Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text

Majdi Sawalha and Eric Atwell

sawalha@comp.leeds.ac.uk eric@comp.leeds.ac.uk

UNIVERSITY OF LEEDS
School of Computing,
University of Leeds, Leeds, LS2 9JT, U.K.
http://www.comp.leeds.ac.uk/sawalha/

Introduction

Morphological analyzers and part-of-speech taggers are key technologies for most text analysis applications. Our aim is to develop a part-of-speech tagger for annotating a wide range of Arabic text formats, domains and genres including both vowelized and non-vowelized text. Enriching the text with linguistic analysis will maximize the potential for corpus re-use in a wide range of applications. We foresee the advantage of enriching the text with part-of-speech tags of very fine-grained grammatical distinctions, which reflect expert interest in syntax and morphology, but not specific needs of end-users, because end-user applications are not known in advance.

In This Research

- Review existing Arabic Part-of-Speech Taggers and tag-sets, and illustrate four different Arabic PoS tag-sets for a sample of Arabic text from the Quran.
- The detailed fine-grained morphological feature tag set of Arabic.
- The fine-grained Arabic morphological analyzer algorithm.
- Applying the morphological analyzer to the 100-million-word Web Arabic Corpus: we had to port the software to the National Grid Service.
- Adapt the analyser to cope with spelling variations and errors
- Utilize a Broad-Coverage Lexical Resource combining 23 traditional Arabic lexicons.
- Outline the construction of a Gold Standard for evaluation

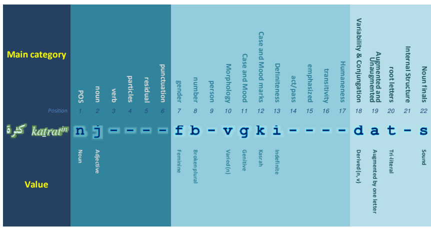
Motivation and Hypothesis

- Arabic is a highly inflectional language, and the traditional classification into nouns, verbs and particles is not enough detail.
- Arabic has many morphological and grammatical features, including sub-categories, person, number, gender, case, mood, etc. (Atwell, 2008).
- More fine-grained tag sets are often considered more appropriate. The additional information may also help to disambiguate the (base) part of speech (Schmid & Laws, 2008).

The Morphological Features Tag Set

- The Morphological Features Tag Set captures long-established traditional morphological features of Arabic, in a compact yet transparent notation.
- A tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature;
  - The dash "–" represents a feature not relevant to a given word,
  - The question mark "?" represents a feature relevant but its attribute value is not known yet.
- http://www.comp.leeds.ac.uk/sawalha/tagset.html

Word		Tag
تم	*tamma*	v-p------s-f-amihdstb-
اعداد	*'i'dādu*	ng----??-vndi---?db3-s
الوثائق	*al-waṯā'iqa*	nq----fb-vafd---ndbt-s
المتوفرة	*al-mutwafrat*	nj----f?-vafd---ndtt-s
ب	*bi*	p--p------------------
كثرة	*kaṯratin*	nj----fb-vgki----dat-s
حول	*ḥawla*	nv-------s-fi----nst-s
أول	*'awali*	n+----ms-vgki----dst-s
رحلة	*riḥlat*	no----fs-vgki----dat-s
طيران	*tayarānin*	ng----??-vgki----dbt-s
عثمانية	*'uṯmāniyyat*	n*----fs-pgki----daq-s
فوق	*fawqa*	nv-------s-fi----nst-s
البلاد	*al-bilādi*	nl----mb-vgkd---ndat-s
العربية	*al-'arabiyyh*	n*----fb-vgkd---hdst-s



Example : Four Arabic Tag Sets

Word	Tag Set	Morphemes		PoS Tag
وَوَصَّيْنَا *wa waaṣṣaynā* And We have enjoined	Morphological Features Tag Set	وَ	*wa*	p--c------------------ [Particle, Conjunction]
		وَصَّيْ	*waṣṣay*	v-p------s-s-amohdst&- [Verb, Perfect, Morphology/Structured, sukūn, Non-emphatic verb, Active voice, Transitive to one object, Human, Derived, Unaugmented, Tri-literal, Separated Lafif]
		نَا	*nā*	r---r-xpfs-f----hn---- [residual, Connected pronoun, Gender/Neuter, Sound plural, First person, Morphology/Structured, *fatḥah*, Human, Non-derived]
	Treebank FULL tag set	وَ	*wa*	wa/CONJ
		وَصَّيْ	*waṣṣay*	waS~ay/VERB_PERFECT
		نَا	*nā*	nA/PVSUFF_SUBJ:1P
	MorphoChallenge Gold Standard	وَ	*wa*	Particle +Conjunction
		وَصَّيْنَا	*waaṣṣaynā*	+Verb +Perf +Act +1P +Pl +Masc/Fem
	Quranic Arabic Corpus	وَ	*wa*	wa+
		وَصَّيْنَا	*waaṣṣaynā*	POS:V PERF (II) ROOT:wSy 1MP
الْإِنسَانَ *al-'insāna* (on) man	Morphological Features Tag Set	الْ	*al-*	r--d------------------ [Residual, Definite Article]
		إِنسَانَ	*'insāna*	nq----mb-pafd---hcbt-s [Noun, Noun of genus, Masculine, Broken plural, Morphology/ Prohibited from variation, Accusative/Subjunctive, fatḥah, Definite, Human, Inert/ Concrete noun, Augmented by two letters, Tri-literal, Noun finals/Sound]
	Treebank FULL tag set	الْ	*al-*	Al/DET
		إِنسَانَ	*'insāna*	+<inosAn/NOUN
	MorphoChallenge Gold Standard	الْإِنسَانَ	*al-'insāna*	+Noun +Triptotic +Sg +Masc +Acc +Def
	Quranic Arabic Corpus	الْ	*al-*	Al+
		إِنسَانَ	*'insāna*	POS:N LEX:<insa`n ROOT:Ans M ACC
بِوَالِدَيْهِ *bi- wālidayhi* His parents	Morphological Features Tag Set	بِ	*bi*	p--p------------------ [Particle, Preposition]
		وَالِدَ	*wālida*	nu----md-vgyd---hdat-s [Noun, Active participle, Masculine, Dual,Morphology/ Varied, Genitive, yā, Definite, Human, Derived, Augmented by one letter, Tri-literal, Noun finals/Sound]
		يْ	*y*	r--u------------------ [Residual, Dual letter]
		هِ	*hi*	r---r-msts-k----hn---- [Residual, Connected pronoun, Masculine, Singular, Third Person, Morphology/Structured, kasrah, Human, Non-derived]
	Treebank FULL tag set	بِ	*bi*	bi/PREP
		وَالِدَ	*wālid*	+wAlid/NOUN
		يْهِ	*ayhi*	+ayo/NSUFF_MASC_DU_ACCGEN+hu/POSS_PRON_3MS
	MorphoChallenge Gold Standard	بِ	*bi*	Prep
		وَالِدَيْهِ	*wālidayhi*	Noun +Triptotic +Dual +Masc +Obliquus +Pron +Dependent +3P +Sg +Masc
	Quranic Arabic Corpus	بِ	*bi*	bi+
		وَالِدَيْهِ	*wālidayhi*	POS:N LEX:wa`liday ROOT:wld MD GEN PRON:3MS
حُسْنًا *ḥusnan* Kindness	Morphological Features Tag Set	حُسْن	*ḥusnan*	ng----xs-vafi----ast-s [Noun, Gerund, Neuter, Singular, Morphology/Varied, Accusative, Fatḥah, Indefinite, Inert/ Abstract noun, Unaugmented, Tri-literal, Noun finals/Sound]
		ً	*an*	r--d------------------ [Residual, Tanween]
	Treebank FULL tag set	حُسْنًا	*ḥusnan*	Huson/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
	MorphoChallenge Gold Standard	حُسْنًا	*ḥusnan*	Noun +Triptotic +Sg +Masc +Acc +Tanwiin
	Quranic Arabic Corpus	حُسْنًا	*ḥusnan*	POS:N LEX:Huson ROOT:Hsn M INDEF ACC

Arabic Morphological Analyzer

Inputs: * **Single Arabic word** *or* * **Arabic text**
vowelized, **partially vowelized**, or **non-vowelized**

Step 1, Tokenization: the tokenizer program uses the **NLTK regular expression** tokenizer Tokenizes the input text into **Arabic words**, **punctuations, currency tokens**, **numbers**, **words written in Latin letters**, and **HTML/XML tags**.

Step 2, Clitics, Affixes and Stems : A special module divides the word into three parts; **proclitics and prefixes**, **stem/root**, *and* **suffixes and enclitics**. Proclitics, prefixes, suffixes and enclitics were collected from **traditional Arabic grammar books**.

Step 3, Root extraction : The second part of the divided word; **stem/root**, is searched in the **broad-coverage lexical resource**. Stem/root is searched in other three linguistic lists; **list of function words (stop words), named entities list** *and* **list of broken plurals.**

Step 4, Pattern Generation : The analyzer has a **list of patterns**, containing **2,730 verb patterns** and **985 noun patterns**. Morphological feature tags are assigned to each pattern in the lists. Two algorithms are used to generate the pattern; (1) Depends on **(word and root).** (2)Depends on **matching the word with one or more patterns** from the pattern lists.

Step 5, Vowelization : The analyzer adds the **short vowels** which appear on the **patterns** to the analyzed word, whether it is partially-vowelized or non-vowelized.

Step 6, Morphological features tag assignment : The analyzer assigns **a tag for each morpheme** of the word. Then the morphemes' tags are combined into **one whole word tag**. The word tag inherits its morphological feature attributes using an algorithm that establish agreements on morphological feature attributes of the word's morphemes.

Outputs : All possible analyses of the word such as, all possible **root, clitics, affixes, stems, lemmas, patterns, different forms of vowelization, and the morphological features** of each analysis represented by a **morphological tag** using the fine-grain morphological feature tag set.

Gold Standard for Evaluation

- **The Problem Domain:** (Evaluating morphological analyzers and part-of-speech taggers)
- **The Corpora:** (Corpora of different text domains, formats and genres of both vowelized and non-vowelized Arabic text. e.g. Qur'an text, and The Corpus of Contemporary Arabic)
- **Gold Standard Format:** (Includes morphological and part-of-speech information for each word of the gold standard in a line separated by tabs. Contains the root and the pattern information of the words. The gold standard will be stored using flat text files, using Unicode utf8 encoding or using XML)
- **Gold Standard Size:** (Relatively large. Covers most cases that morphological analyzers have to handle. Measured by the number of words it contains).

MorphoChallenge 2009
Qur'an Gold Standards

وَوَصَّيْنَا	وصي	يُفَعِّل	وَ +Particle +Conjunction وَصَّيْنَا +Verb +Perf +Act +1P +Pl +Masc/Fem
الْإِنسَانَ	ءنس	فِعلَان	ءِنسَان +Noun +Triptotic +Sg +Masc +Acc +Def
بِوَالِدَيْهِ	ولد	فَاعِل	بِ +Prep وَالِد +Noun +Triptotic +Dual +Masc +Obliquus +Pron +Dependent +3P +Sg +Masc
حُسْنًا	حسن	فُعل	حُسن +Noun +Triptotic +Sg +Masc +Acc +Tanwiin

wawaS~ayonaA	*wSy*	*yufaE~ilu*	wa +Particle +Conjunction waSSaynaA +Verb +Perf +Act +1P +Pl +Masc/Fem
Alo<insaAna	*'ns*	*fiElaAn*	'insaAn +Noun +Triptotic +Sg +Masc +Acc +Def
biwaAlidayohi	*wld*	*faAEil*	b +PrepwaAlid +Noun +Triptotic +Dual +Masc +Obliquus +Pron +Dependent +3P +Sg +Masc
HusonAF	*Hsn*	*fuEl*	Husn +Noun +Triptotic +Sg +Masc +Acc +Tanwiin

Practical Application: Lemmatizing the 100 million-word Arabic Web Corpus

http://smlc09.leeds.ac.uk/query-ar.html

- The lemmatizing part of the morphological analyzer (steps 1 to 3) has been applied to lemmatize a large and real data of the Arabic Web Corpus consisting of 100 million words collected from the web.
- Processing large corpus of is unrestricted text shows other **specific challenges**.

long execution time

We used powerful high performance computing facilities to solve this challenge.
The NGS (National Grid Services *http://www.ngs.ac.uk*) which aims to enable coherent electronic access for UK researchers to all computational and data based resources and facilities required to carry out their research, independent of resource or researcher location.

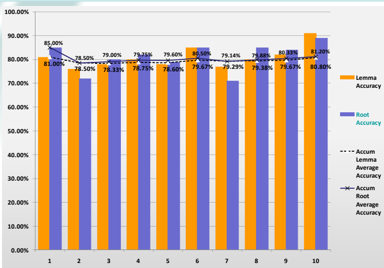
Spelling Errors Detection and Correction

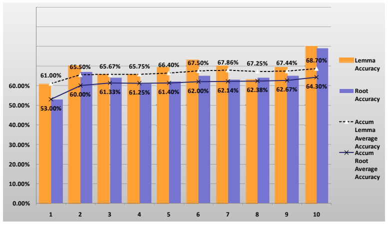
Errors such as:
- Typing more than one short vowel for a letter.
- ***tanwīn*** (ً , ٌ , ٍ) appears on any letter except the last letter is considered as spelling mistake.
- Words starting or ending with ***taṭwīl*** character (-) which is used to extend the length of the word;
- ***šaddah*** (ّ) appears on the first letter of the word or on the letter after the definite particle ***alif-lām*** (ال) needs special processing.

EVALUATION

Evaluation of lemmatized Arabic Web Corpus

- Evaluation was done by selecting 10 100-word random samples from the lemmatized ***Arabic Web Corpus***.
- For Each sample root accuracy and lemma accuracy were computed.
- The accumulative average lemma accuracy is 80.80%.
- The accumulative average root accuracy is 81.20%.



Evaluation of lemmatized Qur'an

- Evaluation was done by selecting 10 100-word random samples from the lemmatized ***Quranic Arabic Corpus***.
- For Each sample root accuracy and lemma accuracy were computed.
- The accumulative average lemma accuracy is 68.70%.
- The accumulative average root accuracy is 64.30%.