Winter School on Arabic Language Processing
Princess Sumaya University for Technology – January 27-29, 2014

MADAMIRA DEMO

Nizar Habash
Center for Computational Learning Systems
Columbia University
MADAMIRA Overview

**Input Arabic Text**

**Morphological Disambiguation (MADA)**
- Extracts Morphological Features (e.g. POS, gender)
- Uses statistical models with context to disambiguate
- **Output**: Fully disambiguated text with full morphology

**Tokenization (TOKAN)**
- General, highly-customizable generator of tokenized forms and/or morphological features
- **Output**: Tokenized text and/or feature forms according to a defined scheme

**Base Phrase Chunking (BPC)**
- Groups and labels sequences of words to form syntactic phrases (e.g. NPs, VPs, ADJPs, ADVPs)
- **Output**: Input text grouped into base phrases

**Named Entity Recognition (NER)**
- Identifies Named Entities (people, organizations, locations, etc.) mentioned in input text
- **Output**: Input text with Named Entities tagged

**User NLP Applications (IR, MT, and others)**
Morphological Disambiguation in English

- Select a morphological tag that fully describes the morphology of a word
- Complete English morphological tag set (Penn Treebank): 48 tags

Verb:

<table>
<thead>
<tr>
<th>VB</th>
<th>VBD</th>
<th>VBG</th>
<th>VBN</th>
<th>VBP</th>
<th>VBZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>go</td>
<td>went</td>
<td>going</td>
<td>gone</td>
<td>go</td>
<td>goes</td>
</tr>
</tbody>
</table>

- Same as “POS Tagging” in English
Morphological Disambiguation in Arabic

- Morphological tag has 14 subtags corresponding to different linguistic categories
  - Example: Verb
    Gender(2), Number(3), Person(3), Aspect(3), Mood(3), Voice(2), Pronominal clitic(12), Conjunction clitic(3)
- 22,400 possible tags
  - Different possible subsets
- 2,200 appear in Penn Arabic Tree Bank Part 1 (140K words)
- Example solution: MADA (Habash&Rambow 2005)
## Inflectional Morphology Terminology

<table>
<thead>
<tr>
<th>Word</th>
<th>A space/punctuation delimited string</th>
<th>lilmaktabapi</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexeme</strong></td>
<td>The set of all inflectionally related words</td>
<td>maktabap, lilmaktabapi, Almaktabapu, walimaktabatihA, etc.</td>
</tr>
<tr>
<td><strong>Lemma</strong></td>
<td>An ad hoc word form used to represent the lexeme</td>
<td>maktabap</td>
</tr>
<tr>
<td><strong>Features</strong></td>
<td>The space of variation of words in a lexeme</td>
<td>Clitics: li_prep, Al_det, Gen:f, num:s, stt:d, cas:g</td>
</tr>
<tr>
<td><strong>Root</strong></td>
<td>The root morpheme of the Lexeme</td>
<td>k-t-b</td>
</tr>
<tr>
<td><strong>Stem</strong></td>
<td>The core root+pattern substring; it does not include any affixes</td>
<td>maktab</td>
</tr>
<tr>
<td><strong>Segmentation</strong></td>
<td>A shallow separation of affixes</td>
<td>li+l+maktab+ap+i</td>
</tr>
<tr>
<td><strong>Tokenization</strong></td>
<td>Segmentation + morpheme recovery</td>
<td>li+Al+maktab+ap+i</td>
</tr>
</tbody>
</table>
# Cliticization Features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>(Some Important) Feature Values</th>
</tr>
</thead>
</table>
| **PRC3**    | Proclitic 3  
Sabqa 3  
<a_ques, 0>  
أداة استفهام، 0 |
| **PRC2**    | Proclitic 2  
Sabqa 2  
fa_conj, wa_conj, 0  
حروف عطف، 0 |
| **PRC1**    | Proclitic 1  
Sabqa 1  
bi_prep, li_prep, sa_fut, 0  
حروف جر، سن الاستقبال، 0 |
| **PRC0**    | Proclitic 0  
Sabqa 0  
Al_det, mA_neg, 0  
ال التعريف، أداة نفي، 0 |
| **ENC0**    | Enclitic  
لاحقة 0  
3ms_dobj, 3ms_poss, …, 0  
ضمير مفعول به مباشر  
مفرد مذكر للغائب،  
ضمير ملكية مفرد مذكر للغائب، …، 0 |
### Inflectional Features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>(Some Important) Feature Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PER</strong></td>
<td>Person</td>
</tr>
<tr>
<td><strong>ASP</strong></td>
<td>Aspect</td>
</tr>
<tr>
<td><strong>VOX</strong></td>
<td>Voice</td>
</tr>
<tr>
<td><strong>MOD</strong></td>
<td>Mood</td>
</tr>
<tr>
<td><strong>GEN</strong></td>
<td>Gender</td>
</tr>
<tr>
<td><strong>NUM</strong></td>
<td>Number</td>
</tr>
<tr>
<td><strong>STT</strong></td>
<td>State</td>
</tr>
<tr>
<td><strong>CAS</strong></td>
<td>Case</td>
</tr>
</tbody>
</table>
MORPHOLOGICAL ANALYZER

• Rule-based
• Human-created

MORPHOLOGICAL CLASSIFIERS

• Multiple independent classifiers
• Corpus-trained

RANKER

• Heuristic or corpus-trained

MADA (Habash & Rambow 2005)
(Habash & Rambow 2007)
(Roth et al. 2008)
**MADA 3.2 (MSA) Evaluation**

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>PATB 3 Blind Test</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>MADA</td>
<td>Error (↓)</td>
</tr>
<tr>
<td>All</td>
<td>74.8%</td>
<td>84.3%</td>
<td>38%</td>
</tr>
<tr>
<td>POS + Features</td>
<td>76.0%</td>
<td>85.4%</td>
<td>39%</td>
</tr>
<tr>
<td>All Diacritics</td>
<td>76.8%</td>
<td>86.4%</td>
<td>41%</td>
</tr>
<tr>
<td>Lemmas</td>
<td>90.4%</td>
<td>96.1%</td>
<td>60%</td>
</tr>
<tr>
<td>Partial Diacritics</td>
<td>90.6%</td>
<td>95.3%</td>
<td>50%</td>
</tr>
<tr>
<td>Base POS</td>
<td>91.1%</td>
<td>96.1%</td>
<td>56%</td>
</tr>
<tr>
<td>Segmentation</td>
<td>96.1%</td>
<td>99.1%</td>
<td>77%</td>
</tr>
</tbody>
</table>

Baseline: most common analysis per word in training
**MADAMIRA Overview**

**Input Arabic Text**

- **Morphological Disambiguation (MADA)**
  - Extracts Morphological Features (e.g. POS, gender)
  - Uses statistical models with context to disambiguate
  - **Output**: Fully disambiguated text with full morphology

- **Tokenization (TOKAN)**
  - General, highly-customizable generator of tokenized forms and/or morphological features
  - **Output**: Tokenized text and/or feature forms according to a defined scheme

- **Base Phrase Chunking (BPC)**
  - Groups and labels sequences of words to form syntactic phrases (e.g. NPs, VPs, ADJPs, ADVPs)
  - **Output**: Input text grouped into base phrases

- **Named Entity Recognition (NER)**
  - Identifies Named Entities (people, organizations, locations, etc.) mentioned in input text
  - **Output**: Input text with Named Entities tagged

**User NLP Applications (IR, MT, and others)**
Tokenization (TOKAN)

- Deterministic, generalized tokenizer
- **Input**: disambiguated morph. analysis + tokenization scheme
- **Output**: highly-customizable tokenized text

```
wsyktbhA = lex:katab-u_1 gloss:write pos:verb prc3:0 prc2:wa_conj prc1:sa_fut prc0:0 enc0:3fs_dobj
```

<table>
<thead>
<tr>
<th>Example</th>
<th>Scheme</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>w+ syktbhA</code></td>
<td>D1</td>
<td>prc3 prc2 REST</td>
</tr>
<tr>
<td><code>w+ s+ yktbhA</code></td>
<td>D2</td>
<td>prc3 prc2 prc1 REST</td>
</tr>
<tr>
<td><code>w+ s+ yktb +hA</code></td>
<td>D3</td>
<td>prc3 prc2 prc1 prc0 REST enc0</td>
</tr>
<tr>
<td><code>w+ syktb +hA</code></td>
<td>ATB</td>
<td>prc3 prc2 prc1 prc0:IA prc0:mA REST enc0</td>
</tr>
<tr>
<td><code>w+•w+•wa+ syktbhA•syktbhA•katab</code></td>
<td>D1-3tier</td>
<td>prc3 prc2 REST ::FORM0 WORD ::FORM1 WORD NORM:AY ::FORM2 LEXEME</td>
</tr>
</tbody>
</table>
• Thank You!