

# 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

V

- *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

Verb

Noun

Adjective

Adverb

None of the above

**In the correct reading of "The cotton clothing is made of grows in Mississippi" "cotton" is a**

Verb

Noun

Adjective

Adverb

None of the above

# What is a word class?

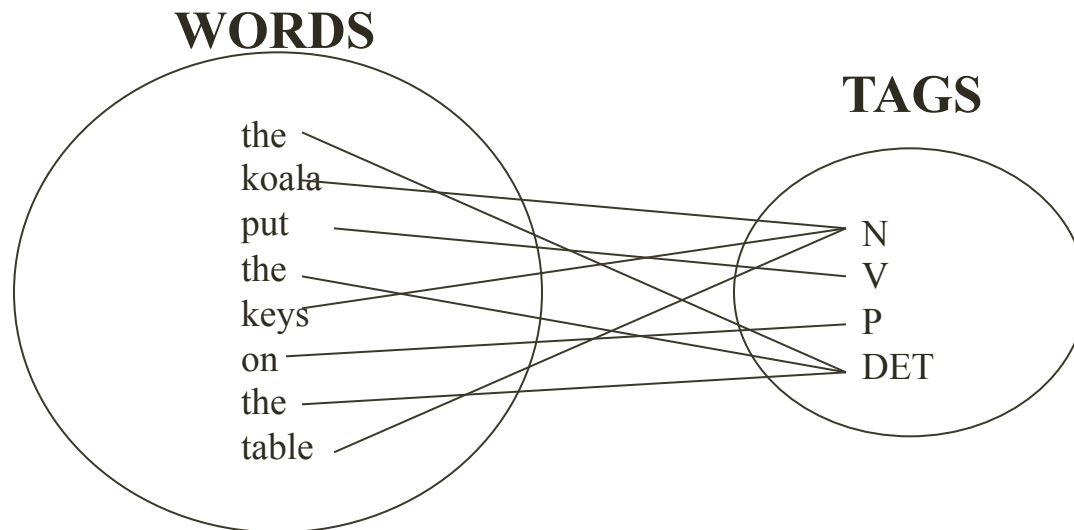
- Words that somehow ‘behave’ alike:
  - Appear in similar contexts
  - Perform similar functions in sentences
  - Undergo similar transformations
- 9 (or so) traditional parts of speech
  - Noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction,

# 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:



# 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

Noun

Adjective

Verb

Adverb

None of the above



# 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  $\Rightarrow$  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 child seat*
    - *A very blue seat*                *\*A very child seat*
    - *This seat is blue*                *\*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

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

# Using the UPenn tagset

- *The/DT grand/JJ jury/NN commmented/  
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/[V,N] flies/[V,N] like/[V,Prep] an/Det arrow/N*
- *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

	Original 87-tag corpus	Trebank 45-tag corpus
<b>Unambiguous (1 tag)</b>	<b>44,019</b>	<b>38,857</b>
<b>Ambiguous (2–7 tags)</b>	<b>5,490</b>	<b>8844</b>
Details:		
2 tags	4,967	6,731
3 tags	411	1621
4 tags	91	357
5 tags	17	90
6 tags	2 ( <i>well, beat</i> )	32
7 tags	2 ( <i>still, down</i> )	6 ( <i>well, set, round, open, fit, down</i> )
8 tags		4 ( <i>'s, half, back, a</i> )
9 tags		3 ( <i>that, more, in</i> )

*Can you think of seven sentences where in each one “well” is used with a different part of speech?*

	Original 87-tag corpus	Treebank 45-tag corpus
<b>Unambiguous (1 tag)</b>	<b>44,019</b>	<b>38,857</b>
<b>Ambiguous (2–7 tags)</b>	<b>5,490</b>	<b>8844</b>
Details:		
2 tags	4,967	6,731
3 tags	411	1621
4 tags	91	357
5 tags	17	90
6 tags	2 ( <i>well, beat</i> )	32
7 tags	2 ( <i>still, down</i> )	6 ( <i>well, set, round, open, fit, down</i> )
8 tags		4 ( <i>'s, half, back, a</i> )
9 tags		3 ( <i>that, more, in</i> )

# 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  $w_1 \dots w_n$ .

# Getting to HMM

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

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n | w_1^n)$$

- Hat  $\hat{\phantom{x}}$  means “our estimate of the best one”
- $\operatorname{Argmax}_x f(x)$  means “the  $x$  such that  $f(x)$  is maximized”

# Getting to HMM

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

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n | w_1^n)$$

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

# Using Bayes Rule

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)}$$

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(w_1^n | t_1^n) P(t_1^n)$$

# Likelihood and prior

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} \overbrace{P(w_1^n | t_1^n)}^{\text{likelihood}} \overbrace{P(t_1^n)}^{\text{prior}}$$

$$P(w_1^n | t_1^n) \approx \prod_{i=1}^n P(w_i | t_i)$$

$$P(t_1^n) \approx \prod_{i=1}^n P(t_i | t_{i-1})$$

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n | w_1^n) \approx \operatorname{argmax}_{t_1^n} \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1})$$

# 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  $P(\text{is} | \text{VBZ})$  by counting in a labeled corpus:

$$P(w_i | t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

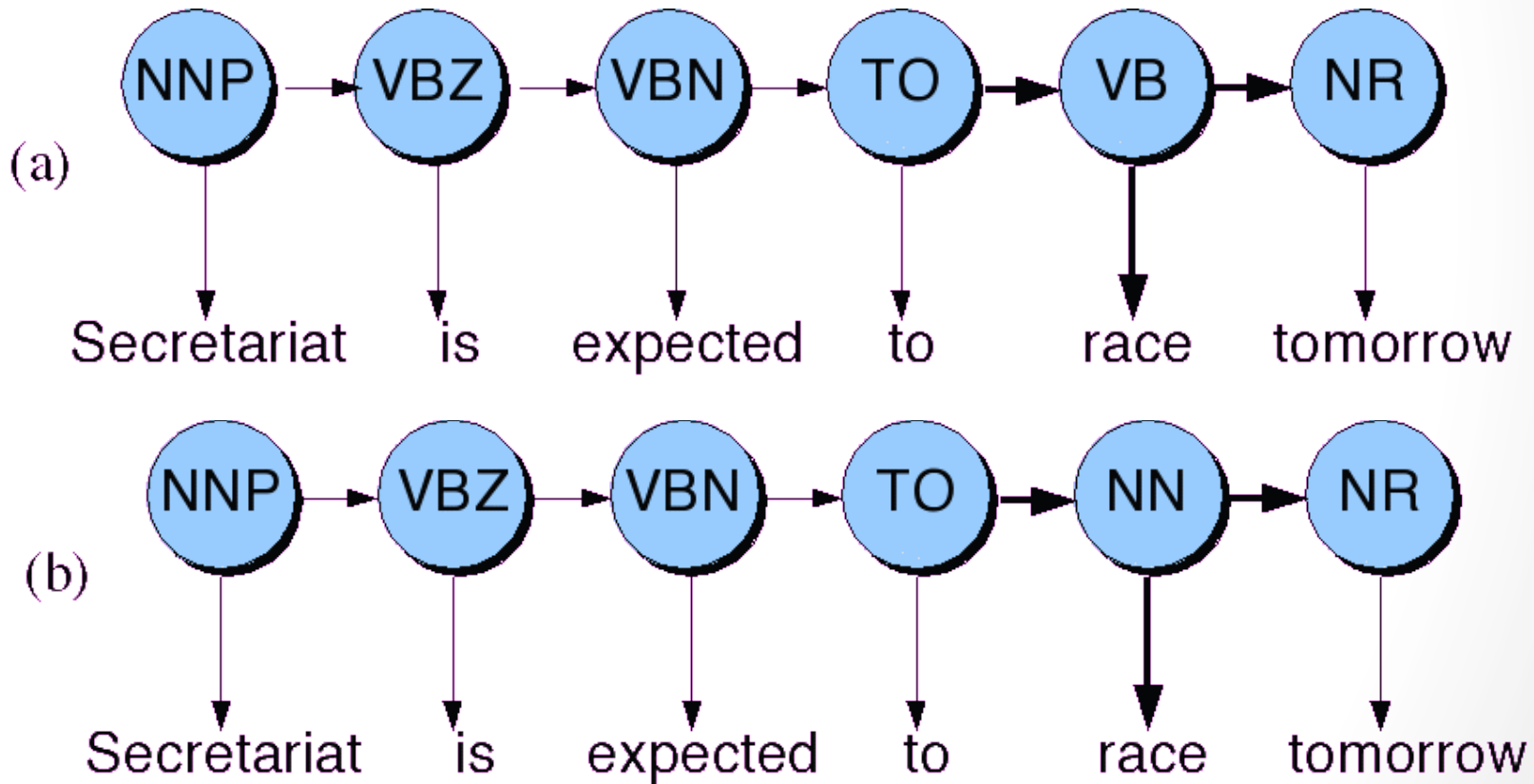
$$P(\text{is} | \text{VBZ}) = \frac{C(\text{VBZ}, \text{is})}{C(\text{VBZ})} = \frac{10,073}{21,627} = .47$$



# 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”



- $P(\text{NN} | \text{TO}) = .00047$
- $P(\text{VB} | \text{TO}) = .83$
- $P(\text{race} | \text{NN}) = .00057$
- $P(\text{race} | \text{VB}) = .00012$
- $P(\text{NR} | \text{VB}) = .0027$
- $P(\text{NR} | \text{NN}) = .0012$
- $P(\text{VB} | \text{TO})P(\text{NR} | \text{VB})P(\text{race} | \text{VB}) = .00000027$
- $P(\text{NN} | \text{TO})P(\text{NR} | \text{NN})P(\text{race} | \text{NN}) = .00000000032$
- 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