Lecture 11:
Parts of speech

Intro to NLP, CS585, Fall 2014
http://people.cs.umass.edu/~brenocon/inlp2014/
Brendan O’Connor (http://brenocon.com)

Some material borrowed from Chris Manning,
Jurafsky&Martin, and student exercises
• Review Next Tues. Which?
  • 12-1:30pm ?
  • 5:30-7pm ?
What’s a part-of-speech (POS)?

- Syntactic categories / word classes
  - You could substitute words within a class and have a syntactically valid sentence.
  - Give information how words can combine.

- I saw the **dog**
- I saw the **cat**
- I saw the {**table**, **sky**, **dream**, **school**, **anger**, ...}
POS is an old idea

- Dionysius Thrax of Alexandria (100 BCE): 8 parts of speech
- Common in grammar classes today: noun, verb, adjective, preposition, conjunction, pronoun, interjection
- Many other more fine-grained possibilities

https://www.youtube.com/watch?v=ODGA7ssL-6g&index=1&list=PL6795522EAD6CE2F7
### Open class (lexical) words

<table>
<thead>
<tr>
<th>Nouns</th>
<th>Verbs</th>
<th>Adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proper</strong></td>
<td><strong>Main</strong></td>
<td><em>old</em> <em>older</em> <em>oldest</em></td>
</tr>
<tr>
<td><em>IBM</em> <em>Italy</em></td>
<td><em>see</em> <em>registered</em></td>
<td></td>
</tr>
<tr>
<td><strong>Common</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cat / cats</em> <em>snow</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Closed class (functional)</th>
<th>Modals</th>
<th>Prepositions</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Determiners</strong></td>
<td><em>can</em> <em>had</em></td>
<td><em>to</em> <em>with</em></td>
<td><em>122,312</em> <em>one</em></td>
</tr>
<tr>
<td><em>the</em> <em>some</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conjunctions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>and</em> <em>or</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pronouns</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>he</em> <em>its</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Adverbs</strong></th>
<th><strong>Particles</strong></th>
<th><strong>Interjections</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>slowly</em></td>
<td><em>off</em> <em>up</em></td>
<td><em>Ow</em> <em>Eh</em></td>
</tr>
</tbody>
</table>

### ...more

- *IBM*
- *Italy*
- *cat / cats*
- *snow*
- *see*
- *registered*
- *old* *older* *oldest*
Open vs closed classes

• Closed
  • Determiners: a, an, the
  • Pronouns: he, she, it, they ...
  • Prepositions: on, over, under, of, ...
  • Why “closed”?
  • Many are “grammatical function words.”

• Open
  • Nouns, verbs, adjectives, adverbs
Many tagging standards

- Brown corpus (85 tags)
- Penn Treebank (45 tags) ... the most common one
- Coarse tagsets
  - Petrov et al. “Universal” tagset (12 tags)
    - Motivation: cross-linguistic regularities
      - e.g. adposition: pre- and postpositions
    - For English, collapsing of PTB tags
  - Gimpel et al. tagset for Twitter (25 tags)
    - Motivation: easier for humans to annotate
    - We collapsed PTB, added new things that were necessary for Twitter
Coarse tags, Twitter

D: It's
D: a
A: great
N: show
V: catch
O: it
P: on
D: the
^: Sundance
N: channel
#: #SMSAUDIO
U: http://instagram.com/p/trHejUML3X/

Proper or common? Does it matter?
Grammatical category??
Not really a grammatical category, but perhaps an important word class
Why do we want POS?

• Useful for many syntactic and other NLP tasks.
  • Phrase identification ("chunking")
  • Named entity recognition
  • Full parsing
  • Sentiment
POS patterns: sentiment

- Turney (2002): identify bigram phrases useful for sentiment analysis

Table 1. Patterns of tags for extracting two-word phrases from reviews.

<table>
<thead>
<tr>
<th>First Word</th>
<th>Second Word</th>
<th>Third Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td>NN or NNS</td>
<td>anything</td>
</tr>
<tr>
<td>RB, RBR, or RBS</td>
<td>JJ</td>
<td>not NN nor NNS</td>
</tr>
<tr>
<td>JJ</td>
<td>JJ</td>
<td>not NN nor NNS</td>
</tr>
<tr>
<td>NN or NNS</td>
<td>JJ</td>
<td>not NN nor NNS</td>
</tr>
<tr>
<td>RB, RBR, or RBS</td>
<td>VB, VBD,</td>
<td>anything</td>
</tr>
<tr>
<td></td>
<td>VBN, or VBG</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. An example of the processing of a review that the author has classified as recommended.\(^6\)

<table>
<thead>
<tr>
<th>Extracted Phrase</th>
<th>Part-of-Speech Tags</th>
<th>Semantic Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>online experience</td>
<td>JJ NN</td>
<td>2.253</td>
</tr>
<tr>
<td>low fees</td>
<td>JJ NNS</td>
<td>0.333</td>
</tr>
<tr>
<td>local branch</td>
<td>JJ NN</td>
<td>0.421</td>
</tr>
<tr>
<td>small part</td>
<td>JJ NN</td>
<td>0.053</td>
</tr>
<tr>
<td>online service</td>
<td>JJ NN</td>
<td>2.780</td>
</tr>
<tr>
<td>printable version</td>
<td>JJ NN</td>
<td>-0.705</td>
</tr>
<tr>
<td>direct deposit</td>
<td>JJ NN</td>
<td>1.288</td>
</tr>
<tr>
<td>well other</td>
<td>RB JJ</td>
<td>0.237</td>
</tr>
<tr>
<td>inconveniently</td>
<td>RB VBN</td>
<td>-1.541</td>
</tr>
<tr>
<td>located</td>
<td></td>
<td></td>
</tr>
<tr>
<td>other bank</td>
<td>JJ NN</td>
<td>-0.850</td>
</tr>
<tr>
<td>true service</td>
<td>JJ NN</td>
<td>-0.732</td>
</tr>
</tbody>
</table>

(plus other sentiment stuff)
POS patterns: simple noun phrases

- Quick and dirty noun phrase identification

<table>
<thead>
<tr>
<th>Tag Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A N</td>
<td>linear function</td>
</tr>
<tr>
<td>N N</td>
<td>regression coefficients</td>
</tr>
<tr>
<td>A A N</td>
<td>Gaussian random variable</td>
</tr>
<tr>
<td>A N N</td>
<td>cumulative distribution function</td>
</tr>
<tr>
<td>N A N</td>
<td>mean squared error</td>
</tr>
<tr>
<td>N N N</td>
<td>class probability function</td>
</tr>
<tr>
<td>N P N</td>
<td>degrees of freedom</td>
</tr>
</tbody>
</table>

Table 5.2  Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.
• Exercises
POS Tagging: lexical ambiguity

Can we just use a tag dictionary (one tag per word type)?

<table>
<thead>
<tr>
<th>Types:</th>
<th>WSJ</th>
<th>Brown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous</td>
<td>44,432 (86%)</td>
<td>45,799 (85%)</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>7,025 (14%)</td>
<td>8,050 (15%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tokens:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous</td>
<td>577,421 (45%)</td>
<td>384,349 (33%)</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>711,780 (55%)</td>
<td>786,646 (67%)</td>
</tr>
</tbody>
</table>

Most words types are unambiguous ...

But not so for tokens!

• Ambiguous word types tend to be very common ones.
  • I know **that** he is honest = IN (relativizer)
  • Yes, **that** play was nice = DT (determiner)
  • You can’t go **that** far = RB (adverb)
POS Tagging: baseline

• Baseline: most frequent tag. 92.7% accuracy
  • Simple baselines are very important to run!

• Why so high?
  • Many ambiguous words have a skewed distribution of tags
  • Credit for easy things like punctuation, “the”, “a”, etc.

• Is this actually that high?
  • I get 0.918 accuracy for token tagging
  • ...but, 0.186 whole-sentence accuracy (!)
POS tagging can be hard for humans

- Mrs/NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG
- All/DT we/PRP gola/VBN do/VB is/VBZ go/VB around/IN the/Dt corner/NN
- Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD
Need careful guidelines (and do annotators always follow them?)
PTB POS guidelines, Santorini (1990)

4 Confusing parts of speech

This section discusses parts of speech that are easily confused and gives guidelines on how to tag such cases.

CD or JJ

Number-number combinations should be tagged as adjectives (JJ) if they have the same distribution as adjectives.

EXAMPLES: a 50–3/JJ victory (cf. a handy/JJ victory)

Hyphenated fractions one-half, three-fourths, seven-eighths, one-and-a-half, seven-and-three-eighths should be tagged as adjectives (JJ) when they are prenominal modifiers, but as adverbs (RB) if they could be replaced by double or twice.

EXAMPLES: one-half/JJ cup; cf. a full/JJ cup
one-half/RB the amount; cf. twice/RB the amount; double/RB the amount
Some other lexical ambiguities

- Prepositions versus verb particles
  - turn into/P a monster
  - take out/T the trash
  - check it out/T, what’s going on/T, shout out/T

- this, that -- pronouns versus determiners
  - i just orgasmed over this/O
  - this/D wind is serious

Careful annotator guidelines are necessary to define what to do in many cases.
- http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports

Test:
- turn slowly into a monster
- *take slowly out the trash
Proper nouns

• Common nouns vs. proper nouns on Twitter

Convinced that Monty\(^\text{python}\)\(^\text{V-VBG}\) doing completely straight faced Shakespeare\(^\text{adaption}\)\(^\text{V-VBD}\) be among the most Monty\(^\text{python}\)\(^\text{V-VB}\) things ever

• Names are multiwords. Their tokens are not always nouns. Many people in the exercises didn’t want to do this. Token-level tagging is a weird abstraction here.

3.4 Names

In general, every noun within a proper name should be tagged as a proper noun (‘\(^\text{^}\)’):

• Jesse/\(^\text{^}\) and/\(^\text{\&}\) the/D Rippers/\(^\text{^}\)

• the/D California/\(^\text{^}\) Chamber/\(^\text{^}\) of/P Commerce/\(^\text{^}\)
Are your tokens too big for tags?

• PTB tokenization of clitics leads to easy tagging
  • I’m ==> I/PRP 'm/VBP

• Twitter: is this splitting feasible? Real examples:
  • hes i’m im ill [http://search.twitter.com]
  • Imma bout to do some body shots

• Gimpel et al.’s strategy: introduce compound tags (I'm = PRONOUN+VERB)
Are your tokens too big for tags?

- Other example: highly inflected languages, e.g. Turkish, have case, gender etc. built into the “tag”

1. Yerdeki izin temizlenmesi gerek. iz + Noun+A3sg+Pnon+Gen
   The trace on the floor should be cleaned.

2. Üzerinde parmak izin kalmış iz + Noun+A3sg+P2sg+Nom
   Your finger print is left on (it).

3. İçeri girmek için izin alman gerekiyor. izin + Noun+A3sg+Pnon+Nom
   You need a permission to enter.

- Our approach for Twitter was to simply treat each compound tag as a separate tag. Is this feasible here?
How to build a POS tagger?

• Key sources of information:
  • 1. The word itself
  • 2. Morphology or orthography of word
  • 3. POS tags of surrounding words: syntactic positions