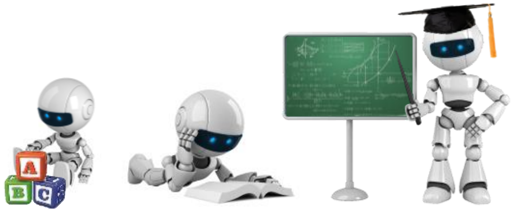
USC Viterbi
School of Engineering
Department of Computer Science

CSCI 544
Applied Natural Language Processing

Mohammad Rostami
USC Computer Science Department



Adapted from slides by Danqi Chen, Michael Collins, and Max Ma

Logistical Notes

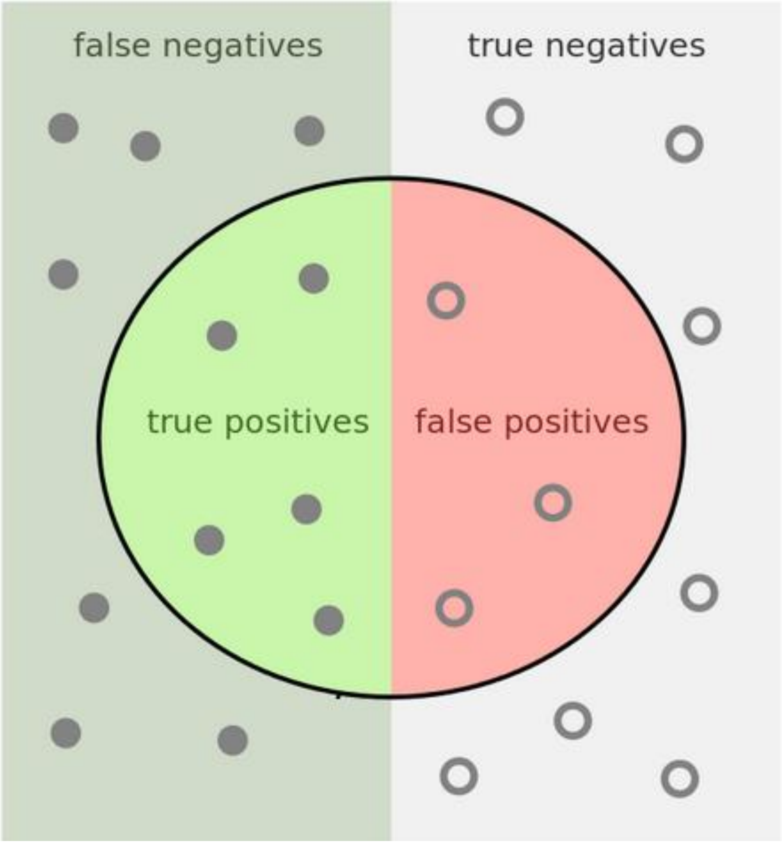
- HW:
  - Submit HW1 on Blackboard.
- Project Group Formation Deadline: 09/12
  - ~3/4 of the class have formed their group.
  - Contact groups with 3 or more members on Excel
  - Do NOT form more than 51 groups
  - Meet weekly, helpful for both HW and project
- Paper Selection Deadline: 09/19
  - Description upload on Blackboard
  - Check and then enter your paper: https://docs.google.com/spreadsheets/d/1_vafG77ijmETCnuVZvKpT35k--5op5wn71GZXgAY7O0
- Project Proposal Deadline: 10/03
  - Pick your paper with an outlook for the project
  - Check YouTube and Arxiv for projects of previous years
  - It is OK you pick something in line with an ongoing project but not past projects

Model Evaluation Process

- We use a training dataset for model selection
- A **good** parametric model along with a **suitable** training algorithm guarantees training a model that works well on the training data
- We need to validate that trained models **generalize** well on unseen data instances
- We need a second testing dataset which is fully independent of the training dataset
- We randomly split the annotated dataset into testing and training splits (sometimes, a validation set is generated as well)

Evaluation Metrics

- **Accuracy**: proportion of correctly classified items
- Accuracy can be dominated by **true negatives** (items correctly classified as not in a class).
- Sensitive with respect to imbalance
- Precision: $\frac{\text{True Positives}}{\text{True Positives+False Positive}}$
- Recall: $\frac{\text{True Positives}}{\text{True Positives+False negative}}$
- Precision and recall are not useful metrics when used in isolation?
- We want our model to have good performance with respect to both metrics
- Implemented in sklearn



Evaluation Metrics

- Why having one measure is helpful?
- $\text{F1} = \frac{2 \text{ Precision Recall}}{\text{Precision} + \text{Recall}}$
- F1 is biased towards the lower of precision and recall:

- harmonic mean < geometric mean < arithmetic mean
- F1=0 when Precision=0 or Recall=0

- Generalized F score:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

Structured Prediction

- Unstructured Prediction

- Output consists of a single prediction: classification, regression

- We may want to predict several outputs

- Examples: image segmentation, sequence tagging
- We have Strong correlations between output components
- Exponential output space: decoding is challenging

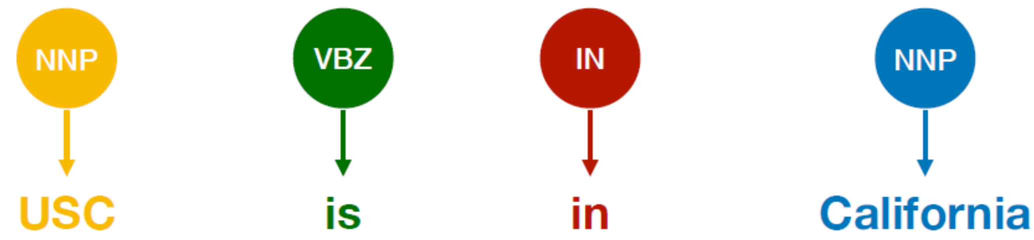
$$y^* = \operatorname{argmax}_{y \in \mathcal{Y}} p(y \mid x)$$

Sequence Labeling

- A structure prediction task

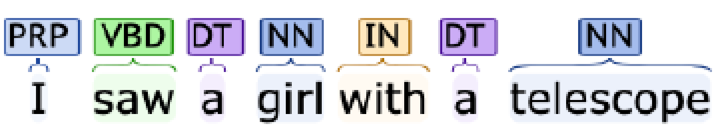
$$Y = < y_i, y_2, \ldots, y_n >$$

$$X = < x_i, x_2, \ldots, x_n >$$

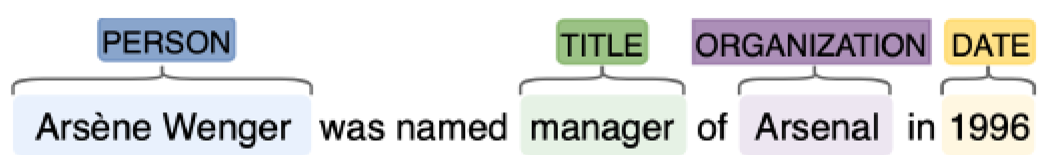


- Goal: assign each token of X, a value from the discrete label-space Y

Part-of-Speech Tagging

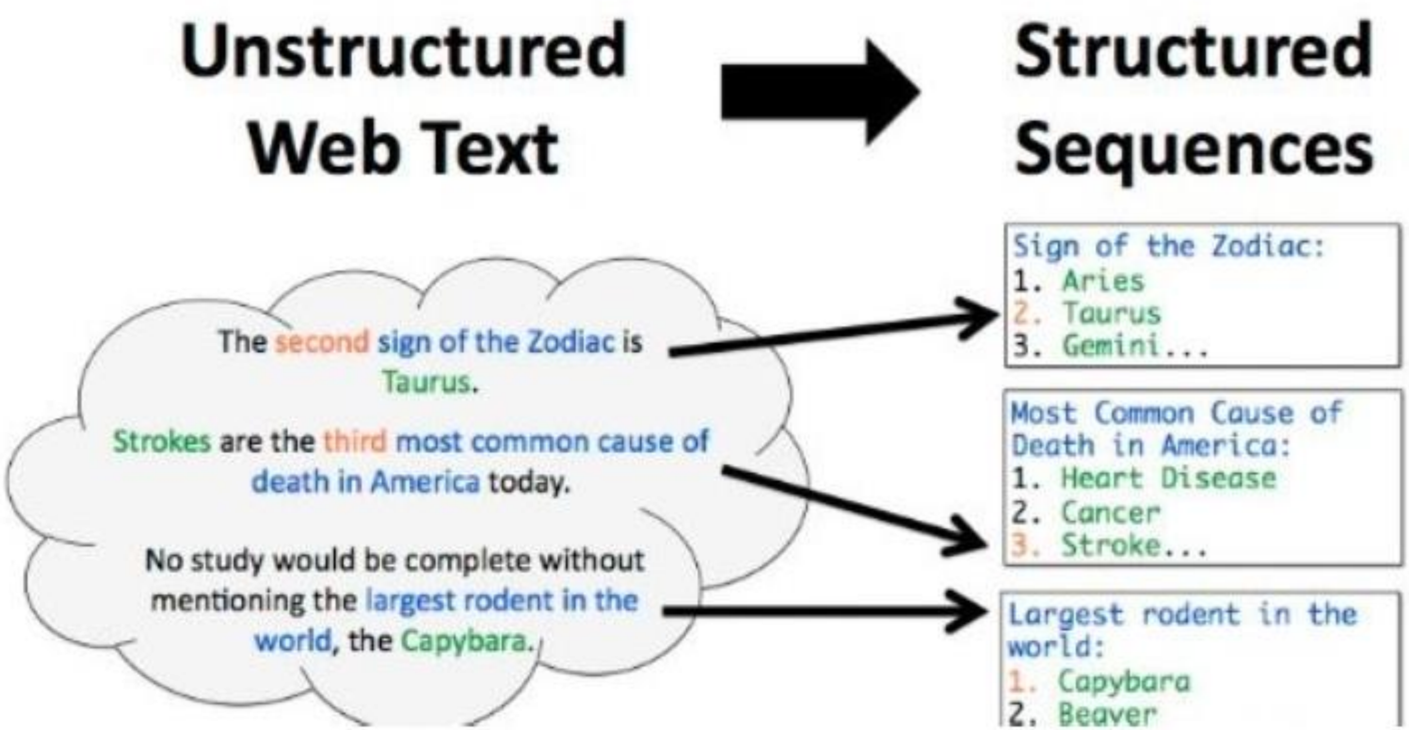


Named Entity Recognition



Why Sequence Labeling?

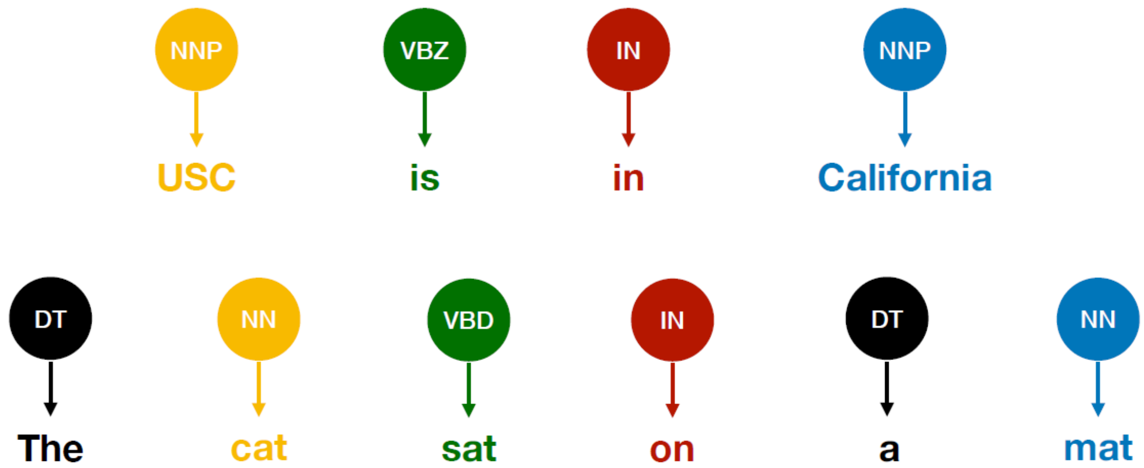
- Helpful to perform downstream information extraction tasks



Part-of-Speech Tagging

- A structured prediction task for NL sequences

- Grammatical word Classes are the label-space
- Reveal useful information about the syntactic role of a word (and its neighbors)



- Closed vs Open classes

Open-class	Noun, verb, adjective, . . .
Closed-class	Preposition, conjunction, determiner, . . .

Part-of-Speech Tagging

- Challenges
  - The same word can have different syntactic functions: duck
  - Ambiguity

Time	**flies**	**like**	**an**	**arrow**	
NN	VBZ	IN	DT	NN	(Penn Treebank tags)
NN	NNS	VBP	DT	NN	
VB	NNS	IN	DT	NN	

  - Long distance dependencies

Flying	planes	can	be	dangerous
$\begin{Bmatrix} \text{VBG} \\ \text{JJ} \end{Bmatrix}$	NNS	MD	VB	JJ

Flying	planes	is	dangerous
VBG	NNS	VBZ	JJ

Flying	planes	are	dangerous
JJ	NNS	VBP	JJ

  - How many tags?

Penn Tree Bank Tagset

- 45 tags

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	: ; ... – -

Named Entity Tagging

- The goal is finding spans of text that constitute proper names and tag the type of the entities
  - Common entity tags: PER (person), LOC (location), ORG (organization), or GPE (geo-political entity).

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	**Mt. Sanitas** is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states	**Palo Alto** is raising the fees for parking.

- Helpful for question answering, linking text to information, etc
- More challenging than POS tagging
  - What is an entity and what is not?
  - The boundary for an entity
  - Ambiguity: JFK

A Simple Baseline for POS Tagging

- Many words might be easy to disambiguate
- Most Frequent Class: Assign each token (word) to the class it occurred most in the training data. (e.g. student/NN)
  - Entirely discarding contextual information

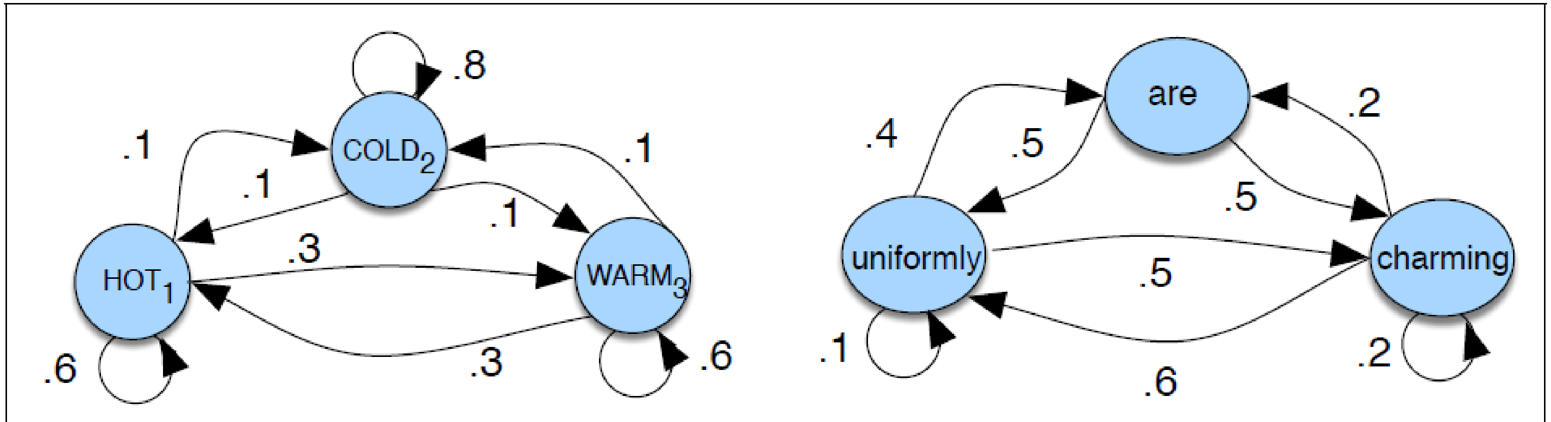
- How accurate do you think this baseline would be at tagging words?
  - 92.34% on WSJ corpus
- Is this a good performance:
  - The average English sentence has 14 words
- It is an unsolved task:
  - SOTA: 97%
  - Highly depends on the domain

Sequence Labeling for POS

- The function (or POS) of a word depends on its context
  - The/DT back/ADJ door/NN
  - On/IN my/PRP$ back/NN
  - Win/VB the/DT voters/NNS back/RP
- Certain POS combinations are extremely unlikely
  - *<JJ, DT>* ("good the")
  - *<DT, IN>* ("the in")
- Better to make predictions on entire sentences instead of individual words
  - Sequence labeling modeling: hidden Markov models
  - 96% on POS Tagging

Markov Chain

- A model for probabilities of sequences of random variables (states), each of which can take on values from some set and transition from one to another
- Model Parameters: transition probabilities (A) and initial probability distribution (π)



(a) (b)

- The future state only depends on the current state

$$P(q_i = a | q_1 \ldots q_{i-1}) = P(q_i = a | q_{i-1})$$

Markov Sequence

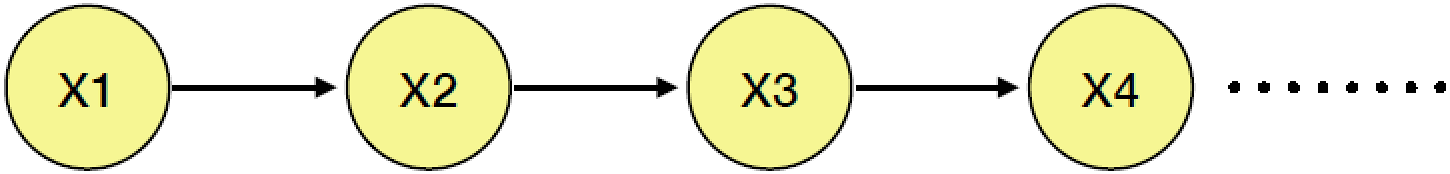
- Consider a sequence of random variable with length m: $X_1, X_2, \ldots, X_m$
- Each variable can take a value from a discrete set with the size K
- Each variable depends on the previous variables
- We want to model the joint probability

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m)$$

Markov Assumption

- Limited conditional dependence

$$
\begin{aligned}
& P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) \\
= \quad & P(X_1 = x_1) \prod_{j=2}^{m} P(X_j = x_j | X_1 = x_1, \ldots, X_{j-1} = x_{j-1}) \\
= \quad & P(X_1 = x_1) \prod_{j=2}^{m} P(X_j = x_j | X_{j-1} = x_{j-1})
\end{aligned}
$$



- A generative model

Markov Model for Sequence Labeling

- We need a pair of sequences

$$S = S_i, S_2, \ldots, S_n$$

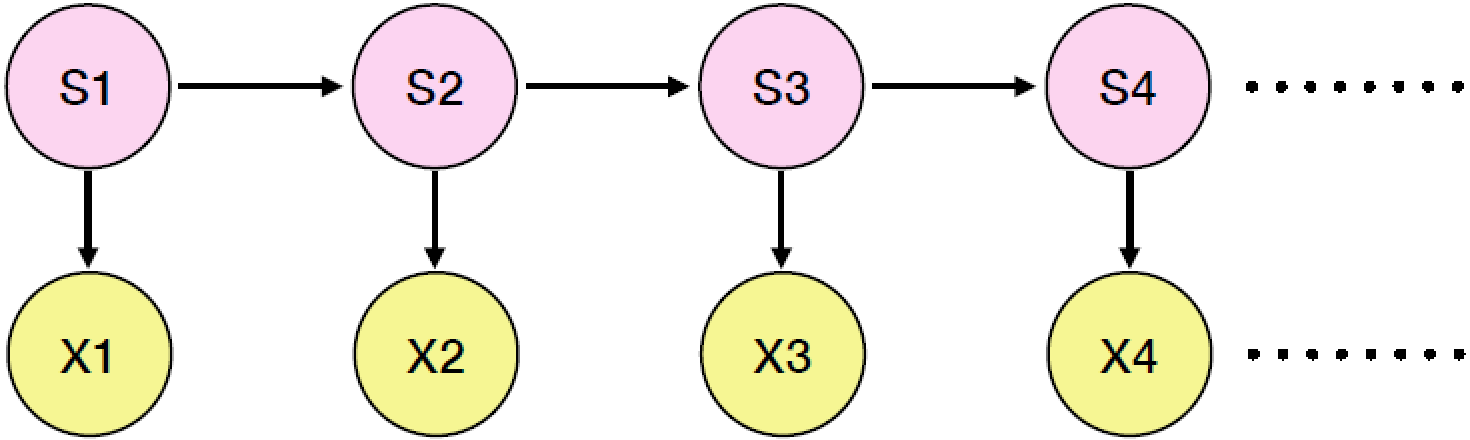
$$X = X_i, X_2, \ldots, X_n$$

- Hidden Markov models: allow us to jointly reason over both X and S

Hidden Markov Model

- Given the two sequences of random variables $X_1, X_2, \ldots, X_m$ and $S_1, S_2, \ldots, S_m$, where **X** corresponds to “observations” and **S** corresponds to the underlying “states” that generate the observations, model the joint probability:

$$P(X_1 = x_1, \ldots, X_m = x_m, S_1 = s_1, \ldots, S_m = s_m)$$



HMM Assumptions

- Markov Assumption on **S**

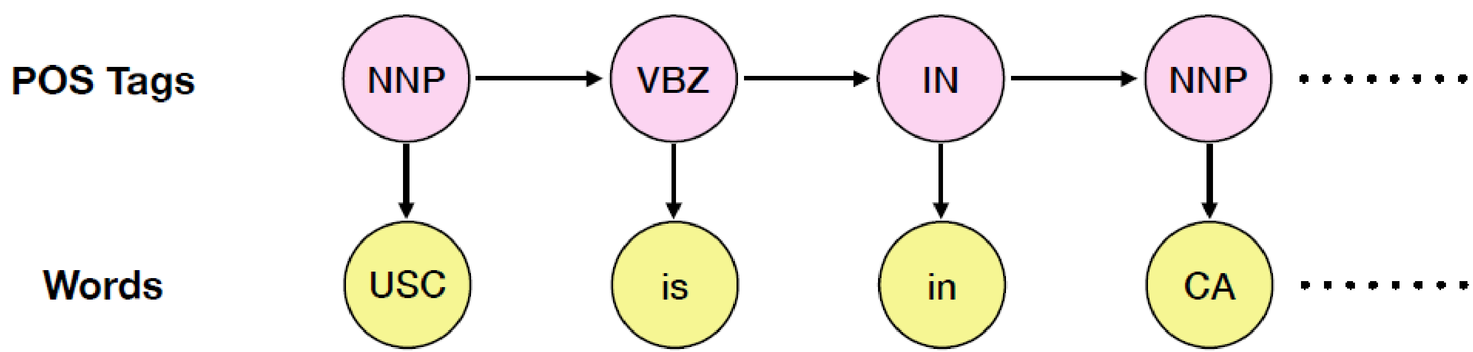
$$P(S_j = s_j \mid S_{j-1} = s_{j-1}, \ldots, S_1 = s_1) = P(S_j = s_j \mid S_{j-1} = s_{j-1})$$

Transition Probabilities

- Conditional independence of **X** and **S**

$$P(X_1 = x_j, \ldots, X_m = x_m \mid S_1 = s_1, \ldots, S_m = s_m) = \prod_{j=1}^{m} P(X_j = x_j \mid S_j = s_j)$$

Emission Probabilities



$$P(S_3 = \text{IN} \mid S_2 = \text{VBZ}, S_1 = \text{NNP}) = P(S_3 = \text{IN} \mid S_2 = \text{VBZ})$$

$$P(\text{USC is in CA} \mid \text{NNP VBZ IN NNP}) = P(\text{USC} \mid \text{NNP})P(\text{is} \mid \text{VBZ})P(\text{in} \mid \text{IN})P(\text{CA} \mid \text{NNP})$$

HMM Assumptions

- Joint Distribution of Sequence Pairs in HMMs

$$P(X_1 = x_j, \ldots, X_m = x_m, S_1 = s_1, \ldots, S_m = s_m)$$

$$= \boxed{P(X_1 = x_j, \ldots, X_m = x_m \mid S_1 = s_1, \ldots, S_m = s_m)}$$

Output Independence

$$\boxed{\times P(S_1 = s_1, \ldots, S_m = s_m)}$$

Markov Assumption

$$= \prod_{j=1}^{m} P(X_j = x_j \mid S_j = s_j)$$

How to model $P(X_j = x_j \mid S_j = s_j)$
and $P(S_j = s_j \mid S_{j-1} = s_{j-1})$?

$$\times P(S_1 = s_1) \prod_{j=1}^{m} P(S_j = s_j \mid S_{j-1} = s_{j-1})$$

Homogenous HMM

- We include an additional assumption

$$P(S_j = s_j \mid S_{j-1} = s_{j-1}) = t(s_j \mid s_{j-1})$$

$$P(X_j = x_j \mid S_j = s_j) = e(x_j \mid s_j)$$

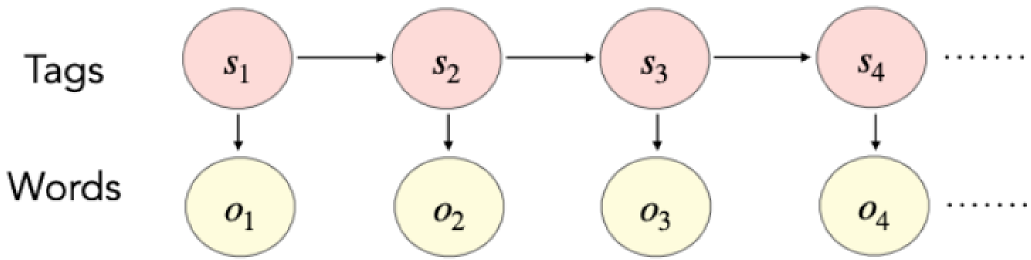
- the transition and emission probabilities do not depend on the position in the Markov chain (do not depend on the index) *j*

$$p(x_1 \ldots x_m, s_1 \ldots s_m) = t(s_1) \prod_{j=2}^{m} t(s_j|s_{j-1}) \prod_{j=1}^{m} e(x_j|s_j)$$

- Initial state parameters $t(s)$ for $s \in \{1, 2, \ldots, k\}$
- Transition parameters $t(s'|s)$ for $s, s' \in \{1, 2, \ldots, k\}$
- Emission parameters $e(x|s)$ for $s \in \{1, 2, \ldots, k\}$ and $x \in \{1, 2, \ldots, o\}$

HMM Example

Sequence probability



Dummy start state

s_{t+1}

s_t		DT	NN
	∅	0.8	0.2
	DT	0.2	0.8
	NN	0.3	0.7

o_t

	the	cat
DT	0.9	0.1
NN	0.5	0.5

What is the joint probability P(the cat, DT NN)?

A) $(0.8 * 0.8) * (0.9 * 0.5)$

B) $(0.2 * 0.8) * (0.9 * 0.5)$

C) $(0.3 * 0.7) * (0.5 * 0.5)$

HMM Learning

- We collect a fully observed dataset $\{X_i, S_i\}_{i=1}^{N}$

Training set:

1 Pierre/NNP Vinken/NNP ,/, 61/CD years/NNS old/JJ ,/
join/VB the/DT board/NN as/IN a/DT nonexecutive/JJ di
Nov./NNP 29/CD ./.
2 Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsev
N.V./NNP ,/, the/DT Dutch/NNP publishing/VBG group/
3 Rudolph/NNP Agnew/NNP ,/, 55/CD years/NNS old/JJ
chairman/NN of/IN Consolidated/NNP Gold/NNP Fields/N
,/, was/VBD named/VBN a/DT nonexecutive/JJ director/
this/DT British/JJ industrial/JJ conglomerate/NN ./.
...
38,219 It/PRP is/VBZ also/RB pulling/VBG 20/CD peopl
of/IN Puerto/NNP Rico/NNP ,/, who/WP were/VBD help
Huricane/NNP Hugo/NNP victims/NNS ,/, and/CC sendin
them/PRP to/TO San/NNP Francisco/NNP instead/RB ./

Maximum Likelihood Estimate:

$$\max_{t(\cdot|\cdot),e(\cdot|\cdot)} \prod_{i=1}^{N} P(X_i, S_i)$$

$$t(s' \mid s) = \frac{\text{count}(s \rightarrow s')}{\text{count}(s)}$$

$$e(x \mid s) = \frac{\text{count}(s \rightarrow x)}{\text{count}(s)}$$

HMM Learning Example

1. the/DT cat/NN sat/VBD on/IN the/DT mat/NN

2. Princeton/NNP is/VBZ in/IN New/NNP Jersey/NNP

3. the/DT old/NN man/VB the/DT boats/NNS

$$t(\textbf{NN} \mid \textbf{DT}) = \frac{3}{4}$$

$$e(\textbf{cat} \mid \textbf{NN}) = \frac{1}{3}$$

Maximum Likehood Estimate:

$$\max_{t(\cdot|\cdot), e(\cdot|\cdot)} \prod_{i=1}^{N} P(X_i, S_i)$$

$$t(s' \mid s) = \frac{\text{count}(s \rightarrow s')}{\text{count}(s)}$$

$$e(x \mid s) = \frac{\text{count}(s \rightarrow x)}{\text{count}(s)}$$

Challenge of Unknown Words

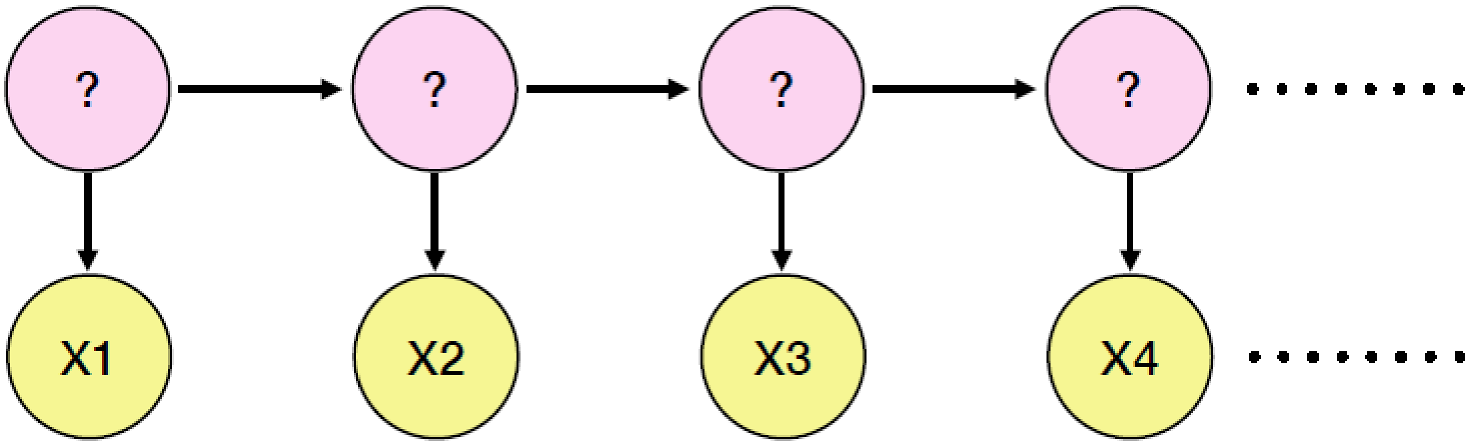
- Unknown words: Zero probabilities!
- Pseudo words: 1993 -> 4digitword, Jago -> initCAP

Word class	Example	Intuition
twoDigitNum	90	Two digit year
fourDigitNum	1990	Four digit year
containsDigitAndAlpha	A8956-67	Product code
containsDigitAndDash	09-96	Date
containsDigitAndSlash	11/9/89	Date
containsDigitAndComma	23,000.00	Monetary amount
containsDigitAndPeriod	1.00	Monetary amount,percentage
othernum	456789	Other number
allCaps	BBN	Organization
capPeriod	M.	Person name initial
firstWord	first word of sentence	no useful capitalization information
initCap	Sally	Capitalized word
lowercase	can	Uncapitalized word
other	,	Punctuation marks, all other words

The mapping to pseudo words used by Bikel et. al (1999).

Decoding with HMM

- Given an input sequence $x_1, \ldots, x_m$ compute:



$$S^* = \arg\max_{s_1,\ldots,s_m} p(x_1,\ldots,x_m,s_1,\ldots,s_m) = t(s_1)\prod_{j=2}^{m} t(s_j \mid s_{j-1}) \prod_{j=1}^{m} e(x_j \mid s_j)$$

How can we maximize this over all state sequences?

- Bruteforce search: 45^14