Sequence Labeling

Prof. Sameer Singh

CS 295: STATISTICAL NLP

WINTER 2017

January 31, 2017

Outline

- Sequence Labelling and POS Tagging
- Generative Modeling: HMMs
- Inference in HMMs: Viterbi and F/B
- Unsupervised Tagging using EM

Outline

Sequence Labelling and POS Tagging

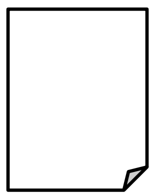
Generative Modeling: HMMs

Inference in HMMs: Viterbi and F/B

Unsupervised Tagging using EM

Classification

$\vec{x} = x_1 \ldots x_n$



$y \in \{1 \ldots L\}$

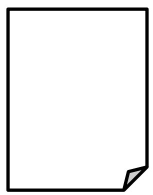
Sentiment Analysis $y \in \{+, -\}$

Identify Topic $y \in \{\text{sports, politics}, \ldots\}$

Language Model $y \in \{1 \ldots V\}$

Sequence Labeling

$x_1 \ldots x_n$



X

$y_1 \ldots y_n$

$y_i \in \{1 \ldots L\}$

Y

$\vec{y} \in Y^n$

Part~~s~~ of Speech

	x_1	x_2	.	. .	x_5	x_6
$\vec{x}$	This	is	a	simple	sentence	.
$\vec{y}$	DET	VB	DET	ADJ	NOUN	.
	⋮					

$y_i \in Y$

Applications:

- Text to speech: record, lead, …
- Machine translation: run, walk, …
- Noun phrases: `grep {JJ | NN}* {NN | NNS}`
- and many others…

Parts of Speech: Tags

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two*	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, sing.	*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*			

PTB

45

“Open classes”

Nouns, verbs, adjectives, adverbs, numbers

“Closed classes”

- Modal verbs
- Prepositions (on, to)
- Particles (off, up)
- Determiners (the, some)
- Pronouns (she, they)
- Conjunctions (and, or)

Named Entity Recognition

Barack Obama spoke from the White House today .

PER PER O O O LOC LOC O O

Field Segmentation: Ads

3BR flat in Bruntsfield , near main roads . Bright , well maintained ...

SIZE TYPE O LOC O LOC LOC LOC O FEAT O FEAT FEAT ...

Field Segmentation: Citations

drucker h., schapire r., and simard r. improving performance in neural networks using a boosting algorithm. advances in neural information processing systems 5, san mateo, ca. morgan kaufmann.1993 pages 42-49, in hanson, s. j., cowan, j. d., and giles, c. l., editors,

Authors

Title

drucker h., schapire r., and simard r. improving performance in neural networks using a boosting algorithm . advances in neural information processing systems 5 , san mateo, ca. morgan kaufmann.1993 pages 42-49, in hanson, s. j., cowan, j. d., and giles, c. l., editors.

Publication Venue

Outline

Sequence Labelling and POS Tagging

Generative Modeling: HMMs

Inference in HMMs: Viterbi and F/B

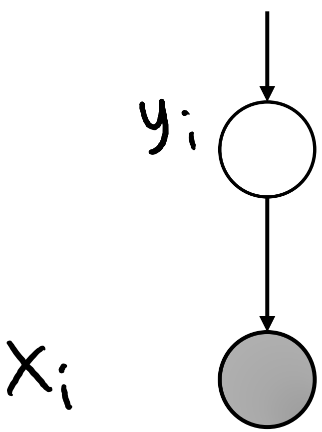
Unsupervised Tagging using EM

Naïve Bayes Classifier

$$p(\vec{y},\vec{x}) = \prod_i p(y_i, x_i)$$

$$= \prod_i \underbrace{p(y_i)}_{\frac{\# y_i}{N}} \; \underbrace{p(x_i \mid y_i)}_{\frac{\# x_i, y_i}{\# y_i}}$$

From labeled data



$\boxed{\sim 3.5\%}$

$\sim 90\%$

WSJ

“Transitions” matter

VBD		VB			
VBN	VBZ	VBP	VBZ		
NNP	NNS	NN	NNS	CD	NN
Fed	raises	interest	rates	0.5	percent

“Impossible” Transitions

- Two determiners never follow each other
- Two base form verbs never follow each other
- Determiner is followed by adjective or noun

Based on semantics

Fruit flies like a bird.

VB IN

Fruit flies like bananas.

N V

How do we select a “consistent” set of POS tags?

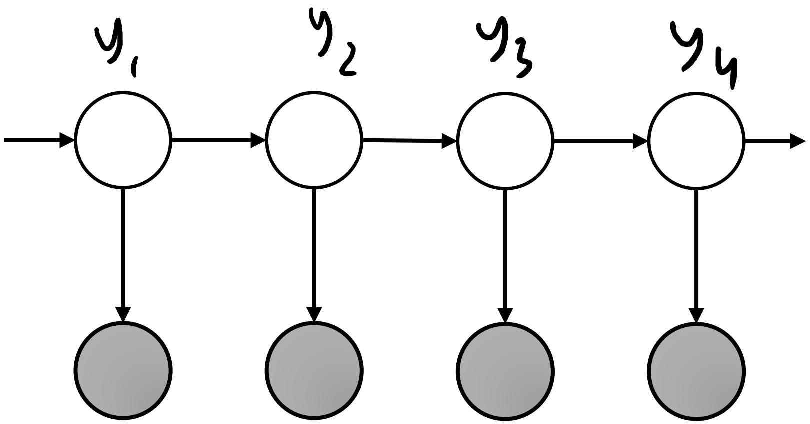
“Transitions” matter

"like", "bananas"

Types:		WSJ		Brown	
Unambiguous	(1 tag)	44,432	(**86%**)	45,799	(**85%**)
Ambiguous	(2+ tags)	7,025	(**14%**)	8,050	(**15%**)
Tokens:					
Unambiguous	(1 tag)	577,421	(**45%**)	384,349	(**33%**)
Ambiguous	(2+ tags)	711,780	(**55%**)	786,646	(**67%**)

x_i = "like"
in "I like bananas"

“Transitions” matter

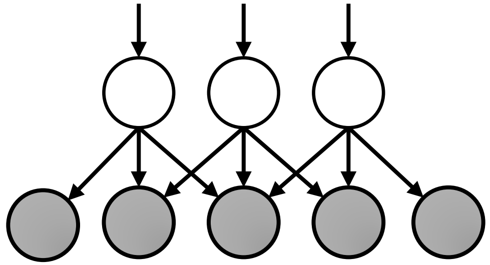


$$p(\vec{y},\vec{x}) = \prod_i p(y_i)\, p(x_i \mid y_i)$$

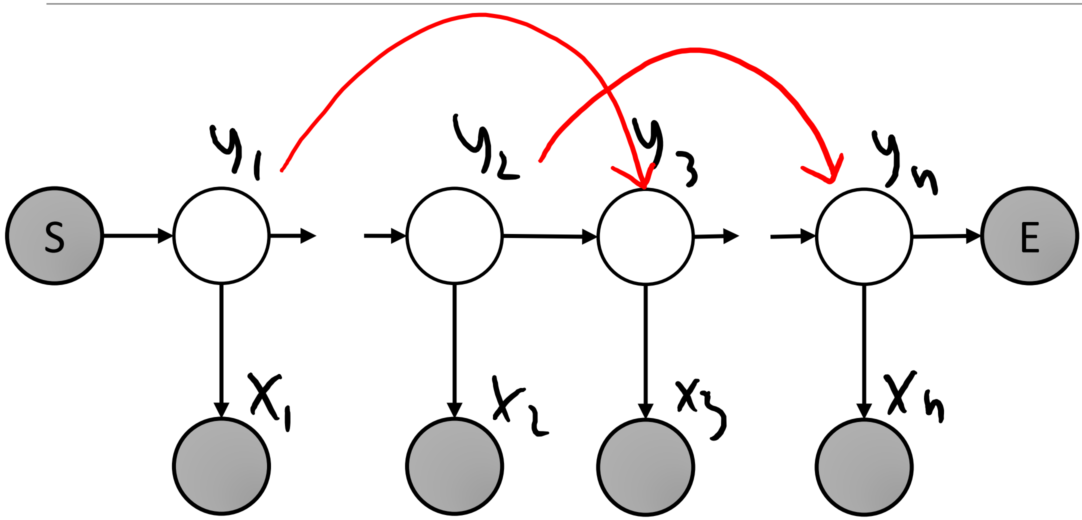
$$p(\vec{y},\vec{x}) = \prod_i p(y_i \mid y_{i-1})\, p(x_i \mid y_i)$$

Transition on Words versus Tags

- Too many words, learn the same thing again
- Support for unseen words: “I like tenguizino!”



Hidden Markov Models



$$p(\vec{y}, \vec{x}) = \prod_i e(x_i | y_i)\, t(y_i | y_{i-1})$$

emission

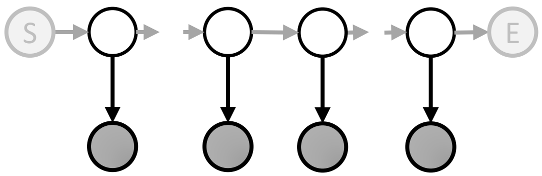
Transition

Example Sentence

x		This	is	a	simple	sentence	
y	S	DET	VB	DET	ADJ	NOUN	E

$$p(\vec{y}, \vec{x}) = e(\text{This}|\text{DET})\, e(\text{is}|\text{VB})\, e(\text{a}|\text{DET}) \cdots$$
$$t(\text{DET}|\text{S})\, t(\text{VB}|\text{DET}) \cdots t(\text{E}|\text{NOUN})$$

Estimating Emissions



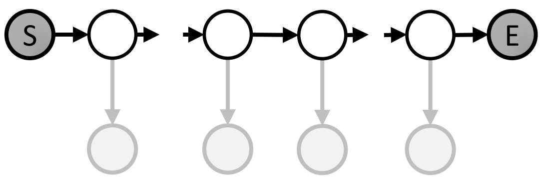
$$e(x_i \mid y_i) = \frac{\# x_i \wedge y_i}{\# y_i}$$

Smoothing

- Unknown/rare words get inaccurate probabilities
- Reminder: Laplace Smoothing (Add-k)
- Next lecture: we will look at "features"

$$e(x_i \mid y_i) = \frac{\# x_i \wedge y_i + k}{\# y_i + kV}$$

Estimating Transitions



$$t(y_i|y_{i-1}) = \frac{\# y_{i-1} \wedge y_i}{\# y_{i-1}}$$

N V

Interpolation

- If there are too many tags, or too little data, some combinations are too rare
- Same as N-gram language models, “backoff” to simpler models

$$t(y_i|y_{i-1}) = \lambda\, P(y_i|y_{i-1}) + (1-\lambda)\, P(y_i)$$

Outline

Sequence Labelling and POS Tagging

Generative Modeling: HMMs

Inference in HMMs: Viterbi and F/B

Unsupervised Tagging using EM

Predicting from HMMs

$$p(\vec{y}, \vec{x})$$

$\vec{x} = \text{" .... "}$ $\qquad$ $\vec{y} = ?$

$$\overset{*}{y} = \operatorname*{argmax}_{y \in \mathcal{Y}^n} p(\vec{y} \mid \vec{x}) = \operatorname*{argmax}_{y} \frac{p(\vec{y}, \vec{x})}{p(\vec{x})}$$

$$= \operatorname*{argmax}_{y} p(\vec{y}, \vec{x})$$

Brute Force Inference

$$\vec{y}^* = \underbrace{\arg\max_{\vec{y} \in \mathcal{Y}^n}}_{O(|\mathcal{Y}|^n)} p(\vec{y} \mid \vec{x})$$

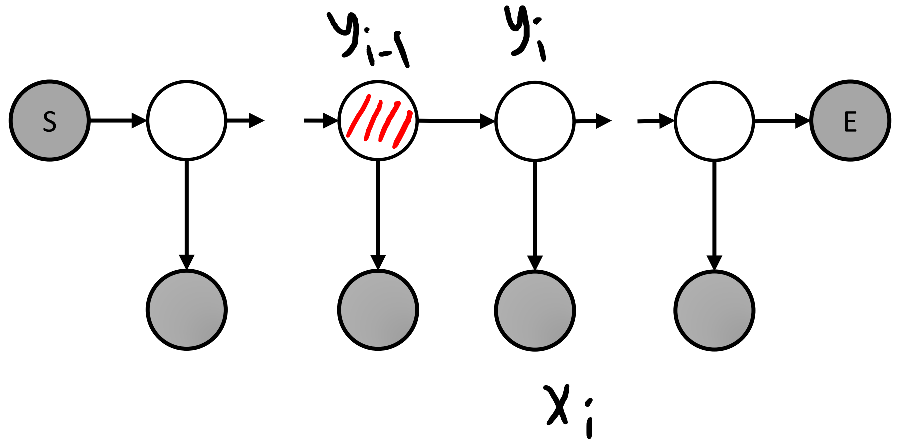
$$O(|\mathcal{Y}|^n)$$

15

45

$$(45)^{15}$$

Conditional Independence



$$y_i \perp y_1 \ldots y_{i-2} \mid y_{i-1}$$

$$y_i \perp y_{i+2} \ldots y_n \mid y_{i+1}$$

Dynamic Programming

$$\overset{1\ldots n}{i},\quad \overset{1\ldots L}{y_i}$$

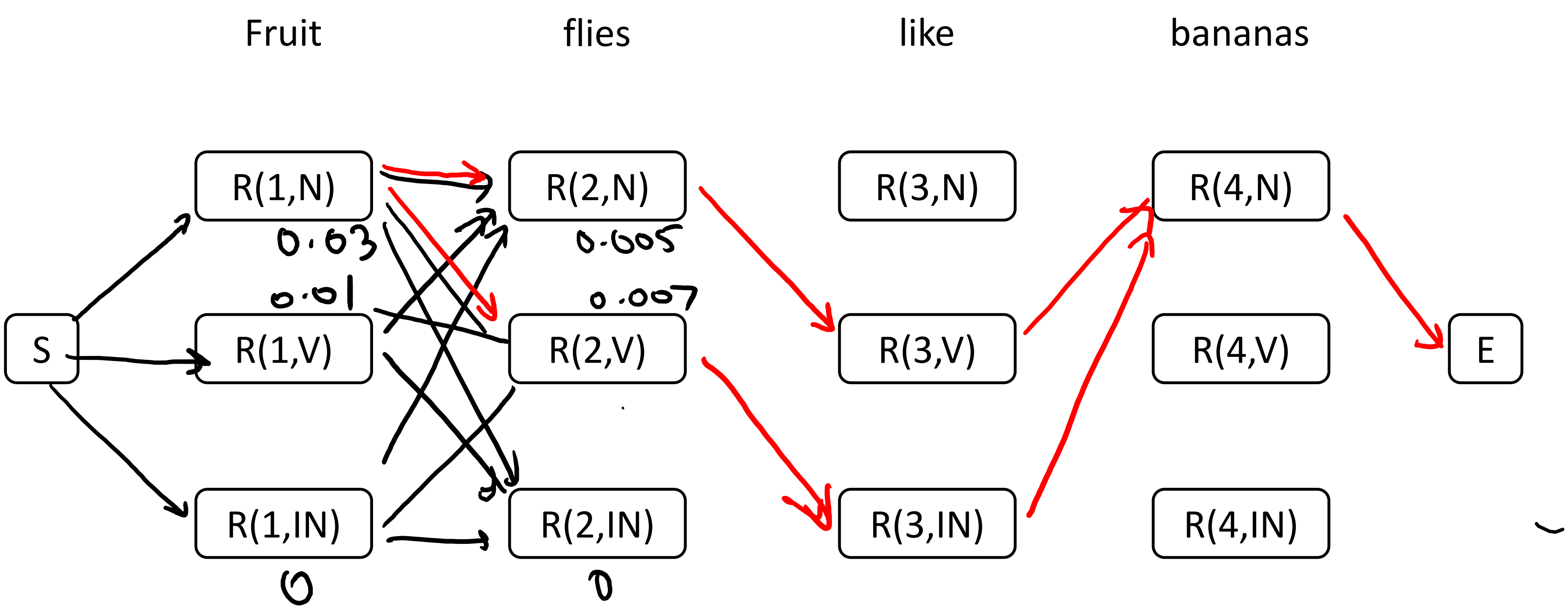
$$R(i, y_i) = \max_{y_1 \ldots y_{i-1}} p(x_1 \ldots x_i, y_1 \ldots y_i)$$

max probability
of setting y_i
before i

$$= \max_{y_{i-1}} e(x_i|y_i)\, t(y_i|y_{i-1}) \max_{y_1 \ldots y_{i-2}} p(x_1 \ldots x_{i-1}, y_1, \ldots y_{i-1})$$

$$= \max_{y_{i-1}} e(x_i|y_i)\, t(y_i|y_{i-1})\, R(i-1, y_{i-1})$$

State Lattice



Viterbi Decoding Algorithm

Initialization

$$R(0, S) = 1$$

Iterative Computation (forward)

$$R(i, y_i) = \max_{y_{i-1}} e(x_i \mid y_i)\, t(y_i \mid y_{i-1})\, \underline{R(i-1, y_{i-1})}$$

argmax

Follow pointers (backward)

Computational Complexity

$$R(i, y_i) = \max_{y_{i-1}} \ldots$$

$1 \ldots n$ $\quad$ $1 \ldots |Y|$ $\quad$ $1 \ldots |Y|$

$$O(n|Y|^2)$$

Outline

- Sequence Labelling and POS Tagging
- Generative Modeling: HMMs
- Inference in HMMs: Viterbi and F/B
- **Unsupervised Tagging using EM**

Unsupervised Tagging

Supervision is not always appropriate

- Linguist has to read and understand each sentence
  - Time consuming and expensive
- Contains domain specific signal in the labels
  - WSJ doesn't generalize to Twitter, for example
- Difficult to agree on the universal part-of-speech tags (C5 tags: 61, Brown: 87)
- Want to apply it to low-resource/unknown languages

Generalize the notion of "clustering" to sequence labeling.

Expectation Maximization

K-Means

Initialization

K "pos" tags

$e(x_i | y_i)$, $t(y_i | y_{i-1})$

Pick K random centroids

Compute Expectations

$p(\vec{y}\ \vec{x})$ given e, t

Cluster all the points

Update Parameters

e, t given $p(\vec{y}, \vec{x})$

Update centroids

Upcoming…

Homework

- Homework 2 is due (~10 days): February 9, 2017
- Write-up, data, and code for Homework 2 is up
- Ask questions early!

Project

- Proposal is due in a week: February 7, 2017
- Only 2 pages

Summaries

- Paper summaries: February 17, February 28, March 14
- Only 1 page each