CS 371N Lecture 15
HMMs, Viterbi

Announcements

- A4 due in a week
- Midterm due next Thurs
- OPTIONAL: Independent FP proposals due after midterm

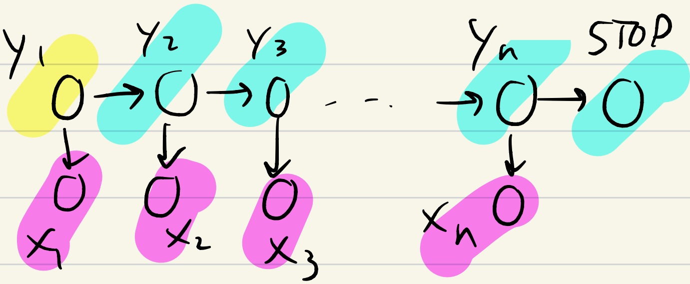
Recap POS tagging

N	N	N	N
V	V	V	V
Fed	raises	interest	rates

There are constraints on what makes a well-formed tag sequence (e.g., rare to have V-V) (discrim.: $P(\bar{y}|\bar{x})$)

<u>HMMs</u> Generative model of sequences

$$P(\bar{y}, \bar{x}) = P(y_1)\, P(x_1|y_1)\, P(y_2|y_1)\, P(x_2|y_2) \cdots$$



Parameters:

Initial $P(y_1)$ vector

Transitions $P(y_i | y_{i-1})$ matrix

Emissions $P(x_i | y_i)$ matrix

$P(\text{go} | V) = 0.2$

$P(\text{is} | V) = 0.2$ ...

$P(\text{eat} | V) = 0.1$

Goal: compute $P(\bar{y}|\bar{x})$ given a sequence $\bar{x}$

Ex $\mathcal{T} = \{N, V, STOP\}$

$\mathcal{V} = \{they, can, fish\}$

Initial $P(y) = \begin{cases} 1.0 & N \\ 0 & V \\ 0 & STOP \end{cases}$

Transitions $P(y_i | y_{i-1}) =$

	N	V	STOP
N	1/5	3/5	1/5
V	1/5	1/5	3/5

Emissions $P(x_i | y_i) =$

	they	can	fish
N	1	0	0
V	0	1/2	1/2

① Compute the probability of

$\begin{pmatrix} N & V & V & STOP \\ they & can & fish & \end{pmatrix}$

$P(STOP|V)$

$P(y_1 = N) \cdot P(y_2 = V | y_1 = N) \cdot P(V|V) \cdot$

$P(x_1 = they | y_1 = N) \; P(x_2 = can | y_2 = V) \; P(fish|V)$

$$1.0 \cdot \tfrac{3}{5} \cdot \tfrac{1}{5} \cdot \tfrac{3}{5} \cdot 1.0 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \frac{9}{500}$$

② Is there a higher-scoring tag sequence for "they can fish"?

Goal of HMMs

HMMs model $P(\bar{y}, \bar{x})$

They are not good generative models of text

What we use them for is $P(\bar{y} \mid \bar{x})$

$$P(\bar{y} \mid \bar{x}) = \frac{P(\bar{y} \mid \bar{x}) \cdot P(\bar{x})}{P(\bar{x})} = \frac{P(\bar{y}, \bar{x})}{P(\bar{x})}$$

$$\sum_{\bar{y}} P(\bar{y}, \bar{x})$$

$$P(\bar{y} \mid \bar{x}) \propto P(\bar{y}, \bar{x})$$

proportional to

What this means:

$$\underset{\tilde{y}}{\operatorname{argmax}}\ P(\tilde{y} \mid \bar{x}) = \underset{\tilde{y}}{\operatorname{argmax}}\ P(\tilde{y}, \bar{x})$$

$$= \underset{\tilde{y}}{\operatorname{argmax}}\ \log P(\tilde{y}, \bar{x})$$

$$\tilde{y} = \tilde{y}_1, \tilde{y}_2, \tilde{y}_3 \ldots \tilde{y}_n$$

$$= \underset{\tilde{y}_1, \tilde{y}_2 \ldots \tilde{y}_n}{\operatorname{argmax}}\ \log P(\tilde{y}_1) + \log P(x_1 \mid \tilde{y}_1) + \log P(\tilde{y}_2 \mid \tilde{y}_1) + \log P(x_2 \mid \tilde{y}_2) + \ldots$$

Viterbi Algorithm

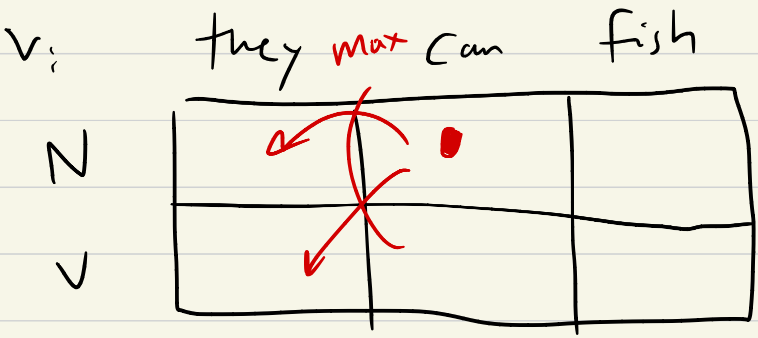
Define $V_i(\tilde{y}_i)$ as the chart

i is index from 1 to n

$\tilde{y}_i$ is a tag in $\mathcal{T}$

$n \times |\mathcal{T}|$ matrix

$v_i(\tilde{y}_i)$ = log prob of the best sequence of tags ending in $\tilde{y}$ at index i



$V_2(N)$

max over $\tilde{y}_1 \in \{N, V\}$

compute V_i $\longrightarrow$

Initial

$$v_1(\tilde{y}_1) = \underbrace{\log P(x_1 \mid \tilde{y}_1)}_{} + \log P(\tilde{y}_1)$$

emission initial

Recurrent Compute v_i using v_{i-1}

$$v_i(\tilde{y}_i) = \log P(x_i \mid \tilde{y}_i) \text{ emission}$$

$$+ \max_{\tilde{y}_{prev}} \left[\log P(\tilde{y}_i \mid \tilde{y}_{prev}) + v_{i-1}(\tilde{y}_{prev}) \right]$$

$$v_2(\tilde{y}_2) = \log P(x_2 \mid \tilde{y}_2)$$

$$+ \max_{\tilde{y}_1} \left[\log P(\tilde{y}_2 \mid \tilde{y}_1) + v_1(\tilde{y}_1) \right]$$

End $v_n(\tilde{y}_n)$ = recurrent formula + $\log P(STOP \mid \tilde{y}_n)$

Given v chart, extract best sequence with backpointers

$$\log P(y_1) = \begin{cases} N & -1 \\ V & -1 \end{cases} \leftarrow \text{log probs}$$

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

	N	V	STOP
N	-2	-1	-1
V	-1	-1	-2

$\log P(x_i \mid y_i) =$

	they	fish	can	dog	a
N	-1	-1	-3	-	-
V	-3	-1	-1	-	-

$\bar{x} =$ they can fish

	1	2
V	Initial	Recurrent
N	-1 $P(N)$ -1 $P(\text{they} \mid N)$ $\boxed{-2}$	$P(\text{can} \mid N) + \max \begin{cases} P(N \mid N) + v_1(N) = \boxed{-4} \\ \quad -2 \qquad -2 \\ P(N \mid V) + v_1(V) = -5 \\ \quad -1 \qquad -4 \end{cases}$ $\boxed{-3}$ $\boxed{-7}$
V	-1 $P(V)$ -3 $P(\text{they} \mid V)$ $\boxed{-4}$	$P(\text{can} \mid V) + \max \begin{cases} P(V \mid N) + v_1(N) \\ \quad -1 \qquad -2 \\ P(V \mid V) + v_1(V) \\ \quad -1 \qquad -4 \end{cases}$ -1 $\boxed{-4}$

STOP

	3 fish
N	$-6 \Rightarrow -7$ w/ STOP $y_{prev} = V$
V	$-6 \Rightarrow -8$ w/ STOP $y_{prev} = V$

Sequence: N V N $\log P = -7$

Markov property allowed us to do this efficiently!

Parameter Estimation

Suppose we have labeled data

$\bar{y}^{(1)}, \bar{x}^{(1)}$ Estimate params by

$\bar{y}^{(2)}, \bar{x}^{(2)}$ counting + normalizing

$\vdots$

$\bar{y}^{(N)}, \bar{x}^{(N)}$

$$\text{Initial prob}(N) = \frac{\text{number of } \bar{y}^{(i)} \text{ w/ } y_1 = N}{\text{total num exs.}}$$

$$\text{Transition prob}(N \to V) = \frac{\text{number of times we saw } N \to V}{\text{number of times we saw } N}$$

These maximize $\sum_{i=1}^{N} \log P(\bar{y}^{(i)}, \bar{x}^{(i)})$

Data: English Penn Treebank

44 tags

Assign each word its most frequent tag = 90% acc.

Trigram HMM tagger: 95%

BERT: 97.5%