Developing an Effective Light Stemmer for Arabic Language Information Retrieval

Sohair Al hakeem*, Ghazi Shakah, Belal abu Saleh
Computer Science Department
Ajloun National Univesity
Ajloun, Jordan
*Email: drsohair [AT] gmail.com

Nisreen Jaber Thalji
Computer Science Department
Hail University
Hail, KSA

Abstract—**Arabic language is one of the top 10 Most Spoken Languages in the World. It belongs to Semitic group of languages. Technology has been slow in development for Arabic due to morphological and structural complexity in the language. Arabic language requires good stemming for effective information retrieval. Many light stemmers have been developed but still suffer weaknesses and high percentage of errors. No standard approach has been emerged yet. In this paper a new effective light stemmer algorithm has been developed overcoming many limitations of previous approaches. The new technique taking into account truncate the word infixes in addition to prefixes and suffixes based on simple rules. Proposed stemming method was found to supersede the other stemming ones. It has been tested and compared with root-based stemmers developed by Khoja [11]. Correctness, strength and similarity of both stemming algorithms are reported.**

Keywords-: Arabic stemmer; Stemming Algorithm; Light stemmer; Information Retrieval

I. INTRODUCTION

Due to the complicated morphological structure embedded in the Arabic language, text processing is hard to perform compared with other languages. Text processing is the main step shared among Information Retrieval (IR), text mining, natural language processing and many other applications. The efforts to improve Arabic information search and retrieval processes compared to other languages are limited and modest, thus there is an urgent need for effective Arabic information search and retrieval tools [4]. Arabic word stemming has been a central topic of many researchers in Arabic (IR). Stemmers are basic elements in query systems, indexing, web search engines and information Retrieval systems (IRs). The two most successful approaches to Arabic stemming have been a root-extraction stemmer developed by Khoja [11] and the light affix removing stemmer developed by Larkey [14], [6]. Larkey has shown that the Khoja and Light stemmers, as well as co-occurrence analysis-aided stemmers, perform information retrieval tasks with statistically equivalent precision [6]. Light stemmer has proved effective for the task of IR. Yet no complete stemmer for the Arabic language is available. In this paper a new stemmer has been developed. The presented stemmer does not need to use dictionary and gives far better results than the existing stemmers. It has been compared with Khojas and showed convenient results and less errors. In the context of this paper a stem does not necessary map the word to its root. A stem is the shortest form of a word among syntactically related words in a document. A root is the original form of a word that cannot be analyzed [3].

Because stemming reduces the vocabulary size by reducing variant words to a single stem, we saw it being equivalent to clustering syntactically related. words. For example, the root of the Arabic word (المعلمون , the teachers) is (علم, science). While a stem is simply defined as a word without a prefix or/and suffix. For example, the stem of the Arabic word (معلمون, teachers) is (معلم, teacher). The words ”loves” and ”loving” are syntactically related, as are the words ”act” and ”acting”. In this case, both ”loves” and ”act” are stems. Although the word ”loves” is not the root, it is the stem for our document.

The proposed stemmer removes the suffixes and prefixes based on a set of rules. In addition, it is introducing a set of rules to remove infixes for the resulting words.

II. ARABIC MORPHOLOGY

Most Arabic words are morphologically derived from a list of roots. The majority of these roots are bare verbs form made up of three consonants. Letters are added at the beginning, middle or end of the root to derive different patterns of words. These patterns generate nouns and verbs. There are about 11,347 roots distributed as follows [12]:

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

Affixes in Arabic are: Prefixes, suffixes (or postfixes) and infixes (morphemes) [1]. Prefixes are attached at the beginning of the words, suffixes are attached at the end, and Infixes are found in the middle of the words. For example, the Arabic word (alkatebat) which means the female writers, consists of the elements as shown in Table I.

TABLE 1. ARABIC WORD ORIGINAL LETTERS AND AFFIXES

Word	Root	Prefix	Suffix	Infix

الكاتبات	كتب	ال	ات	ا
The female writers	A male has written	The	Feminine indicator	Noun formation

III. Related Work

Stemming Arabic documents was performed manually prior to TREC (Text Retrieval Conference) and only applied on small corpora. Later, many researchers both native and non-native Arabic speakers created a considerable amount of Arabic stemming Algorithms [15]. Arabic stemming algorithms can be classified, according to the desirable level of analysis: root-based approach [11] and stem-based approach [6]. Root-Based approach uses morphological analysis to find the root of a given Arabic word. Many algorithms have been proposed for this approach [2], [12] and [10]. Root-based indexing is aggressive in the sense that it reduces words to their three-letters roots. This affects the semantics as several words with different meanings might have the same root [4].With light stemming, words are reduced to their stems. Light stemming removes frequently used prefixes and suffixes in Arabic words. Light stemming does not produce the root and therefore might fail to group all the semantically related words with a common syntactical form.

A. Khoja Root-based Stemmer:-

Among the successful approaches for Arabic stemming, a root-based stemmer has been developed by [11]. Based on predefined root lists and morphological analysis, Khojas algorithm attempts to extract the true root. However, more than one root can be found in an isolated word without diacritics [7]. The Khoja stemmer needs constant maintenance to track newly added words [4]. The main weaknesses of this approach as mentioned in [13] and [8] are:-

- The root dictionary requires an update to ensure that the new terms detected are correctly stemmed.
- The Khoja Stemmer replaces a weak letter with (و) which produces incorrect root. For example,

the word (عامة) which mean (general with feminine indicator (ة) is stemmed to (عوم) which means (float) instead of (عام) which means (general without the feminine indicator.

- Third, the Khoja Stemmer fails to remove all of affixes.

B. Larkey Light Stemmer:-

Larkey [10] created a group of light stemmers including Light 1,2,3,8, and 10, the latest light 10 stemmer is shown to outperform the previous light stemmer versions. Larkey and Connell [6] showed that his light 8 stemmer outperforms the Khoja stemmer. Darwish [9] presented a modified light stemmer "Al-stem" with an extended prefixes and suffixes lists. In a monolingual IR environment, the light 10 had a higher average precision in comparison to the Al-stem [9]. Both Arabic root-based and stem-based algorithms suffer from generating stemming errors. Unfortunately, the unguided removal of a fixed set of prefixes and suffixes causes many stemming errors, especially where it is hard to distinguish between an extra letter and a root letter [4].

IV. METHOLODOGY

A new stemming algorithm is presented in this paper to overcome the weaknesses in the existing stemmers. A set of suffixes, prefixes and infixes have been identified shown in table II, III and IV. Removing infixes was introduced in this algorithm unlike other light stemmers. It has been mentioned in [16] that one of the weak points of existing light stemmers is that they have not dealt with infixes such as Al-Shalabi [12] and larkey [6].

TABLE I. PREFIXES

Length6	والأست والإست والاست والست
Length5	وبالا وبالإ وبالأ والمو الأست الإست الاست وبأست وبإست وباست الست
Length5	ولإل ولأل بأست وبال باست والت وكال بإست والم المت ولال وللت وللإوللأ والإ والأ والا فكال فبال وللا وكال واست وإست وأست واست
Length3	كال فلي فال است إست أست بال للت للإ للا للأ الإ الأ الا الم لإل لأل لال الت وأل وسأ فلن فلأ فلإ فلا ولن ولت ولي ولإ ولأ ولا وسن وسا وسي فلل فلإ فلأ فلا ولل فسأ فسا فسن فسي وست فست فلت
Length2	فن لي لت ان فت سي لأ وم وي وت لن با سن فا كا ست سي لل ال وا بم فت كم لم ون ول فك ست بت فس تت لي يت فب في وك
Length1	ا ك ف ل ب ن م ت ي

TABLE II. SUFFIXES

Length7	تموينها
Length6	تموهما تماهما
Length5	تمونا تماهم انهما تموها تماني تيكما ونهما ناكما اتكما تاكما تماها تموني تهما تموهن انكما تمانا تماهن تموهم ناهما تيهما ونكما ينهما تهما تاهما
Length4	تاكم اتنا ناهم ونكن يننني انكم انها ونكم تيها تننا تاكن يكما الت ناها والت وننا ينهن ينها تموه تاني ناكم تها تيهن اننا اتكم تنهن وكما تانا ناكن تنني تيهم انني اتكن تاهم اتهم ونها نكماكما ينهم انهن تيكن ونهم تماه و همانا هن وتني تينا انهم اتها انكن ونهن يننانا هماتاهاتنهم نهما يهما اهماتيكم تكما يون اتهن
Length3	تيه ونك تما اتك انه تاك يات تنا تين ننا يكم ونا يكن يون انا انكا تها اها تاه ننا ونه تهم نهن وكن يهااكم نكن نهم نهن وني نكم كما اتى تني نها تيك يناانه اني تان تنه نني ينه يهن وهم تاي وكم تهن تكن اهم ناك وها تكم اهن ناه يهم وهن
Length2	هم ان ني تم ين ون تن اء يك اك تك نك كم وك اه ية ته ات تا تي ناكن وه ها أي يه
Length1	ي ة ه

Table III. INFEXES

Length1	ت ي و ا

Choosing the suffixes, prefixes and infixes sets were based on grammatical functions of the affixes. In addition to their occurrence frequencies among the Arabic words found in the Arabic document collection. A set of stop words (or functional words) has been identified. Stop words do not carry a particular and useful meaning for IR. The algorithm is set to remove the stop words during the first phase to reduce the size of the corpus. Some of these stop words are found in Table V.

TABLE IV. STOP WORDS SAMPLE

Stop words	**Meaning**

في	In
الى	To
على	On

No dictionary was needed to get to the right stem. This is one of the advantages of the proposed algorithm over Khojas algorithm. It saves space and time to return the required stem. Unlike other stemmer Al-Shalabi [12] and larkey [6], matching the words against a certain Arabic pattern has not been used.

The process of developing the proposed stemmer passes through different phases:

• Remove all stop words

• Check the length of the word. Return all words with three characters or less without truncating them. Removing any character from a three characters Arabic word will cause a huge ambiguity.

• If the word is more than three characters long the algorithm will truncates a word at the two ends. Making sure that the returned word is at least three characters long. Starting with removing the longest Prefix and check the length of the word. After removing the longest Prefix, if the word contains more than three characters the longest suffix will be truncated. For example the word "الاستشارات" (consulting) is shown in table VI. The decision of starting with Prefixes then suffixes was based on trial and statistics when running the algorithm separately on both cases. After truncating the words the output will be saved in a data file.

• Starting with the saved data file. Check for the words with four or more characters long to remove infixes. In case more than one infix characters were found in the word a priority will be given to removing ي, و, ا then ت in order. Priority was given based on the most common characters occur as an additional character and not as an original character of the word. For example the word (القادمون ,Arrivals) will be processed as shown in table VII.

TABLE V. ARABIC WORD ORIGINAL LETTERS AND AFFIXES

Word	Longest prefix	Longest suffix	Returned word
الاستشارات	الاست	ات	شار

TABLE VI. ARABIC WORD INFIXES REMOVAL IN ADDITION TO SUFFEXES AND PREFIXES

Word	Longest prefix	Longest suffix	Remaining word	Remove infix	Returned word
القادمون	ال	ون	قادم	ا	قدم

V. EVALUATION AND COMPARISON

A good stemmer is defined as one which merges as many as possible pair of words that are different in form, but are semantically equivalent; and avoid merge as many as possible pair of words that are different in form and are semantically distinct.

Different criteria are used to evaluate the performance of the proposed stemmer. The system performance evaluation was based on two testing manners; the former manner focused on measuring the number of acceptable (meaningful) words as an output of applying the stemming algorithms on each test group. The second manner is based on Paice's [17] evaluation methodology to evaluate the stemmer efficiency.

For the first testing scheme, A test has been carried out executing both the proposed and Khoja algorithms. Khojas algorithm is available to download from the following website (http://zeus.cs.pacificu.edu/shereen/research.htm). Both algorithms have been tested on a randomly selected collection of Arabic documents used in the

Khaleej-2004 corpus. Khaleej-2004 corpus is a collection of different topics Arabic documents as shown in the following table VIII.

TBLE VII. KHALEEJ2004 DOCUMENTS

Topic	Corpus Size (Number of documents)
International News	953
Local News	2398
Economy	909
Sports	1430
Total number of docs	5690

Each text document belongs to one of four categories (International News 953, Local News 2398,

Economy 909, and Sports 1430) out of 5690 documents. The corpus is available publically at (https://sites.google.com/site/mouradab) The test was carried out on different topics samples. Each test document sample contains about 1000 words. The test document has been saved in MSOffice excel sheet file after removing the stop words. The actual roots were manually extracted for the test documents words to compare results from different stemming systems. Roots extracted have been checked by Arabic Language scholars who are experts in the Arabic Language. The generated results are shown in Fig. 1

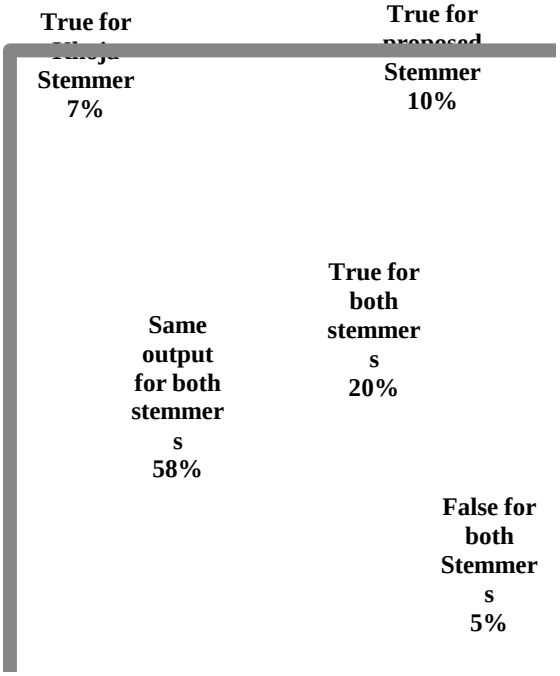

Figure 1: Both Algorithms outputs after execution in percentages

Fig. 1 shows that the proposed stemmer shares about 58 percent with Khoja stemmers output. For comparison purposes, the error percentage was calculated ignoring the same output data. Khoja stemmer gave a 10 percent of errors over the proposed algorithm with 7 percent of errors. This proves that the proposed stemmer gave less incorrect output words in comparison to Khoja's stemmer. Table IX shows sample output words of each stemmer.

TABLE VIII. SAMPLE OUTPUT WORDS OF EACH STEMMER

Word	Khoja stemmer	Proposed stemmer
الامة	لوم	امة
الانتاجية	توج	نتج
البنك	بني	بنك
المالي	ملأ	مال
الحواسيب	حواسيب	حسب
المحور	محر	محور
ايجاد	جود	يجد
اموال	مول	أمل
بعضها	بعض	عضه
بالنشر	نشر	نشر

The correct stem is underlined. The double underlined words are the words that are morphologically and syntactically related to the original input word but not the exact stem.

Unfortunately, stemming can cause errors known as over-stemming and under-stemming, or false- positive and false-negatives respectively. Over-stemming and under-stemming are stemming errors that usually weaken the accuracy of stemming algorithms [5].

Over-stemming occurs when two words with different stems are stemmed to the same root. Merging together the words "probe" and "probable" after stemming would constitute an over-stemming error.

Under-stemming occurs when two words that should be stemmed to the same root are not. For example, if the words "adhere" and "adhesion"are not stemmed to the same root. Paice [17] has introduced the stemming weight (SW) as an indicator to the stemmer efficiency. Stemming weight is the ratio between the under-stemming (UI) errors and the over-stemming (OI) errors.

For the second testing scheme, a sample of 442 words (W) was selected and manually divided into

a 100 conceptual groups. A concept group contains forms which are both semantically and morphologically related one to another see table X.

TABLE IX. CONCEPT GROUPS SAMPLE. INCORRECT STEM IS UNDERLINED

Groups	Khoja stemmer	Proposed stemmer
البنت	بني	بنت
البنات	بني	بنا
بنتان	بني	بنتن
البنتين	بني	بنت
البنك	بني	بنك
البنكين	بني	بنك
البنكان	بني	بنك
البنوك	بني	بنك

For each group g containing ng words, the number of pairs of different words defines the desired merged total (DM T_g):

$$DMT_g = 0.5n_g\ (n_g - 1) \quad (1)$$

Since a perfect stemmer should not merge any member of a group with other group words, for every

group there is a desired non-merge total (DN T_g):

$$DN\ T_g = 0.5n_g\ (W - n_g) \quad (2)$$

When summing these two totals over all groups, the global desired merged total (GDMT) and the global desired non-merge total (GDNT) have been obtained respectively. Thus, stemming errors are calculated as follows:

Conflation Index (CI): proportion of equivalent word pairs which were successfully grouped to the same stem; Distinctness Index (DI): proportion of non-equivalent word pairs which remained distinct after stemming The under-stemming index (UI) and the over-stemming index (OI) are given by "Eq. (3) "and "Eq. (4)"

$$U\ I = (1 - C\ I) \quad (3)$$

$$OI = (1 - DI) \quad (4)$$

The stemming weight (SW) is then given by " Eq.(5)"

SW = OI /U I (5)

The results are listed in Table XI below.

TABLE X. UNDER-STEMMING AND OVER-STEMMING ERRORS FOR BOTH STEMMERS

	UI	OI	SW
Khoja stemmer	0.238095	0.071429	0.3
The proposed stemmer	0.428571	0.026786	0.0625

Although, the proposed stemmer was not able to outperform the Khojas stemmer in producing less under stemming errors, The SW rate for the proposed stemmer was lower than Khojas. This clearly shows the strength of the proposed stemmer.

Another weakness of Khoja algorithm is the need to process a large dictionary, and Arabic word patterns list during runtime which can result in extra requirements for storage space and processing time.

On the other side the proposed algorithm does not have a root dictionary or Arabic word patterns list.

VI. CONCLUSION

Stemming has a large effect on Arabic information retrieval, at least in part due to the highly inflected nature of the language.

In this paper A new Arabic language stemmer algorithm has been proposed. The new approach has been evaluated successfully on the Arabic language. It has been compared to Khoja stemming algorithm using two different methods of evaluation. The number of acceptable (meaningful) words has been measured as an output after applying the stemmer algorithms on each test group. In addition to stemming weight introduced by Paice [17].

We observed that the results of the tests show the proposed algorithm has less incorrect output stemmed words compared to Khoja's algorithm.

The proposed stemmer has a tendency to generate more under stemming error than the Khojas algorithm, and the Khojas algorithm has a tendency to generate more over stemming errors than the proposed algorithm. In general, it can be seen that it is a conflicting task to try reducing the two types of error. The proposed algorithm showed a promising future for the stemming approach, which encourage the research into less under stemming errors. Future research could investigate the stemmers on an information retrieval system to assess its impact over recall and precision.

REFERENCES

[1] Abu Ata B., Mohd T. Sembok. "Arabic word stemming algorithms and retrieval effectiveness,". Proceeding of the World Congress on Engineering, 3:978–988, 2013.

[2] Al-Fedaghi S. and Al-Anzi F. "A new algorithm to generate arabic root-pattern forms," In proceedings of the 11th national Computer Conference and Exhibition, pages 391–400, March 1989.

[3] Al Khaleel M. and George. , A dictionary of arabic syntax terms, Library of Lebanon, 1990.

[4] Al-Shammari,E. "Improving Arabic Text Via Stemming with Application to Text Mining and Web Retrieval," PhD thesis, George Mason University, 2010.

[5] Baeza-Yates R.A., "Text-retrieval: Theory and practice," . In 12th IFIP World computer Congress, Elsevier Science , volume 1, pages 465–476., September 1992.

[6] Ballesteros L. , Larkey L and Connell M., "Improving stemming for arabic information retrieval: Light stemming and co-occurrence analysis," SIGIR, page 269-274, 2002.

[7] Brahmi A., and Ech-Cherif A., " Arabic texts analysis for topic modeling evaluation," Springer Science+Business Media, pages 33–53, June 2011.

[8] Coombs J., Taghva K., and Elkhoury R. "Arabic stemming without a root dictionary," Information Science Research Institute.

[9] Darwish K. and Oard D, " Evidence combination for arabic-english retrieval,". CLIR Experiments at Maryland for TREC-2002, 2002.

[10] Freund G. and Willett P." Online identification of word variants and arbitrary truncation searching using a string similarity measure,". Information Technology: Research and Development, (1):177–187, 1982.

[11] Garside A. Khoja S., " Stemming Arabic Text,"PhD thesis, Lancaster University, UK, 1999.

[12] Kanaan G., Al-Nobani A. , Ababneh M. and Al-Shalabi R., " Building an effective rule-based light stemmer for arabic language to improve search effectiveness,". International Arab Journal of Infromation Technology, 9(4):368–372, July 2012.

[13] Lachkar A., Hadni M., and Ouatik S.," Effective arabic stemmer based hybrid approach for arabic text categorization, "International Journal of Data Mining and Knowledge Management Process (IJDKP), 3(4), 2013.

[14] Leah S. Larkey and Connell M., "Arabic inforamation retrieval". In Proceedings of TREC 10, 2002.

[15] Lin J. and Al-Shammari E.," Towards and error-free arabic stemming," Proceedings of the 2nd ACM workshop on Improving non english web searching iNEWS '08 10 2008.

[16] Otair M., "Comparative analysis of arabic stemming algorithms," IJMIT, 5(2):1–12, May 2013.

[17] Paice C.D.," Method for evaluation of stemming algorithms based on error counting," Journal of the American Society for Information Science, 47:632–649, 1996.