Automated Tagging System And Tagset Design For Arabic Text

[1]Shihadeh Alqrainy,[2] Alladdin Ayesh,Jordan,[1] Hasan Almuaidi
[1]Faculty of Science and IT, Albalqa' Applied University
Jordan
{alqrainy,alserhan}@bau.edu.jo

[2]School of Informatics
De Montfort University
UK
aayesh@dmu.ac.uk

ABSTRACT: *This paper presents diacritics rule-based part-of-speech (POS) tagger which automatically tags a partially vocalized Arabic text. The aim is to remove ambiguity and to enable accurate fast automated tagging system. A tagset is being designed in support of this system. Tagset design is at an early stage of research related to automatic morphosyntactic annotation in Arabic language. Preliminary results of the tagset design have been reported in this paper. Arabic language has a valuable and important feature, called diacritics, which are marks placed over and below the letters of Arabic word. This feature plays a great role in adding linguistic attributes to Arabic words and in indicating pronunciation and grammatical function of the words. This feature enriches the language syntactically while removing a great deal of morphological and semantically ambiguities.*

Key words: Arabic Language, Part-Of-Speech (POS), Diacritics, Tagset, Morphological, Syntactical

Received: 12 July 2010, Revised 31 July 2010, Accepted 4 August 2010

1. Introduction

Arabic language is syntactically and morphologically a rich language, which means several words and meanings, can be derived from the same word leading to ambiguity. The ambiguity of Arabic lies on 3 different levels, the core word level, the derived word forms and agglutinative forms of words [1].

In this paper, we exploit the effect of vocalization, which is considered one of the Arabic Language distinctive features, on the tagging process. It is envisaged that the use of vocalization will increase the speed of the tagging process without scarifying accuracy. Indeed, the use of vocalization, as we demonstrate in this paper, will reduce the ambiguity of the parsed text.

The paper starts with a brief summary of the Arabic language overview followed Diacritics in Arabic Language. The tagset design and uses and benefits of tagging systems are highlighted. Then, we present our tagging system architecture and diacritical rule-based as our approach. Finally, analyses of experiment results are presented with future work and conclusion.

2. Arabic Language Overview

Arabic is considered to be the largest member of Semitic branch of the Afro-Asiatic language family and most widely spoken Semitic language today. languages, and the only family of this group spoken in Asia.

Arabic has been a literary language since the 6th century A.D., and is the liturgical language of Islam in its classical form. Its closest living relatives are Hebrew and Aramaic. Arabic is the official languages of more than 20 Arab countries. A substantial

number of Arabic speakers live in Israel, parts of Africa, Iran and France.

Arabic has several varieties, all of which play an important role in Arabic-speaking countries. These are: Classical Arabic, Modern Standard Arabic (MSA) and Colloquial (spoken) Arabic.

3. Diacritics In Arabic Language

Arabic language has a valuable and important feature, called diacritics, which are marks placed over and below the letters of Arabic word. This feature plays a great role in adding linguistic attributes to Arabic words and in indicating pronunciation and grammatical function of the words. It is particularly of interest for the purpose of this paper.

Arabic is 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. Two different words may have identical spelling whereas their pronunciations and meanings are totally different. They also indicate the grammar function of the word within the context of the sentence [2].

The Arabic alphabet consists of 28 consonants, but 3 of these are also used as long vowels. In English there are five vowel letters ie. (a,e,i,o.u). Unlike English, in Arabic there are two types of vowels :
1- Short Vowels, Arabic has three short vowels described in Table-1.

Sign		*Description*	*Example*
Fat ha	/a /	Marks over the Consonant	بَ
Kas ra	/i /	Marks below the Consonant	بِ
Thm ah	/u /	Marks over the Consonant	بُ

Table 1. Short Vowels

In Arabic, short vowels are not a part of the Arabic alphabet, instead they are written as marks over or below the consonant. They are used in both Noun and Verb in Arabic Language. They indicate the case of the noun and the mood of the verb. Examples and detailed explained in Tagset Analysis section.

2- Long Vowels, Arabic also contains three long vowels described with examples in Table-2.

ا ALIF	و WAW	ي YA
/a:/	/u:/	/i:/
C<u>a</u>r	F<u>oo</u>t	W<u>ea</u>k

Table 2. Long Vowels

There are other marked by diacritics used in Arabic language (Table-3).

Sign			*Example*
Consonant Doubling	Shadda	ّ	بَّ
Vowel Absence	Sukun	ْ	بْ
Tanween Alfatha	/an/	ً	بً
Tanween Alkasr	/in/	ٍ	بٍ
Tanween Adamm	/un/	ٌ	بٌ

Table 3. Other Diacritics

Many words are in general ambiguous in their part-of-speech, for various reasons. In English, for example, a word such as {\it “Make”} can be “Verb” or “Noun”.

In Arabic there are ambiguities. For example, the word “ ذهب “ which either means “*go* “ or “ *gold* “ can be Verb or Noun.

Diacritics are used to prevent misunderstandings, to determine the correct pronunciation, reduce the ambiguity, and indicating grammatical functions. These functions play a great role in removing ambiguity and enabling accurate fast automated tagging system.

To remove ambiguity and to determine the correct tag of the word “ ذهب “ in the above example, adding the short vowels (Fat ha sign) to the last letter of the word to become “ ذهبَ “ enough to get the correct tag [Verb] without any ambiguity and without regards to the context.

4. Tagset

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 [19]. The development of a tagset to support diacritical based tagging system is at early stage. The need for such a tageset comes from the fact that there is no standardized and comprehensive Arabic tagset.

EAGLES [16] guidelines outline a set of features for tagsets, these guidelines were designed to help standardise tagsets for what were then the official languages of the European Union. EAGLES tags are defined as sets of morphosyntactic attribute-value pairs (e.g. Gender is an attribute that can have the values Masculine, Feminine or Neuter).

The tagset discussed here is not being developed in accordance with the EAGLES guidelines for morphosyntactic annotation of corpora. Arabic is very different from the languages for which EAGLES was designed, and belongs to the Semitic family rather than the Indo-European one. Following a normalized tagset and the EAGLES recommendations would not capture some of Arabic’s relevant information, such as the jussive mood of the verb and the dual number that are integral to Arabic. Another important aspect of Arabic is inheritance, where all subclasses of words inherit properties from the classes from which they are derived. For example, all subclasses of the noun inherit the Tanween “*nunation*” when in the indefinite which is one of the main properties of the noun [16].

5. Previous Work In) Pos) Tagsets

There are small number of popular tagsets for English, such as: 87-tag tagset used Brown Corpus, 45-tag Penn Treebank tagset and 61-tag C5 tagset [3]. For Arabic also very small number of tagset had been built, El-Kareh S, Al-Ansary [10] described the tagset, they classifying the words into three main classes, Verbs are subclassified into 3 subclasses; Nouns into 46 subclasses and Particles into 23 subclasses. Shereen Khoja [14] described more detail tagset. Her tagset contains 177 tags, 57 Verbs, 103 Nouns, 9 Paricles, 7 residual and 1 punctuation.

6. Arabic Tagset

We have based our Arabic tagset on inflectional morphology system. The traditional description of Arabic grammarians consider as a base to create the linguistic categories of Arabic tagset. Arabic grammarians describe Arabic as being derived from three main categories: noun, verb and particle. (Figure-1).

The verb in the Arabic language implies a state or action and a notion of time combined with them. The verb in the Arabic language has several aspects: Perfect, Imperfect and Imperative.

Nouns are also divided into the following types: (Common, Demonstartive, Relative, Personal, Adverb, Diminutive, Instrument, Conjunctive, Interrogative, Proper, Adjective).

In Arabic, particles are classified as one of the three main categories as part of speech, Some of the particles activate the verb (i.e. Subjunctive, Jussive), some of them activate the noun (i.e. Preposition, Exception), and some activate both the noun and

the verb (i.e. Conjunction).

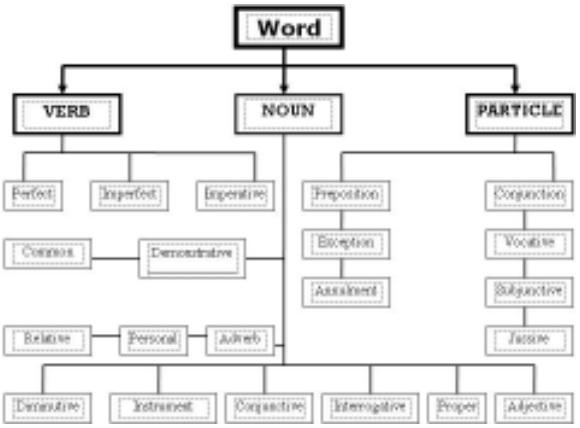

Figure 1. Tagset Hierarchy

The tagset has the following main formula:

[T , S , G , N , P , M , C , F] , Where:

T (Type) = {Verb, Noun, Particle}
S = Sub-Class {(Shown in figure-1)}
G (gender)= {Masculine, Feminine, Neuter}
N (Number) = {Singular, Plural, Dual}
P (Person) = {First, Second, Third}
M (Mood) = {Indicative, Subjunctive, Jussive}
C (Case) = {Nominative, Accusative, Genitive}
F (State) = {Definite, Indefinite}

Table-4, described the Abbreviations which was used to define the words in our tagset.

Let us try to explain the symbols of the tagset formula for a moment.

The symbols **[T, S, G , N , P , M]** consider as linguistic attributes for class Verb, while the symbols **[T, S ,G , N , P ,C ,F]** consider as linguistic attributes for class Noun. For example,the word"كَتَبَ" which means"*he wrote*" has the following tag **[VePeMaSnThSj]**, which means [*Perfect Verb , Masculine Gender , Singular Number , Third Person , Subjunctive Mood*].

7. Tagging Benefits And Related Work

Part-of-speech tagging is the process of assigning a part- of-speech or other syntactic class marker to each word in a corpus [3]. Tagger is necessary for many applications, such as : speech synthesis system, speech recognition system, informational retrieval (IR) and parsing system.

Many techniques have been used to tag English and other European languages corpora. Greene and Rubin [4] developed the first Rule-Based technique to tag Brown Corpus. Eric Brill's [5] interest in rule-based tagger. Garside [15] used hidden Markov Model to develop CLAWS tagger. More recently, taggers that use combination of both Statistical and rule-based [6], Machine learning [7] and Neural Network [8][9] have been developed.

In terms of Arabic, small number of popular Part-of-Speech (POS) tagger have been developed. El-Kareh and Al-Ansary [10] describe a hybrid semi-automatic tagger that uses both morphological rules and statistical techniques in the form of hidden Markov models. Abuleil and Evens [11] describe a system for building an Arabic lexicon automatically by tagging Arabic

Word	Abb	Word	Abb
Verb	Ve	Annulment	An
Noun	Nu	Subjunctive	Sb
Particle	Pr	Masculine	Ma
Perfect	Pe	Feminine	Fe
Imperfect	Pi	Neuter	Ne
Imperative	Pm	Singular	Sn
Common	Cn	Plural	Pl
Adjective	Aj	Dual	Du
Demonstrative	De	First	Fs
Relative	Re	Second	Sc
Personal	Ps	Third	Th
Diminutive	Dm	Indicative	Dc
Instrument	Is	Subjunctive	Sj
Proper	Pn	Jussive	Js
Adverb	Ad	Nominative	Nm
Interrogative	In	Accusative	Ac
Conjunction	Cj	Genitive	Ge
Preposition	Pp	Definite	Df
Vocative	Vo	Indefinite	Id
Conjunction	Co		
Exception	Ex		

Table 4 . Tagset Abbreviations

newspaper text. Shereen Khoja [12] describes an Arabic part-of-speech called APT that uses statistical and rule-based techniques. Diab, Mona et al.[13] present a Support Vector Machine (SVM) based approach to automatically tokenize, part-of-speech tag in Arabic text.

8. Arabic Tagging System

Our tagger system consists of many modules as shown in figure-2.

The input of the tagger system is designed to be Partial Diacritized Arabic text. The user interacts with the system and handles the input/output through Interface Module.

The tokenizer Module locates a document and isolates the words (Token) in the document and stores words in special list.

The syntactical module gets the token from tokenizer module and applies syntactical rules directly to find the part-of-speech tag of the word without return to database (lookup tables). If the module fails to tag the word, it pass the token to the Affixes Analyzer Module.

Affixes Analyzer Module is responsible to find the stem of the word after analyzing the affixes attached the word. Affixes in this module are of two types, prefixes; the extra letters added to the beginning of the word, suffixes; the extra letter at the end of the word.

The Morphological Module is responsible to find the pattern that exactly matches the word by performing the steps of an algorithm, which described later in this paper.

9. Diacritical Rule-Based Approach

Diacritical Rule-based Approach is an our technique which uses Syntactical Information and Morphological Information to assign most likely tag to each unknown and ambiguous word in the text.

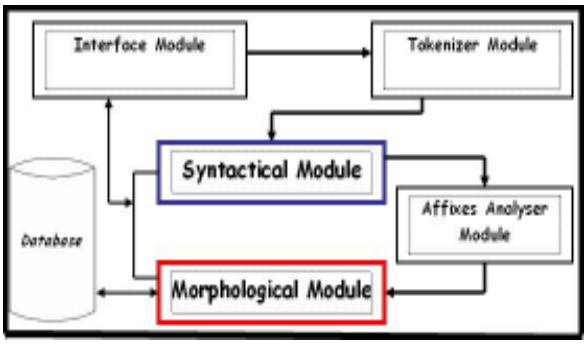

Figure 2. System Model

The proposed approach consists of two types of rules: Syntactical *Rules* and *Morphological Rules.*
Syntactical Rules, based on the diacritics, we applying these rules without regard to context and lookup tables to assign most likely tag of the word.

Some of syntactical rules as examples are listed below:

Consider **W** = The word , **T** = The Tag , **L** = Length of W.

Rule-1 : If **W** end with **"** ِيّ **"** or **"** يّة **", then T = [NuRe].**

For Example, the words **"** مصريّة **",** "مصريّ" **"**

Rule-2 : If **W** begin with " م " followed by kas ra and end with Tanween Adamm, then **T** = **[NuIs]**.

For Example, the words **"** مِفتاحٌ **",** "مِكنسةٌ" **"**

Rule-3 : If L (**W**) = 3 and end with Fat ha, then **T** = **[VePeMaSnThSj]**

For Example, the word **"** لعبَ **"**

Rule-4 : If L (**W**) = 4 and end with Thm ah, then **T** = **[VeImMaSnThDc]**

For Example, the word **"** بكتبُ **"**

Morphological Rules, based on patterns with diacritics. Arabic language has a rich morphological system that contains a lot of patterns. These patterns assign part-of-speech tag of the Arabic word. Some of patterns belong to Verb class, while the others belong to Noun class. Particle has no patterns in Arabic language

In order to applying the Morphological Rules, We need an algorithm describe how we match the correct pattern with the inflected word. The steps of an algorithm with examples described below.

An examples of how an algorithm works, described in figure-3

10. Result

We tested our system to tag the words using partial-diacritization documents from the holly Qur'an and another set chosen randomly from the proceedings of the Saudi Arabian National Computer Conference and other resources. We ran our system on a group of these documents contains 7500 word. The accuracy of our system has been calculated for tagging the words. It achieved 92 \% correctly tagged, 8 \% in errors.

Some errors came from Arabized words which are translated as pronounced from other international languages, such as the word **"**كمبيوتر *computer.* These words do not have a root and a pattern.

Others came from irregular verbs such as the word **"** ضلّ **"**. Also some words in Arabic language consider as primitive verbs, such as**,"**بئس" , "نعم **"** . These words not tagged correctly and need a special treatment.

Pattern-Match-Algorithm

Step-1 : Let N = the number of patterns (P) that have the same length of word (W) ... Call each P
Step-2 : For i = 1 to N
Compute Sim (W, P (i))
Sim = Similarity, the identical letters between W and P
Step-3 : Return P which has the Max (Sim)
Step-4 : Return the tag (T) of (W)
Step-5 : Exit

.

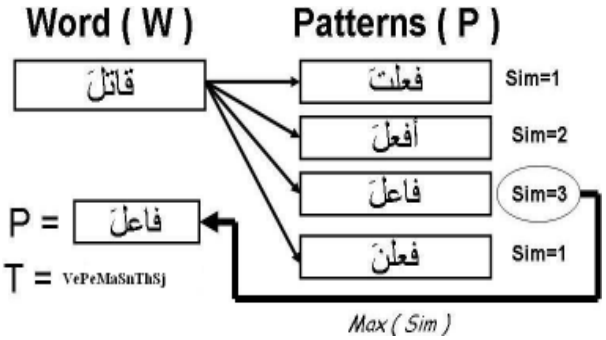

Figure 3. Matching-Pattern Example

11. Conclusion & Future Work

In this paper, we presented diacritics rule-based part-of-speech (POS) tagger which automatically tags a partially vocalized Arabic text. Also, we describe a morphosyntactic tagset that is derived from the ancient Arabic grammar, which is based on Arabic system of inflectional morphology.

The tagset does not follow the traditional Indo-European tagset that is based on Latin but is instead based on the Semitic tradition of analyzing language.

These tags contain a large amount of information and add more linguistic attributes to the word. Also, we are currently collecting much rules to reduce the amount of errors and expanding our tagset to cover most categories word in Arabic language.

It's clear that an overall ambiguity in a vocalized text is quite lower than in an unvocalized text. Diacritics are used to prevent misunderstandings and reduce the ambiguity; diacritics play a great role to speed the tagging process without scarifying accuracy and remove a great deal of morpho-lexical ambiguity when the text is partial diacritization.

References

[1] Mol, M. V (1994). The semi-automatic tagging of arabic corpora," COLING 94, USA.

[2] Elaraby, M. A. (2000). A large scale computational processor of the arabic morphology and application.

[3] Martin, D. J. J. H. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice-hall, USA.

[4] Greene, B., Rubin, G (1971). Automatic grammatical tagging of English, Department of Linguistics, Brown University, Providence, R.I. USA.

[5]Brill, E. (1992). A simple rule-based part of speech tagger, *In*: Proceedings of the Twelfth International Conference on AI. (AAAI- 94),Seattle,bWA.

[6] DeRose., S. J. (1988). Grammatical category disambiguation by statistical optimization, *Computational Linguistics* 14 (1) 3139.

[7] Daelemans, B., Gills (1996). A memory-based part of speech tagger generator, *In*: Proceedings of the Fourth Workshop on Very Large Corpora, Copenhagen, Denmark, p. 1427.

[8]. Marques, N. G. (1996). A neural network approach to part-of-speech tagging," Proceedings of the second workshop on spoken and written Portuguese, Curitiba, Brazil, p. 1-9.

[9] Schmid, H. (1994). Part-of-speech tagging with neural networks, *In*: Proceeding of COLING-94. p. 172-176.

[10] El-Kareh., Al-Ansary. (2000). An Arabic interactive multi-feature pos tagger, *In*: Proceedings of the, ACIDCA conference, Monastir, Tunisia, p. 204- 210.

[11] Abuleil, S., Evens, M (1998). Discovering lexical information by tagging arabic newspaper text, *In*: Workshop on Semitic Language Processing. COLING-ACL.98, University of Montreal, Montreal, PQ, Canada, Aug 16, p. 1-7.

[12] Khoja, S. (2001). Apt: Arabic part-of-speech tagger, *In*: 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. No. 2,.

[13] Diab, K. H., Jurafsky, Mona and D. (2004). Automatic tagging of Arabic text: From raw text to base phrase chunks, *In*: Proceedings of HLTNAACL.

[14] Khojah, G., Knowels. (2001). A tagset for the morphosyntactic tagging of arabic, *In*: Paper presented at Corpus Linguistics 2001, Lancaster University, Lancaster, UK, March 2001, and to appear in a book entitled "A Rainbow of Corpora: Corpus Linguistics and the Languages of the World", edited by Andrew Wilson, Paul Rayson, and Tony McEnery, Lincom-Europa, Munich.

[15] Garside, Roger., Leech, Geoffrey., Sampson, Geoffrey (1987). The Computational Analysis of English: a corpus-based approach. Longman Group UK Limited.

[16] Leech G, Wilson A 1996 Recommendations for the Morphosyntactic Annotation of Corpora EAGLES Report. http://www.ilc.pi.cnr.it/EAGLES96/annotate/

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