Journal of King Saud University – Computer and Information Sciences (2015) 27, 94–103

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

A novel root based Arabic stemmer

Mohammed N. Al-Kabi [a], Saif A. Kazakzeh [b], Belal M. Abu Ata [b], Saif A. Al-Rababah [c], Izzat M. Alsmadi [d,*]

[a] Faculty of Sciences and IT, Zarqa University, P.O. Box 2000, 13110 Zarqa, Jordan
[b] CIS Department, IT & CS Faculty, Yarmouk University, 21163 Irbid, Jordan
[c] Information Systems Department, IT Faculty, Al-albayt University, Jordan
[d] Computer Science Department, Boise State University, Boise, ID 83725, USA

Received 26 December 2012; revised 7 December 2013; accepted 3 April 2014
Available online 21 March 2015

KEYWORDS

Natural Language Processing (NLP);
Computational intelligence;
Stemming;
Information retrieval

Abstract Stemming algorithms are used in information retrieval systems, indexers, text mining, text classifiers etc., to extract stems or roots of different words, so that words derived from the same stem or root are grouped together. Many stemming algorithms were built in different natural languages. Khoja stemmer is one of the known and widely used Arabic stemmers. In this paper, we introduced a new light and heavy Arabic stemmer. This new stemmer is presented in this study and compared with two well-known Arabic stemmers. Results showed that accuracy of our stemmer is slightly better than the accuracy yielded by each one of those two well-known Arabic stemmers used for evaluation and comparison. Evaluation tests on our novel stemmer yield 75.03% accuracy, while the other two Arabic stemmers yield slightly lower accuracy.

1. Introduction

Semitic languages are mainly used in the Middle East, and North Africa. The Arabic language is currently the most used Semitic language, since it is the native language for more than 290 million people Worldwide (Arabic language, 2015). These Semitic languages use the writing style from right to left. Most Semitic scripts use Abjad style. Abjad is a type of alphabet that omits some or all vowels. Not all Semitic languages use a cursive style (Abjad, 2012; Semitic languages, 2012) like the Arabic language (Arabic language, 2015). Semitic languages use non-concatenative (i.e. discontinuous) morphology to form words which represent a modified version of roots (Non-concatenative morphology, 2012; Semitic languages, 2012). Most of Semitic roots consist of three consonants (Triliteral) (Semitic languages, 2012). Affixes are used by Semitic languages. However, most of the words are formulated by vowels between the root consonants (Semitic languages, 2012). Therefore extracting the Semitic roots of different Semitic words is usually not a trivial process.

The official Arabic language also called Modern Standard Arabic (MSA) or Literary Arabic is widely used in schools, universities, academic establishments, newspapers, radio, TV stations, government agencies...etc. Arabic language is

* Corresponding author.
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.04.001

based on 28 letters, where the shapes of some of these letters are changed according to their location in the word. In addition, these letters can be joined together or written separately based on their location in the word. Several vowel diacritics are used especially in the holy Qu'ran and in classical poetry.

Not all Arabic words used in MSA are native Arabic words which are derived from Arabic three consonants' (i.e. Triliteral) origin. These include for example, the following Arabic words which lack authentic Arabic roots, since they are not derived from native Arabic roots while they are phonetically modified Arabic versions from their origins in other languages: (e.g. Television, "تلفاز، تلفزيون"), (Programmer, "مبرمج"), (Telephone, "تلفون"), (Computer, "كومبيوتر"), (Dictionary, "قاموس"), (Chemistry, "كيمياء"), (Physics, "فيزياء"), (Geography, "جغرافية"), (Lemon, "ليمون"), (Orange, "برتقال").

In natural languages it is normal to find a number of words derived from the same root or stem. Stemming is the process of extracting the root of each word, in order to treat a group of words that are derived from the same root as synonyms, since they suppose to refer to the same concept. However, in reality not all words which are derived from the same root may refer to the same concept. Stemming process is widely used in information retrieval, text mining, text classification. . .etc.

The following four Arabic words (Written, "مكتوب"), (Writings, "كتابات"), (Writer, "كاتب"), (Book, "كتاب") are derived from the same Arabic three consonants triliteral with origin verb (Wrote, "كتب"). They also refer to the same concept. Therefore stemming these four Arabic words is useful for some relevant tasks. On the other hand, the stemming of the following two Arabic words (accountant, "محاسب") and (computer, "حاسوب") which are derived from the same Arabic triliteral verb (counted, "حسب") shows that stemming is not beneficial, since these two Arabic words are not synonyms, and refer to two different concepts. Further, the following four Arabic Words: (Books, "كتب"), (Office, "مكتب"), (Library, "مكتبة"), (Writing, "كتابه") represent four different concepts that are derived from the same Arabic triliteral verb (Wrote, "كَتَبَ"). These examples show that Arabic stemming is not always straightforward where even if an automatic extraction tool is very accurate, when evaluating the semantics, some of the stemming activities are not relevant.

There are two types of stemming, the first type is light stemming which is used to remove affixes (i.e. prefixes and suffixes), while the second type is called heavy stemming (i.e. root stemming) which is used to extract the root of the words and include implicitly light stemming.

In this study, a novel Arabic stemming algorithm is proposed, implemented, and tested. The algorithm applies both the light and heavy (root) stemming techniques on Arabic words to extract the triliteral roots of words. Our Arabic stemming algorithm is not dictionary based. The conducted tests on this stemming algorithm reveal an accuracy of 75.03%. The results are compared with two Arabic stemmers described in previous research papers.

The rest of this article is organized as follows: Section 2 presents the related work, Section 3 presents the methodology adopted in this study, Section 4 presents experiments conducted to demonstrate the validity of the proposed algorithm. Section 5 presents an analysis and a comparison between our stemmer and two known Arabic stemmers. Finally Section 5 presents conclusion and future work.

2. Related Work

Several research papers and projects are proposed developing Arabic stemmers (e.g. Al-Shalabi and Evens, 1998; Khoja and Garside, 1999; Abu-Salem et al., 1999). There are many studies that present examples of Arabic Stemming algorithms and their effectiveness. Most of these studies claim an accuracy which exceeds 85%. It is impossible to verify these claims due to the lack of source codes and the datasets which were used in the testing process.

Chen and Gey (2002) study is not purely dedicated to the construction of Arabic stemming, since it aims to study English-Arabic cross-language retrieval (CLIR). Therefore the paper constructed two Arabic stemmers beside an Arabic stop word list. They used a simple program which is restricted to removing major Arabic prefixes: The ('definite article' (Alif-laam), الـ), and four plural suffixes: (Alif-taa, "ات"), (Alif-nuun, "ان"), (Waaw-nuun, "ون") and (Taa, "ة"). Then they built two stemmers, the first one is called MT-based Arabic stemmer, which uses online Ajeeb machine translation system to translate Arabic words to English. These words are then partitioned into groups or clusters, where each group of Arabic words has a common English stem. Next, the MT-based Arabic stemmer selects the shortest word in the cluster and considers it as an Arabic stem for all the Arabic words in the cluster. The second Arabic stemmer is called light stemmer, where its main task is to remove the top frequently used Arabic prefixes and suffixes. In their study Larkey et al. (2002) constructed and tested a number of Arabic light stemmers. Their tests showed that the effectiveness of information retrieval systems (IRSs) which use the best light stemmers yield much better effectiveness than those that use morphological stemmers attempting to find the Arabic root. They also concluded that using the best light stemmer within an IRS is better than avoiding stemming or using co-occurrence analysis to produce stem classes or using very light stemmers. Many think that light stemming is much easier and more accurate than heavy (root-based) stemming, since light stemming is restricted to strip off predetermined Arabic affixes (prefixes and suffixes) from Arabic words. In reality, in many situations the Arabic affix could be part of the root (e.g., (Governor, "والي")). Therefore the light stemmer should decide whether to remove the affix if it is really an affix, or to keep the affix if it is part of the Arabic root. Nwesri et al. (2005) exhibited in their study three novel techniques to remove Arabic prefixes (i.e. Arabic prepositions and conjunctions) from Arabic words inputted to their light stemmers. Those are Arabic light stemmers which could not be benchmarked with our new root-based stemmer.

Most of the Arabic words are derived from triliteral Arabic roots. However, there are very few quadri-literal Arabic roots relative to the number of triliteral Arabic Roots. Kanaan et al. (2004) presented a novel stemming algorithm dedicated to Arabic words derived from quadri-literal Arabic roots only and used a limited set consisting of 145 Arabic words. Stemmer of Kanaan et al., 2004 yields 95% accuracy. Our study is completely different Kanaan et al., 2004 study in data size which is much larger, and their study is restricted to Arabic words derived from quadri-literal Arabic roots, while this one designed for Arabic words is derived from triliteral Arabic roots.