ARABIC LIGHT STEMMER: ANEW ENHANCED APPROACH

Hayder K. Al Ameed, Shaikha O. Al Ketbi, Amna A. Al Kaabi, Khadija S. Al Shebli,
Naila F. Al Shamsi, Noura H. Al Nuaimi, Shaikha S. Al Muhairi
Software Engineering Dept., College of Information Technology, UAE University,
PO.Box 17555, Al-Ain, UAE
Email: {HayderA,ShaikhaO,200002533,200002104,200002027,200002358,200002124}@
uaeu.ac.ae

ABSTRACT

In general, word stemming is one of the most important factors that affect the performance of information retrieval systems. The optimization issues of Arabic light stemming algorithm as a main component in natural language processing and information retrieval for Arabic language are based on root-pattern schemes. Since Arabic language is a highly inflected language and has a complex morphological structure than English, it requires superior stemming algorithms for effective information retrieval.

This paper reports on the enhancement of a TREC-2002 Arabic light stemmer presented by Kareem Darwish, University of Maryland. Five stemming algorithms are proposed that result in significantly better Arabic stemming outcomes in comparison with the TREC-2002 algorithm.

Keywords: Arabic stemmer, Arabic light stemmer, Arabic morphological analyzer, Arabic retrieval, suffixes and prefixes stripping, and Arabic corpus.

1. INTRODUCTION

A morpheme is the smallest element that has a meaning. Morphemes cannot be split into smaller ones, and they should impart a function or a meaning to the word which they are part of [1]. The root is the original form of the word before any transformation process, and it plays an important role in language studies [2]. A stem is a morpheme or a set of concatenated morphemes that can accept an affix [3]. An affix is a morpheme that can be added before (prefix) or after (suffix), or inserted inside (infix) a root or a stem to form new words or meanings [4].

The direction of the writing of the script is not the only difference between Arabic and many other languages. The major difference is that Arabic is mainly derivational while others are concatenative [5]. The removal of prefixes in English is usually harmful because it can reverse or otherwise alter the meaning or grammatical function of the word. This is not so in Arabic, since the removal of prefixes does not usually reverse the meaning of words [6].

Stemming is the process of removing all of a word's prefixes and suffixes to produce the stems or the root [7]. The importance of the stemming process comes in the classification and index builders/searchers because it makes the operations fewer dependants on particular form of words, and it reduces the potential size of vocabularies which might otherwise have to contain all possible forms.

The stemming algorithm is a computational process that gathers all words that share the same stem and have some semantic relation [8]. The main objective of the stemming process is to remove all possible affixes and thus reduce the word to its stem [9]. It is normally used for document matching and classification by using it to convert all likely forms of a word in the input document to the form in a reference document [10].

Arabic stemming algorithms can be classified, according to the desired level of analysis, as either stem-based or root-based algorithms. Stem-based algorithms, remove prefixes and suffixes from Arabic words, while root-based algorithms reduce stems to roots [11]. Light stemming refers to the process of stripping off a small set of prefixes and/or suffixes without trying to deal with infixes or recognize patterns and find roots [6].

Researchers concluded that Arabic information retrieval can be enhanced when the roots or stems are used in indexing and searching [6]. Al-Jlayl and Frieder showed that stem-based retrieval is more effective than root-based retrieval [12]. The main problem of the root-based algorithm in information retrieval is that many surface word variants do not have similar semantic interpretations. Although these surface words are different in meaning, they originate from the same root. Thus, using the root-based algorithms in information retrieval increases the word ambiguities. Word-sense disambiguation is essential to improve any Arabic information retrieval system [6].

2. THEORETICAL BACKGROUNG

Arabic language belongs to the Semitic language group. The grammatical system of Arabic language is based on a root-and-pattern structure and considered as a root-based language with not more than 10000 roots and 900 patterns [13]. The root is the bare verb [14]; root can be tri-literal, which is the overwhelming majority of words (85%) [15], and to a lesser extent, quad-literal, pen-literal, or hexa-literal, each of which generates increased verb forms and noun forms by the addition of derivational affixes (i.e. prefixes and suffixes) [16].

In addition to singular and plural constructs, Arabic has a form called "dual" that indicates precisely two of something. For example, a pen is "qalam", two pens are "qalamaan", and pens are "aqlaam". As in French, Spanish, and many other languages, Arabic nouns are either feminine or masculine, and the verbs and adjectives that refer to them must agree in gender.

2.1 Arabic Roots and Patterns

Arabic patterns are part of the Arabic grammar. They are formed based on the Arabic root [17]. A root is the base form of a word which can not be further analyzed without the loss of the word's identity, or it is that part of the word left when all the affixes are removed.

An Arabic root is an ordered sequence of three (فعل) or four letters (فعلل) from alphabet [18]. The root has a general, basic meaning which forms the basis of many related meanings. These related meanings are represented by the root consonants put in different forms called patterns [19]. They are generated from the process of vocalization and affixation [20]. Table 1 shows a sample of the Arabic Patterns (Three-Consonant root).

TABLE 1: Arabic patterns sample (Trilateral roots)

Arabic Patterns									
fa'ala ↓	فعل	mustafa'ael	مستفعل	mafa'ael	مفاعيل	tafa'alon	تفعلون	fea'altan	فعالتان
maf'al	مفعل	mustafa'aelat	مستفعلات	afa'al	افعل	tafa'aln	تفعلن	fa'a:la	فعالى
mafa'aloon	مفعلون	mustafa'alon	مستفعلون	afa'aela	افعلاء	tafa'aelat	تفعيلات	fa'ala	فعلى
mafa'aleen	مفعلين	mutafa'ael	متفاعل	ifti'a:l	افتعل	Fa:'il	فاعل	fea'ali	فعلي
mafa'alan	مفعلان	mutafa'alat	متفاعلات	eftea'al	افتعال	faa'aelan	فاعلان	fa'aol	فعول
mafa'aool	مفعول	mufa'ael	مفاعل	tafa'aul	تفاعل	fa'ael	فعائل	yafa'al	يفعل
maf'ala	مفعلة	mufa'aeloon	مفاعلون	tafa'aulan	تفاعلان	fi'a:l	فعال	yafa'alan	يفعلان
mefa'aal	مفعال	mufa'aelat	مفاعلات	tafa'alen	تفعلين	fea'ala	فعالة	yafa'alon	يفعلون

Variations of the root and patterns determine the actual meaning of the word. For example, the root (ktb كتب) with the addition of the letters (i ,a) gives the word (kita:b كتاب), which means book,

---

↓ The combination (a'a) represents the letter ' ' of the Arabic alphabet.

while the root pattern combination of (ka:tib **كاتب**) means "one who writes" or "clerk". There are also some prefixes and suffixes which determine whether a word is a subject marker, pronoun, preposition, or a definite article. Table-2 illustrates set of derivatives patterns, its corresponding English word, the position in the language and its Arabic patterns from the Arabic trilateral verbal root 'k t b'.

TABLE 2: Derivatives of the Arabic trilateral root 'k t b'.

Arabic	English	POS	Pattern	Arabic	English	POS	Pattern
Ktb	Write	V	Fa'ala	Maktab	Office	N	Maf'al
Kita:b	Book	N	Fi'a:l	Maktabah	Library	N	Maf'ala
Kita:bah	Writing	N	Fi'a:lah	Muka:tabah	Correspondence	N	Mufa:'alah
Ka:tib	Writer	N	Fa:'il	Iktita:b	Subscription	N	Ifti'a:l
Ka:taba	correspond	V	Fa:'ala	Kita:bi	Clerical	Adj	Fi'a:li

2.2 Prefixes and Suffixes in Arabic Language

Arabic language, unlike English, both prefixes and suffixes are removed for efficient result, but Arabic provides the additional difficulty of infixes [21]. The difficulty arises because Arabic has two genders, feminine and masculine; three numbers, singular, dual, and plural; and three grammatical cases, nominative, genitive, and accusative. A noun has the nominative case when it is a subject; accusative when it is the object of a verb; and genitive when it is the object of a preposition. The form of an Arabic noun is determined by its gender, number, and grammatical case [22].

2.3 Arabic Light Stemmers

Word stemming in Arabic is the process of removing all of a word's prefixes and suffixes to produce the stem or root [23]. Simply, it is a conversion of plurals to singulars, or derivation of a verb from the gerund form. There are also other possibilities such as deriving the root from the pattern words. The importance of the stemming process is in the classification and index builders/searchers because it makes the operations less dependant on particular forms of words and reduces the potential size of vocabularies, which might otherwise have to contain all possible forms.

There are several stemming approaches that are applied for Arabic language; one of them is light stemmer algorithm. It is not an aggressive practice as the root-based algorithm. The aim of this approach is not to produce the linguistic root of a given Arabic surface form; rather, it is to remove the most frequent suffixes and prefixes. In Arabic, unlike English, both prefixes and suffixes are removed for efficient results, but Arabic provides the additional difficulty of infixes [24]. The difficulty arises because Arabic has two genders, feminine and masculine; three cardinality, singular, dual, and plural; and three grammatical cases, nominative, genitive, and accusative. A noun has the nominative case when it is a subject; accusative when it is the object of a verb; and genitive when it is the object of a preposition. The form of an Arabic noun is determined by its gender, cardinality, and grammatical case.

Arabic verbs have two tenses: perfect and imperfect. Perfect tense denotes actions completed, while imperfect denotes uncompleted actions. The imperfect tense has four moods: indicative, subjective, jussive, and imperative. Arabic verbs in perfect tense consist of a stem and a subject marker. The subject marker indicates the person, gender, and number of the subject. The form of a verb in perfect tense can have subject marker and pronoun suffix. The form of a subject-marker is determined together by the person, gender, and number of the subject [18].

This classification of Arabic words exhibits some difficulties for the stemming techniques. As a result, several attempts have been made to improve the Arabic light stemmers in the last few years. Recently, a number of light stemmers were developed. Three of them are: Al-Stem,

3

developed at Maryland and modified by Leah Larkey at University of Massachusetts [25]. Second was described by Larkey, and third one called TREC-2001 stemmer, it is a modified version of the Larkey's stemmer in which two additional prefixes identified [26]. Lastly, TREC-2002 developed as an improved version of TREC-2001 [27].

TREC Light Stemmers remove only prefixes and suffixes. Five pre-defined groups of removable prefixes and suffixes were offered. For prefixes there are three groups (one, two and three-characters), and two groups (two and three-characters) for suffixes, Table 3 shows the TREC-2002 predefined set of prefixes and suffixes. The developer of the TREC-2002 light stemmer identified 9 three-character, 14 two-characters, and 3 one-character prefix terms that are supposed to be removed in stemming, and 17 two-characters, and 4 one-character suffix terms that should be removed by the stemmer. The algorithm designed to non-recursively remove the prefixes from the given word based on the pre-defined set of prefixes, and to recursively remove the suffixes from the given word based on the predefined set of suffixes [28].

TABLE 3: Comparison between TREC-2002 and proposed suffixes and prefixes terms

TREC Prefixes			TREC Suffixes		Proposed Prefixes				Proposed Suffixes	
PREFIX 1	PREFIX 2	PREFIX 3	SUFFIX 1	SUFFIX 2	PREFIX 1	PREFIX 2	PREFIX 3	PREFIX 4	SUFFIX 1	SUFFIX 2
و	كا	كال	ه	ها	ي	ال	وال	وبال	ه	ان
ل	وس	وال	ة	يه	ت	لل	بال		ة	ين
ب	سي	فال	ي	هم	ن	سي	فال		ك	ون
	لا	بال	ت	نا		سا	كال		و	ات
	وب	ولل		ما		ست	ولل		ي	هم
	وت	مال		وا		سن	وسي		ن	هن
	وم	اال		يا		كا	وست		ا	ها
	لل	سال		ني		فا	وسن		ت	كم
	يا	لال		هن		با	وسا			كن
	وا			كم		ب	ولا			نا
	ال			كن		ل	ولي			وا
	فا			تم		لي	ولت			تم
	ول			تن		لت	ولن			ني
	وي			ان		لن				تن
				ين		فت				ته
				ون		في				يه
				ات		فن				ما
										يا
										تا
										تك

3. NEW PROPOSED APPROACH

At the beginning, we carefully analyzed the Text Retrieval Conference 2002 (TREC-2002) light stemmer algorithm. Later, we tried to design new algorithms to get better stemmed results. A system prototype was developed to judge against the algorithms outcome. The enhancement attempts move toward into two main approaches; the first covers the TREC pre-defined removable suffixes and prefixes groups including the contents of terms in each set and the number of terms. The second approach focuses on modifying the sequence of algorithm components execution. In the first approach (Affixation list), we considered the pre-defined TREC-2002 groups of suffixes and prefixes terms. In addition new affixes terms were added including single four-character affixation term which is not encountered in the TREC algorithms.

Our proposed groups include a single four-character, 13 three-characters, 17 two-characters and 3 one-character prefix terms; the suffixes groups contain 20 two-characters and 8 one-character terms, table-3 illustrates the differences between the TREC and proposed suffix and prefix terms. In table-3, PREFIX columns hold prefix affixes terms, while SUFFIX used for suffix affixes. The number indicates the number of term's characters.

In the second algorithm enhancement approach (execution mechanism), we designed five light stemming algorithms in addition to the traditional pre-defined TREC-2002; the six developed algorithms are mainly categorized into two implementation sorts, the *Suffix-Prefix* (SP first remove suffixes terms reclusively while prefix terms removed non-recursively at the end), and *Suffix-Prefix-Suffix* (SPS removes single largest available suffix term first, after largest single prefix term, at the end it removes a single largest remaining suffix if any).

For TREC algorithm, we developed the original defined algorithm (removes suffixes recursively and a single prefix non-recursively), we called it SP_TREC. The same defined affixation terms list were used but with a modified execution step via Suffix_Prefix_Suffix truncating process, that modified algorithm is called SPS_TREC.

We noticed that most of Arabic words use (الـ Alef Lam) prefix as a declarative term (e.g., Al Kitab الكتاب (The Book) and Al Mua'alem المعلم (The Teacher)) Therefore, we proposed two new major categories in classifying of the designed algorithms; Without-AlefLam (WOAL the stemmer accepts the non-stemmed words after removing the prefixed AlefLam) and With-AlefLam (WAL stemmer acquires the whole non-stemmed word without any pre-processing).

The SP and SPS implementation sorts' algorithms were fashioned with the three different implementations TREC-2002, with Alef-Lam (WAL) and Without Alef-Lam (WOAL). The resulted six algorithms called *SP_TREC*, *SPS_TREC*, *SP_WAL*, *SPS_WAL*, *SP_OWAL*, and *SPS_WOAL*.

4. RESULTS ANALYSIS AND DISCUSSION

To evaluate the performance of the developed stemmer algorithms, more than 1450 Arabic words were prepared that fall into three main groups: singular, dual and plural including verb and noun words.

The system performance evaluation was based on two testing manners; the former manner focused on measuring the number of acceptable (meaningful) produced words as an output of applying the stemmer algorithms on each test group. While the second manner is based on measuring the frequency of removing affixation terms from the test words.

For the first testing scheme, Table 4 illustrates the assessed percentage of the produced words.

TABLE 4: Comparison between the developed algorithms (accepted words)

	SP_WAL	*SPS_WAL*	*SP_WOAL*	*SPS_WOAL*	*SP_TREC*	*SPS_TREC*	*Non-Searchable*
Dual	81.75%	79.37%	62.70%	79.37%	45.24%	45.24%	3.17%
Singular	60.07%	63.13%	69.80%	62.04%	32.49%	32.71%	1.75%
Plural	69.16%	73.79%	74.67%	73.57%	36.56%	37.00%	1.98%

From Table 4, we clearly notice that the proposed algorithms improve the stemmed words by 30-40% more than the traditional TREC-2002 approach. Figures 1, 2 and 3 show the statistical charts for the results of singular, dual and plural words group in sequence.
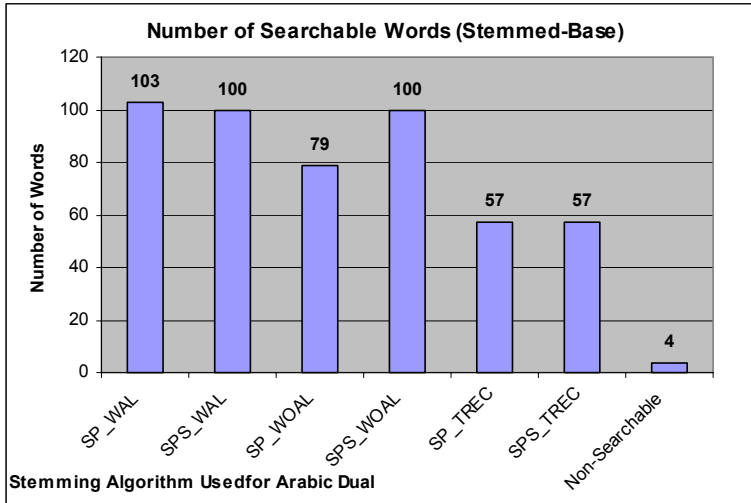
Number of Searchable Words (Stemmed-Base)

Figure 1: Searchable Arabic Dual words after applying the six algorithms (for 126 words).

Number of Searchable Words (Stemmed-Base)
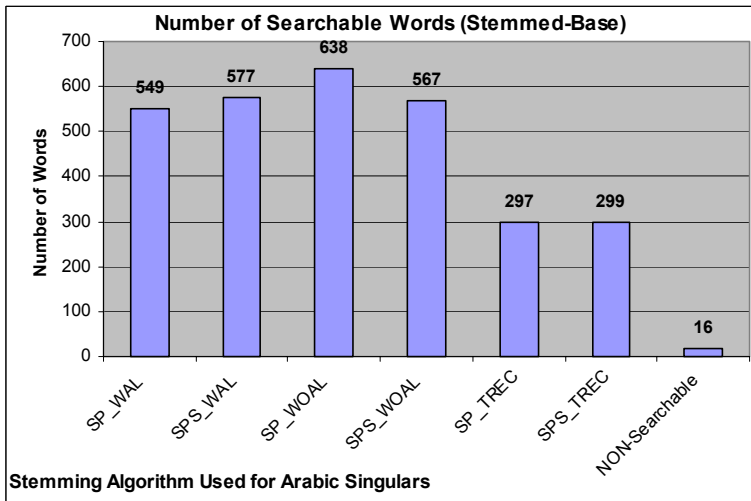
Figure 2: Searchable Arabic Singular words after applying the six algorithms (for 775 words).
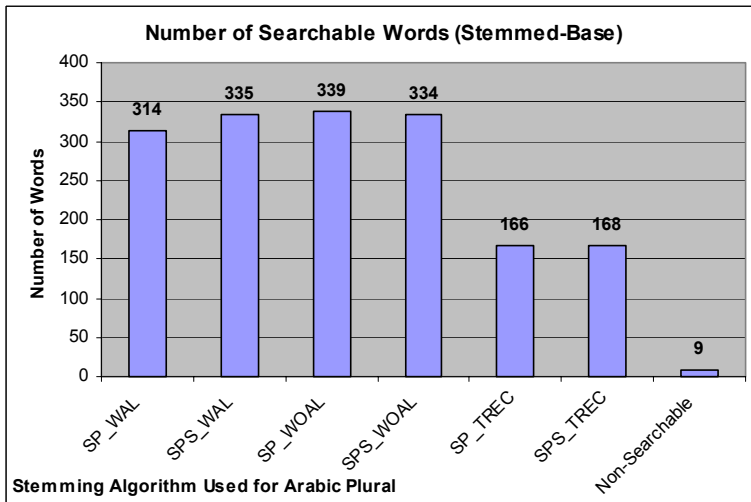
Number of Searchable Words (Stemmed-Base)

**Figure 3:** Searchable Arabic Plural words after applying the six algorithms (for 606 words).

From the above figures, we concluded that the SP_WOAL algorithm provides the highest number of functional truncated affixes terms over TREC and other developed algorithms. The percentage of the non-searchable words is less than 16% in any test group.

By applying the subsequent test manner (removed affixation terms frequency), we noticed that SP-WOAL algorithm removes less average number of suffix and prefix terms than other algorithms that provide evidence of understemming judging stemmers criteria. Table-5 demonstrates the counted truncated affixation terms for the three test groups after applying the six stemming algorithms.

TABLE 5: Comparison between the developed algorithms (number of removed terms)

Stemmer Algorithm	Plural test group (606 words)		Singular test group (775 words)		Dual test group (126 words)	
	Suffix Terms Removed	Prefix Terms Removed	Suffix Terms Removed	Prefix Terms Removed	Suffix Terms Removed	Prefix Terms Removed
SP_TREC	384	44	638	157	135	7
SPS_TREC	383	45	645	154	140	7
SP_WAL	425	242	800	484	181	39
SPS_WAL	401	279	733	545	163	47
SP_WOAL	278	242	311	470	114	42
SPS_WOAL	**400**	**280**	**736**	**542**	**163**	**47**

The values of Table 5 imitated as a statistical charts into figures 4, 5 and 6. The removed suffix terms from the plural test group words are shown in figure 4a, while the frequency of removed prefix terms shown in 4b. For the number of removed affixation terms of the other two test groups figures 5 and 6 are prepared.

(a)   (b)
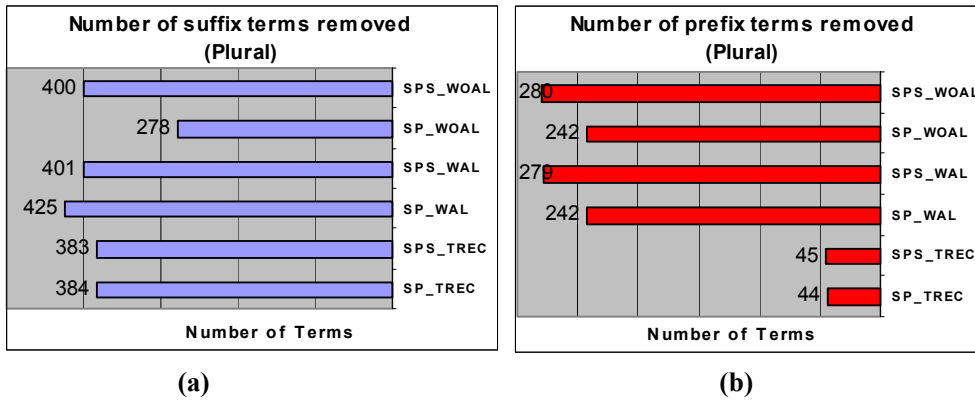
Figure 4: Charts indicating the number of truncated affixes terms (Plural group).

(a)   (b)
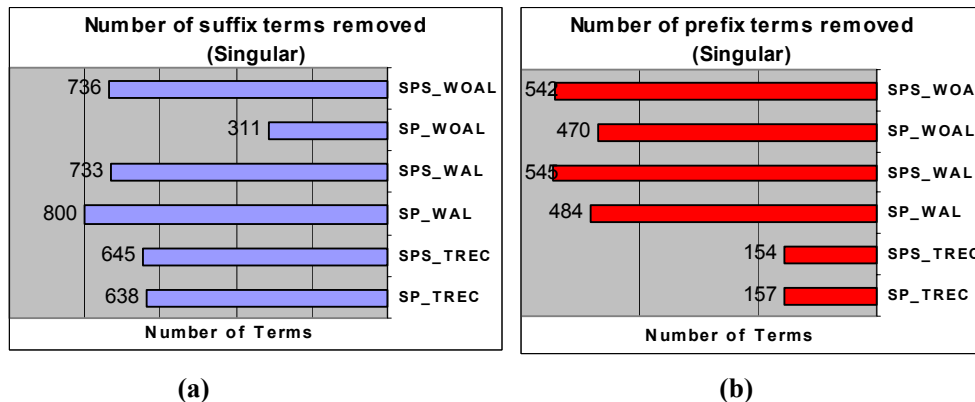
Figure 5: Charts indicating the number of truncated affixes terms (Singular group).

7

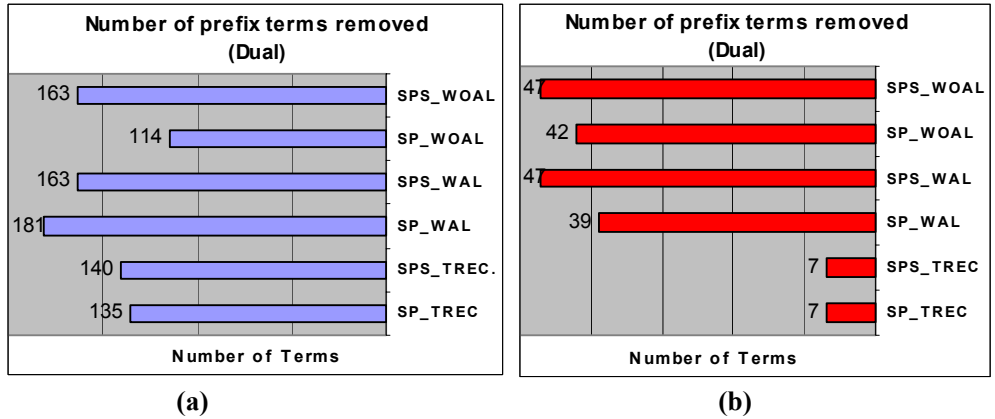**(a)**   **(b)**

Figure 6: Charts indicating the number of truncated affixes terms (Dual group).

CONCLUSION

We have presented five proposed algorithms for improving Arabic light stemmers. The proposed light stemmer algorithms were assessed by using more than 1450 Arabic words including different set of affixation, patterns, as well as hollow verbs and various types of strong verbs. The proposed algorithms outcomes were compared with TREC-2002 algorithm results. Our proposed approaches provide better accepted (meaningful) outcomes of Arabic words with up to 30-50% more than TREC stemmer outcomes.

The assembled statistics identified the new defined SP_WOAL light stemmer approach as an aggressive algorithm in producing the highest percentage of searchable words. In addition to the results mentioned above, the number of non-searchable words turns lower; generally speaking this research offers approaches that enhance the traditional light stemmer engines to be fitted properly into common Arabic information retrieval systems within few modification steps.

REFERENCES

[1] Andrew Spencer. "*Morphological Theory: An introduction to word structure in generative grammar*". Oxford: Blackwell Publishers. 1991.

[2] Metri George. "*Al Khaleel: A dictionary of Arabic syntax terms*". Beirut: Library of Lebanon, 1990.

[3] Muhammad Al-Khuli. "*A Dictionary of Theoretical Linguistics: English-Arabic with an Arabic-English Glossary*". Published by Library of Lebanon, 1991.

[4] B. Thalouth, and Abdulla Al-Dannan. "*A Comprehensive Arabic Morphological Analyzer/Generator*". IBM Kuwait Scientific Center. Kuwait, 1987.

[5] Imed Al-Sughaiyer, and Ibrahim Al-Kharashi. "*An Efficient Arabic Morphological Analysis Technique for Information Retrieval Systems*". In ACIDCA'2000 International Conference. Monastir, Tunisia, March 2000.

[6] Imed Al-Sughaiyer, and Ibrahim Al-Kharashi. "Arabic *Morphological Analysis Techniques: A Comprehensive Survey*", Journal of the American Society for Information Science and Technology, Vol 55, Issue 3, PP. 189 - 213, February 2004.

[7] Shereen Khoja. (no date). "*APT: Arabic Part-if-speech Tagger*". [online]. Available from: http://archimedes.fas.harvard.edu/mdh/arabic/NAACL.pdf [Accessed 17/07/03]

[8] Chris Paice. "*A Method for the Evaluation of Stemming Algorithms Based on Error Counting*". Journal of the American Society for Information Science, 47, 632-649, 1996

[9] John Dawson. "*Suffix Removal and Word Conflation*". ALLC (Association for Literacy & Linguistic Computing) Bulletin, 2(3), 33-46, 1974.

[10] Peter Burden. (2000). "*Stemming Algorithms and their Use*". [online]. http://www.scit.wlv.ac.uk/. Available from: http://www.scit.wlv.ac.uk/seed/docs/mypapers/stemalg.html [Accessed 25/02/04]

[11] Jinxi Xu, Alexander Fraser, and Ralph M. Weischedel. "*Empirical Studies in Strategies for Arabic Retrieval*". In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), pp. 269-274. Tampere, Finland, 2002.

[12] Mohammed Al-Jlayl, and Ophir Frieder. "*On Arabic Search: Improving the Retrieval Effectiveness via Light Stemming Approach*". In Proceedings of the 11th ACM International Conference on Information and Knowledge Management, Illinois Institute of Technology, pp. 340-347, New York: ACM Press, 2002.

[13] Pierre MacKay. "*Computers and the Arabic language*". New York : Hemisphere Pub. Corp., 1990.

[14] Hani Abu-Salem, Mohammed Al-Omari, and Martha W. Evens. "*Stemming methodologies over individual query words for an Arabic information retrieval system*", Journal of American society for information science, volume 50, Issue 6, pp 524-529, 1999.

[15] Sabah Al-Fedaghi, and Fawaz Al-Anzi. "*A new algorithm to generate Arabic root-pattern forms*". Proceeding of the 11th National Computer Conference, king Fhad University of Petroleum & Minerals, Dhahran, Saudi Araibe. , pp04-07.

[16] Saliba, and Abdulla Al-Dannan. "*Automatic morphological analysis of Arabic: A study on content word analysis*". Proceeding of the first Kuwait computer conference, pp. 3-8, 1989.

[17] Arabic Grammer [online]. www.mesiti.it. Available from: http://www.mesiti.it/arabic/grammar/lessons/lesson2/roots.html [Accessed 18/05/05]

[18] David Crystal. "*A Dictionary of Linguistics and Phonetics*" (Language Library). 4th ed. Blackwell Publishers, 1997.

[19] Jonathan Owens. "*The foundations of grammar: an introduction to medieval Arabic grammatical theory*". John Benjamins Publishing Company, 1988.

[20] Joseph Dichy, and Ali A. Farghaly. "*Roots & Patterns vs. Stems plus Grammar-lexis Specifications: On What Basis Should a Multilingual Database Centered on Arabic be Built?*" MT Summit IX -- workshop: Machine Translation for Semitic Languages, New Orleans, USA, 2003.

[21] Monica Rogati, Scott McCarley, and Yiming Yang. (2003). "*Unsupervised Learning of Arabic Stemming Using a Parallel Corpus*". [online]. Available from: http://acl.ldc.upenn.edu/acl2003/main/pdfs/Rogati.pdf [Accessed 26/02/04]

[22] Aitao Chen, and Fredric C. Gey. (no date). "*Building an Arabic Stemmer for Information Retrieval*". [online]. Available from: http://metadata.sims.berkeley.edu/papers/trec2002.pdf [Accessed 26/02/04]

[23] Shereen Khoja. "*APT: Arabic Part-of-Speech Tagger*". Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), Pittsburgh, Pennsylvania, pp. 20-26, 2001.

[24] Monica Rogati, Scott McCarley and Yiming Yang. "*Unsupervised Learning of Arabic Stemming using a Parallel Corpus*". Proceedings of the 41st Annual Meeting of the Association, for Computational Linguistics, pp. 391-398, July 2003.

[25] Kareem Darwish, (2002). "*Al-stem: A light Arabic stemmer*" [Online]. Available: http://www.glue.umd.edu/~kareem/research.

[26] Alexander Fraser, Jinxi Xu, and Ralph Weischedel. "*TREC 2002 Cross-lingual Retrieval at BBN*". In TREC 2002 Proceedings, 2002.

[27] Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell. "*Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis*". In SIGIR'02, August 11-15, 2002, Tampere, Finland, pp. 275-282, 2002.

[28] Kareem Darwish, and Douglas W. Orad. 2003 "*CLIR Experiments at Maryland for TREC-2002: Evidence Combination for Arabic-English Retrieval*". In TREC 2002. Gaithersburg: NIST, pp 703-710, 2002.