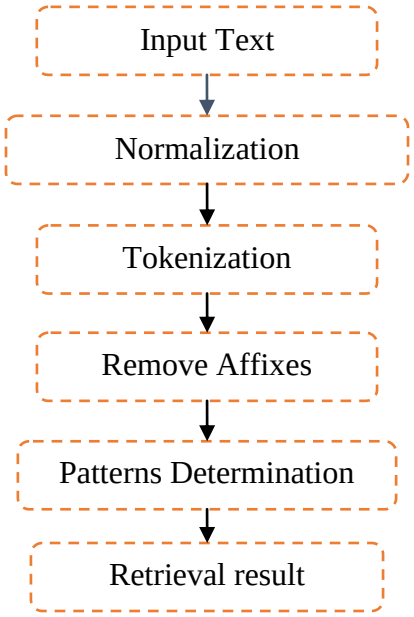

Fig. 1: Stemming Architecture system

The normalization approach involves the following steps:

- Remove punctuation;
- Remove diacritics (primarily weak vowels);
- Remove de definite article ال.
- Remove non-letters (symbols and numbers);
- Replace the initial إ or أ by Alif nu ا;
- Replace the آ by the ا;

3.1. Tokenization

Tokenization is very important step in natural language processing. The only task to do is to decompose the input text into set of states (words and punctuation). The result of this decomposition is illustrated by the example of figure 2:

Input =عَمَّ يَتَسَاءَلُونَ عَنِ النَّبَإِ الْعَظِيمِ الَّذِي هُمْ فِيهِ مُخْتَلِفُونَ

Input_without_vowels =عم يتساءلون عن النبإ العظيم الذي هم فيه مختلفون

Output =

'عم'

'يتساءلون'

'عن'

'النبا'

'العظيم'

'الذي'

'هم'

'فيه'

'مختلفون'

Fig. 2: Example of the Tokenizer

3.2. Arabic patterns and affixes.

We have three lists: Patterns, prefixes and suffixes list, which we have based on to determine the root as follows:

- Arabic prefixes

S_B={'ال' 'وال' 'فال' 'أال' 'أفال' 'بال' 'وبال' 'فبال' 'أبال' 'لبال' 'أفبال' 'لل' 'ولل' 'فلل' 'أفلل' 'كال' 'وكال' 'فكال' 'أكال' 'لكال' 'أفكال'};

- Arabic suffixes

S_E={'اها' 'اهما' 'اهم' 'اه' 'اك' 'اكا' 'اكما' 'اكم' 'اهن' 'اكن'};

- Patterns list example

Patterns={'مفعول' 'مفعولين' 'مفعولان' 'مفعولون' 'فاعلة' 'فاعلتان' 'فاعلتين' 'فاعلات' 'فاعل' 'فاعلان' 'فاعلين' 'فاعلون' 'مفعلة' 'مفعلتين' 'مفعلتان' 'مفعلات'...}

Here is an example of segmentation of the word فسيعلمون Table II.

TABLE II: Segmentation example

word	Suffix	word without suffix	pathern	root
فسيعلمون	فس	يعلمون	يفعلون	علم

3.3. Stemming Technique

To find the root of the word this technique is based on the pattern matching [6, 7, 8]. The root is extracted after removing the affixes attached to the given word. The root extraction process is started comparing a given word with the patterns list with the same length. We give "1" for similar letters and "0" otherwise, than we take the pattern with has the maximum mark. The Figure 1 describes the pattern matching process in Arabic. It illustrates the full process of root extraction for the word wa-khalaqnakum (وخلقناكم, And created you) according to the corresponding pattern fGalnakum (فعلناكم). The root Khalaqa خلق is extracted from the word wa-

khalaqnakum (وخلقناكم) by taking the letters of the word that are in the positions of the main letters (ف,ع,ل) of the pattern fGalnakum (فعلناكم). Extraction of these letters produces the root khalaqa (خلق, create) figure 3.

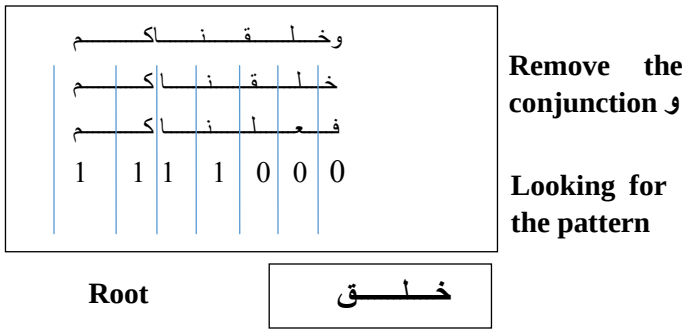

Fig. 3: Tokenizer Root extraction process of the word (وخلقناكم , And created you)

3.4. Stemming Rules

The stemming algorithm needs some rules to solve problems specific word. So if the extracted root contain the letters و, ي and all kinds of Hamza (The letter Hamza takes different forms depending on its position in words and the preceded letter (e.g لأ ئ ؤ إ أ ء).

Replace the Final أ with ى

- Replace the{'ؤها' 'ئهم' 'ءتهم' 'ءين' 'ءون' 'ؤون' 'ءت' 'ءان' 'ءتان' 'ءات' 'ءهم' 'ءكن' 'ءكما' 'ءهما' 'ءكم' 'ءهن' 'ءك' 'ءها' 'ءيه' 'ءه'
'ئهما' 'ئهن' 'ئها' 'ؤهما' 'ؤهم' 'ؤهن'} with ء
- Replace 'اها' 'اهما' 'اهم' 'اه' 'اك' 'اكا' 'اكما' 'اكم' 'اهن' 'اكن' with ى
- Replace 'يات' 'يان' 'يين' 'اوات' 'اوان' 'واو' 'واوين' 'وانان' 'وانين' 'وين' 'تان' 'تين' 'تات' 'هات' with 'اة'

Example: The extracted roots of the words:

- ومرعاها and أرساها is مرعى and أرسى respectively.
- ماءها is ماء
- الحيوان is حياة

The Table III shows some results obtained at the output of the software and their appearance frequencies. This task is very important to have the input of lexicons.

Table III: software output words and frequencies for the Hizb 59 and 60 of the Holy Qur'an.

Occurrences	AM_Stemmer output
10	كذب
9	قال
9	ذكر
8	خلق
7	علم
5	جعل
5	خرج
5	يوم
5	أمر
5	نفس

4. Results and Conclusion

To evaluate our system we have collect a database for the experimental retrieval system consists of the Quran collection which contains the Hizb 59 and 60. We have storted the original word with the extracted output of the AM_Stemmer. We have done the mark one (1) for the correct extracted word and zero (0) elsewhere Table 2.

Table IV: evaluation results

occurrences	output	words	mark
1	نبأ	النبإ	1
1	عظيم	العظيم	1
17	الذي	الذي	1
10	هم	هم	1
1	في	فيه	1
1	مختلف	مختلفون	1

We obtain 88 % as correct extraction. This result push us to evaluate this system for future work.

5. References

[1] Khoja, Shereen (2001), Stemming Arabic Text. http://zeus.cs.pacificu.edu/shereen/research.htm Larkey

[2] Larkey, L. Ballesteros, and M. Connell (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. SIGIR. pp. 275-282.
https://doi.org/10.1145/564376.564425

[3] Darwish, K. Building a Shallow Morphological Analyzer in One Day. ACL Workshop on Computational Approaches to Semitic Languages.
https://doi.org/10.3115/1118637.1118643

[4] Chen, A., Gey, (2002). Building an Arabic Stemmer for Information Retrieval. TREC-2002

[5] Syed A. Barakat(1985)., Introduction to Qur'anic Script, Curzon Press, London.

[6] Lazrek A (2002), Vers un système de traitement du document scientifique arabe. Thèse de Doctorat, Université Cadi Ayyad Marrakech Maroc.

[7] Mohamed haCm al-XTaT (1986)., Les règles de la calligraphie arabe, Ensemble calligraphique des styles d'écritures arabes, Univers des livres, Beyrouth, Liban.

[8] Kiraz G. A. (1996). Analysis of the Arabic Broken Plural and Diminutive, In Proceedings of the 5th International Conference and Exhibition on Multi-Lingual Computing (ICEMCO96), Cambridge, UK

[9] S. Baloul, M. Alissali, M. Baudry (2002), P. Boula de Mareüil : Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe, 24es Journées d'Étude sur la Parole, Nancy, pp.329-332.

[10] Ghazali .S (1981), La coarticulation de l'emphase en arabe. Dans Etudes de linguistique arabe, Arabica Journal, Paris, France, Vol. 28, n° 2-3, pp. 251-277.
https://doi.org/10.1163/157005881x00258

[11] Al-Sughaiyer, I. and Al-Kharashi (2004), I. "Arabic Morphological Analysis Techniques: A Comprehensive Survey". Journal of the American Society for Information Science and Technology. Vol 55. Issue 3. pp. 189 – 213, 2004.
https://doi.org/10.1002/asi.10368

[12] Beesley. K (1998). Arabic Morphological Analysis on the Internet. In the proceedings of the 6th International Conference and Exhibition on Multilingual Computing, Cambridge.

[13] Khoja. S (1999). Stemming Arabic Text. Computing Department, Lancaster University. Lancaster, U.K.