Journal Computer Science

Original Research Paper

Novel Prefix Tri-Literal Word Analyser: Rule-Based Approach

[1]Mohammed M. Abu Shquier and [2]Khaled M. Alhawiti

[1]Department of Information Science, University of Tabuk, Tabuk, KSA
[2]Department of Computer Science, University of Tabuk, Tabuk, KSA

Article history
Received: 31-07-2014
Revised: 29-11-2014
Accepted: 17-5-2015

Corresponding Author:
Mohammed M. Abu Shquier
Department of Information Science, University of Tabuk, Tabuk, KSA
Email: shquier@gmail.com

Abstract: Arabic stemming is a technique to find the stem or lexical root for Arabic words through the process of eliminating affixes (preffixes, infixes and suffixes) attached to their roots. Several approaches have been implemented to generate the stem of Arabic words according to a certain level of analysis, i.e., root-based approach, stem-based approach and statistical approach. Arabic language is a Semitic language which means that it is a derivational rather than a concatinative language. In this study we designed and implemented an Arabic triliteral Morphological Analyser that is capable of analysing the classical and Modern Standard Arabic (MSA) effectively with the capability of analysing vowelised, semi-vowelised and nonvowelised text. The system is integratable with other applications so that vast number of people can get benfited from. One shortcomming for the developed system is that the output obtained from the morphological analyser may contain several alternative solutions which leads to extraction ambiguity.

Keywords: Morphological Analyser, Stemmer, Semitic, Hamzated, Doubled, Hollow, Defective, Roots, Stems

Introduction

Arabic verbs are constructed on the root فعـل that uses three consonants ف, ع and ل that is know by Arabic grammarians as Morphological Balance (MB), the result of mapping root letters to MB forms is verbal or nominal stems. The stem is used to construct verbs or nouns through prefixing and suffixing inflectional prefixes and suffixes to those stems (Attia, 2008). The Arabic three consonants in the root-verb (فعـل) are represented as (C_1), (C_2) and (C_3) respectively, while the supscript followed the consonant represents the sequence of these consonants However, the multifarious vowels and affixes are attached to the root verbs to create the desired inflection of the meaning. Each root can generate a vast number of meanings. Arabic roots can be classified into two classes as shown in Fig. 2; the vowelized roots and non-vowelized Roots (Al-Omari, 1995; Al-Dahdah, 1985). This classification was made in accordance with the availability of the Arabic vowels in the roots.

The previous studies in the Arabic language research explained that the greater portion of the Arabic root verbs are of trilateral origin, while the remaining are of quadlateral and biliteral origin (Al-Fedaghi and Al-Sadoun, 1990). Arabic language plays a crucial role with the root ($C_1aC_2aC_3a$) (To clarify the structure of Morphological forms we have used the corresponding CV array of each form alongside. C_ns corresponds to radical letters and represent the consonants of فعـل) to add subtle variations to the meaning.

Arabic is considered as one of the Semitic language based on roots. A root is the original form of a word which can not be further analysed. Arabic roots are verbs only. The majority of Arabic roots are triliteral (George, 1990; Al-Najem, 1998; Al-Momani, 2010). Al-Fedaghi and Al-Anzi (1989) claimed that there are around ten thousand independent roots. Each root may be attached to prefixes, suffixes, infixes to derive nouns and adjectives. The addition of infixes is based on certain structurers. Words constructed from the same root are not related semantically in general (Rafea and Shaalan, 1993).

Stemmer or morphological analyser are widely used by researchers dealing with languages with complicated Many challenges may face the construction of well

guided Arabic rule-based stemmers, it is worthy stressing to mention some of these difficulties; the existence of irregular/broken plurals جمع رىالتكـــس., i'lal الإعلال and ibdal الإبدال, the huge number of Arabic roots, differentiating between affixes and original letters is ambiguous, un-vocalised Arabic representation, the existance of and the semantic ambiguity is also another challenge to the Arabic stemmers.

An affix is a morpheme that can be added before or after, or inserted within a root or a stem as a prefix, suffix or infix, respectively, to form new words or meanings (Al-Khuli, 1991; Thalouth and Al-Dannan, 1987). Arabic prefixes and suffixes are sets of letters and articles attached to the beginning and the end of the lexical word and written as part of it respectively (Al-Atram, 1990). English has 75 prefixes and about 250 suffixes (Salton, 1989). Arabic has fewer affixes to concatenate with each other in predefined linguistic rules. This feature increases the overall number of affixes (Ali, 1992). The removal of prefixes in English requires further analysis since it can alter the meaning or grammatical function of the word. This is not the case in Arabic, since the removal of prefixes does not usually reverse the meaning of words.

Literature Review

Several methods were developed to represent text in Natural Language Processing (NLP) and Information Retrieval (IR) fields. For Arabic Language, there are three different Stemming approaches: The root-based approach (Khoja and Garside, 1999); the light stemmer approach (Larkey *et al*., 2002) and the statistical stemmer approach (N-Gram (Khreisat, 2006; Mustafa and Al-Radaideh, 2004)).

Al-Shammari (2010) stated that both Arabic root-based and stem-based algorithms are lacking from generating errors. The removal of prefixes and suffixes generate many errors, especially when the algorithm is expected to distinguish between an extra letter and a root letter. Al-Shammari claimned that stemming process can return with errors known as over-stemming and under-stemming respectively.

Hawas (2013) presented a novel Arabic words root-extraction approach, he tried to assign a unique root for each Arabic word without having an Arabic roots list, a words patterns list, or even the Arabic words prefixes/suffixes list, his algorithm predicts the letters positions using rules based on the relations between the Arabic word letters and their position in the word. The proposed approach was composed of several corporate modules. Hawas tested the proposed approach using the Holy Quran words and he claimed that the total success ratio for the proposed algorithm was about 93.7% but she considered the root is correct if it has one correct letter.

Boudlal *et al*. (2011) provided a new way to find the system that assigns, for every non-vowel word a unique root depending on the context of the word on the sentence. The proposed system is composed of two modules. These modules start by segmenting the words of the sentence into its elementary morphological units in order to identify its possible roots.

Momani and Faraj (2007) proposed a novel algorithm to extract triliteral Arabic roots. The first step of their algorithm was to eliminate the stop words and then the prefixes and suffixes of each word are removed until only three letters remained. Finally, the remaining letters are arranged according to their order in the original word, which form the root of the original word. The researchers tested their algorithm on two types of Arabic text documents. The researchers claimed that the results of both runs were very promising and satisfactory enough to score over 73% of accuracy.

Khoja's stemmer is a root-based Arabic stemmer (Khoja and Garside, 1999). The Khoja's algorithm removes prefixes, infixes and suffixes and uses patterns to extract the roots using a dictionary. Although the algorithm suffers from some issues with proper nouns, broken plurals التكسير جمع and nouns, the Khojas algorithm showed superiority over previous work in root detection algorithms (Khoja and Garside, 1999).

In this study we propose an algorithm for word analyser that accepts the non-article trilateral words and finds out their roots. The word analyser module is shown in Fig. 1.

The word analyser process starts with the prefix/suffix analyser modules that determine whether the particular word is preceded by prefix(es) or attached with suffix(es) or not. The output of this module is the longest prefix/suffix list generaed, then we further invoke the stem generator module that generates all the permutations of the possible stems and then matches template(s) that represent the corresponding stem(s). Afterward, the triliteral root processor recodes the generated root to their original form.

Overview of Arabic Affixation

Essentially, the Arabic word can be described (Abu Shquier, 2013) as follows:

[Prefix1] [prefix1] stem [infixes] suffix1] [suffix2]

The stem is the minimal meaning-bearing unit in a language. Affixes in Arabic can be categorized into three types, prefixes, suffixes (or postfixes) and infixes (Saliba and Al-Dannan, 1990). The prefixes are added at beginning of the stem while the suffixes are attached to the end, Table 1 shows some affix conjugation for the verb ضـرب.

DOI: 10.3844/jcssp.2015.627.638

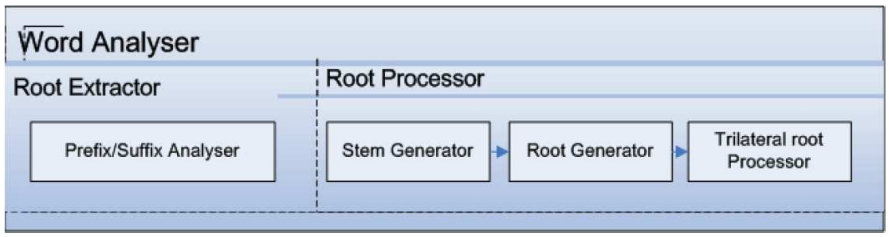

Fig. 1. Arabic word analyser module

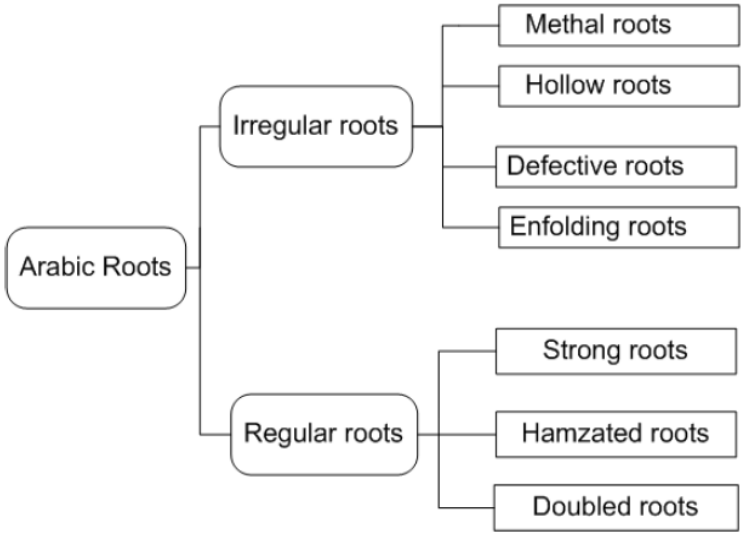

Fig. 2. Arabic roots classification

Table 1. An Arabic affixes example adopted from (Abu Shquier, 2013)

Suffixes	stem	Infex	Prefixes	Arabic	Structure
	ضرب			ضـرب	$C_1aC_2aC_3$
	ضرب	ي		يضرِبُ	$yaC_1C_2iC_3u$
ا	ضرب	ي		يضارِب	$yuC_1AC_2iC_3u$
هم	ضرب	ي		يضرِبهم	$yaC_1C_2iC_3uhum$
هم	ضرب	ي	س	سيضرِبهم	$syaC_1C_2iC_3uhum$
اهم	ضرب	ي	س	ضـاىس رِبهم	$syuC_1AC_2iC_3uhum$
اونهم	ضرب	ي	س	سيضارِبونهم	$syuC_1AC_2iC_3unahum$

Table 2. Arabic Suffixes for the regular verb hit ضـرب adopted from (Abu Shquier, 2013)

Suffix	Person, Gender and Number Features	Suffix Category	Example	Transliteration
نـي	(3rd, N, S)	Verb	ضـــربني	drabny
ك	(2nd, N, S)	Verb and Noun	ضـــربك	drabk
ه	(3rd, M, S)	Verb and Noun	ضـــربه	drabhu
ها	(3rd, F, S)	Verb and Noun	ضـــربها	drabha
هم	(3rd, M, P)	Verb and Noun	ضـــربهم	drabhum
هنّ	(3rd, F, P)	Verb and Noun	ضربهنّ	drabhuna
هما	(3rd, N, D)	Verb and Noun	ضـــربهما	drabahuma
كـم	(2nd, M, P)	Verb and Noun	ضـــربكم	drabakum
كنّ	(2nd, F, P)	Verb and Noun	ضربكنّ	drabkunna
كمـا	(2nd, N, D)	Verb and Noun	ضـــربكما	drabakuma

1st: Denotes the 1st Person (المتكلـــم)
2nd: Denotes the 2nd Person (المخاطـب)
3rd: Denotes the 3rd Person (الغائـــب)
S,D,P: Denotes Number features, Singularity, Duality and Plurality
M,N,F: Denotes Gender features, Masculine, Nuteral and Feminine

Table 3. Morphological balance significant derived forms

Form	Structure	Arabic balance
Form I	$yaC_1C_2aC_3$	يفعَل
Form II	$yuC_1aC_2C_2iC_3$	يفعِّل
Form III	$yuC_1aC_2iC_3$	يُفاعِل
Form IV	$yuC_1C_2iC_3$	يُفعِل
Form V	$ytaC_1aC_2C_2aC_3$	يتفعّل
Form VI	$ytaC_1AC_2aC_3$	يَتفاعَل
Form VII	$yanC_1aC_2iC_3$	يَنفَعِل
Form VIII	$yaC_1taC_2iC_3$	لِيفتَعِ
Form IX	$yaC_1taC_2iC_3$	يَستَفعِل

(1) Forms II and IV can have the meaning of carrying out an action to someone/something else
(2) Forms II and IV are making the verb transitive or causative
(3) Form II can also give a verb the meaning of doing something intensively and/or repeatedly
(4) Form III often carries the meaning of doing something with someone else: Or the meaning of trying to do something (Wightwick and Gaffar, 2007)

Suffixes in Arabic can be categorized into two basic categories, the suffixes that are attached to the verbs and the suffixes that are added to the nouns (Yusif, 2007). Furthermore, some of the suffixes can be attached to both the noun and verb stem. Nevertheless, Arabic permits the use of up to three suffixes simultaneously to be attached to the end of the same stem (Abu-Ata, 2001). Furthermore, Arabic words are built from roots rather than stems and involve diacritization. Written Arabic is also characterized by the inconsistent and irregular use of punctuation marks (Attia, 2008). Table 2 presents a wide range of suffixes example for the verb hit (ضـرب).

Arabic language plays a crucial role with the root ($C_1aC_2aC_3a$) (To clarify the structure of Morphological forms we have used the corresponding CV array of each form alongside. C_ns corresponds to radical letters and represent the consonants of ل)فـعـ) to add subtle variations to the meaning. There are nine significant derived forms (for the singular masculine 3rd person in the present tense) as shown in Table 3:

Arabic Roots Classification

Arab grammarians Al-Dahdah (1985) classifies Arabic roots as shown below in Fig. 2.

Regular roots: The non vowelized roots. This type of roots is sub-divided into the following categories:

- Strong roots: The root that contains neither vowels nor ء (*hamzah* and its second and third consonants are not identical, i.e., (لعـب ، ضـرب ، سـرع)
- Hamzated roots: The root that contain ء (hamzah i.e., (أخذ ، سـئل ، بـرأ)
- Doubled roots: the root in which its second and third consonant are identical i.e., (مرر، عدد ، مدد)

Irregular roots: The vowelized roots. This type of roots is classified into four types depending on which of the root letters is affected:

- Roots with ’waaw’ or ’yaa’ as the first root letter (Mithal roots) (االجـذر المثـال)
- Roots with ’waaw’ or’ yaa’ as the second root letter (Hollow roots) (الجذر الأجوف)
- Roots with ’weaw’ or ’yaa’ as the third root letter (Defective roots) (الجـذر النـاقص)
- Roots that have two weak letters in their roots (Enfolding roots) (الجذر اللّفيف)

Enfolding roots are categorized into two groups; the first group has a middle and final weak original letters, while the second group has a first and final weak original letters:

- The first group enfolds the definitions of both hollow defective roots, yet it is always treated as a defective only and the middle weak letter is treated as if it were a regular letter i.e., (روي، عوي)
- The second group enfolds the definitions of both Mithal and defective roots. These roots get the dealing of both Mithal and Defective verbs together. i.e., (وقى، وعى)

These classifications are general. In our paper, we conduct more analysis for the roots since roots of the same category may act differently during the morphological process. For instance, the verb promised وعد will be changed to promise عدى in the present tense form, while the root facilitated سرى will be to facilitate سـريـى in the same derivational form. Thus, the roots classification takes into account the following considerations: **First:** The category of the root and **second:** The vowels that are involved in root formulation. During the morphological analysis, a word might be represented in many forms.

For example, the root قـول may have many derivational forms. Let us shed light on the generation of the hollow verb said for all person, gender and tenses with singularity, duality and plurality conjugational cases respectively as shown in Table 4.

Table 4. Derivation for the second root hollow-verb say قـولى adopted from (Abu Shquier, 2013; Abu Shquier *et al*., 2012)

	Singular			Dual			Plural		
Features	Past	Pres	Imp	Past	Pres	Imp	Past	Pres	Imp
1st-M	قلـت	أقـول		قلنـــا	نقـول		قلنـــا	نقـول	
1st-F	قلـت	أقـول		قلنـــا	نقـول		قلنـــا	نقـول	
2nd-M	قلتَ	تقـــول	قـل	قلتمـــا	تقـــولا	قـولا	قلتـــم	تقولـــون	قولـوا
2nd-F	قلتِ	تقولـــي	قولـي	قلتمـــا	تقـــولا	قـولا	قلتنّ	تقلنَ	قلنَ
3rd-M	قـال	قـولى		قـالا	قـولاى		قـالوا	قولـونى	
3rd-F	قالـت	تقـــول		قالتـــا	تقـــولا		قلنَ	يقلنَ	

Table 5. Arabic roots representation

Form	1st Letter	2nd Letter	3rd Letter	Example
XXX	X	X	X	كتـــب
XVX	X	و	X	كـون
XVX	X	ي	X	رىط
XXV	X	X	ؤ	بطـؤ
VXX	و	X	X	وجد
VXX	ي	X	X	سـرى
HXX	ء	X	X	كـلأ

X: Denotes the non vowel character, no (ي, و, ا) characters
V: Denotes the vowel (ي, و, ا) character
H: Denotes the Z character, i.e., (أ، ؤ، ئ)

From Table 4 above we can conclude that verbs of the form C_1awaC_3 have the perfective stem patterns C_1uC_3 and C_1uwC_3 and the imperfective stem patterns C_1uC_3 and C_1uwC_3. For example, qaAl قـال (from [qawal]) قـول to visit has the perfective qul قـل and qaAl قـال and the imperfective stems qul قـل and quwl. قـول E.g.: perfect: Qultu I said and qaAlat she said imperfect: Yaqulna they (fem) said قلـــنى and yaquwlu he says. one can conclude that based on the person, number and gender; hollow verbs are realised by two stems in both perfect (simple past) and imperfect tenses (simple present, simple future), one long and one short; the long stems occurs with a weak middle letter, while the long stem cause the middle letter to disappear. It is worth stressing at this point that the words that derived from roots contain ء (hamzah) i.e., (أ، ؤ،ء،ئ) as one of their consonants might also change during the morphological process. For instance, the word to take (S, M, 3rd) ؤخذى is derived from the root أخذ. In such cases, we consider all the other forms that might a root appears in, Table 5 categories the trilateral roots based on the position of ء (hamzah), vowel and non-vowel letters.

This classification will be very helpful in identifying the original root form during the analysis process. Table 5 illustrates a portion of the roots classification that we will adopt.

Arabic Prefix/Suffix Analyser

As a preprocess of the prefix/suffix analyser, we have to check whether a word is an article or not. However, when the word is not an article the system passes the word to the word analyser for further analysis.

Table 6. Arabic attachable prefixes

Prefix	Meaning	Prefix	Meaning
بـال	in the	ف	and, therefore
ال	the	فـس	then will
فـال	and the, therefore	ك	like, as
كـال	like the	ل	for, to
لـل	for the, to the	ول	and (for, to)
ولـل	and (for the, to the)	و	and
وال	and the	س	will
فبـــال	therefore in the	ولال	and for the
وبـال	and in the	فـب	and in, therefore
وكـال	and like the	وب	and in
ول	and will	فـل	and for, therefore
ول	and for and to	ب	in, with

Table 7. Arabic attachable prefixes

Prefix	Meaning	Prefix	Meaning
بـال	in the	ف	and, therefore
ال	the	فـس	then will
فـال	and the, therefore	ك	like, as
الك	like the	ل	for, to
لـل	for the, to the	ول	and (for, to)
ولـل	and (for the, to the)	و	and
وال	and the	س	will
فبـــال	therefore in the	ولال	and for the
وبـال	and in the	فـب	and in, therefore
وكـال	and like the	وب	and in
ول	and will	فـل	and for, therefore
ول	and for and to	ب	in, with

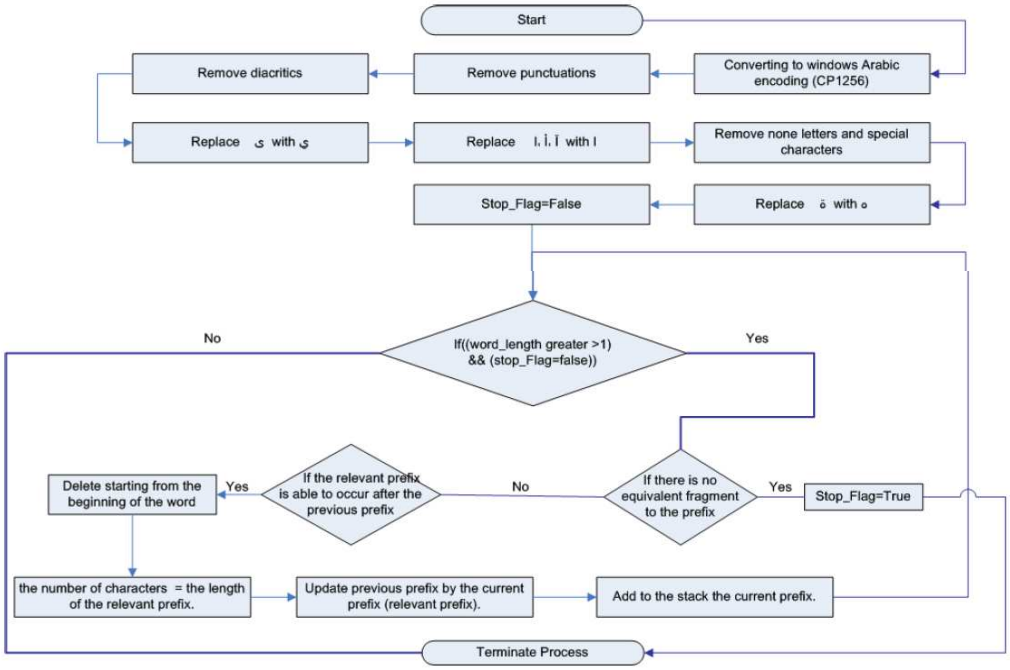

Fig. 3. Prefix extraction flowchart

This particular process starts by executing the prefix analyser module which determine whether a word is preceeded by prefix(es) or not. Prefixes with Arabic language form a closed list. Arabic allows up to three prefixes to precede the word within certain rules. Table 6 and 7 illustrates these prefixes with their associated meanings.

When the prefix analyser processes a word; it requires certain information to decide what to process and where to stop. Table 8 lists the prefixes and their corresponding combining rules based on Table 7.

Prefix Extraction Process

The prefix analyser starts after matching a certain word against a set of possible patterns to handle its prefix/suffix sequence ambiguity, then we start parsing the word from its beginning to extract the longest-possible-prefix, The process stops when there is no more prefix(es) left for extraction. The output of the prefix analyser will be stored in a separate file for further processing. In Arabic text, the analysis of the word is much more complicated. A word can be pronounced differently based on the chosen possible root, this proves that the absence of diacritics can result in ambiguities. Figure 3 represents the prefix extraction module, the module starts with converting the word to the Arabic encoding system, then we remove all punctuations, diacratics, non letters and the special characters, we continue to replace the hamzated letters, أ، ئ،ؤ with alif ا and replace the Alif Al-Maqsorah ى with ي and replace the ة Taa Al-Marbotah with ه; the remainder of the module is illustrated in the Fig. 3.

Table 8. Arabic prefixes joining RULES

Prefix	Meaning
ب	ال
ك	ال
ل	ب، ل
ف	ب، ل، ال ، و ، س
أ	ب، ك ، ل ، ف ، س
و	ب، ك ، ل ، ال ، و، س ، لـل ، لال

After determining the prefix/suffix that will be extracted, the analyser checks the entry of the previously extracted prefix/suffix to ensure that the order of the extracted prefix/suffix is correct, moreover, the stem generator finds a template that matches the proposed stem and then it checks if the extracted prefix is allowed to be concatenated with the generated stem by a certain template.

On the other hand, the suffix analyser parses the word from the end through the beginning of the word, bearing the following condition during the extraction process, first: The suffix has to match the comparable fragment of the word, second: The suffix has to suit the suffix representation of the CFG and third: The suffix should satisfy the prefix/suffix joining rules (Al-Omari, 1995; Abu Shqeer, 2002).

Suffixes can be attached to the end of the word according to certain rules. Table 9 represents a sample of the Arabic suffixes combining rules.

Suffix Extraction Process

This section presents the algorithm embedded in the suffix analyser module (Fig. 4). The algorithm expects a stream of characters as an input.

Table 9. Arabic suffix joining rules

Case	Suffix	Attachable Suffix
1	نى	ت ، ي
2	ان	ت ، ي
3	ات	ت ، ي ، و
4	ون	ت ، ي ، و
5	كم	ان ، ات ، ون ، ت ، ن ، ا ، و
6	هم	ان ، ات ، ون ، ت ، ن ، ي ، ا ، و
7	ن	ان ، ات ، نو ، كم ،هم ، تم ، ت ، ا ، ك ، ه ، و
8	ي	تا ، ت ، ن
9	أ	كم ،هم ، تم، ت ، ا ، ه
10	ك	ان ، ات ، نو ، ت ، ن ، ا ، و
11	ه	ان ، ات ، نو ، ت ، ن ، ي ، ا ، و
12	و	كم ،هم ، تم، ت

Table 10. Generated triliteral roots representation

Case	Form	Case	Form	Case	Form
1	أعل	2	ؤعل	3	فؤل
4	أعا	5	ؤال	6	فؤء
7	أعو	8	وؤل	9	لىف
10	أعي	11	ؤلى	12	ئىف
13	ىأع	14	ؤوي	15	يىف
16	األ	17	أوؤ	18	ئىف
19	أول	20	إعل	21	يىف
22	إعو	23	اىف	24	ألى
25	اى ل	26	فأل	27	وأا
28	ىوع	29	فال	30	أىف
31	وعي	32	فاء	33	فعأ
34	وعأ	35	فول	36	فعا
37	وعئ	38	فوي	39	ىفع
40	وعل	41	فعئ	42	فاو
43	فعي	44	لىو	45	فوء
46	فعو	47	على	48	فوأ
49	فع ء	50	ىعى	51	فؤو
52	فعؤ	53	ولى	54	فؤ
55	فعإ	56	الى	57	فئل
58	فئي	59	أول	60	فعل

It produces a list of parameters which express the extracted suffixes. After the extraction of prefixes and suffixes, the remaining part of the word obtained is called the stem. Table 4 exhibts the procedures of extracting the suffix from a certian word.

Notice that *P+1* means the number of possible prefixes including the null prefix and *S+1* denotes the number of possible suffixes including the null suffix. Due to the possibility of the improper prefixes/suffixes extraction. The morphological analyser should be smart enough to generate all possible stems as well as the joining rules of prefixes and templates.

Arabic Roots

The Arabic roots can be classified into two classes; the Vowelized roots and Non-Vowelized Roots (Al-Dahdah, 1985). This classification was made in accordance with the availability of the Arabicvowels in the roots.

The root extraction process matches the stem with the corresponding template based on the verb ($C_1aC_2aC_3a$) فعل. The system will recode the root and then decide whether it is a correct not. An enhanced structure of the Arabic words has been shown in Fig. 5; For example, the word ضربونىف can be simplified to the following components: Prefixes ف root prefixes ي root ضرب (no embedded infix), suffixes ون there is no root suffixes for the word ضربونىف.

Generation of Arabic Roots

The root generation algorithm expects three arguments as input: Prefix, suffix and stem. The algorithm finds all the template(s) that are related to the stem according to the rules mentioned in Table 9.

As shown in Fig. 6, the root generation process aims to find a template that can represent the stem under certain conditions, first: Both of the template and the stem must be of the same length. Second: The template must be a valid form for the extracted possible prefix and Third: The template is attachable to the associated possible suffix (Al-Omari, 1995).

Triliteral Root Processor

The three letters root processor aims to refer the generated root to their original root form (Arabic Orthography). Previously, we classify the roots according to two characteristics. First: The positions of the vowels and ء (*hamzah*). Second: The vowels and the forms of the written ء (*hamzah*) which are involved in the formulation of the root. Here, we use these classifications to recode the root to its original root form, however, regular root الجذرالسالم is the only type of root that need not any recoding process since it does not contain any vowel or ء (*hamzah*). Furthermore, in some cases, a vowel might be converted to a non vowel which cause the root to be recoded.

Table 10 shows the generated trilliteral root representation, a special recoding process is conducted for each form listed below: We have used the Morphological Balance (MB) ($C_1aC_2aC_3a$) for all the form representation, the Arabic three consonants ف, ع and ل in the root-verb (فعل) are represented as (C_1), (C_2) and (C_3) respectively, however, vowels and hamzah (ؤ ، ئ ، أ ، ى، ئـ ، ا ، و ، ي، ء) have replaced their corresponding consonents ف , ع and ل in the root-verb (فعل). For each form represented in Table 10, there is a corresponding recoding process implemented, we will discuss the usage of Table 10 throughout the following examples. Let us take the word أنضرب as an example. There would be two possible stems for this word إنضرب i.e., (بنضر) and إضرب).

DOI: 10.3844/jcssp.2015.627.638

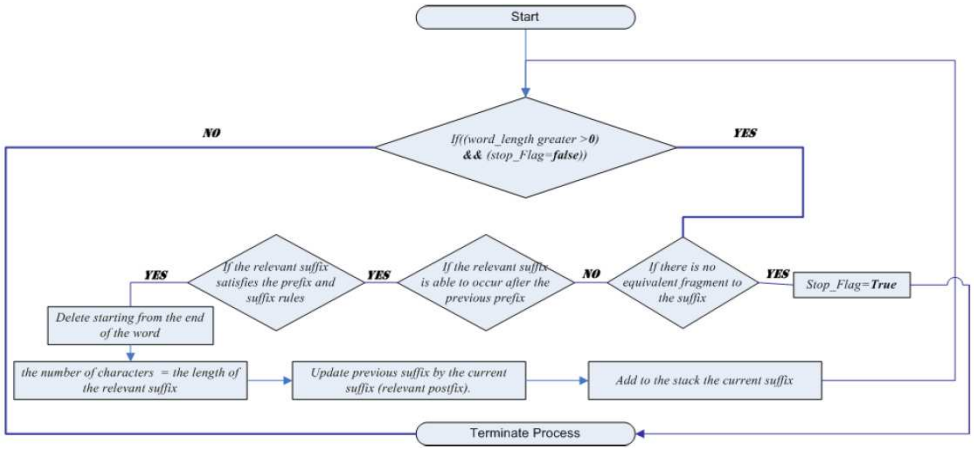

Fig. 4. Suffix extraction flowchart

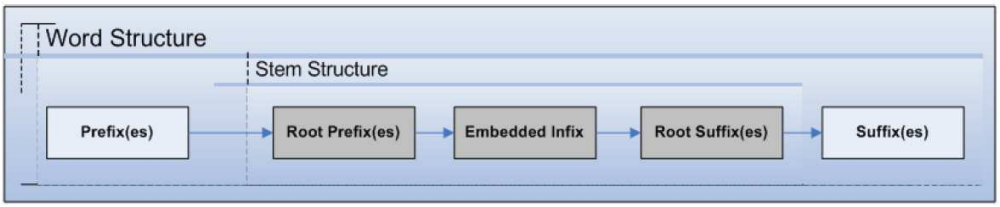

Fig. 5. Enhanced Structure of Arabic word

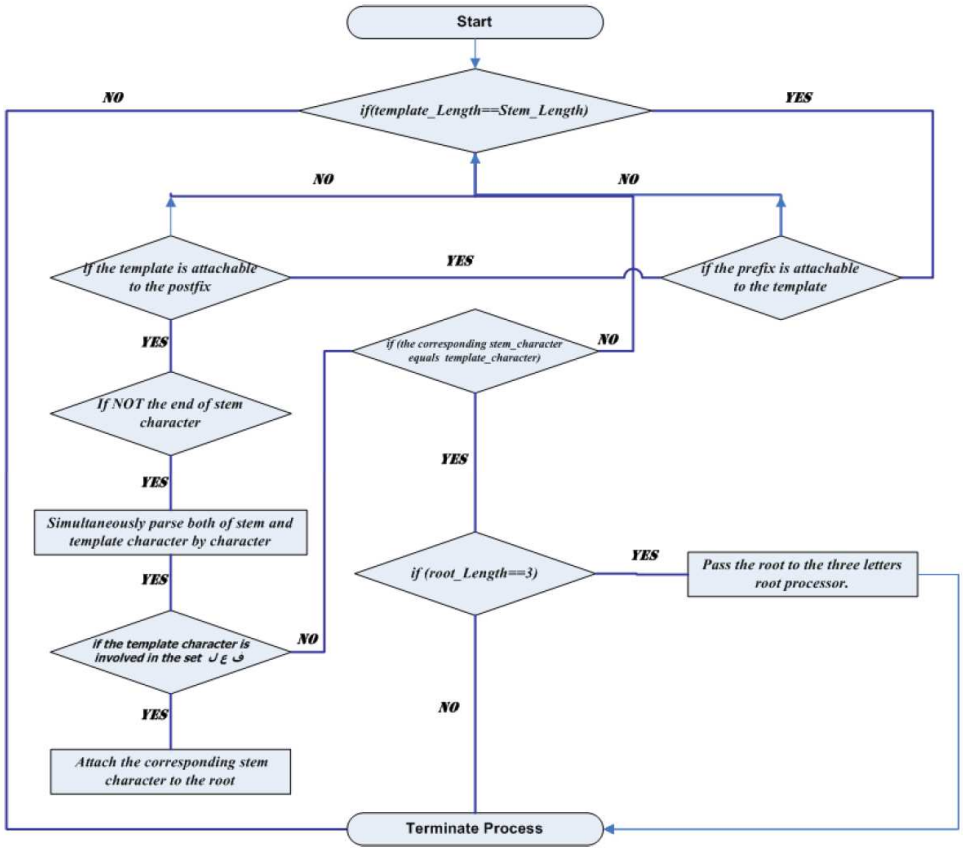

Fig. 6. Root generation flowchart

The recoding process of إنضـــرب is shown below:

- Input word: نضـــرب after removing the prefix
- Prefix: إ
- Stem: نضـــرب
- Template Form: نفعـــل
- Generated Root: ضـرب
- Recoded Root (1): ضـرب

As presented above the stem will be analysed and the root ضـرب will be generated. The root will remain as it is during the recoding process.

We may have another result of the word إنضـــرب to be analysed as follows:

- Stem: إنضـــرب
- Template Form: إ نفعـــل past present of the verb نفعـــلى of yanC$_1$aC$_2$iC$_3$ Table 3 - Form 7

The second result should be discarded since the word is not used in Arabic despite the correct analysis of the word.

Method Limitation

When the system is integrated with some applications like Machine Translation (MT) where the template affects the Part of Speech (POS) (Part of Speech (POS) is the method of classification of words according to their meaning, functions and categories such as noun, verb and adjective. The POS tagging occurs during the Syntactic Analysis phase and it involves assigning of words to their proper part-of-speech tag), in this case the generation of the correct root leads to correct solution, however, in some cases of our method where a particular templates starts with a character that is considered as a prefix. i.e., if the template أفعـــل was used to derive the word أكـــبر, the analyzer will consider the character أ as a prefix and produces the root كـــبر that matches the template فعـــل and that cause ambiguity, however, such issues occur when there are more than one correct analysis for a particular word, in other cases we may obtain three correct roots with respect to the morphological process, while semantically, one of them only is correct.

Experiments and Results

In this section we will be testing the performance of the developed system, we will not be able to conduct a precise evaluation of the system, since the system has not yet been integrated with any other system. However, the test will help in understanding the capabilities of the system better. The test data is taken from one poem تعـــب كلهـــا ةاىالـح for Abu Elalaa Al أبوالعلاءالمعـــــرى, which contains 641 tokens. Figure 7 shows a pie chart for the breakdown of articles and words in the text.

The proposed testing technique of the developed system consists of two main steps to evaluate the performance of the morphological analyser:

- Neither using the roots dictionary nor the root decision table
- With using roots dictionary but not the roots decision table
- With using both the roots dictionary and the roots decision table

A. *The First Test*

In this test, the system is used to process the text using neither dictionary nor the root decision table. However, the system was not able to return the correct analyses of the triliteral words.

After removing the 94 article of the test data, 547 words remaining. In this test the number of analyses returned is 1034 with only 345 correct analysis. Figure 8 shows the percentage of errors obtained from the first test.

The absence of the roots dictionary and the roots decision table are the main reasons behind this result. Another reason might be due to the type of the texts. The texts that contain less vowelized roots will have smaller percentage of errors since vowelized-derived words may have more analyses. Therefore, this factor should be taken into consideration in the evaluation of the system. To reduce the errors we may need the roots dictionary and the roots decision table. Figure 9 shows the analysis of the factors affecting the result.

As shown above, most of the errors occur due to the absence of the roots dictionary. Some of these errors can also be due to the morphological rules of the system which can be reduced when applying a roots dictionary. Three percent of the errors returned as a result of the misuse of the morphological rules. These rules can be reconstructed to eliminate this percentage. Ten percent of the errors are due to the absence of the roots decision table. The correct roots obtained from this test can be classified into two categories as follows:

- Exact root: This occurs when there is only one analysis for a given word. For example, from the word حىقـب we will obtain the root قبـــح from the system
- Ambiguous root: This occurs when there is more than one correct analysis for a particular word. For example, from the word كـنى the system will return three different roots. i.e., (كـان ، كـون and كـن). These roots are all correct with respect to the morphological process, while it is only one correct root when considering semantics. Figure 10 shows the analysis of the correct results obtained from the first test.

DOI: 10.3844/jcssp.2015.627.638

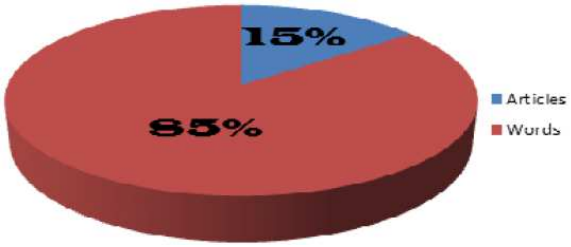

Fig. 7. Words and articles in poem تعـب كلهـا ةاىالـح

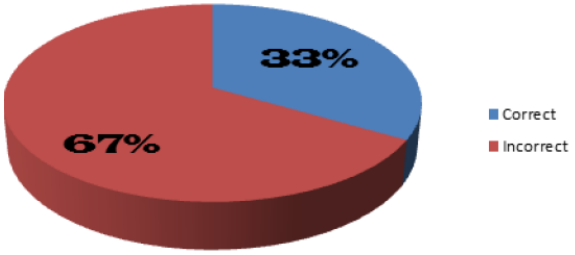

Fig. 8. Percentage of errors retruned from the first test

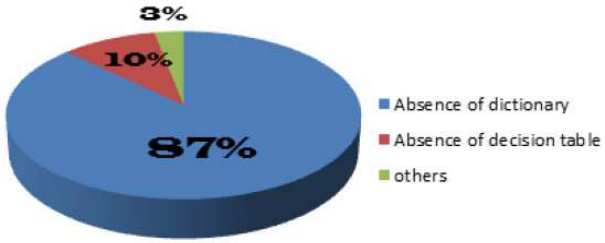

Fig. 9. Analysis of Errors in the first test

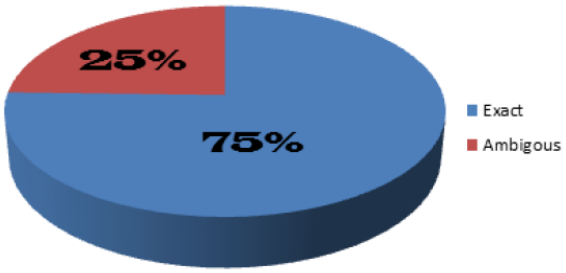

Fig. 10. Analysis of the correct results

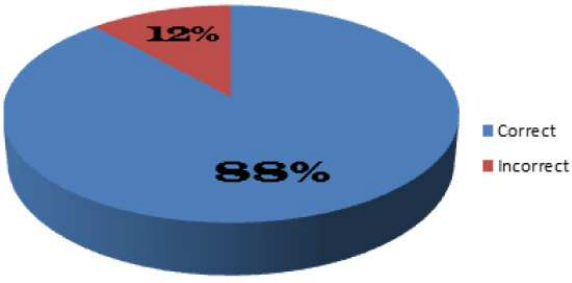

Fig. 11. Percentage of Errors returned from the second test

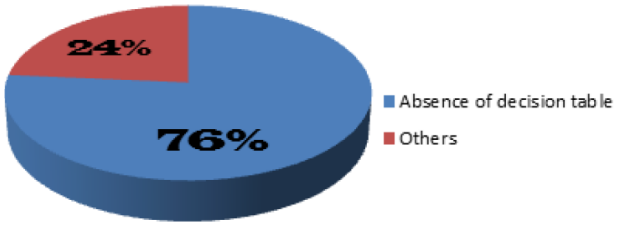

Fig. 12. Errors analysis for the second test

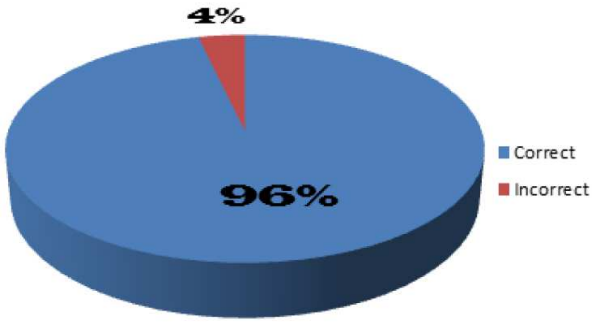

Fig. 13. Percentage of Errors returned from the Third test

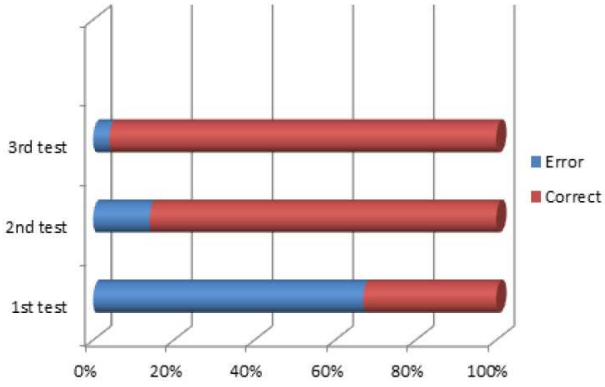

Fig. 14. Experiment results

The ambiguous analysis can be due to the following factors:

- The root types
- The Proper usage of the template: Templates that starts with character that can be considered as a prefix. For example, if the template أفعـل was used to derive the word doubled أقرّ, the analyzer will consider the character أ as a prefix and produces the root قرّ with the template فعّ which will be matching the template فعـل after the separation of the doubled letter. Since the system will be integrating with other application, such as machine translation, the determination of the correct root is the main part of the correct solution.

On the other hand, the system has rejected some words due to different reasons these are:

- The word was derived from quadrilateral root (e.g., مانىسل)
- The words which have no Arabic root (e.g., الهواء)
- The words written in different way because of letter-dropping grammars (e.g., صاح) which is originally came from (ىصاحب)
- some other morphological rules that are not manipulated in the system

The Second Test

This test considers that there is a roots dictionary as a component of the morphological analyzer. Figure 11 shows the percentage of errors encountered in this test. As we can see, the percentage of errors has been reduced from 67% to 12%. This emphasizes the urgent need for the roots dictionary. The error occurred in this test can be reduced further if the roots decision table is included in the system (as we will see in the third test). Figure 12 shows the analysis of the errors encountered in the second test.

The Third Test

This test has been carried out manually since the roots decision table is not yet available. This test proves that the errors that occurred in the second test can be reduced. Figure 13 shows errors that has been eliminated to 4%.

Conclusion

Stemmers and word analysers usually help in resolving the lexical ambiguity, The goal of this paper is to develop a stemmer for the triliteral words of Arabic Language. However, as we analysed Arabic morphology deeply, we realised that the problem is not just a matter of truncating affixes to obtain the stem; the analysis requires heavy computational processes and the usage of large amount of information; on the other hand, the system might be used as an Arabic morphological analyser for general domain since the database can be updated to cover all the Arabic trilateral roots. The three conducted tests prove that the morphological rules used in the system has reduced the errors by 94% when using both the roots dictionary and the roots decision table that we implement. Figure 14 shows a Bar chart comparing the three tests results. In fact, building practical stemmers or morphological analysers requires fully understanding of the language morphology structure. To enhance the output of the morphological analyser, we recommed conducting the following steps: Reducing the rules number and increasing language coveragance while keeping the same level of performance and functionality. Merging rules is very helpful for enhancing the pattern-based stemmer. At present, designing a fully-automated Arabic morphological analyser might not be possible. Instead, analysers should be application-oriented or for specific domain.

Funding Information

The research has been self sponsored.

Author's Contributions

The author's contributions played a significant role in the following categories as shown below:

Abu Shquier: Conception, design, analysis and Interpretation. Final approval of the article. Statistical analysis and Overall responsibility.

Alhawiti: Data collection and Critical revision of the article and obtained funding.

Ethics

The manuscript has not been previously published or accepted for publication elsewhere, either in whole (including book chapters) or in part (including paragraphs of text or exhibits), whether in English or another language.

References

Abu-Ata, B., 2001. An Arabic stemming algorithm on ERA for information retrieval. PhD. Thesis, Universiti Kebangsaan Malaysia.

Abu Shqeer, O., 2002. Handling agreement and word reordering in English to Arabic machine translation. Master Thesis, University of Sains Malaysia.

Abu Shquier, M. and O. Abu Shqeer, 2012. Words ordering and corresponding verb-subject agreements in English-Arabic machine translation: Hybrid-based approach. Int. J. Soft Comput. Software Eng., 2: 49-60. DOI: 10.7321/jscse.v2.n8.5

Abu Shquier, M.M., 2013. Computational approach to the derivation and inflection of Arabic irregular verbs in English-Arabic machine translation. Int. J. Advance. Comput. Technol., 5: 1-21.

Al-Atram, M.A., 1990. Effectiveness of natural language in indexing and retrieving Arabic documents [in Arabic] (King Abdulaziz City for Science and Technology Project number AR-8-47). Riyadh, Saudi Arabia.

Al-Dahdah, A., 1985. Arabic language grammar dictionary. Lebanon Library, Lebanon.

Al-Fedaghi, S.S. and F. Al-Anzi, 1989. A new algorithm to generate Arabic root-pattern forms. Proceedings of the 11th National Computer Conference and Exhibition, (CCE 89), pp: 391-400.

Al-Fedaghi, S.S. and H.B. Al-Sadoun, 1990. Morphological compression of Arabic text. Inform. Proc. Manage., 26: 303-316. DOI: 10.1016/0306-4573(90)90033-X

Ali, N., 1992. Parsing and automatic diacritization of written Arabic: A breakthrough. Proceedings of the 13th National Computer Conference (NCC' 92), King Abdul-Aziz City for Science and Technology, Riyadh, KSA, pp: 794-812.

Al-Khuli, M., 1991. A dictionary of theoretical linguistics: English-Arabic with an Arabic-English glossary. Library of Lebanon.

Al-Momani, I., 2010. Does the VP node exist in modern standard Arabic?. J. Langu. Literature, 2: 76-76.

Al-Omari, H., 1995. ALMAS: An Arabic language morphological analyzer system. Malaysian J. Comput. Sci., 8: 30-50.

Al-Shammari, E.T. 2010. Improving Arabic text processing via stemming with application to text mining and web retrieval. PhD Thesis, George Mason University, USA.

Attia, M., 2008. Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation. PhD Thesis, University of Manchester.

Boudlal, A., R. Belahbib, A. Belahbib and A. Mazroui, 2011. A markovian approach for Arabic root extraction. Int. Arab J. Inform., 8: 91-98.

George, M., 1990. Al Khaleel: A dictionary of Arabic syntax terms. Beirut: Library of Lebanon.

Hawas, F.A., 2013. Exploit relations between the word letters and their placement in the word for Arabic root extraction. Comput. Sci., 14: 327-431. DOI: 10.7494/csci.2013.14.2.327

Khoja, S. and R. Garside, 1999. Stemming Arabic text. Lancaster, UK, Department of Computing, Lancaster University.

Khreisat, L., 2006. Arabic text classification using N-Gram frequency statistics a comparative study. Proceedings of the International Conference on Data Mining (CDM' 06), Las Vegas, NV: USCCM, pp: 7882.

Larkey, L., L. Ballesteros and M.E. Connell, 2002. Improving stemming for Arabic information retrieval: Light Stemming and co-occurrence analysis. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 11-15, ACM, New York, USA, pp: 275-282. DOI: 10.1145/564376.564425

Momani, M. and J. Faraj, 2007. A novel algorithm to extract tri-literal Arabic roots. Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications. May 13-16, IEEE Xplore Press, Amman, pp: 309-315. DOI: 10.1109/AICCSA.2007.370899

Mustafa, S.H. and Q.A. Al-Radaideh, 2004. Using n-grams for Arabic text searching. J. Am. Society Inform. Sci. Technol., 55: 1002-1007. DOI: 10.1002/asi.20051

Rafea, A.A. and K.F. Shaalan, 1993. Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network. Software Practice Experience, 23: 567-588. DOI: 10.1002/spe.4380230602

Saliba, B. and A. Al-Dannan, 1990. Automatic morphological analysis of Arabic: A study of content word analysis. Proceedings of the First Kuwait Computer Conference (KCC' 90), pp: 231-243.

Salton, G., 1989. Automatic text processing: The transformation, analysis and retrieval of information by computer. Addison-Wesley, Reading, ISBN-10: 0201122278, pp: 530.

Thalouth, B. and A. Al-Dannan, 1987. A comprehensive Arabic morphological analyzer/generator. IBM Kuwait Scientific Center, Kuwait.

Wightwick, J. and M. Gaffar, 2007. Arabic Verbs and Essintials of Grammar. 2nd Edn. McGraw-Hill, New York, ISBN-10: 0071596038, pp: 160.

Yusif, J., 2007. Automatic part of speech tagger for Arabic language using neural network. PhD Thesis, National University of Malaysia.