

A Large-Scale Arabic POS Tagger Based on a Compact Arabic POS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words

Muhammad Atiyya^{1,2}, Mohsen A. A. Rashwan^{1,2}

¹Dept. of Electronics & Electrical Communications, Faculty of Engineering, Cairo University

Faculty of Engineering, Cairo Univ., Giza, Egypt

²The Engineering Company for the Development of Computer Systems; RDI.

171 Al-Haram main St., 6th floor, Giza, Egypt

{m_Atteya, mRashwan}@RDI-eg.com

Abstract

Part-Of-Speech (POS) tagging is an essential process for generating POS tags that convey the basic context-free syntactic features of surface text words. Besides many linguistic processing tasks for whom POS tagging is quite useful, POS tags are the most essential input features for all kinds of natural language computational syntax parsers which are in turn one step higher in the ladder towards language understanding and machine translation as well.

This paper describes a readily implemented industrial quality – i.e. large-scale – Arabic POS tagger, and more importantly introduces the underlying compact Arabic POS tags set along with the tagging structure whom are both proposed as standards for this essential process to Arabic NLP. Moreover, we show how the resulting POS tags sequences are statistically utilized to train and infer the syntactic diacritics as part of the process of automatic Arabic diacritization, which in turn is vital for many speech processing technologies esp. Arabic Text-To-Speech.

1.Introduction

Due to the highly derivative and inflective nature of the Arabic language, it is much more effective and economic to deal with its compact set of basic building entities; i.e. morphemes, than its unmanageably huge generable vocabulary.

Following the former morpheme-based approach and leaving the latter word-based one, we formalized our morphological canonical structure of any Arabic word w as a quadruple;

$$w \rightarrow \underline{q} = (t : p, r, f, s) \quad (1)$$

where t is Type Code (with possible types are *Regular Derivative*, *Irregular Derivative*, *Fixed*, *Arabized*), p is Prefix Code, r is Root Code, f is Pattern Code, and s is Suffix Code.

With a dynamic coverage ratio exceeding 99.5% coverage ratio; the knowledge base of our lexical analyzer *ArabMorpho*[®] based on this model are composed from only about 7,500 morphemes with their full agent-oriented linguistic description. (RDI's *ArabMorpho*[®], 2000), (Attia, M., 2000)

2.Compact Arabic POS Tags set

Composing an Arabic POS tags set necessitates scanning the lexico-syntactic features of each possible word of the Arabic vocabulary which is apparently infeasible. Instead, thanks for the morpheme-based approach, the features of each morpheme in the relatively compact *ArabMorpho*[®] knowledge base have been scanned, then digested through

several iterations of decimation into a non redundant compact Arabic POS tags set.

During that scanning process the following criteria has been adhered to:

- 1-All the *existing* lexico-syntactic features must be named and registered, which aims to the *completeness* of the resulting POS tags set.
- 2-All the named and registered features must be *atomic*, which aims to the *compactness* and *avoids redundancy* in the resulting tags set. This in turn is vital for the effectiveness of the based upon POS tagging process - which is essentially an abstraction process - and all higher processing layers as well. (Schutze & Manning, 2000), (Jurafsky & Martin, 2000)
- 3-All the named and registered features can be *ensured* upon the POS labeling of the morphemes in our Arabic lexical knowledge base. (More on this point in section 3 below)

Table 1 displayed below shows our Arabic POS tags set along with the meaning of each tag verbalized in both English and Arabic. Moreover, the 62 tags in the set are functionally categorized in order to maximize clarity.

While some tags in the following table may have corresponding one in other languages; e.g. English, others do not have such counterparts and are specific to the Arabic language.

Cat.	Mnemonic	Meaning in English	Meaning in Arabic
Start of word marker	SOW	Start-Of-Word marker	بداية كلمة
Padding string	Padding	Padding string	حشو
Features of noun and verb prefixes	NullPrefix	Null prefix	لا سابق
	Conj	Conjunctive	عطف
	Confirm	Confirmation by <i>Laam</i>	لام التوكيد
	Interrog	Interrogation by <i>Hamza</i>	همزة الاستفهام
Features of noun and verb suffixes	NullSuffix	Null suffix	لا لاحق
	ObjPossPro	Object or possession pronoun	ضمير نصب أو جر
Verb and noun syntactic cases	MARF	1 st Arabic syntactic case	مرفوع
	MANSS	2 nd Arabic syntactic case	منصوب
Features of noun-only prefixes	Definit	Definitive article	"ال" التعريف
Features of noun-only stems	Noun	Nominal	اسم
	NounInfinit	Nouns made of infinitives	مصدر
	NounInfinitLike	"NounInfinit" like	اسم مصدر
	SubjNoun	Subject noun	اسم فاعل
	ExaggAdj	Exaggeration adjective	صيغة مبالغة
	ObjNoun	Object noun	اسم مفعول
	TimeLocNoun	Noun of time or location	اسم زمان أو مكان
	NoSARF	<i>An Arabic feature of a specific class of nouns</i>	متنوع من الصرف
Features of noun-only suffixes	PossessPro	Possessive pronoun	ضمير جر
	RelAdj	Relative adjectives maker	نسب
	Femin	Feminine	تأنيث
	Masc	Masculine	مذكر
	Single	Singular	مفرد
	Binary	Binary	مثنى
	Plural	Plural	جمع
	Adjunct	Adjunct	مضاف
	NonAdjunct	NonAdjunct	غير مضاف
	MANSS_MAGR	2 nd or 3 rd Arabic syntactic case	منصوب أو مجرور
	MAGR	3 rd Arabic syntactic case	مجرور

Features of verb-only prefixes	Present	Present tense	مضارع
	Future	Future tense	استقبال
Features of verb-only stems	Active	Active sound	مبني للمعلوم (للفاعل)
	Passive	Passive sound	مبني للمجهول (للمفعول)
	Imperative	Imperative	أمر
	Verb	Verb	فعل
	Transitive	Transitive verb	لازم
	MAJZ	4 th Arabic syntactic case	مجزوم
	Past	Past tense	ماض
	PresImperat	Present tense, or imperative	مضارع أو أمر
Features of verb-only suffixes	SubjPro	Subject form pronoun	ضمير رفع
	ObjPro	Object form pronoun	ضمير نصب
	MANS_MAJZ	2 nd or 4 th Arabic syntactic case	منصوب أو مجزوم
Features of, mostly functional fixed words, and scarcely affixes	Prepos	Preposition	حرف جر
	Interj	Interjection	حرف نداء
	PrepPronComp	Preposition-Pronoun Compound	جازر ومجرور
	RelPro	Relative pronoun	اسم موصول
	DemoPro	Demonstrative pronoun	اسم إشارة
	InterrogArticle	Interrogation article	أداة استفهام
	JAAZIMA	For specific articles that make the consequent verb in the 4 th Arabic syntactic case	جازمة
	CondJAAZIMA	Feature of a class of Arabic conditionals	شرطية جازمة
	CondNotJAAZIMA	Feature of a class of Arabic conditionals	شرطية غير جازمة
	LAA	<i>Arabic specific article</i>	لا
	LAATA	<i>Arabic specific article</i>	لات
	Except	Article of exception	استثناء
	NoSyntaEffect	A class of articles that have no syntactic effect	غير عاملة
	DZARF	Feature for certain kind of Arabic adverbs	ظرف
	ParticleNAASIKH	A class of particles that make the subject of the consequent nominal sentence in 2 nd Arabic syntactic case	حرف ناسخ
	VerbNAASIKH	A class of auxiliary verbs that make the predicate of the consequent verbal sentence in 2 nd Arabic syntactic case	فعل ناسخ
	ParticleNAASSIB	Arabic specific class of particles that make the consequent verb in 2 nd Arabic syntactic case	ناصب
MASSDARIYYA	<i>Arabic specific article</i>	مصدرية	
For words beyond our morphological model	Translit	Transliterated Arabic string	كلمة أجنبية مكتوبة بحروف عربية

Table 1; Arabic POS tags set.

3. Arabic POS Labeling

Having the Arabic POS tags set been designed, *labeling* the morphemes of the lexical knowledge base comes as the next job which is a straight forward one given that the following three main points are carefully considered:

- 1-For morphologically analyzed words; the *f* part of the quadruples gives the Arabic POS tagging of stems, while the *p* and *s* parts give the Arabic POS tagging of affixes. Hence, the root morphemes of all kinds which do not participate to tagging are not Arabic POS labeled.
- 2-Due to the *atomicity* of the tags in the Arabic POS tags (see section 2 above) and in same time the compound nature of Arabic morphemes in general, POS labels of Arabic morphemes are vectors not simple scalars.
- 3-Only ensured Arabic POS tags are considered in the Arabic POS labeling of morphemes. i.e. When an Arabic POS tag is a possible - or even a highly probable – but not a strict feature of a given morpheme, it is not included in its Arabic POS label vector.

The following few morpheme labeling examples (from RDI's *ArabMorpho*[®]) are listed below in order to concretely illustrate the process:

Morpheme type and code	Morpheme shown as an Arabic string	Arabic POS vector label
Prefix; 9	الـ	[Definit]
Prefix; 125	سيـ	[Future,Present,Active]
Regular derivative pattern; 482	مُفَاعِلٍ	[Noun,SubjNoun]
Regular derivative pattern; 67	اسْتِفْعَالٍ	[Noun,NounInfinit]
Irregular derivative pattern; 29	مَلَائِكٍ	[Noun,NoSARF,Plural]
Fixed pattern; 8	هُوَ	[Noun,SubjPro]
Fixed pattern; 39	ذُو	[Noun,Masc,Single,Adjunct,MARF]
Suffix; 27	ـات	[Femin,Plural]
Suffix; 427	ـونَهُمْ	[Present,MARF,SubjPro,ObjPro]
Suffix; 195	ـيَّتَانِ	[RelAdj,Femin,Binary,NonAdjunct,MARF]

Table 2; Example Arabic POS labels from RDI's *ArabMorpho*[®] knowledge base.

4. Arabic POS Tagging

The Arabic POS tagging process are implemented in the following steps:

- 1-The Arabic strings sequence to be POS tagged are morphologically analyzed and combinatorially disambiguated using *ArabMorpho*[®]. (RDI's *ArabMorpho*[®]), (Atiyya, 2000), (Atiyya, et al, 2002) This results in a disambiguated quadruples sequence where each string is substituted by either one quadruple or a mark of Transliterated string.
- 2-For the prefix, pattern, and suffix morphemes of each quadruple in the sequence, the Arabic POS labels; $\underline{APOS}(p)$ $\underline{APOS}(t:f)$ $\underline{APOS}(s)$ are retrieved from the Arabic lexical knowledge base.
- 3-The Arabic POS tags vector of each word in the sequence is then composed using the formula:

$$\underline{APOS}(w) = \text{Concat}(\underline{APOS}(p), \underline{APOS}(t:f), \underline{APOS}(s)) \quad (2)$$

where the *Concat* function simply concatenates the POS sub vectors of the constituting morphemes after eliminating any mutual redundancy among their tags.

The resulting Arabic POS tags vectors by RDI's *ArabTagger*[®] of the words in a real-life phrase are shown in the table 3 below:

Phrase words		Most likely Arabic POS tags vectors
وقد	وَقَدْ	[SOW, Conj, NoSyntaDiac, NullSuffix]
صرحت	صَرَحْتُ	[SOW, NullPrefix, Verb, Past, Single, Femin]
رئيسة	رَئِيسَةٌ	[SOW, NullPrefix, Noun, ExaggAdj, ObjPossPro]
الوزراء	الْوَزَرَاءُ	[SOW, Definit, Noun, Plural, NoSARF, NullSuffix]
في	فِي	[SOW, NullPrefix, Prepos, NullSuffix]
نيوزيلندا	نِيوزِيلَنْدَا	[SOW, Translit]

Table 3; The resulting Arabic POS tagging of a real-life phrase using *ArabTagger*[®].

5. An application; Statistically Inferring the Syntactic Diacritics

Inferring the syntactic diacritics of words in Arabic text is an important complementary process for automatic Arabic text diacritization, which in turn is reflected on the elegance of the resulting phonetic transcription for Arabic speech processing systems esp. Arabic Text-To-Speech ones.

The traditional procedure for inferring the syntactic cases hence the syntactic diacritics of Arabic words in a given text is essentially as follows:

- 1-Run a POS tagger – like RDI's *ArabTagger*[®] - to get the Arabic POS tags of the Arabic words. (As we did in the 4 previous sections).
- 2-Fed these Arabic POS tags sequences to an Arabic syntactic parser.

- 3-Run that parser to on the POS tags sequences to get the *most likely* Arabic *parsing tree* corresponding to the input Arabic text.
- 4-The syntactic case – hence syntactic diacritic if any – of each word can then be easily inferred from back tracking the path of each terminal node up in the parsing tree.

While the development of an industrial quality Arabic syntactic parser is a very expensive mega project regarding the required time, workforce, and money, the existing ones that can be considered of industrial quality are extremely scarce, and are moreover inaccessible beyond their original developers. e.g. *Sakhr*'s; visit www.Sakhr.com

Consequently, we tried the following less expensive statistical procedure to directly infer the syntactic diacritics without the deployment of any Arabic syntactic parser:

I. Offline part of the procedure:

- 1-A large-enough corpus (see the next section of this paper for round figures) of supervised lexically analyzed and POS tagged Arabic text is afforded, along with the correct syntactic diacritics selected manually where necessary.
- 2-A statistical database of sequences; i.e. m-grams composed of “Arabic POS tags” and “syntactic diacritics”, is built and stored in a suitable format for efficient retrieval.

II. Online part of the procedure:

- 1-The input surface Arabic text w_1, w_2, \dots, w_n is morphologically analyzed and disambiguated using RDI's *ArabMorpho*[©] which produces the most likely quadruple of each input Arabic string q_1, q_2, \dots, q_n along with whether this analyzed word is missing a syntactic diacritic or not. In the latter case, the possible syntactic diacritics are also produced for each quadruple d_1, d_2, \dots, d_n .
- 2-Knowing the most likely quadruples at a high-enough accuracy ($> 95.5\%$; see the next section of this paper); RDI's *ArabTagger*[©] can easily extract the corresponding most likely POS tags vector of each word $APOS(w_1), APOS(w_2), \dots, APOS(w_n)$.
- 3-Compose the following trellis - as illustrated in figure 1 shown below – to statistically decide the most likely syntactic diacritics – where missing - of text words given their Arabic POS tags sequences.
- 4-The end-to-end path on the search lattice with maximum likelihood probability is then obtained online using the admissible and optimal A^* search algorithm (Nilsson, 1971), and using a combination of *Bayes'-Good-Turing discount-Back-off* techniques to estimate the probability of m-gram path segments from the sparse statistical database built in the offline part of the procedure (Atiyya et al, 2002), (Jurafsky & Martin, 2000), (Katz, 1987), (Nadas, 1985), (Schutze & Manning, 2000).
- 5-The syntactic diacritics on the obtained most likely path are considered the missing ones.

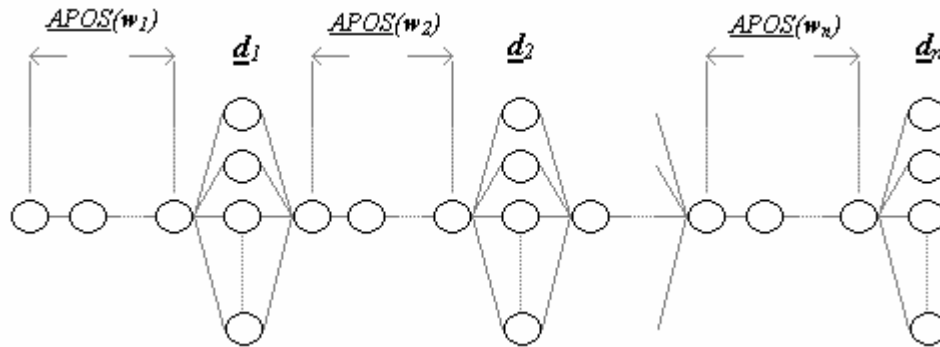


Figure 1; The Arabic POS tags – syntactic diacritics search trellis.

The main explanation of the promising results obtained after our experiments on news domain corpus with this procedure (see the next section of this paper) is that a majority of the syntactic cases – hence syntactic diacritics where missing - in the modern Arabic written text depend on shallow syntactic structures not the deep ones. These

relatively simple structures (like *prep-noun*, *verb-subject*, ..., etc.) are limited horizon ones composed of few words (2, 3, or 4 ones) whose “POS tags”-“syntactic diacritics” m-grams are statistically learnable from a corpus of adequate size, and using long enough maximum size of m-grams.

6. Performance evaluation

Two related systems have been presented in this paper, namely; Arabic POS tagger (RDI's *ArabTagger*[©], 2004) and Arabic syntactic diacritizer deployed in (RDI's

ArabDiac[©], 2004) built over the former one. Two major experiments have been conducted in RDI labs to evaluate the performance of both systems, and the results are briefed in the two tables presented below.

Size of training corpus	Max length of m-gram	Size of language model	Size of test sample	Accuracy of Lexical Disambiguation, hence Lexical Diacritization and POS Tagging
100,151 words; i.e. 400,604 morphemes, news-domain corpus which are lexically analyzed in <i>ArabMorpho</i> [©] format with manual supervision.	8 morphemes	20.4 M.Bytes	10,085 words; i.e. 40,340 morphemes, news-domain corpus.	92.3%
200,044 words; i.e. 800,176 morphemes, news-domain corpus which are lexically analyzed in <i>ArabMorpho</i> [©] format with manual supervision.	8 morphemes	33.5 M.Bytes	10,085 words; i.e. 40,340 morphemes, news-domain corpus.	95.5%
365,100 words; i.e. 1,460,400 morphemes, news-domain corpus which are lexically analyzed in <i>ArabMorpho</i> [©] format with manual supervision.	8 morphemes	45.1 M.Bytes	10,085 words; i.e. 40,340 morphemes, news-domain corpus.	96.8%

Table 2; The accuracy of Lexical Disambiguation, hence Lexical Diacritization, and POS Tagging estimated after 3 training and evaluation field experiments of *ArabMorpho*[©] and *ArabTagger*[©].

Size of training corpus	Max length of m-gram	Size of language model	Size of test sample	Accuracy of Syntactic Diacritics Inference	Accuracy of Full Word Diacritization
534,355 Arabic POS tags from 100,000 news-domain words corpus which are lexically analyzed in <i>ArabMorpho</i> [©] format and labeled with the correct syntactic diacritics, all with manual supervision.	16 Arabic POS tags and syntactic diacritics	25.2 M.Bytes	10,085 words total with only 6,130 words calls for syntactic diacritics.	79.8%	81.0%
1,080,140 Arabic POS tags from 200,000 news-domain words corpus which are lexically analyzed in <i>ArabMorpho</i> [©] format and labeled with the correct syntactic diacritics, all with manual supervision.	16 Arabic POS tags and syntactic diacritics	38.8 M.Bytes	10,085 words total with only 6,130 words calls for syntactic diacritics.	84.3%	86.4%

Table 3; The accuracy of inferring the syntactic diacritics estimated after 2 training and evaluation field experiments of *ArabDiac*[©] based on the output of the 1st two experiments of table no. 2.

The output of the 3rd experiment (lightly shaded) in table no. 2 has not been labeled completely yet (at the time of submitting this paper) with correct syntactic diacritics. Currently this labeling process is being carried out and

using extrapolation we *guess* this will enhance the accuracy of Syntactic Diacritics Inference up to (86.5%) at least which corresponds to an Accuracy of Full Word Diacritization of (89.61%).

References

- An online trial version of the mentioned system; RDI's *ArabDiac*[©] is found at: <http://www.RDI-eg.com> under the sub menu item Arabic NLP under the main menu item Technologies, (2004). (MS-Explorer[®] version 6 or later, and Arabic enabled MS-Windows[®] are needed)
- An online trial version of the mentioned system; RDI's *ArabMorpho*[©] is found at: <http://www.RDI-eg.com> under the sub menu item Arabic NLP under the main menu item Technologies, (2000). (MS-Explorer[®] version 6 or later, and Arabic enabled MS-Windows[®] are needed)
- An online trial version of the mentioned system; RDI's *ArabTagger*[©] is found at: <http://www.RDI-eg.com> under the sub menu item Arabic NLP under the main menu item Technologies, (2004). (MS-Explorer[®] version 6 or later, and Arabic enabled MS-Windows[®] are needed)
- An online trial version of an Arabic Text-To-Speech system; RDI's *ArabTalk*[©] which is based on the Arabic diacritizer mentioned in this paper; RDI's *ArabDiac*[©] is found at: <http://www.RDI-eg.com> under the sub menu item Speech under the main menu item Technologies, (2004). (MS-Explorer[®] version 6 or later, and Arabic enabled MS-Windows[®] are needed)
- Atiyya, M., (2000) A Large-Scale Computational Processor of The Arabic Morphology, and Applications, MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University. This thesis is also downloadable from the following web pages: <http://www.NEMLAR.org/ScientificPapers/Index.htm> , and <http://www.RDI-eg.com> under the menu sub item Arabic NLP under the main menu item Technologies.
- Attia, M., Rashwan, M., Khallaaf, G., (2002) *On Stochastic Models, Statistical Disambiguation, and Applications on Arabic NLP Problems*, The Proceedings of the 3rd Conference on Language Engineering; CLE'2002, the Egyptian Society of Language Engineering (ESLE). This paper is also downloadable from the following web pages; <http://www.NEMLAR.org/ScientificPapers/Index.htm> , and <http://www.RDI-eg.com> under the menu sub item *Arabic NLP* under the main menu item *Technologies*.
- Hifny, Y., Qurany, S., Hamid, S., Rashwan, M., Attia, M., Ragheb, A., Khallaaf, G., (2003) *ArabTalk[®]; An Implementation for Arabic Text To Speech System*, The proceedings of the 4th Conference on Language Engineering; CLE'2003, the Egyptian Society of Language Engineering (ESLE), and published also in the News Letter of Evaluation of Language Resources and Distribution Agency (ELDA) May 2004 issue. This thesis is downloadable from the web pages: <http://www.NEMLAR.org/ScientificPapers/Index.htm> , and <http://www.RDI-eg.com> under the menu sub item *Arabic NLP* under the main menu item *Technologies*.
- Jurafsky, D., Martin, J. H., (2000) *Speech and Language Processing; An Introduction to Natural Language Processing*, Computational Linguistics, and Speech Processing, Prentice Hall.
- Katz, S.M., (1987) Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-35 no. 3, March 1987.
- Nadas, A., (1985) *On Turing's Formula for Word Probabilities*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-33 no. 6, December 1985.
- Nilsson, N.J., (1971) *Problem Solving Methods in Artificial Intelligence*, McGraw-Hill.
- Schutze, H., Manning, C.D., (2000) *Foundations of Statistical Natural Language Processing*, the MIT Press.
- Sproat, R., (1998) *Multilingual Text-To-Speech Synthesis*, Kluwer Academic Publishers.
- Van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J., (1998) *Progress in Speech Synthesis*, Springer Publishers.