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

Majdi Sawalha, Eric Atwell
School of Computing,
University of Leeds, Leeds, LS2 9JT, UK
E-mail: sawalha@comp.leeds.ac.uk , eric@comp.leeds.ac.uk

Abstract

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 paper we 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. We describe the detailed fine-grained morphological feature tag set of Arabic, and the fine-grained Arabic morphological analyzer algorithm. We faced practical challenges in 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, and utilise a Broad-Coverage Lexical Resource combining 23 traditional Arabic lexicons. Finally we outline the construction of a Gold Standard for comparative evaluation.

1. Introduction

Morphological analyzers and part-of-speech taggers are essential technologies for most text analysis applications. The most obvious applications are in lexicography and NLP/computational linguistics. Further applications include using the tags in data compression; more fine-grain information about the compressed text will help in more compression for the data (Teahan, 1998); and as a possible guide in the search for extra-terrestrial intelligence (Elliott & Atwell, 2000). Other specific uses that make use of part-of-speech tag information are: searching and concordancing, grammatical error detection in Word Processing, training Neural Networks for grammatical analysis of text, or training statistical language processing models (Atwell, 2008). Part-of-Speech tagging is a key technology in discovering suspicious events from text, and processing Arabic is a key task to discover these suspicious events.

Many automatic taggers have been made. TAGGIT, achieved an accuracy of 77% tested on the Brown corpus. CLAWS1, data-driven statistical tagger had scored an accuracy rate of 96-97%. Hidden Markov Model (HMM) taggers have been made for several languages. Brill's tagger (1995) is an example of data-driven symbolic tagger. The ENGCG and EngCG-2 are based on a Constraint Grammar (CG) framework. Currently, many new systems based on Markov Model and Machine Learning (ML) techniques appeared for many languages. Hybrid solutions have been investigated (Voulainin, 2003). ACOPOST [1], A Collection Of POS Taggers, consists of four taggers of different frameworks; Maximum Entropy Tagger (MET), Trigram Tagger (T3), Error-driven Transformation-based Tagger (TBT) and Example-based tagger (ET). A SNoW-based Part of Speech Tagger[2] makes use of the Sequential Model. LBJ Part of Speech Tagger [3] is substantially the same as SNoW-based POS tagger, except that it accepts raw, natural language text as input. NLTK[4] - Natural Language Toolkit has implemented many POS tagger such as; Regexp Tagger, N-Gram Tagger, Brill Tagger and HMM Tagger, in addition to some documentations on tagging. RelEx [5] - provides English-language part-of-speech tagging, entity tagging, as well as other types of tags (gender, date, money ...). Spejd[6] - Shallow Parsing and Disambiguation Engine a GPL tool for simultaneous rule-based morphosyntactic disambiguation and partial parsing. VISL Constraint Grammar [7] rule based disambiguation (GPL).

1.2 Arabic Part-of-Speech Taggers

Arabic part-of-speech tagging development started in describing some of the initial findings of the development of Arabic part-of-speech taggers; the APT tagger (Khoja, 2001). Hurdles of the development of Arabic part-of-speech taggers described when developing Brill's "transformation-based" or "rule-based" part-of-speech tagger for Arabic (Freeman, 2001). An early stage of the architecture of a web-based Arabic tagger has been developed (Harmain, 2004). Support Vector Machines (SVM) a supervised learning algorithm achieved an accuracy of 95.49% (Diab et al, 2004). Another SVM-based Yamcha; which uses Viterbi decoding, was developed by (Habash & Rambow, 2005). HMM part-of-speech tagging for Arabic achieved accuracy of 97% for modern standard Arabic (Al-Shamsi & Guessoum, 2006) and 69.83% when tested on Egyptian Colloquial Arabic and the Levantine Arabic (Duh & Kirchhoff,

[1] http://acopost.sourceforge.net/
[2] http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=POS
[3] http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=FLBJPOS
[4] http://www.nltk.org/
[5] http://opencog.org/wiki/RelEx
[6] http://nlp.ipipan.waw.pl/Spejd/
[7] http://beta.visl.sdu.dk/cg3.html

2005). Applications of Memory-based learning to morphological analysis and part-of-speech tagging of written Arabic have been explored (Masri et al, 2005). Besides, combinations of rule based and machine learning methods for tagging Arabic words (Tlili-Guiassa, 2006). The multi-agent architecture was adopted for the conception and the realization of the part-of-speech tagging system of Arabic text with vowel marks (Zribi et al, 2006). Arabic Morphosyntactic Tagger AMT, uses the pattern-based technique and lexical and contextual technique. The accuracy of the AMT tagger reported was 91% (Alqrainy, 2008).

2. 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).

Very fine-grain distinctions may cause problems for automatic tagging if some words can change grammatical tag depending of function and context (Atwell, 2008); on the other hand, fine-grained distinctions may actually help to disambiguate other words in the local context. A practical experiment of using fine grain morphological tag set was reported by Schmid and Laws (2008). Their experiments were carried out using German and Czech as examples of highly inflectional languages. Their HMM part-of-speech tagger makes good use of the fine-grain tag set; it splits the part-of-speech into attribute vectors and estimates the conditional probabilities of the attribute with decision trees. This method achieved a higher tagging accuracy than two state-of-the-art general-purpose part-of-speech taggers (TnT and SVMTool). We believe that this kind of approach may yield better results for an Arabic part-of-speech tagging including fine-grained morphological features.

3. The Morphological Features Tag Set

The Morphological Features Tag Set captures long-established traditional morphological features of Arabic, in a compact yet transparent notation. Each feature and possible values of the Morphological Features Tag Set were explained and illustrated in detail. 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, and the question mark "?" represents a feature relevant but its attribute value is not known yet. The first character shows the main Parts of Speech, from: noun, verb, particle, punctuation, and residual; these last two are an extension to the traditional three classes to handle modern texts. The characters 2, 3, and 4 are used to represent subcategories; of noun, verb and particle. Residuals and punctuations are represented in letters 5 and 6 respectively. The next letters represent traditional morphological features: gender (7), number (8), person (9), morphology (10) case or mood (11), case and mood markers (12), definiteness (13), voice (14), emphasize (15), transitivity (16), humanness (17), Variability and Conjugation (18). Finally there are four characters representing morphological information which is useful in Arabic text analysis, although not all linguists would count these as traditional features: augmented and unaugmented (19), number of root letters (20), verb internal structure (21), noun finals (22). The Morphological Features Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

Tag-assignment is significantly more complex for Arabic. An Arabic lemmatiser program can extract the stem or root, but this is not enough for full PoS-tagging; words should be decomposed into five parts: proclitics, prefixes, stem or root, suffixes and enclitics. The morphological analyzer should then add the appropriate linguistic information to each of these parts of the word; in effect, we need a sub tag for each part (and possibly multiple sub tags if there are multiple proclitics, prefixes, suffixes or enclitics). The word tag inherits its morphological feature attributes using an algorithm that establish agreements on morphological feature attributes of the word's morphemes and combining morpheme tags into word tag.

Figures 1 and 2 show a sample of tagged text using the morphological feature tag set taken from the Arabic Treebank تم اعداد الوثائق المتوفرة بكثرة حول أول رحلة طيران عثمانية فوق البلاد العربية *tamma 'i'dād al-waṯāiqa al-mutwafra^{t} bi- kaṯrat^{in} ḥawla 'awali riḥla^{t} tayarān^{in} 'uṯmāniyya^{t} fawqa al-bilādi al-'arabiyy^{h}* 'Many available documents relate to the first Ottoman's flight over the Arab countries', and from the Qur'an the sentence is وَوَصَّيْنَا الْإِنسَانَ بِوَالِدَيْهِ حُسْنًا *wa waṣṣaynā al-'insāna biwālidayhi ḥusnan* 'We have enjoined on man kindness to parents'.

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-mutwafra^{t}*	nj----f?-vafd---ndtt-s
ب	*bi*	p--p----------------
كثرة	*kaṯrat^{in}*	nj----fb-vgki----dat-s
حول	*ḥawla*	nv-------s-fi----nst-s
أول	*'awali*	n+----ms-vgki----dst-s
رحلة	*riḥla^{t}*	no----fs-vgki----dat-s
طيران	*tayarān^{in}*	ng----??-vgki----dbt-s
عثمانية	*'uṯmāniyya^{t}*	n*----fs-pgki----daq-s
فوق	*fawqa*	nv-------s-fi----nst-s
البلاد	*al-bilādi*	nl----mb-vgkd---ndat-s
العربية	*al-'arabiyy^{h}*	n*----fb-vgkd---hdst-s

Figure 1: Sample of Tagged document of non- vowelized newspaper text using the Morphological feature tag set

Word		Tag
وَ	*wa*	p--c----------
وَصَّيْ	*waṣṣay*	v-p------s-s-amohdst&-
نَا	*nā*	r---r-xpfs-f----hn----
الْ	*al-*	r--d----------
إِنسَانَ	*'insāna*	nq----mb-pafd---hcbt-s
بِ	*bi*	p--p----------
وَالِدَ	*wālida*	nu----md-dgyd---hdat-s
يْ	*y*	r--u----------
هِ	*hi*	r---r-msts-k----hn----
حُسْن	*ḥusn*	ng----xs-vafi----ast-s
اً	*an*	r--d----------

Figure 2: Sample of Tagged document of vowelized Qur'an Text using the Morphological feature tag set

Figures 3 shows sample sentence وَوَصَّيْنَا الْإِنسَانَ بِوَالِدَيْهِ حُسْنًا *wa waṣṣaynā al-'insāna biwālidayhi ḥusnan,* taken from the Qur'an, and tagged using different tag sets. These tagging schemes are; our detailed and fine-grain morphological tag set, the Penn Treebank FULL tag set, the morphochallenge2009 Qur'an gold standard tagging scheme, and the Quranic Arabic Corpus tagging scheme.

In figure 3, the sentence is tagged using our detailed and fine-grained morphological features tag set. It also shows the sentence tagged using Tim Buckwalter morphological analyser and the Penn Arabic Treebank FULL tag set. The Penn Arabic Treebank tag set is the most widely used tag set for Arabic. It is used to annotate the Penn Arabic Treebank (PATB) with part-of-speech tags. The Penn Arabic Treebank model postulates a FULL tag set which compromises over 2000 tag types (Diab, 2007). The FULL tag set includes combinations of 114 basic tags. The FULL tag set exhibits the morphological features; case, gender, number, definiteness, mood, person, voice, tense, aspect and other features. Figure 3 also shows the sentence tagged using morphochallenge2009 Qur'an gold standard morphological tagging scheme which has developed using the data of Morphological Tagging of the Qur'an database (Talmon & Wintner, 2003; Dror et al, 2004). The word analysis is shown after each word where the morphological features are separated by space and "+" sign. These features include the part-of-speech of the word, number, gender, person, case, definiteness, voice and others. The figure shows sentence tagged using the Quranic Arabic Corpus (http://corpus.quran.com/) morphological tagging scheme. The tagged example shows that the words are divided into three parts; prefixes, stem, and suffixes. Each part is assigned morphological features such as, part-of-speech, person, number, gender, definiteness and mood. The construction of the Quranic Arabic Corpus is mainly depends on the Tim Buckwalter morphological analyser and Morphological Tagging of the Qur'an database (Talmon & Wintner, 2003; Dror et al, 2004). The tag set used combines the morphological features of the words derived from Buckwalter morphological analyzer and the tagged database of the Qur'an into new tag set.

The tags of the Full Arabic Treebank, MorphoChallenge 2009 Qur'an gold standard and the Quranic Arabic Corpus, are word tags. The "+" sign is used to separate the tag information of the word's morphemes. In figure 3, we showed the morpheme tags instead of the whole word's tag, we manually separated the morpheme tags as shown in the figure. On the other hand, our morphological features tag set, gives each word's morpheme a specified tag. These tags can be combined into one tag for the whole word using the same tag structure.

By looking to the different tagging schemes and tag sets, we conclude that the morphological features tag set is more detailed; consisting of 22 morphological features of the word, which captures most of the linguistic information of the word and its clitics and affixes. Moreover, the morphological features tag set has a fixed structure where the tag consist 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. This structure makes the morphological features tag set to be more readable than the existing tag sets for Arabic.

4. Arabic Morphological Analyzer

Many challenges face the implementation of Arabic morphology, the rich "root-and-pattern" nonconcatenative (or nonlinear) morphology and the highly complex word formation process of root and patterns, especially if one or two long vowels are part of the root letters. Moreover, the orthographic issues of Arabic such as short vowels {(ً) (ٌ), (ٍ)}, *hamza*^{h} (ئ ؤ أ إ ء), *tā' marbūṭa*^{h} (ة) and *hā'* (ه), *yā'* (ي) and *alif maqṣūra*^{h}(ى) , *šadda*^{h} (ّ) or gemination, and *madda*^{h} (آ) or extension which is a compound letter of *hamza*^{h} and *alif* (ا ء).

Our morphological analyzer uses linguistic knowledge of the language as well as corpora to verify the linguistic information. It uses a broad-coverage lexicon constructed by analyzing 23 established Arabic language dictionaries. The broad-coverage lexicon contains correct vowelized words and roots. It will be developed to contain multi-word expressions, idioms, collocations requiring special part-of-speech assignment, and words with special part-of-speech tags. The morphological analyzer depends on comprehensive lists of affixes, clitics and patterns extracted from authoritative Arabic grammar books. These lists were then cross-checked by analyzing words of three corpora: the Qur'an, the Corpus of Contemporary Arabic and Penn Arabic Treebank (as well as our lexicon, considered as a fourth cross-check corpus). The morphological analyzer uses novel algorithms that generate the correct pattern of the words, deal with the orthographic issues of the Arabic language and other word derivation issues, such as the elimination or substitution of root letters, tokenize the word into proclitics, prefixes, stem or root, suffixes and enclitics, generate all possible vowelizations of the processed word, and assign morphological features tags for the word's morphemes (Sawalha & Atwell, 2009a; Sawalha & Atwell, 2009b).

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, *fatḥa*^{h}, 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, *fatḥa*^{h}, 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, *kasra*^{h}, Human, Non-derived]
	Treebank FULL tag set	ب	*bi*	**bi/PREP**
		وَالِدَ	*wālid*	**+wAlid/NOUN**
		يْهِ	*ayhi*	**+ayo/NSUFF_MASC_DU_ACCGEN+hu/POSS_PRON_3MS**
	MorphoChallenge Gold Standard	ب	*bi*	**Prep**
		وَالِدَيْهِ	*wālidayhi*	**Noun +Triptotic +Dual +Masc +Obliquus +Pron +Dependent +3P +Sg +Masc**
	Quranic Arabic Corpus	ب	*bi*	**bi+**
		وَالِدَيْهِ	*wālidayhi*	**POS:N LEX:wa`liday ROOT:wld MD GEN PRON:3MS**
حُسْنًا *ḥusn*^{an} Kindness	Morphological Features Tag Set	حُسْنَ	*ḥusn*^{an}	**ng----xs-vafi----ast-s**
				[Noun, Gerund, Neuter, Singular, Morphology/Varied, Accusative, Fatḥah, Indefinite, Inert/ Abstract noun, Unaugmented, Tri-literal, Noun finals/Sound]
		ًا	*an*	**r--d----------**
				[Residual, Tanween]
	Treebank FULL tag set	حُسْنًا	*ḥusn*^{an}	**Huson/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF**
	MorphoChallenge Gold Standard	حُسْنًا	*ḥusn*^{an}	**Noun +Triptotic +Sg +Masc +Acc +Tanwiin**
	Quranic Arabic Corpus	حُسْنًا	*ḥusn*^{an}	**POS:N LEX:Huson ROOT:Hsn M INDEF ACC**

Figure 3: Cross-tagset comparison of a sample sentence taken from the Qur'an

4.1 Arabic Morphological Analyzer Algorithm

The morphological analysis of Arabic text involves many processing steps. These steps, described below, are executed sequentially where each step depends on the previous step. Intermediate results can be obtained from each processing step.

Inputs: the morphological analyzer accepts single Arabic word or Arabic text, whether they are vowelized, partially vowelized, or non-vowelized, as inputs to the system.

Outputs: the outputs of morphological analyzer are the full analyses of the words from the analyzed text. The full analyses means 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.

Step 1, Tokenization: the tokenizer program uses the NLTK regular expression tokenizer to tokenize the input text into Arabic words, punctuations, currency tokens, numbers, words written in Latin letters, and HTML/XML tags. The regular expression tokenizer uses regular expression patterns that suite the Arabic text.

Step 2, Clitics, Affixes and Stems: for each tokenized Arabic word, a special module divides the word into three parts; proclitics and prefixes, stem/root, and suffixes and enclitics. The first part is searched in a list of proclitics and prefixes consisting of 220 entries, and the third part is searched with a list of suffixes and enclitics consisting of 341 entries. Only the analyses that match both of the lists of clitics and affixes are taken as candidate analysis.

Step 3, Root extraction: for each candidate analyses from step 2, the second part of the divided word; stem/root, is searched in the broad-coverage lexical resource. If the non-vowelized stem/root is found in the lexicon then all vowelized words-root combinations are retrieved and attached to that analysis, which assumed as candidate analysis.

In this step each stem/root is searched in other three linguistic lists; list of function words (stop words) which have fixed syntactic analysis in any context (Diwan, 2004), named entities list (Benajiba et al, 2008) and list of broken plurals[8]. If the stem/root of any analysis matches one of these lists, then a new analysis entry along with its morphological analysis is added to the candidate analyses list of the word.

Step 4, Pattern Generation: we provided the analyzer with a list of patterns, containing 2,730 verb patterns and 985 noun patterns. Morphological feature tags are assigned to each pattern in the lists. The pattern generation module uses two algorithms; the first depends of the word itself and the extracted root from step 3, while the second algorithm depends on matching the word with one or more patterns from the pattern lists.

For each candidate analysis of step 3, both pattern generation algorithms are applied. Then the generated patterns are searched in the pattern lists, and the fully vowelized pattern and its morphological features tag are assigned to that analysis.

Step 5, Vowelization: After matching the patterns and the analyzed word, in the previous step, taking into account that the patterns are fully vowelized, the analyzer adds the short vowels which appear on the patterns to the analyzed word, whether it is partially-vowelized or non-vowelized. The result is a correctly fully vowelized list of the possible analyses.

Step 6, Morphological features tag assignment: most Arabic words are complex words consisting of multiple morphemes. For example the morphological analyzer will specify the morphemes of the word وسيكتبونها *wasayaktubūnahā* 'and they will write it' as follows: preclitic و *wa* 'and' (conjunction), prefixes س *sa* 'will' and ي *ya* (progress letter), the stem كتب *kataba* 'write', the suffix ون *ūn* 'they' and the enclitic ها *hā* 'it' (relative pronoun). The word consists of 6 morphemes. Each morpheme carries morphological features and belongs to a specific part of speech category. Our morphological analyzer assigns a tag for each morpheme of the word; given that the linguistic lists used by the morphological analyzer are all have the morphological feature tags assigned to each entry in these lists. 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.

4.2 Practical Application: Lemmatizing the 100 million-word Arabic Web Corpus[9]

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.

4.2.1 Challenges of Lemmatizing the Arabic Web Corpus

One challenge is long execution time. We used powerful high performance computing facilities to solve this challenge. The NGS[10] (National Grid Services) 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. We divided the Arabic Web Corpus into several 1-million words files. Then we wrote a program that generates a specified script that is required to run the lemmatizer program for each file in parallel. Then the output files are combined in one lemmatized Arabic Web Corpus.

Another challenge appeared while processing the Arabic Web Corpus is the spelling errors because of typing mistakes. Such errors are; typing more than one short

[8] http://sites.google.com/site/elghamryk/arabiclanguageresources

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

[10] http://www.ngs.ac.uk/

vowel for a letter. *tanwīn* (ً, ٌ, ٍ) appears on any letter except the last letter is considered as spelling mistake. And words starting or ending with *taṭwīl* character (-) which is used to extend the length of the word; such as the characters between the letters ت *t* and ا *ā* كتـــــــاب. More over, *šaddah* appears on the first letter of the word or on the letter after the definite particle *alif-lām* (ال) needs special processing.

To solve these spelling mistakes and orthographic issues of Arabic word, we developed a special algorithm that verify the correct letter-diacritics combinations and corrects the typing errors found. The algorithm is described below.

4.2.2 Spelling Errors Detection and Correction

The algorithm divides the Arabic word into three parts; front part consisting of the first letter and any diacritics appeared on it, the middle part consisting of the letters starting from the second letter till the letter before the last and their diacritics, and the rare part consists of the last letter and its diacritics. Each part has its own valid letter-diacritics combinations.

The front part is checked if it matches the following letter-diacritic valid combinations {(*letter* + *šaddah* + *a short vowel*[11]*), (letter* + *a short vowel), (letter)}*. The *šaddah* appears on the first letter does not indicate that the letter is a doubled letter, it orthographic mark to indicate incorporation with the last letter of the previous word, and it appears in the text of Qur'an. *Sukūn* and *tanwīn* do not appear on the first letter of the word. Also, *taṭwīl* can not be the first letter of the word.
Each letter-diacritic combination from the middle part is checked if it matches the following letter-diacritic valid combinations; {(*letter* + *šaddah* + *a short vowel), (letter* + *a short vowel), (letter* + *sukūn), (letter), (taṭwīl)}*. Any diacritic appear on *taṭwīl* is a spelling mistake. The *šaddah* appears on any letter of the middle part indicates a doubled letter or orthographic mark if the letter belongs to the solar letters and proceeded by the definite article *alif-lām* (ال).

The rare part is checked if it is matches one of the following letter-diacritic valid combinations {(*letter* + *šaddah* + *a short vowel), (letter* + *šaddah* + *tanwīn), (letter* + *a short vowel), (letter* + *sukūn), (letter* + *tanwīn), (letter)}*. The *šaddah* appears on the last letter of the word indicates a doubled letter. And *taṭwīl* can not be the last letter of the word.

5. Broad-Coverage Lexical Resource

Broad-coverage language resources which provide prior linguistic knowledge must improve the accuracy and the performance of automatic language processing applications. We have constructed a broad-coverage lexical resource to improve the accuracy of morphological analyzers and part-of-speech taggers of Arabic text. Twenty three traditional Arabic lexicons have been collected from different resources from the web where all of them are freely available. *maktaba*^{t} *al-mišk*$\bar{a}^{t}$ *al-'islāmyya*^{h}[12] مكتبة المشكاة الاسلامية provides most of these lexicons which are written in MS-Word files. Each lexicon is written in a different format and has its own arrangement methodology of its lexical entries. After manually converting each lexicon text into a unified format by choosing the most common format for all the root entries in the lexicon, information such as roots, words and meaning are automatically extracted using specialized programs. The results are stored in separate dictionaries which include roots, words, and meanings. A combination algorithm combines the disparate lexicon information into one large broad-coverage lexical resource (Sawalha & Atwell, 2010).

The coverage of the constructed broad-coverage lexical resource showed that about 85% of the words processed using the lemmatizer referenced the broad-coverage lexicon and retrieved correct analyses of the analyzed words. The coverage has been computed using three text samples; the Qur'an, the Corpus of Contemporary Arabic and 1 million words from the Arabic Web Corpus (Sawalha & Atwell, 2010).

6. Gold Standard for Evaluation

Gold standards are used to evaluate and measure the actual accuracy of the morphological analyzer and the part-of-speech tagger. To build a widely used general purpose gold standard, we have to select corpora of different text domains, formats and genres of both vowelized and non-vowelized Arabic text. First, we selected the Qur'an corpus to be used in the construction of the gold standard. We have two versions of the Qur'an text, vowelized Qur'an text, where diacritics appear above or below each letter of Qur'an text, and a non-vowelized one, where diacritics are omitted from the vowelized text of Qur'an. Second, we want to use the Corpus of Contemporary Arabic (Al-Sulaiti & Atwell, 2006). This corpus contains 1 million words taken from different genres collected from newspapers and magazines.

The gold standard will include morphological and part-of-speech information for each word. The analysis divides the words into their morphemes; conjunctions, prepositions, prefixes, stem or root, suffixes and relative pronouns. For each morpheme, part-of-speech tag is assigned, as well as a compound part-of-speech tag of the whole word. Moreover, the gold standard will contain the root and the pattern information of the words. The gold standard will be stored using flat text files, where each word and its morphological and part-of-speech information in a line separated by tabs, and using XML technology to store the gold standard in structured format.

We developed a gold standard of the Qur'an to be used to evaluate morphological analyzers in the Morphochallenge 2009 competition, which aims to develop unsupervised morphological analyzers to be used for different languages including Arabic, see MorphoChallenge site: http://www.cis.hut.fi/morphochallenge2009/datasets.shtml . The gold standard size is 78,004 words. The gold standard of Qur'an contains the full morphological analysis for

[11] Short vowels are *fatḥa*^{h}, *ḍamma*^{h} and *kasra*^{h} {(◌َ) (◌ُ), (◌ِ)}

[12] http://www.almeshkat.net

each word, according to the Morphological Tagging of the Qur'an database (Talmon & Wintner, 2003; Dror et al, 2004) but reformatted to match other Morphochallenge test sets in other languages. Figure 4 shows a sample of a tagged sentence taken from morphochallenge 2009 Qur'an gold standard. Arabic script used in the first part of the figure and Romanized script using Tim Buckwalter transliteration scheme used on the second part of the figure.

وَوَصَّيۡنَا	وصي	يُفَعِّلُ	وَ +Particle +Conjunction وَصصَينَا +Verb +Perf +Act +1P +Pl +Masc/Fem
الۡإِنسَانَ	ءنس	فِعلَان	ءِنسَان +Noun +Triptotic +Sg +Masc +Acc +Def
بِوَالِدَيۡهِ	ولد	فَاعِل	ب +Prep وَالِد +Noun +Triptotic +Dual +Masc +Obliquus +Pron +Dependent +3P +Sg +Masc
حُسۡنًا	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

Figure 4: A sample of tagged sentence taken from the MorphoChallenge 2009 Qur'an Gold Standard, the first part uses Arabic script and the second one uses romanized letters using Tim Buckwalter transliteration scheme.

7. Conclusions

Morphological analyses and part of speech (PoS) tagging are very important and basic applications of Natural Language Processing. In this paper we highlighted the importance of morphological analyses and part of speech tagging in wide range of NLP applications. We listed the most recent part of speech taggers and the key technologies used so far for part of speech taggers for Arabic text.

Our hypothesis is based of the fact that 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).

For the implementation of a fine-grain morphological analyzer and part of speech tagger for Arabic text, we designed a detailed morphological feature tag set that captures long-established traditional morphological features of Arabic, in a compact yet transparent notation. Each feature and possible values of the Morphological Features Tag Set were explained and illustrated in detail. 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, and the question mark "?" represents a feature relevant but its attribute value is not known yet.

Then we described the Arabic morphological analyzer algorithm which consists of six sequential steps. The morphological analyzer accepts different types of text whether the text is fully vowelized, partially vowelized or non-vowelized. The morphological analyzer outputs all possible analyses of the word. the analysis include all possible roots, clitics, affixes, stems, lemmas, patterns, different forms of vowelization, and the morphological features represented by detailed and fine-grain morphological feature tag.

A practical application of part of the morphological analyzer has been applied to lemmatize large and real data of the Web Arabic Corpus consisting of 100 million words. The lemmatization process highlighted some challenges when applying NLP application to large data. Such challenges are speed and spelling mistakes. We used high performance computing power of the NGS to run the lemmatizer in reduced time. And we added a spell-check procedure which detects and corrects the spelling errors caused by typing mistakes. The spell-check procedure is specific for Arabic.

The morphological analyzer uses linguistic lists of functional words, named entities and broken plural lists. It also used the broad-coverage lexical resource constructed by analyzing 23 traditional Arabic lexicons. The coverage of the constructed broad-coverage lexical resource showed that about 85% of the words processed using the lemmatizer referenced the broad-coverage lexicon and retrieved correct analyses for the analyzed words.

The evaluation plan of the morphological analyzer and part of speech tagger depends on constructing widely used general purpose gold standard for evaluation of different text domains, formats and genres of both vowelized and non-vowelized Arabic text. We showed the MorphoChallenge2009 Qur'an gold standard as example of constructing gold standard for evaluation of Arabic text processing applications.

8. References

Alqrainy, S. (2008). A Morphological-Syntactical Analysis Approach for Arabic Textual Tagging, by Shihadeh Alqrainy. School of Computing, De Montfort University. Doctor of Philosophy in Computer Science: 197.

Al-Shamsi, F., Guessoum, A. (2006). *A Hidden Markov Model-Based POS Tagger for Arabic*. 8es Journees internationales d'Analyse statistique des Donnees Textuelles.

Al-Suliti, L., Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics* 11, pp.135--171.

Atwell, E. (2008). Development of tag sets for part-of-speech tagging. In Anke Ludeling and Merja Kyto (eds.), *Corpus Liuistics: An International Handbook Volume 1* 501-526 Mouton de Gruyter.

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C., Wilcock, S. (2000). A comparative evaluation of modern English corpus grammatical

annotation schemes. *ICAME Journal, International Computer Archive of Modern and medieval English, Bergen* 24, pp. 7--23.

Benajiba, Y., Diab, M., Rosso, P. (2008) Arabic named entity recognition using optimized feature sets. *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii: Association for Computational Linguistics, pp.284--293

Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics* 21, pp.543--565.

Diab, M. (2007). Towards an Optimal POS Tag Set for Arabic Processing. *Proc RANLP*.

Diab, M., Hacioglu, K., Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. *Proceedings of HLT-NAACL*

Diwan, A. (2004) *al-mu'ğam an-naḥwī li-mufradāt al-luğa^h al- 'arabiyya^h* المعجم النحوي لمفردات اللغة العربية , Aleppo, Syria: fusselat lil-dirasāt wa at-tarğamah wa an-našir.

Dror, J., Shaharabani, D., Talmon, R., Wintner, S. (2004). Morphological Analysis of the Qur'an. *Literary and Linguistic Computing* 19, 431-452.

Duh, K., Kirchhoff, K. (2005). POS Tagging of Dialectal Arabic: A Minimally Approach. *ACL-05, Computational Approaches to Semitic Languages Workshop Proceedings,* pp.55--62. University of Michigan Ann Arbor, Michigan, USA.

Elliott, J., Atwell, E. (2000) Is anybody out there?: the detection of intelligent and generic language-like features. *Journal of the British Interplanetary Society*, 53, pp.13--22.

Freeman, A. (2001). Brill's POS Tagger and a Morphology Parser for Arabic. *NAACL 2001 Student Rersearch Workshop.*

Habash, N., Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.* Ann Arbor, Michigan, Association for Computational Linguistics.

Harmain, H. (2004). Arabic Part-of-Speech Tagging. *The Fifth Annual U.A.E. University Research Conference*. United Arab Emirats.

Khoja, S. (2001). APT: Arabic Part-of-Speech Tagger. *Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)***,** Carnegie Mellon University, Pittsburgh, Pennsylvania.

Marsi, E., Bosch, A., Soudi, A. (2005). Memory-based morphological analysis generation and part-of-speech tagging of Arabic. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages* 1-8. Ann Arbor: Association for Computational Linguistics.

Sawalha, M., Atwell, E. (2010). Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic. *Language Resource and Evaluation Conference LREC 2010* Valleta, Malta: 19-21 May 2010.

Sawalha, M., Atwell, E. (2009a). Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. *Proceedings of the 5th International Corpus Linguuistics Conference CL2009* Liverpool, UK.

Sawalha, M., Atwell, E. (2009b). توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية(Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language). *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy.* Damascus, Syria.

Sawalha, M., Atwell, E. (2008). Comparative evaluation of Arabic language morphological analysers and stemmers. *Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics*.

Schmid, H., Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. *COLING'08* Manchester, UK.

Talmon, R., Wintner, S. (2003). Morphological Tagging of the Qur'an. *In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop* Budapest, Hungary: April 2003.

Teahan, B. (1998). *Modelling English Text.* PhD Thesis, Department of Computer Science, University of Waikato, New Zealand.

Tlili-Guiassa, Y. (2006). Hybrid Method for Tagging Arabic Text. *Journal of Computer Science* 2(3): pp.245--248

Voutilainen, Atro (2003). Part-of-Speech Taging. In Ruslan Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*. Oxford University Press. pp. 219--232

Zibri, C., Torjmen, A., Ahmad, M. (2006). An Efficient Multi-agent system Combining POS-Taggers for Arabic Texts. *CICLing 2006,* LNCS 3878.