CS4705
Part of Speech tagging

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

Announcements and questions

- The accuracy threshold is 75%. Be sure to experiment with designing features, model selection and parameter tuning, and feature selection You will lose two points of the grade for every five points accuracy loss. See the homework pdf for details.

- We have received some questions about hard coding the test and dev file names in hw1.py. This is fine. Just leave the files in the same directory as your code and do not change the filenames.

- Because of student questions, we have changed the arguments of analyze.py to be the pickled model and your vectorized test data. Vectorize the test data for a given model the same as you do for training; this can happen in hw1.py before you use analyze.py. You may want to consider creating functions in hw1.py that prepare data for a specific model. You may use whatever format you want, as long as it is consistent and reasonable (e.g., you can even save it in a text file and read it back out in analyze.py, as long as it is consistent).

- Remember to tag you submission once it is complete; see @155 for details.Threshold for HW1: 75 and above: 10 points of credit, 2 points loss per 5 points of accuracy

- Late days for those who came off the waitlist last week
  - 5 days for HW0 (but don't use them all!)
  - 2-5 days for HW1 (2 for Tuesday… up to … 5 for Friday)

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.*

Response

What is a word class?

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

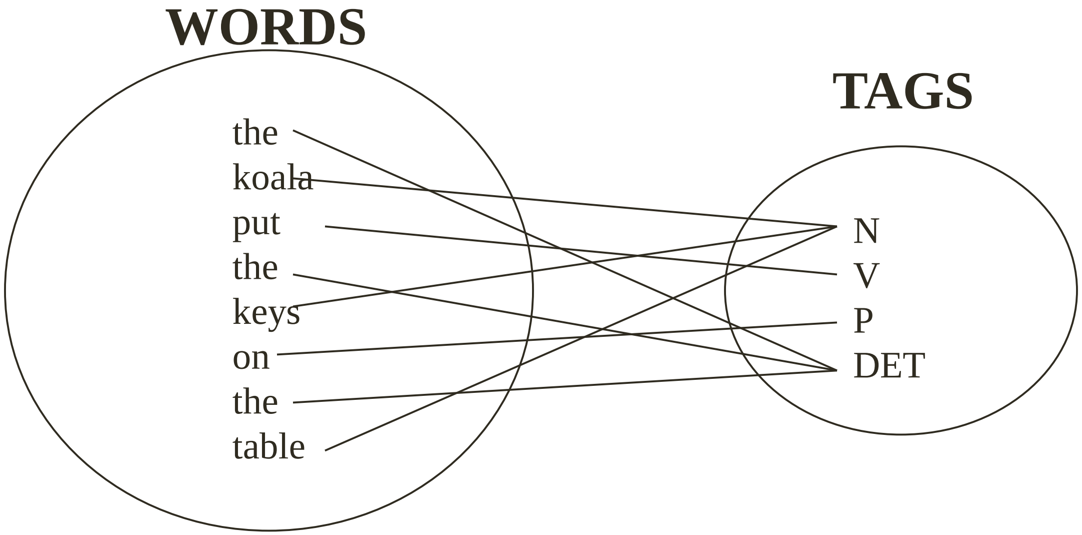
- 9 (or so) traditional parts of speech
  - Noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction,

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

Response

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?*

Response

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

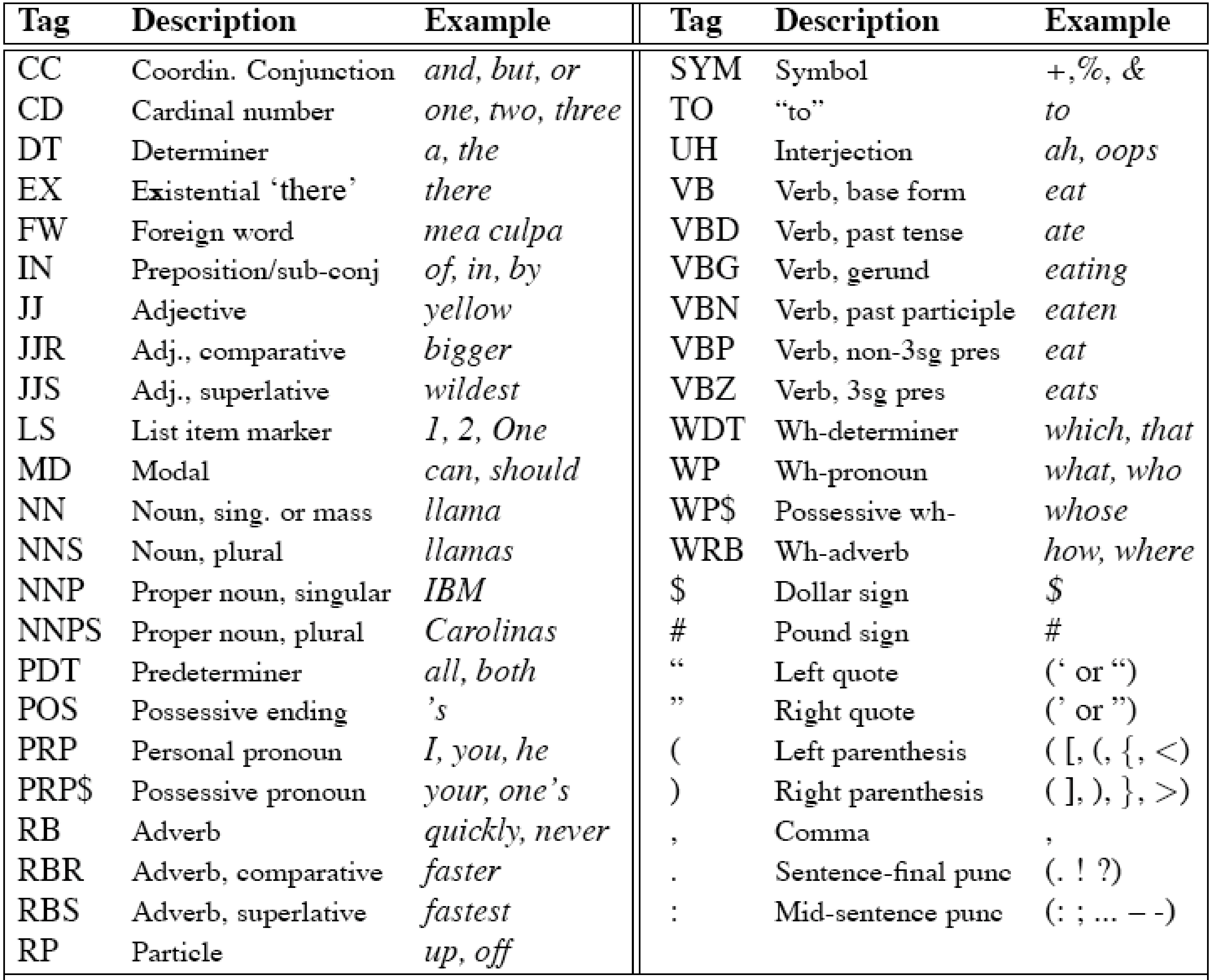
Closed Class Words

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

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

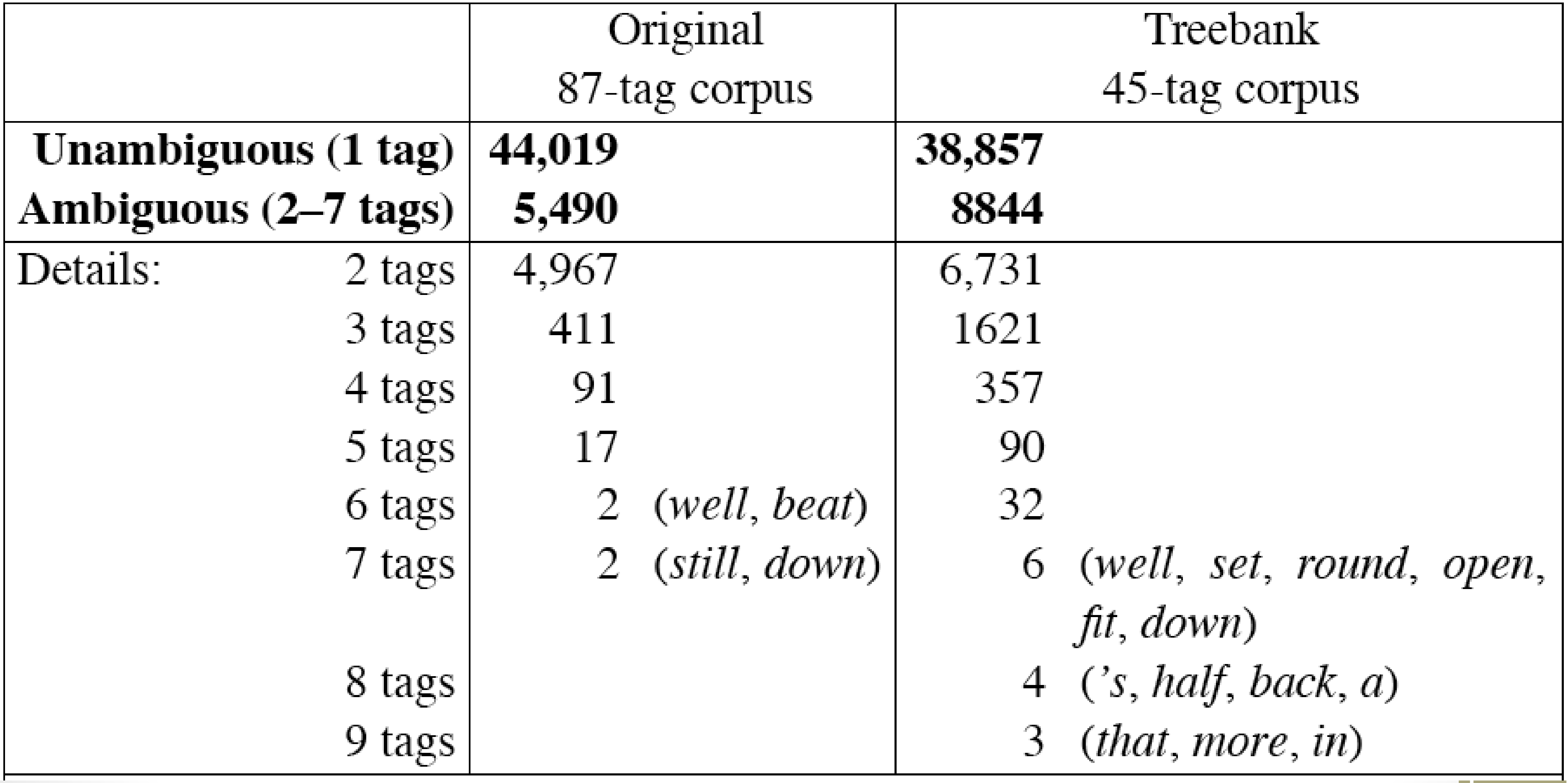
How do we assign POS tags to words in a sentence?

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*

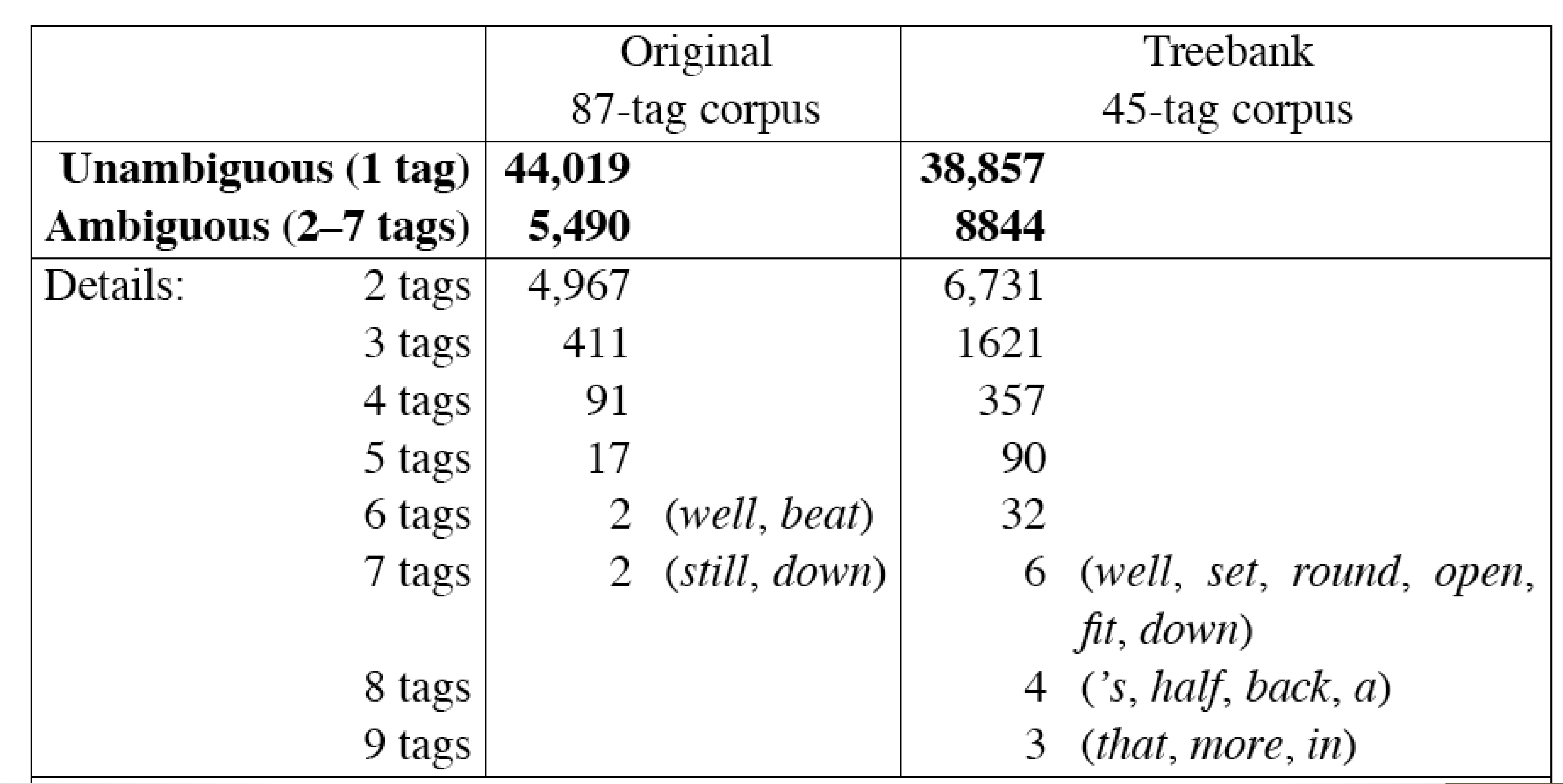
Response

How hard is POS tagging? Measuring ambiguity



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

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

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 w1…wn.

Getting to HMM

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

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

- Hat ^ 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

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

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

Using Bayes Rule

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

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

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

9/18/2017

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(is|VBZ) by counting in a labeled corpus:

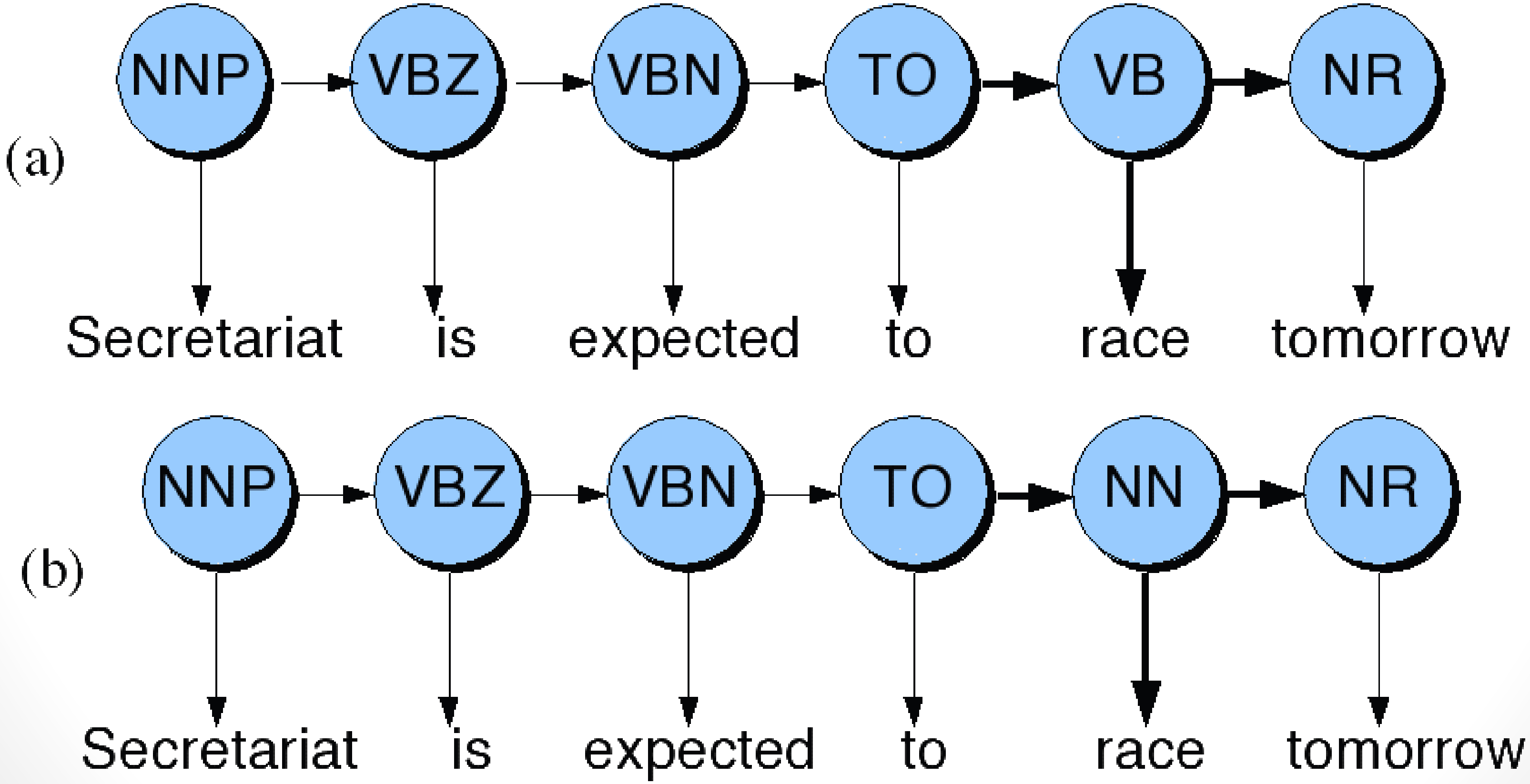
$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

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

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”



- P(NN|TO) = .00047
- P(VB|TO) = .83
- P(race|NN) = .00057
- P(race|VB) = .00012
- P(NR|VB) = .0027
- P(NR|NN) = .0012
- P(VB|TO)P(NR|VB)P(race|VB) = .00000027
- P(NN|TO)P(NR|NN)P(race|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