Motivation

Data released for Arabic Treebank by Linguistic Data Consortium in several formats, including:
- "vocalized" (with vowel)
- "unvocalized" (without vowel)

Parsing work so far on unvocalized data
- More closely represents "real-world"?

What do unvocalized and vocalized really mean in ATB?

Diacritics

What are diacritics?

- Superscript and subscript marks in the Arabic orthographic system
- Represents the three short vowels (ا, ل, ي)

Mark vocally of four letters (ا، ل، ل، ي) use of diacritical marks. vlw.

And what do they indicate?

- Short vowels
- MSA grammatical functions, such as verbs, passive forms and irregular noun plurals forms
- Long vowels: ملک (ملك) ملك (ملك)
- Cases for nouns: ملك (ملك)
- Orthographic normalization

Non-vocalic diacritics

- Shadda (consonantal length or gemination) is used for the derivation of new words
- Hamza marks the existence of the glottal stop
- Complex graphic support

Sukun (a small superscript zero-shaped grapheme) is nothing more than the absence of a vowel. The sukun is used for syllabic identification and to mark the imperative, the passive verb forms and 5 nouns

Diacritics and ambiguity: Lexical senses

The loss of the internal diacritics (such as short vowels, hamza, or shadda) leads to the following types of ambiguity, as exemplified in a given MSA lemma: إيم

1. An ambiguity within 'core' part of speech (POS) tags, distinguishing different lexical senses.
- إيم (a noun meaning 'science, learning'
- إل (another noun meaning 'tag

Diacritics and ambiguity: Different POS

2. A second type of 'core' POS tag ambiguity, distinguishing different core POS tags (lexically and semantically connected) Example:
- إلما (3rd Person Masculine, Singular: Perfective Verb (MSA Verb Form I) meaning 'the learned')
- إلما (3rd Person Singular: Passive Verb (MSA Verb Form I) meaning 'the learned')
- إلما (for the transitive, causative, derivative verb (MSA Verb Form I) meaning 'taught')

Diacritics and ambiguity: Inflectional endings

3. Structural/grammatical level, where the use of short vowels is correlated with case (nominal) and mood/aspect (verbal) information.

Role of diacritization and the life of a token

Source Tokens: Source text consists of words treated as whitespace-delimited tokens, usually lacking diacritic information
POS Tokens: Annotation's choice of Buckwalter Arabic Morphological Analyzer solution, includes morphological segmentation of the word and vocalization/diacritization of each segment
Treebank Tokens: POS tokens including separation of clitics as necessary

Conclusion

- To evaluate a parser on "real-world" data, the unvocalized form is insufficient
- Already tokenized
- Orthographically normalized
- Future work on parser evaluation must
- Take both issues above into account
- Disclose what degree of diacritization is chosen for parsing and parser evaluation

Example:

- Added consonants in unvocalized data
- Orthographic normalization

Linking unvocalized and vocalized trees in ATB

- Modified release format of ATB with explicit the links between unvocalized and vocalized trees
- Maximum flexibility for experimentation
- Points to original source file to relate different annotation levels
- Tries with complex terminals including Source token, vocalized and unvocalized forms, lemma and gloss
- Brings together information previously available only by accessing multiple release formats

Example:

- White-space delimited input string (source token) is &lt;DIA + Data Model suffix + &lt;DIA + Data Model suffix
- Preposition: Vocalized tokens take &lt;DIA + Data Model suffix, which not only adds the short vowels and &lt;DIA + Data Model suffix, but also corrects the "missing hamza" problem by normalizing &lt;DIA + Data Model suffix
- Unvocalized form of the latter results from stripping out the short vowels and &lt;DIA + Data Model suffix
- Not correct for the fuller form, which do not have the correct hamza presence, with &lt;DIA + Data Model suffix

Diacritic Annotation in the Arabic Treebank and Its Impact on Parser Evaluation

Mohamed Maamouri, Seth Kulick, Ann Bies
Linguistic Data Consortium
University of Pennsylvania, USA
{maamouri,skulick,bies}@ldc.upenn.edu