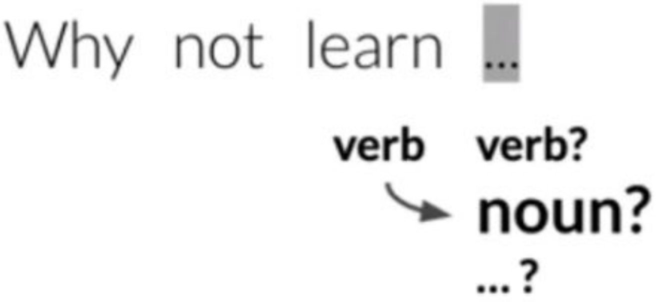
Parts of Speech Tagging



Next word prediction

The man ***fans*** the flame
The ***fans*** watch the race

I will ***book*** my ticket
I read the ***book***

Word level ambiguity

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	

Bender 2013

What is the Exact Aim?

Let’s Annotate Some Examples?

I hate pizza

I hate eating pizza

You used to speak politely

Austin worked really hard for the assignment

Let’s find nouns, proper nouns, verbs, adverbs, adjectives, determiners, and prepositions

Penn Treebank Tagset

From the Wall Street Journal and Brown corpora
Dependency grammars (introduced later) have another tagset

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	*$*
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	*#*
JJS	superlative adj	*wildest*	SYM	symbol	*+,%, &*	“	left quote	‘ or “
LS	list item marker	*1, 2, One*	TO	“to”	*to*	”	right quote	’ or ”
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Part of Speech Tagging Challenge

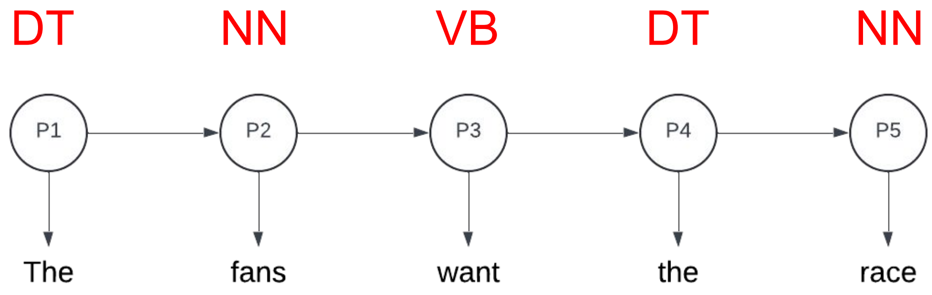
- Many words can take multiple tags depending on context
  - $\sim$ 14–15% of the words in the Wall Street Journal and Brown corpora

Adjective	earnings growth took a back/JJ seat
Mass noun	a small building in the back/NN
Verb present tense	a clear majority of senators back/VBP the bill
Verb	Dave began to back/VB toward the door
Particle	enable the country to buy back/RP about debt
Adverb	I was twenty-one back/RB then

- Simple baseline: most frequent class

Viterbi Algorithm

Example: The fans watch the race



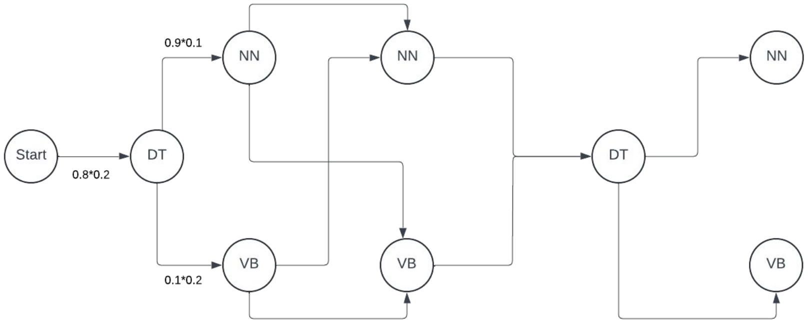
Hidden Markov Model (HMM)

Emission	The	fans	watch	race
DT	0.2	0	0	0
NN	0	0.1	0.3	0.1
VB	0	0.2	0.15	0.3

Transition	DT	NN	VB
(Start)	0.8	0.2	0
DT	0	0.9	0.1
NN	0	0.5	0.5
VB	0.5	0.5	0

Viterbi Algorithm

The fans watch the race

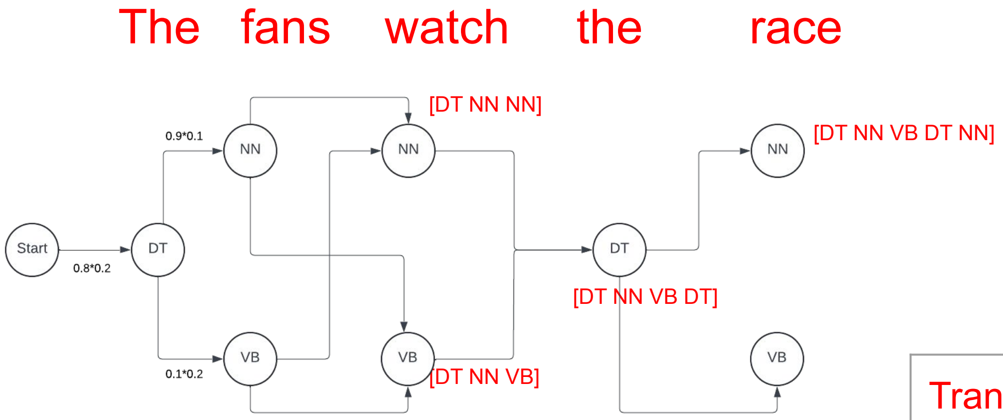


Let's try probabilistic state machine?

Emission	The	fans	watch	race
DT	0.2	0	0	0
NN	0	0.1	0.3	0.1
VB	0	0.2	0.15	0.3

Transition	DT	NN	VB
(Start)	0.8	0.2	0
DT	0	0.9	0.1
NN	0	0.5	0.5
VB	0.5	0.5	0

Viterbi Algorithm

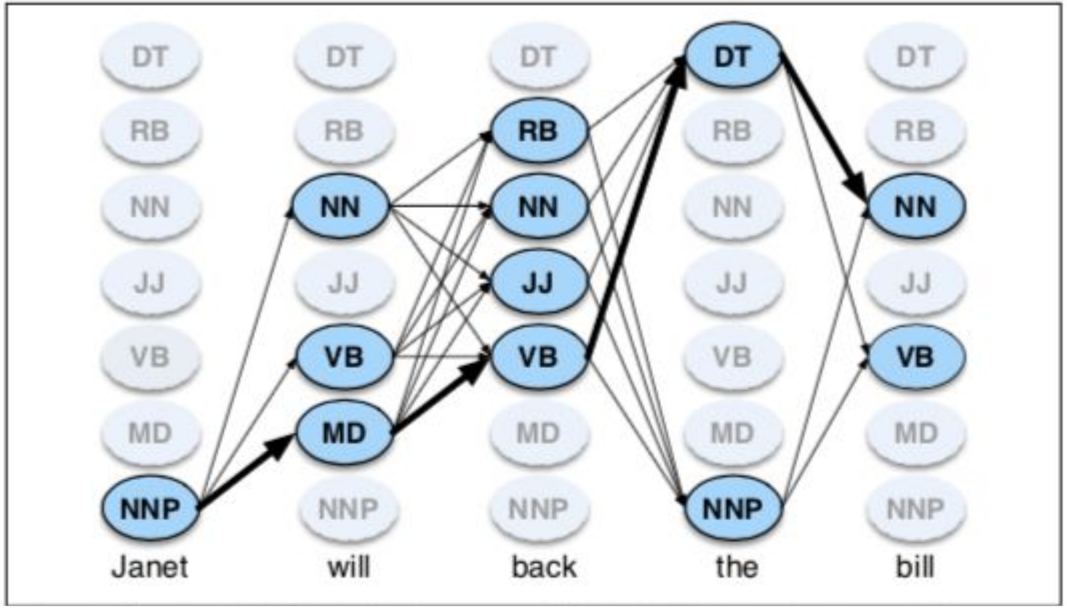


We maximize the product of probabilities

Emission	The	fans	watch	race
DT	0.2	0	0	0
NN	0	0.1	0.3	0.1
VB	0	0.2	0.15	0.3

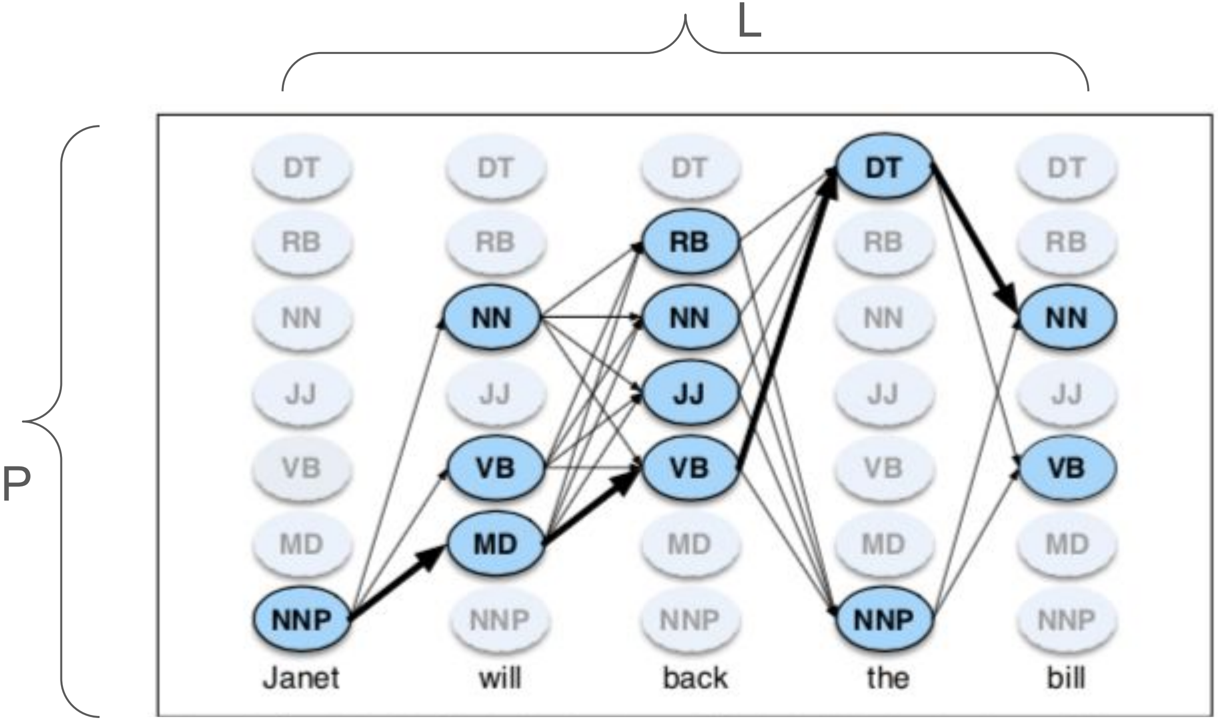
Transition	DT	NN	VB
(Start)	0.8	0.2	0
DT	0	0.9	0.1
NN	0	0.5	0.5
VB	0.5	0.5	0

Time complexity



How will you implement it?

Time complexity



Dynamic programming!

Brute force: P^L;   Viterbi: LP^2

Performance on Penn Treebank

How many words in the test set are tagged correctly?
Answer: 97%

Baseline is 93.7% based on P(t/w)

Statistical models are also powerful!

Parts of speech can be used as features of the model

Categories	Features	Explanation
Linguistic	Definite/indefinite articles	Occurrences normalized by $len(c)$
Linguistic	1st/2nd person pronouns	Occurrences normalized by $len(c)$
Linguistic	Hedges	Use the list of hedge words created by Hyland [18]
Linguistic	Sentiment	VADER compound scores [17]
Linguistic	Biased language	Occurrences of each subtype of biased text [31]
Linguistic	Examples	Occurrences of “for example” and alternative expressions
Linguistic	Questions	Count of question marks
Linguistic	Links	Count of “http” and “https” marks

Zhen Guo, Zhe Zhang, and Munindar Singh. 2020. In Opinion Holders’ Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In Proceedings of The Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, NY, USA, 2388–2399. https://doi.org/10.1145/3366423.3380302

Other Algorithms

Transformers: BERT, RoBERTa, and XLNet

Using LLMs

Acknowledgements

Dr. Munindar Singh at NC State

Dr. David Bamman at UC Berkeley