

Conditional Arabic Light Stemmer: CondLight

Yaser A. M. Al-Lahham

Computer Science Department

Zarqa University

Zarqa – Jordan

Khawlah Matarneh

Computer Science Department

Zarqa University

Zarqa – Jordan

Mohammad A. Hassan

Computer Science Department

Zarqa University

Zarqa – Jordan

Abstract

Arabic language has a complex morphological structure, which makes it hard to select index terms for an IR system. The complexity of the Arabic morphology caused by multimode terms, using diacritics, letters have different forms according to its location in the word and affixes can be added at all locations in a word. Many methods were proposed to overcome this problem; such as root extraction and light stemming. Light stemming show better retrieval efficiency, Light10 is the best stemmer among a series of light stemmers, it simply removes suffixes and prefixes if it is listed in a predefined table. Light10 has no restrictions on the affixes, so it is possible to have two different terms having the same token while they have different meanings. This paper proposes adding extra prefixes and suffixes to the table, and imposes some conditions on removing these affixes. The implementation and testing of the proposed method show better precision than the Light10 stemmer.

Keywords: Arabic IR, Light stemming, term selection, Arabic document indexing.

1. Introduction

Arabic language has complex morphological and orthographic forms making it difficult to develop efficient Arabic information retrieval systems. Moreover the written Arabic letters have different forms depending on its location in the word, for example; the letter (س) has the following forms: **س** at the end of the word, **سـ** in the middle of the word, and **سـ** at the beginning of the word. For information retrieval, this complex nature result a greater likelihood of mismatch between the form of a word in a query and the forms found in documents relevant to the query[1]. On the other hand, words preceded by some prepositions (ب، ك، ل) are agglutinated to them making no distinction between the word and the preposition. In fact many affixes, other than prepositions, have no distinction to the words they agglutinated to; for example: one letter prefixes indicate the gender of the person in the present tense verbs; such as **يكتب** and **تكتب**.

Most of Arabic words are derived from about 10000 roots, which are basic linguistic constructs (mostly 3 letters, few roots of 4 or 5 letters) having a semantic meaning.

Words can be derived by applying some morphological rules. Actually, thirty words can be formed out from a three letter root, in the mean while, not all of these permutations are valid Arabic words[2].

In written Arabic, the vowels (diacritics) are omitted, so several words of different meaning will have the same shape. For example, the word (حسب) can mean (calculate), (think that), or (according to). This ambiguity makes a crucial problem in information retrieval, since an Arabic word can have several meanings.

Another problem of the Arabic language, is the plural form of the irregular nouns, also called broken plural, such words have different forms than its singular form. A solution to this problem is to have a dictionary of such words, which is not always available[2].

It is clear that words having the same origin and attached to different affixes in Arabic (and other languages) need to be equated in a normalized form. One of the tools that used to fulfill this requirement is stemming, for example; in English and many other western European languages, stemming is primarily a process of suffix removal[3]. In Arabic, many stemmers were developed, but the most effective stemmers for information retrieval are the light stemmers, and the best empirical results were recorded by the Light10 stemmer [4].

Light10 stemmer is a table-driven stemmer; i.e. it has a list of affixes such that they are removed from any word that its affix matches any table entry or entries.

The main objective of this paper is to improve Arabic information retrieval by improving the effectiveness of the Light10 stemmer by imposing some rules that

satisfy the nature of the morphology of the Arabic language.

These rules attempt to overcome some problems that light stemmers cause; such as over-stemming, miss-stemming and under-stemming. These drawbacks decrease the effectiveness of stemming algorithms, however these problems are mutually correlated, that means an attempt to reduce the effect of one type may has negative consequence to the other[3].

The proposed method involves exploring important words (instead of excluding less important ones) using definite articles of Arabic words, and other properties such as plural suffix for both male and female.

The rest of the paper is organized as follows: section 2 presents a review of the literature related to Arabic term selection; light stemming in particular. Section 3 introduces the proposed enhancement over Light10 stemmer. Section 4 includes the implementation and evaluation details, and finally section 5 concludes the results of this research.

2. Related Work

Arabic language term selection include four types of stemming techniques, namely: manually constructed dictionaries, root-based stemming (morphological analyzers), statistical stemmers and light stemmers. In fact each of these techniques represents a level in a scale of analysis.

Root-based stemmers use morphological analysis to extract the root of a given Arabic word. One of the most popular root-based stemmers is the *Khoja's* stemmer [5]. Implementation of this stemmer depends on removing suffixes, infixes and prefixes and uses pattern matching to extract the roots. It also includes lists of diacritic characters, punctuation characters, definite articles and 168 stop-words that used to normalize search terms. The algorithm suffered from some drawbacks especially with broken plurals.

Khoja algorithm was improved by *Taghvaetal.* [6], by eliminating the need for a dictionary to support their root stemmer, thus eliminating the intensive maintenance and system requirements associated with a dictionary of the entire Arabic language.

Al-Omarietal. [7] introduced a new Arabic light/heavy stemmer called Arabic Rule-Based Light Stemmer, which is not based on Arabic root patterns. Instead, it depends on mathematical rules and some relations between letters. A series of tests were conducted on Arabic Rule-Based Light Stemmer to compare the effectiveness of this new Arabic stemmer with the effectiveness of *Khoja* stemmer. Experimental results

show clearly that Arabic Rule- Based Light Stemmer is more effective than the other tested Arabic stemmer.

Al-Kabi [8] indicated that many Arabic stemming algorithms were presented during the last fifteen years. *Khoja* Arabic stemmer is considered by many researchers as a standard stemmer for Arabic language. Therefore his study attempts to improve the effectiveness of this standard stemmer, through adoption of new additional (Patterns, Forms). Adding those patterns help to improve its accuracy by around 5%.

This paper focuses on light stemming, so this category will be presented in more detail in the rest of this section.

The light stemming approach, or affix removal, is a process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognize patterns and find roots [4].

Rule based stemmers apply conditions to differentiate between the added affixes and the original part of the word, for example; *Al-Shalabietal.* [9] proposed a new rule-based light stemmer that depends on a set of possible prefix and suffix, in prefix set they combined all possible antefixes and prefixes, on the other hand, the suffix set combines all possible suffixes and postfixes. In this method, a word is compared against an Arabic word patterns list collected from different resources, if there is a match, the prefix is part of the word and will not be truncated. Otherwise, if the word does not match any Arabic pattern, then the relationship between the suffix and prefix will be examined, if there is no relationship between the suffix and prefix then at least one of them is a part of the word.

AlAmeedetal. [10] analyzed TREC-2002 list of affixes, and proposed a sequence of patterns, that applied to improve the effectiveness of light stemmer.

Regarding statistical methods, *Darwish* [11] presented a method for generating morphologically related tokens from Wikipedia hypertext to page title pairs, by altering the spellings of transliterated named entities.

Boudlal et al. [12] presented a morphological analysis system for unvoweled Arabic sentences. The morphological analysis process often gives multiple solutions. They showed that by introducing the context, an approach based on hidden Markov models can have very good results in choosing the correct root of the word.

Hybrid stemming methods that combine stemming techniques with statistical string matching techniques, especially n -gram based approaches[13]. *Alhanini* [14] proposed enhanced Arabic stemming algorithm that uses both the light stemming and dictionary-based stemming that designed to overcome the disadvantages of both. *Hadni et al*[15] proposed a hybrid method which incorporates three different techniques for Arabic stemming to overcome some of the existing stemming algorithms weaknesses. The three techniques are: Khoja Stemmer, Light Stemmer and N-Gram. *Al-Nashashibiet al.*[16] proposed a hybrid stemmer that included rule-based, such stemmer removes prefixes and suffixes from the word according to specific rules. The pattern-based infix remover removes infixes from the word according to specific patterns. *Khedr et al.* [17] introduced a new algorithm that provides an extension for a new set of rule and patterns by extending light stemmer of *EL-Beltagy*[18] and Light10 stemmer.

Affix removal algorithms[19],[20] and [4] introduce a very close set of rules with certain conditions to remove prefixes and suffixes from words. Each of them define a set of rules that applied to words in order to remove prefixes or suffixes; such as removing definite articles for prefixes, or removing pronouns for suffixes.

Light10 proposed by *Larkey et al.* [4] and [21] as a developed version of a series of Arabic light stemmers. Light10 is the most effective of this series. They concluded that light stemming improves retrieval without providing correct morphological roots. Light10 stemmer applies the following procedure:

- Remove the letter و ("and") if the remainder of the word length exceeds, or equal to three characters. Even the removal of the letter 'و' is helpful, it may cause some problems, since many Arabic words begin with this letter as an original part of it.
- Remove any of the definite articles, if two or more characters left, and finally,
- Remove affixes from a word that match any item in a predefined list, if the length of the word left is at least two characters. The list of prefixes is: ال, وال, بال, كال, فال, لل, و. The list of suffixes is: يه, ية, ه, ة, ي, ها, ان, ات, ون.

This paper proposed an enhancement of Light10 by adding additional affixes to remove according to some conditions, so the proposed method is called

CondLight.

3. Conditional Light stemmer: CondLight

Light stemming is used to unify Arabic terms in a simple way without dealing with the complex morphological nature of these words. The unified terms (or tokens) can be used to index Arabic documents in an IR system, for example: different documents that use different forms of a term that has the same meaning will be indexed, even if they don't contain that form. Moreover these documents will be returned as a result of a user need, even they don't include the form of a term entered by the user, table-1 presents an example of this type of retrieval.

Table-1 IR Indexing without stemming, and with stemming

Without Stemming: the user must use the same form of a term to retrieve relevant documents					
المعلم	User query				
semantic	والمعلم	فالمعلم	كالمعلم	المعلمون	المعلمات
Index term	والمعلم	فالمعلم	كالمعلم	المعلمون	المعلمات
With stemming: the user could use any form of a term and get the relevant documents					
المعلم	User query				
semantic	والمعلم	فالمعلم	كالمعلم	المعلمون	المعلمات
Index term	معلم	معلم	معلم	معلم	معلم

This research proposes adding extra prefixes to the list that has been used in Light10 stemmer, which is: (ت, ب, ف, ل, س, ي) and new suffixes (ه, هـ, و). In order to remove these prefixes and suffixes, a set of conditions should be satisfied; these conditions are derived from the morphological properties of the Arabic language, and could be stated as follows:

- Condition1: To delete the suffixes: (هـ, ه, و) from the end of a token, it should not begin with one of those

(ال, فال, كال, وال), since in these cases, these suffixes are an original part of the word.

- Condition2: To delete the prefix (ب) from the token, it must end with one of this set of suffixes (ة, ون, هـ, ه, و), in this case this token is a present verb, for example: (يفعلها, يفعلون, يفعلان, يفعل, يفعلوا). The same condition is applicable for the tokens that begin with (ت).

- Condition 3: To delete the prefix (ب) from the token it must end with one of following set of suffixes (هـ, ه, و, ين, ات, ان, ة, هـ, هـ), since this token is a noun and this prefix is not

4

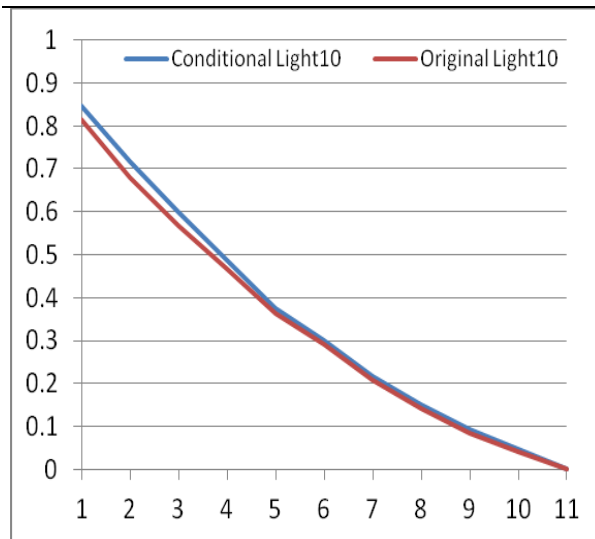


Figure-1: 11-recall levels precision-recall curve for Light10 and CondLight stemmers

Experimental results show that the conditional light stemmer gains about 5% enhancement of retrieval over the original Light10 stemmer. The enhancement occurs at lower recall levels (as shown in Figure-1), which satisfies users' need, such as retrieving relevant documents in the first page while searching web engines.

This slight enhancement could be justified as some terms (of different meaning) will have the same morphological shape after affixes removal, making the system retrieve more irrelevant documents; i.e., increasing the true negative retrieval, which resulted in a lower precision. When applying the proposed conditions, this situation becomes less probable, hence the chance to have true negative retrieval becomes lower, and consequently enhancing the precision ratio.

5. Conclusion

In this paper, a conditional light stemmer is proposed, such that a set of new affixes are proposed to be removed if they satisfied one or more of a set of proposed conditions. The application of the proposed light stemmer shows that adding some conditions to light stemmers enhances the retrieval at the lower recall levels. As a future work it is worth to test the proposed stemmer against the correctness of the resulted words after affixes removal.

REFERENCES

- [1] A. Souidi, Abdelhadi, N. Günter and V. Antal, "Arabic computational morphology: knowledge-based and empirical methods," *Arabic Computational Morphology*, Springer Netherlands, 2007, pp. 3-14.
- [2] Y. Kadri and J. Nie, "Effective stemming for Arabic information retrieval," *proceedings of the Challenge of*

- Arabic for NLP/MT Conference, Londres, Royaume-Un, 2006.
- [3] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, issue (3), pp. 130-137, 1980.
- [4] L. Larkey, L. Ballesteros and M. Connell, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland, 2002.
- [5] S. Khoja, R. Garside and G. Knowles, "A Tagset for the Morphosyntactic Tagging of Arabic Proceedings of the Corpus Linguistics. Lancaster University (UK), Volume 13 - Special issue, 341.
- [6] K. Taghva, R. Elkhoury and J. Coombs, "Arabic stemming without a root dictionary," *International Conference on Information Technology: Coding and Computing, ITCC 2005*, 4-6 April 2005.
- [7] A. Al-Omari, B. Abuata and M. Al-Kabi, "Building and Benchmarking New Heavy/Light Arabic Stemmer," in *International Conference on Information and Communication Systems*, Irbid, Jordan, 2013.
- [8] M. Al-Kabi, "Towards improving Khoja rule-based Arabic stemmer," *Proceedings of Applied Electrical Engineering and Computing Technologies (AEECT)*, IEEE Jordan Conference on, 2013.
- [9] R. Al-Shalabi, G. Kanaan, J. M. Jaam, A. Hasnah and E. Hilat, "Stop-word removal algorithm for Arabic language," *Proceedings of 1st International Conference on Information and Communication Technologies: From Theory to Applications*, pp. 545-550, 2004.
- [10] H. Al Ameer, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi and S. Al Muhairi, "Arabic light stemmer: a new enhanced approach," in *Proceedings of the second international conference on innovations in information technology*, Al Ain, UAE, 9 - 11 November, 2005.
- [11] K. Darwish and A. M. Ali, "Arabic Retrieval Revisited: Morphological Hole Filling," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 2012.
- [12] A. Boudlal, B. Rachid, L. Abdelhak, M. Azzeddine, M. Abdelouafi and M. Behah, "A Markovian approach for Arabic root extraction," *The International Arab Journal of Information Technology*, vol. 8, no. 1, pp. 13-20, 2009.
- [13] J. Kim, J. Taylor, "Fast string matching using an n-gram algorithm," *Journal of Software: Practice and Experience*, Volume 24, Issue 1, January 1994, Pages 79-88.
- [14] Y. Alhanini and M. J. A. Aziz, "The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming," *Journal of Software Engineering and Applications*, vol. 4, no. 9, pp. 522-52, 2011.
- [15] M. Hadni, A. Lachkar and S. A. Ouattik, "A new and efficient stemming technique for Arabic Text Categorization," *Proceedings of International Conference on Multimedia Computing and Systems (ICMCS)*, Tangier, 10-12 May 2012.



-
- [16] M. Y. Al-Nashashibi, D. Neagu and A. Yaghi, "Stemming techniques for arabic words: A comparative study," Proceedings of International Conference on Computer Technology and Development (ICCTD), Cairo, 2-4 Nov. 2010.
 - [17] S. Khedr, D. Sayed and A. Hanafy, "Arabic Light Stemmer for Better Search Accuracy," World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering 10.11 (2016), pp:3410-3418.
 - [18] S. R. El-Beltagy and A. Rafea, "An Accuracy Enhanced Light Stemmer for Arabic Text," ACM Transactions on Speech and Language Processing, Volume 7 Issue 2, February 2011.
 - [19] A. Chen and F.C. Gey, "Building an Arabic Stemmer for Information Retrieval," *TREC*. Vol. 2002. 2002.
 - [20] M. Aljlayl and O. Frieder, "On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach," in Proceedings of the eleventh international conference on Information and knowledge management, 2002.
 - [21] L. Larkey, L. Ballesteros, M. Connell, "Light stemming for Arabic information retrieval," *Arabic computational morphology*. 2007:221-43.
 - [22] F. Gey, and D. W. Oar, "Searching Arabic using English, French or Arabic queries," Text Retrieval Conference (TREC): TREC-2001 Cross- Language Information Retrieval Track. Gaithersburg, 20, 16-23.