CZ4045 Natural Language Processing

Part-of-speech Tagging (Chapter 5)

Topics

- Word classes
- Part of speech tagging
- Use HMMs for POS tagging

Word Classes: Parts of Speech

- Traditional parts of speech
  - Noun, verb, adjective, preposition, adverb,
  - Article, interjection, pronoun, conjunction, etc.

- Also Called:
  - parts-of-speech, lexical categories,
  - word classes, morphological classes, lexical tags...

- Examples on next slide

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 sequence.

WORD	tag
the	DET
koala	N
put	V
the	DET
keys	N
on	P
the	DET
table	N

Why is POS Tagging Useful?

- First step of a vast number of practical tasks
- Information extraction
  - Finding names, e.g., people, organization -- N.
- Machine Translation
  - E.g. result/N, result/V -> kyol-kwa/N, kyol-kwa-lul-ne-da/V
- Parsing
  - Helpful to know parts of speech before you start parsing, e.g. subject-verb-object
- Many Others

POS Tagging: Choosing a Tagset

- To do POS tagging, we need to choose a standard set of tags to work with

- Could pick very coarse tagsets (e.g., N, V, Adj, Adv.)

- More commonly used set is the finer grained, “Penn TreeBank tagset”, 45 tags
  - See next slide

- Even more fine-grained tagsets exist

Penn TreeBank POS Tagset

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 Penn 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 "to" is just marked "**TO**".

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*			

POS Tagging

- Words often have more than one POS: ***back***
  - The ***back*** door = JJ (adj)
  - On my ***back*** = NN
  - Win the voters ***back*** = RB (adv)
  - Promised to ***back*** the bill = VB (verb, base form)

- The POS tagging problem is to determine the POS tag for a particular **instance** of a word.

How Hard is POS Tagging? Measuring Ambiguity

- Number of word types with different levels of POS ambiguity from the Brown corpus

		87-tag Original Brown		45-tag Treebank Brown	
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*)

Many **ambiguous** words appear **frequently** in corpus

Two Methods for POS Tagging

- Rule-based tagging
  - (ENGTWOL; Section 5.4)

- Stochastic
  - Probabilistic sequence models
  - HMM (Hidden Markov Model) tagging
    - Using an HMM to do POS tagging is a special case of ***Bayesian inference***

  - MEMMs (Maximum Entropy Markov Models)

POS Tagging as Sequence Classification

- **Input**: We are given a sentence (an "observation" or "sequence of observations")
  - Secretariat is expected to race tomorrow
- **Output**: What is the best sequence of tags that corresponds to this sequence of observations?

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

- Probabilistic view:
  - Consider all possible sequences of tags
  - Out of this universe of sequences, choose **the tag sequence** which is **most probable** given the observation sequence of n words $w_1 \ldots w_n$.

Road to HMMs

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

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

- Hat ^ means "our estimate of the best one"
- $arg \max_x f(x)$ means "the x such that $f(x)$ is maximized"

Road to HMMs

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

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

- But how to compute this value?

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

Using Bayes Rule

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

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

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

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

Likelihood and Prior

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

The**/DT** yellow**/JJ** hat**/NN**

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

The probability of a word appearing depends only on **its own POS tag** P (the | DT)

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

The probability of a tag appearing depends only **on the previous tag** P(NN|JJ)

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

Two Kinds of Probabilities

- Tag transition probabilities $p(t_i|t_{i-1})$ $$P(t_i|t_{i-1}) = \frac{C(t_{i-1},t_i)}{C(t_{i-1})}$$
  - Determiners likely to precede adjectives 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 not $P(DT|JJ)$ to be high
  - Compute P(NN|DT) by counting in a labeled corpus:

$$P(NN|DT) = \frac{C(DT,NN)}{C(DT)} = \frac{56,509}{116,454} = .49$$

Two Kinds of Probabilities

- Word likelihood probabilities $p(w_i|t_i)$

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

  – VBZ (3rd person singular present verb) likely to be "is"
  – Compute $P(is|VBZ)$ by counting in a labeled corpus:

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

Example: The Verb “race”

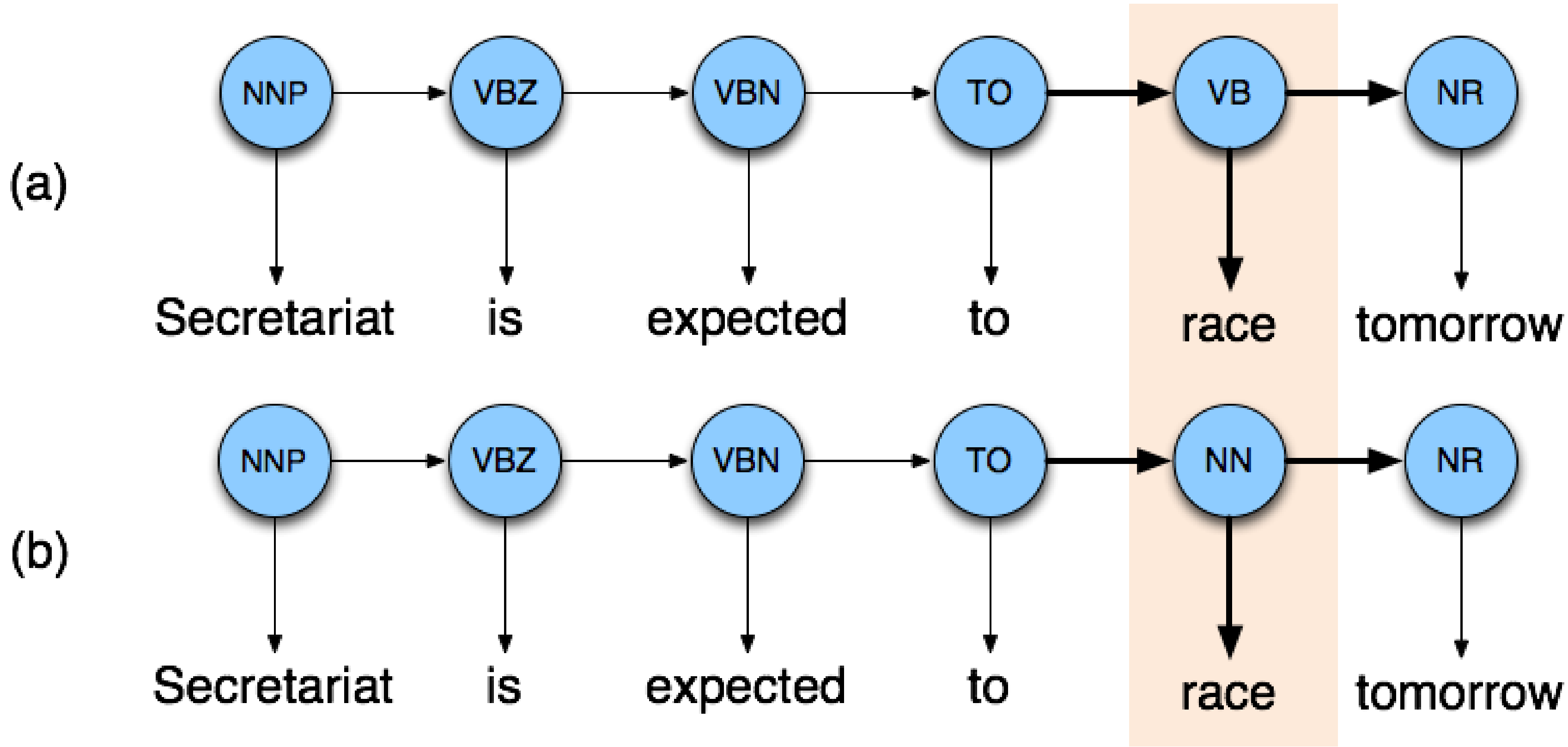
- Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** **race**/**VB** tomorrow/**NR**

- 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?

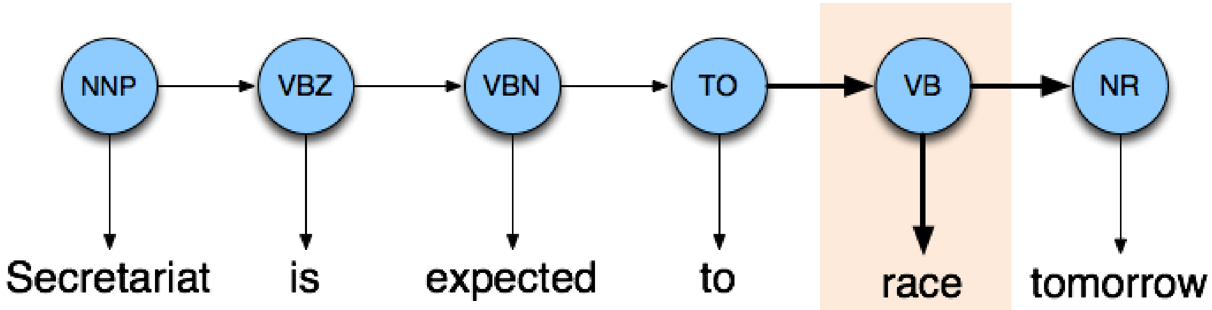
Disambiguating "race"

- for now, assuming tags of other words are known)



Calculating estimated probability

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



$$\begin{aligned}
& p(Secretariat|NNP) * p(NNP|Start) \\
& * p(is|VBZ) * p(VBZ|NNP) \\
& * p(expected|VBN) * p(VBN|VBZ) \\
& * p(to|TO) * p(TO|VBN) \\
& * p(race|VB) * p(VB|TO) \\
& * p(tomorrow|NR) * p(NR|VB)
\end{aligned}$$

Example

- From a training corpus, we know
  - $P(NN|TO)$ = .00047 $P(VB|TO)$ = .83 $P(race|NN)$ = .00057
  - $P(race|VB)$ = .00012 $P(NR|VB)$ = .0027 $P(NR|NN)$ = .0012

- How to use the above information to do POS tagging?
  - $P(VB|TO)P(NR|VB)P(race|VB)$ = .00000027
  - $P(NN|TO)P(NR|NN)P(race|NN)$ = .00000000032

- So we (correctly) choose the verb reading

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

Exercise: Compute missing probabilities

- Also called: build the HMM model
- Training corpus:
  - People/**NNS** continue/**VB** the/**DT** race/**NN**
  - Secretariat/**NNP** inquire/**VB** reason/**NN**

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

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

- Compute transition probabilities
  - P(VB|NNS) = 1/1 P(DT|VB) = ½ P(NN|VB) = ½
  - P(NN|DT) = P(VB|NNP) =

- Compute Word likelihood probabilities
  - P(people|NNS) = 1/1 P(continue|VB) = ½ P(inquire|VB) = ½
  - P(Secretariat|**NNP**) = P(race|**NN**) = p(reason|**NN**) = P(the|**DT**) =

Summary

- Parts of speech
- Tagsets
- Part of speech tagging
- HMM Tagging

- Next:
  - Markov Chains
  - Hidden Markov Models
    - Viterbi decoding

- The next several slides are about linguistics and are for your references

Open and Closed Classes

Skip

- Closed class: a small(ish) fixed membership
  - Usually function words (short common words which play a role in grammar)

  - 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, …*

- 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!
  - Nouns are typically where the bulk of the action is with respect to new items

Open Class Words

Skip

- Nouns
  - Proper nouns (Boulder, Granby, Beyoncé) -- 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)
  - With differing patterns of regularity