International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184, Volume 8, Number 2 (April 2019), pp.17-23
© MEACSE Publications
http://www.meacse.org/ijcar

Arabic Text Light Stemmer

Jaffar Atwan, Mohammad Wedyan and Hadeel Al-Zoubi
Prince Abdullah Ben Ghazi Faculty of Information and Communication Technology
Al-Balqa Applied University, Al-Salt, Jordan

Abstract

Stemming defines as transferring the modified words to their origin instead of their current status. In respect of languages that are extremely modified such as Arabic, stemming plays an important role in enhancing the restored performance through decreasing the words alternatives. The present paper aims at clarifying the efficiency of light stemmer in restoring the Arabic data. Estimating the light stemmer is implemented by employing TFIDF since it considers as a prominent weighting scheme in line with Linguistic Data Consortium (LDC) Arabic Newswire data collection is compared to the primary system without stemming. The suggested light stemmer has to be applied in order to obtain the utmost performance.

Keywords*:* Arabic, Information Retrieval, Stemming, Light Stemmer, Arabic Text.

Introduction

The obstacles that Arabic language poses are countless particularly in the research field due to its outstanding properties like short vowels, the inexistence of capital letters, and its complicated morphology [1]. Indeed, the abovementioned properties complicating the process of data Retrieval in general and Arabic Information Retrieval (AIR). Various methods have been presented in order to promote the AIR. In general abovementioned methods, namely stemming and lemmatization, query emendation, and query expansion play an important role in both restoring information and language flow.

Stemming concept is related to a combination method that seeks to search for a mutual stem for a number of terms that emerge in a text. However, Table 1 clarifies that. Such method denotes that a single stem is used for a number of words that are common in having a specific shape that might be existed without no need to have a right morphological origin.

Table 1.0 shows that several terms are emanated from the same root Shhedشهد

Arabic Word	English Meaning	Arabic Root	English Root
مشهد	Scene	**شهد**	Shhed
شاهد	Witness	**شهد**	Shhed
مشاهدون	Audience	**شهد**	Shhed

In fact, Arabic language has improved several ways for stemming. Although they contain various errors, but they are employed in various IR systems. However, such stemmers are divided into two groups as follows: first, root derivation stemmer such as the stemmer provided by [2]. The second one represents the light stemmers resemble the stemmers that are presented in [3].

The core of this research presents the suggested light stemmer for Arabic text. Moreover, it shows that algorithm does not resemble the aggressiveness of the root extraction algorithm. We reveal that the suggested light stemming algorithm stimulates the restoration surpasses the light10 stemmer algorithm in some inquires. The adopted inquiry is TREC-2001 Arabic group of data and in order to restore the related documents [4].

Other parts of the research are ordered as follows. For instance, the first part sheds the light on the relevant work while the second focuses on describing the experiment. Part 4 discusses the finding. In conclusion, part 5occurs at the end of the research.

Related work

First of all, stemming methods can be classified into three general categories, such as lookup-based, rule-based and probabilistic [5]. The methods of the first category look for the term stem in the text in a lookup table that involves a group of words and their origins. The certain stem is retrieved if the search is successful. Although such method achieves extremely precise findings, but it still suffers from major shortcomings, such as the demand for linguistics specialists, a shaping process of a labour-intensive list, and the system difficulty [6].

The second category defines as a ‘light stemmer’ that utilizes a group of principles that are unintentionally utilized in order to eliminate either suffixes or prefixes. Anyhow, such kind of stemmer does not talk about the broken plurals stemming while morphological analyzers employ both lexicons and morphological rules in order to eliminate the proper affixes (prefix, infix, and suffix), as indicated in Table 2.0. The analyzers of morphological tackle all the corpus of initial and final letters of a word and employ rules in order to verify the combination between such letters and the remaining stem for a given word. In fact these systems not only generates very precise stems, but also they commonly return a number of reasonable stems for the same term, Leading to a difficulty in identifying the most suitable stem that represents the term. These systems do not have the same efficiency as the light stemmers, although in some cases they resembling the findings of the light stemmers [3]. Light stemmers do not make many mistakes like aggressive root-based stemmers. Furthermore, the function of aggressive stemmers lies in decreasing the group size remarkably. The primary reason behind the existence of stemming errors in root-based algorithms for both Arabic and English languages refers to the stemmer’s inability to identify the specific speech acts that fall under each utterance, such as noun, verb, and preposition.

Table 2.0 shows the affixes of Arabic word

		Prefix	Infix	Suffix
Arabic Word	السباحون	ال	ا	ون
English Word	Swimmers	AL	A	Won

Moreover, the finding that emerges after making comparative analysis of context-sensitive morphology and non-context-sensitive morphology denotes that the previous morphology considers more influential than the later morphology. In addition, [7] indicated that utilizing a root-extraction stemmer in Arabic language resembles the findings of Khoja stemmer i.e. without employing a stem glossary. Also, the root-extraction stemmer is identical to the light stemmers in monolingual document Retrieval missions. Nonetheless, Khoja stemmer employs a root-base stemmer that eliminates suffixes, infixes, and prefixes (as indicated below in Table 3.0) while employs pattern matching in order to extract the stem by utilizing root dictionary table [2].

Every existed word in such table returns to the stem, but for non-existent word, it returns the word to its origin without adjusting it.

Table 3.0 Arabic word root extraction by Khoja stemmer

	Word	Root
Arabic	**الخليجي**	**خلج**
English	Al-Khaliji	Khalj

It should be noted that Arabic morphology in IR is intended to find terms that have either similar or related meanings. Moreover, it has been clearly shown that by the purpose of indexing Arabic text, the effectiveness of restoring the terms or stems might be considerably increased by employing stems [8, 9, 10].
Anyhow, in spite of the light stemming capacity to precisely combine many variations of terms into large groups of stems, but it does not the capacity to combine other forms [11]. As indicated in [12], this confirms that stem-based IR considers more influential than root-based IR.
To illustrate, the primary obstacle that confronts the researchers in root-based IR is the terms variants that do not transfer identical meanings or explanations. Although every variant has its specific meaning they emanate from the same root. Thus, it is extremely possible for researchers or document analysts to struggle with vague and blurred words. Therefore, it is extremely vital to stimulate the AIR system by combining the meaning of disambiguated word.

Table 4.0 shows that Suffixes and prefixes are eliminated by light10

Prefixes	Suffixes
ال، وال، بال، كال، فال، لل، و	**ها، ان، ات، ون، ين، يه، ية، ه، ة، ي**

The purpose of the light stemmer Light10 [3], lies in eliminating both prefixes and suffixes of the term that are existed in table contains various length of suffixes and prefixes. As indicated below in Table 4.0.
Such work seeks to search for the suggested light stemmer as well as additional group of suffixes that might be suitable for promoting AIR efficiency.

Experiment

The present study touches upon clarifying both the usage and the impacts of light stemmer. It contrasts the usage of two different runs by employing TF.IDF term weighting scheme. Such light stemmers methods were inspected by employing a huge set that has not been existed prior the rendering of Arabic Cross-Language Restoraction at TREC 2001.
The study estimates the word weighting scheme performance trough reducing the stop words with the light stemming that has similar stop words lists. In fact, both accuracy and recalling have been employed. Most importantly, the efficiency of the light stemming methods was estimated to identify which one yields the utmost performance for Arabic language Retrieval.

Data Set

Such research employs one Arabic test set that is produced in the Linguistic Data Consortium in Philadelphia as well as employed in in the recent TREC tests. The Arabic Newswire A set was generated by both David Graff and Kevin Walker at the Linguistic Data Consortium [11]. It contained articles that were taken from Agence

France Presse (AFP) Arabic Newswire. The origin of the material was tagged by employing TIPSTER style SGML and was converted to Unicode (UTF-8). The set contains articles since 13 May 1994 until 20 December 2000. The information is in 2,337 compressed Arabic text information folders. The number of compressed information is 209 Mbytes while the number of uncompressed information is 869 Mbytes, around 383,872 documents consisted of 76 million tokens over approximately 666,094 unique words.
The purpose of creating the query group that is related to LDC set refers to TREC 2001 and 2002 [13, 14,15] Moreover, Arabic, English, and French contain 25 subjects that are consisting of Title, Description, and Narrative fields. Both titles and descriptions are employed in Arabic as queries in our monolingual test [3].

Stopwords lists

Many researchers, such as [3, 4, 8, 9] and others have employed Khoja Stop words table. Such stop words table is regarded as the most common words in Arabic because it primarily contains Arabic preposition and conjunctions as well as it is considered inadequate. Therefore, they are insufficient for any AIR whether in natural language processing or in any other system because of its limitation.

Light stemmer

The usage of light stemmer represents in stemming process. Indeed, the adopted mechanism in such study is Larkey's. If the rest of the word contains three or more letters, the additions such: "و" or 'and' that occur at the beginning of the term shall be taken out.

The definite articles shall be removed in case of remaining two or more characters.
Skim the suffixes list located in the (right to left) order illustrated in Table 5, taking out any that are existed at the end of the word in case of remaining two or more characters.
Table 5.0 demonstrates the strings that have to be taken. The given "prefixes" mean both definite articles and conjunctions. Every string that considers from Arabic prefixes shall not be taken out by the light stemmers.

Table 5.0 shows both prefixes and suffixes that are taken out by light stemmer

Prefixes	Suffixes
ال، وال، بال، كال، فال، لل، و	**ها،ان،ات،ون،ين،هن،هم،ته،تي،ني،يه،ية،ه،ة،ي**

The adopted standard for formatting the light stemmers consider indicative. The light stemmer is not a dictionary. Therefore, it is unable to determine an application specialized in eliminating affixes for the remaining Arabic words. light stemmer success might attributed to the fact of their words changing regardless their existences in a word list. The purpose of the trial is embodied in removing strings that might be existed dependably as affixes much more than existence either the beginning or the end of an Arabic stem without affixes.

Experiment Process

Both data and query group regarding the tests were tackled out below:

The number of files 383.872 existed in the information group was transferred from UTF-8 template to Windows.
Title and specification per each 25 queries were taken from the main inquiry group.
The set and inquiries were adjusted in accordance with the next steps, the same adjustment is employed by [3]:

Taking out the punctuation

Taking out diacritics (mainly weak vowels) that are existed in some inputs. Such elimination leads to the consistency

Taking out non-letters

Substituting ا،أ and إ with ا

Substituting the last ى withي

Substituting the last ة with ي

Removing the stop words per each run

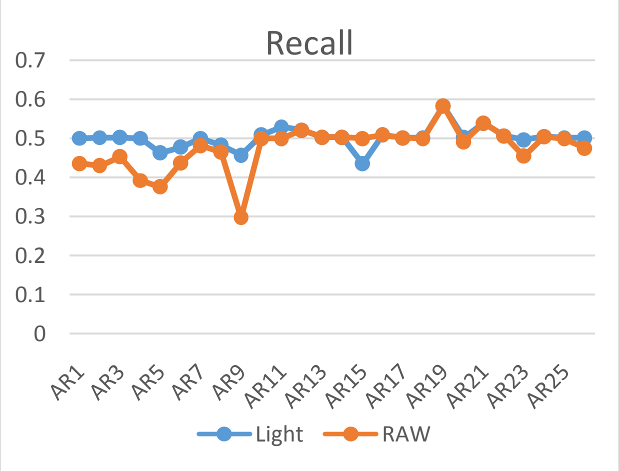
Symbolizing by employing a white area.

Stemming through the usage of light stemmer.

Estimation through employing TFIDF the average uncompleted accuracy for the 25 inquires as well as the compression ratio regarding the system index per each run.

Evaluation and Result

After an around two runs were implemented where RAW through both returning the text to its origin and omitting the stop words without derivation. The aim of creating the baseline is to measure the system performance after constructing each phase. According to the TREC-2001 relevance judgment each of the evaluation measures were employed whether accuracy or recalling in order to estimate the performance of the baseline system. The working mechanism of the suggested light stemmer is embodied in normalizing text, omitting the stop words, and light stemming. The findings of output Retrieval were analysed through counting the average variation among both runs. However, non-interpolated accuracy as well as recall and the compression rate score as indicated in in table 6.0.



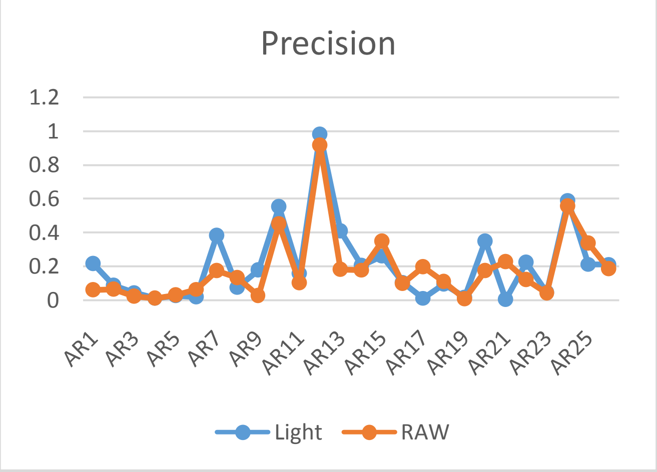


Figure 1.0 the average recall for the TREC 25 queriesFigure 2.0 the average precision for the TREC 25 queries

Table 6. Average Precision, Recall and compression rate for the two runs.

	RAW	Light Stemmer
Average Precision	0.1855	0.21
Average Recall	0.4754	0.5010
Compression Rate	0.40	0.48

The present study demonstrates the effect of light stemmer in promoting the efficiency of AIR concerning the average accuracy, recalling, and the ratio of compression index volume. Figure 1.0 demonstrates the average recalling per 25 employed queries in such test. Clarifying the improvement in the recalling ratio for most of the quires by employing the light stemmer. Figure 2.0 demonstrates the accuracy ratio for 25 inquires was the light stemmer that surpasses the raw system for the most inquires that have been completed by the text edit. Next, the paper will be prepared for the template afterward. After that, doubling the template file through employing “Save As” option. With regard to naming the paper you have to employ the naming convention stipulated in the conference. Such recently established files shed the light on whole contents as well as supply the developed text file. Now you can start now formatting your paper through the usage of the scroll down window inserted on the left of the MS Word Formatting toolbar.

Conclusion and Future Work

Such research talks about the impacts of stemming (light stemming) and its effect of enhancing Arabic monolingual IR by utilizing extra suffixes comparing with Larkey Ref light built. Moreover, the research provides the result of an experimental study that contrasts the performance of both abovementioned experiments for Arabic language in data Retrieval by utilizing standard estimation regarding data Retrieval. The estimation method for both of them as follows: the recalling measure estimates AIR while the accuracy measure estimates restoring the demanded the real information and guarantees that users are convinced with the finding. Consequently, contrasting both Arabic systems may clarify the impacts of the stemmer on enhancing the efficiency of information Retrieval for Arabic documents

The recent light stemmer suffixes denote their impact on enhancing Arabic Retrieval and efficiency. Such test demonstrates that additional investigation is necessary for Arabic Information Retrieval. In the upcoming work, further affixes list that contain various light, morphological analysis, or hybrid will be our goal for promoting the efficiency of AIR.

Acknowledgment

Foremost, we would like to express our sincere gratitude to linguistic Data Consortium (LDC) for supplying us with LDC2001T55 Arabic Newswire Part 1 at no-cost as one of the student recipients of the Autumn LDC Data Scholarship program.

References

[1] Abouenour, L., Bouzouba, K. & Rosso, P.   2010.   An Evaluated Semantic Query Expansion and Structure-Based Approach for Enhancing Arabic Question/Answering. International Journal on Information and Communication Technologies 3(3): 37-51.

[2] S. Khoja, "APT: Arabic part-of-speech tagger," 2001, pp. 20-25.

[3] L. Larkey, L. Ballesteros, and M. Connell, "Light stemming for Arabic information retrieval," Arabic computational morphology, pp. 221-243, 2007.Chen, A. & Gey, F.   2002.   Building an Arabic Stemmer for Information Retrieval. Proceedings of TREC 2002, pp. 631-639.

[4] Ahmed, F. & Nürnberger, A.   2009.   Evaluation of N- Gram Conflation Approaches for Arabic Text Retrieval. Journal of the American Society for Information Science and Technology 60(7): 1448-1465.

[5] Morris, C.   2010.   A Review of Recent Developments in Term Conflation Approaches for Arabic Text Information Retrieval.

[6] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on, 2005, pp. 152-157.

[7] Abu-Salem, H., Al-Omari, M. & Evens, M. W. 1999. Stemming Methodologies over Individual Query Words for an Arabic Information Retrieval System. Journal of the American Society for Information Science 50(6): 524-529.

[8] B. Alhadidi, and M. Alwedyan. "Hybrid Stop-Word Removal Technique for Arabic Language." Egyptian Computer Science Journal 30.1 (2008): 35-38.

[9] M. Wedyan, B. Alhadidi, and Adnan Alrabea. "The effect of using a thesaurus in Arabic information retrieval system." Int. J. Comput. Sci 9 (2012): 431-435.

[10] Kanaan, G., and M. Wedyan. "Constructing an automatic thesaurus to enhance Arabic information retrieval system." The 2nd Jordanian International Conference on Computer Science and Engineering, JICCSE. 2006.

[11] Atwan, J., Mohd, M. & Kanaan, G. 2013. Enhanced Arabic Information Retrieval: Light Stemming and Stop Words. Soft Computing Applications and Intelligent Systems, pp. 219-228. Springer.

[12] M. Aljlayl and O. Frieder, "On Arabic search: improving the retrieval effectiveness via a light stemming approach," 2002, pp. 340-347.

[13] N. I. o. S. a. Technology. (2002). TREC 2002 cross language topics in Arabic. Available: http://trec.nist.gov/data/topics_noneng/

[14] N. I. o. S. a. Technology. (2001). Data - Non-English Relevance Judgements File List. Available: http://trec.nist.gov/data/qrels_noneng/.