

Mieleni minun tefewi,
Xiwoni ajattelewí,
Mieli ruweta runoille,
Laatiua iaulamahan.
5 Meli kulta meiffoseni.

Natural Language Processing

Info 159/259

Lecture 10: Part-of-speech tagging (Feb 18, 2021)

David Bamman, UC Berkeley

context

everyone likes

a bottle of

is on the table

_____ makes you drunk

a cocktail with

and seltzer

from last time

Distribution

- Words that appear in similar contexts have similar representations (and similar meanings, by the distributional hypothesis).

Parts of speech

- Parts of speech are categories of words defined **distributionally** by the morphological and syntactic contexts a word appears in.

Morphological distribution

POS often defined by distributional properties; verbs = the class of words that each combine with the same set of affixes

	-s	-ed	-ing
walk	walks	walked	walking
slice	slices	sliced	slicing
believe	believes	believed	believing
of	*ofs	*ofed	*ofing
red	*reds	*redded	*reding

Morphological distribution

We can look to the function of the affix (denoting past tense) to include irregular inflections.

	-s	-ed	-ing
walk	walks	walked	walking
sleep	sleeps	slept	sleeping
eat	eats	ate	eating
give	gives	gave	giving

Syntactic distribution

- Substitution test: if a word is replaced by another word, does the sentence remain **grammatical**?

Kim saw the	elephant	before we did
	dog	
	idea	
	*of	
	*goes	

Syntactic distribution

- These can often be too strict; some contexts admit substitutability for some pairs but not others.

Kim saw the elephant before we did

*Sandy

both nouns but common
vs. proper

both verbs but transitive
vs. intransitive

Kim *arrived the elephant before we did

Nouns	People, places, things, actions-made-nouns (“I like swimming ”). Inflected for singular/plural
Verbs	Actions, processes. Inflected for tense, aspect, number, person
Adjectives	Properties, qualities. Usually modify nouns
Adverbs	Qualify the manner of verbs (“She ran downhill extremely quickly yesteray ”)
Determiner	Mark the beginning of a noun phrase (“ a dog”)
Pronouns	Refer to a noun phrase (he, she, it)
Prepositions	Indicate spatial/temporal relationships (on the table)
Conjunctions	Conjoin two phrases, clauses, sentences (and, or)

Open class

Nouns

fax, affluenza, subtweet, bitcoin, cronut, emoji, listicle,
mocktail, selfie, skort

Verbs

text, chillax, manspreading, photobomb, unfollow, google

Adjectives

crunk, amazeballs, post-truth, woke

Adverbs

hella, wicked

OOV? Guess Noun

Determiner

Pronouns

Prepositions

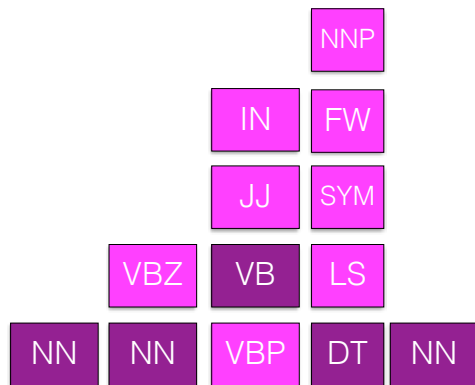
English has a new preposition, because internet [Garber
2013; Pullum 2014]

Conjunctions

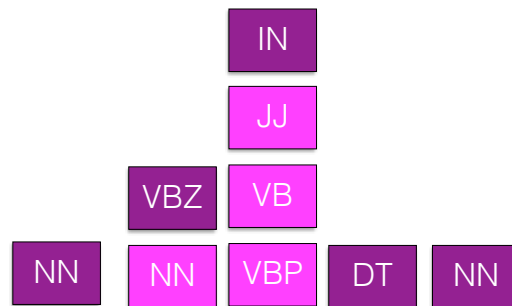
Closed class

POS tagging

Labeling the tag that's correct **for**
the context.



Fruit **flies like** a banana



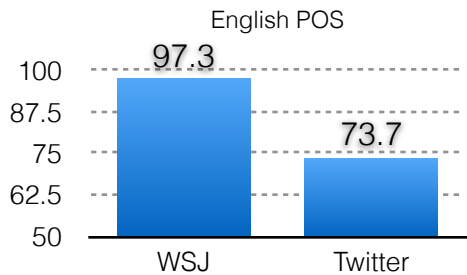
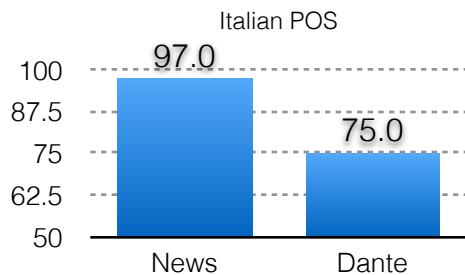
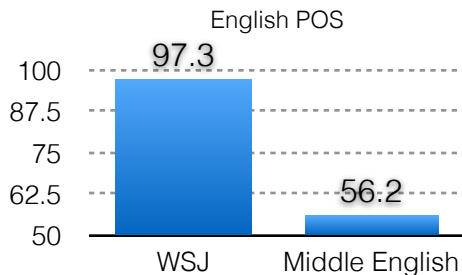
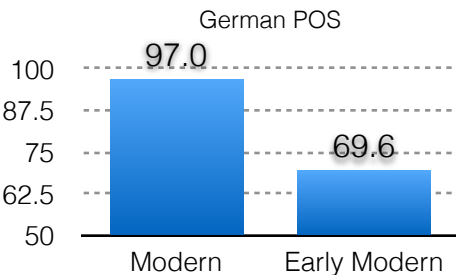
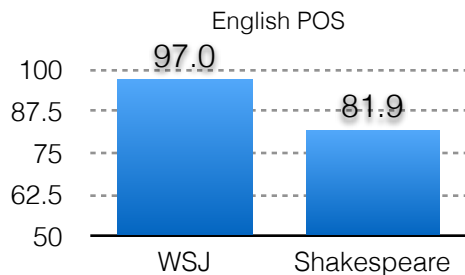
Time **flies like** an arrow

(Just tags in evidence within the Penn Treebank — more are possible!)

State of the art

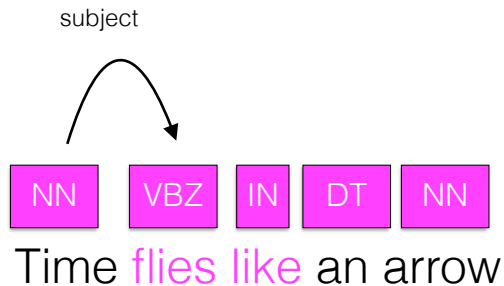
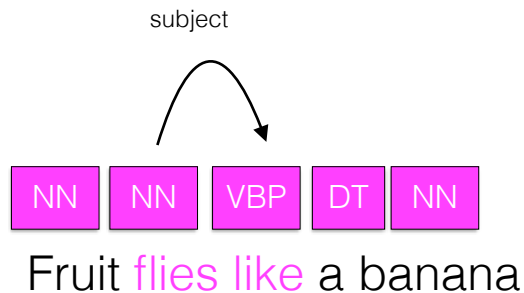
- Baseline: Most frequent class = 92.34%
- Token accuracy: 97% (English news)
[Toutanova et al. 2003; Søgaard 2010]
 - Optimistic: includes punctuation, words with only one tag (deterministic tagging)
 - Substantial drop across domains (e.g., train on news, test on literature)
- Whole sentence accuracy: 55%

Domain difference



Why is part of speech tagging useful?

POS indicative of syntax



POS indicative of MWE

at least one adjective/noun or noun phrase

and definitely
one noun

$$((A \mid N)^+ \mid ((A \mid N)^*(NP))(A \mid N)^*)N$$

AN: linear function; lexical ambiguity; mobile phase

NN: regression coefficients; word sense; surface area

AAN: Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase

ANN: cumulative distribution function; lexical ambiguity resolution; accessible surface area

NAN: mean squared error; domain independent set; silica based packing

NNN: class probability function; text analysis system; gradient elution chromatography

NPN: degrees of freedom; [*no example*]; energy of adsorption

POS is indicative of pronunciation

Noun	Verb
My conduct is great	I conduct myself well
She won the contest	I contest the ticket
He is my escort	He escorted me
That is an insult	Don't insult me
Rebel without a cause	He likes to rebel
He is a suspect	I suspect him

Tagsets

- Penn Treebank
- Universal Dependencies
- Twitter POS

Verbs

tag	description	example
VB	base form	I want to like
VBD	past tense	I/we/he/she/you liked
VBG	present participle	He was liking it
VBN	past participle	I had liked it
VBP	present (non 3rd-sing)	I like it
VBZ	present (3rd-sing)	He likes it
MD	modal verbs	He can go

Nouns

non-proper

proper

tag	description	example
NN	non-proper, singular or mass	the company
NNS	non-proper, plural	the companies
NNP	proper, singular	Carolina
NNPS	proper, plural	Carolinas

DT (Article)

- Articles (a, the, every, no)
- Indefinite determiners
(another, any, some, each)
- That, these, this, those when preceding noun
- All, both when not preceding another determiner or possessive pronoun

65548	the/dt
26970	a/dt
4405	an/dt
3115	this/dt
2117	some/dt
2102	that/dt
1274	all/dt
1085	any/dt
953	no/dt
778	those/dt

JJ

(Adjectives)

- General adjectives
 - *happy* person
 - *new* mail
- Ordinal numbers
 - *fourth* person

2002 other/jj
1925 new/jj
1563 last/jj
1174 many/jj
1142 such/jj
1058 first/jj
824 major/jj
715 federal/jj
698 next/jj
644 financial/jj

RB (Adverb)

- Most words that end in -ly
- Degree words (quite, too, very)
- Negative markers: not, n't, never

4410	n't/rb
2071	also/rb
1858	not/rb
1109	now/rb
1070	only/rb
1027	as/rb
961	even/rb
839	so/rb
810	about/rb
804	still/rb

IN (preposition, subordinating conjunction)

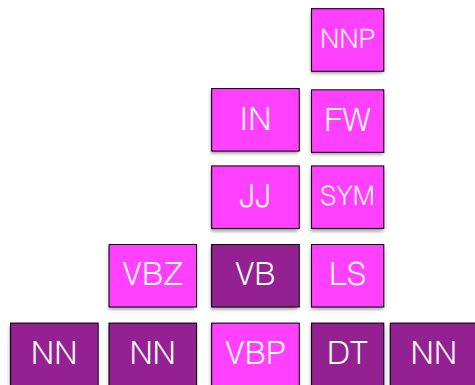
- All prepositions (except *to*) and subordinating conjunctions

- He jumped **on** the table
because he was excited

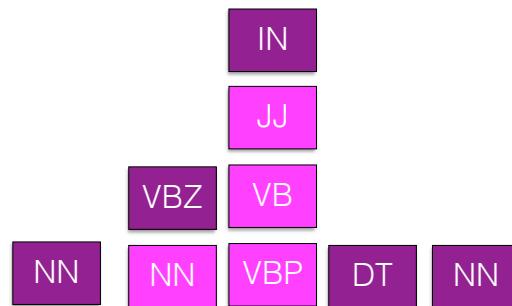
31111 of/in
22967 in/in
11425 for/in
7181 on/in
6684 that/in
6399 at/in
6229 by/in
5940 from/in
5874 with/in
5239 as/in

POS tagging

Labeling the tag that's correct **for**
the context.



Fruit **flies** **like** a banana



Time **flies** **like** an arrow

(Just tags in evidence within the Penn Treebank — more are possible!)

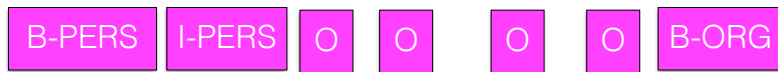
Sequence labeling

$$x = \{x_1, \dots, x_n\}$$

$$y = \{y_1, \dots, y_n\}$$

- For a set of inputs x with n sequential time steps, one corresponding label y_i for each x_i

Named entity recognition



tim cook is the ceo of apple

3 or 4-class:

- person
- location
- organization
- (misc)

7-class:

- person
- location
- organization
- time
- money
- percent
- date

Supersense tagging

O B-artifact I-artifact B-motion O B-time O O O B-group

The station wagons arrived at noon, a long shining line

O B-motion O O B-location I-location

that coursed through the west campus.

1	person	7	cognition	13	attribute	19	quantity	25	plant
2	communication	8	possession	14	object	20	motive	26	relation
3	artifact	9	location	15	process	21	animal		
4	act	10	substance	16	Tops	22	body		
5	group	11	state	17	phenomenon	23	feeling		
6	food	12	time	18	event	24	shape		

POS tagging training data

- Wall Street Journal (~1M tokens, 45 tags, English)
- Universal Dependencies (universal dependency treebanks for many languages; common POS tags for all)

<https://github.com/UniversalDependencies>

Majority class

- Pick the label each word is seen most often with in the training data

fruit	flies	like	a	banana
NN 12	VBZ 7	VB 74	FW 8	NN 3
	NNS 1	VBP 31	SYM 13	
		JJ 28	LS 2	
		IN 533	JJ 2	
			IN 1	
			DT 25820	
			NNP 2	

Naive Bayes

- Treat each prediction as independent of the others

$$P(y \mid x) = \frac{P(y)P(x \mid y)}{\sum_{y' \in \mathcal{Y}} P(y')P(x \mid y')}$$

$$P(\text{VBZ} \mid \text{flies}) = \frac{P(\text{VBZ})P(\text{flies} \mid \text{VBZ})}{\sum_{y' \in \mathcal{Y}} P(y')P(\text{flies} \mid y')}$$

Logistic regression

- Treat each prediction as independent of the others but condition on much more expressive set of features

$$P(y \mid x; \beta) = \frac{\exp \left(x^\top \beta_y \right)}{\sum_{y' \in \mathcal{Y}} \exp \left(x^\top \beta_{y'} \right)}$$

$$P(y = \text{VBZ} \mid x; \beta) = \frac{\exp \left(x^\top \beta_{\text{VBZ}} \right)}{\sum_{y' \in \mathcal{Y}} \exp \left(x^\top \beta_{y'} \right)}$$

Discriminative Features

Features are scoped over
entire observed input

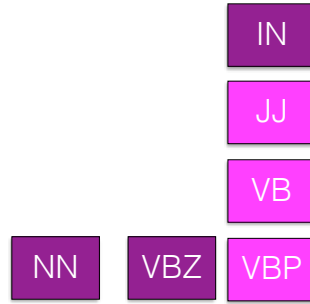
Fruit **flies** like a banana

feature	example
$x_i = \text{flies}$	1
$x_i = \text{car}$	0
$x_{i-1} = \text{fruit}$	1
$x_{i+1} = \text{like}$	1

Sequences

- Models that make independent predictions for elements in a sequence can reason over expressive representations of the **input** x (including correlations among inputs at different time steps x_i and x_j).
- But they don't capture another important source of information: correlations in the **labels** y .

Sequences



Time flies like an arrow

Sequences

Most common tag bigrams in
Penn Treebank training

DT	NN	41909
NNP	NNP	37696
NN	IN	35458
IN	DT	35006
JJ	NN	29699
DT	JJ	19166
NN	NN	17484
NN	,	16352
IN	NNP	15940
NN	.	15548
JJ	NNS	15297
NNS	IN	15146
TO	VB	13797
NNP	,	13683
IN	NN	11565

Sequences

x	time	flies	like	an	arrow
y	NN	VBZ	IN	DT	NN

$$P(y = \text{NN VBZ IN DT NN} \mid x = \text{time flies like an arrow})$$

Generative vs. Discriminative models

- Generative models specify a joint distribution over the labels and the data. With this you could **generate** new data

$$P(x, y) = P(y) P(x | y)$$

- Discriminative models specify the conditional distribution of the label y given the data x . These models focus on how to **discriminate** between the classes

$$P(y | x)$$

Generative

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{\sum_{y' \in \mathcal{Y}} P(x \mid y')P(y')}$$

$$P(y \mid x) \propto P(x \mid y)P(y)$$

$$\max_y P(x \mid y)P(y)$$

How do we parameterize these probabilities when x and y are sequences?

Hidden Markov Model

Prior probability of label
sequence

$$P(y) = P(y_1, \dots, y_n)$$

$$P(y_1, \dots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-1})$$

- We'll make a first-order Markov assumption and calculate the joint probability as the product of the individual factors conditioned **only on the previous tag**.

Hidden Markov Model

First-order HMM

$$P(y_1, \dots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-1})$$

Second-order HMM

$$P(y_1, \dots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-2}, y_{i-1})$$

Third-order HMM

$$P(y_1, \dots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-3}, y_{i-2}, y_{i-1})$$

Hidden Markov Model

$$\begin{aligned} P(y_1, \dots, y_n) &= P(y_1) \\ &\times P(y_2 \mid y_1) \\ &\times P(y_3 \mid y_1, y_2) \\ &\dots \\ &\times P(y_n \mid y_1, \dots, y_{n-1}) \end{aligned}$$

- Remember: a Markov assumption is an approximation to this **exact** decomposition (the chain rule of probability)

Hidden Markov Model

$$P(x \mid y) = P(x_1, \dots, x_n \mid y_1, \dots, y_n)$$

$$P(x_1, \dots, x_n \mid y_1, \dots, y_n) \approx \prod_{i=1}^N P(x_i \mid y_i)$$

- Here again we'll make a strong assumption: the probability of the word we see at a given time step is only dependent on **its own** label, no matter the Markov order used for $P(y)$.

NNP VBZ

is	1121
has	854
says	420
does	77
plans	50
expects	47
's	40
wants	31
owns	30
makes	29
hopes	24
remains	24
claims	19
seems	19
estimates	17

NN VBZ

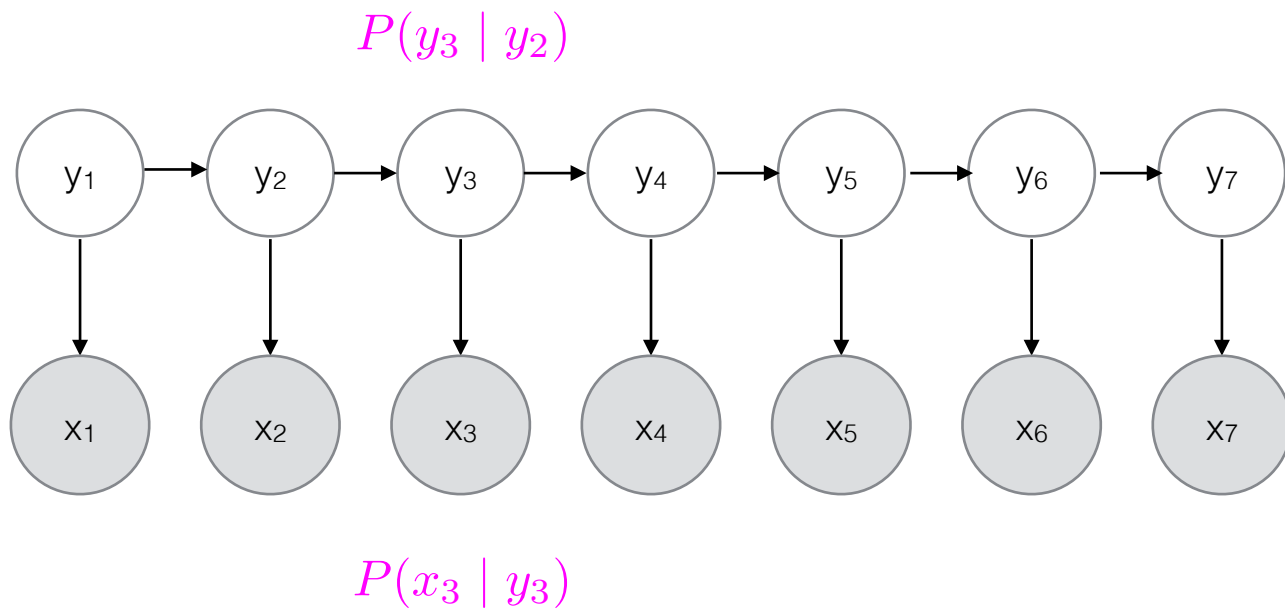
is	2893
has	1004
does	128
says	109
remains	56
's	51
includes	44
continues	43
makes	40
seems	34
comes	33
reflects	31
calls	30
expects	29
goes	27

$$P(x_i \mid y_{i-1}, y_i)$$

HMM

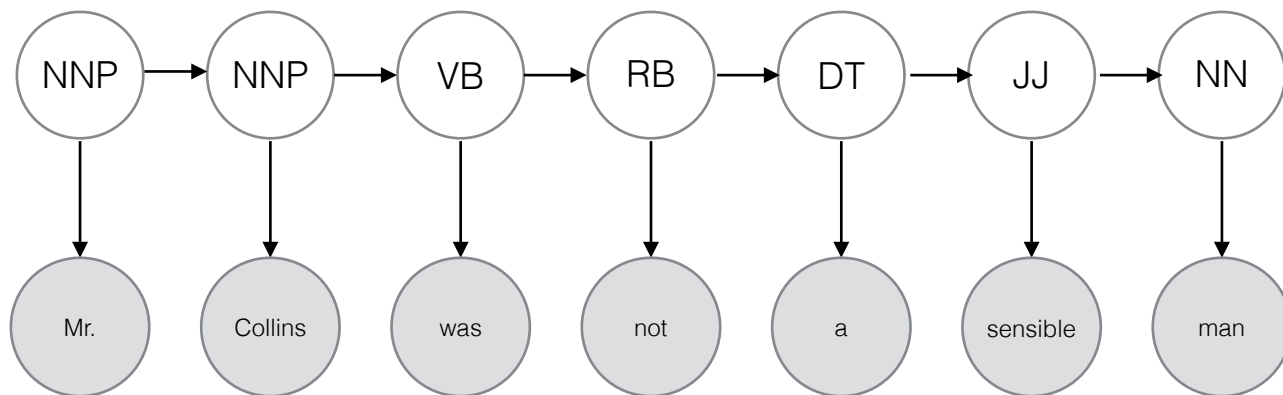
$$P(x_1, \dots, x_n, y_1, \dots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-1}) \prod_{i=1}^n P(x_i \mid y_i)$$

HMM



HMM

$$P(VB \mid NNP)$$



$$P(was \mid VB)$$

Parameter estimation

$$P(y_t \mid y_{t-1}) \qquad \frac{c(y_1, y_2)}{c(y_1)}$$

MLE for both is just counting
(as in Naive Bayes)

$$P(x_t \mid y_t) \qquad \frac{c(x, y)}{c(y)}$$

Transition probabilities

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

Figure 10.5 The A transition probabilities $P(t_i|t_{i-1})$ computed from the WSJ corpus without smoothing. Rows are labeled with the conditioning event; thus $P(VB|MD)$ is 0.7968.

Emission probabilities

	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0.000097	0
NN	0	0.000200	0.000223	0.000006	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Figure 10.6 Observation likelihoods B computed from the WSJ corpus without smoothing.

Smoothing

- One solution: add a little probability mass to every element.

maximum likelihood
estimate

$$P(x_i | y) = \frac{n_{i,y}}{n_y}$$

$n_{i,y}$ = count of word i in class y
 n_y = number of words in y
 V = size of vocabulary

smoothed estimates

$$P(x_i | y) = \frac{n_{i,y} + a}{n_y + Va}$$

same a for all x_i

$$P(x_i | y) = \frac{n_{i,y} + a_i}{n_y + \sum_{j=1}^V a_j}$$

possibly different a for each x_i

Decoding

- Greedy: proceed left to right, committing to the best tag for each time step (given the sequence seen so far)

Fruit flies like a banana

NN	VB	IN	DT	NN
----	----	----	----	----

Decoding

DT NN VBD IN DT NN ???

The horse raced past the barn fell

Decoding

DT NN VBD IN DT NN ???

The horse raced past the barn fell

DT NN VBN IN DT NN VBD

Information later on in the sentence can influence the best tags earlier on.

All paths

END							
DT							
NNP							
VB							
NN							
MD							
START							
	^	Janet	will	back	the	bill	\$

Ideally, what we want is to calculate the joint probability of **each path** and pick the one with the highest probability. But for N time steps and K labels, number of possible paths = K^N

5 word sentence with 45 Penn Treebank tags

$45^5 = 184,528,125$ different paths

$45^{20} = 1.16e33$ different paths

Viterbi algorithm

- Basic idea: if an optimal path through a sequence uses label L at time T, then it must have used an optimal path to get to label L at time T
- We can discard all non-optimal paths up to label L at time T

END							
DT							
NNP							
VB							
NN							
MD							
START							
	^	Janet	will	back	the	bill	\$

- At each time step t ending in label K , we find the max probability of any path that led to that state

END		
DT		$v_1(\text{DT})$
NNP		$v_1(\text{NNP})$
VB		$v_1(\text{VB})$
NN		$v_1(\text{NN})$
MD		$v_1(\text{MD})$
START		

Janet

What's the HMM probability of ending in Janet = NNP?

$$P(y_t \mid y_{t-1})P(x_t \mid y_t)$$

$$P(\text{NNP} \mid \text{START})P(\text{Janet} \mid \text{NNP})$$

END		
DT		$v_1(\text{DT})$
NNP		$v_1(\text{NNP})$
VB		$v_1(\text{VB})$
NN		$v_1(\text{NN})$
MD		$v_1(\text{MD})$
START		

Janet

Best path through time step 1
ending in tag y (trivially - best
path for all is just START)

$$v_1(y) = \max_{u \in \mathcal{Y}} [P(y_t = y \mid y_{t-1} = u) P(x_t \mid y_t = y)]$$

END			
DT		$v_1(\text{DT})$	$v_2(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$
START			

Janet will

What's the **max** HMM probability of ending in will = MD?

First, what's the HMM probability of a single path
ending in will = MD?

END			
DT		$v_1(\text{DT})$	$v_2(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$
START			

Janet will

$$P(y_1 \mid \text{START})P(x_1 \mid y_1) \times P(y_2 = \text{MD} \mid y_1)P(x_2 \mid y_2 = \text{MD})$$

END			
DT		$v_1(\text{DT})$	$v_2(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$
START			

Janet will

Best path through time step 2
ending in tag MD

$$P(\text{DT} \mid \text{START}) \times P(\text{Janet} \mid \text{DT}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{DT}) \times P(\text{will} \mid y_t = \text{MD}))$$

$$P(\text{NNP} \mid \text{START}) \times P(\text{Janet} \mid \text{NNP}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{NNP}) \times P(\text{will} \mid y_t = \text{MD}))$$

$$P(\text{VB} \mid \text{START}) \times P(\text{Janet} \mid \text{VB}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{VB}) \times P(\text{will} \mid y_t = \text{MD}))$$

$$P(\text{NN} \mid \text{START}) \times P(\text{Janet} \mid \text{NN}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{NN}) \times P(\text{will} \mid y_t = \text{MD}))$$

$$P(\text{MD} \mid \text{START}) \times P(\text{Janet} \mid \text{MD}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{MD}) \times P(\text{will} \mid y_t = \text{MD}))$$

Let's say the best path ending $y_2=MD$ includes $y_1=NNP$, with probability 0.0090.

Under our first-order Markov assumption, *if* $y_2=MD$ is in the best path for the complete sequence, *$y_1=NNP$ must be as well.* That means we can forget every other path ending in $y_2=MD$ that does not have *$y_1=NNP$.*

END			
DT		$v_1(DT)$	$v_2(DT)$
NNP		$v_1(NNP)$	$v_2(NNP)$
VB		$v_1(VB)$	$v_2(VB)$
NN		$v_1(NN)$	$v_2(NN)$
MD		$v_1(MD)$	$v_2(MD)$
START			

Janet will

0.0003	$P(DT \mid START) \times P(Janet \mid DT) \times P(y_t = MD \mid P(y_{t-1} = DT) \times P(will \mid y_t = MD)$
0.0090	$P(NNP \mid START) \times P(Janet \mid NNP) \times P(y_t = MD \mid P(y_{t-1} = NNP) \times P(will \mid y_t = MD)$
0.0001	$P(VB \mid START) \times P(Janet \mid VB) \times P(y_t = MD \mid P(y_{t-1} = VB) \times P(will \mid y_t = MD)$
0.0045	$P(NN \mid START) \times P(Janet \mid NN) \times P(y_t = MD \mid P(y_{t-1} = NN) \times P(will \mid y_t = MD)$
0.0002	$P(MD \mid START) \times P(Janet \mid MD) \times P(y_t = MD \mid P(y_{t-1} = MD) \times P(will \mid y_t = MD)$

None of the grey out paths could *possibly* be in the final optimal path, so we can forget them going forward.

To calculate this full probability, notice that we can reuse information we've already computed.

$$\underbrace{P(\text{DT} \mid \text{START}) \times P(\text{Janet} \mid \text{DT}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{DT})) \times P(\text{will} \mid y_t = \text{MD})}_{v_1(\text{DT})}$$

$$\underbrace{P(\text{NNP} \mid \text{START}) \times P(\text{Janet} \mid \text{NNP}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{NNP})) \times P(\text{will} \mid y_t = \text{MD})}_{v_1(\text{NNP})}$$

$$\underbrace{P(\text{VB} \mid \text{START}) \times P(\text{Janet} \mid \text{VB}) \times P(y_t = \text{MD} \mid P(y_{t-1} = \text{VB})) \times P(\text{will} \mid y_t = \text{MD})}_{v_1(\text{VB})}$$

...

END			
DT		$v_1(\text{DT})$	$v_2(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$
START			

Janet will

$$v_t(y) = \max_{u \in \mathcal{Y}} [v_{t-1}(u) \times P(y_t = y \mid y_{t-1} = u) P(x_t \mid y_t = y)]$$

END			
DT		$v_1(\text{DT})$	$v_2(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$
START			

Janet will

Every y at time step t may have a different u at time step $t-1$ that leads to its max.

Once we've determined that u for each y , we can forget all of the other values of u for that each y , since we know **they cannot be on the optimal path for the entire sequence.**

$$v_t(y) = \max_{u \in \mathcal{Y}} [v_{t-1}(u) \times P(y_t = y \mid y_{t-1} = u) P(x_t \mid y_t = y)]$$

END				
DT		$v_1(\text{DT})$	$v_2(\text{DT})$	$v_3(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$	$v_3(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$	$v_3(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$	$v_3(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$	$v_3(\text{MD})$
START				

Janet will back

25 paths ending in back = VB

[illegible]

$P(x_3 = \text{back} \mid y_3 = \text{VB})P(y_3 = \text{VB} \mid y_2 = \text{DT})P(x_2 = \text{will} \mid y_2 = \text{DT})P(y_2 = \text{DT} \mid y_1 = \text{DT})P(x_1 = \text{Janet} \mid y_1 = \text{DT})P(y_1 = \text{DT} \mid \text{START})$
 $P(x_3 = \text{back} \mid y_3 = \text{VB})P(y_3 = \text{VB} \mid y_2 = \text{NNP})P(x_2 = \text{will} \mid y_2 = \text{NNP})P(y_2 = \text{NNP} \mid y_1 = \text{DT})P(x_1 = \text{Janet} \mid y_1 = \text{DT})P(y_1 = \text{DT} \mid \text{START})$
 $P(x_3 = \text{back} \mid y_3 = \text{VB})P(y_3 = \text{VB} \mid y_2 = \text{MD})P(x_2 = \text{will} \mid y_2 = \text{MD})P(y_2 = \text{MD} \mid y_1 = \text{NNP})P(x_1 = \text{Janet} \mid y_1 = \text{NNP})P(y_1 = \text{NNP} \mid \text{START})$
 $P(x_3 = \text{back} \mid y_3 = \text{VB})P(y_3 = \text{VB} \mid y_2 = \text{NN})P(x_2 = \text{will} \mid y_2 = \text{NN})P(y_2 = \text{NN} \mid y_1 = \text{NN})P(x_1 = \text{Janet} \mid y_1 = \text{NN})P(y_1 = \text{NN} \mid \text{START})$
 $P(x_3 = \text{back} \mid y_3 = \text{VB})P(y_3 = \text{VB} \mid y_2 = \text{VB})P(x_2 = \text{will} \mid y_2 = \text{VB})P(y_2 = \text{VB} \mid y_1 = \text{MD})P(x_1 = \text{Janet} \mid y_1 = \text{MD})P(y_1 = \text{MD} \mid \text{START})$

In calculating the best path ending in $x_3=\text{back}$ and $y_3=\text{VB}$, we can forget every other path that we've already determined to be suboptimal.

END			
DT		$v_1(\text{DT})$	$v_2(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$
START			

Janet will

END				
DT		$v_1(\text{DT})$	$v_2(\text{DT})$	$v_3(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$	$v_3(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$	$v_3(\text{VB})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$	$v_3(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$	$v_3(\text{MD})$
START				

Janet will back

So for every label at every time step, we only need to keep track of which label at the previous time step $t-1$ led to the highest joint probability at that time step t .

END						
DT		$v_1(\text{DT})$	$v_2(\text{DT})$	$v_3(\text{DT})$	$v_4(\text{DT})$	$v_5(\text{DT})$
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$	$v_3(\text{NNP})$	$v_4(\text{NNP})$	$v_5(\text{NNP})$
VB		$v_1(\text{VB})$	$v_2(\text{VB})$	$v_3(\text{VB})$	$v_4(\text{MD})$	$v_5(\text{MD})$
NN		$v_1(\text{NN})$	$v_2(\text{NN})$	$v_3(\text{NN})$	$v_4(\text{NN})$	$v_5(\text{NN})$
MD		$v_1(\text{MD})$	$v_2(\text{MD})$	$v_3(\text{MD})$	$v_4(\text{MD})$	$v_5(\text{MD})$
START						

Janet

will

back

the

bill

END							$v_T(\text{END})$
DT		$v_1(\text{DT})$	$v_2(\text{DT})$	$v_3(\text{DT})$	$v_4(\text{DT})$	$v_5(\text{DT})$	
NNP		$v_1(\text{NNP})$	$v_2(\text{NNP})$	$v_3(\text{NNP})$	$v_4(\text{NNP})$	$v_5(\text{NNP})$	
VB		$v_1(\text{VB})$	$v_2(\text{VB})$	$v_3(\text{VB})$	$v_4(\text{MD})$	$v_5(\text{MD})$	
NN		$v_1(\text{NN})$	$v_2(\text{NN})$	$v_3(\text{NN})$	$v_4(\text{NN})$	$v_5(\text{NN})$	
MD		$v_1(\text{MD})$	$v_2(\text{MD})$	$v_3(\text{MD})$	$v_4(\text{MD})$	$v_5(\text{MD})$	
START							

Janet

will

back

the

bill

$v_T(\text{END})$ encodes the best path through the entire sequence

END							$v_T(\text{END})$
DT							
NNP							
VB							
NN							
MD							
START							
		Janet	will	back	the	bill	

For each timestep t + label, keep track of the max element from $t-1$ to reconstruct best path

function VITERBI(*observations* of len T , *state-graph* of len N) **returns** *best-path*

create a path probability matrix $viterbi[N+2, T]$

for each state s **from** 1 **to** N **do** ; initialization step

$viterbi[s, 1] \leftarrow a_{0,s} * b_s(o_1)$

$backpointer[s, 1] \leftarrow 0$

for each time step t **from** 2 **to** T **do** ; recursion step

for each state s **from** 1 **to** N **do**

$viterbi[s, t] \leftarrow \max_{s'=1}^N viterbi[s', t-1] * a_{s',s} * b_s(o_t)$

$backpointer[s, t] \leftarrow \operatorname{argmax}_{s'=1}^N viterbi[s', t-1] * a_{s',s}$

$viterbi[q_F, T] \leftarrow \max_{s=1}^N viterbi[s, T] * a_{s,q_F}$; termination step

$backpointer[q_F, T] \leftarrow \operatorname{argmax}_{s=1}^N viterbi[s, T] * a_{s,q_F}$; termination step

return the backtrace path by following backpointers to states back in time from $backpointer[q_F, T]$

Figure 10.8 Viterbi algorithm for finding optimal sequence of tags. Given an observation sequence and an HMM $\lambda = (A, B)$, the algorithm returns the state path through the HMM that assigns maximum likelihood to the observation sequence. Note that states 0 and q_F are non-emitting.