Stemming the Qur'an

Naglaa Thabet
School of English Literature, Language and Linguistics
University of Newcastle
Newcastle upon Tyne, UK, NE1 7RU
n.a.thabet@ncl.ac.uk

Abstract

In natural language, a stem is the morphological base of a word to which affixes can be attached to form derivatives. Stemming is a process of assigning morphological variants of words to equivalence classes such that each class corresponds to a single stem. Different stemmers have been developed for a wide range of languages and for a variety of purposes. Arabic, a highly inflected language with complex orthography, requires good stemming for effective text analysis. Preliminary investigation indicates that existing approaches to Arabic stemming fail to provide effective and accurate equivalence classes when applied to a text like the Qur'an written in Classical Arabic. Therefore, I propose a new stemming approach based on a light stemming technique that uses a transliterated version of the Qur'an in western script.

1 Introduction

Stemming has been widely used in several fields of natural language processing such as data mining, information retrieval, and multivariate analysis. Some applications of multivariate analysis of text involve the identification of lexical occurrences of word stems in a text. Such lexical analysis, in which the frequency of word occurrences is significant, cannot be done without some form of stemming.

In morphology, variants of words which have similar semantic interpretations are considered to belong to the same stem and to be equivalent for purposes of text analysis and information retrieval. For this reason, a number of stemming algorithms have been developed in an attempt to reduce such morphological variants of words to their common stem.

Various stemming algorithms for a number of languages have been proposed. The structure of these stemmers range from the simplest technique, such as removing suffixes, to a more complicated design which uses the morphological structure of words to derive a stem.

In case of Arabic, several stemming algorithms have been developed. The major inadequacy of existing systems to stem the Qur'an results from the fact that most of them deal with Modern Standard Arabic as their input text; the language of the Qur'an is Classical Arabic. Orthographic variations and the use of diacritics and glyphs in the representation of the language of Classical Arabic increase the difficulty of stemming. In many respects, the Qur'an, with its unique lexicon and orthography requires dedicated attention.

Therefore, I have developed a new light stemmer that uses the Qur'an in western transliteration to improve the effectiveness of the stemming of the text.

2 Stemming in Arabic

Arabic belongs to the Semitic family of languages, and as such differs from European languages morphologically, syntactically and semantically. The Arabic language is somewhat difficult to deal with due to its orthographic variations and its complex morphological structure. Xu et al. provide an overview of the challenges the Arabic language creates for information retrieval [10, 11].

2.1 Arabic Morphology

The grammatical system of the Arabic language is based on a root-and-affix structure and is considered as a root-based language. Most Arabic words are morphologically derived from a list of roots, to which many affixes can be attached to form surface words. Most of these roots are made up of three consonants which convey semantics. In addition to the different forms of the Arabic word that results from the derivational and inflectional process, most prepositions, conjunctions, pronouns, and possession forms are attached to the Arabic surface form.

2.2 Arabic Orthography

Orthographic variations are prevalent in Arabic. Vocalized texts make use of diacritics to represent short vowels. The omission of such diacritics in non-vocalized text gives rise to ambiguity, specifically if words are read out of context. Other spelling variations include changing the letter ي to

ى at the end of a word and replacing إ, أ, and آ with plain ا. A sense of discrimination and a good knowledge of grammar and usage are required if one is to avoid misreading a word.

In terms of multivariate analysis of text as well as information retrieval, the combination of a rich morphology and a pervasively ambiguous writing system results in a degree of complexity such that some sort of pre-processing and classification is required. Therefore, stemming is very important for Arabic text analysis.

2.3 Approaches to Arabic Stemming

Several stemming algorithms for Arabic have been proposed based on different principles; each produces rather different sets of stem classifications. It is possible to evaluate these stemming algorithms by the accuracy of the results they produce. Larkey et al. gives a good summary of stemming approaches for the Arabic language [9]. The most common approaches used in Arabic stemming are the light and the root-based stemmers.

Root-based Stemming is based on removing all attached prefixes and suffixes in an attempt to extract the root of a given Arabic surface word. Several morphological analyzers have been developed, e.g. Buckwalter [3], Khoja and Garside [7] and Darwish [5].

Light Stemming is used not to produce the linguistic root of a given Arabic surface form, but to remove the most frequent suffixes and prefixes. The most common suffixation includes duals and plurals for masculine and feminine, possessive forms, definite articles, and pronouns. Several light stemmers have been developed, all based on suffix and prefix removal and normalization. Examples of light stemmers include: Aljlayl & Frieder's Stemmer [2], Darwish's Al-Stem [6], Chen & Gey's TREC 2002 Stemmer [4], and Larkey et al.'s U Mass Stemmer [8, 9].

All light stemmers adhere to the same steps of normalization and stemming. The main difference among them is the number of prefixes and suffixes removed from each one. During the normalization process, all diacritics, punctuation, and glyphs are removed. The light stemmers had different stopword lists consisting of Arabic pronouns, particles and the like removed after minimal normalization. Test results of previous researchers as in [2, 8], proved that the light stemmer achieved superior performance over the root-based approach since it reduces sense ambiguity by grouping semantically related words into the same class.

Although light stemming can correctly classify many variants of words into large stem classes, it can fail to classify other forms that should go together. For example, broken plurals for nouns and adjectives do not get conflated with their singular forms, and past tense verbs do not get conflated with their present tense forms, because they retain some affixes and internal differences.

3 Stemming the Qur'an

My main objective for stemming the Qur'an is to prepare the text as data for multivariate analysis of the lexical semantics of the Qur'an using self-organizing maps in which words with similar meanings are placed at the same or neighbouring points so that the topological relations among them represent degrees of semantic similarity. This work requires the construction of vector space models of the suras (chapters) of the Qur'an such that each sura is represented by a vector indicating the occurrence frequency of variables. This involves counting the occurrences of lexical items in the Qur'an. Such a task cannot be done accurately without some sort of stemming of words in the text.

The Qur'an has two significant textual features. The first is that the Classical Arabic language in which the Qur'an is written has created difficulty in reading and understanding it, even for the Arabs themselves. Its lexicon, morphology and grammar are more complicated than Modern Standard Arabic. It, therefore, requires specific attention.

The second significant point is the wide use of vocalization. Diacritics (ْ, ّ, ٌ, ٍ, ُ, َ, ِ, ً, ٓ) representing short vowels are prevalent in the Qur'an. Every word, even every letter is marked with a diacritic. The meanings of the words in the Qur'an require the use of such diacritical marks; otherwise it becomes very difficult to comprehend their meanings especially when out of context.

Vocalized text, in Arabic includes diacritics for short vowels and other details. Thus, a word could have several meanings when marked with different diacritics. (see Table 1).

Word	Transliteration	Meaning
مُلْك	mulk	reign
مَلِك	malik	king
مَلَك	malak	angel
خُلُق	khuluq	morals
خَلْق	khalq	creation
اَمَة	amah	female slave
اُمَّة	ummah	nation

Table 1. Orthographic variations of words

For those reasons stemming the Qur'an is not an easy task. In principal, the way existing Arabic stemmers are structured indicates that they will not work reliably on the stemming of the Qur'an. Most of the existing stemmers rely on Modern Standard Arabic as their input script. This modern form of

Arabic is a simplified form of Classical Arabic. The main differences between both forms are that Modern Standard Arabic has less orthographic variation, a less complicated lexicon and a more modern vocabulary. The following two points are also significant regarding the use of existing stemmers to stem the Qur'an.

First, the root-based algorithm increases word ambiguity. The root algorithm stems the surface form to a base form from which the word variants are derived. A major problem with this type of stemmer is that many word variants are different in meaning, though they originate from one identical root. For example words like *hasib* (he thought), *hasaba* (he counted), and *hasab* (of noble origin) are all derived from the same root *hsb.* Therefore, the over-stemming of the root algorithm results in the deterioration of the retrieval performance as compared to the light stemming algorithm. As noted by Khoja [7], another problem that the stemmer faces is that some of the letters that appear to be affixes are in fact part of the word.

Second, the light stemmers perform better than the root-based algorithms, though not entirely efficiently. All initial steps of the light-based algorithms require normalization which involves the removal of diacritics. Thus, if diacritics were removed from the words listed in Table 1 above, there would be no other way to indicate the difference in meaning of all word variants. The normalization technique, though it appears simple, increases ambiguity. If normalization was applied to the Qur'an, it would leave the text highly ambiguous. As the case with root-based algorithms, some of the suffixes and prefixes to be removed using light stemmers are originally part of the word.

Therefore, I propose a new light stemming approach that gives better results, particularly when applied to a rich vocalized text as the Qur'an. The stemmer is basically a light stemmer to remove prefixes and suffixes and is applied to a version of the Qur'an transliterated into western script.

The use of the transliteration is highly significant for resolving the problem of diacritics in the Qur'an. Given that the transliteration of the Qur'an is available in western script, the problem of diacritics is resolved, since in the transliterated version of the Qur'an, each diacritic is translated into a letter in Roman script. Thus, the ambiguity that arises when removing the diacritics from the Arabic text is avoided. So, while the word ملك could have three different meanings when it appears without diacritics in Arabic, in transliteration each meaningful word has a single representation. (see Table 1).

Another advantage of using transliteration is avoiding the removal of suffixes and prefixes that sometimes could be part of the word. The prefix بـ (pronounced as "bi") is very common in Arabic. This preposition resembles the letter ب of the Arabic alphabet. Thus, removing this letter indistinguishably would cause ambiguity if the letter is part of a word. For example, in words as بَحر (sea), بُرهان (proof), the letter بـ is part of the word, whereas, in بِقلم (with a pen) the بـ is a preposition. If the diacritics that are marking the letter ب were removed, the first letter in each word would be exactly the same, though different in pronunciation. Therfore, stemming the words from the prefix بـ, in general, would be incorrect. When transliterating the same three words (بَحر, بُرهان, بِقلم) the prefix بـ would be represented as ba (bahr), bu (burhan), and bi (biqalam) respectively. The proposed light stemmer would only include "bi" as a prefix thus, avoiding removing the other representations of that letter. A few stems in Arabic begin with "bi"; those are added to a stopword list to be removed before stemming. The same process would be applied to the other prefixes to be removed such as (لـ, كـ, فـ, سـ, ال), (la/li, ka, fa, sa, al).

3.1 Implementation

The stemmer has been developed for the windows environment in Delphi, an object-oriented programming language which creates a graphical user interface to facilitate the presentation of its applications.

a. Preprocessing

Rather than the use of Arabic script, the system uses a Roman transliteration of the Qur'an which is formatted on the Web as HTML. This presents a particular problem that need to be remedied before the text can be stemmed. The problem is that some phonemically important distinctions, i.e., distinctions that are represented by different graphs in Arabic, are shown using HTML tags; when the HTML files are saved as text, these tags disappear, and the distinctions are lost. The Arabic phonemes (آ, ط, ح, ص, ض, ذ, ظ) are represented in the HTML transliteration files as underlined (a, t, h, s, d, th, *th*) respectively.

Preprocessing involves (1) stripping out the entire HTML markup, and (2) before doing so, replacing all the above phonemes with the following characters: a^, t^, h^, s^, d^, z^, z*. The result is a pure text file in ASCII codes.

b. Construction of stopword list

A stopword list of all the words to be excluded from the stemming process was compiled. The list was manually constructed using a concordance of the Qur'anic lexicon compiled by Abd Al-Baqi [1]. It consists of words which begin with the same letters which compose Arabic prefixes. Arabic pronouns, prepositions and names of people and places were also included in the stopword list.

c. Construction of stemmer

The algorithm for the stemmer is as follows:

Step 1. Prefix Stemming

The program reads individual suras from text files, replaces all uppercase letters with lower case letters and constructs a list of word lists, where each word list contains all the words in a single sura. It then reads single words from each word list and compares the current word supplied as a parameter to each successive word in the stopword list. If the word is found in the stopword list, it is excluded from prefix stemming; otherwise it adheres to following procedures:

- Remove prefixes (wa, fa, la, li, lil, bi, ka, sa, s^a, al)
- After stemming, the word is inserted back into the word list.

Step 2. Suffix Stemming

Six groups of suffixes are identified ranging from one-letter suffixes to six-letter suffixes. The system starts stemming the words in the word lists from the longest prefixes (six-letter prefixes) to the three-letter prefixes. Stemming the one and two-letter suffixes causes some ambiguity, since some of the suffixes could sometimes be part of the word stem. To resolve this problem, the stemmer sorts the words alphabetically. In the sorted list of words, if a given sequence displays a variety of suffixes including one and two-letter suffixes, the suffixes are removed and the stem is retained, otherwise the word is left intact.

3.2 Results

Preliminary results for seven long suras selected randomly and representing 6% of the Qur'an show that the stemmer achieves an accuracy of 99.6% for prefix stemming and 97% for suffix stemming. As the stemmer is being used, some inaccuracies were detected, but investigation shows that they are mainly to do with erroneous lexical items in the transliterated Qur'an. An evaluation of the system with accuracy figures should be available shortly for the entire Qur'anic text.

4 Conclusion

Stemming is important for a highly inflected language as Arabic. Existing Arabic stemmers, though produced effective results in some applications, failed to provide good stemming for the Qur'an. Therefore, I have proposed this new method of using transliterated script, which gave good preliminary results. Ongoing work on the system is focused on improving the accuracy of the results either by modifying the algorithms or editing the transliteration of the Qur'an.

References

[1] M.F. Abd Al-Baqi. 1987. *Al-Ma&jam Al-Mufahras li-alfaz Al-Qur'an Al-Karim.* Dar Al-hadith, Cairo.

[2] M. Aljlayl and O. Frieder. 2002. On Arabic Search: Improving the retrieval effectiveness via a light stemming approach. In *Proceedings of CIKM'02*, VA, USA.

[3] T. Buckwalter. 2002. *Buckwalter Arabic Morphological Analyzer Version 1.0*. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49.

[4] A. Chen and F. Gey. 2002. Building an Arabic stemmer for information retrieval. In *Proceedings of TREC 2002,* Gaithersburg, Maryland.

[5] K. Darwish. *An Arabic Morphological analyzer*. http://www.glue.umd.edu/~Kareem/research/

[6] K. Darwish and D. Oard. 2002. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English Retrieval. In *Proceedings of TREC 2002,* Gaithersburg, Maryland.

[7] S. Khoja and R. Garside. 1999. *Stemming Arabic text.* Computing Department, Lancaster University, Lancaster. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps

[8] L. S. Larkey and M. E. Connell. 2001. Arabic information retrieval at UMass. In *Proceedings of TREC 2001*, Gaithersburg: NIST, 2001.

[9] L. S. Larkey, L. Ballesteros and M.E.Connell. 2002. Improving stemming for Arabic information retrieval: Light Stemming and co-occurrence analysis. In *SIGIR 2002*, Tampere, Finland: ACM, 2002.

[10] J. Xu, A. Fraser and R. Weischedel. 2001. TREC 2001 cross-lingual retrieval at BBN. In *TREC 2001*, Gaithersburg: NIST, 2001.

[11] J. Xu, A. Fraser and R. Weischedel. 2002. Empirical studies in strategies for Arabic information retrieval. In *SIGIR 2002*, Tampere, Finland: ACM, 2002.