Warsaw University
Faculty of Mathematics, Informatics and Mechanics

Ahmed Hussein Aliwy

Arabic Morphosyntactic Raw Text Part of Speech Tagging System

summary of PhD dissertation

Supervisor

Dr hab. Jerzy Tyszkiewicz
Institute of Informatics
University of Warsaw

May 2012

Introduction and Overview:

The topic of this dissertation is morphosyntactic part of speech tagging (abbreviated POS tagging) for Arabic. This topic has long and rich history for other languages, mainly for English.

POS Tagging provides fundamental information about word forms used in sentences of natural language. The method of utilizing this information varies depending on the particular NLP application (Information Retrieval, Machine Translation, etc.), in which it is used.

Tagging is a source of many challenges for researchers to date. These challenges depend very much on the language under consideration. In this dissertation we consider Arabic, a highly inflected language. Although Arabic language is generally quite regular and there are very few irregular forms, very rich and complicated structure of inflection, which in many cases changes the structure of the words, causes high degree of complexity of Tagging. The other hard problem is the lack of Arabic language resources, corpora and other tools. We propose a new Tagset in this dissertation and in this case the scarcity of resources makes the work much more difficult. Tokenization schemes[1] are also a source of problems in tagging.

My dissertation consists of six main parts (see figure 1):

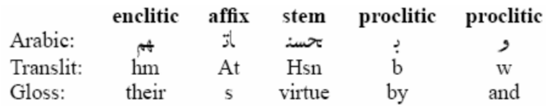
1. Introduction to Arabic language
2. Designing Tagset
3. Constructing special dictionary.
4. Tokenization
5. Lemmatization and Analyzing
6. Tagging
    - Master-slaves tagger
    - Stacking of Master-slaves and rule-based taggers.

The input to the proposed system is the raw text and the output is POS tagging, features and lemma for each word in the text.

Introduction to Arabic language

Arabic is a Semitic language with rich templatic morphology. An Arabic word may be composed of a stem (consisting of a consonantal root and a template), plus affixes and clitics. The affixes include inflectional markers for tense, gender, and/or number. The clitics include some (but not all) prepositions, conjunctions, determiners, possessive pronouns and pronouns. Some are proclitic (attaching to the beginning of a stem) and some enclitics (attaching to the end of a stem). The following is an example of the different morphological segments in the word (وبحسناتهم) which means *and by their virtues*:

[1] See Habash book page tokenization schema.



	enclitic	affix	stem	proclitic	proclitic
Arabic:	ـهم	اتـ	حسنـ	بـ	و
Translit:	hm	At	Hsn	b	w
Gloss:	their	s	virtue	by	and

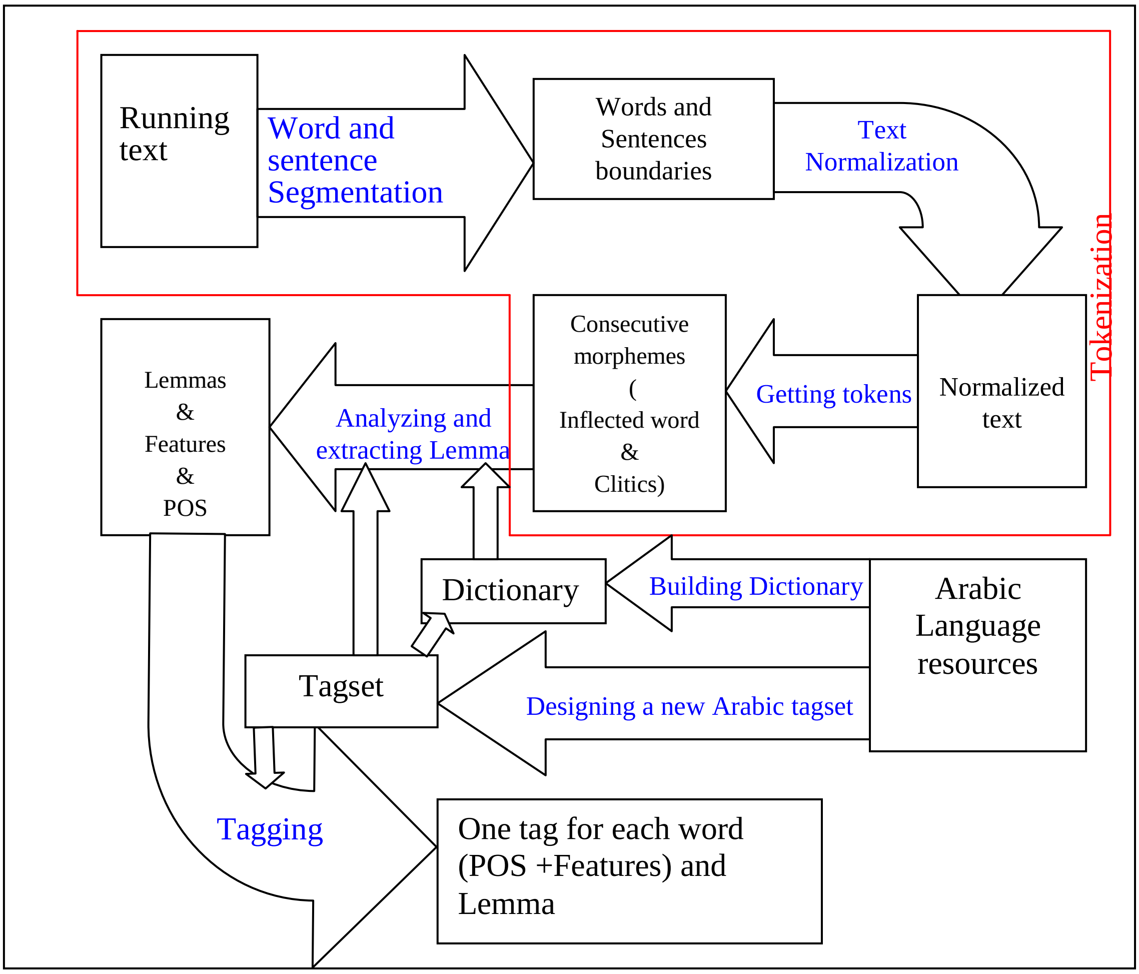

Running text
Word and sentence Segmentation
Words and Sentences boundaries
Text Normalization
Tokenization
Normalized text
Getting tokens
Consecutive morphemes (Inflected word & Clitics)
Analyzing and extracting Lemma
Lemmas & Features & POS
Dictionary
Building Dictionary
Arabic Language resources
Tagset
Designing a new Arabic tagset
Tagging
One tag for each word (POS +Features) and Lemma

Figure 1 the whole tagging system.

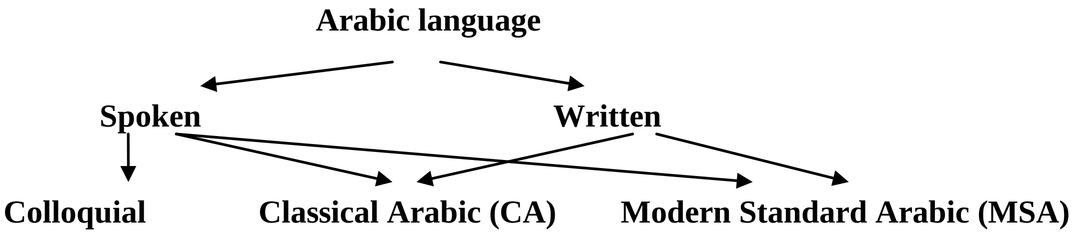
Arabic texts could be either a vowelled text such as the language of Qur'an or children's books; or an unvowelled one used in newspapers, books, and media. For unvowelled Arabic, there are many possible morphological analyses corresponding to alternative vowelings, so tagging is more like full-blown "understanding" of the text.

The level of difficulty of word construction depends on the language type. In English, the word-formation is easy but in Arabic language it is more difficult because the word is inflected according to number, type, gender …etc.

Morphology, which is important in my dissertation, is a branch of grammar which examines the forms of words as well as the principles of word-formation and inflection. Here, word formation is the creation of a new word which can be classified according to rules into:

- Derivational rules: relate a lexeme to another lexeme (changes them from one syntactic category into words of another syntactic category or from one meaning to another)
- Compound: attaches two or more words together to make them one word

Many linguists make a distinction between Classical Arabic (CA), the name of literary language of the previous eras, and the modern form of literary Arabic, commonly known (in English) as Modern Standard Arabic (MSA). In terms of linguistic structure, CA and MSA are largely but not completely similar:



My Tagset:

Tagset: a set of tags (symbols) representing information about part of speech and about values of grammatical categories (case, gender, etc.) of word forms.

The Tagset is the basis of almost all NLP fields. A good Tagset is very important in NLP fields and is the foundation stone in these fields. Some researchers of the Arabic language were constructing Tagsets depending on the English language and were missing some of the important features in the Arabic language. Other researchers made special Tagsets depending on the Arabic language and taking some features from other languages, but these Tagsets didn't take in all the important Arabic language features. Some others summarized all Arabic language features including tags which were not useful.

In this paper, 10 Arabic Tagsets are compared and their limitations indicated. We present a new Arabic Tagset avoiding these limits. The design is intended for Arabic language only and is not based on Tagsets for other languages. It is a multilevel Tagset compatible with traditional and modern standard Arabic. The noun classes have three levels (fixed POS types, grammatical feature and changed POS types), Verbs have two levels (POS types and grammatical features) and particles have two levels (working and meaning). We also introduce the notion of Tagset interleaving. Figures 2 to 7 present my Tagset.

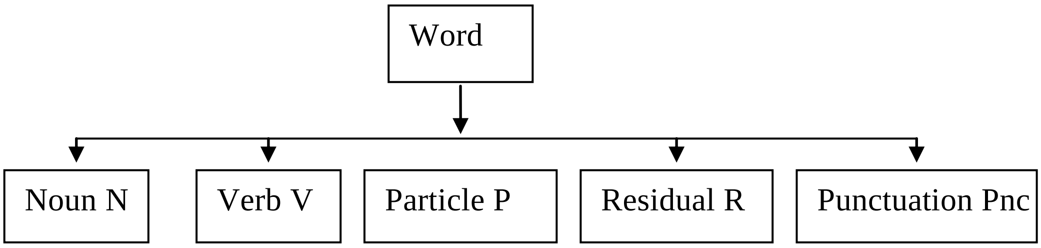


Figure 2: Main POS

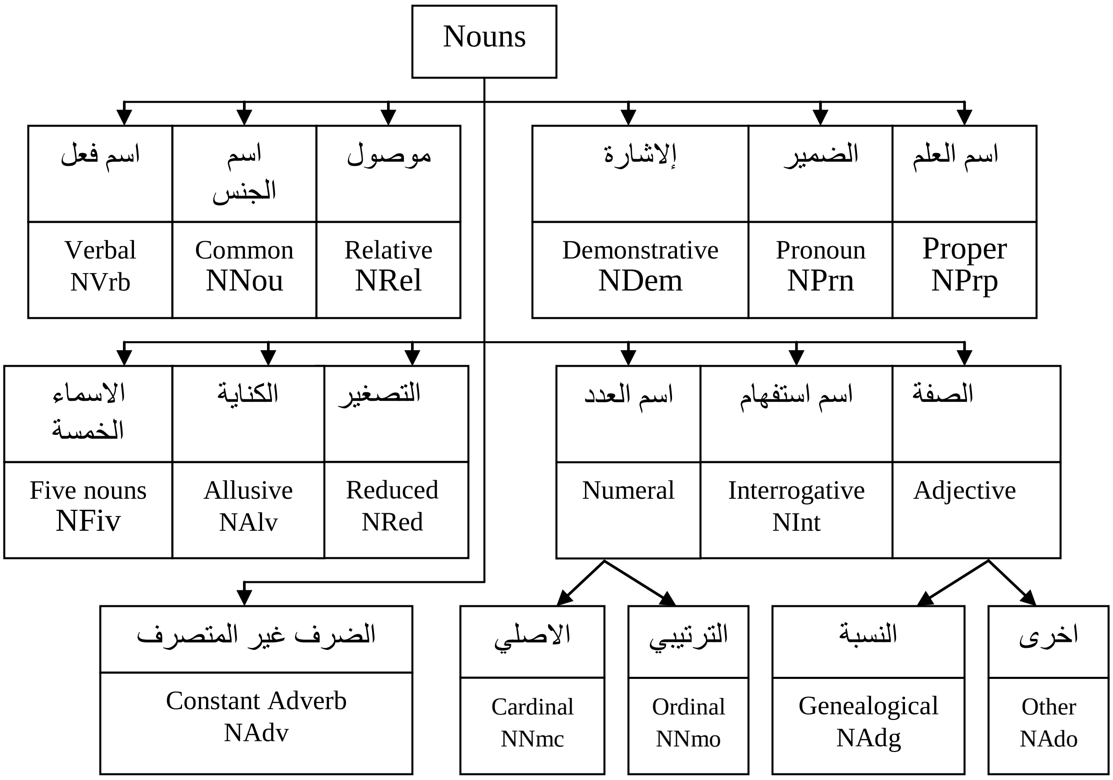


Figure 3a: Arabic Noun Classes in my TAGset

Gender:	Masculine	Feminine	Common
Number:	Singular	Plural	Dual
Case:	Nominative	Accusative	Genitive
Structured[2]	Yes	No	

Figure 3b: features of Noun in my TAGset

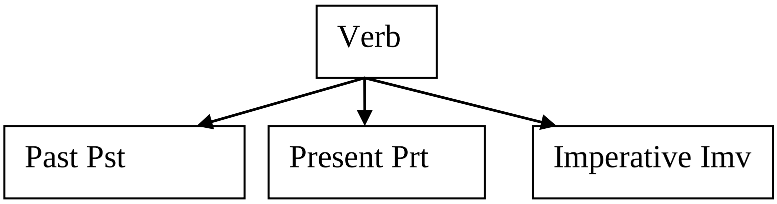


Figure 4a: Verb classes in my TAGset

Gender:	Masculine	Feminine	Common (مشترك)	
Number:	Singular	Plural	Dual	
Person:	First	Second	Third	
Mood:	Nominative	Accusative	Jussive	Non
Certainty	Yes	No		
Structured	Yes	No		
Voice	Passive	Active		

Figure 4b: Verbal attributes in my TAGset

[2] The word ending will be changed (letter or diacritics) according to the case of the word (nominative accusative …). In the case of structured word, the word ending will be constant at all word cases (nominative, accusative …)

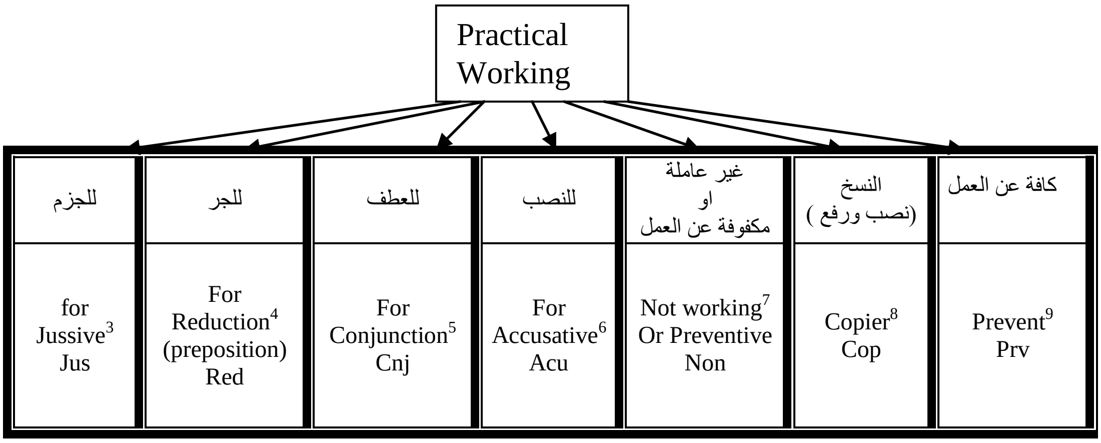


Figure 5a: The classes of particles (working) in my TAGset

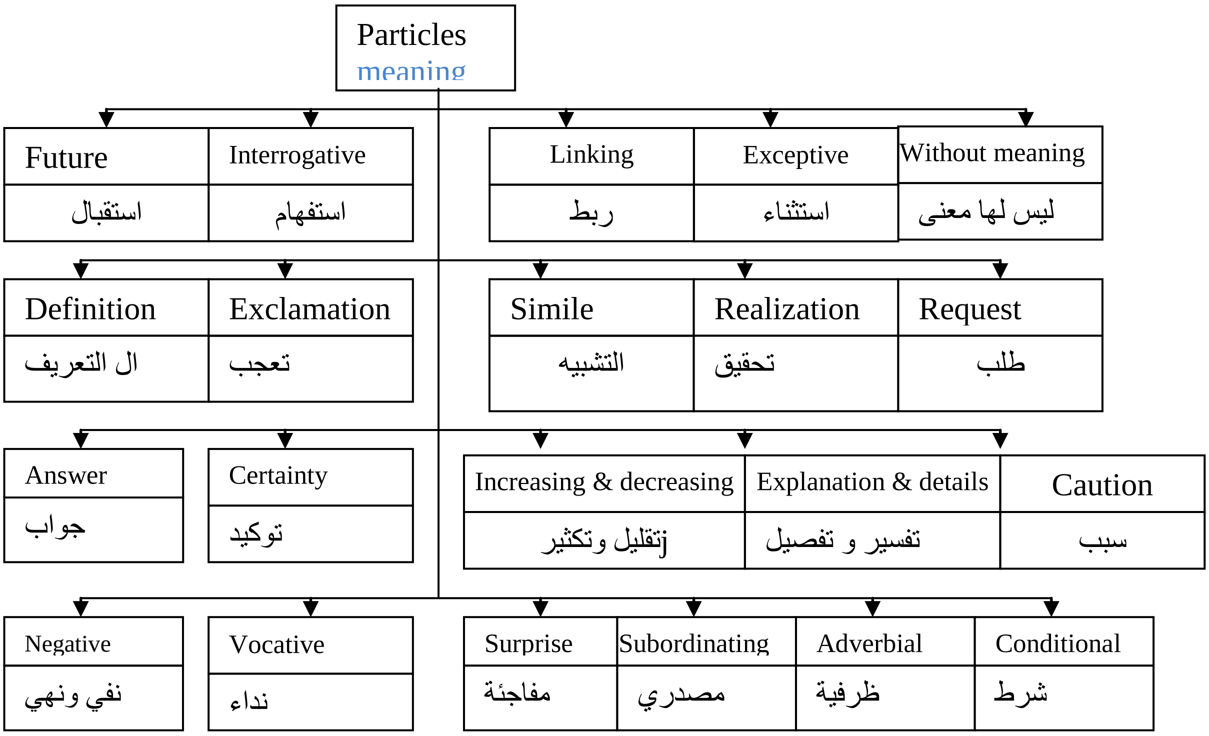
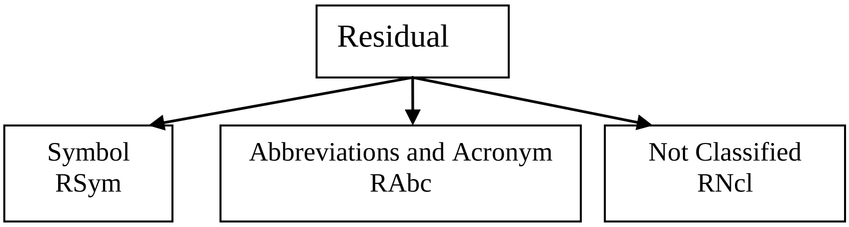


Figure 5b: Particles meaning in my TAGset (features).



[3] The present after these particles is in Jusive mood.
[4] The Noun after these particles is in Genetive case.
[5] The conjucted nouns or verbs by these particles are having the same case.
[6] The nouns or verbs after these particles are in accusative case.
[7] They did not have any effect on the following word.
[8] They have dual affect on the following words. One of the following word in nominative and the other in accusative
[9] Any particle after this particl will be not working (i.e. prevented from working).

Figure 6 residuals classes

فاعل (Subject of a verb)	مفعول مطلق (Cognate)	مستثنى (Excepted)
نائب فاعل (Passive subject representative)	مفعول لأجله (Accusative of purpose)	منادى (Vocative)
مبتدأ (subject)	مفعول معه (Commutative object)	مضاف اليه (Possessive construction)
خبر (Predicate of a subject)	حال (Circumstantial accusative)	بدل, نعت (Apposition)
مفعول به (Object of a verb)	تمييز (Specification),	X NOT USED[10]

Figure 7 syntactic classes of noun (level three).

Constructing special dictionary

As preprocessing stage we construct several dictionaries. These dictionaries have similar rôle as the dictionaries used in Buckwalter Analyzer, and provide lemma in addition to POS and Features.

For verbs: This dictionary consists of 6000 verbs inflected in all possible forms according to the templates used by AL-Dahdah with adding certainty and jussive case. Then all these inflections are sorted and encoded in away such that we can find them efficiently. The input to the dictionary is an inflected verb in any tense or case and the output are its lemma and features. We use this large dictionary for one reason which is to get rid the problem of the changing which happened in the inflected verb.

We deal with this dictionary by constructing dictionaries of prefixes, suffixes and stems as in the Buckwalter analyzer, and additionally the lemmas coded in the dictionary of stems. In this case we must induce the stem from the inflection of the verb. For example, when the verb "قال" "say" is inflected, we get "قال", "قول", "قيل" and "قل" as possible stems from 163 inflections of this verb.

In case of particles we have a list of all particles with their workings and meanings, and therefore the analyzing process is a search problem in cases of verbs and particles.

In case of nouns, adjectives and so on, we collected them from many resources. We added inflections and derivations as feminine (if applicable), numbers, genealogically (Yaa Alnasabi) and reduced noun.. There are many classes of nouns which are closed sets, for example question nouns, numerals nouns and so on. The resulted dictionary is updatable.

Tokenization and segmentation

Tokenization is the task of separating out words (morphemes) from running text [1]. It is also sometimes called segmentation. It refers to the division of a word into clusters of consecutive morphemes, one of which typically corresponds to the word stem, usually including inflectional morphemes [2]. We can use blanks (white space) to help in this task, but there are hard cases. This definition is for English language but for Arabic the situation is different. While discussing

[10] My plaining is designing Tagset and building POS Tagger for arabic. The Level three is outside the range of my Tagger therefore I used this letter to indicate non used level (level three) as future work for tagging.

tokenization, it is important to remember that there is no single optimal tokenization. What is optimal for Information Retrieval (IR) may not be true for Statistical Machine Translation (SMT). Also, what is optimal for a specific SMT implementation may not be the same for another [2].

This task accomplished by following:

2. Orthographic Normalization: It is a basic task that researchers working on Arabic NLP always apply with a common goal in mind: reducing noise in the data [2]. It can be unification of variants of letters, deleting Tatweel and the like.
3. Sentence segmentation: it is the process of splitting running text into sentences.
4. Word segmentation: it is the process of splitting sentences into words.
5. Getting tokens: it is the process of splitting words into morphemes.

We propose a hybrid method of unsupervised methods for Arabic tokenization considered as a stand-alone problem. After getting words from sentences by segmentation, we use the author's analyzer to produce all possible tokenizations for each word. Then, written rules and statistical methods are applied to solve the ambiguities. The output is one tokenization for each word. The statistical method was trained using 29k manually tokenized words (the raw text was taken from Al-Watan 2004 corpus). The final accuracy was 98.83%.

Lemmatization and Analyzing

Lemmatization and analyzing is the second preprocessing step of the whole tagging system which we propose. The output of this stage will be lemma, POS and features in case of nouns and verbs, meaning and working in case of particles. Each word may have more than one analysis.

Analyzing is the process of extracting all possible parts of speech and features for each word. But, Lemmatization is the process of extracting lemma from the inflected word where the lemma is the **canonical form**, **dictionary form**, or **citation form** of a set of words.

We suppose that the input to lemmatization and analyzing process is word without clitics (inflected word alone) or clitcs alone. Lemmatization and analyzing deal with known and unknown words in different ways:

1. Known words processing (found in the dictionary): no processing is needed because the lemma and features are in the dictionary.
2. Unknown words processing: Unknown words are processed by using rule-based approach. The rules depend on clitics, affixes, context, word pattern and word structure.

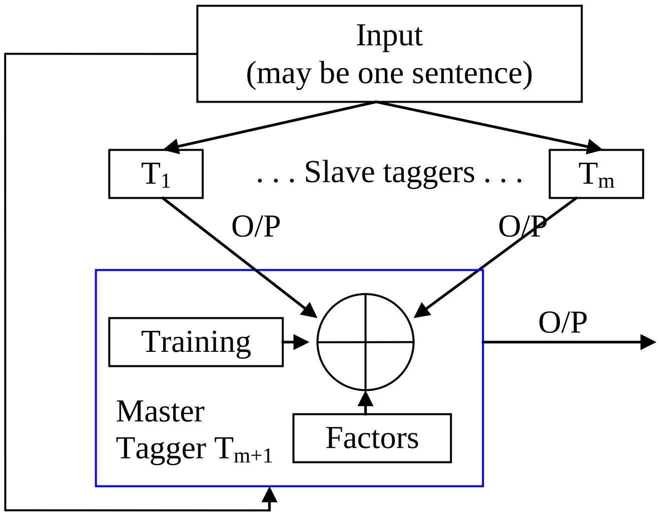
Unknown words are more likely to be nouns, because we use a large and fairly complete database of inflected verbs in the dictionary. As we know there are many classes of nouns which are closed sets (like e.g. relative nouns). The open classes of nouns are: proper, common, adjectives including genealogical and reduced nouns.

Tagging (disambiguation)

We implemented an efficient Arabic POS Tagger based on multitagger technique.

We propose a new method to combine taggers, which we call master-slave technique. In our approach, HMM tagger is used as the master tagger and Brill and MaxMatch (MM) taggers as slaves. The main property of our method is that the master tagger will process each sentence with different probabilities (different knowledge).

Any number of slave taggers can be used with one master. Assume that we have m+1 taggers (T_1 … T_{m+1}). T_{m+1} is HMM tagger and will be used as a master, the other will be used as slave taggers. The master tagger is trained for estimating its probabilities. Then the input sentence is tagged by each of the slave taggers T_1… T_m. The outputs of all slave taggers are fed to master in parallel for each sentence. Then the master changes its probabilities according to the outputs of the slaves for this sentence and a factor *f*. Then master does the tagging for this sentence according to the new probabilities. The important thing, in this method, is that using different probabilities for estimating each sentence. Figure 7-1 shows a block diagram for the proposed master-slaves tagger.



We have done two tests where HMM has been the master tagger. In the first test the Brill tagger has been the only slave, and in the second we have added MM as the second slave. The factor has been constant 0.29 for all tests. It was selected in a few other tests as the most effective one. The data set was Brown corpus which is freely available as a part of the NLTK package under Python environment. Also Brill tagger is a built-in tagger in Python. We built very simple implementations of MM and HMM taggers. We gained 0.26 % by using Brill tagger as the only slave and 0.42 % by using both Brill and MM as slaves. When annotating a corpus of 2 million words, it means correcting the tagging of about 8400 words. The other data set for Arabic consists of 45 files (29k words) annotated by hand with our new tagset. We gained 1.24 % by using Master-slaves.

The second tagging technique was “written rule-based tagging”. A few hundred of written rules were used in this tagger. Most of these rules were taken from Arabic traditional books. It was used for eliminating unwanted tags for specific words, and making the next tagger more accurate.

The two taggers were combined as stacking multi-taggers system where the output of the first tagger is fed to the second. The accuracy after adding the rule-based tagger increased from 90.05 to 92.86 %.