Toward An Efficient Arabic Part of Speech Tagger

Ahmed Abdelali
Qatar Computing Research Institute
QCRI. Doha, Qatar.
aabdelali@qf.org.qa

Yahya O. Mohamed Elhadj
Center for Arabic and Islamic
Computing, Al Imam Mohammad Ibn
Saud Islamic University.
Riyadh 11432, Kingdom of Saudi Arabia
yelhadj@ariscom.org

Rachid Bouziane
Arabic Language Department, College of
Arts & Science, Qatar University, Qatar,
rachid.bouziane@qu.edu.qa

***Abstract*—The task of tagging and allotting the correct Part of Speech (POS) to text given its context is not obvious and requires expertise and use of considerable resources. Automating such task and building tools that can carry such job is crucial and imperative to advance in major areas of natural language processing. A limited numbers of Part of Speech Taggers exist currently for Arabic and their availability is not trivial. In this paper we present an effort to design and build a POS tagger that would take into consideration the richness of the language as well as the efficiency in processing volumes of text. The Light Arabic Part of Speech Tagger (LAPOST) current output is very comparable to existing system but more effective from the processing perspective.**

Keywords—Natural Language Processing, Part of Speech, Tagging, Arabic Language, Linguistic Features, Syntax, Morphology.

I. PROPOSED ARABIC TAGSET

While the approach implemented in LAPOST balances between the usage of Arabic native traditional grammatical analysis and modern work; the proposed system is founded on new assumptions:

A. *Verb Tenses:* Traditionally, Arabic grammarians classify Arabic verbs into three categories (Perfect, Imperfect and Imperative). The Imperative category "الأمر", is not more than a mood for issuing orders rather than a tense; Therefore we propose to use the three categories commonly used by linguists: Past, Present, and Future.

B. *Demonstrative and Relative Pronouns*: These two types of pronouns are traditionally categorized as Nouns rather than pronouns; Even if pronouns are much related to nouns, we consider these are pronouns only.

According to the above assumptions, we can classify Arabic word into the following categories: *Noun (الأسماء)*, *Verbal-Noun (المصادر)*, *Verb (الأفعال)*, *Adjective (الصفات)*, *Particle (الحروف)*, *Adverb (الظروف)*. In addition to this categorization, we need to add another class for *Markers*(العلامات). As our objective is to provide comprehensive analysis for Arabic word, hence we add a number of features that would accompany any of the above categories.

The POS Tagger uses a number of morphological and syntactic features that are crucial for knowing the role of the word within the text.

II. LAPOST DESIGN

The system is fully coded in Java and the source code is very small (70kb of code with additional 60kb for additional files of affixes, special words, etc). It is platform independent; as the code can be compiled and run in any environment that has Java or JVM.

TABLE I. Output of sample text from ATP.

LAPOST	
ذهبت	ذهب/Vت/PRON_3FS|2MS
الطالبة	ال/Det+طالب/Nة/F
الصغيرة	ال/Det+صغير/Nة/F
الى	الى/PREP
المدرسة,	ال/Det+مدرس/Nة/F
ودرست	و/Conj+درس/Vت/PRON_3FS|2MS
الدروس	ال/Det+دروس/N
جميعها,	جميع/Nها/PRON_3FS
وحين	و/Conj+حين/ADV
جاء	جاء/V
وقت	وقت/N
الاختبار,	ال/Det+اختبار/N
نجحت	نجح/Vت/PRON_3FS|2MS
طالبتنا	طالب/Nت/PRON_3FS|2MS+نا/PL
بامتياز!	ب/Prep+امتياز/N

III. LAPOST EVALUATION

TABLE II. Tagging agreements between LAPOST and MADA

Category	**# of Tokens**	**Agreement**
All	5113	0.4836
Verbs	491	0.3380
Nouns	2672	0.6096
Others	1950	0.3651

TABLE III. Performance of LAPOST versus MADA+TOKAN[1]

N. of Words	**LAPOST**	**MADA+TOKAN**
46	0.5s	1m15.950s
3515	1.26s	2m4.783s
4331	1.33s	2m29.776s
5113	1.93s	2m32.612s

IV. CONCLUSION AND FEATURE WORK

LAPOST is a Part of Speech Tagger that was designed using linguistic features for Arabic. The system in the current state is comparable in terms of accuracy to extensive and large systems. The system still requires important updates which allow to produce more accurate and precise results.

[1] Results obtained using Mac Book Pro (OS X 10.6.8) 2.66GHz Intel Core 2 Due with 8GB 1333MHz memory.