Part of Speech Tagging

FSNLP, chapters 9 and 10

Christopher Manning and Hinrich Schütze
© 1999–2003

The problem of POS ambiguity

- Structures for: Fed raises interest rates 0.5% in effort to control inflation (NYT headline 17 May 2000)

Part-of-speech ambiguities

<table>
<thead>
<tr>
<th>VB</th>
<th>VBP</th>
<th>VBZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>NNS</td>
<td>NN</td>
</tr>
</tbody>
</table>

Fed raises interest rates 0.5% in effort to control inflation

Part-of-speech examples

<table>
<thead>
<tr>
<th>NN</th>
<th>noun</th>
<th>baby, toy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VB</td>
<td>verb</td>
<td>see, kiss</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>tall, grateful, alleged</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>quickly, frankly, ...</td>
</tr>
<tr>
<td>IN</td>
<td>preposition</td>
<td>in, on, near</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>the, a, that</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun</td>
<td>who, what, which, ...</td>
</tr>
<tr>
<td>CC</td>
<td>conjunction</td>
<td>and, or</td>
</tr>
</tbody>
</table>

POS ambiguity

- 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.

Why should we care?

- The first statistical NLP task
- Been done to death by different methods
- Easy to evaluate (how many tags are correct?)
- Canonical sequence (finite-state model) task
- Can be done well with methods that look at local context
- Though should $\neg$really $\neg$ do it by parsing!
- Fast linear task of considerable value
The task of part of speech tagging

- A lightweight (usually linear time) processing task, which can usefully empower other applications:
  - Knowing how to pronounce a word: récord [noun] vs. recórd [verb]; lead as noun vs. verb
  - Matching small phrasal chunks or particular word class patterns for tasks such as information retrieval, information extraction or terminology acquisition (collocation extraction). E.g., just matching nouns, compound nouns, and adjective noun patterns:
    ▶ {A|N}^n N
  - POS information can be used to lemmatize a word correctly (i.e., to remove inflections):
    ▶ saw [n] → saw; saw [v] → see

Part of speech tagging

Information sources:

- Sequence of words: syntagmatic information
  - Surprisingly weak information source
  - Many words have various parts of speech – cf. the example above
- Frequency of use of words
  - Surprisingly effective: gets 90+% performance by itself (for English)*
    ▶ This acts as a baseline for performance

Hidden Markov Models – POS example

- Top row is unobserved states, interpreted as POS tags
- Bottom row is observed output observations
- We normally do supervised training, and then (Bayesian network style) inference to decide POS tags

*Even up to 93.7%, based on the results of Toutanova et al. (2003).