Segmentation in Arabic NLP

Rudy Khalil

Typology of Morphosyntax

25.01.2019

Overview

- POS Tagging
- Arabic NLP Ambiguity
- Segmentation

POS Tagging

ذهب الطالب إلى المدرسة

The scholar went to the school

V det N PP det N

POS Tagging

ذهب الطالب إلى المدرسة

The scholar went to the school

V det N PP det N

Easy, right?

POS Tagging

ذهب الطالب إلى المدرسة

The scholar went to the school

V det N PP det N

Easy, right? Thing again!

POS Tagging in The Wild

/wakabiyutina/و كبيوتنا

و + ك + بيوت + نا

Wa+ka+biyut+na

And+like+houses+our

And like our houses

Remember this?

POS Tagging in The Wild

/wakabiyutina/و كبيوتنا

و + ك + بيوت + نا

Wa+ka+biyut+na

And+like+houses+our

And like our houses

CONJ+ particle + N+ poss

POS Tagging in The Wild - Ambiguity (1)

/wakabiyutina/و كبيوتنا

نا + بيوت + ك + و

Wa+ka+biyut+na

And+like+houses+our

And like our houses

Adjunt + particle + N+ poss

For a good POS Tagger we need a good segmenter.

POS Tagging in The Wild - Ambiguity (2)

- For a good POS Tagger (and other NLP tasks) we need a good segmenter

- Arabic word could contains clitics (proclitics and enclitics) that should be tagged properly

Segmentation

- **Conditional Random Fields (CRF)**
  Souhir Gahbiche-Braham et. el.
- **Neural Networks based Segmentation**
  Nizar Habash et. el.
- **Word Segmentation with Domain Adaptation (for dialects)**
  Monroe et. el.

Conditional Random Fields (1)

- Combinations of Affixes and roots depend on the main category of the word
  Ex. سيحاربنا(Nsubj) بيتنا(Poss)
- Approach proposed : Joint morphological decomposition and POS tagging using CRF (statistical model)
- POS tagger includes unigram, bigrams and trigrams test in a window of 7, 5 and 3 words and limited prefix and suffix tests.
- Segmentation Prediction:
  - SEG: prefixes are predicted without any POS feature
  - POS-then-SEG: the POS tags are first predicted and the used to predict prefixes
  - POS+SEG: POS tags and prefixes are predicted simultaneously

Conditional Random Fields (2) - Features

pr1, pr2, pr3 encode the presence/absence and type of prefixes

Prefix	Label/Value
pr1	CONJ/*w*+, *f*+ or *none*
pr2	PREP/*b*+, *l*+, *k*+ or SUB/l+ or FUT/*s*+ or *none*
pr3	DET/*Al*+ or *none*

Table 1: Prefixes, labels and values

Conditional Random Fields (3) - Results

Scheme	SEG	POS-then-SEG	POS+SEG
pr1+pr2+pr3	0.78%	0.64%	**0.60%**
pr1	0.22%	**0.18%**	**0.18%**
pr2	0.46%	0.35%	**0.34%**
pr3	0.13%	0.13%	**0.11%**
POS	-	4.20%	**3.72%**
After segmentation	0.55%	0.42%	**0.40%**

Table 3: Segmentation Error Rate of the different schemes

Neural Networks based Segmentation (1)

- Neural Networks and good at sequence labeling (if you have enough data to train)
- LSTM is good for long-sequence tagging. Other approaches fail when there's long dependency (CRF with 3 word window for example)
- The morphological disambiguation task involves choosing the correct morphological analysis from the set of potential analysis obtained from the analyzer
- Train several models for individual morphological features, and use the result to score and rank the different analyses and choose an optimal one
- Word and subword and character embedding used to get morphological information

Neural Networks based Segmentation (2)

Feature	Definition
diac	Diacratization
lex	Lemma
pos	Basic part-of-speech tags (34 tags)
gen	Gender
num	Number
cas	Case
stt	State
per	Person
asp	Aspect
mod	Mood
vox	Voice
prc0	Proclitic 0, article proclitic
prc1	Proclitic 1, preposition proclitic
prc2	Proclitic 2, conjunction proclitic
prc3	Proclitic 3, question proclitic
enc0	Enclitic

Neural Networks based Segmentation (3)

- Morphological Dictionary: Analyzer that encodes all the word inflection in a language

- Lightstemmer: Language-specific Affixes

Model	Embedding	
	Word	Char
No Morphology	96.4	96.7
Fixed Character Affixes	96.6	NA
Lightstemmer	96.7	96.8
Morphological Dictionary	97.5	97.5
+ Fixed Character Affixes	**97.6**	NA
+ Lightstemmer	**97.6**	**97.6**

Word Segmentation with Domain Adaptation (for dialects)

- Monroe et. el use a statistical model to obtain better POS tags for egyptian dialect
- Character-level Conditional Random Fields to get a better segmenter for Arabic clitic
- As an Input, there's the sequence to be tagged and other features that specify the egyptian dialects
- The model perform better than other segmenters that are designed toward MSA.
- Improved Speed

References

- Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier - Souhir Gahbiche-Braham et. el.
- Don’t Throw Those Morphological Analyzers Away Just Yet - Habash et. el.
- Word Segmentation of Informal Arabic with Domain Adaptation - Monroe et. el.

Thanks