Arabic Information Extraction Methods: A Survey

Mazen El Sayed, George Lebbos & Haissam Hajjar

Lebanese University

ABSTRACT

The IR systems developed for western languages, such as English, have high performances when used in their own languages, but they don't have this same performance when used for eastern languages such as Arabic. This is due to the fact that the Arabic language has a different and complex structure and morphology: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short (diacritics) and long vowels. In addition, an Arabic word is derived from a root by concatenating some affixes based on regular set of word patterns. To address these problems, several methods have been proposed. The aim of this paper is to propose a survey of these methods. Although we not claim that this an exhaustive study, this work covers near 20 different methods. The main approaches applied in these methods are morphological or statistical analyses. To extract information from an Arabic document, the involved methods based on both approaches must answer the following question: "How can we find the root of the word we search". To find a word in an Arabic dictionary, first we must extract the root of this word and then find this root in the dictionary, due to the fact that the vocabulary of the Arabic language is essentially built from the roots derivation. The roots are words composed of three to five consonants letters. This work will contribute to the enhancement of the Arabic information retrieval system performance, due to the fact that Arabic information extraction methods are the kernel of such system.

Keywords: arabic langue; dictionary; information extraction; morphological analysis; n-gram; statistical analysis; stemmer.

Classification: FOR Code: 091599

Language: English

London Journals Press

LJP Copyright ID: 392852
Print ISSN: 2631-8474
Online ISSN: 2631-8482

London Journal of Engineering Research

Volume 19 | Issue 2 | Compilation 1.0

Arabic Information Extraction Methods: A Survey

Mazen El Sayed[α], George Lebbos[σ] & Haissam Hajjar[ρ]

ABSTRACT

The IR systems developed for western languages, such as English, have high performances when used in their own languages, but they don't have this same performance when used for eastern languages such as Arabic. This is due to the fact that the Arabic language has a different and complex structure and morphology: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short (diacritics) and long vowels. In addition, an Arabic word is derived from a root by concatenating some affixes based on regular set of word patterns. To address these problems, several methods have been proposed. The aim of this paper is to propose a survey of these methods. Although we not claim that this an exhaustive study, this work covers near 20 different methods. The main approaches applied in these methods are morphological or statistical analyses. To extract information from an Arabic document, the involved methods based on both approaches must answer the following question: "How can we find the root of the word we search". To find a word in an Arabic dictionary, first we must extract the root of this word and then find this root in the dictionary, due to the fact that the vocabulary of the Arabic language is essentially built from the roots derivation. The roots are words composed of three to five consonants letters. This work will contribute to the enhancement of the Arabic information retrieval system performance, due to the fact that Arabic information extraction methods are the kernel of such system.

Keywords: arabiclangue; dictionary; informa-tionextraction; morphologicalanalysis; n-gram; statistical analysis; stemmer.

Author α σ ρ: Computer sciences Lebanese university, faculty of technology Saida, Lebanon.

I. INTRODUCTION

Information retrieval (IR) is a communication process that relies on language to perform its functions [57]. All human languages have vocabularies, corpora of words whose elements constitute the building blocks from which meaningful communication constructions can be formed. If we consider of a document as a collection of words, then it is easy to contemplate the role played by the structure of a language in providing access to information within this document. Words are formed according to specific rules and guidelines that differ among languages, creating IR problems and potential solutions that need to be investigated with the language involved in mind [30, 45]. The performance of an IR system is mainly affected by the efficiency of the information extraction method which constitutes the kernel of these systems [8, 1, 23, 42]. The IR systems developed for western languages, such as english, have high performances when used in their own languages, but they don't have this same performance when used for eastern languages such as arabic.

Arabic language is used by more than 330 million arabic speakers that are spread over 22 countries [16, 24]. However, the performance of information retrieval in arabic language is very problematic due to the specific morphological and structural changes in the language: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short (diacritics) and long vowels, most of the arabic words contain affixes (Table 1, 2) [6, 10, 11, 21, 29, 49, 52, 59]. To address these problems, several methods have

been proposed. The main approaches applied in these methods are morphological or statistical analyses [7]. To find a word in an arabic dictionary, first we must extract the root of this word and then find this root in the dictionary [32]. This is because the vocabulary of the arabic language is essentially built from the roots derivation. The roots are words composed of three to five consonants letters. The arabic language has about ten thousand roots, 85% of them are trilateral. The derivation of words is done by adding affixes (prefix, infix, or suffix) to the root according to several patterns that are around 120 [9]. For example, let us take the root (كتب "ktb"); the words (مكتوب "makatob", كاتبة "kateba", كاتب "kateb") are respectively derived from this root according to the patterns (فاعل, فاعلة, مفعول) (Table 3). To extract information from an arabic document, the involved methods must answer the following question: "How can we find the root of the word we search".

The aim of this paper is to propose a survey of arabic information extraction methods. Although we not claim that this an exhaustive study, this work covers near 20 different methods. This work will contribute to the enhancement of the arabic IR systems performance.

Table 1: Arabic Diacritics and Letters Transcription. Empty Case Means No Writing Change in the Corresponding Letter And Position. X-case Means No Existing of the Corresponding Letter

Letter	Transcription	Writing		
		At Begin	*In Middle*	*At End*
ً	Tanween Fatha			
ٌ	Tanween Dama			
ٍ	Tanween Kasra			
َ	Fatha			
ُ	Dama			
ِ	Kasra			
ْ	Sokon			
ّ	Shedda			
~	Maada			
ء	Hamza			
ا	Alef			
إ	Alef with Hamza on bottom			
أ	Alef with Hamza on top			
آ	Alef with Maada			
ب	Baa	بـ	بـ	ب
ة	Taa Marbouta	X	X	X
ت	Taa	تـ	تـ	ت
ث	Tha	ثـ	ثـ	ث
	Jeem	جـ	جـ	ج
ح	H'a	حـ	حـ	ح
خ	Khaa	خـ	خـ	خ
ر	Raa	ر	ر	ر
ز	Thal	ز	ز	ز
س	Seen	سـ	سـ	س
ش	Sheen	شـ	شـ	ش
ص	Saad	صـ	صـ	ص
ض	Daad	ضـ	ضـ	ض
ط	T'aa	طـ	طـ	ط
ظ	Zha	ظـ	ظـ	ظ
ع	Ain	عـ	عـ	ع
غ	Jain	غـ	غـ	غ
ف	Faa	فـ	فـ	ف
ق	Qaf	قـ	قـ	ق
ك	Kaf	كـ	كـ	ك
ل	Lam	لـ	لـ	ل
م	Meem	مـ	مـ	م
ن	Noon	نـ	نـ	ن
ه	Haa	هـ	هـ	ه
و	Waw	و	و	و
ؤ	Hamza on waw	X	ؤ	ؤ
ى	Alif Makzora	X	X	
ي	Yaa	يـ	يـ	ي
ئ	Hamza on yaa	ئـ	ئـ	ئ

Table 2: Example Arabic Affix Transcription Cited In This Paper And Their Transcription

Affix	Transcription		Affix	Transcription
ال	Alef Alef Lam		يون	Yaa Waw Noon
ات	Alef Taa		ات	Alef Taa
ال	Alef Lam		الم	Alef Lam Meem
ان	Alef Noon		ان	Alef Noon
بال	Baa Alef Lam		با	Baa Alef
تم	Taa Meem		بال	Baa Alef Lam
دعي	Dal Ain Yaa		تان	Taa Alef Noon
سْ	Saa Sokon		تما	Taa Meem Alef
سال	Saa Alef Meem		تين	Taa Yaa Noon
عين	Ain Yaa Noon		سَ	Seen with Fatha
فال	Faa Alef Lam		سي	Seen Yaa
كال	Kaf Alef Lam		فا	Faa Alef
كن	Kaf Noon		كال	Kaf Alef Lam
لا	Lam Alef		كما	Kaf Meem Alef
لال	Lam Alef Lam		وال	Waw Alef Lam
لإ	Lam Alef with Hamza on bottom		وبال	Waw Baa Alef Lam
لل	Lam Lam		يية	Yaa Yaa Taa Marbouta
مال	Meem Alef Lam		ار	Alef Raa
مود	Meem Waw Dal		إس	Alef with Hamza on bottom Seen
ها	Haa Alef		تم	Taa Meem
هما	Haa Meem Alef		تمل	Taa Meem Lam
هما	Haa Meem Lam		را	Raa Alef
وال	Waw Alef Lam		ري	Raa Yaa
وب	Waw Baa		سُ	Seen with Dama
وت	Waw Taa		ست	Seem Taa
ودع	Waw Dal Ain		كا	Kaf Alef Lam
وس	Waw Seen		مر	Meem Raa
ولل	Waw Lam Lam		وا	Waw Alef
وم	Waw Meem		ول	Waw Lam
ون	Waw Noon		ون	Waw Noon
وي	Waw Yaa		ية	Yaa Taa Marbouta
ين	Yaa Noon		يه	Yaa Haa

Table 3: A Sample of Arabic Pattern Cited in this Paper and their Transcription

Pattern	Transcription		Pattern	Transcription
افتعل	Eftaala		مستفعل	Mostafeel
افعلل	Afaalal		مفعالة	Mafaala
أفعال	Afaal		مفعلل	Mafaalal
فاعلة	Faaela		استفعل	Iestafaal
فعول	Faaol		افعلال	Afaalal
مفعول	Mafool		تفاعل	Tafaool
افتعال	Efteaal		تفعلل	Tafaalal
متفعلل	Moftaeel		فاعل	Faeel

II. Arabic information extraction methods

To extract information from an arabic document, we must find the root of the word we search. To answer this problem, we can perform morphological or statistical analyses. For this reason, the methods that we describe in this work are based on one of these approaches (Fig. 1).

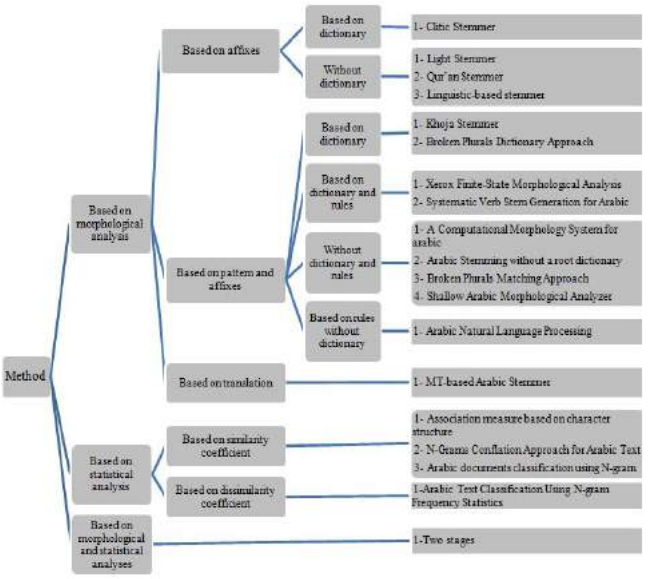

Fig. 1: Arabic information extraction methods

2.1 Common Specific Treatements

To treat arabic text, we must specify the text encoding and clean the arabic text. For that, two main approaches are applied in most arabic information extraction methods. These approaches are: normalization and transcription.

2.1.1 Normalization

The diacritics and the variation of the letter forms according to its positions take an important role in the arabic reading and writing complexity and reduce the Arabic Information Extraction methods performance. To resolve these problems, the normalization phase is applied before applying these methods, the text normalization

takes a character string as input and tries to remove or replace some characters under the predefined rules to converts it into a string of letters (Fig. 2). Every method has the specific rules, in general a text is normalized by removing (the tatweel character "-", the diacritics and the shedda " " (Table 1), the punctuations, the non letters, the stop words, the specials characters, and the numbers) and replacing ("إ" ,"أ" and "آ" by "ا" ,"ى" by "ي" at the end of the words, "ة" by "ه" at the end of the words, "يء" by "ؤ" ,"ئ" by "ء و", and ....)(Table 1) [17, 18, 22, 35, 40, 41, 42].

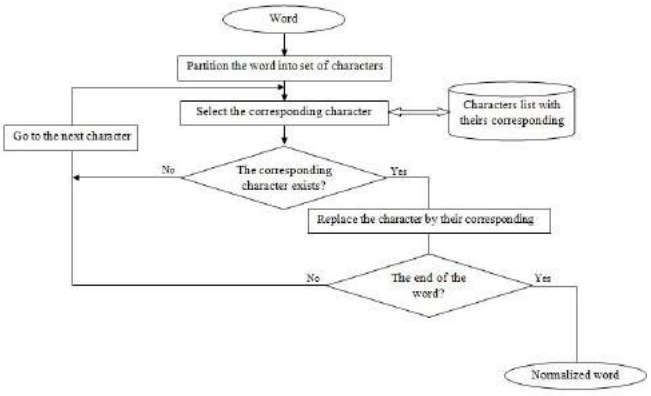

Fig. 2: Normalization Process

2.1.2 Transcription

The transcription consists in replacing each arabic word by its english phonetics, for example, كتاب is transcribed to "ketab". The use of the transcription is highly significant for resolving the problem of diacritics. So, while the word ملك "malek" could have three different meanings when it appears without diacritics in arabic, in transcription each meaningful word has a single representation. Another advantage of using transcription is avoiding the removal of suffixes and prefixes that sometimes could be part of the word. The prefix ب (pronounced as "bi") is very common in arabic (Table 1). This preposition resembles the letter ب of the arabic alphabet (Fig. 3). The untranscription is the inverse of transcription, consists in replacing each Arab word written in english phonetics, by the arabic form, for example, "ketab" is untranscribed to "كتاب" [37, 55].

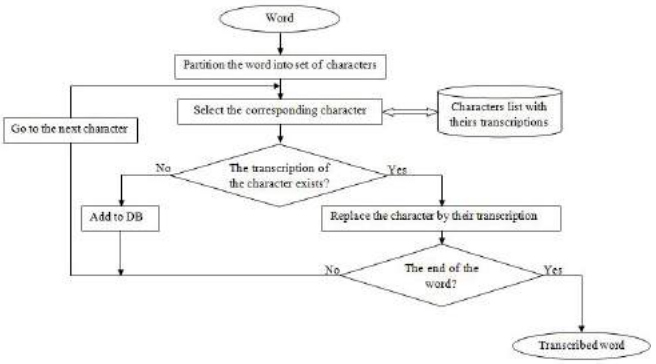

Fig. 3: Transcription Process

2.2 Morphological analysis based methods

In the arabic language, a morphological analysis consists to identify the morphemes of a word (Stem): the affixes (prefix, infix, and suffix) and the root. A stem can be a noun, verb or particle. It can be composed of one part, root, for example: ت ك ب "k t b", two parts, root and pattern, for example: كَ تِ بَ "k u t e b": Root (ك ت ب "k t b") and pattern (CuCiC where C is the consonants of the root (the radicals)), and three parts, root, pattern, and affixes, for example: ال ك ا ت ب و ن "al k a t e b o n": Root (ك ت ب "k t b") , pattern (CaCiC), and affixes (prefix (ال "al"), infix (alef "ا"), and the suffix (ون "waw non") (Table 2). This analysis can be based on affixes, translation, or pattern and affixes.

2.2.1 Morphological analysis based on affixes

Several Stemmer algorithms use the predefined rules to remove the affixes (prefix, infix, suffix ...) from the word to extract the root, after applying the normalization phase. This class allows remarkably good information retrieval without providing correct morphological analysis. In this class, we can find a part of method which uses a dictionary to validate the extracted root, and the other part one does not use any dictionary.

2.2.1.1 Morphological analysis based on affixes and dictionary

In several methods, the Stem is reducing by removing these affixes, and then the Stem not affixed is checked against the full dictionary, if the

word was found that means the algorithm find the root.

2.2.1.1.1 Clitic Stemmer

The process of the morphological analysis is beginning with the tokenization (separating the input stream into a graph of words), and then apply the "simple word lookup" of the tokenized strings in the dictionary. If the word was not found, an existing orthographical alternative lookup (looking for differently accented forms, alternative hyphenisation, concatenated words, and abbreviation recognition) was also used in order to find lexical entries for unvoweled or partially voweled words. At this point in the processing, a word that contains clitics will not have been found in the dictionary since they had decided not to include word forms including clitics. They introduced, here, a new processing step for arabic: Clitic Stemmer. This stemmer uses the following linguistic resources: The full form dictionary (5.4 million entries), containing for each word form its possible part-of-speech tags and linguistic features (gender, number, etc.). The proclitic dictionary (77 entries) and the enclitic dictionary (65 entries), having the same structure of the full form dictionary with voweled and unvoweled versions of each valid combination of clitics (Grefenstette et al., 2005; Semmar et al., 2005). The clitic stemmer proceeds as follows: First steps is the normalization: remove ّ, replace أ, إ and آ by ا, and final ة, ى, or ؤ by ء ى ه, ي, or و ء)(Table 1). In the second step, a radical, computed by removing these clitics, by using proclitics and enclitics dictionaries, and then this radical is checked against the full form lexicon. If it does not exist in the full form lexicon, re-write rules (such as those described in [18]) are applied, and the altered form is checked against the full form dictionary (Fig. 4).

Example: the token "wahawahon" وهواهم and the included clitics (و, هم) (Table 1, 2), the computed radical "hawa" هوا does not exist in the full form lexicon but after applying one of the re-write rules, the modified radical "hawa" هوى is found the dictionary and the input token is segmented into root and clitics as: وهواهم = و + هوى +هم.

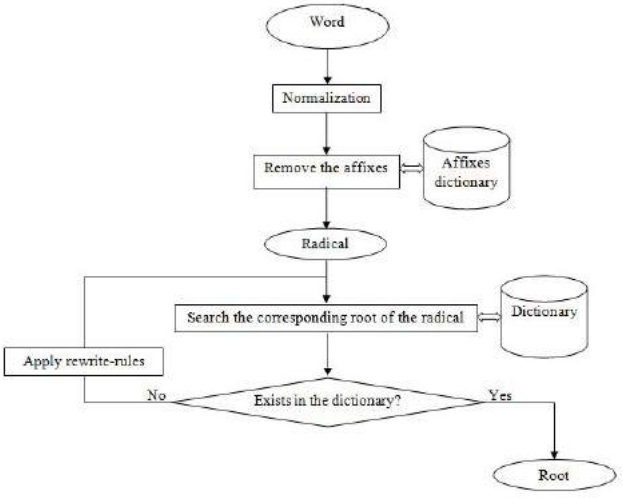

Fig. 4: Clitic stemmer process.

2.2.1.2 Morphological analysis based on affixes without dictionary

The algorithms including in this class, refers to the process of stripping off affixes from arabic words, without using of a dictionary. Several algorithms have been developed under this class.

2.2.1.2.1 Light Stemmer

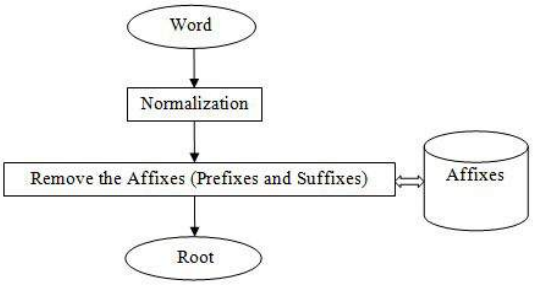

Fig. 5: Light stemmer process.

[12] proposed an algorithm for Arabic texts which is called light stemmer, this approach is mainly based on normalization, and affixes removal if the length of this word is greater than or equal to 3 characters (Fig. 5).

[17] identified other sets of prefixes and suffixes. To remove the prefixes and suffixes in the pre-defined sets, each algorithm proposes their own rules (Fig. 5). For example, they apply the following rules: If the word is at least five-character long, remove the prefixes of three

characters: ال, بال, فال, كال, ولل, مال, اال, سال, لال. If the word is at least four-character long, remove the first two characters: ال, وا, با, لل, وم, وت, وب, لا, سي, وس, وي, ول, كا, فا. If the word is at least four-character long and begins with و remove the initial letter و. If the word is at least four-character long and begins with either ب orل remove ب or ل (Table 2).

[36] presented a new stemming algorithm to extract quadrilateral arabic roots. The algorithm starts by excluding the prefixes, and then checks the word characters starting from the last letter backward to the first one (Fig. 5). A temporary matrix is used to store the suffix letters of the arabic word, and another matrix is used to store the roots. Algorithm checks the letters of any word, also checking whether the tested letter is included within the general standard arabic word.

[5, 42] developed the light stemmer which is based on the suppression of "و" if it is initial at the beginning of the word, of the prefixes (ال, وال, بال, كال, فال, لل), and of the suffixes (ها, ان, ات, ون, ين, يه, ية, ه, ة, ي) (Table 2). They divide the algorithm into several categories (SPS_TREC, SP_WAL, SPS_WAL, SP_WOAL, SPS_WOAL): SP (Suffix-Prefix): remove suffixes terms reclusive while prefix terms removed non-recursively at the end. SPS (Suffix-Prefix-Suffix) removes single largest available suffix term first, after largest single prefix term, at the end it removes a single largest remaining suffix if any. We noticed that most of arabic words use (ال Alef Lam) prefix as a declarative term. Therefore two new categories: WOAL (without ال), WAL (with ال) (Fig. 5).

2.2.1.2.2 Qur'an Stemmer

The Morphology and grammar of the Qur'an are more complicated than Modern Standard arabic [1, 34, 55]. Therefore, [55] proposes a new stemming approach based on a light stemming technique that uses a transliterated version of the ur'an in western script.

The large-scale use of diacritics (ً, ٌ, ٍ, َ, ُ, ِ, ّ, ْ) (Table 1) representing short vowels are prevalent in the Qur'an. Every word, even every letter is marked with a diacritic. (For example: مُلك Mulk "reign", مَلِك Malik "king" ...).

This stemmer is basically a light stemmer to remove prefixes and suffixes and is applied to a version of the Qur'an transliterated into western script. The algorithm of this method is given by the figure 6.

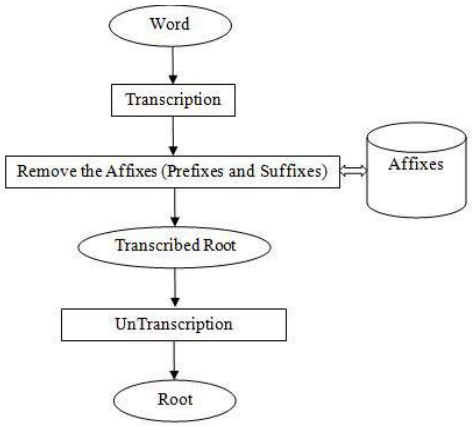

Fig. 6: Qur'an stemmer process.

The prefix stemming reads individual suras from texts files, replaces all the uppercase letters with the lower case letters and constructs a list of word lists, where each word list contains all the words in a single sura. If the word is found in the stopword list, it is excluded from prefix stemming; otherwise, it removes prefixes (wa, fa, la, li, lil, bi, ka, sa, s^a, al), after stemming, the word is inserted back into the word list.

In the suffix stemming, six groups of suffixes are identified by (one to six letters). The system starts stemming the words in the word lists from the longest prefixes (six-letter prefixes) to the three-letter prefixes.

2.2.1.2.3 Linguistic-based Stemmer

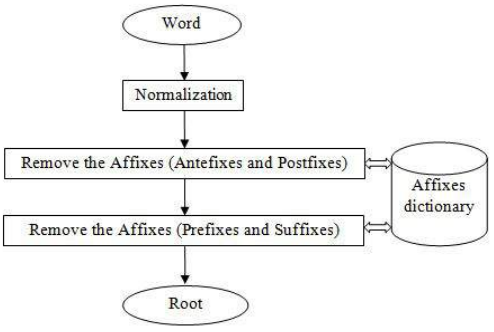

Fig. 7. Linguistic-Based stemmer process.

[35] defined that the arabic words are usually formed as a sequence of antefixes are generally prepositions joined to words at the beginning (كال, فال, بال, وال, وبال,...)(Table 2), prefix are usually represented by only one letter and indicate the conjugation person of verbs in the present tense (ا, ت, ي, ن...)(Table 1), core, suffixes are the conjugation terminations of verbs and they are the dual/plural/female marks for the nouns (تما, ان, ات, تان, تين, يون...), (Table2) and postfixes represent pronouns attached to the end of the words (كن, هما, كما...)(Fig. 7).

For example: in the word ليفاوضونهم "liyofawidonahom": Antefix: لـ (pour), Prefix: يـ, Suffix: ـونـ, Postfix: ـهم ☐ Core: فاوض "fawada" (Table1, 2).

2.2.2 Morphological analysis based on patterns and affixes

Several algorithms based on patterns and affixes have been developed, to find roots with three letters, four and five letters. In these types of morphological analysis, we can find some ones use the dictionaries and the specific rules, to achieve the information extraction process.

2.2.2.1 Morphological analysis based on patterns, affixes, and dictionary

The dictionary is a best reference in any language. The methods of this class validate the extracted root in a dictionary. Several methods are based on this concept.

2.2.2.1.1 Khoja Stemmer

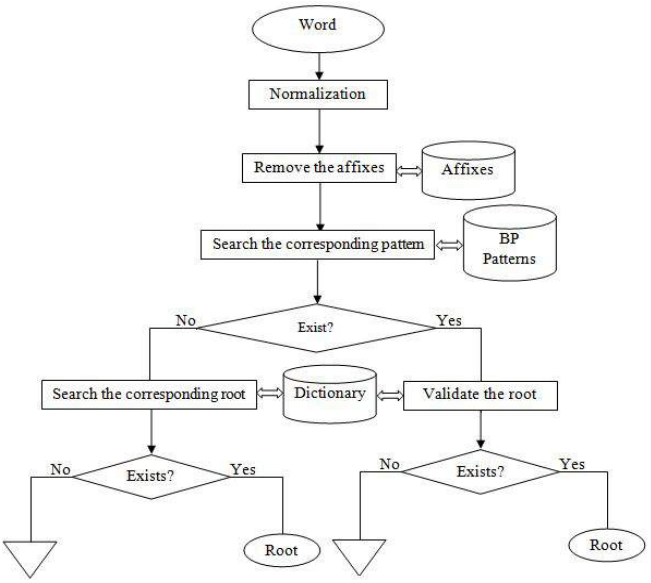

Fig. 8: Khoja stemmer process.

[38] have proposed a method that involves removing diacritics representing vowelization, the stop words, the punctuation, the numbers, the definite article (ال), the inseparable conjunction (و), and the longest prefix and suffix (Fig. 8). Then, the result is compared to a list of patterns. If a match is found, the characters representing the root in the pattern are extracted, and Match the extracted root against a list of known roots (dictionaries) (Table 1, 2).

[3] proposed the QARAB system which based on the Khoja Stemmer [38]. [12] propose more pattern for Khoja Stemmer [38].

[46] modified the Khoja stemmer to check whether there is a match between a word and a list of patterns after stemming without further checking against the root dictionary. If there is no match, the word is considered a foreign word. Also, They added 37 new patterns (افعوعل, افعلة, افعلال, افعلاء,....) (Table 3).

2.2.2.1.2 Broken Plurals Dictionary Approach

The fastest way to detect BPs is to use a look-up table which lists all BP stems. It is quite clear that it will be fairly difficult to build look-up tables listing either BP stems or full words from language dictionaries (Fig. 9). The dictionary was

built as follows: Prepare the all broken plural BPs. The manually restricted BP matching system [25] was run on the 127,000 stem types, extracted from a large corpus. To retrieve all types that match BP patterns. A list of roughly 3,600 BP stems, alphabetically ordered and categorized according to each BP pattern, was extracted.

The list was further revised in collaboration with a linguist, who is an Arabic native speaker. The revised list contained exactly 3,580 BP stems.

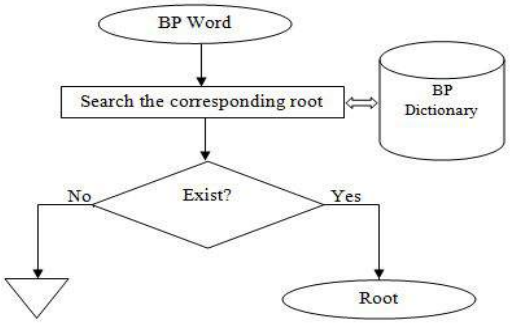

Fig. 9: Broken Plurals Dictionary Approach process

2.2.2.2 Morphological Analysis based on patterns, affixes, rules, and dictionary

The principle of this class is the generation of a stem dictionary using roots, patterns, rules, and affixes. To extract the root of a given word they directly search in this dictionary. Several methods are based on this concept.

2.2.2.2.1 Xerox Finite-state Morphological Analysis

In 1996, the Xerox Research Centre Europe produced a morphological analyzer for Modern Standard Arabic is based on dictionaries. In 1997 a Java-applet interface was added to allow testing on the Internet. In 2001, the system was extensively redesigned and rebuilt using Xerox finite-state technology [14].

The Finite State Technology research concentrates on tools for specifying and manipulating finite state automata (acceptors, transducers, and multi-tape machines). The tools (xfst, twolc, lexc) are built on top of a software library that provides algorithms for creating automata from regular expressions and equivalent formalisms and contains both classical operations such as union and composition and also new algorithms such as replacement and local sequentialisation. Over the years, the products of this research have come to be used all over the world in many linguistic applications such as morphological analysis, tokenisation, and shallow parsing of a wide variety of natural languages. The xfst tool has been licensed to over 70 universities world-wide. Many components have been incorporated into commercial software.

Xerox has several lexicons. The first is a lexicon of ROOTS, which contains 4,930 entries. Each ROOT-entry is manually coded and associated with PATTERNS. The second is a dictionary of PATTERNS, which includes about 400 entries. The manual association of ROOTS and PATTERNS produces about 90,000 Arabic stems (Fig. 10). When these stems combine with possible prefixes, suffixes and clitics by composition, 72 million abstract words are generated [14].

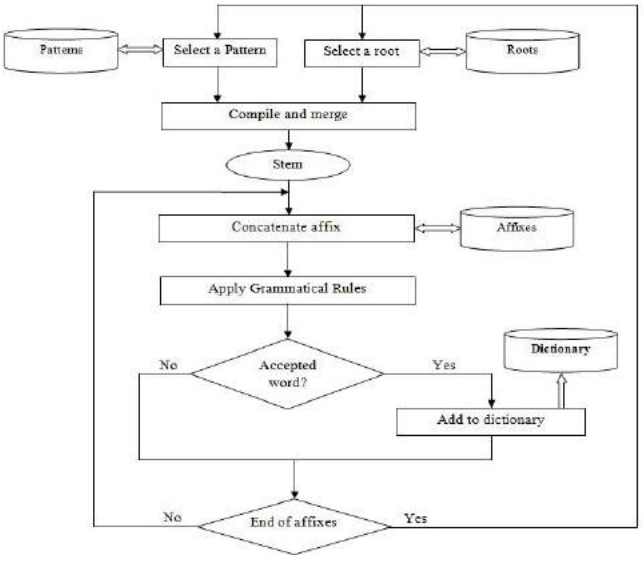

Fig. 10: Xerox Finite-State Morphological Analysis process

To find the root of a specific word, we use the dictionary built in figure 10, according to the procedure described by the figure 11.

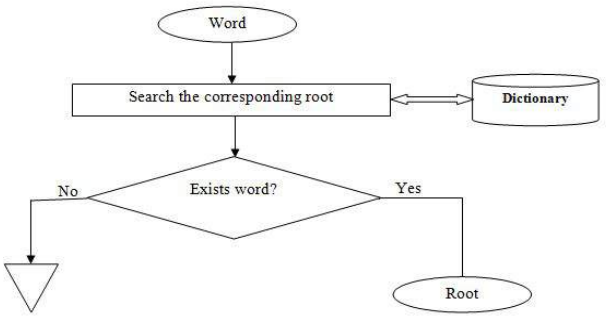

Fig. 11: Root Extraction process using built dictionary.

This method work as the following example: Take the words كتبت "katabat",بنت "banat", and قال "qaala", the first step consist to merge the roots (ktb, bny, qwl) with the pattern (CaCaC, CaCaC, CaCuC) to build the Stems (katab, banaya, qawul) after application of the MERGE and COMPILE-REPLACE algorithms. The second step consists to add the suffixes to the stems (katabat, banayat et qawula). The third step consists to compile and apply the alternation rules. E.g. In the case of the word katabat, it does not apply any rules, which is essentially finished كَتَبَت "katabat". Otherwise, in the case of the word banayat apply the rule consist to disappear the letter y and to map them into their final orthographical forms بَنَت "banat". And in the case of qawula the wu is replaced by a to find the form قال "qaala".

2.2.2.2.2 Systematic Verb Stem Generation For Arabic

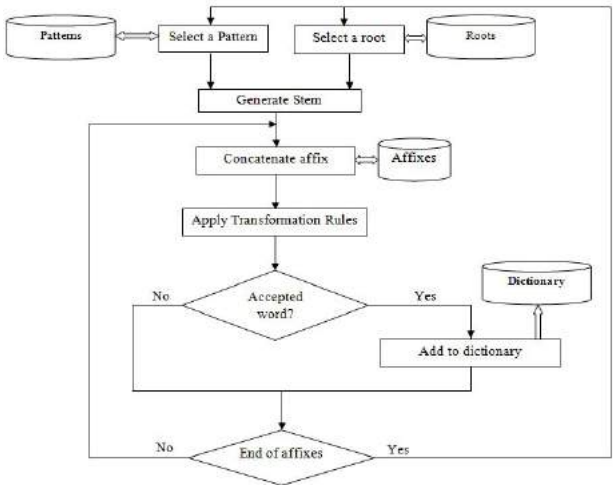

Fig. 12: Systematic Verb Stem Generation process.

This method present an arabic stem generation engine with the required database. The roots are represented in terms of their ordered sequence of three or four radicals in a set notation, i.e., {F, M, L, Q}. They stand for First Radical (F), Medial Radical (M), Last Radical in a trilateral root (L), and Last Radical in a quadrilateral root (Q) (Yagi AND Yagi, 2004).

Arabic stems can be generated if lists of all roots and all morphological patterns (about 44 patterns) are provided. It is necessary that this data be coupled with a database that links the roots with their morphological patterns so that only valid stems are generated for each root (Fig. 12).

Stems should be defined in terms of the root radicals, Pattern, and morphophonemic alterations. Since originally Arabic words can have a maximum of four root radicals, a root radical set R is defined in terms of the ordered letters of the root as follows:R = {rF, rM, rL, rQ}

The text s is defined in terms of the letters and diacritics of the template in sequence (x1...xn) and the radical position markers or place holders (hF, hM, hL, hQ), that indicate the positions that letters of the root should be slotted into

$$s = x1x2...hF...hM...hL...hQ...xn$$

Table 4: Example of the word رسم "Rasama"

Root	Pattern	intermediate template	Stem
{ر,س,م} ({r, s, m})	hF hM hL (FaMaLa)	hF َ hMسhLم (FraMsaLma)	"rasama"رَسَمَ

In this method many transformation rules are used in the stem generation process. For example, TR1: replaces ت withذ , TR12: geminate duplicate letters (ّ), TR20: change rF to a duplicate of the next letter (ت), TR31: delete diacritic after the ي in position hL (Table 1).

To find the root of a specific word, we use the dictionary built in figure 12, according to the procedure described by the figure 11.

2.2.2.3 Morphological Analysis based on patterns, affixes, and rules without dictionary

To extract a root, these methods apply morphological rules, remove affixes, and search for the corresponding pattern. Several methods are based on this concept.

2.2.2.3.1 Arabic Natural Language Processing

In this method, the types of part of word are defined: prefix, suffix, circumfix, templatic root, and pattern (Habash, 2005). The functions necessaries to create a word are: Derivational allow creating new words, inflectional which allow modifying words features (tense, number, person, mood, aspect) (Table 5).

From the taken root, the Derivational Morphology let to create a Lexeme: Root + Pattern, and the Inflectional Morphology allow to derive a Word = Lexeme + Features (Fig. 13).

Table 5: Example of extraction the root of the word وللمكتبات "wael almaktaba " in Arabic Natural Language Processing extraction method.

وللمكتبات "waalelmaktabate"				
Word				
ات "ate"	مكتبة "maktaba"	ال "al"	ل "la"	و "wa"

Plural	Noun	Article	Prep	conj
مكتبة "maktaba"				
Lexeme				
كتب "kataba"			ةa3a21aم	
Root			Pattern	
			ة321م	a.a.a
			Pattern	Vocalisme

The Features are the Part-of-speech (Noun, Verb, Particle, N, PN, V, Adj, Adv, P, Pron, Num, Conj, Det, Aux, Pun, IJ, and others), and Noun-specific (Number: singular, dual, plural, collective, Gender: masculine, feminine, Neutral, Definiteness: definite, indefinite, Case: nominative, accusative, genitive, Possessive clitic) [39, 44].

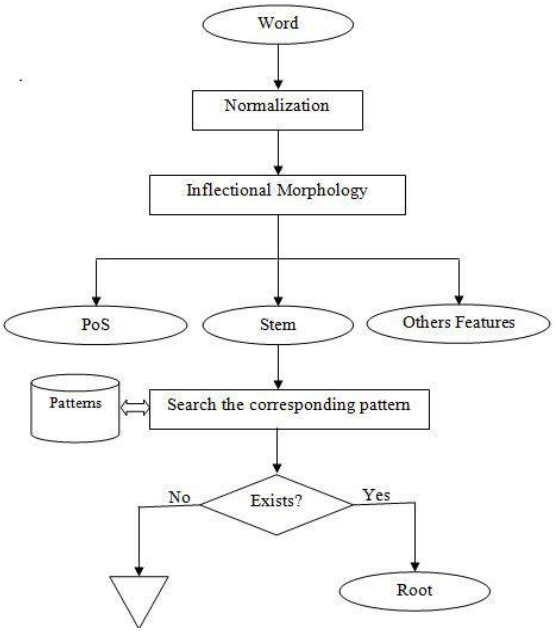

Fig. 13: Arabic Natural Language Processing.

2.2.2.4 Morphological Analysis based on patterns and affixes without dictionary and rules

To extract a root, these methods remove affixes and search for the corresponding pattern. This

treatment can be repeated according to the word morphology. Several methods are based on this concept.

2.2.2.4.1 Computational Morphology System For Arabic

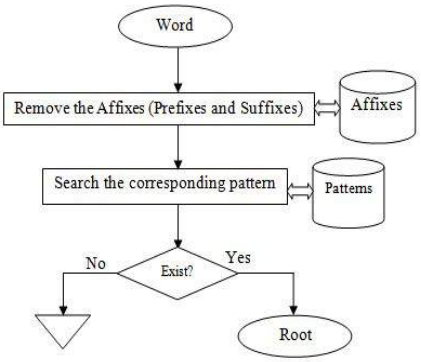

Fig. 14: A Computational Morphology System for Arabic process

[10] proposed an approach of Stemmer, whose the first step is to remove the longest possible prefix. The three letters of the root must lie somewhere in the first four or five characters of the remainder. They checked all the possible trigrams within the first five letters of the remainder. That is, we check the following six possible trigrams: First, second, and third letters. First, second, and fourth. First, second, and fifth, etc. To test the algorithm, they prepared two files: a file of roots and a file of prefix (Fig. 14).

2.2.2.4.2 Arabic Stemming Without A Root Dictionary

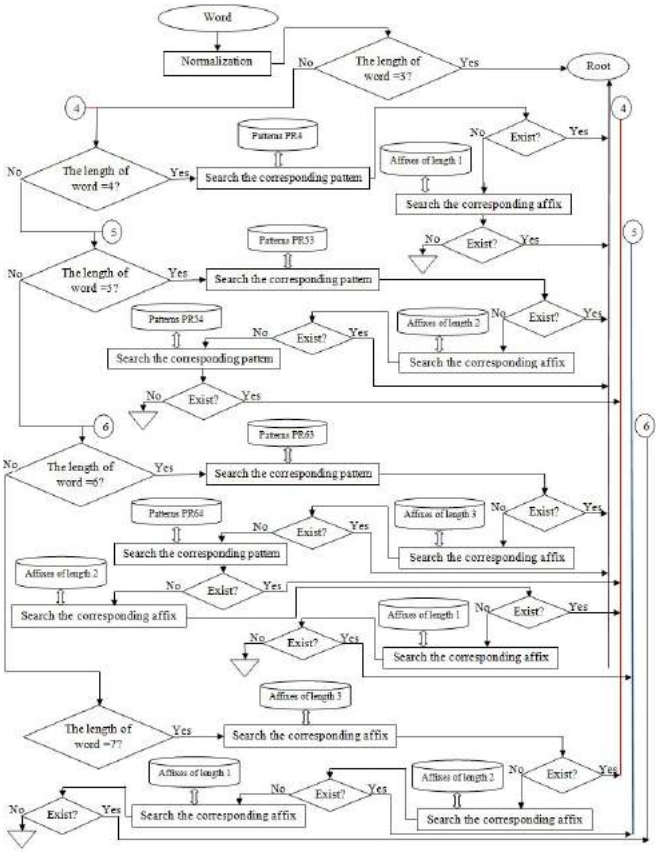

Fig. 15: Arabic Stemming without a root dictionary

[54] implemented a root-extraction stemmer for arabic which has a performance equivalent to the Khoja stemmer [38] and the "Light Stemmer" [5, 42].

To implement this algorithm, they have defined several sets of the affixes (D diacritic: سَ, سِ, سُ, سْ (Table 1). P3 prefix of length 3: وال, ولل. P2 prefix of length 2: ال, لل. P1 prefix of length 1: لَ, ب, ف. S3 suffix of length 3: تان, همل, تمل. S2 suffix of length 2: ون, ات, ان. S1 suffix of length 1: ة, ه, ي (Table 2)), and several sets of models (PR4 model of length 4: فاعل, فعول, فعلة. PR53 model of length five and a root of length 3: تفاعل, افتعل, أفعال.... PR54 model of length five and a root of length four: تفعلل, افعلل, مفعلل. PR63 model of length 6 and a root of length 3: استفعل مفعالة افتعال.... PR64 model of length 6 and a root of length 4: افعلال, متفعلل... (Table 4)).

This algorithm proceeds in the following steps (Fig. 15):

Remove diacritics representing vowels, normalize the hamza which appears in several distinct forms in combination with various letters to one form (ا),remove the prefixes of length three and two, remove connector "و" if it precedes a word, normalize □, إ, أ to ا, and return the stem it is less than or equal to three (Table 1).

Then, the extraction of the root is based on the length of the stem. If the length is 4, and the word matches one of the patterns from PR4, extract the relevant stem and return it. Otherwise, attempt to remove length-one suffix and prefix from S1 and P1 in that order the word provided is not less than length three. If the length is 5, Extract stems with three characters for words that match patterns from PR53. If none are matched, attempt to remove suffixes and prefixes, otherwise the relevant length-three stem is returned. If the word is still five characters in length, the word is matched against PR54 to determine if it contains any stems of length 4. The relevant stem is returned if found. And so on.

The authors tested their method on the 'Arabic Trec 2001 ", which contains 383,872 documents. They also compared their method with Khoja Stemmer [38] and Light Stemmer [5, 42], according to them, their method gave a higher precision.

2.2.2.4.3 Broken Plurals Matching Approach

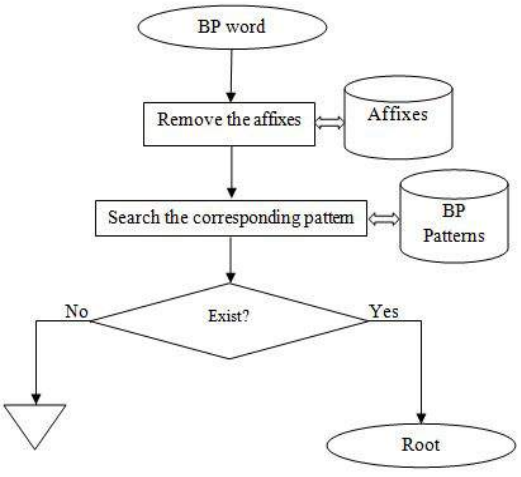

Fig. 16: Broken Plurals Matching Approach

This method detects the broken plural using a list of patterns. Broken plurals (BPs) are formed by altering the singular through an application of interdigitating patterns on stems.

In the first time, this method use light stemmer to produce morphological information such as stem, prefix and suffix, and returns TRUE if the stem matches one of 39 BP patterns found in grammar books. The stem matches a BP pattern if and only if they have the same number of letters and the same letters in the same positions, excluding the consonants f (ع) 3 ,(ف), and l (ل) of the basic root f3l(فعل) found in the pattern(Table 1, 4) (Fig. 16).

2.2.2.4.4 Shallow Arabic Morphological Analyzer

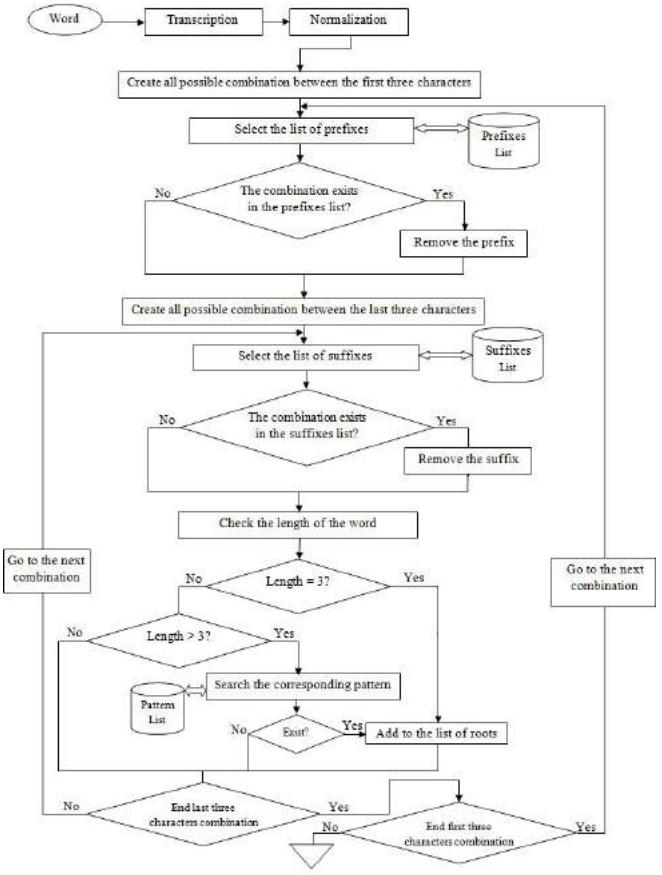

Fig. 17: Shallow Arabic Morphological Analyzer process

This method presents an arabic morphological analyzer. The analyzer will be concerned with generating the possible roots of any given arabic word. The analyzer is based on automatically derived rules and statistics. It is based on collecting statistics from word-root pairs, to build

morphological rules for deriving roots from words, to construct a list of prefixes and suffixes, and to estimate the probability that a rule will be used or a prefix or suffix will be seen. For that, a system module called build-model utilizes a list of arabic word-root pairs to derive a list of prefixes and suffixes, to construct stem templates, and to compute the likelihood that a prefix, a suffix, or a template would appear. Second, another system module called detect-root accepts arabic words as input, attempts to construct possible prefix-suffix-template combinations, and outputs possible roots [37] (Fig. 17). For example, the arabic word “yAktobon” have the possible prefixes (“#”, “y”, and “yA”) and the possible suffixes (“#”, “n”, “on”). The resulting possible stems are (Table 6):

Table 6: Example of extraction the list of roots of the word “yAktobon”

Stem	Pre fix	Template	Suf fix	Suf fix
yAktobon	#	yAXXoXon	#	Ktb
Aktobon	Y	AXXXbon	#	Kto
Aktob	Y	AXXoX	On	Ktb
…				

2.2.3 Stemmer based on translation

The algorithms of english Stemmer have better performance than arabic [43, 47]. For that, several methods use the translation technique [48], to allow any languages (arabic) to use the Stemmer of another language (english) to extract the root of a word [33].

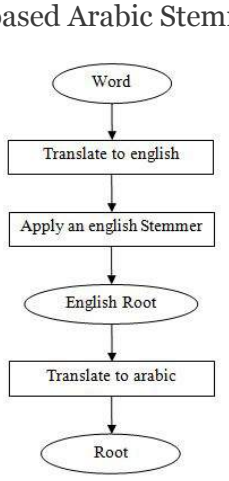

Fig. 18: MT-based Arabic Stemmer process.

[17] have built a MT-based arabic stemmer from the arabic words found in the arabic documents, and their english translations are using the online “Ajeeb” machine translation system. They divided the arabic words into clusters based on the english translations of the arabic words. The arabic words whose english translations, after removing english stop words, are conflated to the same english stem that made from one cluster. All the arabic words in the same cluster are conflated to the same arabic word, which is the shortest arabic word in the cluster. For example: أطفالنا ¨atfalona¨ (our children), remove "our" is a word parasite, أطفالنا ¨atfalona is apparent that in relation to "child". So أطفالنا ¨atfalona¨ is related to طفل ¨tofol¨ (Fig. 18).

2.3. Statistical analysis based methods

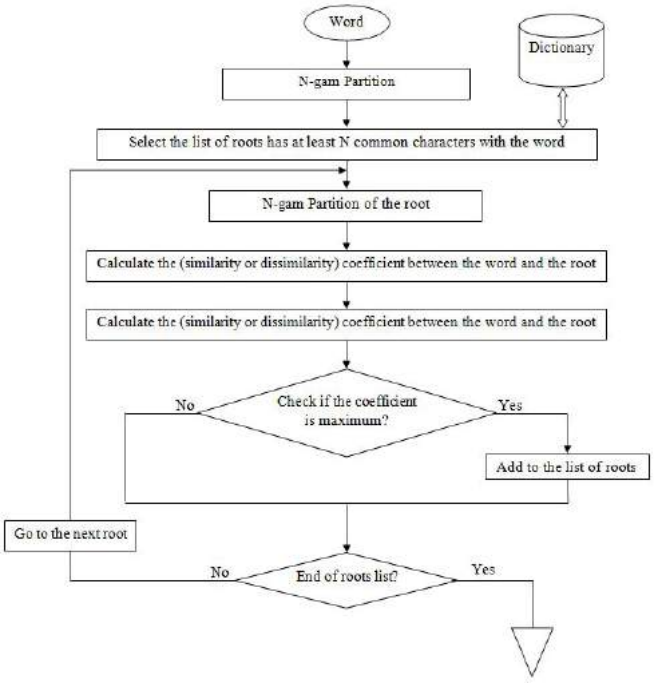

Fig. 19: Statistical analysis process (N-gram).

To find the root of a word, the second approach is based on statistical analysis. This approach consists to determine the semantic similarity or dissimilarity of words set. Two words are considered similar, if they have in common several substring of N characters, this is done by calculating a coefficient on these two words (Fig. 19).

2.3.1 Statistical analysis based on Similarity Coefficient

[2] has been developed the first automatic classification technique based on the character structure of words. Dice's Similarity Coefficient is computed from the number of matching bigrams (2-gram) in pairs of character strings, and used to cluster sets of character strings (Fig. 20).

[53] assesses the performance of two N-gram matching techniques for arabic root-driven string have prefixes and suffixes which make more searching: contiguous N-grams and hybrid N-grams, combining contiguous and non-contiguous.

The two techniques were tested using three experiments involving different levels of textual word stemming, a textual corpus containing about 25 thousand words (with a total size 160KB), and a set of 100 queries textual words. The results of the hybrid approach showed significant performance improvement over the conventional contiguous approach.

[3] have been presented the n-gram model which can be used to compute the similarity between two strings by counting the number of similar n-grams they share. The more similar n-grams they found between the two strings exist the more similar they are (Ahmed et al., 2007; Buscaldi et al., 2006). Based on this idea the similarity coefficient can be derived (Fig. 19). The similarity coefficient δ is defined by the following equation:

$$\delta_n(a,b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|}$$

Where α and β are the n-gram sets.
For Example: shows an example of two arabic words: الإستمرارية¨alestemrareya¨ and استمرار ¨estemrar¨(Fig. 20):

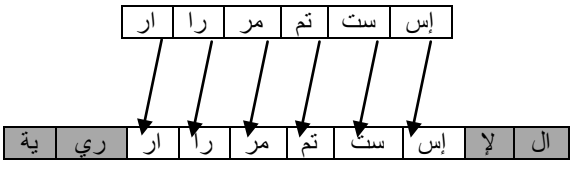

Fig. 20: Bigram similarity measure between two words الإستمرارية ¨alestemrareya¨and استمرار ¨estemrar¨

[51] presented an approach that uses N-gram based on the word and characters. Four basic types have been explored, sometimes separately and sometimes in combination: Word, lexical root, root, and N-gram. In general, N-grams based on the stems are better than those based on words, because the N-grams based on words could have prefixes and suffixes which make more mistakes in the Similarity between the document and query.

2.3.2 Statistical Analysis bBased on Dissimilarity Coefficient

[40] presented the N-Gram Frequency Statistics technique for classifying arabic text documents. The technique employs a dissimilarity measure called the "Manhattan Distance", and "Dice's measure". A corpus of arabic text documents was collected from online arabic newspapers, 40% of the corpus was used as training classes and the remaining 60% of the corpus was used for classification. All documents, whether training documents or documents to be classified went through a preprocessing normalization phase that remove the punctuation marks, the stop words, the diacritics, and the non letters. For the training documents, the N-gram (N=3) (the trigrams of the word المودعين ¨almodeoon¨are: الم, مود, ودع, دعي, عين) (Table 2), was generated for each document and saved in text files. Then for each document to be classified, the N-gram frequency profile was generated and compared against the N-gram frequency profiles of all the training classes.

2.4 Morphological and statistical analyses based methods

A. N. De Roeck and W. Al-Fares (2000) presented a method based on both approaches. In first step, they applied the Light Stemmer to remove affixes, the second step is based on the Adamson algorithm [2] with some modifications (Fig. 21). These modifications consist on assigning for each bi-gram a weight (0.25 for bi-gram containing low letter, 0.5 for bi-gram containing the non-low letter, 1 for all other bi-gram). Then, the oefficients are calculated in the same way.

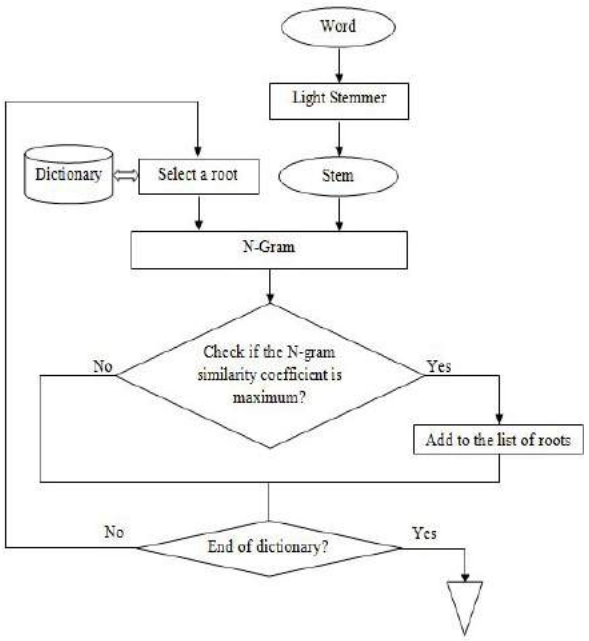

Fig. 21: Two Stage process.

III. CONCLUSION

In this paper we have presented a survey of near 20 different arabic information extraction methods. The main approaches applied in these methods are morphological or statistical analyses. The morphological analyses consist to identify the morphemes of a word, the affixes (prefix, infix, and suffix), the pattern, and the root. The statistical approach consists to determine the semantic similarity or dissimilarity coefficient between two words, two words are considered similar, if they have in common several substrings of N characters. In the two approaches, the dictionary has an important role, it use as look-up reference, to extract the root or to check the validity of the extracted root. The advantages of statistical approaches are that they do not require a preliminary knowledge of the language, do not require predefined rules, and do not require the construction of a vocabulary database. The advantages of morphological approaches that are more appropriate for complex words which not included in the dictionary and they are more rapid.

This work will contribute to the enhancement of the Arabic information retrieval system performance, due to the fact that arabic information extraction methods are the kernel of such system. The next step will be the making of a detailed comparative study of the early described methods by examining their performances, stabilities, usability, advantages, and disadvantages. Another possible extension of the present work is to evaluate these categories in similar conditions.

ACKNOWLEDGEMENTS

This work has been done as a part of the project “Arabic Web Intelligence” supported by the Lebanese National Centre of Scientific Research (CNRSL).

REFERENCE

1. Abd Al-Baqi, M.F. 1987. "Al-Ma&jam Al-Mufahras li-alfaz Al-Qur’an Al-Karim". Dar Al-hadith, Cairo.
2. Adamson, G.W., AND Boreham, J. 1974. "The use of an association measure based on character structure to identify semantically related pairs of words and document titles". Information Storage and Retrieval, Vol. 10, pp 253-260.
3. Ahmed, F., AND Nürnberger, A. 2007. “N-grams Conflation Approach for Arabic”. ACM SIGIR Conference, Amsterdam.
4. Ahmed, F., De Luca, E.W., AND Nürnberger. A. 2007. "MultiSpell: an N-Gram Based Language-Independent Spell Checker". In: Poster Postproc of Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007), Mexico City, Mexico, IEEE CS Press, 2008.
5. Al Ameed, H., Al Ketbi, S., Al Kaabi, A., Al Shebli, K., Al Shamsi, N., Al Nuaimi, N., AND Al Muhairi, S. 2005. "Arabic Light Stemmer: A new Enhanced Approach", The Second International Conference on Innovations in Information Technology (IIT’05).
6. Al Fedaghi, S., AND Al-Anzi, F. 1989. "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.
7. Al Hajjar, A., Hajjar, M., AND Zreik, K. 2009. "Classification of Arabic Information Extraction methods", 2nd International Conference on Arabic Language Resources and Tools Cairo (Egypt), 22 - 23 April 2009.
8. Al Kharashi, I., AND Evens, M. 1994. "Comparing words, stems, and roots as index terms in an Arabic Information Retrieval system". Journal of the American Society for Information Science, Volume 45, Issue 8 , Pages 548 – 560.
9. Al Kharashi, I. 1999. "A Web Search Engine for Indexing, Searching and Publishing Arabic Bibliographic Databases".
10. Al Shalabi, R., AND Evens, N. 1998. "A Computational Morphology System for Arabic", Proceedings of COLING-ACL, New Brunswick, NJ.
11. Al Sughaiyer, I., AND Al-Kharashi, I. 2004. "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.
12. Aljlayl, M., AND Frieder, O. 2002. "On arabic search: Improving the retrieval effectiveness via a light stemming approach". In Proceedings ofACM Eleventh Conference on Information and Knowledge Management, Mclean, VA.
13. Attia, A.M. 2007. "An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks". The Challenge of Arabic for NLP/MT.
14. Beesley, K. R. 2001. "Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001". In The ACL 2001Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France.
15. Buscaldi, D., Manuel, J., AND Sanchis, E. 2006. "N-gram vs. Keyword-based Passage Retrieval for Question Answering". TIN2006-15265-C06-04 research project.
16. Censure de l'internet dans les pays arabes, Tribune des Droits Humains - Genève 2006 - www.humanrights-geneva.info, 2006.
17. Chen, A., AND Gey, F. 2002. "Building an Arabic stemmer for information retrieval". TREC-11 conference 2002.
18. Darwish, K. 2002. "Al-stem: A light Arabic stemmer" [Online]. Available:http:// www. glue.umd.edu/~kareem/research.
19. Darwish, K., Hassan, H., AND Eman, O. 2005. "Examining the Effect of Improved Context Sensitive Morphology on Arabic Information Retrieval". Computational Approaches to Semitic Languages, University of Michigan, Ann Arbor, Michigan, USA, June 29, 2005.
20. De Roeck, A.N., AND Al-Fares, W. 2000. "A morphologically sensitive clustering algorithm for identifying Arabic roots". In Proceedings ACL-2000. Hong Kong.
21. Dichy, J., AND Farghaly, A. 2003. "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.
22. Douzidia, F., AND Lapalme, G. 2005. "Un système de résumé de textes en arabe", 2ème Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la langue, Alger.
23. El-Halees, A. M. 2007. "Arabic Text Classification Using Maximum Entropy". The Islamic University Journal (Series of Natural Studies and Engineering) Vol. 15, No.1, pp 157-167, ISSN 1726-6807, http//www .iugzaza.edu.ps/ara/research/.
24. Ghosn, Z. 2003. Les sites Internet gouvernementaux au Moyen-Orient 2003, The Arab Advisors Group, www.arabadvisors.com/ Pressers/presser-230101.htm.
25. Goweder, A., Poesio, M., De Roeck, A., AND Reynolds, J. 2003. "Identifying Broken Plurals in Unvowelised Arabic Text". Proceedings of EMNLP, pp. 246-253.
26. Grefenstette, G., Semmar, N., F., AND Fluhr, C. 2005. "Modifying a Natural Language Processing System for European Languages to

Treat Arabic in Information Processing and Information Retrieval Applications". ACL computational approaches to semitic languages workshop ANN ARBOR USA, June 29 2005.
27. Habash .N. 2005. "Arabic Natural Language Processing: Words". Summer School on Human Language Technology Johns Hopkins University, Baltimore. July 6th, 2005
28. Habash, N., AND Rambow, O. 2005. "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop". In Proceedings of the 43rd Annual Meeting of the ACL, pages 573–580. Ann Arbor, Michigan.
29. Habash, N., AND Rambow, O. 2006. "MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects". In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 681–688. Sydney,Australia.
30. Haddad, H. 2006. "Arabic Natural Language Processing for Information Retrieval". The Seventh Annual U.A.E. University Research Conference CIT – 75.
31. Hammo, B., Abu-Salem, H., Lytinen, S., AND Evens, M. 2002. "A Question Answering System to Support the Arabic Language". Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages Philadelphia, Pennsylvania Pages: 1 – 11.
32. Ibn Manzour, 2008. Lisan Al-Arab. www.muhaddith.org.
33. Ibrahim, B., Zbib, R. AND. Glass, J. 2008. "Segmentation for English-to-Arabic Statistical Machine Translation". In Proceedings of the Conference of Association for Computational Linguistics (ACL), Columbus, Ohio. 2008.
34. Judith Dror, J., Shaharabani, D., Talmon, R., AND Wintner, S. 2008. "Morphological Analysis of the Qur'an". CiteSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States).
35. Kadri, Y., AND Nie, J. 2006. "Effective Stemming for Arabic Information Retrieval". proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni.
36. Kanaan, G., Al-Shalabi, R., Jaarn, J., Al-Kabi, M., AND Hasnah, A. 2004. "A New Stemming Algorithm to Extract Quadri-Literal Arabic Roots". Information and Communication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on Volume , Issue , 19-23 April 2004 Page(s): 543.
37. Kareem, D. 2002. "Building a Shallow Arabic Morphological Analyzer in One Day". In The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, USA.
38. Khoja, S., AND Garside, R. 1999. "Stemming Arabic text". Computing Department, Lancaster University, Lancaster, www. Comp. lancs.ac.uk/computing/users/khoja/stemmer. ps.
39. Khoja, S., Garside, R., AND Knowles G. 2001. "A Tagset for the Morphosyntactic Tagging of Arabic". Proceedings of the Corpus Linguistics. Lancaster University (UK), Volume 13 - Special issue, 341.
40. Khreisat, L. 2006. "Arabic Text Classification Using N-gram Frequency Statistics a Comparative Study". The 2006 International conference on Data Mining Part of the 2006 World Congress in Computer Sciences DMIN 2006: 78-82.
41. Larkey, L. S., Ballesteros, L. S., AND Connel, M. E. 2002. "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis", in Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275 – 282.
42. Larkey, L., Ballesteros, L., AND Connell, M. 2005. "Light Stemming for Arabic IR" Arabic Computational Morphology: Knowledge-based and Empirical Methods, A.Soudi, A. van en Bosch, and Neumann, G., Editors. Kluwer/Springer's serieson Text, Speech, and Language Technology.

43. Les Hatton., L., 2006. "An implementation of a word stemming algorithm for English". latest version of the GNU public license.
44. Marsi, E., Bosch, A.v.d., AND Soudi, A. 2005. "Memory-based morphological analysis generation and part-of-speech tagging of Arabic". In Computational Approaches to Semitic Languages Workshop Proceedings 29 June 2005 University of Michigan Ann Arbor, Michigan, USA.
45. Moukdad, H. 2006. "Stemming and root-based approaches to the retrieval of Arabic documents on the Web". Webology, 3(1), Article 22. Available at: http://www.webology.ir/2006/v3n1/a22.html
46. Nwesri, A.F.A., Tahaghoghi, S.M.M., AND Scholer, F. 2006. "Capturing Out-of-Vocabulary Words in Arabic Text ". In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), Sydney, Australia, 22-23 July, 2006.
47. Porter, M.F. 1980. "An algorithm for suffix stripping". Published in \Program\, \14\ no. 3, pp 130-137, July 1980.
48. Rogati, M., McCarley S. AND Yiming Y. 2003. "Unsupervised Learning of Arabic Stemming using a Parallel Corpus". Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 391-398.
49. Ryan, R., Rambow, O., Habash, N., Diab, M. AND Rudin, C. 2008. "Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking". In Proceedings of Association for Computational Linguistics (ACL), Columbus, Ohio. 2008.
50. Semmar, N., Elkateb-Gara, F., AND Fluhr, C. 2005. "Using a Stemmer in a Natural Language Processing system to treat Arabic for Cross-language Information Retrieval". International conference on Machine Intelligence, Tozeur, TUNISIE, November 05-07 2005.
51. Sinane, M., Rammal, M., AND Zreik, K. 2008. "Arabic documents classification using N-gram". Conference ICHSL6, Toulouse.
52. Sonbol, R., Ghneim, N., AND Desouki, M.S. 2008. "Arabic Morphological Analysis: a New Approach". In Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008.
53. Suleiman, M. H. 2004. "Character contiguity in N-gram based word matching: the case for Arabic text searching". Information Processing and Management.41 (4), 819-827.
54. Taghva, K., Elkoury, R., AND Coombs, J. 2005. "Arabic Stemming without a root dictionary". International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I pp. 152-157.
55. Thabet, N. 2004. "Stemming the Qur'an". WORKSHOP ON Computational Approaches to Arabic Script-based Languages, University of Geneva, Geneva, Switzerland.
56. Xu, J., and Croft, W.B. 1996. "Corpus-Based Stemming using Co-occurrence of Word Variants". In ACM TOIS, Jan. 1998, vol. 16, no. 1, pp. 61-81, Computer Science Technical Report TR96-67.
57. Xu, J., Fraser, A., AND Weischedel, R. 2002. "Empirical studies in strategies for Arabic retrieval". Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, August 11-15, 2002, Tampere, Finland [doi>10.1145/ 564376.564424]
58. Yaghi. J AND Yagi. S, 2004. "Systematic Verb Stem Generation For Arabic". Workshop On Computational Approaches To Arabic Script-Based Languages.
59. Zitouni, I., Sorensen, J., Luo, X., AND Florian, R. 2005. "The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution". In ord المودعين ¨almodeoon¨are: عين ,دعي ,ودع ,مود ,الم) (Table 2), was generated for each document