POS

based on Jurafsky and Martin Ch. 5

Miriam Butt

October 2014

Parts of Speech

There are ten parts of speech and they are all troublesome.

Mark Twain
The awful German Language

The definitions [of the parts of speech] are very far from having attained the degree of exactitude found in Euclidean Geometry.

Otto Jespersen
The Philosophy of Grammar

Parts of Speech

A gnostic was seated before a grammarian. The grammarian said, 'A word must be one of three things: either it is a noun, a verb or a particle.' The gnostic tore his robe and cried, 'Alas! Twenty years of my life and striving and seeking have gone to the winds, for I laboured greatly in the hope that there was another word outside of this. Now you have destroyed my hope.' Though the gnostic had already attained the word which was his purpose, he spoke thus to arouse the grammarian.

Rumi
The Discourses of Rumi
(from J&M p. 157)

Parts of Speech

Go back to early Greek grammar (techne by Thrax).

8 POS: noun, verb, pronoun, preposition, adverb, conjunction, participle, article.

CL Applications:

- 45/36 (Penn Treebank)
- 61 (CLAWS, for the BNC)
- 54 (STTS, German standard)

POS Tags

- Why so many POS Tags in CL?

  Machines (and humans) need to be as accurate as possible.

  (Though ADV tends to be a garbage category).

- Why the Differences?

  Different Languages have different requirements.

Compare the Penn Tagset with STTS in detail.

On-going work: Universal Tag Set (e.g., Google)

Word Classes

Open Class: Nouns, Verbs, Adjectives, Adverbs

vs.

Closed Class: Auxiliaries, Articles, Conjunctions, Prepositions/Particles

Because languages have open word classes, one cannot simply list word+tag associations.

What to do?

POS Tagging

Methods:

1. Manual Tagging
2. Machine Tagging
3. A Combination of Both

Manual Tagging

Methods:

1. Agree on a Tagset after much discussion.
2. Chose a corpus, annotate it manually by two or more people.
3. Check on inter-annotator agreement.
4. Fix any problems with the Tagset (if still possible).

Machine Tagging

Methods:

1. Rule based tagging.
2. Stochastic tagging.
3. A combination of both.

Rule Based Tagging

Mostly used by early applications (1960s-1970s)

Methods:

1. Use a lexicon to assign each word potential POS.
2. Disambiguate POS (mostly open classes) via rules:

   to race/VB vs. *the race/NN*

   This entails some knowledge of syntax (patterns of word combination).

Rule Based Tagging: ENGTWOL

ENGTWOL (Voutilainen 1995)

Methods:

1. Morphology for lemmatization.
2. 56 000 entries for English word stems (first pass)
3. 3744 handwritten constraints to eliminate tags (second pass)

Rule Based Tagging: ENGTWOL

Example: First Pass

Pavlov	**PAVLOV N NOM SG PROPER**
had	**HAVE V PAST VFIN SVO**
	HAVE PCP2 SVO
shown	**SHOW PCP2 SVOO SVO SV**
that	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation	**N NOM SG**

Rule Based Tagging: ENGTWOL

Example: Second Pass

Adverbial-that rule

Given input "that"

if

(+1 A/ADV/QUANT); /* if next word is one of these */

(+2 SENT-LIM); /* and following is a sentence boundary */

(NOT -1 SVO/A); /* and previous word is not a verb like */

/* consider (object complements) */

/* "I consider that odd." */

then eliminate non-ADV tags

else eliminate ADV tag

Machine Tagging

Wide-spread Today

Methods:

1. Use a lexicon to assign each word potential POS.
2. Disambiguate POS (mostly open classes) via learned patterns: what type of word is most likely to follow a given POS? *to race/VB* vs. *the race/NN*

   This entails machine learning.

Machine Learning

Methods:

1. Take a hand tagged corpus
2. Have the machine learn the patterns in the corpus.
3. Give the machine a lexicon of word+tag associations.
4. Give the machine a new corpus to tag.
5. The machine uses the initial information in the lexicon and the patterns it has learned to tag the new corpus.
6. Examine the result and correct the output.
7. Give the corrected output back to the machine for a new round.
8. Keep going until the machine is not learning any more.

Machine Tagging

- Example in J+M: HMM (Hidden Markov Models)
- Others also possible, e.g. Neural Nets

Probability of Tag Assignment
P(word|tag) * P(tag|previous *n* tags)

If we are expecting a tag (e.g., V), how likely is it that this word would appear (e.g., race)?

Bigram or Trigram Strategy is commonly used.

Machine Tagging

Simplified Example from J+M 176-178

(1) Secretariat/NNP is /VBZ expected/VBN to/TO **race/??** tomorrow/NN

(2) People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN
for/IN the/DT **race/??** for/IN outer/JJ space/NN

race: VB or NN?

Bigram Analysis

P(race|VB)*P(VB|TO) vs. P(race|NN)*P(NN|TO)
P(race|VB)*P(VB|DT) vs. P(race|NN)*P(NN|DT)

Machine Tagging

Likelihoods from Brown+Switchboard Corpora

P(race|VB) = .00003 P(VB|TO) = . 34
P(race|NN) = .00041 P(NN|TO) = . 021

Result for first sentence: **race/VB**
P(race|VB)*P(VB|TO) = .00001
P(race|NN)*P(NN|TO) = .000007

Combination Tagging

• Most taggers today use a combination of some rules plus learned patterns.

• The famous *Brill Tagger* uses a lexicon, and handwritten rules plus rules learned on the basis of a corpus (previous errors in tagging).

• Accuracy of today's taggers: 93%-97%.

So, they are accurate enough to be a useful first step in many applications.

Common Tagging Problems

- Multiple Words
- Unknown Words

Very good German tagger is the TreeTagger by Helmut Schmid (IMS).

Common Problem:

Das bedachte/V *ich*. vs. *Das bedachte*/ADJ *Haus*

Treebanks

- Machine learning can only be done on the basis of a huge corpus.

- Treebanks store these types of corpora (mostly initially tagged by hand).

- Examples: Penn Treebank, BNC, COSMAS, TIGER

Online Taggers

- http://www.infogistics.com/posdemo.htm
- https://open.xerox.com/Services/fst-nlp-tools/Pages/Part%20of%20Speech%20Tagging
- https://open.xerox.com/Services/fst-nlp-tools/