Part-of-speech (POS) Tagging

- "Classes" of words
- 8 parts of speech: noun, verb, pronoun, preposition, adverb, conjunction, participle, article
  - Verbs are actions
  - Adjectives are properties
  - Nouns are things
  - Mad Libs??

Why do POS tagging?

- One of the most basic NLP tasks
  - Nicely illustrates principles of statistical NLP
  - Useful for higher-level analysis
  - Needed for syntactic analysis
  - Needed for semantic analysis
  - Sample applications that require POS tagging
    - Machine translation
    - Information extraction
    - Lots more…

Why is it hard?

- Not only a lexical problem
  - Remember ambiguity?
  - Better modeled as sequence labeling problem
    - Need to take into account context!
How do we define POS?

• By meaning
  - Verbs are actions
  - Adjectives are properties
  - Nouns are things
• By the syntactic environment
  - What occurs nearby?
  - What does it act as?
• By what morphological processes affect it
  - What affixes does it take?
• Combination of the above

Unreliable! Think back to the comic!

Parts of Speech

• Open class
  - Impossible to completely enumerate
  - New words continuously being invented, borrowed, etc.
• Closed class
  - Closed, fixed membership
  - Reasonably easy to enumerate
  - Generally, short function words that “structure” sentences

Open Class POS

• Four major open classes in English
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs
• All languages have nouns and verbs... but may not have the other two

Nouns

• Open class
  - New inventions all the time: muggle, webinar, ...
• Semantics:
  - Generally, words for people, places, things
  - But not always (bandwidth, energy, …)
• Syntactic environment:
  - Occurring with determiners
  - Pluralizable, possessivizable
• Other characteristics:
  - Mass vs. count nouns

Verbs

• Open class
  - New inventions all the time: google, tweet, …
• Semantics:
  - Generally, denote actions, processes, etc.
• Syntactic environment:
  - Intransitive, transitive, ditransitive
  - Alternations
• Other characteristics:
  - Main vs. auxiliary verbs
  - Gerunds (verbs behaving like nouns)
  - Participles (verbs behaving like adjectives)

Adjectives and Adverbs

• Adjectives
  - Generally modify nouns, e.g., tall girl
• Adverbs
  - A semantic and formal potpourri…
  - Sometimes modify verbs, e.g., sang beautifully
  - Sometimes modify adjectives, e.g., extremely hot
Closed Class POS

• Prepositions
  • In English, occurring before noun phrases
  • Specifying some type of relation (spatial, temporal, …)
  • Examples: on the shelf, before noon
• Particles
  • Resembles a preposition, but used with a verb ("phrasal verbs")
  • Examples: find out, turn over, go on

Particle vs. Prepositions

He came by the office in a hurry  (by = preposition)
He came by his fortune honestly  (by = particle)
We ran up the phone bill  (up = particle)
We ran up the small hill  (up = preposition)
He lived down the block  (down = preposition)
He never lived down the nicknames  (down = particle)

More Closed Class POS

• Determiners
  • Establish reference for a noun
  • Examples: a, an, the (the articles), that, this, many, such, …
• Pronouns
  • Refer to person or entities: he, she, it
  • Possessive pronouns: his, her, its
  • Wh-pronouns: what, who

Closed Class POS: Conjunctions

• Coordinating conjunctions
  • Join two elements of "equal status"
  • Examples: cats and dogs, salad or soup
• Subordinating conjunctions
  • Join two elements of "unequal status"
  • Examples: We'll leave after you finish eating. While I was waiting in line, I saw my friend.
  • Complementizers are a special case: I think that you should finish your assignment

POS Tagging: What's the task?

• Process of assigning part-of-speech tags to words
• But what tags are we going to assign?
  • Coarse grained: noun, verb, adjective, adverb, …
  • Fine grained: (proper, common) noun
  • Even finer-grained: (proper, common) noun ± animate
• Important issues to remember
  • Choice of tags encodes certain distinctions/non-distinctions
  • Tags will differ across languages!
  • For English, Penn Treebank is the most common tagset

Penn Treebank Tagset: 45 Tags
Why is POS tagging hard?

- Not only a lexical problem
  - Remember ambiguity?
- Better modeled as sequence labeling problem
  - Need to take into account context!

Why is it hard?*

<table>
<thead>
<tr>
<th></th>
<th>87-tag Original Brown</th>
<th>45-tag Treebank Brown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous (1 tag)</td>
<td>44,019</td>
<td>38,857</td>
</tr>
<tr>
<td>Ambiguous (2-7 tags)</td>
<td>5,490</td>
<td>8,844</td>
</tr>
</tbody>
</table>

Details:

- 2 tags: 4,997
- 3 tags: 411
- 4 tags: 91
- 5 tags: 90
- 6 tags: 2 (well, beat)
- 7 tags: 2 (still, down)
- 8 tags: 4 (‘s, half, back, a)
- 9 tags: 3 (that, more, in)

Automatic POS Tagging

- Rule-based POS tagging
- Transformation-based learning for POS tagging
- Hidden Markov Models (next week)
- Maximum Entropy Models (CMSC 773)
- Conditional Random Fields (CMSC 773)

Rule-Based POS Tagging

- Dates back to the 1960’s
- Combination of lexicon + hand crafted rules
  - Example: EngCG (English Constraint Grammar)

EngCG Architecture

<table>
<thead>
<tr>
<th>w_1</th>
<th>w_2</th>
<th>...</th>
<th>w_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon Lookup</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sentence:

- 56,000 entries
- 3,744 rules

Disambiguation using Constraints

Overgenerated tags (w_1, w_2, ..., w_n)

Final tags

EngCG: Sample Lexical Entries

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Additional POS features</th>
</tr>
</thead>
<tbody>
<tr>
<td>smaller</td>
<td>ADJ</td>
<td>COMPARATIVE</td>
</tr>
<tr>
<td>first</td>
<td>ADV</td>
<td>SUPERLATIVE</td>
</tr>
<tr>
<td>that</td>
<td>DET</td>
<td>CENTRAL, DEMONSTRATIVE, SG</td>
</tr>
<tr>
<td>all</td>
<td>DET</td>
<td>PREDETERMINER, SOPL, QUANTIFIER</td>
</tr>
<tr>
<td>dog’s</td>
<td>N</td>
<td>GENITIVE, SG</td>
</tr>
<tr>
<td>furniture</td>
<td>N</td>
<td>NOMINATIVE, SG, NOINDEF, DETERMINER</td>
</tr>
<tr>
<td>our third</td>
<td>NUM</td>
<td>SG</td>
</tr>
<tr>
<td>she</td>
<td>PRON</td>
<td>PERSONAL, FEMININE, NOMINATIVE, SG</td>
</tr>
<tr>
<td>show</td>
<td>V</td>
<td>PRESENT, SG, VFIN</td>
</tr>
<tr>
<td>show</td>
<td>N</td>
<td>NOMINATIVE, SG</td>
</tr>
<tr>
<td>shown</td>
<td>PCP</td>
<td>SV</td>
</tr>
<tr>
<td>occurred</td>
<td>PCP</td>
<td>SV</td>
</tr>
<tr>
<td>occurred</td>
<td>V</td>
<td>PS8T, VFIN, SV</td>
</tr>
</tbody>
</table>
EngCG: Constraint Rule Application

Example Sentence: Newman had originally practiced that ...

```
Newman
NEWMAN N NOM SG PROPER
had
HAVE <SVO> V PAST VFIN
originally
ORIGINAL ADV
practiced
PRACTICE <SVO> <SV> V PAST VFIN
that
ADV
PRON DEM SG
DET CENTRAL DEM SG
CS
```

ADVERBIAL-THAT Rule
Given input: that
if
(+1 A/ADV/QUANT);
(+2 SENT-LIM);
(NOT -1 SVOC/A);
then eliminate non-ADV tags
else eliminate ADV tag

disambiguation constraint

overgenerated tags

I thought that you...
(subordinating conjunction)

That
determiner)

You can go that far.
(adverb)

EngCG: Evaluation

• Accuracy ~96%*
  
  A lot of effort to write the rules and create the lexicon
  
  Try debugging interaction between thousands of rules!
  
  Recall discussion from the first lecture?
  
  Assume we had a corpus annotated with POS tags
  
  Can we learn POS tagging automatically?

Supervised Machine Learning

• Start with annotated corpus
  
  Desired input/output behavior
  
  Training phase:
    • Represent the training data in some manner
    • Apply learning algorithm to produce a system (tagger)
  
  Testing phase:
    • Apply system to unseen test data
    • Evaluate output

Agenda: Summary

• HW2 – assigned today, due next Thursday (9/29)
  
  Questions, comments, concerns?
  
  Part-of-speech Tagging