Sequence Tagging

• **Today**
  – Part-of-speech tagging
    • Introduction

Part of speech tagging

“There are 10 parts of speech, and they are all troublesome.”
- *Mark Twain*

• POS tags are also known as word classes, morphological classes, or lexical tags.

• Typically much larger than Twain’s 10:
  – Penn Treebank: 45
  – Brown corpus: 87
  – C7 tagset: 146

Part of speech tagging

• **Assign the correct part of speech (word class) to each word/token in a document**
  “The/DT planet/NN Jupiter/NNP and/CC its/PPS moons/NNS are/VBP in/IN effect/NN a/DT mini-solar/JJ system/NN ./, and/CC Jupiter/NNP itself/PRP is/VBZ often/RB called/VBN a/DT star/NN that/IN never/RB caught/VBN fire/NN ./.”

• **Needed as an initial processing step for a number of language technology applications**
  – Answer extraction in Question Answering systems
  – Base step in identifying syntactic phrases for IR systems
  – Critical for word-sense disambiguation
  – Information extraction
  – …

Why is p-o-s tagging hard?

• **Ambiguity**
  – He will *race/VB* the car.
  – When will the *race/NOUN* end?
  – The boat *floated/VBD*.
  – The boat *floated/VBD* down the river.
  – The boat *floated/VBN* down the river sank.

• **Average of ~2 parts of speech for each word**

• The number of tags used by different systems varies a lot. Some systems use < 20 tags, while others use > 400.
Hard for Humans

- **particle vs. preposition**
  - He talked *over* the deal.
  - He talked *over* the telephone.
- **past tense vs. past participle**
  - The horse *walked* past the barn.
  - The horse *walked* past the barn fell.
- **noun vs. adjective?**
  - The *executive* decision.
- **noun vs. present participle**
  - *Fishing* can be fun.

To obtain gold standards for evaluation, annotators rely on a set of tagging guidelines.

From Ralph Grishman, NYU

---

Among easiest of NLP problems

- **State-of-the-art methods achieve ~97% accuracy.**
- **Simple heuristics can go a long way.**
  - ~90% accuracy just by choosing the most frequent tag for a word (MLE)
  - To improve reliability: *need to use some of the local context.*
- **But defining the rules for special cases can be time-consuming, difficult, and prone to errors and omissions**

---

Penn Treebank Tagset

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordination and, but, or</td>
<td>NVM Symbol +, &amp;</td>
<td>CD</td>
<td>Cardinal number</td>
<td>TO “to”</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td>one, two, three</td>
<td>DT</td>
<td>Determiner</td>
<td>at the</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>there</td>
<td>EX</td>
<td>Existential “there”</td>
<td>there</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
<td>meta culpa</td>
<td>IN</td>
<td>Preposition/subconj</td>
<td>of, in, by</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>yellow</td>
<td>JJR</td>
<td>Adj., comparative</td>
<td>bigger</td>
</tr>
<tr>
<td>JJR</td>
<td>Adj., comparative</td>
<td>bigger</td>
<td>JJR</td>
<td>Adj., comparative</td>
<td>smaller</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
<td>1, 2, One</td>
<td>MD</td>
<td>Modal</td>
<td>can, should</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td>can, should</td>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td>Huma</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td>Huma</td>
<td>NNS</td>
<td>Noun plural</td>
<td>HumaX</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>IBM</td>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>Cardinals</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td>’s</td>
<td>PDT</td>
<td>Predeterminer</td>
<td>all, both</td>
</tr>
<tr>
<td>PP</td>
<td>Possessive pronoun</td>
<td>’s, your, he’s</td>
<td>RB</td>
<td>Adverb</td>
<td>quickly, never</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td>quickly, never</td>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>faster</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>faster</td>
<td>RBS</td>
<td>Adverb, superlative</td>
<td>fastest</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
<td>up, off</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Approaches**

  1. **rule-based**: involve a large database of hand-written disambiguation rules, e.g. that specify that an ambiguous word is a noun rather than a verb if it follows a determiner.

  - HMM tagger

  2. **probabilistic**: resolve tagging ambiguities by using a training corpus to compute the probability of a given word having a given tag in a given context.

    - HMM tagger

  3. **hybrid corpus-rule-based**: E.g. transformation-based tagger (Brill tagger); learns symbolic rules based on a corpus.

  4. **ensemble methods**: combine the results of multiple taggers.