Winter School on Arabic Language Processing

Princess Sumaya University for Technology – January 27-29, 2014

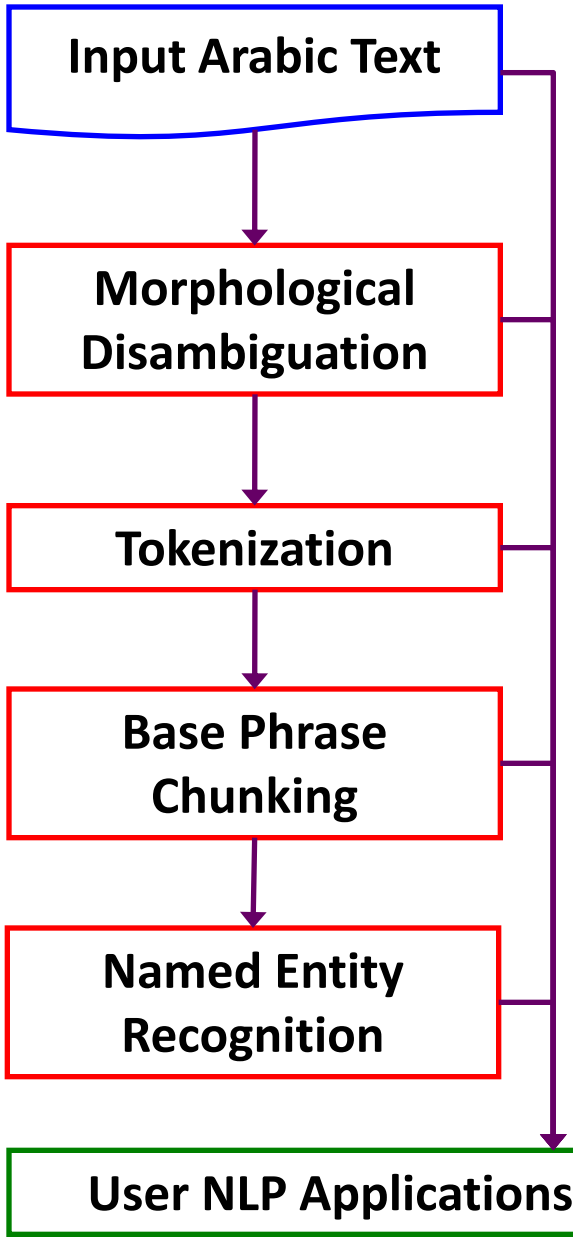
MADAMIRA DEMO

Nizar Habash

Center for Computational Learning Systems

Columbia University

MADAMIRA Overview



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

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:

VB	VBD	VBG	VBN	VBP	VBZ
go	went	going	gone	go	goes

- 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

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

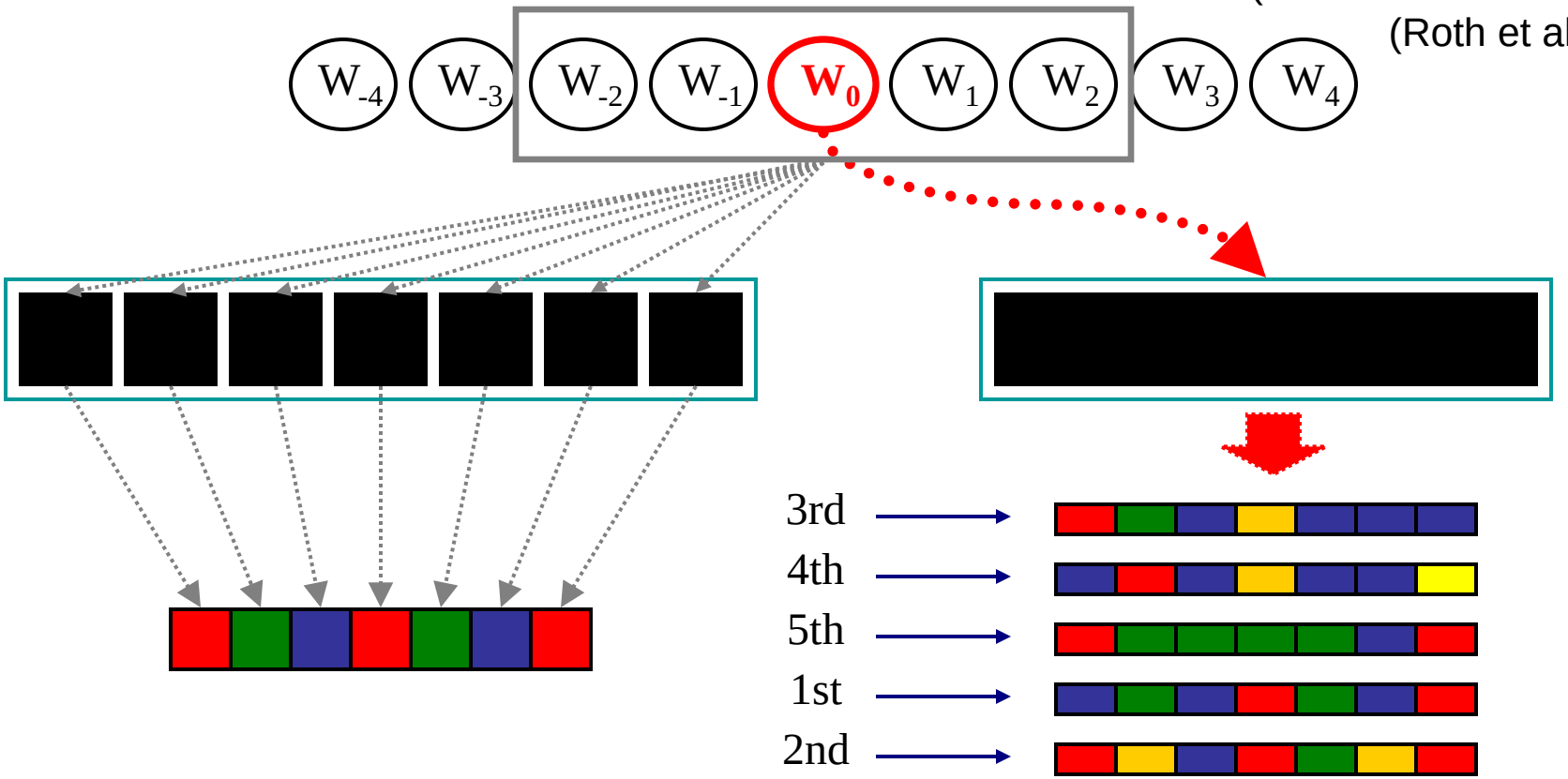
Cliticization Features

Feature Name			(Some Important) Feature Values	
PRC3	Proclitic 3	سابقة 3	<a_ques, 0	أداة استفهام،0
PRC2	Proclitic 2	سابقة 2	fa_conj, wa_conj, 0	حروف عطف،0
PRC1	Proclitic 1	سابقة 1	bi_prep, li_prep, sa_fut, 0	حروف جر، سين الاستقبال، 0
PRC0	Proclitic 0	سابقة 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

	Feature Name		(Some Important) Feature Values	
PER	Person	الشخص	1st, 2nd, 3rd, na	متكلم، مخاطب، غائب، غ/م
ASP	Aspect	الزمن	perfect, imperfect, command, na	ماضي، مضارع، أمر، غ/م
VOX	Voice	البناء	active, passive, na	للمعلوم، للمجهول، غ/م
MOD	Mood	الصيغة	indicative, subjunctive, jussive, na	مرفوع، منصوب، مجزوم، غ/م
GEN	Gender	الجنس	feminine, masculine, na	مؤنث، مذكر، غ/م
NUM	Number	العدد	singular, dual, plural, na	مفرد، مثنى، جمع، غ/م
STT	State	التعريف	indefinite, definite, construct, na	نكرة، معرفة، مضاف، غ/م
CAS	Case	الحالة	nominative, accusative, genitive, na	مرفوع، منصوب، مجرور، غ/م

MADA (Habash&Rambow 2005)
(Habash&Rambow 2007)
(Roth et al. 2008)



MORPHOLOGICAL CLASSIFIERS	RANKER	MORPHOLOGICAL ANALYZER
• Multiple independent classifiers • Corpus-trained	• Heuristic or corpus-trained	• Rule-based • Human-created

MADA 3.2 (MSA) Evaluation

Accuracy	PATB 3 Blind Test		
	Baseline	MADA	Error ↓
All	74.8%	84.3%	38%
POS + Features	76.0%	85.4%	39%
All Diacritics	76.8%	86.4%	41%
Lemmas	90.4%	96.1%	60%
Partial Diacritics	90.6%	95.3%	50%
Base POS	91.1%	96.1%	56%
Segmentation	96.1%	99.1%	77%

Baseline: most common analysis per word in training

وكاتب wkAtb

and (the) writer of

wakAtibu

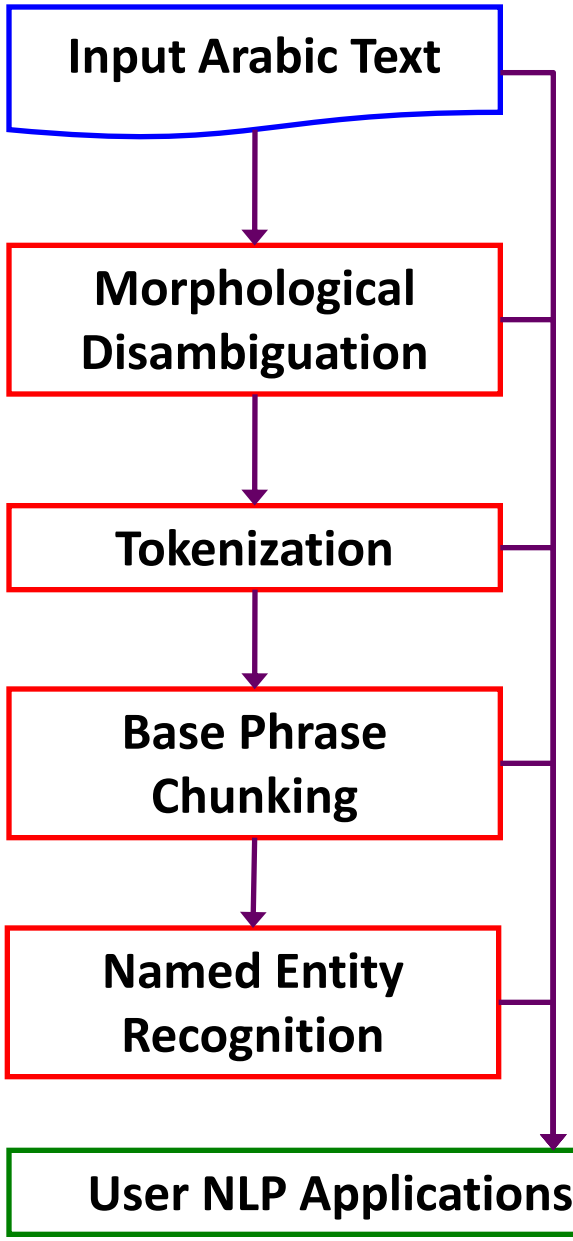
kAtib_1

pos:noun

prc3:0 prc2:wa_conj prc1:0 prc0:0 per:3 asp:na vox:na mod:na gen:m num:s stt:c cas:n enc0:0

w+ kAtb

MADAMIRA Overview



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

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
```

Example	Scheme	Specification
w+ syktbhA	D1	prc3 prc2 REST
w+ s+ yktbhA	D2	prc3 prc2 prc1 REST
w+ s+ yktb +hA	D3	prc3 prc2 prc1 prc0 REST enc0
w+ syktb +hA	ATB	prc3 prc2 prc1 prc0:lA prc0:mA REST enc0
w+•w+•wa+ syktbhA•syktbhA•katab	D1-3tier	prc3 prc2 REST ::FORM0 WORD ::FORM1 WORD NORM:AY ::FORM2 LEXEME

- Thank You!