NATURAL LANGUAGE PROCESSING

LESSON 5: PART OF SPEECH (POS) TAGGING

OUTLINE

- **POS Tagging Methods**
  - Rule-Based
  - Stochastic
  - Hybrid (Transformation-based)
- **Treebanks and POS Tags**
- **POS Taggers**
  - Stanford Log-linear Part-of-Speech Tagger
  - LingPipe POS Tagger
  - NLTK POS Taggers
  - CLAWS POS Tagger for English
  - TSCorpus POS Tagger

PART OF SPEECH TAGGING METHODS

Generally three models are mentioned:

- Rule-Based Part-of-Speech Tagging
- Stochastic Part-of-Speech Tagging
- Hybrid (Transformation-based) Part-of-Speech Tagging

PART OF SPEECH TAGGING METHODS

Rule-Based Part-of-Speech Tagging

- Starts from 1960s with two stage architecture.
- First stage used a dictionary to assign each word a list of potential parts-of-speech.
- Second stage used large lists of hand-written disambiguation rules to winnow down this list to a single POS-Tag for each word.

PART OF SPEECH TAGGING METHODS

Rule-Based Part-of-Speech Tagging

- May not be practical for active natural languages
- Rules will never cover all situations
- Turkish has complex evolution history:
  - Ottoman Turkish era: Rich interaction with Arabic (Semitic) and Persian(Indo-European) languages.
  - Turkey Turkish era: Interaction with French, English and German.

PART OF SPEECH TAGGING METHODS

Stochastic Part-of-Speech Tagging

- Probabilistic approaches begins at late 1960s
- Hidden Markov Model (HMM) Taggers
- For a given sentence or word sequence, HMM taggers choose the tag sequence that maximizes the following formula:

$$P(word|tag) * P(tag|previous\ n\ tags)$$

PART OF SPEECH TAGGING METHODS

Stochastic Part-of-Speech Tagging

- Markov Models
  - Sequence of random variables that aren't independent
  - Have a limited relation (only with previous and next word)
  - Define with a transition matrix and initial state probabilities.
  - For a given sentence S with words ($w_1, \dots w_t$), Markov Model can be defined as:

$$P(w_1, \dots w_t) = P(w_1)P(w_2|w_1)P(w_3|w_2w_1) \dots P(w_t|w_{t-1} \dots w_1)$$
$$= P(w_1)P(w_2|w_1)P(w_3|w_2) \dots P(w_t|w_{t-1})$$

PART OF SPEECH TAGGING METHODS

- Assume a finite set of words V and a finite set of tags K. Define S to be the set of all sequence/tag-sequence pairs $(x_1 \dots x_n, y_1 \dots y_n)$ such that $n \geq 0, x_i \in V \; for \; i = 1 \dots n \; and \; y_i \in K \; for \; i = 1 \dots n$
- For any $(x_1 \dots x_n, y_1 \dots y_n) \in S$, $\mathrm{p}(x_1 \dots x_n, y_1 \dots y_n) \geq 0$
- and $\sum_{(x_1 \dots x_n, y_1 \dots y_n) \in S} \mathrm{p}(x_1 \dots x_n, y_1 \dots y_n) = 1$
- Given a generative tagging model, the function from words to tag sequences is defined as:

$$f(x_1 \dots x_n) = \arg \max_{y_1 \dots y_n} \mathrm{p}(x_1 \dots x_n, y_1 \dots y_n)$$

PART OF SPEECH TAGGING METHODS

- Trigram Hidden Markov Model
  - A finite set V of possible word and a finite set K of possible tags

    q(s|u,v)
  - for any trigram (u,v,s) such that s ∈ K ∪ {STOP}, and u,v ∈ V ∪ {*}. The value for q(s|u,v) can be interpreted as the probability of seeing the tag s immediately after the bigram of tags (u,v).

    e(x|s)
  - for any x ∈ V, s ∈ K. The value for e(x|s) can be interpreted as the probability of seeing observation x paired with state s.

PART OF SPEECH TAGGING METHODS

- Trigram Hidden Markov Model
  - Define S to be the set of all sequence/tag-sequence pairs $(x_1 \dots x_n, y_1 \dots y_{n+1})$ such that $n \geq 0, x_i \in V \ for \ i = 1 \dots n,$ $y_i \in K \ \ for \ i = 1 \dots n, \ and \ y_{n+1} = STOP$
  - We then define the probability for any $(x_1 \dots x_n, y_1 \dots y_{n+1}) \in S$ as:

$$p(x_1 \dots x_n, y_1 \dots y_{n+1}) = \prod_{i=1}^{n+1} q(y_i | y_{i-2}, y_{i-1}) \prod_{i=1}^{n} e(x_i | y_i)$$

(assuming $y_0 = y_{-1} = *$)

PART OF SPEECH TAGGING METHODS

- As one example, if we have n = 3, $x_1 \ldots x_3$ equal to the sentence «the dog laughs.»

and $y_1 \ldots y_4$ equal to the tag sequence «D N V STOP»

Then,

$$\boldsymbol{p(x_1 \ldots x_n, y_1 \ldots y_{n+1}) = Q \times E}$$

where

$$\boldsymbol{Q = q(D|*,*) \times q(N|*,D) \times q(V|D,N) \times q(STOP|N,V)}$$
$$\boldsymbol{E = e(the|D) \times e(dog|N) \times e(laughs|V)}$$

PART OF SPEECH TAGGING METHODS

- The quantity of $\boldsymbol{Q}$ is the prior probability of seeing the tag sequence «D N V STOP»

$$\boldsymbol{q(D|*,*) \times q(N|*,D) \times q(V|D,N) \times q(STOP|N,V)}$$

- The quantity $\boldsymbol{E}$ can be interpreted as the conditional probability. Here, p(the dog laughs|D N V STOP)

$$\boldsymbol{e(the|D) \times e(dog|N) \times e(laughs|V)}$$

PART OF SPEECH TAGGING METHODS

Stochastic Part-of-Speech Tagging

- Conditional Random Field (CRF) Model Taggers
  - For sentence ***S***: «*This new movie is totally awesome*.», the position of a word: ***i***
  - The label of the current word: $\boldsymbol{label_i}$ and The label of the previous word: $\boldsymbol{label_{i-1}}$
  - Define feature functions like:

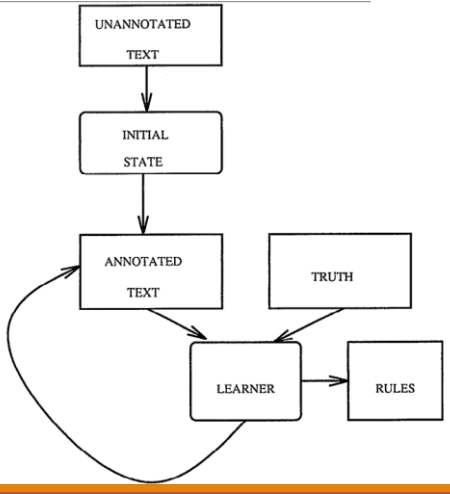
$$f_1(S, i, label_i, label_{i-1}) = \begin{cases} 1, & label_i = ADVERB \text{ and } i^{th} \text{ word ends in "} -ly\text{"} \\ 0, & otherwise \end{cases}$$

  - Every HMM model can be built as a CRF model. On the other hand, vice versa may not be true.

PART OF SPEECH TAGGING METHODS

Transformation-based Tagging (Hybrid)

- Transformation-Based Learning approach to machine learning (Brill, 1995)
- Inspires from both the rule-based and stochastic taggers.
- Uses broadest rule first than goes for narrower rule till all of them applied.



TREEBANKS AND POS TAGS

- Brown Corpus & Lancaster-Oslo-Bergen (LOB) Corpus
  - has 85 tags
- British National Corpus
  - has 61 tags
- PENN Treebank:
  - has 45 tags

PART OF SPEECH TAGGERS

Some of POS Taggers in the literature

- Stanford Log-linear Part-of-Speech Tagger
- LingPipe POS Tagger
- NLTK POS Taggers
- CLAWS POS Tagger for English
- TSCorpus POS Tagger

PART OF SPEECH TAGGERS

- Stanford Log-linear Part-of-Speech Tagger
  - Written by Kristina Toutanova, improved by Dan Klein, Christopher Manning, William Morgan, Anna Rafferty, Michel Galley, and John Bauer
  - Using «Maximum Entropy» model for POS Tagging
  - It has default 3 trained taggers for English, and one for Arabic, Chinese, French and German.
  - It can be retrained with a POS-annotated training text for any other language.

PART OF SPEECH TAGGERS

- LingPipe POS Tagger
  - Commercial product, closed source statistical model for POS tagging
  - Works default with English, but can be trained with any other language (it has no extra data for other languages).

PART OF SPEECH TAGGERS

- NLTK POS Taggers (some tools)
  - FeaturesetTagger (Stochastic)
  - NGramTagger (Stochastic)
  - BrillTagger (Transitional-hybrid)
  - CRFTagger (Stochastic)
  - HiddenMarkovModelTagger (Stochastic)
  - PerceptronTagger (Default POS Tagger of NLTK)

PART OF SPEECH TAGGERS

- Default NLTK POSTagger for English(PerceptronTagger):

```
$ python
Python 3.6.3 (v3.6.3:2c5fed8, Oct  3 2017, 17:26:49) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> sentence = "This is a simple test sentence for part of speech tagging in English."
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['This', 'is', 'a', 'simple', 'test', 'sentence', 'for', 'part', 'of', 'speech', 'tagging', 'in',
 'English', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged
[('This', 'DT'), ('is', 'VBZ'), ('a', 'DT'), ('simple', 'JJ'), ('test', 'NN'), ('sentence', 'NN')
, ('for', 'IN'), ('part', 'NN'), ('of', 'IN'), ('speech', 'NN'), ('tagging', 'VBG'), ('in', 'IN')
, ('English', 'NNP'), ('.', '.')]
>>>
```

PART OF SPEECH TAGGERS

- CLAWS POS Tagger for English
  - CLAWS (the Constituent Likelihood Automatic Word-tagging System), has been continuously developed since the early 1980s.
  - The latest version of the tagger, CLAWS4, was used to POS tag c.100 million words of the British National Corpus (BNC).
  - CLAWS4 uses both probabilistic and rule-based methods mixed, calls itself as a hybrid POSTagger.
  - CLAWS has consistently achieved 96-97% accuracy.
  - It has some versions:
    - CLAWS1 has 132 tags, while CLAWS2 has 166 tags.

PART OF SPEECH TAGGERS

- TSCorpus POS Tagger
  - Uses a perceptron-based morphological disambiguator for Turkish text
  - Has 2 corpus source: TS Corpus & TrMorph lexicon source
  - It has extra POSTags for internet language:
    - İntabbr (Internet Abbreviations) -> slm (selam)
    - Emoticon -> ☹
    - intSlang -> «slangs in tweets»
    - YY (Misspelling) -> qibi (gibi)
    - tinglish -> feysbuk (facebook), tivit (tweet)

PART OF SPEECH TAGGING ISSUES

- Words can be ambigous between multiple tags.
  - In order to solve it, the researchers still try to improve methods.
- Unknown words: new words not included in actual dictionary data
  - For all living languages, an updated dictionary is needed.

SOME REFERENCES

- **Daniel Jurafsky and James H. Martin, 1999**. Speech and Language Processing. Prentice Hall, New Jersey.
- **Christopher D. Manning and Hinrich Schütze, 1999.** Foundations of Statistical Natural Language Processing, MIT Press, Massachusetts.
- https://en.wikipedia.org/wiki/Treebank
- https://en.wikipedia.org/wiki/Brown_Corpus
- https://en.wikipedia.org/wiki/Part-of-speech_tagging
- http://dev.tscorpus.com/postagger/index.php
- https://nlp.stanford.edu/software/tagger.shtml
- http://alias-i.com/lingpipe/demos/tutorial/posTags/read-me.html
- http://www.nltk.org/