A Rule-Based Approach for Tagging Non-Vocalized Arabic Words

Ahmad Al-Taani and Salah Abu Al-Rub
Department of Computer Sciences, Yarmouk University, Jordan

Abstract: *In this work, we present a tagging system which classifies the words in a non-vocalized Arabic text to their tags. The proposed tagging system passes through three levels of analysis. The first level is a lexical analyzer that composed of a lexicon containing all fixed words and particles such as prepositions and pronouns. The second level is a morphological analyzer which relies on word structure using patterns and affixes to determine word class. The third level is a syntax analyzer or a grammatical tagging which relies on the process of assigning grammatical tags to words based on their context or the position of the word in the sentence. The syntax analyzer level consists of two stages: the first stage depends on specific keywords that inform the tag of the successive word, the second stage is the reversed parsing technique which scans the available grammars of Arabic language to get the class of a single ambiguity word in the sentence. We have tested the proposed system on a corpus consists of 2355 words. Experimental results showed that the proposed system achieved a rate of success approaching 94% of the total number of words in the sample used in the study.*

Keywords: *Part-of-speech tagging, lexical analyzer, morphological analyzer, Arabic language processing.*

Received July 3, 2008; accepted September 3, 2008

1. Introduction

Arabic language ranks sixth in the world's league of languages with an estimated 200 million native speakers and is widely used throughout the Muslim world [8]. Some archeological evidence shows that Arabic may be the oldest language [27].

Arabic texts could be either a vocalized text such as the language of the holy Quran or a non-vocalized text which is used in newspapers, books, and media. Handling the non-vocalized texts is confusing since the non-vocalized word may have more than one meaning. For instance, the non-vocalized Arabic word (كتب) has three possible interpretations: "kataba" (he wrote), "kutiba" (has been written), and "kutubun" (books).

To understand what a word class is, we must understand the idea of putting similar things together into groups or categories. We usually use three categories to classify all the words used in Arabic: Nouns, verbs, and particle [17]. This classification is not perfect in non-vocalized Arabic text. Sometimes, it is hard to tell which category a word belongs to. Moreover, the same word may belong to different categories depending on how it is used.

Word classification is the process of assigning tags to words and it is often only one step in a text processing application then the tagged text could be used for deeper analysis. A tagged corpus is more useful than an untagged corpus because there is more information than in the raw text alone. Once a corpus is tagged, it can be used to extract information from the corpus. This can then be used for creating dictionaries and grammars of a language using real language data. Tagged corpora are also useful for detailed quantitative analysis of text and it is preparation for higher level natural language understanding tasks such as parsing, semantic, and translation [20]. The parser must know the tag for each word. Most previous approaches used manual tagging but having an automatic tagging will increase the efficiency and performance of the parser. Information about the category of the word is very helpful in understanding the full meaning of the word and knowing how to use it. For example, a machine translation system processes the input text in stages: de-formatting, morphological analysis, word classification tagging disambiguation, shallow structural transfer, lexical transfer, morphological generation, and re-formatting. Word classification is a basic stage that is needed in any machine translation system. Having an automatic tagging system will increase the efficiency and performance of the translation system.

2. Problem Statement

Some problems of using affixes in word classification are encountered by some researchers. Some letters that appear to be affixes are in fact part of the word such as in the word (التقى ELTAGA -meet). We treat the first two letters (ال) as an article and the word is classified as a noun, but in fact some of these letters are part of the word and the word is a verb [19]. Some letters (for instance the long vowels) may change to other letters when an affix is added and so the letters should be changed back when that affix is removed.

Some words in non-vocalized text may have more than one tag (ambiguous words and unknown words) in which the classification may depend on the word meaning. The ambiguity of Arabic lies on different levels. We give a few examples of possible combinations of grammatical categories of non-vocalized words.

Many verbs have the same shape as nouns (especially in roots that have no affixes). For example, the word (ذهب) may mean go and is classified as a past verb or may mean gold and is classified as a noun [7].

Another word pattern which covers both nouns and adjectives is the pattern of both active and passive participles (فاعل، مفعول) and derivatives. These cases are sometimes even more complicated because they can also be classified from time to time as a preposition as in the word (داخل within) and as a participle with the function of a verb such as in the sentence (هو داخل) "he is going inside" [18].

Many verbs have the same shape as adjectives. Often a non-vocalized verb with three radicals has the same pattern as an adjective. For example, the three radicals "فرح" can both stand for the verb "فَرِحَ" and the adjective " فَرِحْ" [18].

The most important mingling of word patterns between verbs and nouns occurs in the verbal nouns "masdar". The verbal nouns of the fifth and the sixth form often raise confusion. For example, the word "تدخّل" (fifth form) can be both a verb (to meddle) and a noun (interference). Also the word "تعاون" (sixth form) can be both a verb (to help) and a noun (cooperation) [18].

The pattern (أفعل) is even more complicated. This pattern offers at least three possibilities: a noun, an adjective or a verb. The word (أبيض) for instance mean both white as a white (a member of the white race) and it can also have the function of a verb in the sentence (ما أبيض وجهه) which means "what is his face white!" [18].

The proposed approach deals with nouns, verbs and particles since they are the main three parts of the Arabic speech and no Arabic words classified outside of these parts and all other classes are branches of these parts. In this study, we have concentrated on ways to completely distinguish between these main three parts which will enable the system to tag more classes with higher rates in future works.

3. Previous Work

Many methods have been proposed for word-class tagging. Most works used the affixes of the words and their patterns for this purpose. Abuleil *et al.* [1, 2, 3, 4] proposed four approaches for Arabic language processing. In 1998 [1], they built an automatic Arabic lexicon for tagging Arabic newspaper texts. In 2002 [2], they proposed a rule-based system that uses suffix analysis and pattern analysis to analyze Arabic nouns to produce their morphological information with respect to both gender and number. In 2004 [3], he built a database and graphs to represent the words that might form names and the relationships between them. In 2006 [4], they described a learning system that analyzes Arabic nouns to produce their morphological information with respect to both gender and number based on suffix analysis and pattern analysis.

Many other rule-based techniques are proposed. Diab *et al.* [10] designed an automatic tagging system to tokenize part-of-speech tag in Arabic text. Habash *et al.* [11] proposed a morphological analyzer for tokenizing and morphologically tagging Arabic words. Khoja [14] developed a tagging system by combing statistical techniques with rule-based techniques. The tag set used is extracted from the BNC English tag set but modified with some concepts from traditional Arabic grammar. The tag set contains 131 tags assigned to words. A corpus of 50,000 words from the Saudi newspaper Al-Jazira was used to train the tagging system.

Freeman [13] described an Arabic part-of-speech tagging system based on the Brill tagging system which is a machine learning system that can be trained with a previously-tagged corpus. Freeman used a tag set consists of 146 tags extracted from Brown corpus for English. Also, Lee *et al.* [23] used a corpus of manually segmented words which appears to be a subset of the first release of the ATB (110,000 words). They obtained a list of prefixes and suffixes from this corpus which is apparently augmented by a manually derived list of other affixes. Maamouri *et al.* [15] presented a part-of-speech tagging system for Arabic. The authors based their work on the output of Tim Buckwalter's morphological analyzer. This tagging system is tested on a corpus consisted of 734 files extracted from the "Agence France Press" which was developed by Maeda and Hubert Jin.

Many researchers search for new methods to resolves the ambiguity in Arabic text. Marsi *et al.* [16] explored the application of memory based learning to morphological analysis and part-of-speech tagging of written Arabic based on data from the Arabic Treebank. Al Shamsi *et al.* [5] resolved Arabic text part-of-speech tagging ambiguity through the use of a statistical language model developed from Arabic corpus as a Hidden Markov Model (HMM).

Most of the Arabic researches are processed and analyzed non-vocalized text. But many other researchers processed and analyzed vocalized text and construct rules on short vowels (Fatha, Damma, Kasra, Sukun, Tanween-Fateh, Tanween-Damm, Tanween-Kasir) to classify the word and identify the group that it belongs to. Alqrainy *et al.* [7] presented a rule-based part-of-speech tagging system which automatically tags a partially vocalized Arabic text. The aim was to remove ambiguity and to enable accurate fast automated tagging system. A tag set has been designed

in support of this system. Tag set design is at an early stage of research related to automatic morph-syntactic annotation in Arabic language. Talmon *et al.* [22] presented a computational system for morphological tagging of the holy Quran for research and teaching purposes. The core of the system is a set of finite-state based rules which describe the morphological and morph-syntactic phenomena of the Quran language. The system is currently being used for teaching and research purposes. Safadi *et al.* [21] presented a method to supply vocalized Arabic text by using unsupervised machine methods.

Recent work by Chiraz *et al.* [9] addressed the problem of part-of-speech tagging of Arabic texts with vowel marks. The system consists of five agents work in parallel in order to determine a suitable tag for each word in a sentence.

4. Methods

In this study, we present a tagging system which classifies the words in a non-vocalized Arabic text to their tags. The system processes a non-vocalized text which is a text without short vowels that are normally omitted from Arabic text such as newspapers, books and media. Figure 1 shows the architecture of the proposed system. It consists of three levels: the first level is the lexicon analyzer which contains all Arabic particles including prepositions, adverbs, conjunctions, interrogative Particles, exceptions, and interjections. The second level is a morphological analyzer which uses morphological information such as the patterns of the word and its affixes to presume the class of the words. The last level is a syntax analyzer which consists of two stages; the first stage depends on specific keywords that inform us the tag of the successive word, and the second stage is the reversed parsing technique.

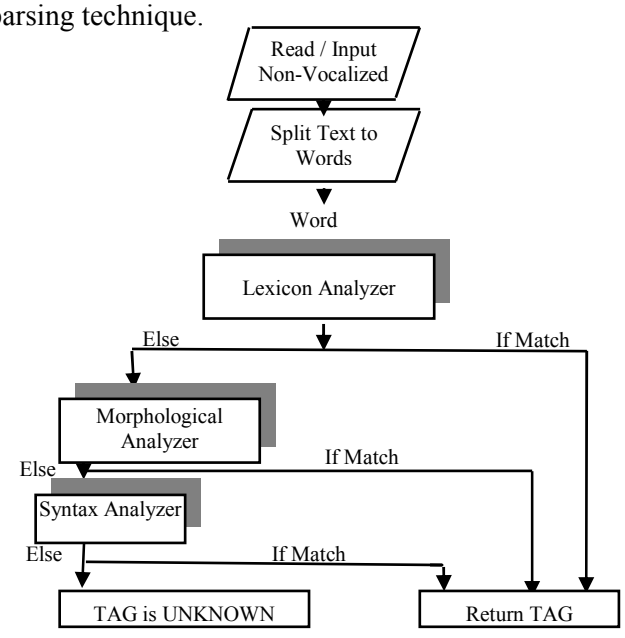

Figure 1. The Architecture of the tagging system.

The proposed system reads a non-vocalized Arabic text and divides it into separate words, then we take each word and enter it into the first level (Lexicon Analyzer), if it exists we return the corresponding TAG, if not; we transfer the word to the second level (Morphological Analyzer). After processing the word; if it matches we return the presumed TAG, if not; we transfer it to the final level (Syntax Analyzer). After testing words positions, if the TAG is found then the TAG is returned, otherwise there is no presumption about the corresponding TAG or the TAG is UNKNOWN.

4.1. The Lexicon Analyzer

Lexicons are the heart of any natural language processing system. The initial tagging level is a lexicon analyzer. The system has a lexicon which stores all Arabic fixed words and particles (prepositions, adverbs, conjunctions, interrogative particles, exceptions, questions and interjections, see Appendix A). Each word in the reading non-vocalized text is explored in the lexicon, if it is found; then the corresponding TAG is returned. But if it is not found, we transfer the word to the second level of the system i.e. the morphological analyzer. The process of the lexicon analyzer can be summarized by the following algorithm:

```
Begin
    read text
    tokenization
    take word
    search for the word in the lexicon
    if found then
        return the corresponding tag
    else
        transfer word to the morphological
        analyzer
End
```

4.2. The Morphological Analyzer

A morphological system is the backbone of a natural language processing system. Building a morphological analyzer for Arabic has its own distinct motivations and challenges that add to those shared for all morphological analyzers. Arabic language is a highly inflectional and a highly derivational language. These are respectively attributable to the large number of possible affixes (especially prefixes and suffixes), Arabic possesses, and the large number of derivational forms (patterns) of a certain word that can have a unique root system. Since there are multitude and diversity of rules of Arabic morphology, many researches can apply more than one approach using these rules. For many classical Arabic morphology operations, linguists have different ways for working out with the same operation.

There are several signs in the Arabic language that indicate whether the word is a noun or a verb. One sign is the pattern of the word, some of the patterns are used with verbs and others are used with nouns. But when we deal with non-vocalized words many words are ambiguous since their patterns are used with both verbs and nouns. Part-of-speech of a word can also be found by using affixes. Some affixes are used with verbs only, some are used with nouns only and some are used with verbs and nouns. Many researchers listed and defined some prefixes and suffixes that identified the class of a given word for vocalized or non-vocalized.Since most of Arabic language words are trilateral, the morphological experts consider that the origins of Arabic words are three letters. For this reason, if we want to balance a word to know the origin of it (affixes: prefix, infix, and suffix) we have to face the original letters of the word by the letters of the word (فعل). If the length of the word is greater than three letters we face the original letters for the balanced word by the letters of the word (فعل) and the additional letters are faced by its pronunciation like (المنازل ...المفاعل)، (استشعر ...استفعل)، (انتظار ...افتعال) [24] as in the following example:

انتظارهم	ا	ن	ت	ظ	ا	ر	ه	م		
افتعالهم	ا	ف	ت	ع	ا	ل	ه	م		

Affixes are always a subset of the word "سألتمونيها" which come in the word in four positions [25]: before the 'fa' of the word (قبل فاء الكلمة) which called prefix, after the 'fa' of the word (بعد فاء الكلمة) which called infix, after the 'ayn' of the word (بعد عين الكلمة) which called infix, and after the 'lam' of the word (بعد لام الكلمة) which called suffix. Prefixes and suffixes can be used not only to extract information like tense and subject/object features, but also to help in identifying the type of the token (noun or verb). This is because some prefixes/suffixes are attached to a specific type of tokens as shown in Figure 2 and hence can be used in the disambiguation process. Sometimes one affix can determine the tag of a word: for example, if the prefix is ((ال - التعريف then the word is a noun [17].

The body of the word is its main part. It is called the stem of the word (ساق/جذع الكلمة). It is the inner part surrounded by some prefixes and suffixes. So the proposed system concentrates to find the stem of the word as the first step. Then, from the stem we presume the prefixes, suffixes and infixes which may map the word to the corresponding predefined TAG. The Morphological Analyzer uses a simple approach to divide the Arabic word into three parts:

Prefix: Consist of as many as five concatenated prefixes or could be null (Appendix B).

Stem: It is composed of root and pattern morphemeswhich derived from nouns and verbs patterns [25] (Appendix D).

Suffix: Consist of as many as three concatenated suffixes or could be null (Appendix C). For example, the word "syktbwnha سيكتبونها" would be analyzed as follows:

Prefix	Stem	Suffix
sy سي	ktb كتب	wnhA ونها

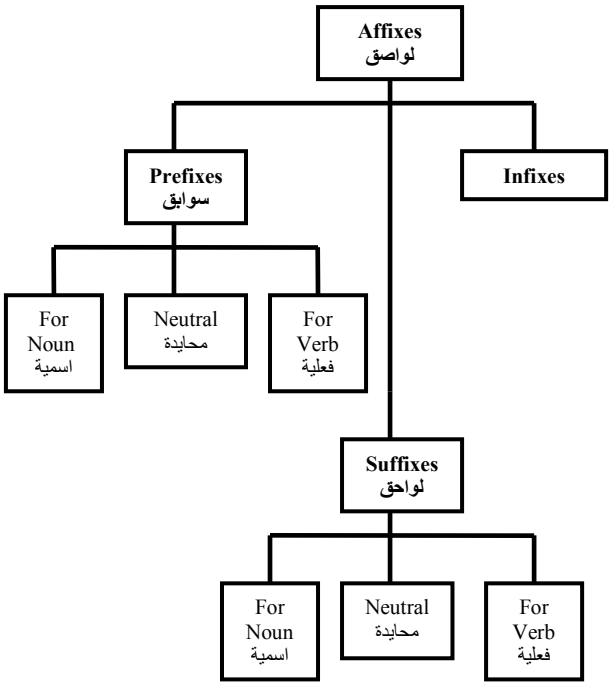

Figure 2. Specific class affixes.

We have extracted the longest common prefix from a given word by comparing the word with prefixes in Table 4 (see appendix B). Longest common suffix is extracted from a given word by comparing the word with the suffixes in Table 5 (see appendix C), then we compare the remaining letters with existing stems patterns (listed in appendix D) and ignoring the three letters corresponding word فعل as a root and then retrieve the Infixes. For example, for the word استعبادهم (Figure 3) we first extract the longest common prefix (است) after matched with prefixes in Table 4, the we extract the suffix (هم) after matched with suffixes in Table 5. Then we compare the remaining letters with existing stems patterns (see appendix D), it will be matched with the stem pattern (فعال), then we ignore the letters of word فعل which is a root and extract the remaining letters as infixes.

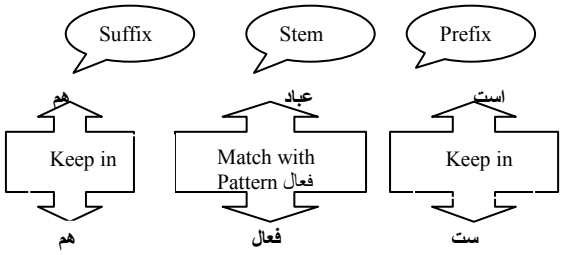

Figure 3. An Example.

We believed that the information provided by grammatical affixes would be useful but not sufficient to determine the exact word classification within the two major categories nouns and verbs. Certain prefixes, suffixes or infixes come with certain classes of words, so we try to collect our own rules of grammatical prefixes, suffixes and infix that Identifying verbs and nouns.

The proposed morphological analyzer constructs a list of distinct prefixes, a list of distinct suffixes, a list of distinct infixes, and a list of relations between prefixes, suffixes and infixes. All of these lists may enable the system to identify the class of a given word.

We made intensive morphological studies and concentrated on the affixes that construct unique patterns of nouns and verbs to recognize what affixes make. As a result we developed a list of rules that recognize the prefixes, suffixes, infixes, and the relations between them to identify the class of the word. These rules are:

Rule 1: The following prefixes (or part of prefix) map the word to NOUN class: AL, FL, LL, M, Preposition affixes (b, k). (ال التعريف، فل، لل، م، زوائد أحرف الجر ب، ك)

*Rule 2:*The following suffixes (or part of suffix) map the word to NOUN class: T, AT, AA (ة، ا ت، اء)

Rule 3: The following suffixes (or part of suffix) map the word to NOUN class with the condition of not existing of the imperfect tense letters (المضارعة أحرف)**:** WN, YN, AN, Y (ون، ين، ان، ي **)**

*Rule 4:*The following infixes map the word to NOUN class with the condition of satisfying the corresponding position within the stem pattern determined between parentheses:
A, Y, AW, AWY (after the ayn of the word) ا، ي، و، او، اوي (بعد عين الكلمة) WA (after the fa' of the word) وا (بعد فاء الكلمة)

*Rule 5:*The following prefixes (or part of prefix) map the word to VERB class: Y, N, A', Future S (ي، ن، أ، س الاستقبال)

Rule 6: The following prefixes (or part of prefix) map the word to VERB class with the condition of the Rules that map the word to the NOUN class did not satisfy: A (ا)

Rule 7: The following suffixes map the word to VERB class: Opening T (ت المفتوحة)

We can see from these rules that the system should extract the prefix, suffix, stem and the pattern of the stem from the word in order to use these rules. Once a rule is satisfied; the word class is identified. For example, in the word ALMADRASA (المدرسة); we have a part of prefix is AL (ال التعريف) and the other part is M (م) and both mapping the word to the noun class by rule 1. the suffix is T (ة) which map the word to the noun class by rule 2. After extracting the stem and plotting its pattern we can recognize the position of infixes after the fa' or the 'ayn' of the word (بعد فاء أو عين الكلمة). For example, in the word ALMADROSA, we have the infix W (و) after the 'ayn' of the word (بعد عين الكلمة) which map the word to the noun class by rule 4. The previous rules give us a high chance to presume the class of a given word, but some words may not have any affixes to guide us to the corresponding TAG, in such cases we pass the word to the final phase i.e. the syntax analyzer. The morphological analyzer can be summarized by the following algorithm:

```
Begin
    take word; extract the stem;
    extract all affixes with defined positions
    test the seven rules
    if one of them satisfied then
        return the corresponding tag
   else transfer word to the syntax analyzer
End
```

4.3. Syntax Analyzer

This phase is used if the word does not have any affixes to guide the morphological analyzer. Syntactic analysis is probably the most well-studied and well-understood aspects of language processing. We used two rules to map ambiguous word to the corresponding TAG: sentence context and reverse parsing.

4.3.1. Sentence Context

This stage depends on the relations with adjacent and related words in a phrase or sentence. In Arabic language, the position of the word in the sentence is a good indicator to identify a noun from a verb.

Prepositions (حروف الجر) and interjections (حروف النداء) are always followed by nouns such as in the word: "fe almadrasa" (في المدرسة) and in the word "ya Mohamed" (يا محمد). Some words are always followed by nouns such as the words: السيد، الشيخ، المملكة، الرئيس [6].

4.3.2. Reverse Parsing

The morphological analyzer succeeds in solving almost non-vocalized words in Arabic corpus, but there are some words that have ambiguity structure which prevents the morphological analyzer from guessing its class. For example the word KTB (كتب) may be a verb which means "write" or a noun which means "books".

In this study we have developed an Arabic context-free grammar to determine the class of the words of this type. For example, in the sentence:

(ذهب الولد الى المدرسة)
(??? اسم حرف جر اسم)
(thahaba alwalado ela almadrasate)
(??? NOUN preposition NOUN)

The rule with (verb, noun, preposition, and noun) is matched and the class of the word verb (thahaba ذهب) classified as a verb. In the sentence " ذهب الولد الى

المدرسة", the word ذهب is ambiguous and it has been failed to identify its TAG. While the other words in the sentence has been succeeded in identifying its TAG. The word الولد is NOUN which satisfies the Rule1, the word الى is a preposition particle which stored in lexicon, and the word المدرسة is NOUN which satisfies the Rule1. When we compared the above sentence with the stored Arabic language rules, we found that the rule (verb فعل ، noun اسم، prepositionحرف جر، noun اسم) is matched with the sentence. When we ignored the word ذهب and traced the matched Arabic language rule and the sentence alternatively; we can guess and return the TAG which is matched the corresponding ambiguous word, so we return verb (فعل) TAG to the ambiguous word ذهب. The number of rules used in reverse parsing is the 10 most frequently used Arabic rules which contain VP or NP or both. These rules cover the following sequences: *verb noun, verb noun noun, verb noun particle noun, verb noun particle noun noun, verb noun noun particle noun, noun noun, noun verb noun, noun verb, particle noun verb noun, noun verb particle noun.* If there is more than one rule that matches with the sentence to analyze we ignore the word which being unanalyzed word.

We can summarize the process of reverse parsing by the following algorithm:

Begin
List sequence of tags corresponds to each word
Ignore the tag of ambiguity word
Compare a sequence of tags with a stored cfg
When one grammar rule matched
Trace the sentence with the matched rule
Return the tag of the ambiguous word
End

5. Experiment and Evaluation

We have tested the accuracy of the proposed approach using data set consisting of 2355 non-vocalized Arabic words in 10 randomly selected newspaper articles. Table 1 shows the number of words in each article, the number of successful guessing TAGs that map to words and the percentage of successful guessing TAGs. It can be seen from Table 1 that the system succeeded to analyze 2211 words and map them to the corresponding TAGs, and failed to analyze 144 words, i.e., it got a successful rate of 94%.

Some of the unanalyzed words gave incorrect results and the others are unanalyzed; the percentages of incorrect results and unanalyzed words can be measured by recall and precision.

We have calculated the precision and the recall of the proposed approach using the following formulas:

$$Recall = \frac{Correct}{Correct + UnAnalyzed}$$

$$Precision = \frac{Correct}{Correct + Incorrect}$$

Table 1. Accuracy of word classification system.

Articles	Number of Words	Number of Successful TAGs	Percentage of Successful
Article1	240	226	94%
Article 2	173	164	95%
Article 3	254	238	94%
Article 4	396	369	93%
Article 5	127	119	94%
Article 6	147	138	94%
Article 7	208	198	95%
Article 8	361	339	94%
Article 9	282	263	93%
Article 10	167	157	94%

The results of this stage are given in Table 2. It can be seen from this Table that the system obtained about 98% overall precision for the analyzed words. This Table also shows that the system obtained about 96% overall recall for analyzed and unanalyzed words. Words that did not match the correct TAG are ignored. The proposed word classification system misclassified 4% of tested words, gave 2% incorrect results from analyzed words and succeeds in analyzing and got correct results of 94% from tested words, as shown in Figure 4.

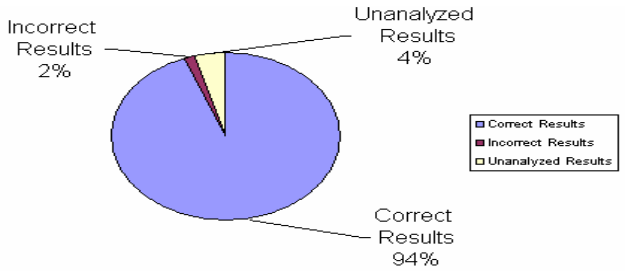

Figure 4. Final results.

6. Conclusions

We have designed and implemented a rule-based classification system to solve the problem of automatically annotating non-vocalized Arabic text with tags. We store all particles and fixed words in the lexicon, and we have revealed how to use a morphological analyzer for tokenization by extracting prefix and suffix and extracting infixes from the pattern of the stem, then trace the rules until one of them is matched.

We have demonstrated how to use a sentence context and the structure of Arabic language to construct a reverse parsing for solving most unanalyzed words in the morphological analyzer. All three analyzers in the proposed system can be used

successfully for determining a high percentage of word classes.

Most previous works used prefix and suffix analysis, but the proposed approach has the advantage of using prefix, suffix and infix analysis. Adding infix analysis helped in solving many ambiguous cases of nouns and verbs that have similar prefixes and suffixes.

The position of the word in the sentence is a good indicator in identifying nouns. Many researchers used these phenomena to construct a rule to help in identifying nouns in the text like in [22] and others used them to identify personal names in the text like QARAB system [12]. In contrast, our approach is capable of providing full coverage to identify both nouns and personal names. Our approach also used these phenomena to construct a new technique i.e. the reversed parsing technique which scans the available grammars of Arabic language to get the class of a single ambiguity word in the sentence based on its position.

Table 2. Accuracy of word classification system.

Articles	No. of Correct Result	No. of Incorrect results	No. of Misclassified Words	Precision	Recall
Article1	226	4	10	98%	96%
Article 2	164	4	5	98%	97%
Article 3	238	4	12	98%	95%
Article 4	369	8	19	98%	95%
Article 5	119	4	4	97%	97%
Article 6	138	3	6	98%	96%
Article 7	198	4	6	98%	97%
Article 8	339	8	14	98%	96%
Article 9	263	6	13	98%	95%
Article 10	157	3	7	98%	96%

7. Future Work

Many techniques have been proposed to tag English and other European language corpora. One of these techniques developed was the rule-based technique and all other techniques are extended to it. Rule-based technique is the technique we used in our system, so we can utilize from our rules in the morphological analyzer to construct a new technique like statistical model or semantic analysis to map a given word to the corresponding TAG.

References

[1] Abuleil S. and Evens M., "Discovering Lexical Information by Tagging Arabic Newspaper Text," CSAM, Illinois Institute of Technology, Chicago, 1998.

[2] Abuleil S., Alsamara K., and Evens M., "Acquisition System For Arabic Noun Morphology," *in Proceedings of the Workshop on Computational Approaches to Semitic Languages*, USA, pp. 1-8, 2002.

[3] Abuleil S., "Extracting Names from Arabic Text for Question Answering Systems," *MIS Department*, Chicago State University, 2004.

[4] Abuleil S. and Alsamara K., "Enhance the Process of Tagging and Classifying Proper Names in Arabic Text," *in Proceedings of the International Arab Conference on Information Technology (ACIT'2006)*, Jordan, pp. 43, 2006.

[5] Al Shamsi F. and Guessoum A., "A Hidden Markov Model-Based POS Tagger for Arabic," *in Proceedings of the 8th International Conference on the Statistical*, 2006.

[6] Al-Shalabi R. and Kanaan G., *Constructing an Automatic Lexicon for Arabic Language*, Yarmouk University, Jordan, 2004.

[7] Alqrainy S. and Ayesh A., "Word-Class Tagger and Tagset Design for Vocalized Arabic Text," *in Proceedings of the 2nd Jordanian International Conference on Computer Science and Engineering (JICCSE 2006)*, Jordan, pp. 278-283, 2006.

[8] A Web of Morphology, http://angli02.kgw.tuberlin.de/call/webofdic/morph.html, 2007.

[9] Chiraz Z., Aroua T., and Mohamed A., "A Multi Agent System for POS-Tagging Vocalized Arabic Texts," *International Arab Journal of Information Technology (IAJIT)*, vol. 4, no. 4, pp. 322-329, 2007.

[10] Diab M., Hacioglu K., and Jurafsky D., "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," *Linguistics Department*, Stanford University, 2004.

[11] Habash N. and Rambow O., "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop," *in Proceedings of the Annual Meeting on Association for Computational Linguistics*, Michigan, pp. 573-580, 2005.

[12] Hammo B., Abu-Salem H., and Lytinen S., "QARAB: A Question Answering System to Support the Arabic Language*," in Proceedings of Workshop on Computational Approaches to Semitic Language (ACL)*, Philadelphia, pp. 55-65 2002.

[13] Freeman A., "Brill's POS Tagger and a Morphology Parser for Arabic*," Department of Near Eastern Studies*, Michigan, USA, 2001.

[14] Khoja S., "APT: Arabic Part-of-Speech Tagger," *Computing Department*, Lancaster University, Lancaster, 2003.

[15] Maamouri M. and Cieri C., "Resources for Arabic Natural Language Processing at the Linguistic Data Consortium*," in Proceedings of the International Symposium on Processing of Arabic language Faculté des Lettres*, Tunisia, 2002.

[16] Marsi E. and Soudi A., *Memory-based Morphological Analysis Generation and Part-of-Speech Tagging of Arabic*, Tilburg University, 2006.

[17] Mohamed A., "A Large-Scale Computational Processor of the Arabic Morphology and Applications," *Master Thesis*, Faculty of Engineering, Cairo University, Egypt, 2000.

[18] Mol V, "The Semi-Automatic Tagging of Arabic Corpora," *Katholieke University*, ILT, 2004.

[19] Mustafa S. and Awwad S., "Arabic Word Class Tagging Based on the Analysis of Affix Structure," *in Proceedings of the International Arab Conference on Information Technology (ACIT'2006)*, Jordan, pp. 145-145, 2006.

[20] Nachum D., "Part of Speech Tagging," *Seminar in Natural Language Processing and Computational Linguistics*, USA, 2007.

[21] Safadi H., Dakkak O., and Ghneim N., *Computational Methods to Vocalize Arabic Texts*, Syria, 2006.

[22] Talmon R. and Wintner S., *Morphological Tagging of the Qur'an*, University of Haifa, Israel, 2001.

[23] Young-Suk L., Papineni K., and Roukos S., "Language Model Based Arabic Word Segmentation*," in Proceedings of the Annual Meeting on Association for Computational Linguistics,* Japan, pp. 399- 406, 2003.

[24] الدكتور عبد العزيز عتيق، "المدخل الى علم الصرف"، دار النهضة العربية، ١٩٧١.

[25] الدكتور زين كامل الخويسكى، "الزوائد في الصيغ في اللغة العربية"، دار المعرفة الجامعية، ١٩٧١.

[26] ابن القطاع الصقلي، "أبنية الأسماء و الأفعال و المصادر"، دار الكتب و الوثائق القومية، ١٩٩٩.

[27] فارس محمد عيسى، "علم الصرف منهج في التعلم الذاتي"، دار الفكر للطباعة و النشر و التوزيع، عمان، الأردن، ٢٠٠٠.

Appendix A

Table 3. Lexicons.

حروف الجر Prepositions	من، الى، إلى، حتى، حتّى، على، عن، في، ربّ، مذ، منذ، خلا، عدا، حاشا
حروف العطف Conjunctions	و، ثم، ثمّ، أو، أم، بل
حروف النفي Negation Particles	لم، لما، لن، ما، لا، لات
حروف الجواب Answering Particles	نعم، بلى، إي، أجل، كلّا
حرف تفسير Explanation Particle	أي
حروف الشرط Conditional Particles	إن، إذما، لو، لولا، لوما، أمّا، لمّا
حروف تحضيض و تنديم	هلّا، ألّا
حروف العرض	ألا، أما
حروف مصدرية Verbal Noun Particles	أنّ ، أن، كي
حرف استقبال Future Particle	سوف
حروف التوكيد Emphasis Particles	إنّ، قد
حرف استفهام Interrogative Particles	هل
حرف تمني Wishing Particle	ليت
حرف ترجّي و إشفاق	لعل، لعلّ
حروف النداء Interjections	يا، أيا
إن حرف مشبه بالفعل و نا ضمير مستتر	إنّا
حرف استثناء Exceptive Particle	إلّا
إن الشرطية و ما زائدة	إمّا
كأن الشرطية و ما زائدة	كأنّما
الكاف حرف جر و ما زائدة	كما
ربّ حرف جر و ما زائدة	ربّما
ضمائر المتكلم First Person Pronouns	أنا، نحن
ضمائر المخاطب Second Person Pronouns	أنت، أنتما، أنتم، أنتن
ضمائر الغائب Third Person Pronouns	هو، هم، هن، هما
أسماء الإشارة Demonstrative Pronouns	هذا، ذلك، هذي، تلك، هذان، هاتان، هؤلاء، أولئك، هنا، هنالك، هذه
أسماء موصولة Relative Pronouns	الذي، التي، اللذان، اللتان، الذين
ظروف المكان Nouns of Place	فوق، تحت، أمام، وراء، يمين، شمال، خلف، إثر
ظروف الزمان Nouns of Time	حين، وقت، ساعة، يوم، شهر، سنة، عام، زمان، أوان، بكرة، ضحوة، ليلة، مساء، عشية، غدوة
ظروف زمان أو مكان Nouns of Place or Time	دون، بين، وسط، عند، قبل، بعد
إنّ و أخواتها Inna and its sisters	أن، إنّ، لكنّ، كأنّ، لعل، ليت
كان و أخواتها Kaana and its sisters	كان، أمسى، أضحى، أصبح، ليس، بات، مالبث، مابرح، ما انفك

Appendix B

Table 4. Common prefixes.

No. of Letters	Prefixes
5	"المست"
4	"سنست" ،"لتست" ،"فيست"، "سيست" ،""ليست" "بالت" ،"فنست" ،"سنست"، "لنست" ،"فتست"
٣	"است" ،"تست"، "فال" ،"بال" ،"تست" ،"يست" "وال" ،"للت" ،"للأ" ،"سيت"، "كال" ،"مست"
٢	"لت" ،"لن" ،"لل"، "ال" ،"سن" ،"ست" ،"سي" ،"سأ" "تت" ،"يت" ،"نت"، "لا" ،"لي"
١	"ا"،"ي"،"ن"،"أ"،"ت"،"ل"،"ف"

Appendix C

Table 5. Common suffixes.

No. of letters	Suffixes
5	"اتهما"
4	"وهما"، "تهما"
٣	"وها"، "تما" ،"تهم" ،"تنا" ،"تني" ،"وهم"
٢	"ات" ،"كن" ،"ين"، "هن" ،"كم" ،"تن" ،"هم" ،"تم" ،"ها"، "نا" ،"ون" ،"وك" ،"تك"، "ان" ،"وا"
١	"ا""، "ي" ،"ت" ،"ه" ،"ن" ،"ك" ،"ة"

Appendix D

The stem patterns or words with infixes after eliminating prefixes and suffixes which extracted and used in the morphological analyzer of our system, can be abstracted by:

فعل، فعال، فتعل، فتعال، فعوعل، فاعل، فاعيل، فعول، فعيل، فعائل، فوعل، فيعل، فواعل، فاعول، فواعيل، فعلل، فعالل.

Ahmad Al-Taani is an associate professor of artificial intelligence at Yarmouk University, Jordan. He received his Bachelor of science in computer science in 1985 from Yarmouk University, Jordan. He received his Master of science in software engineering from National University, USA in 1988. He received his PhD in computer vision from University of Dundee, UK in 1994. His research interests includes image processing, Arabic language processing, machine translation, and Arabic web page classification.

Salah Abu Al-Rub is an oracle developer and maintainer at Amman Stock Exchange. He received his Bachelor of computer science in 2005 from Yarmouk University, Jordan. He received his Master of computer science in 2007 from Yarmouk University, Jordan. His research interests include Arabic stemmers and Arabic part-of-speech tagging.