Sequence Prediction and Part-of-speech Tagging

Instructor: Yoav Artzi
Overview

• POS Tagging: the problem
• Hidden Markov Models (HMM)
  – Supervised Learning
  – Inference
    • The Viterbi algorithm
• Feature-rich models
  – Maximum-entropy Markov Models
  – Perceptron
  – Conditional Random Fields
Parts of Speech

Open class (lexical) words

Nouns
- Proper
  - IBM
  - Italy
- Common
  - cat / cats
  - snow

Verbs
- Main
  - see
  - registered

Adjectives
- old
- older
- oldest

Adverbs
- slowly

Closed class (functional)

Determiners
- the
- some

Conjunctions
- and
- or

Pronouns
- he
- its

Modals
- can
- had

Prepositions
- to
- with

Particles
- off
- up

Interjections
- Ow
- Eh

Numbers
- 122,312
- one

… more

… more
POS Tagging

• Words often have more than one POS: \textit{back}
  – The \textit{back} door = JJ
  – On my \textit{back} = NN
  – Win the voters \textit{back} = RB
  – Promised to \textit{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 regular expressions 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
Penn TreeBank Tagset

- Possible tags: 45
- Tagging guidelines: 36 pages
- Newswire text
Penn TreeBank Tagset

• How accurate are taggers? (Tag accuracy)
  – About 97% currently
  – But baseline is already 90%
    • Baseline is performance of simplest 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!
  – Upperbound: probably 2% annotation errors
Hard Cases are Hard

• 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.
The Tagset

• Wait, do we really need all these tags?
• What about other languages?
  – Each language has its own tagset
## Tagsets in Different Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Source</th>
<th># Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>PADT/CoNLL07 (Hajič et al., 2004)</td>
<td>21</td>
</tr>
<tr>
<td>Basque</td>
<td>Basque3LB/CoNLL07 (Aduriz et al., 2003)</td>
<td>64</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>BTB/CoNLL06 (Simov et al., 2002)</td>
<td>54</td>
</tr>
<tr>
<td>Catalan</td>
<td>CESS-ECE/CoNLL07 (Martí et al., 2007)</td>
<td>54</td>
</tr>
<tr>
<td>Chinese</td>
<td>Penn Chinese Treebank 6.0 (Palmer et al., 2007)</td>
<td>34</td>
</tr>
<tr>
<td>Chinese</td>
<td>Sinica/CoNLL07 (Chen et al., 2003)</td>
<td>294</td>
</tr>
<tr>
<td>Czech</td>
<td>PDT/CoNLL07 (Böhmová et al., 2003)</td>
<td>63</td>
</tr>
<tr>
<td>Danish</td>
<td>DDT/CoNLL06 (Kromann et al., 2003)</td>
<td>25</td>
</tr>
<tr>
<td>Dutch</td>
<td>Alpino/CoNLL06 (Van der Beek et al., 2002)</td>
<td>12</td>
</tr>
<tr>
<td>English</td>
<td>Penn Treebank (Marcus et al., 1993)</td>
<td>45</td>
</tr>
<tr>
<td>French</td>
<td>French Treebank (Abeillé et al., 2003)</td>
<td>30</td>
</tr>
<tr>
<td>German</td>
<td>Tiger/CoNLL06 (Brants et al., 2002)</td>
<td>54</td>
</tr>
<tr>
<td>German</td>
<td>Negra (Skut et al., 1997)</td>
<td>54</td>
</tr>
<tr>
<td>Greek</td>
<td>GDT/CoNLL07 (Prokipidis et al., 2005)</td>
<td>38</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Szeged/CoNLL07 (Csendes et al., 2005)</td>
<td>43</td>
</tr>
<tr>
<td>Italian</td>
<td>ISST/CoNLL07 (Montemagni et al., 2003)</td>
<td>28</td>
</tr>
<tr>
<td>Japanese</td>
<td>Verb mobi l/CoNLL06 (Kawata and Bartels, 2000)</td>
<td>80</td>
</tr>
<tr>
<td>Japanese</td>
<td>Kyoto4.0 (Kurohashi and Nagao, 1997)</td>
<td>42</td>
</tr>
<tr>
<td>Korean</td>
<td>Sejong (<a href="http://www.sejong.or.kr">http://www.sejong.or.kr</a>)</td>
<td>187</td>
</tr>
<tr>
<td>Portuguese</td>
<td>Floresta Sintá(t)ica/CoNLL06 (Afonso et al., 2002)</td>
<td>22</td>
</tr>
<tr>
<td>Russian</td>
<td>SynTagRus-RNC (Boguslavsky et al., 2002)</td>
<td>11</td>
</tr>
<tr>
<td>Slovene</td>
<td>SDT/CoNLL06 (Džeroski et al., 2006)</td>
<td>29</td>
</tr>
<tr>
<td>Spanish</td>
<td>Ancora-Cast3LB/CoNLL06 (Civit and Martí, 2004)</td>
<td>47</td>
</tr>
<tr>
<td>Swedish</td>
<td>Talbanken05/CoNLL06 (Nivre et al., 2006)</td>
<td>41</td>
</tr>
<tr>
<td>Turkish</td>
<td>METU-Sabanci/CoNLL07 (Oflazer et al., 2003)</td>
<td>31</td>
</tr>
</tbody>
</table>

[Petrov et al. 2012]
The Tagset

• Wait, do we really need all these tags?
• What about other languages?
  – Each language has its own tagset
    • But why is this bad?
    • Differences in downstream tasks
    • Harder to do language transfer
Alternative: The Universal Tagset

• 12 tags:
  – NOUN, VERB, ADJ, ADV, PRON, DET, ADP, NUM, CONJ, PRT, ‘.’, and X.
• Deterministic conversion from tagsets in 22 languages.
• Better unsupervised parsing results
• Was used to transfer parsers

[Petrov et al. 2012]
Sources of Information

• What are the main sources of information for POS tagging?
  – Knowledge of neighboring words
    • Bill saw that man yesterday
    • NNP NN DT NN NN
    • VB VB(D) IN VB NN
  – Knowledge of word probabilities
    • *man* is rarely used as a verb….  

• The latter proves the most useful, but the former also helps
Word-level Features

• Can do surprisingly well just looking at a word by itself:
  – Word the: the → DT
  – Lowercased word Importantly: importantly → RB
  – Prefixes unfathomable: un- → JJ
  – Suffixes Importantly: -ly → RB
  – Capitalization Meridian: CAP → NNP
  – Word shapes 35-year: d-x → JJ