FULL AUTOMATIC ARABIC TEXT TAGGING SYSTEM

GHASSAN KANAAN
Yarmouk University
Irbid-Jordan
ghassank@yu.edu.jo

RIYAD AL-SHALABI
Yarmouk University
Irbid-Jordan
shalabi@yu.edu.jo

MAJDI SAWALHA
Yarmouk University
Irbid-Jordan
maj_sawalha@yahoo.com

ABSTRACT

Part-of-Speech tagging is the process of assigning grammatical part-of-speech tags to words based on their context. Many automated tagging systems have been developed for English and many other western languages, and for some Asian languages, and have achieved accuracy rates ranging from 95% to 98%. A tagged corpus has more useful information than untagged corpus; so, tagged corpus can be used to extract grammatical and linguistic information from the corpus. Then, it can be used for many applications such as creating dictionaries and grammars of a language using real language data. Tagged corpora are also useful for detailed quantitative analysis of text.

In this project, we have described a system for recognizing names, verbs, particles, and proper names in Arabic language text through a combination of high precision morphological analysis and a subsequent component which recognizes the named entities. Although highly deterministic and not taking account of context, the morphological analysis component removes a great deal of morpho-lexical ambiguity; yet, has the side-effect (Still, it has demonstrated ???) of demonstrating that the true difficulties in Arabic morphological ambiguity might be limited to specific contexts. We have shown that morphological information is crucially important to effective Arabic name recognition.

We have tested our system by using some vowelized and non-vowelized text documents, and achieved an accuracy rate of about 93%. We have stated the factors of errors and how this accuracy rate can be enhanced.

1.1 INTRODUCTION

A tag is a code which represents some features or set of features and is attached to the segment in a text. Single or complex information are carried by a tag.

A tag has many requirements and characteristics; representation of only one part-of-speech per unit, distribution of the tag, e.g. each element in a tag should mean something by itself, and mnomonicity; the ability to easily recognize the meaning of a tag. [12]

The tag-sets may represent morphological, syntactic and/or phrase structure information. The tag-set varies from one system to another according to the size, content, complexity and level of annotation, words' categories and phrases. [12]

The effectiveness of POS taggers is measured in terms of accuracy, i.e. correctly assigning tags in a text. The perfect system could be characterized as 100% recall and 100% precision. The accuracy rate for the majority of systems is between 95% and 99%, according to Krenn & Samuelsson (1997). Accuracy depends on the following conditions: obtain

- The size of the tag-set, a small tag-set achieves good results when trained on a small corpus;
- the size of the training corpus, e.g. the bigger the tag-set, the less correct the result will be if the size of the training corpus doesn't increase accordingly;
- the type of training and test corpora, e.g. if a tagger is trained on a specific training corpus and is tested on another type of text, the result will be less satisfactory than with a similar test corpus, and lastly;
- the type and the completeness of the vocabulary, e.g. an incomplete lexicon also results in less accuracy.[12]

1.2 BENEFITS AND USES OF TAGGING SYSTEMS

There are many applications where tagger is necessary. For instance, in parsing systems, parsers have to know the part-of-speech for each word in the parsed text. Full automatic tagging system will increase the efficiency and performance of the parser. Taggers are also important for the automatic building or extracting of noun phrase. For automatic building, tagger can be used for creating dictionaries and grammar of a language using real data. It also can be used in text-to-speech systems. In addition, it can be utilized in building inverted indexed files, which uses nouns as index terms rather than other words; the use of tagger is useful for knowing the part-of-speech of a word in order to use it as an index term. Query expansion, building thesaurus, text summarization, and automatic compounds extraction also benefit form tagging systems.

1.3 STATISTICAL (PROBABILISTIC) VS. RULE-BASED TAGGING SYSTEMS

The major problem of automatic POS tagging is that the word may belong to several parts of speech, which leads to ambiguity between two or more classes. in this case, contextual information is needed to distinguish between classes. It is an easy task to identify all the possible tags of a word, but it is difficult to acheive the disambiguation. Two approaches are designed to solve (or realize???) the disambiguation of the tag, statistical/probabilistic approaches and rule based methods. [12]

There are two main methodologies for automatic Part of Speech (POS) tagging; statistical (probabilistic) and rule-based. Almost all automatic POS tagging systems are based on Markov models where training consists of learning lexical probabilities and contextual probabilities. These taggers usually take a large manually annotated corpus from which they extract probabilities. Rule-based taggers on the other hand, work by rules that have been constructed by linguists.[12]

In terms of accuracy, systems based on the statistical approach tend to reach 95-97% correct analysis using tag-sets ranging from a few dozen to about 130 tags (Charniak 1996; Church 1988; DeRose 1988; Garside et al. 1987; Merialdo 1994, etc.), while rule-based systems, such as Constraint Grammar, can achieve slightly more than 99% correct results (Samuelsson & Voutilainen, 1997). In spite of the fact that statistical models are less accurate than rule-based models, most existing POS analyzers have been based on a probabilistic model because of the robustness and automatic training ability of such systems. The limitations of the rule-based taggers are non-automatic, costly and time-consuming. [12]

1.4 NON-ARABIC TAGGERS

Constituent Likelihood Automatic Word-Tagging System (CLAWS), is a POS tagger for English trained on manually analyzed text based on probabilistic approach. The CLAWS system has an accuracy rate of 96%-97%, depending on the type of text. VOLSUNGA algorithm, which is similar to CLAWS, has an accuracy rate of 96% (DeRose 1988). TAGGIT uses context-frame rules to select the appropriate tag for each word. It has achieved an accuracy rate of 77% (Greene and Rubin 1997). Brill's Tagger, developed by Eric Brill in 1992-1994, is a rule-based tagger for English language. This tagger has achieved an accuracy rate of %97.5. [12]

Other tagging systems such as, Hybrid tagger which is a combination of both statistical and rule-based methods has been reported to have an accuracy rate of 98% (Tapanainen and Voultialainen 1994). Artificial intelligence techniques have been used in developing taggers. Daelemans (1996) tagger has achieved an accuracy rate of 97% for English language. Neural networks have been used in developing part-of-speech taggers. The University of Memphis has developed a tagger that achieved an accuracy rate of 91.6%. Another neural network tagger developed for Portuguese achieved an accuracy of 96%. [12]

1.5 ARABIC LANGUAGE TAGGING SYSTEMS

Arabic is a Semitic language, and its basic feature is that most of its words are built up from, and can be analyzed down to roots. Words are built up from these roots by following fixed patterns that add prefixes, suffixes and infixes to the word. For example, the Arabic word **"مـدّرس"** is built up from the root **"درس"** * by following the pattern **"مفعّـل"**. Prefixes and suffixes can be added to the words to add number or gender to these words, Adding the Arabic suffix **"ون"** to the word **"مـدرس"** "teacher" gives the word **"**

"مدرســـون "teachers" which indicates the masculine plural.

Arabic contains three genders; masculine, feminine and neuter. It also contains three persons, one to describe the speaker (first person), one to describe the person being addressed (second person), and one to describe the person that is not present (third person). It contains three numbers; singular, plural, and the dual. Arabic language distinguishes between the three moods of the verb (indicative, subjunctive, and jussive). Arabic contains three case forms of the noun (nominative, accusative, and genitive). It also distinguishes between the three moods of the verb (indicative, subjunctive, and jussive), and the three case forms of the noun (nominative, accusative, and genitive). [6]

The Arabic alphabet consists of 28 consonants but 3 of these are also used as long vowels. These are the same as the **"ي"** "**ea**" in beak, the **"و"** "**oo**" in root, and the **"l"** "**a**" in tar. Arabic also contains three short vowels that are **" ِ " الكسره "** equivalent to the "**i**" in sit, the **" ُ " الـضمه "** "**oo**" in foot, and the **" ً " الفتحه "** "**a**" in bat.

When it comes to the Arabic language, there are problems and challenges that are not present in English or other European languages. [4]

- Newspaper articles are full of proper

* These are sometimes called measures or *bayan* **"البيان"**, and in Arabic are called *awzan*. **"أوزان"**

- Some words in Arabic text begin with one, two, three, or four extra letters that constitute articles or prepositions
- Arabic is highly inflectional and derivational, which makes morphological analysis a very complex task
- The writing direction is from right-to-left and some of the characters change their shapes based on their location in the word.
- There is also a lack of Arabic corpora, lexicons, and machine-readable dictionaries, which are essential to advance research in different areas.

1.6 HISTORICAL OVERVIEW

Abuleil, S. Alsamara, Kh. and Evens, M 2002, describe a learning system that can analyze Arabic noun to produce their paradigms with respect to both gender and number using a rule-base that uses suffix analysis as will as pattern[3]. *Abuleil, S. and Evens*, M 1998, describe a system for building an Arabic lexicon automatically by tagging Arabic newspaper text.[4]

Khoja, Sh. 1996 described some of initial findings in the development of an Arabic part-of-speech tagger [5]. *Shreen Khoja, Roger Garside and Garry Knowles* proposed a tag-set for the morphosyntatic tagging of Arabic described morphosyntactic tag-set that is derived from the ancient Arabic grammar.[6]

Andrew Freeman documents some of the hurdles that were encountered during a long semester project to implement Brill's POS tagger for Arabic [7]

Bassam Hammo, Hani Abu-Salem and Steven Lytinen describe the design and implementation of a question answering (QA) system called QARAB.[8]

John Maloney and Michael describe a fast, high-performance name recognizer for Arabic texts. It combines a pattern-matching engine and supporting data with a morphological analysis component.[9]

Mohamed Attia Mohamed Elaraby Ahmed 2000, in his thesis has implemented an industry-quality computational processor of the Arabic morphology – called Morpho3 – along with a host of dependent applications as well as complementary utilities. [11]

2.1 THE MODEL OF THE ARABIC WORD

An Arabic word is a word either having all its characters bare or with discredited Arabic alphabets. It also may belong to either the original Arabic words, or arabized words. The original Arabic words are divided in turn into two sub categories; Derivative Arabic words, which are the verbs and nouns that are built according to the Arabic derivation rules, and Fixed Arabic words, which are a set of words molded by Arabs, anciently, and do not obey the Arabic derivation rules. The arabized words are nouns borrowed from foreign languages (perhaps with some phonetic adjustments to suit the Arabic

pronunciation) and have become common among the native Arabic speakers.[11]

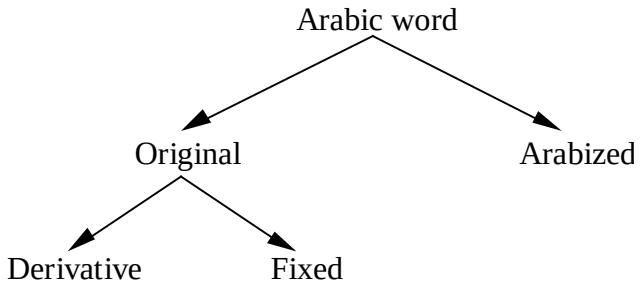

FIGURE [1]: THE CLASSIFICATION OF THE ARABIC WORDS

2.2 ARABIC IS A DIACRITIZED LANGUAGE

Arabic is also a diacritized language, that has the most elaborate diacritization system. The pronunciation of diacritized languages words cannot be fully determined by their spelling characters only; special marks are put above or below the spelling characters to determine the correct pronunciation. These marks are called *diacritics.* In such languages, two different words may have identical spelling whereas their pronunciations and meanings are totally different. Arabic is also a diacritized language that has the most elaborate diacritization system.[11]

Each character in an Arabic word must be assigned two things about diacritics:

1. The shadda state of the character. (With Shadda/Without Shadda.)
2. The diacritic of the character.

These are called the diacritic information of the character

Three diacritization states of an Arabic word:

- **Full diacritization**: It is the assignment of all the diacritic information for each character in the word including the last one.
- **Half diacritization**: It is the same as full diacritization except for that it does not provide the diacritic mark of the last character if it depends on the syntactic analysis of the word.
- **Partial diacritization**: Any other diacritization state of the word that provides less diacritic information than half diacritization is called partial diacritization.

2.3 ARABIC WORD CATEGORIES

Arabic grammarians traditionally classify words into three main categories. These categories are also divided into subcategories, which collectively cover the whole of the Arabic language. These categories are:

2.3.1 NOUNS

A noun in Arabic is a name or a word that names a person, thing, or an idea. Nouns are also divided into the following types:

- An **agent noun** is a derived noun indicating the actor of the verb or its behavior. "**كاتب**" "**writer**".
- A **patient noun** is a derived noun indicating the person or thing that undergoes the action of the verb"**مقتول** " "**killed**".
- An **instrument noun** is a noun indicating the tool of an action
- An **adjective** is considered to be a type of noun in traditional Arabic grammar.
- An **adverb** is a noun that is not derived and that indicates the place or the time of the action.
- A **proper noun** is the name of a specific person, place, organization, thing, idea, event, date, time, or other entity

- Nouns are divided into two classes according to their origins
  - ***Primitive noun*** "**الأسماء الجامدة**", which are substantives having no Arabic root
  - ***Derivative nouns*** "**الأسماء المشتقة**" which can be ***substantive*** or ***adjectives*** "**الصفات**".

- ***The subcategories of the noun are*** [6]
  - **1. Common 2. Proper 3. Pronoun**
  - **4. Numeral 5. Adjective**

- ***The subcategories of the pronoun are*** [6]
  - **1 Personal 2 Relative 3. Demonstrative**

- ***The subcategories of the relative pronoun are*** [6]
  - **1. Specific 2. Common**

- ***The subcategories of the numeral are*** [6]
  - **1.Cardinal 2. Ordinal 3. Numerical**

Adjective

- ***The linguistic attributes of nouns*** [6]

Gender: Masculine Feminine Neuter
Number: Singular Plural Dual
Person: First Second Third
Case: Nominative Accusative Genitive
Definiteness: Definite Indefinite

2.3.2 VERBS

Verbs indicate an action, although the tenses and aspects are different. Verb categories are divided into subcategories such as Perfect, Imperfect, and Imperative.

Verbs are categorized into three main parts:[6]
1. Perfect 2. Imperfect 3. Imperative

The definition of perfect verbs includes:

a. Equivalent of English past tense verbs (i.e. to describe acts completed in some past time).
b. Describes acts which at the moment of speaking have already been completed and remain in a state of completion.
c. Describes a past act that often took place or is still taking place (i.e. commentators are agreed (commentators agree or have agreed??) (have agreed and still agree)).
d. Describes an act which is just completed at the moment by the very act of speaking it (*I sell thee this*).
e. Describes acts which is certain to occur, that is it can be described as having already taken place (mostly used in promises, treaties and so on) *(Wright, 1974)*.

The imperfect does not in itself express any idea of time; it merely indicates a begun, incomplete, or enduring existence either in present, past or future time. While the imperative verbs order or ask for something to be done in the future.

The verbal attributes that have been used in our tagset are:[6]

- **Gender:** Masculine Feminine Neuter
- **Number:** Singular Plural Dual
- **Person:** First Second Third
- **Mood:** Indicative Subjunctive Jussive

2.3.3 PARTICLES

The Particle class includes: Prepositions, Adverbs, Conjunctions, Interrogative Particles, Exceptions and Interjections.

The subcategories of particle are:[6]
1.1. Prepositions 1.2. Adverbial 1.3. Conjunctions
1.4. Interjections 1.5. Exceptions 1.6 Negatives
1.7. Answers 1.8. Explanations
1.9. Subordinates

3. OUR APPROACH

Our approach is aimed to find a part-of-speech tag for all words in an Arabic language document in a most efficient and highly accurate way. Moreover, it is purposed to assigning POS tags to words in text documents, which specifies the three main word categories such as verb, noun and particle.

The approach consists of three parts: Inputs and Outputs of the system, Processors and Processes, and the Project Database. Below is a detailed description for each part of the system.

3.1. INPUTS AND OUTPUTS.

The input of the tagging system is designed to be an Arabic text document either vowelized or non-vowelized. We tested our system using vowelized documents from the holly Qur'an, and another set of non-vowelized 242 Arabic abstracts chosen randomly from the proceedings of the Saudi Arabian National Computer Conference. Outputs of the system are a fully tagged documents, where each tagged word consists of the word followed by a tag assigned to that word, the tag of the word consists of eight digits, each digit decodes a part-of-speech or a linguistic

attribute for the word, table [3.1] below describes each digit in the tag and how it is encoded**.**

TABLE [1]: DESCRIPTION OF TAG FIELDS

Digit Position	Description	Example
1	Word Category	Verb
2,3	Word Sub-Category	Proper Noun
4	Gender	Feminine
5	Number	Singular
6	Person	First
7	Case	Nominative
8	Definiteness	Definitive
9	Mood	Indicative

3.2 PROCESSORS AND PROCESSES.

First, tokenizing process, stemming process, affix extractor process, and pattern recognizer, which are applied in the first part of the algorithm, and we called them the utility functions of the system.

Second, the lexical word matcher, which is responsible for assigning tags to the words in the document that match words stored previously in the lexicon database.

Third, the noun tagger process, which is responsible for applying linguistic rules for nouns to the document's words, if one of the rules matches for any word in the document, then its tag will be assigned and stored.

Fourth, the verb tagger process, which is responsible for applying linguistic rules for verbs to the document's words, if one of these rules matches for any of the untagged words, its tag will be assigned and stored.

Other secondary processes are responsible for result summarization and report generation that display linguistic information about the tagged documents in a well formatted way.

3.2.1 UTILITY FUNCTIONS

a) Tokenizing Process: This process locates a document and isolates words (tokens), each token is stored in the words table in the tagging database.

b) The Stemming Process: this process is responsible for extracting the root of all words in the document. It is an essential part of our program, since many processes following it use its output.

c) Affix Extractor Process: this process extracts affixes of the word. Affixes are of three types, prefixes; the extra letters added to the begging of the word, infixes; the extra letters added to the middle of the word, consist of two parts, the first part between the first and second origin letters which the root of the word consists of, and the other part is between the second and the third part, and suffixes; the extra letters at the end of the word. By extracting the root of the word, we are specifying the original letters of the word, so all other letters forming the word are extra words.

d) The Pattern Recognizer Process: This process extracts the patterns of the Arabic word documents. The Arabic language has many patterns that govern the formation of the word; some of them are known and their attributes are specified (**اوزان قياســية**) *patterns,* and the other are depending on the sound of the word, *word pronunciation* (**اوزان سماعية**).

The pattern recognizer identifies relative pronouns attached to the end of the verbs and the definiteness letters, progress verb letters, order verb letters, conjunctions such as "**و**", "**ف**",.... etc. and prepositions where all are attached to the beginning of the word, those are not part of the word and should be avoided when pattern is recognized.

The Arabic word consists of two groups of letters, *original letters* (**حـروف أصـلية**) and the *extra letters* (**حـروف زائـدة**), original letters form the root of the word which is extracted by the stemming processor. Each letter of the root is changed to its equivalent in the pattern; the first letter is changed to "**ف**", the second to "**ع**", and the third to "**ل**", and the fourth to "**ل**" if the root is constructed of 4 letters.

The pattern is constructed by combining the letters **"ل","ع","ف"** with the affixes of the word according to their order in the word**.**

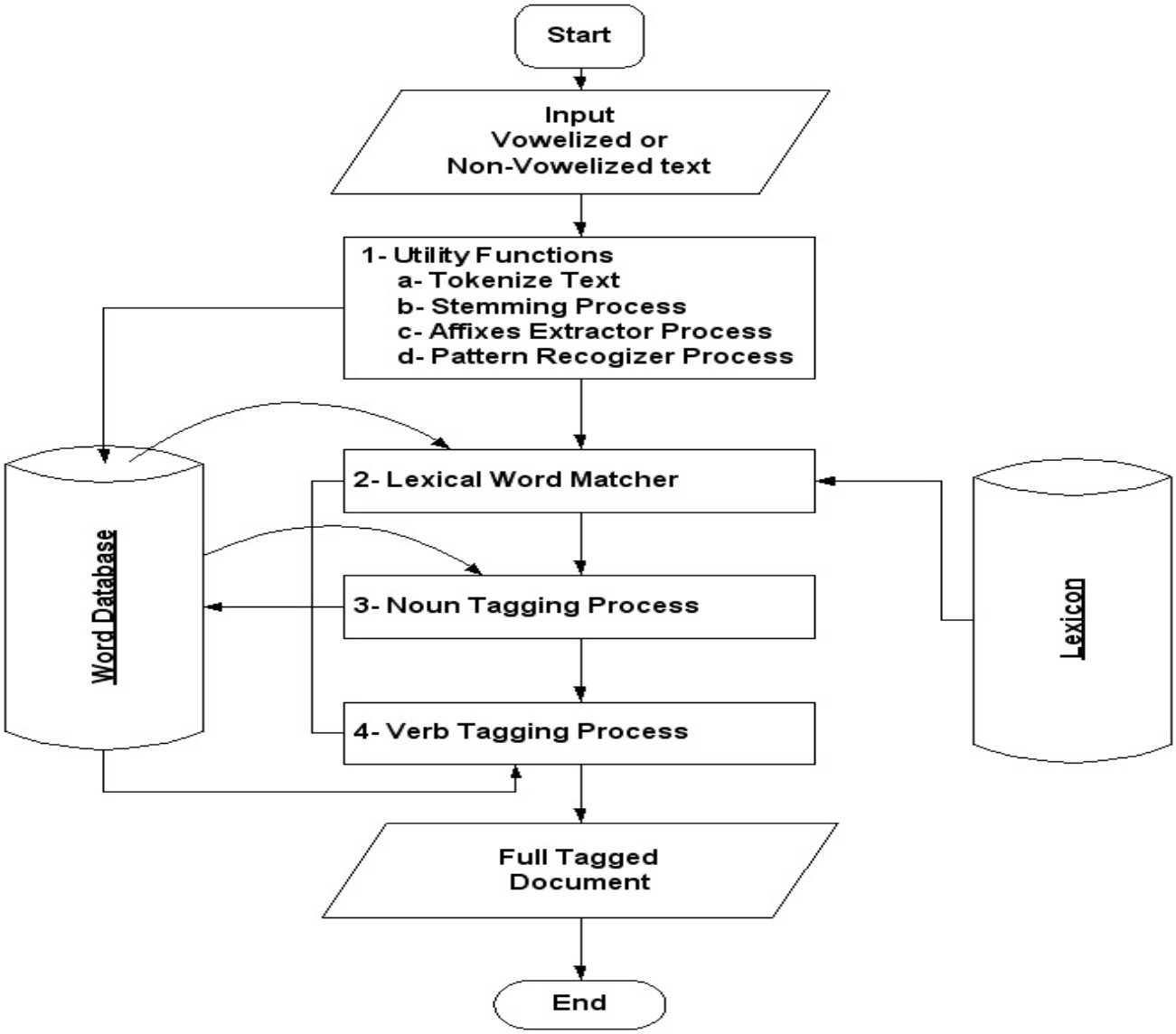

FIGURE [2]: THE FULL AUTOMATIC ARABIC TEXT TAGGING SYSTEM

3.2.2. THE LEXICAL WORDS' MATCHER

The lexical word matcher is constructed of the LEXICON table in the project database which contains many Arabic stop words, compiled by Bonnie Glover Stalls and Yaser Al-Onaizan. [21]

3.2.3 THE NOUN TAGGING PROCESS

We have constructed linguistically many rules that identify the words to be nouns from the behavior of the Arabic language.

Arabic language classifies nouns into two classes:

a. **Declinable Nouns (الاسماء المعربة)**: Nouns where diacritic mark is changed on the last letter of the word according to its position in the sentence.

b. **Indeclinable Nouns (الاســماء المبنــيه):** Nouns where diacritic mark is not changed on the last letter of the word according to its position in the sentence.

Rules for extracting nouns from the documents are constructed according to some special behaviors of the Arabic language, these behaviors might be found in the word's affixes:

- Word's prefixes such as; **"ال", "لـل",** etc.
- Word's suffixes such as; **"ة", "ء", "ى ",** etc.
- Diacritic Marks attached to the first and last letters of the word.

The word position in the sentence is a good indicator to identify nouns, and some words always followed by nouns construct a linguistic rule to identify them in the text such as **" كـان وأخـواتها", "إن واخـواتها"**, and some of these words mainly used when recognizing proper nouns such as, **"الـسيد"** which means *"Mr"*, **"المملكـة"** which means "*kingdom*"**,**...etc.

Patterns of the words; Arabic language has many patterns; some of them indicate nouns only, some indicate verbs only, and the others indicate both nouns and verbs. We recognize some of which used as noun patterns such as **"فاعـل", "مفعول",...**etc

3.2.4 TAGGING PROPER NOUNS

A proper noun is the name of specific person, place, organization, thing, idea, event, date, time, or other entity. They may be solid or derived nouns[3]. It is a big problem to recognize proper nouns in an Arabic text, since Arabic language does not distinguish between lower and upper case letters; this makes it not nearly as easy to locate them in Arabic text as in English text. We use a technique used by Abuleil and Evens 1998, for tagging proper nouns in Arabic text, which depends on the keywords. They have classified these keywords as shown in table[2]. [4]

TABLE[2] : PROPER NOUNS KEYWORDS CLASSIFICATION.

#	Classification	Example	Example
1	Personal names (title)	**Mr**. John Adams	**السيد** جون آدم
2	Personal names (job title)	**President** John	**الرئيس** جون
3	Organization names	Yarmouk **University**	**جامعة** اليرموك
4	Locations (political names)	**Kingdom** of Jordan	**المملكة** الاردنيه
5	Locations (natural names)	Jordan **River**	**نهر** الاردن
6	Times	**Month of** April	**شهر** نيسان
7	Product	IBM **Computers**	**حاسبات** اي بي ام
8	Events	**Exhibition** of Egyptian Books	**معرض** الكتاب المصري
9	Category	Arabic **Language**	**اللغة** العربية

They also design a set of grammatical rules to the proper noun phrases in the text. These rules are listed below and figure [3] shows an example applied to these set of rules. [4]

PNP ⟶ ال K/W-TITLE A
A ⟶ A1 | A2
A1 ⟶ ADFPN
| ADFPN PN-PERSON
| ADFPN A2
A2 ⟶ ADJ
| ADJ PN-PERSON
ADFPN ⟶ ال [ADJ. DERIVED FROM P.N]

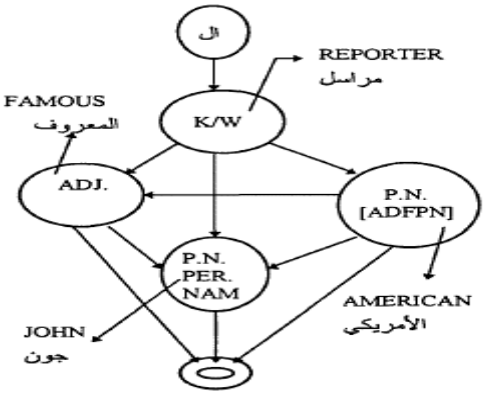

FIGURE [3]: TAGGING PROPER NOUNS EXAMPLE
المراسل الامريكي المعروف جون

3.2.4 VERB TAGGING PROCESS

This process is responsible for identifying verb words in the document. Verb is defined as a word that indicates a meaning by itself that is united with a tense or time. Verbs accepts words or letters as indicators such as the particles **"قـد", "سـوف",** or pronouns or the letters **"س", "ت", "ن".** [1]

Arabic language behaviors which indicate whether the noun is a noun or verb are:

1. PATTERN

Verbs of the Arabic language have roots consisting of 3 or 4 letters, out of these root, many verbs can be generated according to some fixed rules that add letters as **"ا", "أ", "ت", "س", "ي", "و", "هـ", "ن", "م", "ل"** and are combined with the word **"سـألتمونيها"** , these fixed rules which are called patterns, contain 15 different essential patterns

2. AFFIX OF THE WORD

Some affixes are used with verbs and some with nouns while others with both verbs and nouns. We extract 31 groups of affixes that are used with the essential pattern listed in table [3.9], these affixes affect verb semantics, such as verb category (perfect, imperfect, imperative), gender (masculine, feminine, neuter), number (singular, dual, plural), person (first, second, third), and mood (indicative, subjunctive, jussive).

The number property of words that have patterns with no suffixes can not be specified directly. To identify that, we have to refer to the next word; which is the subject of the sentence, where verb and its subject are identical in number property. For example, **"كـتب الطالـب الـدرس"** which means "*the student wrote the lesson*", the verb **"كـتب"** "*wrote*" in this sentence is singular verb where it is dual in the sentence **" كـتب الطالـبان الدرس"**, which means "*the two students wrote the lesson*", and plural in the sentence **" كـتب الطـلاب الـدرس"** which means "*the students wrote the lesson*". By referring to the subject we can decide the number of the verb.

3. RULE-BASED

Rules are extracted from the syntax of the Arabic sentence formation; tags are assigned to verbs according to their position in the Arabic sentence, where some types of pronouns, prepositions and letters are grouped to the verb word.

4.1 RESULTS

We tested our system using vowelized documents from the holly Qur'an and another set of 242 non-vowelized Arabic abstracts chosen randomly from the proceedings of the Saudi Arabian National Computer Conference. We run our system on a group of these documents selected randomly, we obtain results as shown in table[4.1]

The accuracy of our system has been calculated for the main parts of the system, pattern analyzer, noun and verb tagger. The number of correct words, the number of incorrect words along with their percentages have been calculated, we grouped these results and obtained the total accuracy of the system of about %93 and a fault result of 7%.

When we calculated the system's efficiency, errors come from stemming process are discarded and corrected manually since the research's problem is about tagging Arabic text. Other essential parts of the system are analyzed and efficiency of each part is calculated.

Faults in the system caused from some uncontrolled conditions; stemming algorithm used in our program is designed for extracting roots constructed of three letters, however some words'

roots are constructed of four letters, which we don't deal with in our tagging system. Another factor that affects the efficiency of the system is the incorrect roots extracted when some of the word's letters being doubled, the doubled letters are marked with shadda " ّ ", which is not a diacritic but is a mark of doubling the character while pronouncing it.

Some errors of the system come from lexical words not included in the lexicon, this type of errors can be corrected easily by adding these words to the lexicon, incorrect words may include Arabaized words which are translated as pronounced from other international languages. These words do not have a root and a pattern.

These types of errors are accumulated since errors coming from extracting faulty root form their words, and this affects the process of generating the correct pattern of the word, which in turn affects tagging these words with the correct POS tag especially for words which mainly use pattern rule in the tagging process.

TABLE [3]: SYSTEM ACCURACY TABLE

System model	Doc #	# Words	# Correct	# Incorrect	% Correct	% Incorrect
Pattern analyzer	1	309	297	12	96.12	3.88
	2	183	177	7	96.20	3.80
	3	159	153	6	96.22	3.79
	4	176	165	11	93.75	6.25
	Total	**827**	**792**	**36**	**95.76**	**4.35**
Noun Tagger	1	162	151	11	93.21	6.79
	2	153	142	11	92.81	7.19
	3	119	112	7	95.73	5.87
	4	149	139	10	93.29	6.71
	Total	583	544	39	93.31	6.68
Verb Tagger	1	69	63	6	91.30	8.70
	2	30	27	3	90.00	10.00
	3	37	34	3	91.89	8.11
	4	26	25	1	96.15	3.85
	Total	162	149	13	91.98	8.02
Totals	**AllDoc**	638	590	48	92.48	7.52

4.2 CONCLUSION

In this study, we have designed a full automatic tagging system for Arabic language text, and achieved an accuracy rate of about %93. We have stated many grammatical rules for tagging verbs and nouns and assigned some useful linguistic attributes that may be helpful for Arabic language applications in the fields of information retrieval and natural language processing.

We have designed some utility processes; pattern extractor and stemming process that are essential parts in our system, pattern extractor process uses roots of words extracting by stemming process to generate patterns of the words, which are mainly used in tagging verbs and used in some rules of tagging nouns. Lexicon is also used to tag words like particles and some nouns and verbs that have a static tag in the Arabic language.

The error percentage of about %7 percent comes from the faulty roots, patterns and the word missing in lexicon. The system can achieve more accuracy by enhancing the root algorithm, which also enhances the pattern extractor process. Adding lexical words to the lexicon which are not included enhances both the lexicon and the accuracy of the system.

More future work can be done in the field of part-of-speech tagging for Arabic language text by applying other tagging techniques such as probabilistic, hybrid and neural networks tagging techniques.

4.3 FUTURE WORK AND SYSTEM'S ENHANCEMENT

We use a rule-based technique in our system and we achieved accuracy rate of about 93%. To enhance system's accuracy we may apply a hybrid technique by introducing some statistical processing beside linguistic rules used in our system.

Enhancement of root algorithm may increase the system accuracy; root algorithm used in our system has generated some errors that affect the accuracy of the system badly.

Some other techniques may be used in tagging system; such as neural network. It is a good research to build a tagging system based on neural networks technique, and then comparing its results with the other tagging systems based on

other techniques like rule based, probabilistic and hybrid.

5. REFERENCES

[1] **انطوان الدحداح** ، " **معجم قواعد اللغة العربيه – في جداول ولوحات** " ، مكتبة لبنان ، الطبعة الثالثة ، ١٩٨٧

[2] **الشيخ مصطفى غلاييني ،" جامع الدروس العربية"**، المكتبة العصريه، صيدا- لبنان، الطبعة الثامنه عشرة ، ١٩٨٦

[3] **Abuleil S, Alsamara Kh, Evens, M**. 2002 " Acquisition system for Arabic Noun Morphology", Proceedings of the Computational Approaches to Semitic Languages Workshop, University of Pennsylvania, 11th July 2002.

[4] **Abuleil, S. and Evens, M**., 1998. "*Discovering Lexical Information by Tagging Arabic Newspaper Text*", Workshop on Semitic Language Processing. COLING-ACL.98, University of Montreal, Montreal, PQ, Canada, Aug 16 1998, pp 1-7.

[5] **Shreen Khoja**, 2001,"*APT: Arabic Part-of-speech Tagger*". Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), Carnegie Mellon University, Pittsburgh, Pennsylvania. June 2001.

[6] **Shreen Khoja**, **Porger Garside, and Gerry Knowles**. 2001 "*A tagset for the morphosynactic tagging of Arabic".* Paper presented at Corpus Linguistics 2001, Lancaster University, Lancaster, UK, March 2001.

[7] **Andrew Freeman**, "*Brill's POS tagger and a Morphology parser for Arabic*". NAACL 2001 Student Research Workshop, Lancaster University.

[8] **Bassam Hammo, Hani Abu-Salem, and Steven Lytinen**. "*QARAB: A Question Answering System to Support the Arabic Language*" , Proceedings of the Computational Approaches to Semitic Languages Workshop, University of Pennsylvania, 11th July 2002.

[9] **John Maloney and Michael Niv**," *TAGARAB: A Fast, Accurate Arabic Name Recognizer Using High-Precision Morphological Analysis*".

[10] **Shreen Khoja**, "Shreen Khoja web site", http://www.comp.lancs.ac.uk/computing/users/khoja/index.htm , last visited march 2003.

[11] **Mohamed Attia Mohamed Elaraby Ahmed 2000,** " *A Large-Scale Computational Processor of the Arabic Morphology, and Applications*" .

[12] **Beáta Megyesi,** D-level thesis (Master's thesis) in Computational Linguistics, spring 1998. "Brill's Rule-Based Part of Speech Tagger for Hungarian". Computational Linguistics, Department of Linguistics, Stockholm University, Sweden.

[13] **Thalouth B, Al-danan A** 1986 "*A comprehensive Arabic morphological analyser-generator"*. In Proceedings of Arab summer school, Syria.

[14] **B.B. Greene and G.M. Rubin** (1971) "*Automatic Grammatical Tagging of English".* Department of Linguistics, Brown University, Providence, R.I.

[15] **Eric Brill (1992)** ***A*** *"Simple Rule-Based Part of Speech Tagger".* In "Proceedings of the Third Conference on Applied Natural Language Processing", Trento, Italy, pp. 152-155.

[16] **Eric Brill** (1994) *"Some Advances in Transformation Based Part of Speech Tagging*". In "Proceedings of the Twelfth International Conference on Artificial Intelligence" (AAAI-94), Seattle, WA.

[17] **Abuleil, S. and Evens, M**., 2002. Extracting an Arabic Lexicon from Arabic Newspaper Text. *Computers and the Humanities*, 36(2), pp. 191-221.

[18] **Al-Shalabi, R. and Evens, M**., 1998. 9A Computational Morphology System for Arabiclo.Workshop on Semitic Language Processing.COLING-ACL™98, University of Montreal,Montreal, PQ, Canada, Aug 16 1998. pp. 66-72.

[19] **Beesley, K. and Karttunen, L**., 2000. Finite-state Non-Concatenative Morphotacticsli.Proceedings of the 38 th Annual Meeting of the Association for Computational Linguistics.Hong Kong, Oct 1-8, 2000. pp.191-198.

[20] **Nuno Marques and Jos Gabriel Lopes** (1996) Using Neural Nets for Portuguese Part-of-Speech Tagging.In "Proceedings of the Fifth International Conference on The Cognitive Science of Natural Language Processing", Dublin City University.

[21] **Bonnie Glover Salls and Yaser Al-Onaizan**, "Arabic *Stop Word List*", NLP research activities at University of Southern California's , Information Science Institute.