A Comparative Study on Arabic Stemmers

Mohamed Y. Dahab
Computer Science Department,
Faculty of Computing &
Information Technology
King Abdulaziz University,
Jeddah, Saudi Arabia

Asma'a Al Ibrahim
Computer Science Department,
Faculty of Computing &
Information Technology
King Abdulaziz University,
Jeddah, Saudi Arabia

Rihab Al-Mutawa
Computer Science Department,
Faculty of Computing &
Information Technology
King Abdulaziz University,
Jeddah, Saudi Arabia

ABSTRACT
Stemming is considered as a pre-processing step in many applications: text mining, information retrieval, machine translation etc. The Arabic language has many special cases or properties that affect stemming or any automatic method, it depends on both inflectional and derivational morphology to produce the various forms of the language words. Many researchers have proposed algorithms to solve the problems of stemming. This paper aims to make a comparison study among the existing Arabic stemmers, the comparison study is based on the methodologies, the usage, main idea, algorithm, the affixes, limitations, output, and the stemmers' sensitivity for both diacritics and context.

General Terms
Natural Language Processing.

Keywords
Arabic Stemmers, Arabic Morphological Analyzer.

1. INTRODUCTION
Stemming is the process of correlating several terms onto one common representation in the base form [16]. It minimizes the index size because it has the advantage of reducing storage requirements by eliminating the redundant words. This leads to improved results; increasing the matching probability for unifying vocabulary [6]. A controversial issue is whether to use stems or roots as terms for indexing. Most of the modern studies indicate that using stems as index terms outperform roots [13]. Stemming uses morphological heuristics in order to remove affixes from words and the processing cost is relatively low. For those reasons, the stemming is important and highly attractive for many natural language processing (NLP) fields such as: information retrieval (IR), question answering (QA), information extraction (IE), machine translation (MT), text summarizations (TS), Text Classification (TC), Text Clustering (TClu), Text segmentation (TS), Indexing (Ind), and Automatic Speech Recognition (ASR) [16]. There are many developed algorithmic stemming and various morphological analysis approaches to achieve morphologically related forms combined under the same stem using stemmer [14]. The approaches for stemming are three different approaches: the light stemmer; the root-base stemmer; and the statistical stemmer [27].

Stemmers are generally developed for each specific language. Their design requires an understanding of the needs of information retrieval and some linguistic experience in the language. Stemmers have been developed for a wide range of languages including Arabic, English, Chinese, French, German, Italian, and many other languages. Many factors influence the stemming across languages. Stemming either makes small difference in retrieval or it improves performance by a small amount. It improves effectiveness more for the highly inflected languages such as the Arabic language and when documents and/or queries are short [25].

Arabic stemming is more complicated than stemming in any European language such as English [16]. The usage of affixes in English language is less complex compared to Arabic language because for English language, stemmers are only concerned with the removal of the suffixes [4]. Such stemmers do not conflate irregular forms as (write, wrote) [25].

This paper is a comparative study for the most of the existing stemmers that uses different approaches for stemming. These stemmers are about twenty which are: Arabic stemmer [23], AlKhalil morphological analyzer [9], SAFAR [17], Arabic Stemming with two dictionaries [21], new version of the Arabic Morphological analyzer MORPH2 [19], light-stemming algorithm [2], the light stemmer [23], Light8 [25], Light10 [24], Berkeley light stemmer [10], Al-Stem Stemmer [11], SP_WOAL Arabic Light Stemmer [1], Rule-based Light Stemmers [20], Linguistic-based stemmer [18], Domain-Specific Arabic Stemmer [12], The ISRI Arabic Stemmer [28], Enhanced algorithm for extracting the root of Arabic words, Buckwalter Arabic Morphological Analyzer (BAMA), and light stemmer [29]. It compares stemmers based on the differences and the similarities between various stemmers in terms of main ideas behind the development of the stemmers, removed prefixes and suffixes, reason of choosing the affixes, dealing with infixes, and the limitations, more information about Arabic prefixes, suffixes and infixes can be found in [5], the stemmers are also compared in terms of sensitivity for both diacritics and context, usages in the different fields of NLP.

The characteristics (or novelties) of this article are presented as follows. Compared with [3] and[26] published five years ago, this survey introduces and describes more recent works in this area. In addition to compare existing Arabic stemmers based on the sensitivity of both diacritics and context as well as the possibility of usages in the different fields of NLP.

The remainder of the paper is organized as follows. In section 2, Arabic stemmers is illustrated. Section 3 explains the criteria of comparison. The comparisons of stemmers are presented in section 4, and the conclusion is given in sections 5.

2. ARABIC STEMMERS
Arabic belongs to the Semitic family of languages group which also includes languages like Hebrew and Aramaic [1]. It is more complicated than any other language. The changes in the forms of words morphological have the same effective

change in the meaning of words and they can be considered equivalent [26]. The grammatical system of the Arabic language is based on a root-and-pattern structure, it also considered as a root-based language with not more than 10000 roots and 900 patterns [1].

Arabic stemming is an approach that goes after finding the origin of words in the natural Arabic language by removing any additional (affixes) in the words [26]. Since Arabic language is a highly inflected language, it requires a good normalization and stemming for effective information retrieval. Orthography and morphology led to a huge amount of lexical variation [25].

The direction when writing the script is not the only difference between Arabic and many other languages [1]. Unlike the English language, Arabic has much richer morphology, it has two genders, feminine and masculine. Arabic has a form called "dual" in addition to singular and plural constructs, it indicates precisely two of something. It also has three numbers, singular, dual, and plural; and three grammatical cases, nominative, genitive, and accusative. Arabic nouns are either feminine or masculine, it 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 [10].

The form of an Arabic noun is determined by its gender, number, and grammatical case, so the verbs and adjectives that refer to them must agree in gender [1]. The definitive nouns are formed by attaching the Arabic article (ال) to the immediate front of the nouns. Sometimes a preposition, such as (ب) (i.e, by) and (ل) (i.e., to), is attached to the front of a noun, often in front of the definitive article. Besides prefixes, a noun can also carry a suffix which is often a possessive pronoun. The conjunction word (و) (and) is often attached to word. Arabic has two kinds of plurals: sound plurals and broken plurals. The sound plurals are formed by adding plural suffixes to singular nouns [10].

All these factors make the Arabic language a very difficult language to stem. The difficulty arises because Arabic is mainly derivational while others are concatenative [1]. There are four different approaches to Arabic stemming that can be identified: first is the manually constructed dictionaries of words with stemming information, then there is the light stemmers which remove prefixes and suffixes, without trying to deal with infixes, or recognize patterns and find roots. The third one is the morphological analyses which attempt to find roots. The final one is the statistical stemmers, which group word variants using clustering techniques [25].

3. THE CRITERIA OF COMPARISON

This research is considering most of the studies in Arabic light stemming and Arabic root-based stemming which are useful in different fields of the NLP. It compares between the following twenty stemmers: Arabic stemmer [23], AlKhalil morphological analyzer [9], SAFAR [17], Arabic Stemming with two dictionaries [21], new version of the Arabic Morphological analyzer MORPH2 [19], light-stemming algorithm [2], the light stemmer [23], Light8 [25], Light10 [24], Berkeley light stemmer [10], Al-Stem Stemmer [11], SP_WOAL Arabic Light Stemmer [1], Rule-based Light Stemmers [20], Linguistic-based stemmer [18], Domain-Specific Arabic Stemmer [12], The ISRI Arabic Stemmer [28], Enhanced algorithm for extracting the root of Arabic words [15], Buckwalter Arabic Morphological Analyzer (BAMA), and light stemmer [29]. This study is comparing Arabic stemmers based on the following aspects:

- The main ideas behind the each stemmer and limitation.
- The sensitivity for diacritics and context.
- The usages in the different fields of NLP domains.

3.1 THE COMPARISON OF ARABIC STEMMERS

3.2 The main ideas behind the each stemmer and limitation

We summarizes the similarities, and differences between all the stemmers included in this study in terms of main idea, the prefixes and suffixes removal, basis of selecting these affixes, dealing with infixes, and limitations. The summaries are represented in table 1.

3.3 The sensitivity for diacritics and context

The comparison between the stemmers in terms of sensitivity for diacritics is shown in table 2. If the stemmer produce the root or the stem with respect diacritics if any then, it is sensitive to diacritics, otherwise it is not.

Table 1: Main idea and limitations

Stemmer	Main Idea	Limitations
Arabic Stemmer by Shereen Khoja [22]	First, it removes diacritics then it removes stop-words, punctuation, and numbers. After that, it removes definite article (ال). And inseparable conjunction (و) and finally, it removes suffixes and prefixes.	First, the root dictionary requires maintenance to guarantee that newly discovered words are correctly stemmed. Second, it replaces a weak letter with (و) which occasionally produces a root that is not related to the original word. Finally, the stemmer will fail in some cases to remove all affixes by following certain order.
AlKhalil Morphological Analyzer by Boudlal Abderrahim,[9]	First , Segment the text into word then identify them (Segmentation), after that it analysis the steam of the segmentation that was validated (assume an interpretation that corresponds to a word that is non-derivable, a second interpretation that could refer to a noun and a third	Does not differentiate between clitics and affixes

Stemmer	Main Idea	Limitations
	interpretation to a verb) , that last step will be Result screening	
SAFAR (Software Architecture For Arabic Language Processing) by Jafar and Bouzoubaa [17]	Compares the morphological analyzers by Integrate them and save the output as xml file and matched it to meet the standard format of Alecso then convert it into memory objects , each morphological analyzer processes the input text of the evaluation corpus. The results of each morphological analyzer are then retrieved as memory objects , measure the performance of a given morphological analyzer by return a list of metrics	Depends on the stemmer that is going to be compared.
Arabic Stemming with Two Dictionaries by Kchaou and Kanoun [21]	It has two dictionaries one is for roots and the other is for the stems, it remove the character according to the word length	Does not handle irregular plural.
New Version of the Arabic Morphological Analyzer MORPH2 by Kammoun et al [19]	First, divides the text into sentences then extracts clitics and classifies the stem after that it identify the possible root and affixes and determines all possible morph syntactic features, finally it matching the analyzed word with its pattern	Does not handle irregular plural.
Light-Stemming Algorithm by Aljlayl and Frieder's [2]	remove the prefix (و) if the word is greater than or equal to three characters then normalize (آ, أ, إ) from the beginning of the word to (ا) after that remove the suffixes form the stem if the remaining stem is greater than or equal to three characters than if the remaining still three characters, remove the prefixes form the stem an finally Return the stem	Does not handle irregular plural.
The Light Stemmer by Leah Larkey [23]	It does not deal with infixes or patterns; it is generally the process of removing prefixes and/or suffixes	the stemmers does not handle irregular plural and it removes the affixes without any prior knowledge in linguistic rules. It does not deal with infixes or patterns
Light8 by Leah Larkey [25] Light10 by Leah Larkey [24]	The stemmers only removes the prefixes and suffixes. It also tries to remove the strings that would be found as affixes far more often than they would be found as the beginning or end of an Arabic word without affixes.	First, the stemmers does not handle irregular plural and it removes the affixes without any prior knowledge in linguistic rules.
Berkeley Light Stemmer by Chen and Gey [10]	Use the standard Arabic data collection provided by Linguistic Data Consortium removes only prefixes and suffixes. The suffixes removed recursively while the prefixes are not. He Also check if the remaining word is exist in the corpus for some words only	Does not handle irregular plural
Al-Stem Stemmer by Kareem Darwish [11]	Prefixes and suffixes is removed if they do exist at the beginning or the end of the word.	Remove affixes without any prior knowledge (linguistic rules). Second it very aggressive (which could remove a lot wrong strings from the words beginnings and ends). Does not handle irregular plural
SP_WOAL Arabic Light Stemmer by Al Ameed et al [1]	It removes everything may appear as a prefix or suffix in order to be exhaustive.	There was no evaluation for the stemmer against IR tasks. And it does not deal with irregular plural.
Rule-based Light Stemmers by Kanaan [20]	It keeps valid Arabic core words using simple rules or heuristics exist in Arabic language and using a large lexicon that contains all the forms of the Arabic language.	The rules do not assure correctness. There is a need for lexicon which contains all the forms of all the words in Arabic language which is very difficult to achieve. Also, it does not handle irregular plural.
Linguistic-based stemmer by Kadri and Nie [18]	Arabic word contains five parts; antefixes, prefixes, stem, suffixes and postfixes.	It does not deal with irregular plural.
Domain-Specific Arabic	The remaining word; after removing the affixes from the	It has to be used with a particular

Stemmer	Main Idea	Limitations
Stemmer by El-Beltagy and Rafea [12]	words, is looked up in the corpus text and user stems list.	domain. Also, to modify the mistakes done by the stemmer, the stem list that is built during the construction phase must have the user intervention.
The ISRI Arabic Stemmer by Taghva et al [28]	It is a root-extraction stemmer without root dictionary. The stemmer returns a normalized form for un-stemmed words.	It does not deal with irregular plural.
Enhanced Algorithm for Extracting the Root of Arabic Words by Ghwanmeh et al [15]	The stemmer is based on the roots of the Arabic word. It extracts the Arabic roots from the Arabic words.	It does not deal with irregular plural.
Buckwalter Arabic Morphological Analyzer (BAMA) by Tim Buckwalter (2004)	BAMA contains a dictionary of lexicons of Arabic prefixes, stems, and suffixes, with truth tables to indicate a correct combination of these three segments. It offers morphological categories such as Nouns, Function word, and Verbs. And it uses Buckwalter transliteration, which has possibility to get converted directly to Unicode Arabic with least amount of automatic processing.	It does not deal with irregular plural.
Light Stemmer by Naglaa Thabet [29]	It removes everything may appear as a prefix or suffix in order to be exhaustive.	There was no evaluation for the stemmer against IR tasks. And it does not deal with irregular plural.

Table2: Sensitivity for diacritics and context

Stemmer	Sensitivity for Diacritics and Context	
	Diacritics	Context
Arabic Stemmer by Shereen Khoja [22]	X	√
AlKhalil Morphological Analyzer by Boudlal Abderrahim,[9]	√	X
SAFAR (Software Architecture For Arabic Language Procesing) by Jafar and Bouzoubaa [17]	Depends on the stemmer that is going to be compared.	
Arabic Stemming with Two Dictionaries by Kchaou and Kanoun [21]	X	√
New Version of the Arabic Morphological Analyzer MORPH2 by Kammoun et al [19]	X	√
Light-Stemming Algorithm by Aljlayl and Frieder's [2]	X	√
The Light Stemmer by Leah Larkey [23]	X	√
Light8 by Leah Larkey [25]	X	√
Light10 by Leah Larkey [24]	X	√
Berkeley Light Stemmer by Chen and Gey [10]	√	√
Al-Stem Stemmer by Darwish [11]	X	X
SP_WOAL Arabic Light Stemmer by Al Ameed et al [1]	X	√
Rule-based Light Stemmers by Kanaan [20]	X	√
Linguistic-based stemmer by Kadri and Nie [18]	X	√
Domain-Specific Arabic Stemmer by El-Beltagy and Rafea [12]	X	X
The ISRI Arabic Stemmer by Taghva et al [28]	X	√
Enhanced Algorithm for Extracting the Root of Arabic Words by Ghwanmeh et al [15]	X	X
Buckwalter Arabic Morphological Analyzer (BAMA) by Tim Buckwalter	√	√

Stemmer	Sensitivity for Diacritics and Context	
	Diacritics	Context
Light Stemmer by Naglaa Thabet [29]	√	√

The comparison between the stemmers in the terms of sensitivity for context is shown in table 2, the context-sensitive stemmers prevents the production of insensible and invalid roots while the context-free stemmers may end up into some invalid and insensible roots or free morphemes [8].

3.4 The usages in the different fields of NLP domains

As shown in table 3,Light-stemming algorithm by Aljlayl and Frieder's [2] and Khoja [22] stemmers were only used in the information retrieval field from year 2000 to 2004, as for the light8 it was used in the information retrieval, Q&A, machine translation, text indexing and text summarizations fields. Berkeley light stemmer was used in the information retrieval, information extraction, machine translation and text indexing fields. Al-Stem was used in the information retrieval, extraction, machine translation, text indexing, summarizations and automatic speech recognition. Finally, BAMA and light stemmer by Naglaa Thabet [29] were not cited in this time period.

From 2005 until 2009, light-stemming algorithm by Aljlayl and Frieder [2] was used most in information retrieval flowed by text mining then information extraction and finally in both Q&A and machine translation. Khoja [22] was used most in information retrieval then text mining, indexing, Q&A and machine translation. Light8 was used in all the field that have been mentioned except in the text clustering field, information retrieval was the most used field for Light8 followed by text classification then text indexing then information extraction, machine translation and automatic speech recognition, finally text mining, Q&A, text segmentation and summarizations. Berkeley light stemmer was used in information retrieval the most, then text classification, information extraction, and machine translation, finally Q&A and automatic speech recognition. Al-Stem was used the most in information retrieval then machine translation, Q&A and finally information extraction. SP_WOAL Arabic Light Stemmer was only used in information retrieval field. Linguistic-based stemmer by Kadri and Nie [18] was used in two field information retrieval and Q&A. The ISRI Arabic Stemmer was used twice in the information retrieval field followed by information extraction, text mining and classification. Buckwalter was used only in machine translation. Light stemmer by Naglaa Thabet [29] used only in information retrieval and text classification. Light10, Rule-based Light Stemmers by Kanaan [20], Enhanced algorithm for extracting the root of Arabic words by Ghwanmeh [15] were not cited in this time period.

From 2010 until 2014, light-stemming algorithm by Aljlayl and Frieder [2] was used most in the text mining field, followed by information retrieval then text classification, indexing, clustering, then Q&A, finally text segmentation and summarization. Tashaphyne was only used in text mining. Arabic Stemming with two dictionaries used in only two fields, information retrieval and extraction. MORPH2 was used twice in both fields': information extraction and text mining, and once in information retrieval. Khoja [22] was used most in information retrieval followed by text mining and classification then information extraction and finally Q&A. AlKhalil was used twice in machine translation and text segmentation, and once in information extraction, text classification and automatic speech recognition. The light stemmer by Leah Larkey [25] was used the most in information retrieval and text classification then information extraction followed by machine translation and text clustering, then text mining and Q&A. Light8 was used in all the filed that have been mentioned except in the automatic speech recognition field, it was used most in information retrieval then text classification followed by information extraction and text clustering, then machine translation, then indexing and Q&A and finally text mining, segmentation and summarization. Light10 was used in three fields: information retrieval, text indexing and segmentation. Berkeley light stemmer was used most in information retrieval followed by information extraction and text classification, then text mining, then machine translation and text indexing, finally text clustering and summarization. Al-Stem was used most in information retrieval followed by text classification, then text clustering and finally text mining and Q&A. SP_WOAL Arabic Light Stemmer was used most in information retrieval followed by information extraction, then text mining, classification and summarization. Rule-based Light Stemmers by Kanaan [20] was used three time in information retrieval and once in text classification. Linguistic-based stemmer by Kadri and Nie [18] was used most in information retrieval and text classification followed by Q&A, then information extraction and finally text mining. Domain-Specific Arabic Stemmer by El-Beltagy and Rafea [12] used in text classification the most, then text mining, information retrieval and text summarization. The ISRI Arabic Stemmer was used most in information retrieval then text classification, information extraction, text mining, Q&A and text clustering. Enhanced algorithm for extracting the root of Arabic words by Ghwanmeh [15] was used most in information retrieval then text indexing and finally information extraction and text classification. Buckwalter used most in machine translation the most then automatic speech recognition then information retrieval, extraction, text mining and Q&A.

Table 3: The frequency of research citations' usages for each stemmer in the fields of NLP.

Stemmer	**Year**	**NLP Field**										
		IR	**IE**	**TM**	**TC**	**QA**	**MT**	**Ind**	**TS**	**TClu**	**Sum**	**ASR**
light-stemming algorithm by Aljlayl and Frieder's [2]	**2000 -**	5	-	-	-	-	-	-	-	-	-	-

Stemmer	**Year**	**NLP Field**										
		IR	**IE**	**TM**	**TC**	**QA**	**MT**	**Ind**	**TS**	**TClu**	**Sum**	**ASR**
Arabic stemmer by Shereen Khoja [22]	**2004**	9	-	-	-	-	-	-	-	-	-	-
Light8 by Leah Larkey [25]		11	-	-	-	1	1	1	-	-	1	-
Berkeley light stemmer by Chen and Gey [10]		2	1	-	-	-	1	1	-	-	-	-
Al-Stem Stemmer by Kareem Darwish [11]		9	1	-	-	-	2	1	-	-	1	1
Buckwalter Arabic Morphological Analyzer (BAMA) by Tim Buckwalter(2004)		-	-	-	-	-	-	-	-	-	-	-
light stemmer by Naglaa Thabet [29]		-	-	-	-	-	-	-	-	-	-	-
light-stemming algorithm by Aljlayl and Frieder's [2]	**2005 - 2009**	16	2	10	-	1	1	-	-	-	-	-
Arabic stemmer by Shereen Khoja [22]		25	-	6	-	1	1	2	-	-	-	-
The light stemmer by Leah Larkey [23]		1	-	-	-	-	-	-	-	-	-	-
Light8 by Leah Larkey [25]		42	3	2	7	2	3	4	2	-	2	3
Light10 by Leah Larkey [24]		-	-	-	-	-	-	-	-	-	-	-
Berkeley light stemmer by Chen and Gey [10]		14	3		8	1	3					1
Al-Stem Stemmer by Kareem Darwish [11]		16	1	-	-	2	5	-	-	-	-	-
SP_WOAL		2	-	-	-	-	-	-	-	-	-	-

Stemmer	**Year**	**NLP Field**										
		IR	**IE**	**TM**	**TC**	**QA**	**MT**	**Ind**	**TS**	**TClu**	**Sum**	**ASR**
Arabic Light Stemmer by Al Ameed et al [1]												
Rule-based Light Stemmers by Abdusalam Nwesri (2008)		-	-	-	-	-	-	-	-	-	-	-
Linguistic-based stemmer by Kadri and Nie (2006)		2	-	-	-	-	1	-	-	-	-	-
The ISRI Arabic Stemmer by Taghva et al [28]		5	2	1	1	-	-	-	-	-	-	-
Enhanced algorithm for extracting the root of Arabic words by Ghwanmeh et al [15]		-	-	-	-	-	-	-	-	-	-	-
Buckwalter Arabic Morphological Analyzer (BAMA) by Tim Buckwalter (2004)		-	-	-	-	-	6	-	-	-	-	-
light stemmer by Naglaa Thabet [29]		2	-	1	-	-	-	-	-	-	-	-
light-stemming algorithm by Aljlayl and Frieder's [2]	**2010 - 2014**	17	-	23	7	4	-	6	3	6	3	-
SAFAR (Software Architecture For Arabic language Processing) by Jafar and Bouzoubaa [17]		-	-	-	-	-	-	-	-	-	-	-
Tashaphyne by Taha Zerrouki (2012)		-	-	3	-	-	-	-	-	-	-	-
Arabic Stemming with two dictionaries by Kchaou and Kanoun [21]		2	1	-	-	-	-	-	-	-	-	-
new version of the Arabic Morphological analyzer		1	2	2	-	-	-	-	-	-	-	-

Stemmer	**Year**	**NLP Field**										
		IR	**IE**	**TM**	**TC**	**QA**	**MT**	**Ind**	**TS**	**TClu**	**Sum**	**ASR**
MORPH2 by Kammoun et al [19]												
Arabic stemmer by Shereen Khoja [22]		24	4	13	13	1	-	-	-	-	-	-
AlKhalil morphological analyzer by Azzeddine Boudlal Abderrahim [9]		-	1	-	1	-	2	-	2	-	-	1
The light stemmer by Leah Larkey [23]		9	3	1	9	1	2	-	-	2	-	-
Light8 by Leah Larkey [25]		20	5	1	11	2	4	3	1	5	1	-
Light10 by Leah Larkey [24]		1	-	-	-	-	-	1	1	-	-	-
Berkeley light stemmer by Chen and Gey [10]		18	6	3	6	-	2	2	-	1	1	-
Al-Stem Stemmer by Kareem Darwish [11]		9	-	1	3	1	-	-	-	1	-	-
SP_WOAL Arabic Light Stemmer by Al Ameed et al [1]		5	2	1	1	-	-	-	-	-	1	-
Rule-based Light Stemmers by Abdusalam Nwesri (2008)		3	-	-	1	-	-	-	-	-	-	-
Linguistic-based stemmer by Kadri and Nie [18]		6	2	1	6	2	-	-	-	-	-	-
Domain-Specific Arabic Stemmer by El-Beltagy and Rafea [12]		1	-	2	3	-	-	-	-	-	1	-
The ISRI Arabic Stemmer by Taghva et al [28]		15	6	3	7	2	-	-	-	2	-	-

Stemmer	Year	NLP Field										
		IR	IE	TM	TC	QA	MT	Ind	TS	TClu	Sum	ASR
Enhanced algorithm for extracting the root of Arabic words by Ghwanmeh et al [15]		4	1	-	1	-	-	2	-	-	-	-
Buckwalter Arabic Morphological Analyzer (BAMA) by Tim Buckwalter (2004)		1	1	1	-	1	6	-	-	-	-	2
light stemmer by Naglaa Thabet [29]		10	1	3	2	-	-	-	-	-	-	-

4. CONCLUSION

Stemming is common requirement of NLP functions. It is a preliminary step in many applications involving information retrieval, text mining, machine translation, etc. Some natural languages are morphologically complex such as Arabic which is complicated more than any other language. There are many challenges in the NLP areas and this could be a motivating factor for NLP researchers around the world to develop more good stemming approaches. Arabic NLP researches are more developed year after another but there is a need for more open-source development of Arabic stemmers' resources. In this paper, we compared between twenty Arabic stemming approaches; Arabic stemmers, Khoja's stemmer [22], Al-Khalil morphological analyzer, SAFAR, Tashaphyne, Arabic Stemming with two dictionaries, MORPH2, light-stemming algorithm, Larkey's light stemmer, Light8, Light10, Berkeley light stemmer [23],[24] and [25], Al-Stem Stemmer, SP_WOAL Arabic Light Stemmer, Rule-based Light Stemmers, Linguistic-based stemmer, Domain-Specific Arabic Stemmer, ISRI Arabic Stemmer, Enhanced algorithm for extracting the root of Arabic words, Buckwalter Arabic Morphological Analyzer, light stemmer. The comparison between the stemmers are based on the differences and the similarities between various stemmers in terms of the methodologies, the usage in different fields of NLP, main ideas behind the development of the stemmers, algorithm, the affixes, limitations, output, and the stemmers' sensitivity for both diacritics and context. The table of stemmer's usage shows that from the year 2000 to 2014, the stemmers where used mostly in information retrieval, followed by text classification. And in general, light8 is the most used stemmer, Khoja stemmer [22] comes next. However, there are many other factors for comparisonbetween the stemmers that we suggest for future work such as fault rate, and accuracy.

5. REFERENCES

[1] Al Ameed, H., Al Ketbi, S., Al Kaabi, A., Al Shebli, K., Al Shamsi, N., Al Nuaimi, N. H., & Al Muhairi, S. S. (2005, September). Arabic light stemmer: A new enhanced approach. In The Second International Conference on Innovations in Information Technology (IIT'05).

[2] Aljlayl, M., & Frieder, O. (2002, November). On Arabic search: improving the retrieval effectiveness via a light stemming approach. In Proceedings of the eleventh international conference on Information and knowledge management(pp. 340-347). ACM.

[3] Al-Nashashibi, May Y., D. Neagu, and Ali Yaghi. "Stemming techniques for Arabic words: A comparative study." Computer Technology and Development (ICCTD), 2010 2nd International Conference on. IEEE, 2010.

[4] Al-Omari, A., & Abuata, B. (2014). ARABIC LIGHT STEMMER (ARS). Journal of Engineering Science and Technology, 9(6), 702-717.

[5] Aqel, Afnan, Sahar Alwadei, and Mohammad Dahab. "Building an Arabic Words Generator." International Journal of Computer Applications 112.14 (2015).

[6] Al-Shammari, E., & Lin, J. (2008, July). A novel Arabic lemmatization algorithm. InProceedings of the second workshop on Analytics for noisy unstructured text data (pp. 113-118). ACM.

[7] Al Sughaiyer, Imad A., and Ibrahim A. Al- Kharashi. "Arabic morphological analysis techniques: A comprehensive survey." Journal of the American Society for Information Science and Technology 55.3 (2004): 189-213.

[8] Bal, B. K., & Shrestha, P. (2004). A Morphological Analyzer and a stemmer for Nepali. PAN Localization, Working Papers, 2007, 324-331.

[9] Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. O. A. O., & Shoul, M. (2010). Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In *International Arab Conference on Information Technology*.

[10] Chen, A., & Gey, F. C. (2002, November). Building an Arabic Stemmer for Information Retrieval. In TREC (Vol. 2002, pp. 631-639).

[11] Darwish, Kareem. "Building a shallow Arabic morphological analyzer in one day." Proceedings of the

ACL-02 workshop on Computational approaches to semitic languages. Association for Computational Linguistics, 2002.

[12] El-Beltagy, S. R., & Rafea, A. (2011). An accuracy-enhanced light stemmer for arabic text. ACM Transactions on Speech and Language Processing (TSLP),7(2), 2.

[13] Eldesouki, M. I., Arafa, W., & Darwish, K. (2009). Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective. The Egyptian Computer Journal, 36(1).

[14] Fautsch, C. and Savoy, J. (2009), Algorithmic stemmers or morphological analysis? An evaluation. J. Am. Soc. Inf. Sci., 60: 1616–1624. doi: 10.1002/asi.21093

[15] Ghwanmeh, S., Kanaan, G., Al-Shalabi, R., & Rabab'ah, S. (2009, August). Enhanced algorithm for extracting the root of Arabic words. In Computer Graphics, Imaging and Visualization, 2009. CGIV'09. Sixth International Conference on (pp. 388-391). IEEE.

[16] Hammo, B. H. (2009). Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. Information retrieval, 12(3), 300-323.

[17] Jafar, Younes, and Karim Bouzoubaa. "Benchmark of Arabic morphological analyzers challenges and solutions." *Intelligent Systems: Theories and Applications (SITA-14), 2014 9th International Conference on*. IEEE, 2014.

[18] Kadri, Y., & Nie, J. Y. (2006, October). Effective stemming for Arabic information retrieval. In proceedings of the Challenge of Arabic for NLP/MT Conference, Londres, Royaume-Uni.

[19] Kammoun, N. C., Belguith, L. H., & Hamadou, A. B. (2010, June). The MORPH2 new version: A robust morphological analyzer for Arabic texts. InJADT 2010: 10th International Conference on Statistical Analysis of Textual Data.

[20] Kanaan, G., Al-Shalabi, R., Ababneh, M., & Al-Nobani, A. (2008, December). Building an effective rule-based light stemmer for Arabic language to inprove search effectiveness. In Innovations in Information Technology, 2008. IIT 2008. International Conference on (pp. 312-316). IEEE

[21] Kchaou, Z., & Kanoun, S. (2008, December). Arabic stemming with two dictionaries. In Innovations in Information Technology, 2008. IIT 2008. International Conference on (pp. 688-691). IEEE.

[22] Khoja, S., & Garside, R. (1999). Stemming arabic text. Lancaster, UK, Computing Department, Lancaster University.

[23] Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007). Light stemming for Arabic information retrieval. In Arabic computational morphology (pp. 221-243). Springer Netherlands.

[24] Larkey, L. S., & Connell, M. E. (2005). Structured queries, language modeling, and relevance modeling in cross-language information retrieval. Information processing & management, 41(3), 457-473.

[25] Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002, August). 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 (pp. 275-282). ACM.

[26] Otair, M. "Comparative analysis of Arabic stemming algorithms." International Journal of Managing Information Technology (IJMIT) Vol5 (2013): 1-12.

[27] Syiam, M. M., Fayed, Z. T., & Habib, M. B. (2006). An intelligent system for Arabic text categorization. International Journal of Intelligent Computing and Information Sciences, 6(1), 1-19.

[28] Taghva, Kazem, Rania Elkhoury, and Jeffrey Coombs. "Arabic stemming without a root dictionary." *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*. Vol. 1. IEEE, 2005

[29] Thabet, N. (2004, August). Stemming the Qur'an. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages* (pp. 85-88). Association for Computational Linguistics.