CS4705
Part of Speech tagging

Some slides adapted from: Dan Jurafsky, Julia Hirschberg, Jim Martin
Announcements

• Reading for today:
• Reading for next time:
• We are turning to syntax. Today: part-of-speech tagging
• Homework 2 is out
Garden path sentences

- The old dog the footsteps of the young.
- The horse raced past the barn fell.
- The cotton clothing is made of grows in Mississippi.
Garden path sentences

N

• The old dog | the footsteps of the young.

• The horse raced past the barn fell.

• The cotton clothing is made of grows in Mississippi.
Garden path sentences

• The old dog | the footsteps of the young.

• The horse raced past the barn fell.

• The cotton clothing is made of grows in Mississippi.
Garden path sentences

• *The old dog the footsteps of the young.*

  VBD

• *The horse raced past the barn | fell.*

• *The cotton clothing is made of grows in Mississippi.*
Garden path sentences

• *The old dog the footsteps of the young.*

  VBN VBD

• *The horse raced past the barn | fell.*

• *The cotton clothing is made of grows in Mississippi.*
Garden path sentences

• The old dog the footsteps of the young.

• The horse raced past the barn fell.

• The cotton clothing is made of grows in Mississippi.
In the garden path reading of "The cotton clothing is made of grows in Mississippi" "cotton" is a
In the correct reading of "The cotton clothing is made of grows in Mississippi" "cotton" is a

<table>
<thead>
<tr>
<th>Option</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td></td>
</tr>
<tr>
<td>Noun</td>
<td></td>
</tr>
<tr>
<td>Adjective</td>
<td></td>
</tr>
<tr>
<td>Adverb</td>
<td></td>
</tr>
<tr>
<td>None of the above</td>
<td></td>
</tr>
</tbody>
</table>
What is a word class?

• Words that somehow ‘behave’ alike:
  • Appear in similar contexts
  • Perform similar functions in sentences
  • Undergo similar transformations

• 9 (or so) traditional parts of speech
  • Noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction,
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: Definition

• The process of assigning a part-of-speech or lexical class marker to each word in a corpus:

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

N
V
P
DET
```
What is POS tagging good for?

• Is the first step of a vast number of Comp Ling tasks
• Speech synthesis:
  • How to pronounce “lead“?
  • INsult  inSULT
  • OBject  object
  • OVERflow  overFLOW
  • DIScount  disCOUNT
  • CONtent  conTENT
• Parsing
  • Need to know if a word is an N or V before you can parse
• Word prediction in speech recognition
  • Possessive pronouns (my, your, her) followed by nouns
  • Personal pronouns (I, you, he) likely to be followed by verbs
• Machine Translation
If lead is pronounced "led" it is a

Noun

Adjective

Verb

Adverb

None of the above
If "lead" is pronounced as "leed" it is a

<table>
<thead>
<tr>
<th>Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
</tr>
<tr>
<td>Verb</td>
</tr>
<tr>
<td>Adverb</td>
</tr>
<tr>
<td>None of the above</td>
</tr>
</tbody>
</table>
Open and closed class words

• Closed class: a relatively fixed membership
  • 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 all 4, but not all!
  • In Lakhota and possibly Chinese, what English treats as adjectives act more like verbs.
Open class words

- **Nouns**
  - Proper nouns (Columbia University, New York City, Elsbeth Turcan, Metropolitan Transit Center). English capitalizes these.
  - Common nouns (the rest). German capitalizes these.
  - Count nouns and mass nouns
    - Count: have plurals, get counted: goat/goats, one goat, two goats
    - Mass: don’t get counted (fish, salt, communism) (*two fishes)

- **Adverbs:** tend to modify actions or predicates
  - 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)
  - Actions (walk, ate) and states (be, exude)
• Many subclasses, e.g.
  • eats/V ⇒ eat/VB, eat/VBP, eats/VBZ, ate/VBD, eaten/VBN, eating/VBG, ...  
  • Reflect morphological form & syntactic function
How do we decide which words go in which classes?

• **Nouns** denote people, places and things and can be preceded by articles? But...
  
  *My typing is very bad.*
  
  *The Mary loves John.*

• **Verbs** are used to refer to actions, processes, states
  
  • But some are **closed class** and some are **open**
  
  I will have **emailed** everyone by noon.

• **Adverbs** modify actions

• **Is Monday a temporal adverb or a noun?**
Is Monday a temporal adverb or a noun?

Temporal adverb

Noun

Neither
Determining Part-of-Speech

• A blue seat / A child seat: noun or adj?

• Some tests
  • Syntactic
    • A blue seat
    • A very blue seat
    • This seat is blue
    • A child seat
    • *A very child seat
    • *This seat is child
  • Morphological
    • Bluer
    *childer

• Blue is an adjective, but child is a noun
Determining Part-of-Speech

• Preposition or particle?

A. He threw out the garbage.

B. He threw the garbage out the door.

C. He threw the garbage out

D. *He threw the garbage the door out.

• out in A is a particle, in B is a preposition
Closed Class Words

- Idiosyncratic
- Closed class words (Prep, Det, Pron, Conj, Aux, Part, Num) are easier, since we can enumerate them....but
  - Part vs. Prep
    - George eats up his dinner/George eats his dinner up.
    - George eats up the street/*George eats the street up.
  - Articles come in 2 flavors: definite (the) and indefinite (a, an)
POS tagging: Choosing a tagset

• To do POS tagging, need to choose a standard set of tags to work with
• Could pick very coarse tagsets
  • N, V, Adj, Adv.
• Brown Corpus (Francis & Kucera ‘82), 1M words, 87 tags
• [Penn Treebank]: hand-annotated corpus of *Wall Street Journal*, 1M words, 45-46 tags
  • Commonly used
  • set is finer grained,
• Even more fine-grained tagsets exist
# Penn TreeBank POS Tag set

<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>Coordin. Conjunction</td>
<td><em>and, but, or</em></td>
<td>SYM</td>
<td>Symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td><em>one, two, three</em></td>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td><em>a, the</em></td>
<td>UH</td>
<td>Interjection</td>
<td><em>ah, oops</em></td>
</tr>
<tr>
<td>EX</td>
<td>Existential ‘there’</td>
<td><em>there</em></td>
<td>VB</td>
<td>Verb, base form</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
<td><em>mea culpa</em></td>
<td>VBD</td>
<td>Verb, past tense</td>
<td><em>ate</em></td>
</tr>
<tr>
<td>IN</td>
<td>Preposition/sub-conj</td>
<td><em>of, in, by</em></td>
<td>VBG</td>
<td>Verb, gerund</td>
<td><em>eating</em></td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td><em>yellow</em></td>
<td>VBN</td>
<td>Verb, past participle</td>
<td><em>eaten</em></td>
</tr>
<tr>
<td>JJR</td>
<td>Adj., comparative</td>
<td><em>bigger</em></td>
<td>VBP</td>
<td>Verb, non-3sg pres</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>JJS</td>
<td>Adj., superlative</td>
<td><em>wildest</em></td>
<td>VBZ</td>
<td>Verb, 3sg pres</td>
<td><em>eats</em></td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
<td><em>1, 2, One</em></td>
<td>WDT</td>
<td>Wh-determiner</td>
<td><em>which, that</em></td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td><em>can, should</em></td>
<td>WP</td>
<td>Wh-pronoun</td>
<td><em>what, who</em></td>
</tr>
<tr>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td><em>llama</em></td>
<td>WP$</td>
<td>Possessive wh-</td>
<td><em>whose</em></td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
<td><em>llamas</em></td>
<td>WRB</td>
<td>Wh-adverb</td>
<td><em>how, where</em></td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td><em>IBM</em></td>
<td>$</td>
<td>Dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td><em>Carolinas</em></td>
<td>#</td>
<td>Pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td><em>all, both</em></td>
<td>“</td>
<td>Left quote</td>
<td>(‘ or “)</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td><em>’s</em></td>
<td>”</td>
<td>Right quote</td>
<td>(‘ or ”)</td>
</tr>
<tr>
<td>PRP</td>
<td>Personal pronoun</td>
<td><em>I, you, he</em></td>
<td>(</td>
<td>Left parenthesis</td>
<td>([, (, {, &lt;)</td>
</tr>
<tr>
<td>PRP$</td>
<td>Possessive pronoun</td>
<td><em>your, one’s</em></td>
<td>)</td>
<td>Right parenthesis</td>
<td>( ], ), }, &gt;)</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td><em>quickly, never</em></td>
<td>,</td>
<td>Comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td><em>faster</em></td>
<td>.</td>
<td>Sentence-final punctuation</td>
<td>( ! ?)</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
<td><em>fastest</em></td>
<td>:</td>
<td>Mid-sentence punctuation</td>
<td>( : ; ... - -)</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
<td><em>up, off</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Using the UPenn tagset

• *The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.*

• Prepositions and subordinating conjunctions marked IN (“although/IN I/PRP..”)

• Except the preposition/complementizer “to” is just marked “to”.
POS Tagging

• 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 do we assign POS tags to words in a sentence?**

What information do you think we could use to assign POS in the following sentences?

- *Time flies like an arrow.*
- *Time/N flies/V like/Prep an/Det arrow/N*
- *Fruit/N flies/N like/V a/DET banana/N*
- *Fruit/N flies/V like/Prep a/DET banana/N*
- *The/Det flies/N like/V a/DET banana/N*
How hard is POS tagging? Measuring ambiguity

<table>
<thead>
<tr>
<th></th>
<th>Original 87-tag corpus</th>
<th>Treebank 45-tag corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unambiguous (1 tag)</strong></td>
<td>44,019</td>
<td>38,857</td>
</tr>
<tr>
<td><strong>Ambiguous (2–7 tags)</strong></td>
<td>5,490</td>
<td>8844</td>
</tr>
<tr>
<td>Details:</td>
<td></td>
<td></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 (’s, half, back, a)</td>
</tr>
<tr>
<td>9 tags</td>
<td></td>
<td>3 (that, more, in)</td>
</tr>
</tbody>
</table>
Can you think of seven sentences where in each one “well” is used with a different part of speech?

<table>
<thead>
<tr>
<th></th>
<th>Original 87-tag corpus</th>
<th>Treebank 45-tag corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unambiguous (1 tag)</strong></td>
<td>44,019</td>
<td>38,857</td>
</tr>
<tr>
<td><strong>Ambiguous (2–7 tags)</strong></td>
<td>5,490</td>
<td>8844</td>
</tr>
<tr>
<td>Details:</td>
<td></td>
<td></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>1621</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 (’s, half, back, a)</td>
</tr>
<tr>
<td>9 tags</td>
<td></td>
<td>3 (that, more, in)</td>
</tr>
</tbody>
</table>
Potential Sources of Disambiguation

- Many words have only one POS tag (e.g. *is, Mary, very, smallest*)
- Others have a single most likely tag (e.g. *a, dog*)
- But 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
Hidden Markov Model Tagging

• Using an HMM to do POS tagging

• A special case of Bayesian inference

• Related to the “noisy channel” model used in MT, ASR and other applications
POS tagging as a sequence classification task

- We are given a sentence (an “observation” or “sequence of observations”)
  - Secretariat is expected to race tomorrow

- What is the best sequence of tags which corresponds to this sequence of observations?

- Probabilistic view:
  - Consider all possible sequences of tags
  - Choose the tag sequence which is most probable given the observation sequence of n words $w_1...w_n$. 

9/30/19
Getting to HMM

• Out of all sequences of n tags $t_1...t_n$ want the single tag sequence such that $P(t_1...t_n|w_1...w_n)$ is highest.

$$\hat{t}_1^n = \arg\max_{t_1^n} P(t_1^n|w_1^n)$$

• Hat $^\wedge$ means “our estimate of the best one”

• Argmax$_x$ f(x) means “the x such that f(x) is maximized”
Getting to HMM

• This equation is guaranteed to give us the best tag sequence

\[ \hat{t}_1^n = \text{argmax}_{t_1^n} P(t_1^n | w_1^n) \]

• Intuition of Bayesian classification:
  • Use Bayes rule to transform into a set of other probabilities that are easier to compute
Using Bayes Rule

\[ P(x|y) = \frac{P(y|x)P(x)}{P(y)} \]

\[ \hat{t}_1^n = \underset{t_1^n}{\text{argmax}} \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)} \]

\[ \hat{t}_1^n = \underset{t_1^n}{\text{argmax}} P(w_1^n|t_1^n)P(t_1^n) \]
Likelihood and prior

\[ \hat{t}^n_1 = \arg\max_{t^n_1} P(w^n_1|t^n_1) \] 
\[ \times \underbrace{P(t^n_1)}_{\text{prior}} \]

\[ P(w^n_1|t^n_1) \approx \prod_{i=1}^{n} P(w_i|t_i) \]

\[ P(t^n_1) \approx \prod_{i=1}^{n} P(t_i|t_{i-1}) \]

\[ \hat{t}^n_1 = \arg\max_{t^n_1} P(t^n_1|w^n_1) \approx \arg\max_{t^n_1} \prod_{i=1}^{n} P(w_i|t_i) P(t_i|t_{i-1}) \]
Two kinds of probabilities (1)

- Tag transition probabilities \( p(t_i | t_{i-1}) \)
  - Determiners likely to precede adjs and nouns
    - That/DT flight/NN
    - The/DT yellow/JJ hat/NN
    - So we expect \( P(NN | DT) \) and \( P(JJ | DT) \) to be high
    - But \( P(DT | JJ) \) to be:
  - Compute \( P(NN | DT) \) by counting in a labeled corpus:

\[
P(t_i | t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}
\]

\[
P(NN | DT) = \frac{C(DT, NN)}{C(DT)} = \frac{56,509}{116,454} = .49
\]
Two kinds of probabilities (2)

- Word likelihood probabilities $p(w_i | t_i)$
  - VBZ (3sg Pres verb) likely to be “is”
  - Compute $P(is | VBZ)$ by counting in a labeled corpus:
    
    $$P(w_i | t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

    $$P(is | VBZ) = \frac{C(VBZ, is)}{C(VBZ)} = \frac{10,073}{21,627} = .47$$
An Example: the verb “race”

• Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NR

• People/NNS continue/VB to/TO inquire/VB the/DT reason/NN for/IN the/DT race/NN for/IN outer/JJ space/NN

• How do we pick the right tag?
Disambiguating “race”

(a)

Secretariat is expected to race tomorrow

(b)

Secretariat is expected to race tomorrow
• $P(\text{NN} | \text{TO}) = .00047$
• $P(\text{VB} | \text{TO}) = .83$
• $P(\text{race} | \text{NN}) = .00057$
• $P(\text{race} | \text{VB}) = .00012$
• $P(\text{NR} | \text{VB}) = .0027$
• $P(\text{NR} | \text{NN}) = .0012$
• $P(\text{VB} | \text{TO})P(\text{NR} | \text{VB})P(\text{race} | \text{VB}) = .00000027$
• $P(\text{NN} | \text{TO})P(\text{NR} | \text{NN})P(\text{race} | \text{NN}) = .00000000032$
• So we (correctly) choose the verb reading,
Summary

Parts of speech
• What’s POS tagging good for anyhow?
• Tag sets
• Statistics and POS tagging
• Next time:
  • HMM Tagging
Homework 2