CS 6501 Natural Language Processing

Sequence Labeling (II)

Yangfeng Ji

September 17, 2018

Department of Computer Science
University of Virginia

UNIVERSITY of VIRGINIA | ENGINEERING

Important Dates

- Project 1 due: Sept. 23, 11:59PM
- Group project proposal due: Oct. 7, 11:59PM
- Group project proposal presentation: Oct. 10 & Oct. 15

Other things

- Project 1 submission is open on Collab

Overview

1. Conditional Random Fields
2. Inference
3. Parameter Estimation

POS Tagging

Example

`Rain and wind from Florence will pop up on Tuesday`

POS tagging

$$\texttt{Rain}_{\texttt{NN}}\ \texttt{and}_{\texttt{CC}}\ \texttt{wind}_{\texttt{NN}}\ \texttt{from}_{\texttt{IN}}\ \texttt{Florence}_{\texttt{NNP}}\ \texttt{will}_{\texttt{MD}}\ \texttt{pop}_{\texttt{VB}}\ \texttt{up}_{\texttt{RP}}\ \texttt{on}_{\texttt{IN}}\ \texttt{Tuesday}_{\texttt{NNP}}$$

- NN: Noun, singular or mass
- NNP: Proper noun, singular
- IN: Preposition or subordinating conjunction
- CC: Coordinating conjunction
- MD: Modal
- VB: Verb, base form
- RP: Particle

Sequence Labeling

Example

[Atlantis]$_{\text{MSIC}}$ touched down at [Kennedy Space Center]$_{\text{LOC}}$

Atlantis	touched	down	at	Kennedy	Space	Center	.
B_{MSIC}	O	O	O	B_{LOC}	I_{LOC}	I_{LOC}	O

Tag set

- B: beginning
- I: inside
- O: outside

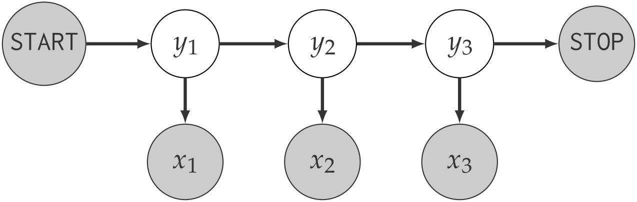
Category

- Person
- Location
- Organization
- Msic

Hidden Markov Models

$$P(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1} \left\{ P(y_i | y_{i-1}) P(x_i | y_i) \right\} \qquad (1)$$

Graphical model



- $\boldsymbol{x}$: observation (e.g., sentences)
- $\boldsymbol{y}$: hidden variables (e.g., POS sequences)

Generative Models

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\boldsymbol{x}|\boldsymbol{y}) \cdot P(\boldsymbol{y}) \tag{2}$$

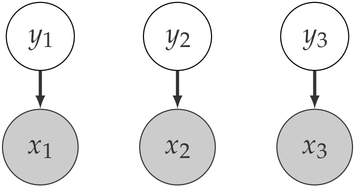
Generative Models

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\boldsymbol{x}|\boldsymbol{y}) \cdot P(\boldsymbol{y}) \tag{2}$$

Factorization

$$P(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1} \underbrace{P(x_i|y_i)}_{\text{Emission probability}} \tag{3}$$

Graphical model



Discriminative Models: Logistic Regression

$$P(y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, y))}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, y'))} \tag{4}$$

Question

How to build a discriminative sequential model $p(\boldsymbol{y}|\boldsymbol{x})$?

- $\boldsymbol{y} \in \mathcal{Y}^T$ is a sequence

Conditional Random Fields

Logistic Regression

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}' \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \tag{5}$$

Huge $\mathcal{Y}^T$ causes the problems on

- decoding $\arg\max_{\boldsymbol{y}' \in \mathcal{Y}^T} P(\boldsymbol{y}'|\boldsymbol{x})$

Logistic Regression

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\underbrace{\sum_{\boldsymbol{y}' \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))}_{\text{partition function } Z}} \tag{5}$$

Huge $\mathcal{Y}^T$ causes the problems on

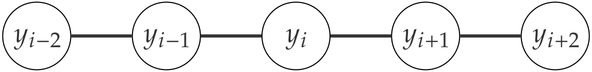
- decoding $\arg\max_{\boldsymbol{y}' \in \mathcal{Y}^T} P(\boldsymbol{y}'|\boldsymbol{x})$
- computing the partition function with $|\mathcal{Y}^T| = K^T$ possible values

Markov Property

Global feature function:

$$f(x, y) \tag{6}$$

Markov assumption:



- Conditional independence
- Factorization over cliques

Decomposition of $f(x, y)$

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \underbrace{f_i(\boldsymbol{x}, y_{i-1}, y_i)}_{\text{local feature function}} \tag{7}$$

- $f_i(\boldsymbol{x}, y_{i-1}, y_i)$ captures the transition from y_i to y_i
- i: the position to be tagged
- $y_i \in \mathcal{Y}$: POS tag at position i
- $y_{i-1} \in \mathcal{Y}$: POS tag at position $i - 1$
- $\boldsymbol{x}$: the entire sentence

Local Feature Function: Example

- standard features

[Lafferty et al., 2001]

Local Feature Function: Example

- standard features
- whether a spelling begins with upper case letter,
  - `IBM`, `Virginia`: PROPER NOUN

[Lafferty et al., 2001]

Local Feature Function: Example

- standard features
- whether a spelling begins with upper case letter,
  - `IBM, Virginia`: PROPER NOUN
- whether it ends in one of the following suffixes:
  - `-ies` e.g., `parties`: PROPER NOUN, PLURAL
  - `-ly` e.g., `extremely, loudly`: ADVERB
  - `-ing` e.g., : VERB, GERUND OR PRESENT PARTICIPLE
  - ...

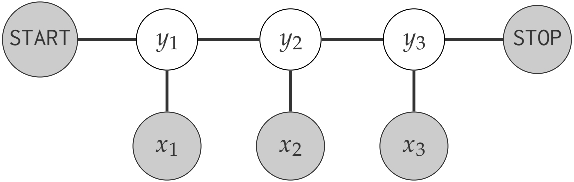
[Lafferty et al., 2001]

Example

$\texttt{Rain}_{\text{NN}}$ $\texttt{and}_{\text{CC}}$ $\texttt{wind}_{\text{NN}}$ $\texttt{from}_{\text{IN}}$ $\texttt{Florence}_{\text{NNP}}$ $\texttt{will}_{\text{MD}}$ $\texttt{pop}_{\text{VB}}$ $\texttt{up}_{\text{RP}}$ $\texttt{on}_{\text{IN}}$ $\texttt{Tuesday}_{\text{NNP}}$

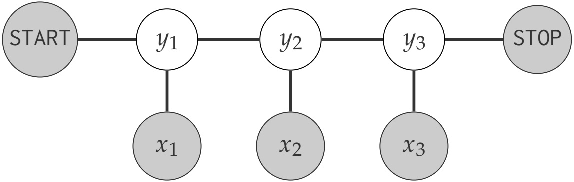
Graphical Model Representation

Conditional Random Fields:

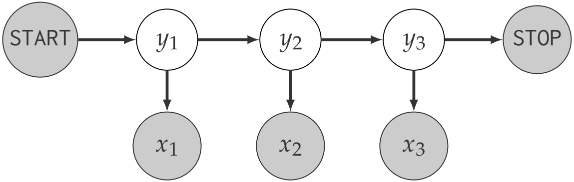


Graphical Model Representation

Conditional Random Fields:



Hidden Markov Models:



CRF vs HMM

HMM factorization

$$P(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1} \left\{ P(y_i | y_{i-1}) P(x_i | y_i) \right\} \tag{8}$$

Let

- $\psi(y_1, y_2) = P(y_1)P(y_2|y_1)P(x_1|y_1)P(x_2|y_2)$
- $\psi(y_{i-1}, y_i) = P(y_i|y_{i-1})P(x_i|y_i), \forall i > 2$

CRF vs HMM

HMM factorization

$$P(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1} \left\{ P(y_i | y_{i-1}) P(x_i | y_i) \right\} \tag{8}$$

Let

- $\psi(y_1, y_2) = P(y_1)P(y_2|y_1)P(x_1|y_1)P(x_2|y_2)$
- $\psi(y_{i-1}, y_i) = P(y_i|y_{i-1})P(x_i|y_i), \forall i > 2$

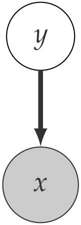
$$P(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{T} \psi(y_{i-1}, y_i) \tag{9}$$

with $Z = 1$

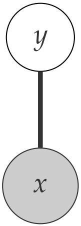
Directed vs. Undirected Graphical Models

- Undirected GMs allow more flexibile definitions
- Directed GMs can always be converted into an undirected GMs

$$P(x, y) = P(x|y)P(y)$$



$$P(x, y) = \psi(x, y)/Z$$

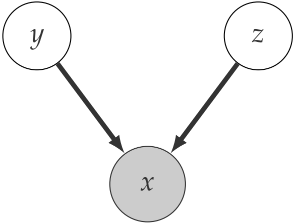


[Bishop, 2006, Chap. 8]

Directed vs. Undirected Graphical Models (II)

- Directed GMs can give more concise definitions

$$P(x, y, z) = P(x|y, z)P(y)P(z)$$

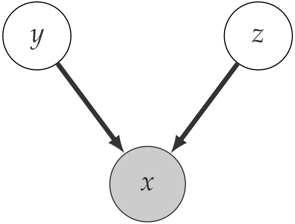


[Bishop, 2006, Chap. 8]

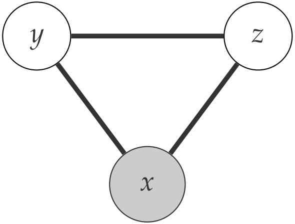
Directed vs. Undirected Graphical Models (II)

- Directed GMs can give more concise definitions

$$P(x, y, z) = P(x|y, z)P(y)P(z)$$



$$P(x, y, z) = \psi(x, y, z)/Z$$



[Bishop, 2006, Chap. 8]

Inference

Decode $P(\boldsymbol{y}|\boldsymbol{x})$

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} f_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{10}$$

Decode $P(\boldsymbol{y}|\boldsymbol{x})$

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{10}$$

$$\begin{aligned}
\arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} P(\boldsymbol{y}|\boldsymbol{x}) &= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}'\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \\
&= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))
\end{aligned}$$

Decode $P(\boldsymbol{y}|\boldsymbol{x})$

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{10}$$

$$\begin{aligned}
\arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} P(\boldsymbol{y}|\boldsymbol{x}) &= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}'\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \\
&= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})) \\
&= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})
\end{aligned}$$

Decode $P(\boldsymbol{y}|\boldsymbol{x})$

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{10}$$

$$\begin{aligned}
\arg\max_{\boldsymbol{y} \in \mathcal{Y}^T} P(\boldsymbol{y}|\boldsymbol{x}) &= \arg\max_{\boldsymbol{y} \in \mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}' \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \\
&= \arg\max_{\boldsymbol{y} \in \mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})) \\
&= \arg\max_{\boldsymbol{y} \in \mathcal{Y}^T} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{y} \in \mathcal{Y}^T} \boldsymbol{\theta}^\top \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i)
\end{aligned}$$

Decode $P(y|x)$

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{10}$$

$$\begin{aligned}
\arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} P(\boldsymbol{y}|\boldsymbol{x}) &= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}'\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}'))} \\
&= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})) \\
&= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \boldsymbol{\theta}^\top \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \\
&= \arg\max_{\boldsymbol{y}\in\mathcal{Y}^T} \sum_{i=1}^{T} \boldsymbol{\theta}^\top \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i)
\end{aligned}$$

Factorization

Factorize $\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ with respect to timestep i

$$
\begin{aligned}
\sum_{i=1}^{T} \boldsymbol{\theta}^\top \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) = & \underbrace{\sum_{j \le i-1} \boldsymbol{\theta}^\top \boldsymbol{f}_j(\boldsymbol{x}, y_{j-1}, y_j)}_{\text{past}} \\
& + \underbrace{\boldsymbol{\theta}^\top \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i)}_{\text{present}} \\
& + \underbrace{\sum_{k \ge i+1} \boldsymbol{\theta}^\top \boldsymbol{f}_k(\boldsymbol{x}, y_{k-1}, y_k)}_{\text{future}}
\end{aligned} \tag{11}
$$

Viterbi Algorithm

$$s_i(k, k') = \boldsymbol{\theta}^\top \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1} = k', y_i = k)$$

Algorithm 11 The Viterbi algorithm. Each $s_m(k, k')$ is a local score for tag $y_m = k$ and $y_{m-1} = k'$.

for $k \in \{0, \ldots K\}$ **do**
  $v_1(k) = s_1(k, \Diamond)$
for $m \in \{2, \ldots, M\}$ **do**
  for $k \in \{0, \ldots, K\}$ **do**
    $v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$
    $b_m(k) = \operatorname{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$
$y_M = \operatorname{argmax}_k s_{M+1}(\blacklozenge, k) + v_M(k)$
for $m \in \{M-1, \ldots 1\}$ **do**
  $y_m = b_m(y_{m+1})$
return $\boldsymbol{y}_{1:M}$

[Eisenstein, 2018]

Parameter Estimation

Parameter Estimation: Logistic regression

When label y is still a single component

$$\frac{\partial \log P(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{f}\left(\boldsymbol{x}, y\right) - \mathbb{E}_{Y|X}[\boldsymbol{f}\left(\boldsymbol{x}, y\right)] \tag{12}$$

where

$$\mathbb{E}_{Y|X}[\boldsymbol{f}\left(\boldsymbol{x}, y\right)] = \sum_{y \in \mathcal{Y}} \left\{ P(y|\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x}, y) \right\} \tag{13}$$

Parameter Estimation: CRFs

When label $\boldsymbol{y}$ is a sequence

$$\frac{\partial \log P(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) - \mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})] \tag{14}$$

where

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{15}$$

and

$$\mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})] = \sum_{\boldsymbol{y} \in \mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \right\} \tag{16}$$

Expectation

$$
\begin{aligned}
\mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})] &= \sum_{\boldsymbol{y} \in \mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \right\} \\
&= \sum_{\boldsymbol{y} \in \mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\} \\
&= \sum_{\boldsymbol{y} \in \mathcal{Y}^T} \sum_{i=1}^{T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\} \\
&= \sum_{i=1}^{T} \sum_{\boldsymbol{y} \in \mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\} \\
&= \sum_{i=1}^{T} \sum_{y_{i-1} \in \mathcal{Y}; y_i \in \mathcal{Y}} \left\{ P(y_{i-1}, y_i|\boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\}
\end{aligned}
$$

Expectation

$$
\begin{aligned}
\mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})] &= \sum_{\boldsymbol{y}\in\mathcal{Y}^T}\left\{P(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\right\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^T}\left\{P(\boldsymbol{y}|\boldsymbol{x})\sum_{i=1}^{T}\boldsymbol{f}_i(\boldsymbol{x},y_{i-1},y_i)\right\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^T}\sum_{i=1}^{T}\left\{P(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x},y_{i-1},y_i)\right\} \\
&= \sum_{i=1}^{T}\sum_{\boldsymbol{y}\in\mathcal{Y}^T}\left\{P(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x},y_{i-1},y_i)\right\} \\
&= \sum_{i=1}^{T}\sum_{y_{i-1}\in\mathcal{Y};y_i\in\mathcal{Y}}\left\{P(y_{i-1},y_i|\boldsymbol{x})\boldsymbol{f}_i(\boldsymbol{x},y_{i-1},y_i)\right\}
\end{aligned}
$$

Basic Operation

$$P(y_{i-1}, y_i|\boldsymbol{x}) = \frac{\sum_{\boldsymbol{y}\backslash\{y_{i-1},y_i\}} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{y}))} \tag{17}$$

Basic operation

$$\sum_{\tilde{\boldsymbol{y}}} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{s}, \tilde{\boldsymbol{y}})) \tag{18}$$

where $\tilde{\boldsymbol{y}}$ could be

- $\tilde{\boldsymbol{y}} = \boldsymbol{y}\backslash\{y_{i-1}, y_i\}$
- $\tilde{\boldsymbol{y}} = \boldsymbol{y}$

$\psi_i(y_{i-1}, y_i)$

$$
\begin{aligned}
\psi(\boldsymbol{y}) &= \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})) \\
&= \exp(\sum_{i=1}^{T} \boldsymbol{\theta}^\top \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \\
&= \prod_{i=1}^{T} \psi_i(y_{i-1}, y_i)
\end{aligned}
$$

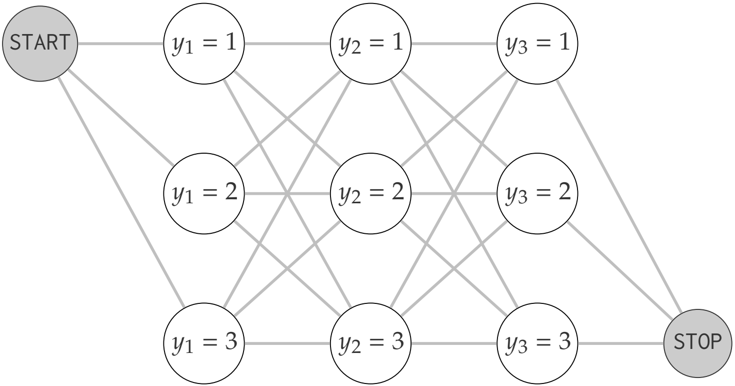
where

$$
\psi_i(y_{i-1}, y_i) = \exp(\boldsymbol{\theta}^\top \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i)) \tag{19}
$$

What We Need?

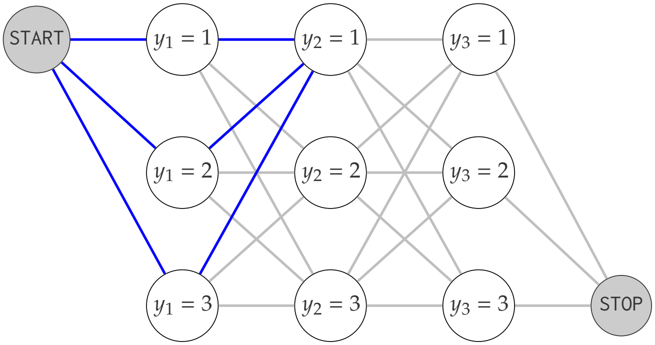
$$
\begin{aligned}
P(y_{i-1}, y_i | \boldsymbol{x}) &= \frac{\sum_{\boldsymbol{y} \setminus \{y_{i-1}, y_i\}} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}} \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{y}))} \\
&= \frac{\sum_{\boldsymbol{y} \setminus \{y_{i-1}, y_i\}} \prod_{i=1}^{T} \psi_i(y_{i-1}, y_i)}{\sum_{\boldsymbol{y}} \prod_{i=1}^{T} \psi_i(y_{i-1}, y_i)}
\end{aligned}
$$

Forward-Backward Algorithm: Basic idea



Forward-Backward Algorithm: Forward term

$$P(y_2 = 1, y_3 = 1) \propto \sum_{y_1} \psi_1(\textsc{start}, y_1)\psi_2(y_1, y_2 = 1) \cdot \psi_3(y_2 = 1, y_3 = 1) \quad (20)$$



Forward-Backward Algorithm: Forward term

Forward term

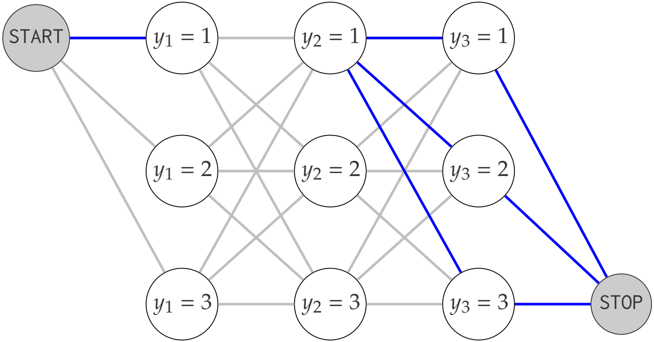
$$\alpha_i(y) = \sum_{y' \in \mathcal{Y}} \alpha_{i-1}(y')\psi_i(y', y) \tag{21}$$

Base case

$$\alpha_1(y) = \psi(\textsc{start}, y) \tag{22}$$

Forward-Backward Algorithm: Back term

$$P(y_1 = 1, y_2 = 1) \propto \sum_{y_3} \psi_1(\textsc{start}, y_1)\psi_2(y_1 = 1, y_2 = 1) \cdot \psi_3(y_2 = 1, y_3) \quad (23)$$



Forward-Backward Algorithm: Backward term

Backward term

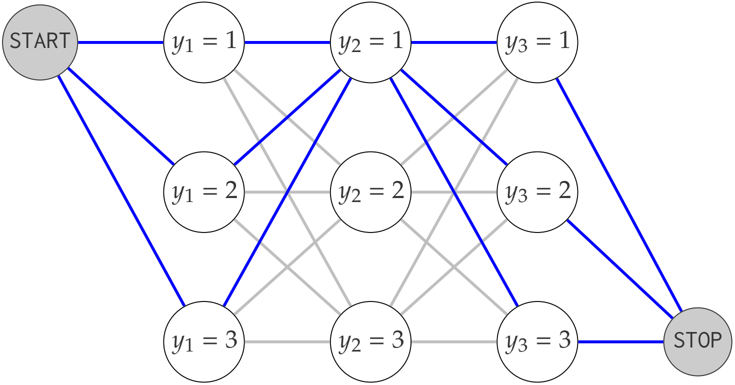
$$\beta_i(y) = \sum_{y' \in \mathcal{Y}} \psi_{i+1}(y, y')\beta_{i+1}(y') \tag{24}$$

Base case

$$\beta_T(y) = 1 \tag{25}$$

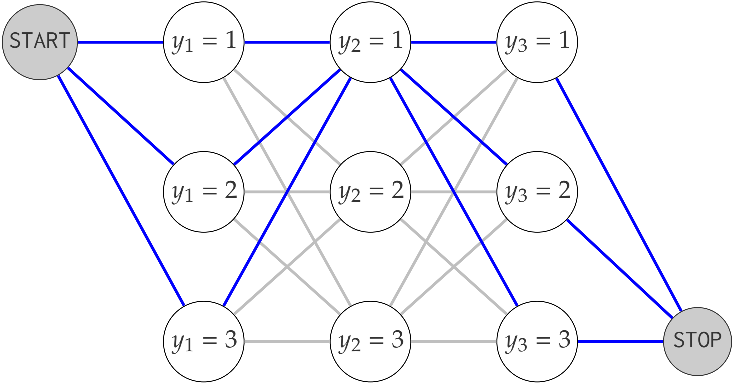
Test

What we compute here?



Test

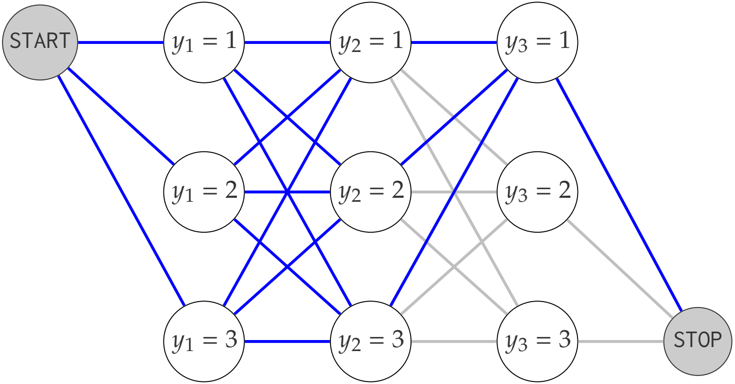
What we compute here?



Answer: $P(y_2 = 1)$

Normalization Term Z

$$Z = \sum_{\boldsymbol{y} \in \mathcal{Y}^T} \psi(\boldsymbol{y}) = \sum_{y \in \mathcal{Y}} \alpha_T(y)\beta_T(y) = \sum_{y \in \mathcal{Y}} \alpha_T(y) \qquad (26)$$



Expectation

$$
\begin{aligned}
\mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})] &= \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \right\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \sum_{i=1}^{T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\} \\
&= \sum_{i=1}^{T} \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\} \\
&= \sum_{i=1}^{T} \sum_{y_{i-1}\in\mathcal{Y};\, y_i\in\mathcal{Y}} \left\{ P(y_{i-1}, y_i|\boldsymbol{x}) \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \right\}
\end{aligned}
$$

Parameter Estimation: CRFs

$$\frac{\partial \log P(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) - \mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})] \tag{27}$$

where

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{T} \boldsymbol{f}_i(\boldsymbol{x}, y_{i-1}, y_i) \tag{28}$$

and

$$\mathbb{E}_{Y|X}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})] = \sum_{\boldsymbol{y}\in\mathcal{Y}^T} \left\{ P(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \right\} \tag{29}$$

Summary

Reference

Bishop, C. M. (2006).
Pattern Recognition and Machine Learning.
Springer-Verlag.

Eisenstein, J. (2018).
Natural Language Processing.
MIT Press.

Lafferty, J., McCallum, A., and Pereira, F. (2001).
Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
In *ICML*.