International Journal of Engineering &Technology, 7 (4) (2018) 6547-6557

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.28688
Research paper



A semantic search engine based on morphology processing to improve Arabic search

Aziz Barbar [1] *, Anis Ismail [2]

[1] *American University of Science & Technology – AUST*
[2]*Lebanese University, Faculty of Technology*
**Corresponding author E-mail: abarbar@aust.edu.lb*

Abstract

In this paper, improving retrieval from Arabic text is tackled. A number of techniques such as truncating, stemming, and morphological analyzers have been introduced into to improve the retrieval performance in search engines. In Arabic search engines, three methods are mainly used: word, stem, and root. The word method is based only on term matching, while the other two methods use morphological anal-ysis. The two methods have different levels of morphological analysis, however, each of these has its limitations. For example, the word and stem methods may miss some relevant records that may contain morphological variations of the targeted word. On the other hand, the root method will always retrieve irrelevant records because it extracts the root from the word, and then searches for all possible morphologi-cal variations of that word. The limitations of the current search methods have motivated this research to investigate a new method to be used in Arabic search engines. This approach is called Semantic Search based on Morphological Processing. This method is based on se-mantic links of the morphological forms. The aim of introducing this method is based on the hope that this method will im-prove the effec-tiveness of the word and stem methods in terms of retrieving more relevant records. At the same time, it is also hoped that the proposed method will improve the root method by rejecting the irrelevant records that may be retrieved by the root method. A morphology analysis algorithm was designed and proposed to provide the needed stem and pattern extraction with high precision. The proposed algorithm targets modern Arabic text that is not discretized and may contain some faults in spelling. The proposed algorithm is rule-based and can solve all Arabic morphological variations.

Keywords: *Use about five key words or phrases in alphabetical order, Separated by Semicolon*

1. Introduction

The web created new challenges for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users unexperienced in the art of web search.
The number of surfers depending on directories such as Yahoo Directory or Google Directory (ODP) is declining. Given that human maintained lists cover popular topics effectively, these lists are subjective, expensive to build and maintain, slow to improve, and cannot cover all topics. People are now likely to surf the web using search engines directly. The search engine is now the gate for the World Wide Web.
Many search engine giants like Google, Bing, and Yahoo tends not to use morphological analysis when indexing. They tend to index the whole text with the use of some semantics. Today, if Google was used and the word المدرسةwas search for, very differ-ent results will be produced when searching for the word مدارسهم knowing that the two words mean the same. And this problem gets bigger, when searching for the word الأكلthe web search engine will not consider words such as مأكولاتknowing that these two words are highly relevant.
This paper discusses the use of morphology analysis in Search Engines. The paper first suggests a new morphology analysis algo-rithm and its implementation in search engines.
Arabic language ranks fifth in the world's league table of lan-guages [5], with an estimated 280 million native speakers [4]. Arabic belongs to the Semitic group of languages which also in-cludes Hebrew and Amharic, the main language of Ethiopia.
There are many Arabic dialects. Classical Arabic – the language of the Qur'an – was originally the dialect of Mecca in what is now Saudi Arabia. An adapted form of this, known as Modern Stand-ard Arabic, is used in books, newspapers, on television and radio, and in conversation between educated Arabs from different coun-tries (for example at international conferences). This study will mainly deal with text written in Modern Standard Arabic.
The Lisan al-Arab (لسان العرب, "The Arab Tongue") by Ibn Manzur is among the most well-known and comprehensive dictionaries of the Arabic language. Ibn Manzur compiled it from other sources. Lisan al-Arab listed the roots of the Arabic language and then gave the meaning and other derivations of this root. Lisan al-Arab will be used as the main and initial Arabic root dictionary in this study.
Arabic words belong to one of two categories Original Arabic Words or Arabized Words. Original Arabic words are divided into two sub-categories.
Derived Arabic Words which are the verbs and nouns produced according to Arabic morphology patterns. Most of the words in Arabic text belong to this category.
Fixed Arabic Words which are a set of words that do not obey any morphology patterns and they do not transform. These are mostly functional words like pronouns, prepositions, and others. For ex-ample, the Arabic word أوis a fixed word it cannot be morpholog-ically transformed.

Copyright © 2018 Badr Nasiri et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Arabized words are nouns or verbs created from other foreign languages' words. These words are increasing in number and appear frequently in modern texts. Examples of Arabized words include مليون، دولار، كمبيوتر، إنترنت. All of these words are not included in dictionaries such as Lisan Al-Arab.

Arabic uses diacritics to remove the ambiguity, since two words may have the same characters but may differ in their meaning and in their pronunciation.

In modern Arabic texts, diacritics are rarely used. People depend on their knowledge of the language and the context to guess the missing diacritics while reading a non-diacritized text. In this paper, diacritics will be ignored and all modern text will be considered non-diacritized.

Arabic words are divided into three types: noun, verb, and particle. Nouns and verbs are derived from a closed set of around 10,000 words, called roots [1]. The roots are commonly three or four letters and are rarely five letters. Lisan al-Arab also lists very few six letter and 7 letter words. Three letter roots compose more than 70% of the total roots in Lisan al-Arab.

An Arabic word may contain more than one entity. An Arabic word constitutes of the main part of the word (a noun or verb or article), a prefix like the gender determiner or tense determiner or a definitive article (example, تذهبthe تin تذهبis a prefix to determine gender) and a suffix like a pronoun or gender determiner (example, كتابهاthe suffix هاis a pronoun).

A root is the base form of a word which cannot be further analyzed without the loss of the word's identity also it is the part of the word left when all the affixes (suffixes, prefixes and infixes) are removed. The root has a general, basic meaning which forms the basis of many related meanings which are formed from different patterns of the roots consonants.

More Arabic nouns and verbs can be derived from the roots. These are called stems. Stems are generated by applying a pattern to the root and then adding prefixes or suffixes. An example of a pattern is XAXX where X represents a letter in the root and A is the additional letter that formed the pattern. When applying the pattern XAXX on the root word عمل, the stem عاملwill be produced. Note that the generated stem is a noun since the pattern XAXX is a noun-generator pattern.

Also, prefixes and suffixes can be introduced to stems to generate more words. In the above example, the prefix الcan be introduced to have العاملor the suffix انcan be introduced to have the عاملان. A combination of more than one prefix and more than one suffix at the same time is also possible, for example, it can be said فالعاملانwhere the two prefixes فand الand the suffix انare added.

The high inflectional nature of the Arabic language arises because Arabic has two genders, feminine and masculine; three numbers, singular, dual, and plural; and three grammatical cases, nominative, genitive, and accusative. A noun has the nominative case when it is a subject; accusative when it is the object of a verb; and genitive when it is the object of a preposition. The patterns, suffixes and prefixes are used to inflect and generate the needed meaning.

Arabic Morphology (علم الصرف) is a branch of Arabic Grammar dealing with word-forms, patterns and stem generation. It is known to be a prime importance to acquire an understanding of word patterns when learning the language.

In this study, the researcher will rely heavily on the rules of Arabic Morphology to design the solution in question.

When building the search engine "Google", Brin and Page had the following objectives [3]. Fast crawling technology is needed to gather the web documents and keep them up to date. Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently. And last queries must be handled quickly, at a rate of hundreds to thousands per second.

Google was built to create a search engine which scales to the growing web [3]. Crawling, indexing, storage and fast querying were becoming increasingly difficult as the Web grows. Google found a solution and reached its goals but Google was poor in understanding the inflections of non-concatenative and highly inflectional languages such as Arabic.

Google approach was primary based on the matching process between the search words and the texts which is indexed when crawling. The matching process treated words as a set of symbols, not words with meanings to human users. Google strictly uses Symbolic Computing when searching, in addition, to using techniques to built relevancy between words.

For a language like Arabic, where the structure of a word can change according to many factors while maintaining the same meaning, "Symbolic Computing" does not retrieve an accurate set of results and there is an immense need for a linguistic approach to handle the search [3].

When indexing Arabic articles, Google indexes the whole Arabic word without doing any morphological analysis. This leads to different search results for different inflected Arabic words. For example, the word مدارسwill lead to a different search results than the word مدرسة. This simply means that the search engine giant Google did not relate between the two words on the meaning level.

Intelligent Search Engines is a search engine that tackles the problem of enhancing the precision and recall for retrieval of documents. The main techniques that they apply here are the use of subsumption information (or behavior based information) and the use of default information. The use of subsumption information allows for the retrieval of documents that include information about the desired topic as well as information about more specific topics [6].

The use of default information allows for retrieving of documents that include typical content information about a topic. The strict and default information are represented in an extension of description logics that can deal with defaults.

While retrieving information based on the root overflows the user with a complete but less relevant set of results, the user of the stem based search not only retrieves results in lesser numbers but also the set of results is less accurate.

Satya Sai Prakash et al, present architecture and design specifications for new generation search engines highlighting the need for intelligence in search engines based on intuition capturing. A knowledge framework has been built to capture and understand intuition [7].

The search engine was depicted by a simulation methodology showing behavior and performance. Simulation studies are conducted using fuzzy algorithm and heuristic search criterion after modeling the user behavior and web dynamics.

Dan Meng, Xu Huang discussed an interactive intelligent search engine model based on user information preference [8]. This model can be an effective and useful way to realize the individuation information search for different user information preference. This model framework used some artificial intelligent methods and technologies to improve the quality and effectiveness of information retrieval.

Xiajiong Shen Yan Xu Junyang Yu Ke Zhang forward an intelligent search engine where Information Retrieval model is found on formal context of FCA (formal concept analysis) and incorporates with a browsing mechanism for such a system based on the concept lattice. Test data validates its feasibility, and implement of the FCA-search engine indicates that the concept lattice of FCA is a useful way of supporting the flexible management of documents according to conceptual relation [9].

On another hand, Hattab M et al (2006) tried to enhance the search using different levels of morphological knowledge [10]. This paper enhances this approach by using related patterns on which the stems of the words are generated.

Due to the complexity of the Arabic language, Arabic text search may use three methods which are based on word, stem, or root. The word method is based only on exact term matching with some prefixes/suffixes truncating algorithms.

The root method is based on morphological analysis where the root is extracted and compared. This method will overflow the search results with irrelevant results since a different stems may be different in meaning although they share the same root. The stem

method will retrieve less irrelevant results but will miss other relevant results.

2. Related works

In this paragraph, we describe the structure of the Arabic language, although more emphasis is placed on those aspects which are important to information retrieval systems, such as affixation, morphology and derivation. This paragraph lists the state of the art in morphology analysis of the Arabic language and the state of the art in augmented search techniques.

2.1. Arabic morphology

Morphology is the identification, analysis and description of the structure of morphemes. A morpheme is the smallest component of a word, or other linguistic unit, that has semantic meaning. Morphemes cannot be split into smaller ones, and they should impart a function or a meaning to the word which they are part of. The root is the original form of the word before any transformation process, and it plays an important role in language studies. Morphology allows us to analyze an Arabic word and retrieve the stem, the root, the prefixes and suffixes.

For example, the word المدرسةin Arabic when undergone by a morphology analysis will result in Root as درس, Stem as مدرسةand Prefix as ال

Morphological typology represents a way of classifying languages according to the ways by which morphemes are used in a language. Arabic language would be classified as highly inflectional with stems being produced from roots based on several patterns. Each pattern creates a different semantic.

In Arabic, additions to the root can be within the root (not only on the word sides) which is called a pattern. This causes a serious issue in stemming Arabic documents because it is hard to differentiate between root particles and affix letters.

For example, for the root درس, adding the letter ا(infix) formed a different work such as دارسwith a different meaning. It can be said that the researcher have introduced the pattern فاعلto produce the new word دارسfrom the root درس.

Here, it is concluded that stems are derived from roots through the application of a set of fixed patterns. Addition of affixes to stems yields words. In addition, words sharing a root are semantically (but not totally) related and root indexing is reported to outperform stem and word indexing on both recall and precision [12].

For the purpose of this paper, a word is any Arabic surface form, a stem is a word without any prefixes or suffixes, and a root is a word without any prefixes, suffixes, or infixes. Roots are the units from which words are derived. However, often irregular roots, which contain double or weak letters, lead to stems and words that have letters from the root deleted or replaced.

There have been several algorithms that did Arabic morphology analysis; in this paragraph the state of the art will be listed and presented.

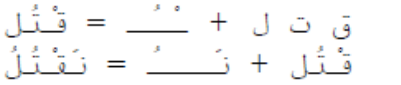
2.2. Incremental substitution method

Incremental Substitution algorithm rose from the idea that every morphological process can be modeled in terms of the composition of regular languages. This method allows for an elegant description of all main morphological processes present in natural languages including non-concatenative ones, mainly Arabic, in strict finite-state terms. The aim of this method is to create a linear rending of the finite state terms [13].

Buckwalter transliteration was used to help in composing clear regular expressions. The syntax of regular expressions used is that of xfst, the Xerox Finite State Tool.

The incremental substitution method works by [13] introducing a regular expression that enlists in a concatenative way all the morphemes (or rather, their lexical representations) which make up a word, in the order in which it should process their ‘merging’ with the string obtained at each phase. The Introduction of a subsequent regular expressions process their ‘merging’ with any intermediate string previously obtained, according to the order of the remaining tags at each point, ‘erasing’ one tag at a time after its surface counterpart has been created and merged to the rest.

Here, the method was able to give a linear rendering of what is assumed to be only a hierarchical representation or incremental creation of bigger building blocks from already elaborated ones.



Running such kind of machine on an Arabic text input will produce an output of all the encountered root bundles classified by the patterns they were found in.

2.3. Concatenative method

This method is based on the concept that Arabic inflectional morphology uses prefixes, suffixes and infixes. This method tries to reduce the complexity of Arabic morphology using an approach based on discrimination trees and transformational rules [14].

This method is based on a computational model that handles Arabic morphology generation concatenatively by separating the infixation changes undergone by an Arabic stem from the processes of prefixation and suffixation [14].

In this method, stems generation are described as being derived from a combination of a root morpheme and a vowel melody; where the two are arranged according to canonical patterns [14].

This approach seeks to reduce the number of rules for generating morphological variants of Arabic verbs by breaking the problem into two parts [14]. The first part is decoupling the problem of stem changes from that of prefixes and suffixes due to the very little interaction between stem changes and the addition of prefixes and suffixes. Since we may have the infix اas in فاعلand the prefixes and suffixes will be added in the same way independent of the infix inserted. فاعلbecomes يفاعلor يفاعلون. Similarly if we have another infix leading to stem تفاعلwe can have the same prefixes and suffixes added يتفاعلand يتفاعلون. This decoupling improves the space efficiency and its maintainability because the number of rules is reduced. The second part is simplifying the rules by isolating the different types of changes.

This method mainly targets developing machine translation systems. The two-step approach significantly reduces the number of morphological transformation rules that must be written, allowing the Arabic generator to be smaller, simpler, and easier to maintain [14].

2.4 Statistical method

Building large-scale morphological analyzers is typically a laborious and time-consuming task. For example, MORPHO3 was developed in 3 man/years [16]. The statistical method presents a quick method for performing shallow morphological analysis for finding the roots of words. The method is based on collecting statistics from word-root pairs for [15] building the rules of Arabic morphology that derives the roots from words, constructing a list of known prefixes and suffixes, and setting the probability that a rule will be used or a prefix or suffix is used.

The Statistical Method uses a list of Arabic word-root pairs to construct a list of prefixes and suffixes, to construct a list of stem templates, and to estimate the probability that a prefix, a suffix, or a template would appear.

The method then accepts Arabic words as input, tries to construct possible prefix-suffix template combinations based on the previously constructed lists and calculated probabilities, and outputs the possible roots [15].

This method takes a list of word-root pairs as input. By comparing the word to the root, the method determines the prefix, suffix, and stem template. For example, given the pair

(وكتابهم, كتب), the system generates "و" as prefix, "هم" as suffix, and "فعال" as stem template.
The method then looks-up the prefix "و" in the list of prefixes. If "و" is not found, "و" is added to the list of prefixes with a number of occurrences equal to one. If "و" is found, then the number of occurrences of the prefix is incremented by one. And this frequency will then be used to calculate the probability [15].
After the lists of prefixes, suffixes, and templates are constructed, the probability to each item on the list is estimated by dividing the occurrence of each item on each list by the total number of words [15].
This method will have an Arabic word as input, it will generate all possible prefixes and all possible suffixes and thus all possible stem patterns and will test these with the probability tables to output the most probable prefix, suffix and stem pattern and thus yielding the most probable root(s) [15].

2.5. Unsupervised learning method

Another approach that was proposed for performing morphology, in general, is the use of unsupervised learning techniques. Goldsmith proposed an unsupervised learning automatic morphology tool called AutoMorphology [17].
This method takes a text file as its input (typically in the range of 1,000,000 words) and outputs a morphological analysis of most of the words of the corpus; the goal is to produce an output that matches as closely as possible the analysis that would be given by a human morphologist. The method performs unsupervised learning in the sense that the program's sole input is the corpus; the program will not be provided with any dictionary or morphological rules particular to any specific language [17].
At present, the output of the program is the correct analysis of words into morphemes, though with only a rudimentary categorical labeling. The technique that this method is based on invokes the principles of the minimum description length (MDL) framework, which provides a helpful perspective for understanding the goals of traditional linguistic analysis. MDL focuses on the analysis of a corpus of data that is optimal by virtue of providing both the most compact representation of the data and the most compact means of extracting that compression from the original data. It thus requires both a quantitative account whose parameters match the original corpus reasonably well and a spare, elegant account of the overall structure [17].
This system is advantageous because it learns prefixes, suffixes, and patterns from a corpus or word-list in the target language without any need for human intervention. However, such a system would not be effective in Arabic morphology, because it does not address the issues of infixation, and would not detect rare prefixes and suffixes.

2.6. Morpho3

Morpho3 is a morphology analysis tool developed in the year 2000. Morpho3 was built to overcome challenges faced by the leading market tools Sakhr and Xerox [19].
Sakhr's Arabic morphological processes is considered till now one of the best tools that serves this purpose and many archiving and intelligence agencies depends on this system.
Nevertheless, Sakhr disadvantages are the patterns for Arabic derivations in Sakhr are fixed and and the statistical disambiguation in Sakhr suffers from lack of updates since the statistics are built from using one corpus ran once in time to produce the table of statistics.
On the other hand, Xerox is another well-known morphology analysis tool with the disadvantages that the patterns for Arabic derivations in Xerox are fixed and Xerox has no disambiguation mechanism.
The study of both systems leads to the development of Morpho3. Morpho3 allows the use of any pattern with any Arabic word. There is no fixed set of patterns for a specific word. Morpho3 algorithm operates the following steps [19] extract all possible prefixes, extract all possible prefixes, extract all possible suffixes, match the possible prefixes with the possible suffixes, extract all possible stems, dissolve the stems into roots and patterns and at last select the highest probable root/pattern (statistically).

2.7. Augmented search

This is the classical method for semantic search techniques where the search keywords are simply augmented with their synonyms, relevant keywords or similar words. The search is then performed using an OR operator to link the augmented keywords [20].
In this method, ontological techniques are used in a multitude of ways to augment keyword search, whether to increase recall or precision. In the following, a variety of approaches is presented.
Many query expansion implementations utilized in keyword search make use of thesaurus ontology navigation as a step in query expansion. Particularly utilized is the large WordNet ontology [20] for English language, defining synonym and meronym sets for words. A meronym denotes a constituent part of; for example, a finger is a meronym of hand.
The technique is simple. All function along the same basic scheme: first, the keywords are located in the ontology, then, various other concepts are located through graph traversal, after which the terms related to those concepts are utilized to either broaden or constrain the search.
In other related algorithms such as those mentioned in [20] and [21], terms are expanded to their synonym and meronym sets using the boolean OR operations available in most search engines. While in Clever Search [22], a particular meaning of a word in the WordNet ontology can be selected, resulting in the clarification text of that meaning being added to the search keywords via the boolean AND operator. This method is more restrictive in terms of search results. In the ontology navigation phase, the implementations differ mostly in which properties of the ontology are navigated and which terms are picked.
A very simple manner of augmenting traditional keyword search results is taken in the "Semantic Search" interface [23] of the TAP infrastructure. Here, in addition to a traditional keyword search targeted at a document database, the keywords are matched against concept labels in an RDF repository (W3C's Resource Description Framework). Matching concepts are then returned in addition to the located documents. The paper also sets a continuation of the search similar to the one described in [22], where, if multiple concepts match the keyword, the user can select his intended meaning to constrain the search. Here, however, the idea is not to expand search terms, but to use some procedure to classify the actual documents as pertaining or not pertaining to concepts, and constraining results based on that semantic annotation.
[24] describes an algorithm for locating additional information relevant to a query given a starting set obtained via text search. First, traditional text search is applied into a document collection. Then, a process of RDF graph traversal is begun from the annotations of those documents. The intention is to find related concepts such as the writer of the document or the project the document refers to in a general manner.
The traversal is done by a spread activation algorithm, for the use of which the arcs in the ontology are weighed according to general interestingness. This is calculated by combining a specificity measure favoring unique connections in the knowledge base, and a cluster measure, which favors links between similar concepts.
In an information retrieval environment it is well known that the relationship between a query and a document is determined primarily by the number and frequency of terms which they have in common. Unfortunately, words have many morphological variants, which will not be recognized by term-matching algorithms without some form of natural language processing. In most cases, these variants have similar semantic interpretations and can be treated as equivalent for information retrieval application.
In order to overcome the morphological variations of the word, a number of tools have been introduced to the information retrieval

environment such as truncation, morphological analyzer, prefix and suffix removal and Boolean operators.
Truncation supports searching on word stems. The use of truncation eliminates the need to specify each word variant, and thus simplifies the search strategies. The truncation technique is particularly useful in natural language information retrieval systems, where word variations are uncontrolled. There are two or three types of truncations: right-hand truncation (suffix removal), left-hand truncation (prefix removal) and middle truncation (infix removal). It is also useful for alternative spellings [39].
The idea of truncation, as [38] states, is based on the fact that, for information retrieval purposes, what matters is the basic concept represented by the first few letters of a word; the ending represents the syntactic function or some other subsidiary property, and is a positive nuisance if it prevents corresponding ideas being matched to one another. Hence, if the endings can be removed or transformed, variant forms can be reduced to a common "stem" and thus treated as equivalent.
Morphological analyzers for Arabic have been developed for various purposes. They differ in many ways. Some of the ways in which they differ are outlined in the following sections. Generally speaking, the existing Arabic morphological analyzers can be broadly divided into two types or categories: rule-based approach and statistical approach.

3. Proposed morphology processing algorithm

Morphology processing algorithm is the main algorithm used to establish the solution needed. In paragraph Two, the different state of the art algorithms have been listed. These algorithms are Incremental Substitution Method, Concatenative Method, Statistical Method, Unsupervised Learning Methods and Morpho3.
In this paragraph, the weaknesses and disadvantages of these methods will be pointed out. The morphology processing algorithm that will be used for the semantic search engine will also be presented.

3.1. Weaknesses and disadvantages of current methods

The incremental substitution does not primary deals with modern text. Arabic words should be properly diacritized for the method to work since the method does not include a disambiguation algorithm. This method cannot be used as part of the solution because the target text is modern non-diacritized text.
The Concatenative Method did not cover all kinds of weak verbs: defective and assimilated verbs. The method claims that other categories of words can be handled in a similar manner but no measurable tests have been done on that level.
The statistical method has several limitations. Some word patterns are rare that they may not appear in the training set. The analyzer lacks the ability to decipher which prefix-suffix combinations are legal. Although deciphering the legal combinations is feasible using statistics, the process would potentially require a huge number of examples to insure that the system would not disallow legal combinations.
The problem here is that the algorithm hardly detects infixation, and would not detect rare prefixes and suffixes. Also, the method requires a sum of training sets to get trained and start giving results.
Morpho3 uses statistical methods to determine the correct root. As per the above mentioned weaknesses, the training set may not be large enough to have all rare patterns included and thus it may not be 100% precise. Morpho3 on the other hand is considered one of the most precise algorithms mentioned above although not the faster.

3.2. Morphology processing algorithm needs

The morphology processing algorithm that is needed as part of the solution should have the following requirements.
There should be high precision in finding the stems and the roots of a word when working on modern Arabic text. There should be high precision in finding the stems and roots of Arabic foreign words. The algorithm should be rule-based to provide higher precision and make it independent on the training sets. And the rules are to be filled dynamically in case of updates. Only when there are two or more possible roots for one word, statistical disambiguation will be used. There should be high precision in rendering some faulty written words since modern written Arabic presents a unique range of orthographic problems. Different regional spelling conventions may appear together in a single text and show interference with spelling errors. For example, the ىcharacter may be written as يand vice versa also the ةmaybe written as a ه character.
The algorithm designed to cater for the above needs is rule-based. The algorithm will be described fully in the next paragraph. The algorithm operates by stripping away the prefixes and suffixes and stems are known according to patterns and rules and with reference to a roots dictionary. Then, if there is more than one possible root produces, statistical disambiguation will be used.

3.3. Building the morphological rules

First, in the rules the Arabic characters will be mapped with English characters. This mapping will make it easy to write the rules. The table below shows the developed mappings.

Table 1: Character Mapping

Arabic Character	English Character
و	W
ف	F
ب	B
ا	A
ن	N
ت	T
ي	Y
ل	L
إ	E
ه	H
م	M
س	S
ك	K
ء	2
ة	P

In this algorithm, a set of rules that maps prefixes and suffixes will be defined. Meaning, words may not have all prefixes and suffixes combinations. There are prefixes/suffixes that work only on verbs and others that work only on nouns. Moreover, some prefixes cannot be assigned simultaneously to the same word. For example, I cannot use the prefix بand the prefix لwith each other. It is morphologically wrong.
The following table shows the rules set for prefixes/suffixes combination for Arabic morphology.

Table 2: Prefixes/Suffixes Combination for Verbs

Prefix Set	Associated Suffix Set
{ "W", "F" }	{ "TN", "NA", "TMA", "N", "WA", "A", "TA", "T" }
{ "Y", "WY", "FY", "SY", "LY", "WSY", "WLY", "FSY", "FLY" }	{ "NAN", "N", "A", "WA", "WN", "AN" }
{ "T", "WT", "FT", "ST", "LT", "WST", "WLT", "FST", "FLT" }	{ "NAN", "Y", "A", "WA", "YN", "N", "WN", "AN" }
{ "N", "WN", "FN", "SN", "LN", "WSN", "WLN", "FSN", "FLN" }	{ "N" }
{ "A", "WA", "FA", "SA", "LA", "WSA", "WLA", "FSA", "FLA" }	{ "Y", "A", "WA", "N" }

Table 3: Prefixes/Suffixes Combination for Nouns

Prefix Set	Associated Suffix Set
{ "AL", "FAL", "KAL", "LL", "BAL", "WAL"}	{ "AT", "WN", "YN" }
{ "F", "K", "L",	{ "Y", "KM", "KMA", "K", "HM", "HMA", "H",

"B", "W" }	"HA", "AT", "WN", "YN", "ATKM", "ATKMA", "ATK", "ATHM", "ATHMA", "ATH", "ATHA" }

The following prefixes/suffixes set are put carefully using reference [25]. These rules are complete. Moreover for any update, the rules are dynamic and can be updated. The following examples explain the above mentioned rules.

Table 4: Examples of Correct and Wrong Affixation

Word	Correct Affixation	Wrong Affixation
بيت	البيت =بيت +ال البيتان =ان +البيت بكتاب =كتاب +ب	X =ه +البيت
كتاب	وبكتاب =بكتاب +و وبكتابهم =هم +وبكتاب	X =وبكتابهم +ف
ذهب	يذهب =ذهب +ي يذهبون =ون +يذهب	X =يذهبون +!

After setting the rules for the prefixes/suffixes, the rules for the patterns that produce the different stems from the root will be set. The patterns rules are described using regular expressions. Regular expressions provide a concise and flexible means for matching strings of text, such as particular characters, words, or patterns of characters. A regular expression is written in a formal language that can be interpreted by a regular expression processor, a program that either serves as a parser generator or examines text and identifies parts that match the provided specification.
The following table lists the regular expression patterns used to describe verb stemming patterns.

Table 5: Verb Patterns Using Regular Expressions

Regular Expression Pattern	Arabic Pattern
"^(\w)A(\w)(\w)$"	فاعل
"^E(\w)T(\w)(\w)$"	إفتعل
"^A(\w)T(\w)(\w)$"	افتعل
"^AN(\w)(\w)(\w)$"	انفعل
"^EN(\w)(\w)(\w)$"	إنفعل
"^T(\w)A(\w)(\w)$"	تفاعل
"^T(\w)(\w)(\w)$"	تفعل
"^ST(\w)(\w)(\w)$"	ستفعل
"^AST(\w)(\w)(\w)$"	استفعل
"^EST(\w)(\w)(\w)$"	إستفعل

Table 6: Noun Patterns Using Regular Expressions

Regular Expression Pattern	Arabic Pattern
"^(\w)A(\w)(\w)$"	فاعل
"^(\w)(\w)Y(\w)$"	فعيل
"^(\w)(\w)A(\w)$"	فعال
"^(\w)(\w)(\w)AN$"	فعلان
"^(\w)(\w)W(\w)$"	فعول
"^(\w)A(\w)W(\w)$"	فاعول
"^(\w)Y(\w)A(\w)$"	فيعال
"^(\w)(\w)(\w)Y$"	فعلي
"^(\w)Y(\w)(\w)$"	فيعل
"^(\w)(\w)(\w)A2$"	فعلاء
"^(\w)A(\w)(\w)Y$"	فاعلي
"^(\w)(\w)A(\w)Y$"	فعالي
"^M(\w)(\w)W(\w)$"	مفعول
"^M(\w)T(\w)(\w)$"	مفتعل
"^MT(\w)(\w)(\w)$"	متفعل
"^M(\w)(\w)A(\w)$"	مفعال
"^M(\w)(\w)(\w)$"	مفعل
"^M(\w)A(\w)(\w)$"	مفاعل
"^M(\w)A(\w)Y(\w)$"	مفاعيل
"^M(\w)(\w)Y(\w)$"	مفعيل
"^MN(\w)(\w)(\w)$"	منفعل
"^MST(\w)(\w)(\w)$"	مستفعل
"^A(\w)(\w)(\w)A2$"	افعلاء
"^T(\w)(\w)Y(\w)$"	تفعيل
"^T(\w)(\w)(\w)$"	تفعل
"^T(\w)(\w)(\w)YP$"	تفعلية
"^T(\w)A(\w)Y(\w)$"	تفاعيل
"^T(\w)W(\w)(\w)$"	تفوعل
"^T(\w)(\w)N(\w)$"	تفعنل
"^T(\w)A(\w)(\w)$"	تفاعل
"^TM(\w)(\w)(\w)$"	تمفعل
"^A(\w)A(\w)(\w)$"	افاعل
"^A(\w)T(\w)A(\w)$"	افتعال
"^A(\w)(\w)A(\w)$"	افعال
"^A(\w)(\w)N(\w)A2$"	افعنلاء
"^AN(\w)(\w)A(\w)$"	انفعال
"^A(\w)(\w)(\w)$"	افعل
"^A(\w)A(\w)(\w)$"	افاعل
"^AST(\w)(\w)A(\w)$"	استفعال
"^E(\w)T(\w)A(\w)$"	إفتعال
"^E(\w)(\w)(\w)A2$"	إفعلاء
"^E(\w)(\w)N(\w)?A2$"	إفعنلاء
"^EST(\w)(\w)A(\w)$"	إستفعال
"^EN(\w)(\w)A(\w)$"	إنفعال

The previous tables list the patterns for verbs and nouns regarding 3-letter roots. Three letter roots (فعل) are the most frequent in Arabic languages. The patterns of the 4-letter roots will not be listed but as previously noted the rules are dynamic and one can easily add 4-letter root and 5-letter roots patterns.
To make the algorithm more precise and to lessen processing needs, a list of articles/prepositions has been fed as rules with their transformations also in regular expressions format. I will mention only some samples of these rules (they are a total of 92 rules).The following table lists the articles and their possible patterns.

Table 7: Articles Patterns

Article	Possible Patterns
أخرى	[W]?(F|B|L|K)?(WORD)
لدن	(W|F|L|K)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
إلا	[W]?(F|L)?(WORD)
إلى	(W|B|F|L)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
أم	[W]?(F)?(WORD)
أما	[W]?(F)?(WORD)
ذلك	[W]?(B|F|L|K)?(WORD)
ذو	[W]?(F|B|L|K)?(WORD)
ذي	[W]?(F|B|L|K)?(WORD)
ريثما	[W]?(F|B|L|K)?(WORD)
شتى	[W]?(F|B|L|K)?(WORD)
على	[W]?(B|F|L|K)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
عن	[W]?(B|F|L|K)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
عند	[W]?(B|F|L|K)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
عندما	[W]?(F|K)?(WORD)
غير	[W]?(B|F|L|K)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
فوق	[W]?(F|B|L|K)?(WORD)
في	[W]?(B|F|L)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
فيما	[W]?(B|F|L)?(WORD)
قد	[W]?(F|L|FL)?(WORD)

The above table shows what each article can have as prefixes or suffixes. This knowledge is important to make the morphology processor more accurate and precise. For example, the article في can have the suffix ها to make the new article فيها but the article أما can't have the suffix ها.
Moreover, a set of rules has also been developed for geographical places, such as countries and cities. Such words are frequently used in news text which is frequently the target of search queries.
An additional set of rules has also been compiled to solve some cases for example the following rule

{"^AT(\w)A(\w)$", "START{و}"}

Has been developed to solve the stemming of roots starting with و for the pattern افتعال. For example, the root بزل with pattern افتعال will produce ابتزال while the root وحد with patter افتعال will produce اتحاد the و will be dropped. This rule solves this case. And as mentioned previously, the rules are dynamic additional rules can be added to solve such cases.

A dictionary of 10,000 roots from [1] has also been prepared for the algorithm to work on. For each entry in the dictionary, a probability value will be attributed. This probability will be used in roots disambiguation. Notice that [1] does not contain new used Arabic words such as دولار. These words should be added to the roots while testing with the test corpus.
In conclusion, the knowledge rules prepared for the algorithm to work effectively are a list of prefixes/suffixes morphologically correct combinations, a list of verb patterns, a list of noun patterns, a list of articles and their associated patterns, a list of geographical places patterns, a list of rules to solve specific stemming cases, a list of roots from [1] and a list of geographical places.

3.4. The algorithm

The algorithm should take as an input any word in any text and should produce a quadruple which is (prefix, suffix, stem, root).
For example, the input يتكاتبونshould produce the output (ي, ون, تكاتب, كتب).
The following are the steps that the algorithm performs to reach the quadruple.
Step 1: Match if the word is an article with any of the article patterns
The word will be tested on all available article patterns checking if this word is an article. If a pattern returns true, the algorithm stops now knowing directly the suffix, prefix, stem and root. In the case of an article, stemming does not exists. The root and the stem are the same. The patterns refer only to the use of prefixes/suffixes and not infixes. For example, if the word is بينهاthe processor directly will match it with the regular expression
(W|B|F|L|K)?(WORD)(H|HA|HMA|HM|HN|K|KMA|KM|NA)?
Here, the quadruple will be (,ها, بين, بين).
Step 2: Get Suggested Verb Stems
The algorithm will start to strip the possible prefixes from the word. The possible prefixes are already loaded in the prefixes list. The possible prefixes are stripped and the results are added as suggested stems if the length of the resulting stem is more than 2.
Then the suffixes pertaining to the list of prefixes stripped above will now be tested on the word in an attempt to strip them. The result of this stripping process is also added to the suggested stems list if the length of the resulting stem is more than 2.
Step 3: Get Suggested Noun Stems
The algorithm will start to strip the possible prefixes from the word. The possible prefixes are already loaded in the prefixes list. The possible prefixes are stripped and the results are added as suggested stems if the length of the resulting stem is more than 2.
Then the suffixes pertaining to the list of prefixes stripped above will now be tested on the word in an attempt to strip them. The result of this stripping process is also added to the suggested stems list if the length of the resulting stem is more than 2. Also in this step, testing on geographical patterns are geographical places list is performed.
Step 4: Find the Correct Roots
Now, the suggested stems of the verbs are tested among the regular expressions of the verb patterns and the suggested stems of the nouns are tested among the regular expressions of the noun patterns. The resulting matches constitute the possible roots of the word. If there is only 1 root, the algorithm will end declaring the quadruple. If there is more than one root, the algorithm should continue with the disambiguation process.
Step 5: Roots Disambiguation
For each root in the dictionary, a probability value will be set by having these roots pass by different training sets. The root with the highest probability will be declared as the correct root and the algorithm will return the quadruple accordingly.
In the following table, an example of the algorithm on the word أنباءwill be implemented.

Table 8: Example of Algorithm Implementation

Step 1: Match if the word is an article with any of the article patterns	No matches for أنباءin the Articles regular expressions
Step 2: Get Suggested Verb Stems	The prefix أis removed and the suggested stem is نباء
Step 3: Get Suggested Noun Stems	No prefixes or suffixes found to strip
Step 4: Find the Correct Roots	The possible verb stems from Step 2 are: أنباءand نباء Testing on the list of verb patterns – no match found. The possible noun stems from Step 3 are: أنباء Testing on the list of noun patterns – there is two matches: The pattern فعلاءfrom the root أنب The pattern أفعالfrom the root نبأ Both of the two roots are found in the root list. Step 4 produced 2 correct roots أنبand نبأ
Step 5: Roots Disambiguation	Statistical Disambiguation shows that نبأ has higher probability than أنب. Then, the quadruple will be (,,أنباء,نبأ)

Figure 1 depicts the algorithm flowchart showing the different five steps of the algorithm. The process starts with the Arabic word as the input. Then the prefix/suffix handling and the testing of the different patterns, the roots extraction and at last the statistical disambiguation process.
In this paper, we propose a new morphological algorithm based on morphological rules. References have been used to collect all Arabic morphological rules and the rules have been related to the different possible patterns and suffixes. The algorithm overcomes the obstacles faced by the state of the art morphological analyzer to produce a solution that will be used in the design of the augmented search algorithm presented in the next paragraph.
The algorithm tries to guess the correct patter by discovering the several suffixes and prefixes and trying to find the correct stem pattern by matching with the list of roots extracted from Lisan Al Arab dictionary. If more than one root is found, statistical disambiguation will be used to select the high probable root.

4. Building the Arabic semantic search engine

In this paragraph, the experiment done that tends to relate the stemming patterns together will be introduced. The experiment will be used to build the semantic search engine based on morphology processing and relating stemming patterns. A pattern relevancy matrix will be built. This matrix will be based on statistical methods to relate words having the same root with different morphological variations.

4.1. Limitations of the current search methods

After the implementation of the morphology process, the design of the Arabic Semantic Search Engine starts. Light stemming is mostly used for indexing and searching in current solutions [37]. Research on Arabic information retrieval tends to treat automatic indexing and stemming separately. Light stemming does not provide a semantic foundation for the search.
Using morphological analysis to support information retrieval in Arabic has lead to some differences of opinion as to the most suitable methods for information retrieval (word, stem or root methods). Some believe that the Arabic information retrieval system should be based on morphological analysis in order to retrieve the word by its root. They justify their view by stating that Arabic is a derivative language; therefore the system should be based on the root of the word.
They give the following example: suppose that the user wants to search for the word كتب(wrote); if the system is based on word search only then it will retrieve the word as it is entered. However if the system was based on root searching, the system will retrieve all forms of the root كتبsuch as: يكتب, نكتبand others.

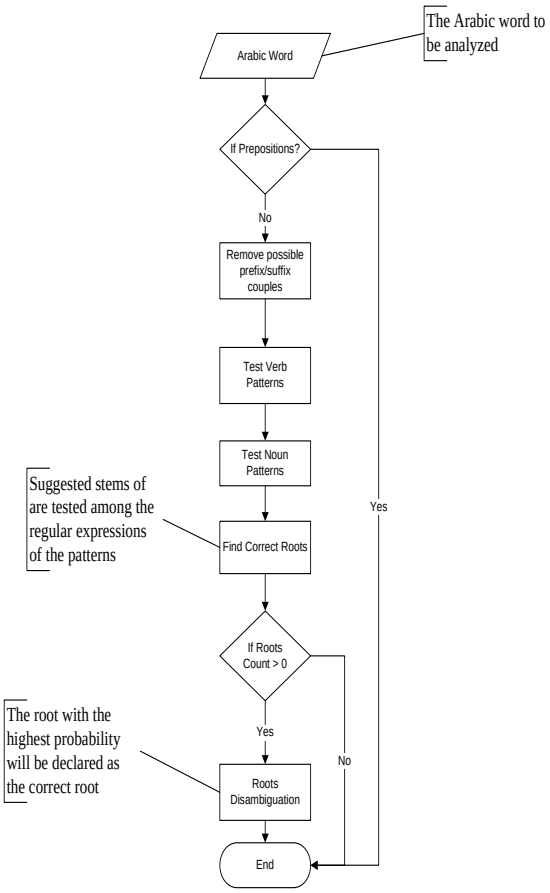


Fig. 1: Algorithm Flowchart.

Some information specialists, on the other hand, do not believe there to be a useful relation between information retrieval and the root of the word. Among examples they give the following: from the root جمع there can be derived a number of words: إجتماع(meeting), مجموع(summation), جمعة(Friday), جماعة(group), جامع (mosque), جامعة(university). Thus, they justify their rejection of the root method by stating that if the information retrieval were to be based on this system it would retrieve all the above words though there is no real need to do so.

In order to overcome the morphological variations of the word, a number of techniques such as truncation, stemming algorithms, and morphological analyzers have been introduced into information retrieval systems to improve the retrieval performance. In Arabic information retrieval systems, three search methods are used: namely, word, stem, and root. The word method is based only on term matching, while the other two methods are based on morphological analysis. These have different levels of morphological analysis. However, each of these has its limitations. For example, the word and stem methods may miss relevant records (because of morphological variations). On the other hand, the root method may retrieve irrelevant records. This is due to the fact that the root method is capable of reducing a given word to its root, and then it will generate all possible morphological variations of that word.

The limitations of the current search methods have motivated the investigation of a novel approach to be used in an Arabic information retrieval system. It is hoped that this method will improve the effectiveness of the word and stem methods in terms of retrieving more relevant records than the previous word and stem methods did. At the same time, it is also hoped that the same method will improve the root method in terms of rejecting irrelevant records that may be retrieved by the root method.

This method is based on augmented search. This means that regarding indexing of text the stem of each word will be indexed and regarding search keywords the stems from the keyword will be retrieved and use it augmented by other relevant stems from the same root. This method basically depends on relating Arabic patterns together to have more relevant search results.

4.2. The experiment

The experiment conducted explains the level of relevancy between the stemming patterns. The experiment will be limited to 3-letter roots with their possible patterns. To conduct this experiment, a corpus containing more than 1,000,000 distinct words was used. For these words, the morphology processing application and attributed the pattern and root for each of the words was implemented. Then, relevancy matrices between the different possible patterns were created.

For example, it was found out that the noun pattern فاعل is very relevant to the root فعل. دارس is always related to درس, كاتب related to كتب, حامد related to حمد. There are few words that may not be considered very relevant such as نائب may not be much related to ناب since ناب is a general word while نائب has somehow a political meaning. Table 9 shows the pattern relevancy matrix where each pattern is related to the other pattern by a statistical percentage.

Table 9: Pattern Relevancy Matrix

	فعل	افتعل	تفاعل	انفعل	استفعل	Relevancy Examples
فاعل	92%					طالب from طالب [Relevant]
فعيل	71%					صغير from صغر [Relevant]
فعال	57%					كليم from كلم [Relevant]
فعلان	23%					مثال from مثل [Relevant]
فعول	42%					سعال from سعل [Relevant]
فاعول	12%					كلاب from كلب [Relevant]
فيعال	25%					غضبان from غضب [Relevant]
فعلي	75%					إنسان from أنس [Not relevant]
فيعل	0%					سلمان from سلم [Not relevant]
فعلاء	81%					كتوم from كتم [Relevant]
فاعلي	0%					رسول from رسل [Not relevant]
فعالي	0%					صعوبة from صعب [Relevant]
مفعول	65%					حاسوب from حسب [Not relevant]
مفتعل	11%	100%				قابوس from قبس [Relevant]
						شيطان from شطن [Not relevant]
						ديوان from دون [Not Relevant]
						ريماس from رمس [Relevant]
						لبناني from لبنان [Relevant]
						فعلي from فعل [Not Relevant]
						كرماء from كرم [Relevant]
						فعالية from فعل [Not Relevant]
						مكتوب from كتب [Not Relevant]
						مدروس from درس [Relevant]
						مستور from ستر [Relevant]
						مبتكر from بكر [Not Relevant]
						مبتكر and ابتكار [Rel-

مفعال	50%			evant] The machine name of مفتاح likeفعل
مفعل	50%			صنع fromمصنع [Relevant] ذهب fromمذهب[Not Relevant]
مفاعل	66%			كبح fromمكابح[Relevant] سمع fromمسامع [Relevant] قبر fromمقابر[Relevant] فعل fromمفاعل[Not Relevant] as in مفاعل نووية
مفاعيل	21%			جمع fromمجاميع [Relevant] نشر fromمناشير [Relevant]
منفعل	0%	100%		قلب fromمنقلب[Not Relevant] انفعل andمنفعل are relevant
مستفعل	0%		100%	استقل andمستقل are relevant
افعلاء	50%			استشهد andمستشهد are relevant صدق fromاصدقاء [Not Relevant] شيء fromاشياء [Relevant]
تفعيل	0%			شغل fromتشغيل[Not Relevant]
تفعل	91%			سمع fromتسمع[Relevant]
تفاعيل	0%			ضرس fromتضاريس [Not Relevant]
تفوعل تفعنل				قلب fromتقولب
تفاعل	0%	100%		
تمفعل افاعل				Relevant to مفعل pattern only
افتعال		100%		
افعال	83%			
انفعال			100%	
افعل	100%			افضل onافضل[Relevant]
افاعل استفعال				100%

According to the above matrix, two types of relevancies can be categorized as: primary, which has a relevancy above 70% and secondary which has a relevancy above 20% and below 70%. Accordingly, the following tables list the relevancies.

Table 10: Patterns with Primary Relevance

فعيل ,افعل ,فعلي ,فعلاء ,تفعل ,افعال ,فاعل ,فعل
مفتعل ,افتعل ,إفتعال
متفاعل ,تفاعل ,تفاعل
انفعال ,منفعل
استفعال,استفعل ,مستفعل

Table 11: Patterns with Secondary Relevance

مفعول ,فعال ,فعلان ,فعول ,مفعال ,مفعل ,مفاعل ,مفاعيل ,افعلاء ,فعل

4.3. Indexing of Arabic text

In this example, the Arabic Text should be indexed by stemming each word. For example, having this Arabic text paragraph retrieved from a daily newspaper:

اعتذرت الخطوط الجوية السعودية عن تجاوزات بعض موظفيها نحو المسافرين خلال الفترة الماضية، متوعدة المخطئين منهم بعقوبات قاسية تصل إلى الفصل من العمل، وأكد مدير عام الخطوط الجوية العربية السعودية، المهندس خالد بن عبدالله الملحم، أن المؤسسة ستتعامل بحزم مع موظفيها الذين يسيؤون للمسافرين، وأنها لن تتوانى عن إيقاع العقوبات النظامية بما فيها الحسم من الراتب والفصل من الخدمة لأي موظف لا يلتزم بمتطلبات العمل، وفي أولوية ذلك وجوب الالتزام بالتعامل الراقي والحضاري مع المسافرين.

This text will be indexed as follows:

اعتذر خطوط جوي سعودي عن تجاوز بعض موظف نحو مسافر خلال فترة ماضي، متوعد مخطئ من عقوبة قاسي تصل إلى فصل من عمل أكد مدير عام خطوط جوي عربي سعودي، مهندس خالد بن عبدالله الملحم أن مؤسس تتعامل حزم مع موظفي الذين سيؤ مسافر أن لن تتوانى عن إيقاع عقوبة نظامي بما في حسم من راتب فصل من خدم أي موظف لا التزم متطلب عمل في أولوي ذلك وجوب التزام تعامل راقي حضاري مع مسافر

This stemming is the result of applying the morphological processor described above. Having the text to search stemmed will make it easy to implement the semantic search engine and search for exact stems or relevant stems.

[31] extended the Arabic text indexing to provide weights for each word in the text based on the rate of occurrence of the word or its morphological variations and how it is spread. In the above example, only retrieving the stems of each word is being retrieved since the proposed search algorithm will search for the exact stem and the stems with related pattern.

4.4. The algorithm

The algorithm relies heavily on the relevancy tables produced previously. The algorithm supposes that the search text contains only stems.

Step 1: Determine the stems and pattern of each word in the search query.

The algorithm loops over all the words in the search query and using the morphological processor produces the stems and patterns of each of the words in the search query. For example, if I have the search query "الأكل البسيط" the algorithm will produce the following in this step.

Table 12: Sample Search Query

Search Word	Stem	Pattern
الأكل	أكل	فعل
البسيط	بسيط	فعيل

Step 2: Construction of a new augmented search query based on the Relevancy Tables.

The algorithm will construct a new augmented search query based on Primary Relevancy Table. The augmented search will use the OR Boolean operator. Concerning the example "الأكل الصيني", the algorithm will produce the following search query:

"(أكل OR أكلي OR (أكيل AND (بسط OR بسيط OR باسط OR تبسط OR بسطاء OR بسطي OR ابسط)"

Step 3: The search engine will perform the query and will produce a list of secondary relevant search results.

The search engine will operate the query on the stemmed texts and will return search results. With the search results, the search engine will also show other relevant search keywords (secondary keywords). In this example, the search engine will show مأكول بسيط and مأكل بسيط andأكول بسيط

and other combinations according to the relevancy tables. Users using the search can see these secondary suggested search results and will use them for their next search if it was relevant to them.

In this paragraph, the semantic search algorithm is designed from the patterns relevancy table where the patterns are related to each other by statistical methods. According to the relevancy table, the search algorithm will provide more relevant search results by linking the morphology patterns to each others. In the next paragraph, testing sample search queries will be performed and the algorithm efficiency will be compared with the other known methods.

5. Testing and validation

In this paragraph, we will test the algorithm and validate the results by comparing the results with three known search methods for Arabic Language, the word method, the stem method and the

root method. This paragraph will prove the efficiency of the algorithm with respect to retrieval of relevant records.

5.1. Testing the algorithm

The previous paragraph was a description of the implementation of the semantic search algorithm. In order to run the experiments and to evaluate the algorithm, a sample of 1,000 Arabic records was used as a database.
The sample was selected to be representative of Arabic texts and was selected from various Arabic news sources including Kuwait News Agency (KUNA) website kuna.net.kw and Al-Arabiya website alarabiya.net.
This section discusses the outcome of the evaluation. It starts with a data representation of the main findings of the evaluation. The study uses four parameters to evaluate the retrieval performance of each method of search: word, stem, root, and the proposed semantic algorithm.
Table 13 shows the total retrieved records by the four methods: word, stem, root and the proposed semantic search. Column 1 shows the number of queries being used in this study.

5.2. Results validation

Table 14 shows the relevant and irrelevant records retrieved by each method. As far as irrelevant records are concerned, the root method retrieved more irrelevant records than the other methods. The root method retrieved 70 irrelevant records out of 165. In other words, 42% of records being retrieved by the root method were irrelevant. The second method, which brings more irrelevant records (after the root method) was the proposed semantic method. Out of 118 records retrieved by the semantic method, it was found that 27 (23%) records were irrelevant.

Figure 2 shows a bar graph depicting the performance of each method by comparing the 4 values: Retrieved, Relevant and Retrieved, Irrelevant and Retrieved, Relevant and Not Retrieved. It could be clearly observed that the Semantic method out-performed the other methods.

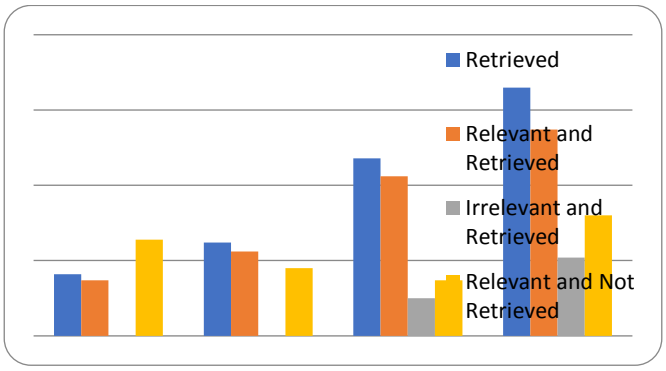


Fig. 2: Methods Comparison Chart.

In this paragraph, an evaluation was also done to compare the retrieval performance of the semantic search method against the three methods of search used in information retrieval for Arabic: namely, word, stem, and root. An important feature of this work is the introduction of the semantic method based on morphology analysis to be used in information retrieval for the first time. The results give a clear indication that the method has great potential for improving the retrieval performance of the word and stem methods.

6. Conclusions and future work

In Arabic information retrieval systems, three search methods are used namely, word, stem, and root. The word method is based only on term matching, while the other two methods are based on morphological analysis, (they have different levels of morphological analysis). However, each method has its limitations. For example, the word and stem methods may miss relevant records (because of morphological variations), while, on the other hand, the root method may retrieve irrelevant records. This is due to the fact that the root method is capable of reducing a given word to its root, and then it will generate all possible morphological variations of that word. In this section, the findings, limitations and future work will be listed.
Having identified some limitations of the current search methods in Arabic, the present study has introduced an approach based on what is called the semantic augmented search method. The aim of this method is to improve the effectiveness of the word and stem methods in terms of retrieving more relevant records and, at the same time, it is hoped that this proposed solution will improve the root method in terms of rejecting irrelevant retrieved records. This approach has used a morphology analysis algorithm that is also depicted in this research.
To implement this, a knowledge representation technique is used which is the semantic relations. Within the semantic relations, a number of links have been made between related morphological forms.
During the evaluation of the semantic search algorithm, the results show that the morphological variations linking improved retrieval of relevant records and reduced the retrieval of irrelevant records.
A main limitation was that, the study used only one approach of linguistic analysis which is morphological analysis. Other linguistic approaches, such as syntactic, semantic, and pragmatic analyses are beyond the scope of the study.
Another limitation is that this approach is only for the Arabic language and cannot be implemented for other languages due to the use of morphological patterns and their relations which are unique to the Arabic language.
The study results of the morphology and semantic method have motivated us to carry out more work in this area. This involves the extension of the work on semantic linking of morphological forms.

As shown in the table, the total number of queries was 10. The second column of the same table shows the query statement in the Arabic language. Column 3 indicates the relevant records available in the database. The remaining columns (i.e. 4, 5, 6 and 7) show the number of records retrieved by all four methods of search.

Table 13: Sample Search Query

No.	Query	Relevant	Word	Stem	Root	Semantic Search
1	الإستثمار	12	5	8	18	10
2	المصارف	26	8	12	20	20
3	الديون	15	6	10	17	17
4	المحاصيل	3	1	1	10	3
5	العاملة	4	3	3	15	12
6	التلوث	8	4	5	10	10
7	الجامعة	12	5	9	25	20
8	حقوق	6	3	6	12	8
9	العقود	5	2	2	10	6
10	الأموال	10	4	6	28	12

Table 14: Relevant/Irrelevant Records Retrieved

	Relevant				Irrelevant			
No.	Word	Stem	Root	Seman-tic	Word	Stem	Root	Seman-tic
1	5	8	18	10	0	0	6	2
2	8	12	20	20	0	0	0	0
3	6	10	17	17	0	0	2	2
4	1	1	10	3	0	0	7	0
5	3	3	15	12	0	0	11	8
6	4	5	10	10	0	0	2	2
7	5	9	25	20	0	0	13	8
8	3	6	12	8	0	0	6	2
9	2	2	10	6	0	0	5	1
10	4	6	28	12	0	0	18	2
	41	62	165	118	0	0	70	27

The prototype which was used in this study, could be improved via in-depth analysis of Arabic morphology. The in-depth analysis should lead to better relations between morphology patterns. Also, it would be important if these relations change according to context or according to other statistical values.
The future work should also include completing the implementation of the morphology analyzer algorithm by understanding and setting all the rules/patterns of Arabic morphology.

References

[1] Ibn Manzur, "Lisan El Arab," Dar Kotob Al Ilmiyah; Latest Edition, 2006.

[2] A. Chen and F. Gey, "Building an Arabic Stemmer for Information Retrieval," in Proc. The Eleventh Text Retrieval conference, National Institute of Standards and Technology (NIST), 2002.

[3] S. Brin and L. Page, "The Anatomy of a Large Hypertextual Web Search Engine," in Proc. Seventh International World-Wide Web Conference, Australia, 1998 https://doi.org/10.1016/S0169-7552(98)00110-X.

[4] Wikipeida Website on Arabic Language [Online], Available: http://en.wikipedia.org/wiki/Arabic_language

[5] G. Weber, "Top Languages - The World's 10 most influential Languages" [Online], Available: http://www.andaman.org/BOOK/reprints/weber/rep-weber.htm.

[6] S. Malik, N. Prakash, S. Marwaha, "Role of Search Engines in Intelligent Information Retrieval on Web," in Proc. The 2nd National Conference, INDIA COM, 2008.

[7] K. Satya Sai Prakash and S. V. Raghavan, "Intelligent Search Engine: Simulation to Implementation," in Proc. 6th International conference on Information Integration and Web-based Applications and Services (iiWAS2004), Jakarta, Indonesia, 2004, pp. 203-212.

[8] D. Meng and X. Huang, "An Interactive Intelligent Search Engine Model Research Based on User Information Preference," in Proc. 9th International Conference on Computer Science and Informatics, 2006. https://doi.org/10.2991/jcis.2006.103.

[9] X. Shen, Y. Xu, J. Yu, K. Zhang, "Intelligent Search Engine Based on Formal Concept Analysis" in Proc. IEEE International Conference on Granular Computing, 2007. https://doi.org/10.1109/GrC.2007.62.

[10] M. Hattab, B. Haddad, M. Yaseen, A. Duraidi, A. Abu Shmais, "Addaall Arabic Search Engine: Improving Search based on Combination of Morphological Analysis and Generation Considering Semantic Patterns", 2007, [Online]. Available: http://www.uop.edu.jo/download/cv/Addlaall-Search-Engine--Hattab-Haddad-Yaseen-UOP.pdf.

[11] S. Kumar and S. Kumar Malik, "Towards Semantic Web Based Search Engines," presented at National Conference on Advances in Computer Networks & Information Technology (NCACNIT-09), 2009.

[12] I. Hmeidi, G. Kanaan, M. Evens, "Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents," Journal of the American Society for Information Science, 48/10, 1997, pp. 867-881. https://doi.org/10.1002/(SICI)1097-4571(199710)48:10<867::AID-ASI3>3.3.CO;2-R.

[13] S. Jaber and R. Delmonte, "Sarrif – The Elegant Arabic Morphology Parser," in Proc The 9th international conference on Computational Linguistics and Intelligent Text Processing, Springer-Verlag Berlin, Heidelberg, 2008.

[14] V. Cavalli-Sforza, A. Soudi, T. Mitamura, "Arabic Morphology Generation Using a Concatenative Strategy," presented at 1st North American chapter of the Association for Computational Linguistics conference, San Francisco, CA, USA, 2000.

[15] K. Darwish, "Building a Shallow Arabic Morphological Analyzer in One Day," in Proc. The ACL-02 workshop on Computational Approaches to Semitic Languages, Stroudsburg, PA, USA 2002 https://doi.org/10.3115/1118637.1118643.

[16] M Attia (2000), "A Large-Scale Computational Processor of the Arabic Morphology, and Applications," [Online], Available: http://www.nemlar.org/Publications/M_A_Thesis2000.pdf.

[17] J. Goldsmith (2000), "Unsupervised Learning of the Morphology of a Natural Language," [Online], Available: http://humanities.uchicago.edu/faculty/goldsmith.

[18] A. N. De Roeck, W. Al-Fares, "A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots", in Proc. The 38th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, 2000 https://doi.org/10.3115/1075218.1075244.

[19] M. Attia, "A Large-Scale Computational Processor of the Arabic Morphology, and Applications," M.Sc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.

[20] D. I. Moldovan and R. Mihalcea, "Using Wordnet and Lexical Operators to Improve Internet Searches," IEEE Internet Computing Journal, Vol. 4, 2000, pp. 34–43 https://doi.org/10.1109/4236.815847.

[21] D. Buscaldi, P. Rosso, E.S. Arnal, "A Wordnet-Based Query Expansion Method for Geographical Information Retrieval," presented in the 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain.

[22] P.M. Kruse, A. Naujoks, D. Roesner, M. Kunze, "Clever Search: A Wordnet Based Wrapper for Internet Search Engines", in Proc. The 2nd GermaNet Workshop, Bonn, Germany, 2005.

[23] R. Guha, R. McCool, E. Miller, "Semantic Search Meets the Web," in Proc. The 12th international conference on World Wide Web, ACM Press, 2003, pp. 700–709 https://doi.org/10.1145/775152.775250.

[24] C. Rocha, D. Schwabe, M.P. de Aragao, "A Hybrid Approach for Searching in the Semantic Web," in Proc. The 13th international conference on World Wide Web, 2004, pp. 374–383. https://doi.org/10.1145/988672.988723.

[25] M. E. Muhammad, "From the Treasures of Arabic Morphology," Zam Zam Publishers, 2005, pp. 1-359

[26] E. Adams, "A Study of Trigrams and their Feasibility as Index Terms in a Full Text Information Retrieval System," PhD Dissertation, George Washington University, USA, 1991.

[27] S. Al-Fedaghi and F. Al-Anzi, "A New Algorithm to Generate Arabic Root-Pattern Forms," in Proc. The 11th National Computer Conference, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, 1989, pp. 04-07.

[28] I. Al-Kharashi and M. Evens, "Comparing Words, Stems, and Roots as Index terms in an Arabic Information Retrieval System," Journal of the American Society for Information Science, 45/8, 1994, pp. 548-560. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<548::AID-ASI3>3.0.CO;2-X.

[29] K.B. Beesley, "Arabic Morphological Analysis on the Internet," in Proc. The 6th International Conference and Exhibition on Multi-Lingual Computing, Cambridge, 1998.

[30] M. F. Porter (1980), "An Algorithm for Suffix Stripping," Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 14/3, pp. 130-137. https://doi.org/10.1108/eb046814.

[31] N. Mansour, R. Haraty, W. Daher, M. Houri, "An Auto-Indexing Method for Arabic Text", Information Processing and Management: An International Journal, Volume 44 Issue 4, July 2008. https://doi.org/10.1016/j.ipm.2007.12.007.

[32] H. Al-Haj and A. Lavie, "The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation," in Proc. The Ninth Conference of the Association for Machine Translation in the Americas (AMTA-2010), Denver, Colorado, 2010.

[33] M. Gridach and N. Chenfour, "Design and Realization of an Arabic Morphological Automaton: New Approach for Arabic Morphological Analysis and Generation," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011. https://doi.org/10.1155/2011/629305.

[34] M. Al-Sadat Hoseini, "Semantic Processing of Arabic Language," Journal of American Science, Vol. 7, No. 4, 2011, pp. 174-178

[35] A. Y. Samarah, "Arabic Linguistics and Sibawaihi," International Journal of Academic Research", Vol. 3, No. 2, 2011

[36] M. Aljlayl, "Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach," in Proc. ACM eleventh conference on Information and Knowledge Management, 2002. https://doi.org/10.1145/584845.584848.

[37] L. S. Larkey, L. Ballesteros, M. E. Connell, (2002), "Light Stemming for Arabic Information Retrieval," [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.877&rep=rep1&type=pdf

[38] C. D. Paice, "Method for Evaluation of Stemming Algorithms Based on Error Counting," Journal of American Society for Information Science, 47 (8), 1996, pp. 632-649. https://doi.org/10.1002/(SICI)1097-4571(199608)47:8<632::AID-ASI8>3.0.CO;2-U.

[39] J. Rowley, "The Electronic Library", London: Library Association Publishing, 1998.