Part-of-Speech Tagging

Chapter 8
(8.1-8.4.6)

Outline

- Parts of speech (POS)
- Tagsets
- POS Tagging
  - Rule-based tagging
  - Probabilistic (HMM) tagging
Garden Path Sentences

- The old dog the footsteps of the young

Parts of Speech

- Traditional parts of speech
  - Noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction, etc
  - Called: parts-of-speech, lexical categories, word classes, morphological classes, lexical tags...
  - Lots of debate within linguistics about the number, nature, and universality of these
    - We’ll completely ignore this debate.
Parts of Speech

- Traditional parts of speech
  - ~ 8 of them

POS examples

- **N** noun  *chair, bandwidth, pacing*
- **V** verb  *study, debate, munch*
- **ADJ** adjective  *purple, tall, ridiculous*
- **ADV** adverb  *unfortunately, slowly*
- **P** preposition  *of, by, to*
- **PRO** pronoun  *I, me, mine*
- **DET** determiner  *the, a, that, those*
POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a collection.

```
the
koala
put
the
keys
on
the
table
```

```
the
DET
koala
put
the
keys
on
the
table
```

POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a collection.

<table>
<thead>
<tr>
<th>WORD</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>koala</td>
<td>N</td>
</tr>
<tr>
<td>put</td>
<td>V</td>
</tr>
<tr>
<td>the</td>
<td></td>
</tr>
<tr>
<td>keys</td>
<td></td>
</tr>
<tr>
<td>on</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
</tr>
<tr>
<td>table</td>
<td></td>
</tr>
</tbody>
</table>

1/21/2020 Speech and Language Processing - Jurafsky and Martin

9
POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a collection.

<table>
<thead>
<tr>
<th>WORD</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>koala</td>
<td>N</td>
</tr>
<tr>
<td>put</td>
<td>V</td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>keys</td>
<td>N</td>
</tr>
<tr>
<td>on</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
</tr>
<tr>
<td>table</td>
<td></td>
</tr>
</tbody>
</table>
### POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a collection.

<table>
<thead>
<tr>
<th>WORD</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>koala</td>
<td>N</td>
</tr>
<tr>
<td>put</td>
<td>V</td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>keys</td>
<td>N</td>
</tr>
<tr>
<td>on</td>
<td>P</td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>table</td>
<td></td>
</tr>
</tbody>
</table>

1/21/2020 Speech and Language Processing - Jurafsky and Martin

---

### POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a collection.

<table>
<thead>
<tr>
<th>WORD</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>koala</td>
<td>N</td>
</tr>
<tr>
<td>put</td>
<td>V</td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>keys</td>
<td>N</td>
</tr>
<tr>
<td>on</td>
<td>P</td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>table</td>
<td></td>
</tr>
</tbody>
</table>

1/21/2020 Speech and Language Processing - Jurafsky and Martin
POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a collection.

<table>
<thead>
<tr>
<th>WORD</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>koala</td>
<td>N</td>
</tr>
<tr>
<td>put</td>
<td>V</td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>keys</td>
<td>N</td>
</tr>
<tr>
<td>on</td>
<td>P</td>
</tr>
<tr>
<td>the</td>
<td>DET</td>
</tr>
<tr>
<td>table</td>
<td>N</td>
</tr>
</tbody>
</table>

Why is POS Tagging Useful?

- First step of many practical tasks, e.g.
- Speech synthesis (aka text to speech)
  - How to pronounce "lead"?
  - Object object
  - CONTENT CONTENT
- Parsing
  - Need to know if a word is an N or V before you can parse
- Information extraction
  - Finding names, relations, etc.
- Language modeling
  - Backoff
Why is POS Tagging Difficult?

- Words often have more than one POS: 
  - *back*
    - The *back* door = adjective
    - On my *back* =
    - Win the voters *back* =
    - Promised to *back* the bill =
Why is POS Tagging Difficult?

- Words often have more than one POS: 
  *back*
  - The *back* door = adjective
  - On my *back* = noun
  - Win the voters *back* = adverb
  - Promised to *back* the bill =

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

POS tagging performance

- How many tags are correct? (Tag accuracy)
  - About 97% currently
  - But baseline is already 90%
    - Baseline is performance of stupidest 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!
Deciding on the correct part of speech can be difficult even for people

- Mrs/NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG

- All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN

- Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD

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
Open vs. Closed Classes

- **Closed class**: *why?*
  - Determiners: a, an, the
  - Prepositions: of, in, by, ...
  - Auxiliaries: may, can, will had, been, ...
  - Pronouns: I, you, she, mine, his, them, ...
  - Usually *function words* (short common words which play a role in grammar)

- **Open class**: *why?*
  - English has 4: Nouns, Verbs, Adjectives, Adverbs
  - Many languages have these 4, but not all!

Open vs. Closed Classes

- **Closed class**: a small fixed membership
  - Determiners: a, an, the
  - Prepositions: of, in, by, ...
  - Auxiliaries: may, can, will had, been, ...
  - Pronouns: I, you, she, mine, his, them, ...
  - Usually *function words* (short common words which play a role in grammar)

- **Open class**: new ones can be created all the time
  - English has 4: Nouns, Verbs, Adjectives, Adverbs
  - Many languages have these 4, but not all!
Open Class Words

- Nouns
  - Proper nouns (Pittsburgh, Pat Gallagher)
  - English capitalizes these.
  - Common nouns (the rest).
  - Count nouns and mass nouns
    - Count: have plurals, get counted: goat/goats, one goat, two goats
    - Mass: don't get counted (snow, salt, communism) (*two snows)
  
- Adverbs: tend to modify things
  - Unfortunately, John walked home extremely slowly yesterday
  - Directional/locative adverbs (here, home, downhill)
  - Degree adverbs (extremely, very, somewhat)
  - Manner adverbs (slowly, slinkily, delicately)

- Verbs
  - In English, have morphological affixes (eat/eats/eaten)

Closed Class Words

Examples:
  - prepositions: on, under, over, ...
  - particles: up, down, on, off, ...
  - determiners: a, an, the, ...
  - pronouns: she, who, I, ...
  - conjunctions: and, but, or, ...
  - auxiliary verbs: can, may should, ...
  - numerals: one, two, three, third, ...
Prepositions from CELEX

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>540,085</td>
<td>through</td>
<td>14,964</td>
<td>worth</td>
<td>1,563</td>
</tr>
<tr>
<td>in</td>
<td>331,235</td>
<td>after</td>
<td>13,670</td>
<td>toward</td>
<td>1,390</td>
</tr>
<tr>
<td>for</td>
<td>142,421</td>
<td>between</td>
<td>13,275</td>
<td>plus</td>
<td>750</td>
</tr>
<tr>
<td>to</td>
<td>125,691</td>
<td>under</td>
<td>9,525</td>
<td>till</td>
<td>686</td>
</tr>
<tr>
<td>with</td>
<td>124,965</td>
<td>per</td>
<td>6,515</td>
<td>amongst</td>
<td>525</td>
</tr>
<tr>
<td>on</td>
<td>109,129</td>
<td>among</td>
<td>5,090</td>
<td>via</td>
<td>351</td>
</tr>
<tr>
<td>at</td>
<td>100,169</td>
<td>within</td>
<td>5,030</td>
<td>amid</td>
<td>222</td>
</tr>
<tr>
<td>by</td>
<td>77,794</td>
<td>towards</td>
<td>4,700</td>
<td>underneath</td>
<td>164</td>
</tr>
<tr>
<td>from</td>
<td>74,843</td>
<td>above</td>
<td>3,056</td>
<td>versus</td>
<td>113</td>
</tr>
<tr>
<td>about</td>
<td>38,428</td>
<td>near</td>
<td>2,026</td>
<td>amidst</td>
<td>67</td>
</tr>
<tr>
<td>than</td>
<td>20,210</td>
<td>off</td>
<td>1,695</td>
<td>sans</td>
<td>20</td>
</tr>
<tr>
<td>over</td>
<td>18,071</td>
<td>past</td>
<td>1,575</td>
<td>circa</td>
<td>14</td>
</tr>
</tbody>
</table>

POS Tagging
Choosing a Tagset

- There are so many parts of speech, potential distinctions we can draw
- To do POS tagging, we need to choose a standard set of tags to work with
- Could pick very coarse tagsets
- More commonly used set is finer grained, the "Penn TreeBank tagset", 45 tags
- Even more fine-grained tagsets exist
Using the Penn Tagset

- The grand jury commented on a number of other topics.
Using the Penn Tagset

- The grand jury commented on a number of other topics.

Recall POS Tagging Difficulty

- 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.

These examples from Dekang Lin
How Hard is POS Tagging?
Measuring Ambiguity

<table>
<thead>
<tr>
<th>Details</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>
<tr>
<td>2 tags</td>
<td>4,967</td>
<td>6,731</td>
</tr>
<tr>
<td>3 tags</td>
<td>411</td>
<td>1,621</td>
</tr>
<tr>
<td>4 tags</td>
<td>91</td>
<td>357</td>
</tr>
<tr>
<td>5 tags</td>
<td>17</td>
<td>90</td>
</tr>
<tr>
<td>6 tags</td>
<td>2 (well, beat)</td>
<td>32</td>
</tr>
<tr>
<td>7 tags</td>
<td>2 (still, down)</td>
<td>6 (well, set, round, open, fit, down)</td>
</tr>
<tr>
<td>8 tags</td>
<td></td>
<td>4 (’, half, back, a)</td>
</tr>
<tr>
<td>9 tags</td>
<td></td>
<td>3 (that, more, in)</td>
</tr>
</tbody>
</table>

Tagging Whole Sentences with POS is Hard too

- Ambiguous POS contexts
  - E.g., Time flies like an arrow.
- Possible POS assignments
  - Time/[V,N] flies/[V,N] like/[V,Prep] an/Det arrow/N
  - Time/N flies/V like/Prep an/Det arrow/N
  - Time/V flies/N like/Prep an/Det arrow/N
  - Time/N flies/N like/V an/Det arrow/N
  - .....
How Do We Disambiguate POS?

- Many words have only one POS tag (e.g. is, Mary, smallest)
- Others have a single *most likely* tag (e.g. Dog is less used as a V)
- Tags also tend to *co-occur* regularly with other tags (e.g. Det, N)
- In addition to conditional probabilities of words $P(w_1|w_{n-1})$, we can look at POS likelihoods $P(t_1|t_{n-1})$ to disambiguate sentences and to assess sentence likelihoods

More and Better Features ➔ Feature-based tagger

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

- Rough accuracies:
  - Most freq tag: ~90% / ~50%
  - Trigram HMM: ~95% / ~55%
  - Maxent P(t|w): 93.7% / 82.6%
  - Upper bound: ~98% (human)

Most errors on unknown words