ARABIC WORDS STEMMING APPROACH USING ARABIC WORDNET

Abdel Hamid Kreaa, Ahmad S Ahmad and Kassem Kabalan

College of Information Engineering, Tishreen University, Latakia, Syria

ABSTRACT

The big growth of the Arabic internet content in the last years has raised up the need for an effective stemming techniques for Arabic language. Arabic stemming algorithms can be ranked, according to three category, as root-based approach (ex. Khoja); stem-based approach (ex. Larkey); and statistical approach (ex. N-Garm). However, no stemming of this language is perfect: The existing stemmers have a low efficiency. In this paper, we introduce a new stemming technique for Arabic words that also solve the problem of the plural form of irregular nouns in Arabic language, which called broken plural. The proposed stem extractor provides very accurate results in comparisons with other algorithms. Consequently the search effectiveness improved.

KEYWORDS

Ontology, Stemming, Arabic WordNet

1. INTRODUCTION

The main goal behind building any stemmer is to improve the search effectiveness so an IR system can match user's queries with relevant documents. Users form their query terms in many different formats. However they are looking for the same thing [1]. Now an IR system should be able to translate all these forms that have the same meaning to a standard form. Thus grouping all these different formats in a singular or standard format on both users queries and index terms sides.

Recently, the big growth of the Arabic internet content has raised up the need to an effective stemming techniques for Arabic language [2]. Arabic stemming algorithms can be ranked, according to three category, as root-based approach (ex. Khoja); stem-based approach (ex. Larkey); and statistical approach (ex. N-Garm). Although many stemming methods have been developed for Arabic language, they suffer from many problems. In this paper, we introduce a new stemming technique for Arabic words that also solve the problem of the plural form of irregular nouns in Arabic language, which called broken plural. The proposed stem extractor provides very accurate results in comparisons with other algorithms. Consequently the search effectiveness improved. The remainder of this paper is organized as follows: Firstly, Section 2 provides the background information regarding Arabic language and Arabic WordNet. Then popular related works of Arabic language stemming methods are discussed in Section 3. In Section 4 a novel stemming Algorithm based on Arabic WordNet is proposed. A comparison with other methods results is presented in Section 5. Conclusions are presented in Section 6.

DOI : 10.5121/ijdkp.2014.4601

2. THE RESEARCH METHOD

2.1. Arabic Language Characteristics

Arabic is a Semitic language which differs from Indo-European languages syntactically, morphologically and semantically. The writing system of Arabic has twenty five consonants and three long vowels that are written from right to left and take different shapes according to their position in the word. In addition, Arabic has short vowels, which are not part of the alphabet, they are written as vowel diacritics above or under a consonant to give it its desired sound, hence they give a word a desired meaning. Texts without vowels are considered more appropriate by the Arabic-speaking community since they are widely used in everyday written and printed materials (books, magazines, newspapers, letters, etc.). However, in Holy Koran, printed collections of classical poetry, school books and some Arabic paper dictionaries, vowel diacritics appear in full. Usually, well-edited books, some printed texts, and manuscripts vowel diacritics partially or randomly are written out where words could be ambiguous or difficult to read [3].

Arabic is a very rich and complex language and the morphological representation of Arabic is rather complex. Arabic language is based on set of roots, Therefore all nouns and verbs are generated from a set of roots. This set contains 11,347 roots distributed as follow [4]:

- 115: Two character roots (and these roots have no derivations from them).
- 7198: Three character roots.
- 3739: Four character roots.
- 295: Five character roots.

These roots join with various vowel patterns to form simple nouns and verbs to which affixes can be attached for more complicated derivations. Patterns play an important role in Arabic lexicography and morphology. For example, the root “لعب” corresponds to the pattern “فعل”, where other letters can be added to form another pattern. For example :the pattern “مفعل” form the word “ملعب” by adding the letter “ م” to the morpheme “لعب” [5].

Dual or plural pattern are composed by adding suffixes like “ان“،”ين” for dual and “ون“،”ين” ،”ات” for plural. There are irregular forms of plural patterns called broken plural. In this case, a noun in plural takes another morphological form different from its initial form in singular, (Table 1, shows singular and plural patterns), In some cases a plural pattern may have more than one singular patterns as shown in table1. For example the word “ أطباء” has a singular “طبيب” which is like “فعيل” but the word “عقلاء” has a singular “عاقل” which is like “فاعل”.

Table 1. Singular and plural patterns.

Plural Pattern	Singular Pattern
مفاعل	مفعل
مفاعيل	مفعول
أفعال	فعل
فعلاء	فعل, فعال, فاعل, فعيل
فعال	فاعل
أفعل	فعل
أفعلة	فعيل, فعال
فواعل	فوعل, فاعل

Some of the broken plural forms like “فعائل” ،“فعالل” ،“فعلة” has many singular patterns. There are no general rules to cover them all, and it is hard to deal with them without having a complete dictionary for these plurals[5].

2.2. Arabic Wordnet

Arabic WordNet (AWN) is based on the design and contents of the Princeton WordNet (PWN) [3]. The success of the PWN for English has motivated similar projects that aim to develop wordnets for other languages. That project aimed to develop a linguistic resource with a deep formal semantic foundation in order to capture the richness of Arabic as described in [6]. The Arabic WordNet has been built following methods that developed Euro WordNet. In addition, word meanings are defined with a machine understandable semantics in first order logic. The basis for this semantics is the Suggested Upper Merged Ontology (SUMO) and its associated domain ontologies, SUMO is being enlarged to provide a formal semantic foundation for AWN.

The AWN project provides a deep semantic underpinning for each concept. The approach that was previously used in mapping all of PWN to a formal ontology is considered (Figure 1).

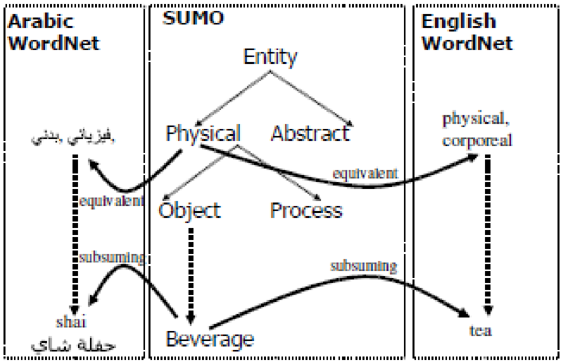


Figure 1: SUMO mapping to wordnets

AWN is not designed to take any Arabic text as an input directly. However we have to transliterate convert the Arabic system text to be fed to as an input to the system. The results must be transliterated converted back into Arabic to be understood. This technique was introduced by Buckwalter (2002) [7], for example the letter "و" is transliterate to "w". The letter "ؤ" is transliterate to "&".

(Table 2, shows some Arabic letters and their transliteration in Buckwalter).

Table 2. Arabic letters representation in Buckwalter.

الشدة	~	ألف ممدودة	|
السكون	o	ألف الوصل	{
حرف الباء	b	ألف مقصورة	Y
حرف العين	E	الفتحة	a
حرف الفاء	f	الضمة	u
		الكسرة	i

The AWN database is freely and publicly available.

3. RELATED WORKS

Stemming is the process for reducing inflected words to their stem or root form (generally a written word form). The stem does not need to be identical to the morphological root of the word. Usually, it is sufficient that related words is mapped to the same stem, even if this stem is not a valid root itself. For example the root of the Arabic word ("المعلمون", the teachers) is ("علم", science). The stem is simply defined as a word without a prefix or/and suffix. For example, the stem of the Arabic word (“المعلمون”, the teachers) is (“معلم”, teacher). Arabic stemming algorithms are classified under three categories [8]:

- Root-Based Approach (Khoja Stemmer).
- Stem-Based Approach (Larkey Light Stemmer).
- Statistical Approach (often involves N-Gram).

In this section, a brief review on stemming approaches for stemming Arabic Text is presented.

3.1.Khoja Stemmer

In 1999 Khoja and Garside produced an effective root-extracting stemmer. (Khoja Stemmer)[9]:

The Khoja stemmer follows this procedure:

1. Remove diacritics representing vowelization.

2. Remove stopwords, punctuation, and numbers.

3. Remove definite article “ال”

4. Remove inseparable conjunction “و”.

5. Remove suffixes.

6. Remove prefixes.

7. Match result against a list of patterns. If a match is found, then extract the characters in the pattern representing the root.

8. Match the extracted root against a list of known "valid" roots.

9. Replace weak letters “ا,و,ي”with “و”

10. Replace all occurrences of hamza “ؤ,ئ,ء”with “أ”

11. Two letter roots are checked to see if they should contain a double character. If so, the character is added to the root.

Figure 2 shows the Flowchart of Khoja Stemmer [9]:

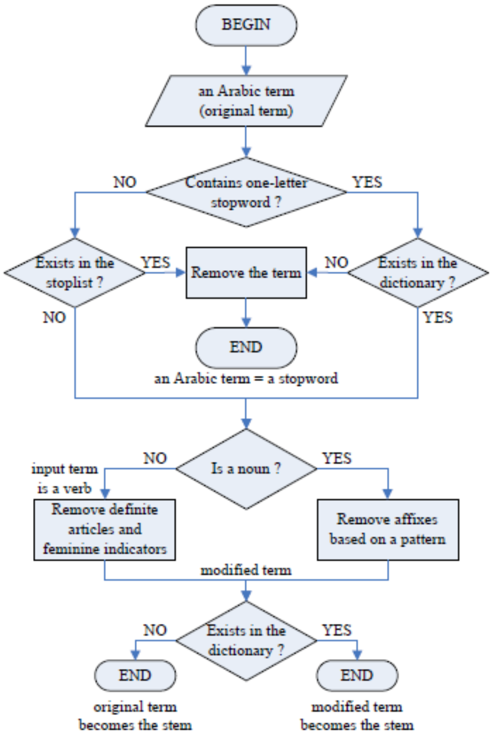

BEGIN
an Arabic term
(original term)
Contains one-letter stopword ?
NO
YES
Exists in the stoplist ?
YES
Remove the term
NO
Exists in the dictionary ?
NO
YES
END
an Arabic term = a stopword
input term is a verb
NO
Is a noun ?
YES
Remove definite articles and feminine indicators
Remove affixes based on a pattern
modified term
NO
Exists in the dictionary ?
YES
END
END
original term becomes the stem
modified term becomes the stem

Figure 2 Flowchart of Khoja Stemmer

3.2.Light Stemmer

In 2002 Larkey et al proposed the Light stemmer. The purpose of the Stem-Based Approach or Light Stemmer is to remove the most frequent suffixes and prefixes. However, this algorithm changes the form of the words in some cases [10].

The light stemmer, light10, strips off initial “و” (and), definite articles:

(”لل“,”وال“،”ال“،”كال“، ”فال“، ”بال“)

and suffixes:

(”ها“,”ان“،”ات“،”ون“،”ين“،”يه “،”ية“،”ه“،”ة“،”ي“)

It was designed to strip off strings that were frequently found as prefixes or suffixes at the beginning or ending of stems. It was not intended to be exhaustive.

Table 3: Strings removed by light stemming

	Remove prefixes	Remove Suffixes
Light1	ال، وال، بال، كال، فال	none
Light2	ال، وال، بال، كال، فال، و	none
Light3	“	ه، ة
Light8	“	ها، ان، ات، ون، ين، يه، ية، ه، ة، ي
Light10	ال، وال، بال، كال، فال، لل، و	“

There are several versions of light stemming, all of them follow the same steps [10]:

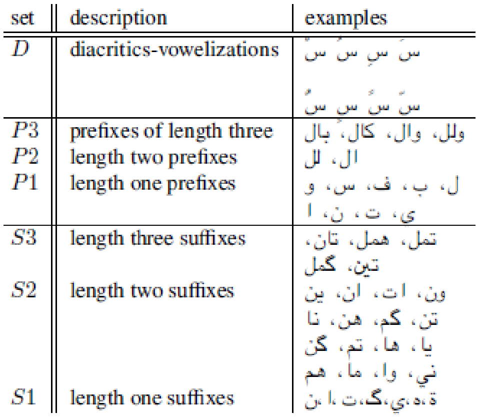
1. Remove "و" from lgiht2, light3, and light8 And light10 if the remainder of the word is 3 or more characters long
2. Remove any definite article if this leaves 2 or more characters
3. Go through the list of suffix once in right to left order (Table 3)

3.3.ISRI Stemmer

In 2005 Kazem T et al proposed an algorithm for Arabic stemming shares many features with the Khoja stemmer [11]. The Information Science Research Institute's (ISRI) stemmer [11] uses a similar approach to word rooting as the Khoja stemmer, but does not employ a root dictionary for lookup. Additionally, if a word cannot be rooted, the ISRI stemmer normalizes the word and returns a normalized form (for example, removing certain determinants and end patterns) instead of leaving the word unchanged.

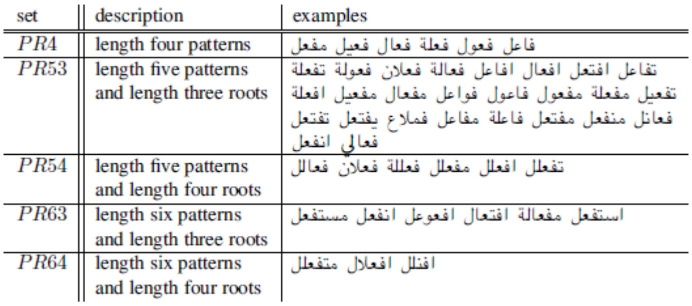
It defines sets of diacritical marks and affix classes as shown in Table 4. These are sets of marks are removed by the stemmer. In addition, it defines some pattern sets as shown in Table 5.

Table 4. Affix Sets



set	description	examples
D	diacritics-vowelizations	سَ سِ سُ سْ سّ سً سٍ سٌ
$P3$	prefixes of length three	ولل، وال، كال، بال
$P2$	length two prefixes	ال، لل
$P1$	length one prefixes	ل، ب، ف، س، و ي، ت، ن، ا
$S3$	length three suffixes	تمل، همل، تان، تين، كمل
$S2$	length two suffixes	ون، ات، ان، ين تن، كم، هن، نا يا، ها، تم، كن ني، وا، ما، هم
$S1$	length one suffixes	ة،ه،ي،ك،ت،ا،ن

Table 5. Arabic Patterns and Roots



set	description	examples
$PR4$	length four patterns	فاعل فعول فعلة فعال فعيل مفعل
$PR53$	length five patterns and length three roots	تفاعل افتعل افعال افاعل فعالة فعلان فعولة تفعلة تفعيل مفعلة مفعول فاعول فواعل مفعال مفعيل افعلة فعائل منفعل مفتعل فاعلة مفاعل فملاع يفتعل تفتعل فعالى انفعل
$PR54$	length five patterns and length four roots	تفعلل افعلل مفعلل فعللة فعلان فعالل
$PR63$	length six patterns and length three roots	استفعل مفعالة افتعال افعوعل انفعل مستفعل
$PR64$	length six patterns and length four roots	افعلال افعلال متفعلل

Stemming proceeds in the following steps:

1- Remove diacritics representing vowels.

2- Normalize the *hamza* which appears in several distinct forms in combination with various letters to one form "أ".

3- Remove length three and length two prefixes respectively.

4- Remove connector “و” if it precedes a word beginning with “و”.

5. Normalize “آ,أ,إ”to “ا”.

6. Return stem if less than or equal to three.

7. Consider four cases depending on the length of the word:

a) Length = 4: If the word matches one of the patterns from PR4 (Table 5), extract the relevant stem and return it. Otherwise, attempt to remove length-one suffixes and prefixes from S1 and P1 in that order to get a word not less than three letters in length.

b) Length = 5: Extract stems with three characters for words that match patterns from PR53. If none are matched, attempt to remove suffixes and prefixes. Otherwise the relevant length-three stem is returned. If the word is still five characters in length, the word is matched against PR54 to determine if it contains any stems of length 4. The relevant stem is returned if found.

c) Length = 6. Extract stems of length three if the word matches a pattern from PR63. Otherwise, attempt to remove suffixes. If a suffix is removed and a resulting term is of length five letters, send the word back through step 7b. Otherwise, attempt to remove one character prefixes. If successful, then send the resulting length-five term to step 7b.

d) Length = 7. Attempt to remove one-character suffixes and prefixes. If it is successful, send the resulting length-six term to step 7c.

The ISRI stemmer has been shown to give good improvements to language tasks such as document clustering, as opposed to a non-stemmed approach

4.PROPOSED STEMMING ALGORITHM

To overcome the weaknesses in the nowadays Arabic Stemming techniques and consequently improves the search effectiveness, we present a hybrid method which incorporates two different techniques: Light Stemmer and Look in Tables.

The proposed algorithm removes the affixes letter by letter (not group of letters at a time) depending on diacritical marks and affix sets (as shown in table 6)in a try to overcome Over Stemming problem.

The proposed algorithm uses Arabic WordNet tables to verify the correctness of the extracted stems and to solve the problem of the broken plural stem.

Our proposed Stemming algorithm (AKK Stemmer) proceeds in the following steps:

1. Remove any character that is not listed in the group L of Table 6.

2. Remove diacritics of the words listed in the group D of Table 6.

3. Remove prefixes listed in group P of Table 6 one by one until the word length reaches 3 and add the resultant words to the word list WL1.

4. For each word in the WL1 remove suffixes listed in the group S of Table 6 one by one until the word length reaches 3 and add the resultant words to the word list WL1.

Table 6: Arabic Patterns and Affix Sets

Set	Description
D	علامات التشكيلَ ◌ِ◌ُ◌ً◌ٍ◌ٌ◌ْ
L	ا آ أ إ ب ت ث ش س ي ل ص ق ف غ ع ه خ ح ض ط ز ة و ؤ ر ء ئ ذ ط م ن ك
P	ك ل ا و س ب ي ن م ت ف
S	ن ك ه ة ت ا ي و

5. Compares the resulting word list WL1 words with the values of the column "Word_AR" in the Arabic WordNet "Word" Table, there are two cases:
   a) There is a match, add the matched words to the word list WL2.
   b) If no match exists then compares the resulting word list WL1 words against the values of the column "Broken_Plural" in the Arabic WordNet "Form" Table, if there is a match, we look for the value of the column "Root" in the Arabic WordNet "Word" Table, and the value of the column "Word_AR" is added to the word list WL2.

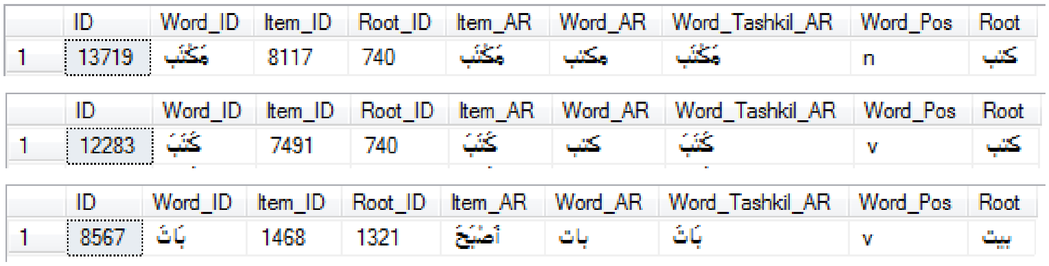
The longest word(s) in the WL2 is the relevant stem.

Table 7 shows the WL1 result from stemming process for the Arabic word “المكتبات”. Figure 3 illustrates the records from "Word" Table matches the words in WL1, and table 8 shows the WL2 as a result to the matching process”**.** The **relevant stem** is the word “**مكتب.**

Table 9, figure 4, figure 5, figure 6 and table 10 are the results for the stemming process of the word “**المبالغات**”.

Table 7. Word list WL1 resulting from the stemming of the word “المكتبات”.

المكتب	المكتبا	بات	تبات	كتبات	مكتبات	لمكتبات	المكتبات



	ID	Word_ID	Item_ID	Root_ID	Item_AR	Word_AR	Word_Tashkil_AR	Word_Pos	Root
1	13719	مَكْتَب	8117	740	مَكْتَب	مكتب	مَكْتَب	n	كتب

	ID	Word_ID	Item_ID	Root_ID	Item_AR	Word_AR	Word_Tashkil_AR	Word_Pos	Root
1	12283	كَتَبَ	7491	740	كَتَبَ	كتب	كَتَبَ	v	كتب

	ID	Word_ID	Item_ID	Root_ID	Item_AR	Word_AR	Word_Tashkil_AR	Word_Pos	Root
1	8567	بَاتَ	1468	1321	أَصْبَحَ	بات	بَاتَ	v	بيت

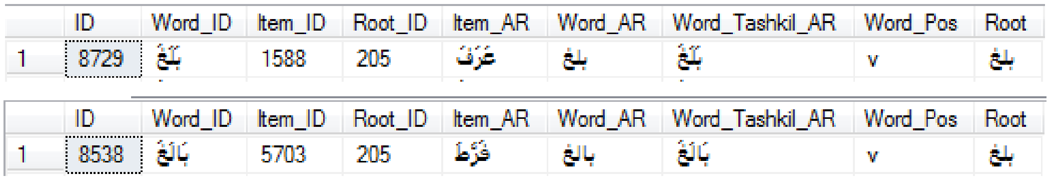
Figure 2. Records from "Word" Table matches the words in WL1 resulting from the stemming of the word “المكتبات”.

Table 8. Word list WL2 resulting from the stemming of the word “المكتبات”.

كتب	بات	مكتب

Table 9. Word list WL1resulting from the stemming of the word “المبالغات”.

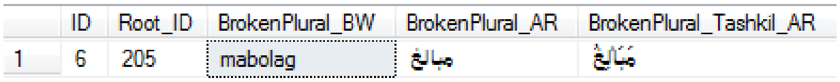
لمبالغا	المبالغ	المبالغا	غات	لغات	الغات	بالغات	مبالغات	لمبالغات	المبالغات
				لغا	الغ	الغا	بالغ	بالغا	مبالغا



	ID	Word_ID	Item_ID	Root_ID	Item_AR	Word_AR	Word_Tashkil_AR	Word_Pos	Root
1	8729	بَلَّغ	1588	205	عَرَّفَ	بلغ	بَلَّغ	v	بلغ

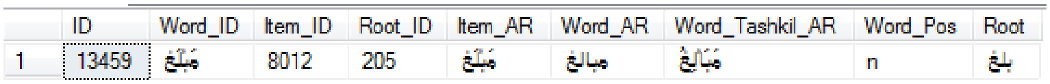
	ID	Word_ID	Item_ID	Root_ID	Item_AR	Word_AR	Word_Tashkil_AR	Word_Pos	Root
1	8538	بَالَغ	5703	205	فَرَّطَ	بالغ	بَالَغ	v	بلغ

Figure 3. Records from "Word" Table matches the words in WL1 resulting from the stemming of the word “المبالغات”.



	ID	Root_ID	BrokenPlural_BW	BrokenPlural_AR	BrokenPlural_Tashkil_AR
1	6	205	mabolag	مبالغ	مَبَالِغ

Figure 4. Records from "Form" Table matches the words in WL1 resulting from the stemming of the word “المبالغات”.



	ID	Word_ID	Item_ID	Root_ID	Item_AR	Word_AR	Word_Tashkil_AR	Word_Pos	Root
1	13459	مَبْلَغ	8012	205	مَبْلَغ	مبالغ	مَبَالِغ	n	بلغ

Figure 5. Records from "Word" Table matches the word from "Form" Table.

Table 10. Word list WL2 resulting from the stemming of the word “المبالغات "

مبلغ	بلغ	بالغ

The **relevant stems** are the words “مبلغ” and “بالغ”

4.1.Algorithm description

Let *T* denote the set of characters of the Input Arabic word

Let *Stem* denote the term after stemming each step

Let *L* denote the set of characters of the Arabic Alphabet

Let *D* denote the set of diacritics

Let *S* denote the set of suffixes

Let *P* denote the set of prefixes

Let *wL1* denote the array of words after stemming in each step

Let *wL2* denote the array of words after searching in search step

```
Step 1: Remove any diacritic in $T$, $Stem = T - D_i$

Step 2: Remove any character not in $L$, $Stem = Stem – L_i$

Step 3: $wL1 = wL1 + Stem$

Step 4: If the length of $Stem$ is greater than 3 characters then,
                $Stem = Stem – S_i$
                $wL1 = wL1 + Stem$

Step 5: For all words in $wL1$,
        If the length of words in $wL1$ is greater than 3 characters then,
                $Stem = Stem – P_i$
  $wL1 = wL1 + Stem$

Step 6: For all words in $wL1$,
        Locate for in AWN.Word Table
        If matched
                $wL2 = wL2$ + AWN.Word.Word_AR

Step 7: For all words in $wL1$,
        Locate for in AWN.Form Table
        If matched
                Locate for in AWN.Word Table
                $wL2 = wL2$ + AWN.Word.Word_AR

Step 8: For all words in $wL2$,
        Locate for the longest word, $Stem$ = Longest word

Step 9: Return the $Stem$
```

Figure 6 shows the Flowchart of AKK Stemmer.

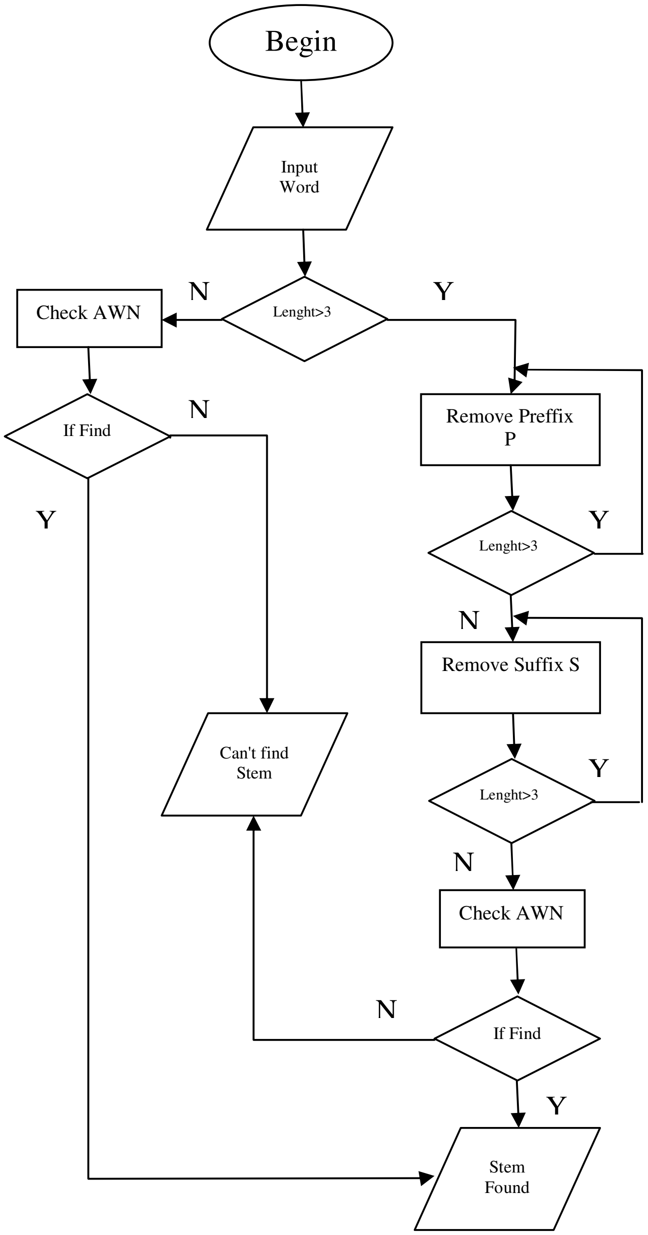


Figure 6 Flowchart of AKK Stemmer

5. COMPARISON AND EVALUATION

We have done two experiments in order to evaluate performance of AKK stemmer. In the first experiment, a sample terms list which used in [5,8] is considered and tested with the most famous stemming methods (Light stemmer, Khoja stemmer, ISRI stemme) in addition to the AKK stemmer. The results are shown in Table 11.

Table 11. A comparison between stemmers.

Term	AKK Stemmer	ISRI	Light Stemmer	Khoja
المكتبات	**مكتب**	كتب	مكتب	كبي
منظمات	**منظم**	نظم	منظم	ظمأ
كامل	**كامل**	كمل	امل	كمل
يبحثون	**بحث**	بحث	يبحث	بحث
ركبتيه	**ركب**	ركبتي	ركبتي	ركبتيه
تستغرق	**غرق**	ستغرق	تستغرق	تستغرق
فلتستعجله	**عجل**	فلتستعجله	فلتستعجل	فلتستعجله
متفاهمون	**فهم**	متفاهم	فاهم	فهم
بسطاء	**بسيط**	بسطا	بسطا	بسط
مبالغات	**مبالغ– مبلغ**	بلغ	مبالغ	بلغ
بخلاء	**بخيل**	بخل	بخلا	بخل

These results reveals that the light stemmer fails to get the correct stem of the word many times. In many cases it produced a completely new word and sometimes it created a wrong word that doesn't exist in Arabic language. Furthermore, it didn't handle the broken plural forms. The Khoja stemmer ,as disclosed, produces a very general words (roots) that are far in their meaning from the original word, while our suggested stemmer succeed (in many cases) to get the expected stems and removed all affixes effectively. Moreover it didn't remove the affixes as they are part of the original word and fairly handles the broken plural forms and generates the correct singular forms.

In the second experiment we evaluated our stemmer in terms of its effectiveness for document retrieval. We applied four stemmers to the Arabic Trec 2001 collection.[12]

Retrieval effectiveness is measured by the usual precision and recall formulas. Recall is the number of relevant documents retrieved divided by the total number of relevant documents. Precision is the number of relevant documents retrieved at a given point divided by the total number of retrieved documents. We compared three approaches to stemming Arabic words, Khoja, ISRI, and Light, with our approach. The results are shown in Table 10.

Table 10. Average Precision for stemmers.

Stemmer	Less than 5 word Queries	5-12 word Queries	More than 12 word Queries
Akk	0.492	0.438	0.304
Light	0.474	0.410	0.291
Khoja	0.463	0.445	0.280
ISRI	0.480	0.424	0.282

Table 10 seems to indicate that the Akk stemmer performs better than the other approaches.

6. DISCUSSION AND CONCLUSION

Generally the purpose of stemming process is to find out the representative indexing form of a word, Arabic as a highly inflected language requires a good stemming process to make information retrieval effective. Currently, there is no standard approach for stemming, Nevertheless, as mentioned earlier, there are two general methods was used. Either by extracting the root of the word like Khoja stemmer, or by just truncation of affixes like Light stemmer. The first method have many problems. Firstly, the root dictionary requires maintenance to guarantee that the newly discovered words are stemmed correctly. Secondly, in some cases it fails to remove the affixes of the word, thus it fails to extract the root. For example the Khoja stemmer will fail to remove affixes in the words "تستغرق"and "ركبتيه" so it will not stem them where they are respectively derived from the roots "غرق"and "ركب" [5]. Thirdly, the most important problem is that root extraction stemmers are not useful for Arabic language from IR system point of view as we stated before, in many cases the resultant word is the root that is very general and thus leading to a poor search effectiveness. For example the word"محامون" will be reduced to the root "حمى". This root is very general and there are a huge number of words that can be derived from this root, therefore, reducing all the forms that can be generated from that root to its basic form will result in a general index terms which has a serious effect on the quality of the results.

The second method is more useful from IR system considering Arabic language. However, the available techniques have two major problems, the first one is in many cases these stemmers truncate a sequence of characters that matches one of the affixes but it is actually part of the original word and this will lead to a completely new word for example Larkey stemmer will truncate the word "ولدين"to "لد" which has no meaning in Arabic [11]. The second problem is that they are not dealing with the plural form of irregular nouns in Arabic language. Therefore in many cases it fails to group words that have the same meaning in one reduced form. In our AKK stemmer, we consider the second method and introduced a new algorithm as described before. This algorithm uses a set of rules to determine if a certain sequence of characters is part of the original word or not. This helped us solving some ambiguity problems. Furthermore, we introduced a way to handle the majority of broken plural forms and to reduce them into their singular form. This method is useful in grouping words of the same meaning in a common form.

REFERENCES

[1] Abdusalam N., Seyed T., and Falk S., "Stemming Arabic Conjunctions and Prepositions," in Proceedings of the 12th international conference on String Processing and Information Retrieval.

[2] Aitao C., "Building an Arabic Stemmer for Information Retrieval," in Proceedings of the Eleventh Text Retrieval Conference, Berkeley, pp. 631-639, 2003.

[3] Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C., (2006). Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

[4] Marwan B., Arabic Language Processing in Information Systems, Springer, 2004.

[5] M Ababneh1, R Al-Shalabi, G Kanaan, and A Al-Nobani " Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness" The International Arab Journal of Information Technology, Vol. 9, No. 4, July 2012.

[6] Elkateb, S. "Design and implementation of an English Arabic dictionary/editor", PhD thesis, Manchester University (2005).

[7] Buckwalter, T (2002). Arabic transliteration. http://www.qamus.org/transliteration.html.

[8] M Hadni, S Ouatik and A Lachkar "EFFECTIVE ARABIC STEMMER BASED HYBRID APPROACH FOR ARABIC TEXT CATEGORIZATION" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013.

[9] Purwitasari D, Najibullah A, Zainal Arifin A "Modification of Khoja Stemmer for Searching Arabic Text "Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Indonesia 2005.

[10] Larkey L., L. Ballesteros, and M. E. Connell."Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis". Proceedings of SIGIR'02. PP 275–282. 2002..

[11] Kazem T., Rania E., and Je.rey C., Arabic Stemming Without A Root Dictionary, Information Science Research Institute, USA, 2005.

[12] David Graff and Kevin Walker. Arabic newswire part 1,. http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T55. 2001.