Introduction to Part of Speech Tagging

Alan Ritter

Many slides adapted from Brendan O’Connor Chris Manning
Where are we going with this?

- Text classification: bags of words

- Sequence tagging
  - Parts of Speech
  - Named Entity Recognition
  - Other areas: bioinformatics (gene prediction), etc...
What’s a part-of-speech (POS)?

• Syntax = how words compose to form larger meaning bearing units
• POS = syntactic categories for words
  • You could substitute words within a class and have a syntactically valid sentence
  • Gives information how words combine into larger phrases

• I saw the **dog**
• I saw the **cat**
• I saw the ***
Parts of Speech is an old idea

- Perhaps starting with Aristotle in the West (384–322 BCE), there was the idea of having parts of speech

- School grammar: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection

- Many more fine grained possibilities

https://www.youtube.com/watch?v=ODGA7ssL-6g&index=1&list=PL6795522EAD6CE2F7
Open vs. Closed classes

- Open vs. Closed classes
  - Closed:
    - determiners: *a, an, the*
    - pronouns: *she, he, I*
    - prepositions: *on, under, over, near, by, ...*
    - Why “closed”?*
  - Open:
    - Nouns, Verbs, Adjectives, Adverbs.
Many Tagging Standards

• Penn Treebank (45 tags) ... this is the most common one
• Brown corpus (85 tags)
• Coarse tagsets
  • Universal POS tags (Petrov et. al. https://github.com/slavpetrov/universal-pos-tags)
  • Motivation: cross-linguistic regularities
What are parts of speech useful for?

- Phrase identification (chunking)
- Named entity recognition
- Information Extraction
- Parsing
Quick and Dirty Noun Phrase Identification

**Grammatical structure:** Candidate strings are those multi-word noun phrases that are specified by the regular expression \(((A \mid N)^+ \mid ((A \mid N)^*(NP)^?)((A \mid N)^*\rangle N,

<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.
POS Tagging

• Words often have more than one POS: back
  • The back door = JJ
  • On my back = NN
  • Win the voters back = RB
  • Promised to back the bill = VB

• The POS tagging problem is to determine the POS tag for a particular instance of a word.
POS Tagging

• Input: Plays well with others
• Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS
• Output: Plays/VBZ well/RB with/IN others/NNS
• Uses:
  • Text-to-speech (how do we pronounce “lead”?)
  • Can write regexps like (Det) Adj* N+ over the output for phrases, etc.
  • As input to or to speed up a full parser
  • If you know the tag, you can back off to it in other tasks
POS tagging performance

• How many tags are correct? (Tag accuracy)
  • About 97% currently
  • But baseline is already 90%
    • Baseline is performance of stupidest possible method
      • Tag every word with its most frequent tag
      • Tag unknown words as nouns
  • Partly easy because
    • Many words are unambiguous
    • You get points for them (the, a, etc.) and for punctuation marks!
Deciding on the correct part of speech can be difficult even for people

• Mrs/Shaefer/never/got/around/to/joining

• All/we/gotta/do/is/go/around/the/corner

• Chateau/Petrus/costs/around/250
How difficult is POS tagging?

• About 11% of the word types in the Brown corpus are ambiguous with regard to part of speech
• But they tend to be very common words. E.g., *that*
  • I know *that* he is honest = IN
  • Yes, *that* play was nice = DT
  • You can’t go *that* far = RB
• 40% of the word tokens are ambiguous
It’s hard for people too!

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