Arabic POS Tagging

Emad Mohamed, Sandra Kübler

Indiana University

The Structure of Arabic Words

- An Arabic word may consist of several segments.
- Possible segments: inflectional affixes, the stem, clitics
- example: `WsyktbwnhA` (Engl.: *and they will write it*):
    - conjunction: `w`
    - future particle: `s`
    - 3rd person imperfect verb prefix: `y`
    - imperfect verb: `ktb`
    - 3rd person feminine singular object pronoun: `hA`

The Structure of Arabic Words

- An Arabic word may consist of several segments.
- Possible segments: inflectional affixes, the stem, clitics
- example: `WsyktbwnhA` (Engl.: *and they will write it*):
  - conjunction: `w`
  - future particle: `s`
  - 3rd person imperfect verb prefix: `y`
  - imperfect verb: `ktb`
  - 3rd person feminine singular object pronoun: `hA`
- POS tag:
  [CONJ+FUTURE_PARTICLE+
  IMPERFECT_VERB_PREFIX+IMPERFECT_VERB+
  IMPERFECT_VERB_SUFFIX_MASC_PLURAL_3RD_PERSON+
  OBJECT_PRONOUN_FEM_SINGULAR]

Tagging Approaches

- whole word tagging: assign complex tag to complete word

- segment-based tagging: segment first; then assign tags to segments

Tagging Approaches

- whole word tagging: assign complex tag to complete word
  `wsyktbwnhA`: CONJ+FUT+IV3MS+IV+IVSUFF_SUBJ:MP_MOOD:I+IVSUFF_DO:3FS
- segment-based tagging: segment first; then assign tags to segments
  - `w`: CONJ
  - `s`: FUT
  - `y`: IV3MS
  - `ktb`: IV
  - `wn`: _SUBJ:MP_MOOD:I
  - `hA`: IVSUFF_DO:3FS

Tagging Approaches

- whole word tagging: assign complex tag to complete word

  wsyktbwnhA:

  CONJ+FUT+IV3MS+IV+IVSUFF_SUBJ:MP_MOOD:I+IVSUFF_DO:3FS

  993 tags

- segment-based tagging: segment first; then assign tags to segments
  - w: CONJ
  - s: FUT
  - y: IV3MS
  - ktb: IV
  - wn: _SUBJ:MP_MOOD:I
  - hA: IVSUFF_DO:3FS

  139 tags

Data Set & Experimental Setup

- Penn Arabic Treebank (after-treebank POS files)
- P1V3 + P3V1: ca. 500 000 words
- non-vocalized version
- reattached conjunctions, prepositions, pronouns, etc. to get text as written
- remove null elements: `{i$otaraY+(null)` / PV+PVSUFF_SUBJ:3MS ⇒ `{i$otaraY` / PV
- 5-fold cross validation
- evaluation: per-segment accuracy (SAR) + per-word accuracy (WAR)

Memory-Based Segmentation

- per character classification: segment-end, no-segment-end
- memory-based learning: TiMBL
- features: focus character, previous 5 characters, and following 5 characters, POS tag for word based on whole word tagging
- TiMBL parameters: IB, overlap metric, gain ratio weighting, nearest neighbors $k = 1$
- two rounds: in second round include class from first round

Segmentation Results

all words:	98.23%
known words:	99.75%
unknown words:	82.22%

Segmentation Results

all words:	98.23%
known words:	99.75%
unknown words:	82.22%

proper noun errors:	33.87% of all errors

% unknown words in data:	8.5%

POS Tagging

- memory-based tagger: MBT
- parameters: Modified Value Difference metric, $k = 25$
- for **known words**: IGTree, 2 words to left, their POS tags, focus word, its ambitag, 1 right context word, its ambitag
- for **unknown words**: IB1, focus word, first 5 + last 3 characters, 1 left context word + its POS tag, 1 right context word + its ambitag
- previous decisions are included

POS Tagging Results

gold standard seg.		segmentation-based		whole words
SAR	WAR	SAR	WAR	WAR
96.72%	94.91%	94.70%	93.47%	94.74%

POS Tagging Results

gold standard seg.		segmentation-based		whole words
SAR	WAR	SAR	WAR	WAR
96.72%	94.91%	94.70%	93.47%	94.74%

POS Tagging Results

gold standard seg.		segmentation-based		whole words
SAR	WAR	SAR	WAR	WAR
96.72%	94.91%	94.70%	93.47%	94.74%

Discussion

- gold standard segmentation: upper bound
- gives best results

- no gold standard segmentation available: whole words better than automatic segmentation

- segmentation → more ambiguity per segment
- small percentage of unknown words

- in segmentation-based tagging, 28% of all errors are results of wrong segementation

Known vs. Unknown Words

	gold std. seg.	seg.-based	whole words
known words	95.90%	95.57%	96.61%
unknown words	84.25%	71.06%	74.64%

Known vs. Unknown Words

	gold std. seg.	seg.-based	whole words
known words	95.90%	95.57%	96.61%
unknown words	84.25%	71.06%	74.64%

Known vs. Unknown Words

	gold std. seg.	seg.-based	whole words
known words	95.90%	95.57%	96.61%
unknown words	84.25%	71.06%	74.64%

Known vs. Unknown Words

	gold std. seg.	seg.-based	whole words
known words	95.90%	95.57%	96.61%
unknown words	84.25%	71.06%	74.64%

Error Analysis

confusion sets:

gold	tagger	% of errors
noun	adjective	7.88%
adjective	noun	7.75%
proper noun	noun	9.10%
noun	proper noun	2.51%

Error Analysis

confusion sets:

gold	tagger	% of errors
noun	adjective	7.88%
adjective	noun	7.75%
proper noun	noun	9.10%
noun	proper noun	2.51%

- no clear distinction between nouns and adjectives in Arabic: adjectives behave morphologically like nouns and can be used as nouns
- proper nouns are normally standard nouns, and are no marked specifically

Comparison to Habash & Rambow

- whole word tagging
- then convert to Habash & Rambow tokenization + reduced tagset: 15 tags

	H&R ATB1	H&R ATB2	whole word tagger
Token. acc.	99.1	–	99.33
POS acc.	98.1	96.5	96.41

Conclusion & Future Work

- whole word tagging has higher accuracy than segmentation based tagging
- no preprocessing necessary
- but Penn Arabic Treebank has low percentage of unknown words

- segmentation quality is bottleneck for improving segmentation-based tagger
- need to find more reliable segmentation
- will integrate vocalization with segmentation