NLP
Introduction to NLP

Part of Speech Tagging
The POS task

• Example
  – Bahrainis vote in second round of parliamentary election

• Jabberwocky (by Lewis Carroll, 1872)
  `Twas brillig, and the slithy toves
  Did gyre and gimble in the wabe:
  All mimsy were the borogoves,
  And the mome raths outgrabe.
Parts of speech

• Open class:
  – nouns, non-modal verbs, adjectives, adverbs

• Closed class:
  – prepositions, modal verbs, conjunctions, particles, determiners, pronouns
# Penn Treebank tagset (1/2)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordinating conjunction</td>
<td>and</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>1</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>the</td>
</tr>
<tr>
<td>EX</td>
<td>existential there</td>
<td>there is</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>d'oeuvre</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/subordinating conjunction</td>
<td>in, of, like</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>green</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative</td>
<td>greener</td>
</tr>
<tr>
<td>JJS</td>
<td>adjective, superlative</td>
<td>greenest</td>
</tr>
<tr>
<td>LS</td>
<td>list marker</td>
<td>1)</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>could, will</td>
</tr>
<tr>
<td>NN</td>
<td>noun, singular or mass</td>
<td>table</td>
</tr>
<tr>
<td>NNS</td>
<td>noun plural</td>
<td>tables</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
<td>John</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Vikings</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td>both the boys</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>friend’s</td>
</tr>
</tbody>
</table>

Penn Treebank tagset (1/2)
### Penn Treebank tagset (2/2)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, he, it</td>
</tr>
<tr>
<td>PRP$</td>
<td>possessive pronoun</td>
<td>my, his</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>however, usually, naturally, here, good</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>better</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td>best</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>give up</td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
<td>to go, to him</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
<td>uhhuhuhhh</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
<td>take</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
<td>took</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund/present participle</td>
<td>taking</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle</td>
<td>taken</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, sing. present, non-3d</td>
<td>take</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, 3rd person sing. present</td>
<td>takes</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-determiner</td>
<td>which</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun</td>
<td>who, what</td>
</tr>
<tr>
<td>WP$</td>
<td>possessive wh-pronoun</td>
<td>whose</td>
</tr>
<tr>
<td>WRB</td>
<td>wh-abverb</td>
<td>where, when</td>
</tr>
</tbody>
</table>
Universal POS

Alphabetical listing

- **ADJ**: adjective
- **ADP**: adposition
- **ADV**: adverb
- **AUX**: auxiliary verb
- **CONJ**: coordinating conjunction
- **DET**: determiner
- **INTJ**: interjection
- **NOUN**: noun
- **NUM**: numeral
- **PART**: particle
- **PRON**: pronoun
- **PROPN**: proper noun
- **PUNCT**: punctuation
- **SCONJ**: subordinating conjunction
- **SYM**: symbol
- **VERB**: verb
- **X**: other

http://universaldependencies.org/u/pos/
Universal Features

Alphabetical listing

- Animacy: animacy
- Aspect: aspect
- Case: case
- Definite: definiteness or state
- Degree: degree of comparison
- Gender: gender
- Mood: mood
- Negative: whether the word can be or is negated
- NumType: numeral type
- Number: number
- Person: person
- Poss: possessive
- PronType: pronominal type
- Reflex: reflexive
- Tense: tense
- VerbForm: form of verb or deverbalative
- Voice: voice

http://universaldependencies.org/u/feat/
Some Observations

- **Ambiguity**
  - count (noun) vs. count (verb)
  - 11% of all types but 40% of all tokens in the Brown corpus are ambiguous.
- Examples
  - *like* can be tagged as ADP VERB ADJ ADV NOUN
  - *present* can be tagged as ADJ NOUN VERB ADV
# POS Ambiguity

<table>
<thead>
<tr>
<th>Types:</th>
<th>WSJ</th>
<th>Brown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous</td>
<td>44,432</td>
<td>45,799</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>7,025</td>
<td>8,050</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tokens:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous</td>
<td>577,421</td>
<td>384,349</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>711,780</td>
<td>786,646</td>
</tr>
</tbody>
</table>

**Figure 9.2** The amount of tag ambiguity for word types in the Brown and WSJ corpora, from the Treebank-3 (45-tag) tagging. These statistics include punctuation as words, and assume words are kept in their original case.
Some Observations

• More examples:
  – transport, object, discount, address
  – content

• French pronunciation:
  – est, président, fils

• Three main techniques:
  – rule-based
  – machine learning (e.g., conditional random fields, maximum entropy Markov models)
  – transformation-based

• Useful for parsing, translation, text to speech, word sense disambiguation, etc.
Example

- Bethlehem/NNP Steel/NNP Corp./NNP,/, hammered/VBN by/IN higher/JJR costs/NNS
- Bethlehem/NNP Steel/NNP Corp./NNP,/, hammered/VBN by/IN higher/JJR costs/VBZ
Classifer-based POS Tagging

• A baseline method would be to use a classifier to map each individual word into a likely POS tag
  – Why is this method unlikely to work well?
Sources of Information

- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/VBN by/IN higher/JJR costs/NNS
- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/VBN by/IN higher/JJR costs/VBZ

- Knowledge about individual words
  - lexical information
  - spelling (–or)
  - capitalization (IBM)

- Knowledge about neighboring words
Evaluation

• **Baseline**
  – tag each word with its most likely tag
  – tag each OOV word as a noun.
  – around 90%

• **Current accuracy**
  – around 97% for English
  – compared to 98% human performance
Rule-based POS tagging

- Use dictionary or finite-state transducers to find all possible parts of speech
- Use disambiguation rules
  - e.g., ART+V
- Hundreds of constraints need to be designed manually
La teneur en uranium des rivières, bien que délicate à calculer, est moyenne en uranium des rivières, bien que délicate à calculer, est moyenne en uranium des rivières, bien que délicate à calculer, est moyenne.
Sample Rules

• BS3 BI1
  – A BS3 (3rd person subject personal pronoun) cannot be followed by a BI1 (1st person indirect personal pronoun).
  – In the example: “il nous faut” (= “we need”) – “il” has the tag BS3MS and “nous” has the tags [BD1P BI1P BJ1P BR1P BS1P].
  – The negative constraint “BS3 BI1” rules out “BI1P”, and thus leaves only 4 alternatives for the word “nous”.

• N K
  – The tag N (noun) cannot be followed by a tag K (interrogative pronoun); an example in the test corpus would be: “... fleuve qui ...” (...river that...).
  – Since “qui” can be tagged both as an “E” (relative pronoun) and a “K” (interrogative pronoun), the “E” will be chosen by the tagger since an interrogative pronoun cannot follow a noun (“N”).

• R V
  – A word tagged with R (article) cannot be followed by a word tagged with V (verb): for example “I' appelle” (calls him/her).
  – The word “appelle” can only be a verb, but “I’” can be either an article or a personal pronoun. Thus, the rule will eliminate the article tag, giving preference to the pronoun.
NLP