A Rule-Based Arabic Stemming Algorithm

TENGKU MOHD T. SEMBOK
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi 43600, MALAYSIA
tmts@ftsm.ukm.my

BELAL MUSTAFA ABU ATA
Department of Computer Science
Bahrain University, BAHRAIN

ZAINAB ABU BAKAR
Faculty of Computer and Mathematical Sciences
Universiti Teknology MARA
Shah Alam, MALAYSIA
zainabcs@salam.uitm.edu.my

Abstract: - Stemming is used in information retrieval systems to reduce variant word forms to common roots in order to improve retrieval effectiveness. As in other languages, there is a need for an effective stemming algorithm for the indexing and retrieval of Arabic documents. The Arabic stemming algorithm developed by Al-Omari is studied and new versions proposed to enhance its performance. The improvements relate to the order in which the dictionary is looked-up and the order in which the morphological rules are applied.

Key-Words: - Stemming, Information Retrieval, Arabic Morphology.

1 Introduction

One of the main modules of a document retrieval system is the text processing and indexing of the input documents to obtain the representation of the documents in the form of indexes. These indexes will be the surrogates to the documents and facilitate the process of retrieving relevant documents with respect to the given query. The process of selecting the representation or index terms constitutes a major operation and technique applied in information retrieval systems. Word stemming is one technique normally applied in the indexing process because it helps in reducing the size of the index terms and also proved to help in improving the degree of relevancy in retrieving documents (van Rijsbergen, 1979). The stemming process constitutes word morphological analysis based on the language used in order to get the words' stems to represent the documents as well as to function as indexes to the documents for efficient and effective retrieval.

Stemming is defined as the conflation of all variations of specific words to a single form called the root or stem. Stemming algorithms for some languages have been published and applied in building of information retrieval systems, among which for English is the famous Porter's algorithm [1], for French we have Savoy's algorithm [2], and for the Malay language we have Fatimah's et al. [3]. Stemming techniques play a vital part in the development of a good document retrieval system. The stemming process will reduce the size of the documents representations by 20-50% compared to full words representations, according to van Rijsbergen [4]. Furthermore, the relevancy of the retrieved documents will be improved and their number will also be increased.

Stemming algorithms for the Arabic language are not widely available and published in journals. The current algorithms reported are either general in nature, or lack in the morphological aspect of getting to the correct Arabic stems. Pioneer works on Arabic stemming have been published by researchers such as Gheith & El-Sadany[5], El-Sadany & Hashish[6], Saliba & Al-Dannan [7], Hilal [8], AlKharashi & Even [9], and Al-Omari [10].

2 Arabic Language Stemming Algorithms

In the previous sections, we mentioned some of the English, Malay and French language stemmers. Approaches adopted by these stemmers are not fully appropriate for the development of Arabic stemmers due to differences in the morphological structures peculiar to each of the languages as well as their semantic differences. The main differences as put forward by El-Sadany & Hashish [6] are as follows:

i. Arabic is one of Semitic languages which differ in structure of affixes from Indo-European type of languages such as English and French;
ii. Arabic is mainly roots and templates dependent in the formation of words;
iii. Arabic roots consonants might be changed or deleted during the morphological process;

Stemmers such as Porter's are mainly developed to improve the retrieval performance of document retrieval systems. As a result, these stemmers does not make use of dictionary that checks for the correctness of the resulted stem. Whereas, for languages such as Malay, French and Arabic, it will be somehow impossible to develop a stemming algorithm that does not make use of such dictionaries for stems and phrases checking. More precisely, if such stemmers are developed, their accuracy and performance will be low [9].

Approaches for the development of Arabic stemmers are restricted due to its complicated structure. These approaches are mainly dependent on the understanding of the Arabic morphology. Hence, Arabic stemming is actually a process of morphological analysis applied for the word in order to extract the correct stem. The stemming approach adopted by most of the previous Arab researchers for the development of morphological analysers is mainly an iterative of the following processes:

i. Analysing the prefixes
ii. Analysing the postfixes
iii. Analysing the stem
iv. Recoding the root
v. Lexicon verification

This approach was used by several researchers such as El-Sadany & Hashish[6], and Hilal [8], and Shahein & Youssef [11].

3 Arabic Word Formation

The grammatical system of the Arabic language is based on a root-and-pattern structure and considered as a root-based language with not more than 10,000 roots [12]. A Root in Arabic is the base verb form which can be trilateral, which is the overwhelming majority of Arabic words, and to a lesser extent, quadrilateral, pentaliteral, or hexaliteral, each of which generates additional verb forms and noun forms by the addition of derivational affixes [13].

A stem is a combination of a root and derivational morphemes to which an affix (or more) can be added [14]. However, when applying this definition to Arabic, the verb roots and their verb and noun derivatives are considered as stems. *Affixes* in Arabic are: prefixes, suffixes (or postfixes) and infixes (morphemes). Prefixes are attached at beginning of the words, where suffixes are attached at the end, and infixes are found in the middle of the words. For example, the Arabic word **الطالبات** (***altalibat***) which means "*female students*"**,** consists of the elements as shown in Table 1**:**

Table 1: Example of Arabic Affixes

Word	prefix	suffix	infix	root
الطالبات	*ال*	*ات*	*ا*	*طلب*

There are 8 prefixes in Arabic language which form a small set of prefixes compared to languages such as Malay and Slovene [3][15]. However, Arabic allows up to three concurrent prefixes to be added to the same word [16][17], such as the word وبالوالدين which contains three prefixes (و , ب, ال). Table 2 contains all the 8 Arabic prefixes and their meanings.

Table 2: Arabic Prefixes and Their Meanings

Prefix	Meaning	Example
ب	with, in, by	بالسيارة
ك	same as,	كالدخان
س	will,	سأذهب
و	and	ورجالهم
ال	the	النساء

Prefix	Meaning	Example
أ	questioning	أأكلت
ف	then	فذهبوا
ل	to, because	لتنام

Arabic prefixes does not follow a systematic standard for their attachment to Arabic words. In order to find all the rules that cover prefixes, rules of their combination and letters of words they are allowed to precede, a thorough and extensive study and analysis of the Arabic words and roots are needed.

4 Arabic Suffixes

There are 15 suffixes in Arabic language which form a small set of suffixes compared to languages such as Malay and the Slovene. However, Arabic allows up to three concurrent suffixes to be attached at the end of the same word, for example, the word ضربناهم contains three prefixes (هم , ا, ن). Arabic suffixes are mostly made of attachable pronouns. Table 3 contains all the 15 Arabic suffixes and their English meanings [16][17].

Table 3: Arabic Suffixes and Their Meanings

Prefix	Attached word	Example
ين	singular female	تلعبين
ان	male dual	يلعبان
و	plural male,	ينمو
ه	missing singular	ضربته
ك	addresser singular	ضربك
ا	male dual	أكلا
ي	singular female	أكلتي
ن	plural	أكلن
ت	singular female	أكلت
ات	female plural	لاعبات
ون	absent male plural	يلعبون
وا	absent male plural	أكلوا
تم	addresser male plural	أكلتم
هم	absent male plural	ضربهم
كم	addresser male plural	ضربكم

The characteristic of Arabic suffixes is similar to that of prefixes, which does not have a systematic rule for their attachment to Arabic words.

5 Rule-based Stemming Algorithm

The stemming algorithm was implemented using Standard C language with Arabic language support. The stemming algorithm consists of the following main modules:

- Prefix and suffix removal module;
- Root generator and checking module;
- Pattern generator and checking module;
- Intensification module which handle double letters (تشديد);
- Recoding module.

5.1 Prefix and Suffix Removal Module

This module will try to find all the valid affixes in a given Arabic word and remove such affixes. Arabic language contains just a few number of affixes, however, affixes attachment rules to words are not easy to list out. After some thorough study of the Arabic morphological structure and word formation, we came out with around 800 rules that cover both Arabic prefixes and suffixes attachment rules to words. The prefix and suffix rules are define accordaing to the following syntax:

1. Prefix rules: **prefix** + ***let(s)***
where *let(s)* is a set of valid letters to follow the prefix,
example: أ + **تأ**
which means أ *is considered a prefix if the next two letters are* تأ *such as in the word* أتأتي

2. Suffix rules: ***let(s)*** + **Suffix**
where let(s) is a set of valid letters preceding the suffix,
example: **ببـ + ت**
which means ت *is considered a suffix if the previous 2 letters are* ببـ *such as in the word* أحببت

Table 4 shows example of prefixes and suffixes in the given words.

Table 4: Examples of Word Letters that match the Arabic Affixes

Word	Letter(s)	Type of Matched Affix
فارس	ف	*Prefix*
لاعبون	ل	*Prefix*
بارد	ب	*Prefix*
بنات	ت	*Suffix*
متم	تم	*Suffix*
القرون	ون	*Suffix*

5.2 Root Generator and Checking Module

This module will try to find all the valid possible roots for a given word. The module will check for the root validity by using the hashing technique to search for it in the root dictionary. This module will invoke the *Intensification submodule* that checks for words of double letters in order to change it to the normal form.

5.3 Pattern Generator and Checking Module

This module will take the word to be stemmed and one possible root (generated by the root module) as an input and then derive a template that match both of them. This process will be repeated for all the possible root generated from the root generator module. The module will also check the resulted template for its correctness by matching it to a set of valid Arabic templates. An example of this is as follows:

- for the word فاسقين, some of the possible roots generated are فسق, فاس, سقي, قين, فاق , where the roots قين, فاق are ignored as they are not valid Arabic roots.

- the templates for the remaining 3 roots are constructed with reference to the word فاسقين, the resulting templates and their validity in Arabic are shown in Table 5.

Table 5: Possible Templates for the Word فاسقين

Root	Generated Template	Template Validity
فسق	فاعلين	*valid*
فاس	فعلقين	*invalid*
سقي	فافعلن	*invalid*

5.4 Intensification Module Handling Double Letters

There are many Arabic words and roots with *double* letters, which means that two similar adjacent letters are combined into one letter. This module will check for such words and its root and reconstruct the word by adding that letter. This will help in obtaining the correct root. Examples of words with *Intensification* are shown in Table 6.

Table 6: Examples of Arabic words with *Intensification*

Root	Generated Template	Template Validity
فسق	فاعلين	*valid*
فاس	فعلقين	*invalid*
سقي	فافعلن	*invalid*

5.5 Recoding Module

The *recoding module* main concern is to change some of the letters to their correct form. These letter changes will most probably occurs during the process of template formation in Arabic when a word is formed from a root. Some letter may dropped, changed or replaced by other letters. Table 7 lists some of the most recoded Arabic letters.

Table 7: Examples of Letter Recoding for Arabic Words

Word	Recoding Rule (from→ to)	Word after Recoding
هزئ	ئ → ؤ	هزؤ
	ئ → أ	هزأ
نبئ	ئ → أ	نبأ
خطئ	ئ → أ	خطأ
خسئ	ئ → أ	خسأ
صبئ	ئ → أ	صبأ
سيئ	ئ → أ	سيأ
نبء	ء → أ	نبأ
دنى	ى → ا	دنا
تؤمن	ؤ → أ	تأمن
يؤمر	ؤ → أ	يأمر
يؤخذ	ؤ → أ	يأخذ
ؤمر	ؤ → أ	أمر
راد	ا → و	رود
حيا	ا → ي	حيي

The stemming algorithm will take as input an Arabic word (not a stop word), and the output will be the extracted root (or stem). In cases where the algorithm cannot find a root for the a specific word, the word itself will be taken as a root. Such cases are few based on the performance of the algorithm

5.6 Flowchart of Stemming Algorithm

The flow chart of the stemming algorithm is shown in Figure 1. The stemming process begins by processing a word and trying to find its correct stem.

In case the word does have a correct stem, then the word without its affixes will be returned.

Figure 1: The steps to find the stem for the word: فسيأكلون:

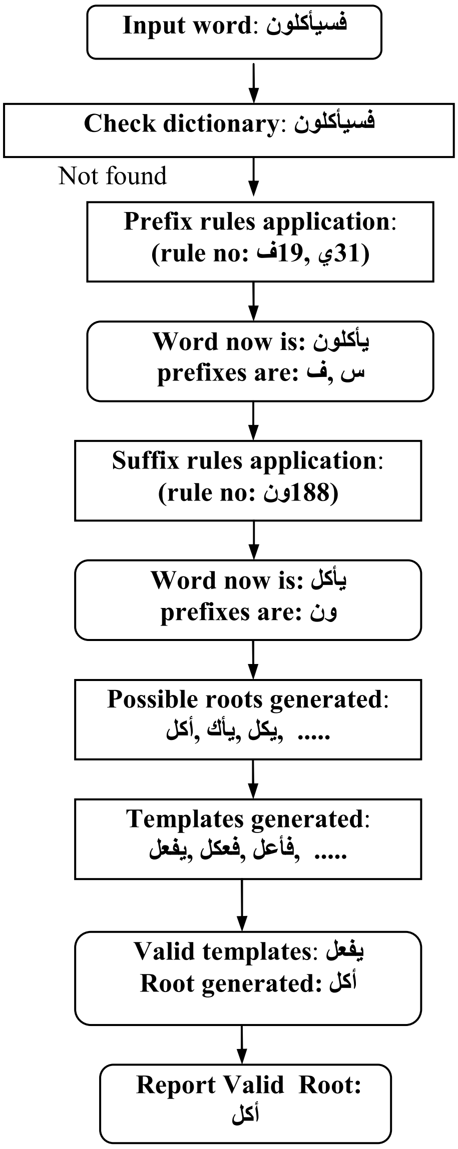
Input word: فسيأكلون
Check dictionary: فسيأكلون
Not found
Prefix rules application: (rule no: ف19, ي31)
Word now is: يأكلون
prefixes are: ف, س
Suffix rules application: (rule no: ون188)
Word now is: يأكل
prefixes are: ون
Possible roots generated: أكل, يأك, يكل, ...
Templates generated: فأعل, فعكل, يفعل, ...
Valid templates: يفعل
Root generated: أكل
Report Valid Root: أكل

Table 8 shows the number of errors obtained by our stemming algorithm compared to the results obtained by Al-Omari's algorithm. Hence, we can conclude that the our algorithm does performance better than Al-Omari's algorithm.

Table 9 shows all the 21 words that has been stemmed wrongly and the types of errors for each word.

Table 8: Results of the Experiments for 10 Chapters of the Quran

		Number of wrongly stemmed words	
		Ours	**Al-Omari's**
Chapter 1	684+	9	25
	126~	6	15
Chapter 2	450+	8	23
	83~	6	17
Chapter 3	444+	6	24
	100~	4	20
Chapter 4	462+	9	27
	103~	5	19
Chapter 5	202+	6	15
	56~	3	12
Chapter 6	278+	5	19
	48~	2	17
Chapter 7	341+	9	29
	73~	4	19
Chapter 8	299+	5	12
	73~	4	10
Chapter 9	341+	4	14
	88~	3	9
Chapter 10	181+	4	7
	46~	2	4
	3682+	**65**	**195**
Totals	**796~**	**39**	**142**
	330*	**21**	**85**

Keys:
+ Total number of all words in the chapter
~ Total number of unique words in the chapter
*Total number of unique words in all the chapters

Table 9: Stemming Errors on Ten Chapters of the Quran

Word	Actual Root	Resulting Root	Error Type
ربه	ربب	ربه	unchanged
موتها	موت	موة	spelling
الظانين	ظنن	ظان	spelling
الرياح	ريح	راح	spelling
بالباطل	بطل	اطل	spelling
وبارك	برك	ارك	spelling
فويل	ويل	يل	overstemming
الفلك	فلك	لك	overstemming
ليبلو	بلو	بل	overstemming
بآل	بآل	آل	overstemming
بأس	بأس	أس	overstemming
بوالديه	ولد	ديه	others
ووقاهم	وقى	قا	others
تنزيل	نزل	زيل	others
كرها	كره	رها	others
والفسوق	فسق	سوق	others
المبين	بين	مبي	others
فتبينوا	بين	تبي	others

Word	Actual Root	Resulting Root	Error Type
مبين	بين	مبي	others
بالهم	بآل	هم	others
فأتنا	أتي	تنا	others

Table 10: Distribution of Unique Errors on Quranic Data Set

Error Type	Number (%)
Overstemming	5 (23.8%)
Understemming	0 (0 %)
Unchanged	1 (4.7%)
Spelling	5 (23.8%)
Others	10(47.6%)

There are a total of 21 unique errors as shown in Table 10. These errors are classified into 5 groups, namely, *understemming, over-stemming, spelling, unchanged*, and *others*. The names of the groups describe the type of errors. Understemming and overstemming indicate that the resultant stems are uder stemmed or over stemmed. The group *spelling* indicates there is one letter in the resulted stem that is different from the correct root. As for unchanged group indicates that the resultant stem is the same as the original word which is not the correct root. Others indicate other types of stemming errors.

6 Conclusions

Our experiments have shown that our new stemming algorithm performs better than that of Al-Omari. Could it be improved further? Our analysis suggests that most of the remaining errors are due to the precise order in which the rules are applied, and we are currently considering ways in which this ordering can be best applied.

References

[1] Porter, M. F. (1980). An algorithm for suffix stripping'. *Program, 14*, 130-137.

[2] Savoy, J. (1993). 'Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science,44*,1-9.

[3] Ahmad, F., Mohammed Yusoff, Sembok, T.M.T. 1996. Experiments with A Malay Stemming Algorithm, *Journal of American Society of Information Science*.

[4] van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths.

[5] Gheith, M.& El-Sadany, T. 1987. Arabic Morphological Analyzer on A Personal Computer. *Arabic Morphology Workshop*. Stanford University.

[6] El-Sadany, T.A. & Hashish, M.A. 1988. Semi-Automatic Vowelization of Arabic Verbs. *Proceedings of 10th National Computer Conference:* 45-56.

[7] Saliba, B., & Al-Dannan, A. 1990. Automatic Morphological Analysis of Arabic: A study of Content Word Analysis. *Proceedings of the First Kuwait Computer Conference*: 231-243.

[8] Hilal, Y. 1990. Automatic Processing of Arabic Language and Applications. *Proceedings of the Arabic Language Processing Using Computer Conference*: 213-219.

[9] Al-Kharashi, I.A. & Evens, M.W. 1994. Comparing words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System. *Journal of the American Society for Information Science*. **45**(8): 548-560.

[10] Al-Omari, H. 1994. *ALMAS: An Arabic Language Morphological Analyzer System.* National University of Malaysia. Bangi, Selangor.

[11] Shahein, H. I. & Youssef, S.A. 1990. A Model for Morphology As A Production System. *Proceedings of the Second Cambridge Conference on Bilingual Computing in Arabic and English.*

[12] Ali, A.Y. 1983. *The Holy Quran: Text, Translation and Commentary*. Maryland: Amana Corp.

[13] Saliba, B., & Al-Dannan, A. 1990. Automatic Morphological Analysis of Arabic: A study of Content Word Analysis. *Proceedings of the First Kuwait Computer Conference*: 231-243.

[14] Gleason, H.A. 1970. *An Introduction to Descriptive Linguistics*. 3rd Edition. New York: Holt. Rinehart and Winston.

[15] Popovic. M. 1991. *Implementations of A Slovene Language-Based Free-Text Retrieval System*. University of Sheffield. UK.

[16] Al-Fedaghi, S.S., & Al-Sadoun, H.B. 1990. Morphological Compression of Arabic Text. *Information Processing & Management*. **26**(2): 303-316.

[17] Dahdah, A. 1985. *Arabic Language Grammar Dictionary. “معجم قواعد اللغة العربية“*.Beirut: Lebanon. Lebanon Library.