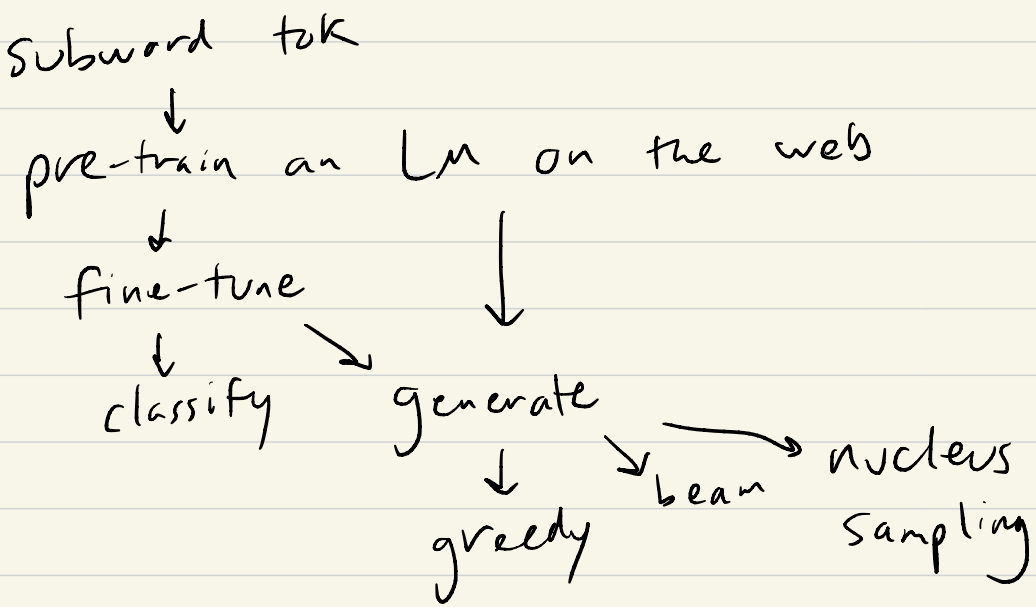
# CS371N Lecture 14
# Sequence Modeling, POS Tagging

## Announcements

- A2 solutions posted
- A3 due today

## Recap Language modeling



Today Structured prediction: sequence modeling
POS tagging

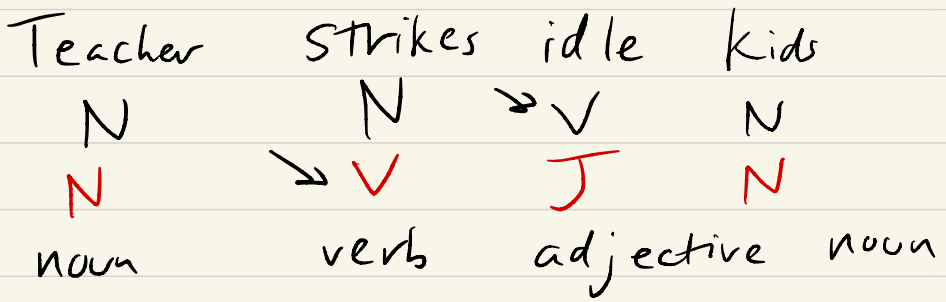
## Part-of-speech tagging

Input sentence: $x_1, \ldots, x_n$

Output: POS tags $y_1, \ldots, y_n$ for each word

Different than sentiment class. One output per word

What are POS tags (and why?)




Text-to-speech: record

N
V

## POS tags

Open-class: new words with these tags are always emerging

Closed-class: known set

## Open-class

(N) Nouns → Proper (Google)
→ common (cat)
plural vs. singular

(V) verbs tense, "person" (inflection)

(J) adjectives idle, yellow

(RB) adverbs swiftly

## Closed-class

(DT) Determiners: articles (the, a)
Some, many

(CD) Cardinal: numbers

Prepositions: up, on, in, ...

Particles: made up

Auxiliaries: had Modals: could/would/should

(1) What tags are possible for each word?

Fed raises interest rates 0.5 percent

Fed
- VBD — I fed the cat
- NNP — proper noun
- VBN — "fed up"

raises
- NNS — plural
- VBZ — (3rd person sg present)

interest
- NN
- VBP — present — "I interest you"
- VB — infinitive — "I want NLP to interest me"

rates
- NNS
- VBZ

Interps: Standard, weird, weirder

0.5
- CD

percent
- NN

## Two ways to tag

- Using classifiers?
- Using Hidden Markov Models

## Classifiers POS tags $y$

Multiclass LR:

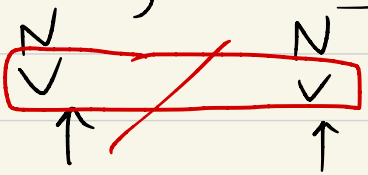
$P(y_i = t \mid \bar{x})$ iterate over the sequence

letter counting in A3

$$P(\bar{y} \mid \bar{x}) = \prod_{i=1}^{n} P(y_i \mid \bar{x})$$

independent classification of each

Why modeling the sequence?




bad combo

Fed raises interest rates

$P(y_2 = V \mid \bar{x})$ is high $\Rightarrow V$

$P(y_3 = V \mid \bar{x})$ is high $\Rightarrow V$

$P(y_2 = V, y_3 = V \mid \bar{x})$ is low

↓ we don't model this

## Hidden Markov Models

⊕ model the sequence $(y_3 \mid y_2)$

⊖ simple generative model

(CRFs: discriminative HMMs, Transformer + HMM)

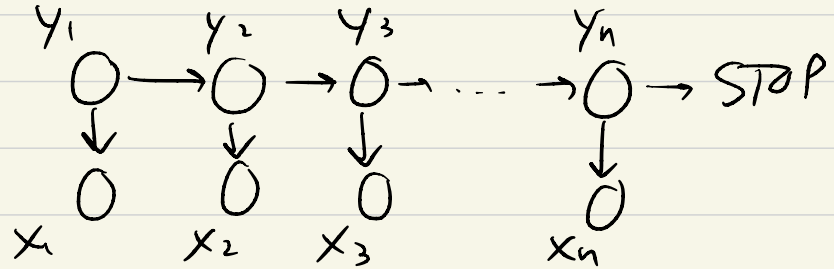
$y_1 \cdots y_n$ $x_1 \cdots x_n$

Generative model $P(\bar{y} | \bar{x})$

Tags $y_i \in \mathcal{T}$ (tagset) words $x_i \in V$

HMM:

$$P(\bar{y}, \bar{x}) = P(y_1) P(x_1 | y_1) P(y_2 | y_1) P(x_2 | y_2) P(y_3 | y_2) P(x_3 | y_3) \cdots P(\text{STOP} | y_n)$$




Assumptions:

1. Each $y$ depends only on the previous $y$ (Markov)
2. Each $x_i$ is indep. of all else given $y_i$

Goal: We model $P(\bar{y}, \bar{x})$ (joint, conditional) but we care about $P(\bar{y} \mid \bar{x})$

(you give me a sentence $\bar{x}$ and I give you tags)

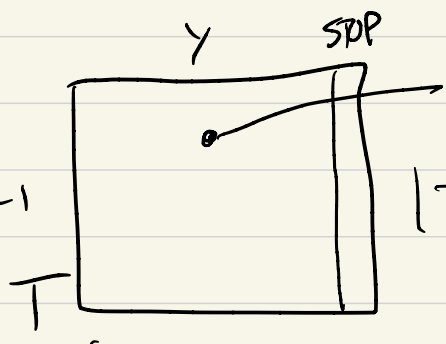
## Parameters Three types

$P(y_1)$
initial dist


$|\tau|$ - len vector

sums to 1

$P(y_i \mid y_{i-1})$
transition probs




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

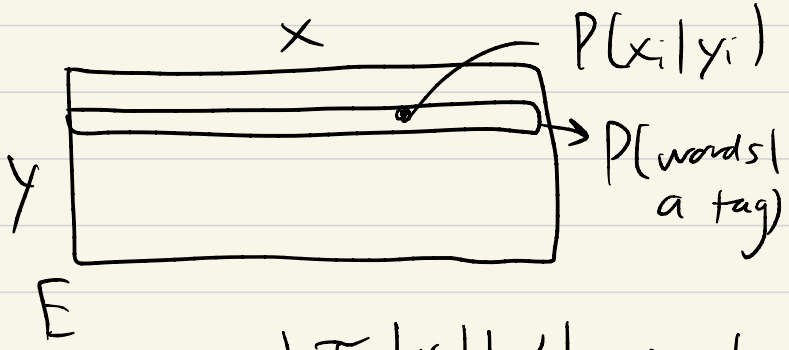
$|\tau| \times |\tau| + 1$ (STOP)

each row is a dist.

P(rates|N)

$P(x_i | y_i)$

emissions




$|T| \times |V|$ vocab

store distributions explicitly