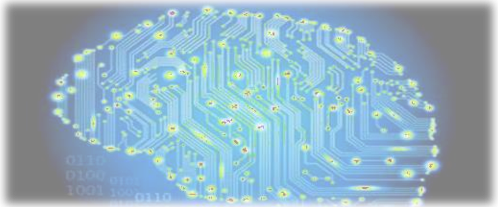
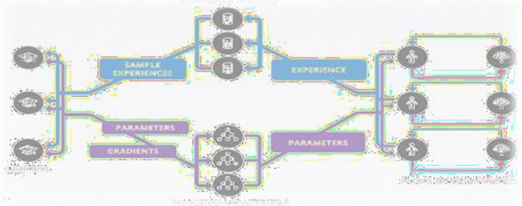
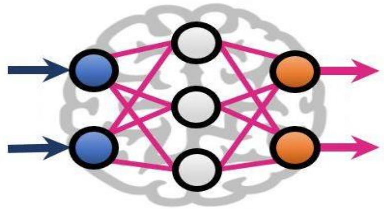
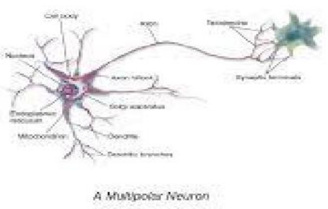
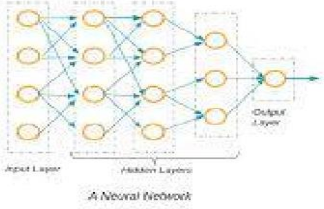










Hidden Markov Model, Maximum Entropy Markov Model and Conditional Random Field

HMM, MEMM, CRF

Geonil Yun,

rjsdlf45@gmail.com

2021. 08. 17.

Introduction

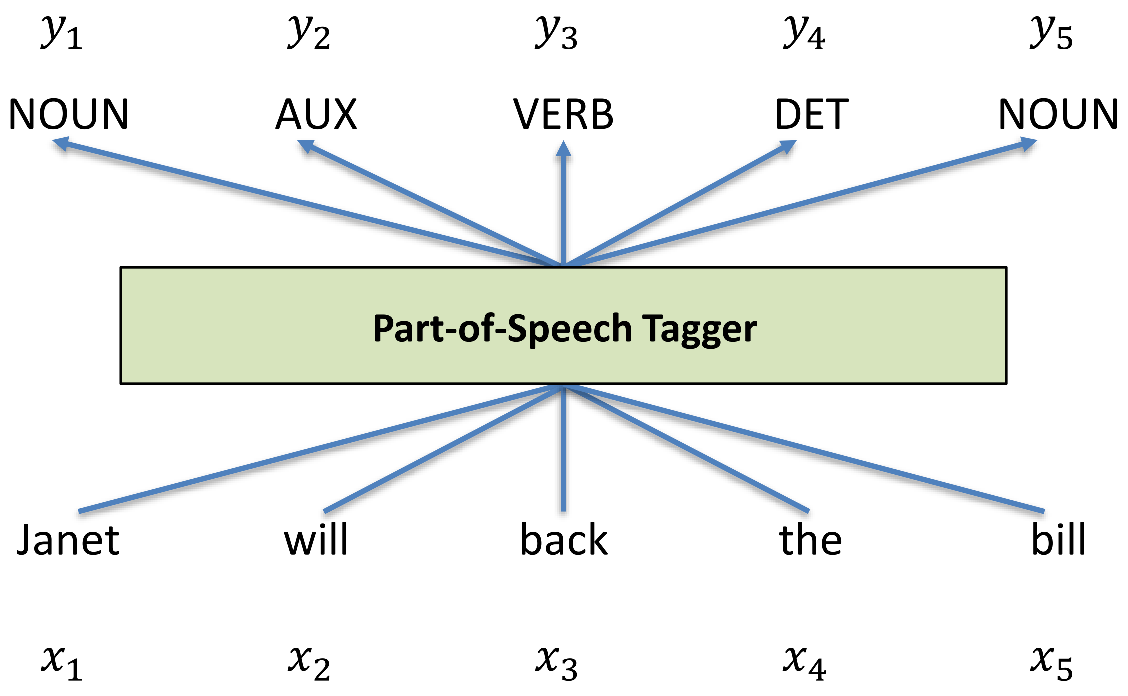
- **Tagging**

  - **Part-of-Speech (POS) tagging**

    - **VERB** : Book that flight
    - **NOUN** : Read a book.

words POS

sequence labeling : from $\{x_1 \ \dots \ x_t\}$ to $\{y_1 \ \dots \ y_t\}$



Introduction

- **Tagging**

  - **Named Entity Recognition (NER) tagging**

    - **PER (Person)** : Janet
    - **LOC (Location)** : New York
    - **ORG (Organization)** : Samsung
    - **TIME (Time)** : Friday

Jane Villanueva of United Airlines Holding discussed the Chicago route.

NER Tagging

Words	**Tag**
Jane	PER
Villanueva	PER
of	
United	ORG
Airlines	ORG
Holding	ORG
discussed	
the	
Chicago	LOC
route	
.	

- **each word does not have one tag**

  2-word : “Jane Villanueva”
  3-word : “United Airlines Holding”
  1-word : “Chicago”

BIOES

- **BIO tagging for NER**

  - **Tagged by one word**

    - **B** : token that begins a span
    - **I** : tokens inside a span
    - **O** : tokens outside of any span

  - **each word gets one tag**

    [Jane *B-PER*, Villanueva *I-PER*]
    [United *B-ORG*, Airlines *I-ORG*, Holding *I-ORG*]
    [Chicago *B-LOC*]

Jane Villanueva of United Airlines Holding discussed the Chicago route.

NER with BIO tagging

Words	Tag	BIO Tag
Jane	PER	B-PER
Villanueva	PER	I-PER
of		O
United	ORG	B-ORG
Airlines	ORG	I-ORG
Holding	ORG	I-ORG
discussed		O
the		O
Chicago	LOC	B-LOC
route		O
.		O

BIOES

- **BIOES tagging (BIO variants)**

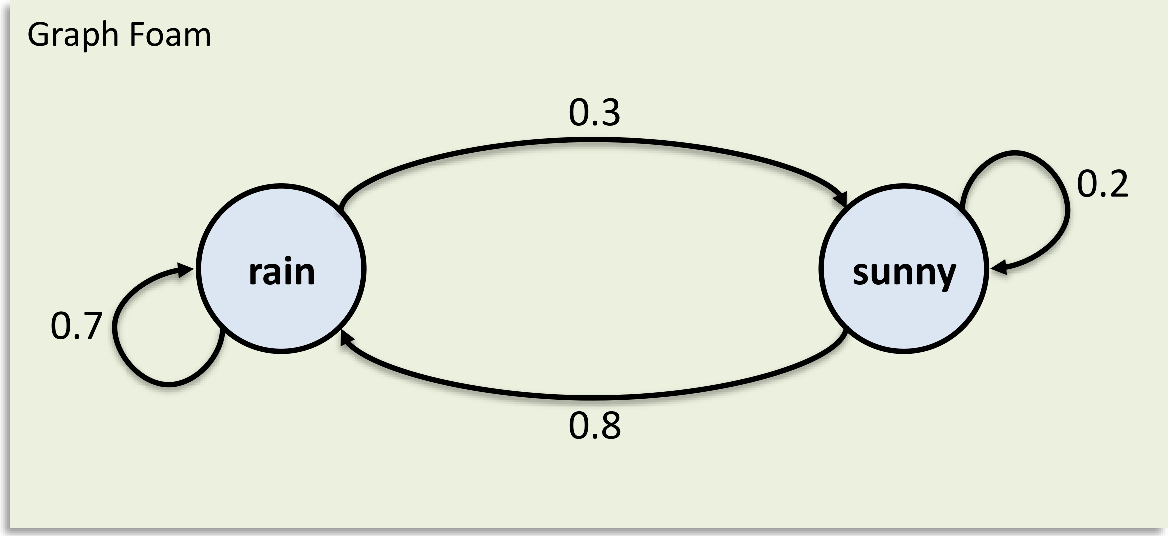
  - **Tagged by one word**

    - **B** : token that begins a span
    - **I** : tokens inside a span
    - **O** : tokens outside of any span
    - **E** : token that ends a span
    - **S** : token that gets single span

Words	Tag	BIO Tag	BIOES Tag
Jane	PER	B-PER	B-PER
Villanueva	PER	I-PER	E-PER
of		O	O
United	ORG	B-ORG	B-ORG
Airlines	ORG	I-ORG	I-ORG
Holding	ORG	I-ORG	E-ORG
discussed		O	O
the		O	O
Chicago	LOC	B-LOC	S-LOC
route		O	O
.		O	O

Conditional Random Field

- **Hidden Markov Model (HMM)**
  - **Based on Markov chain that embodies Markov Assumption**
    - Markov Assumption : $\boldsymbol{P(q_t|q_{t-1}, q_{t-2}, \dots, q_1) = P(q_t|q_{t-1})}$



state transition probability matrix(A)

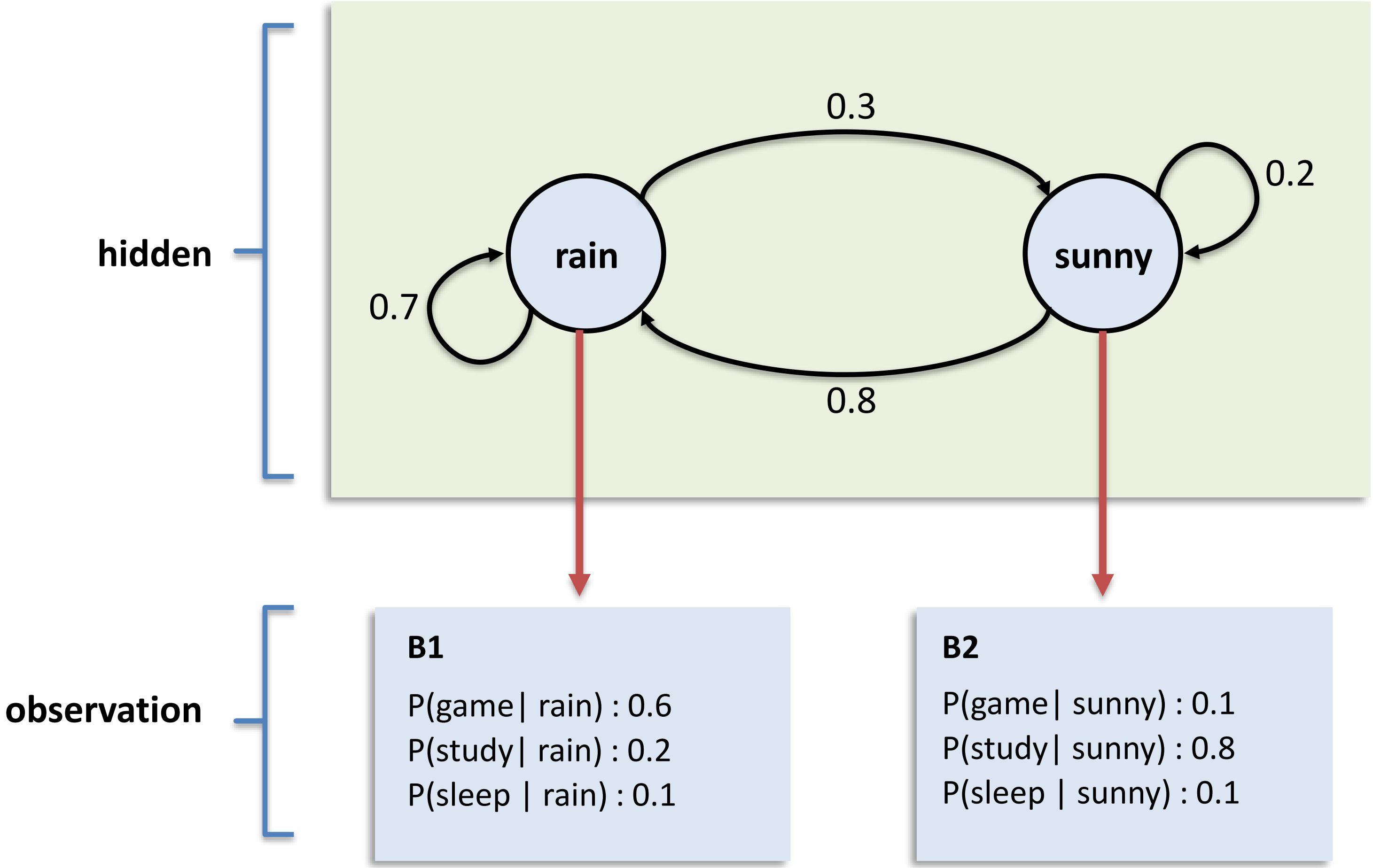
$\boldsymbol{t}$

$\boldsymbol{t-1}$

	rain	sunny
rain	0.7	0.3
sunny	0.8	0.2

Conditional Random Field

- **Hidden Markov Model (HMM)**
  - Predict hidden state(tag) by using observation state(word)



state transition probability matrix (A)

t

$t-1$	rain	sunny
rain	0.7	0.3
sunny	0.8	0.2

emission probability matrix (B)

	game	study	sleep
rain	0.6	0.2	0.1
sunny	0.1	0.8	0.1

Conditional Random Field

- **Hidden Markov Model (HMM)**
  - Parameter of HMM, $\lambda = \{A, B, \pi\}$

A : state transition probability

- $a_{ij} = P\left(q_t = s_j \mid q_{t-1} = s_i\right)$, s: state, q:sequence
- $\sum_j a_{ij} = 1$

B : emission probability

- $b_j(v_k) = P(o_t = v_k \mid q_t = s_i)$, o: observation, v:
- $\sum_j b_j(v_k) = 1$

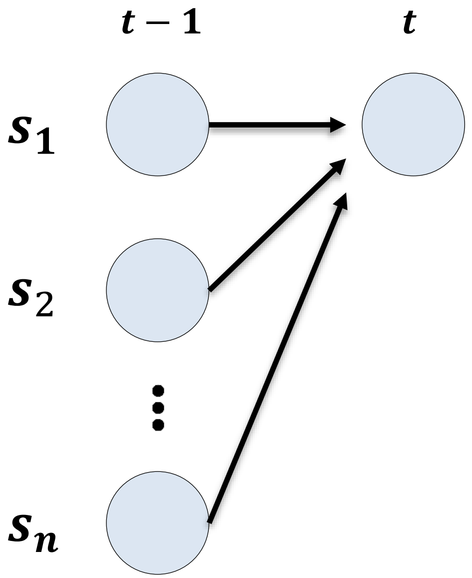
π : initial state probability

- Probability of start with state $s_i = \pi_i$
- $\sum_i \pi_i = 1$

Conditional Random Field

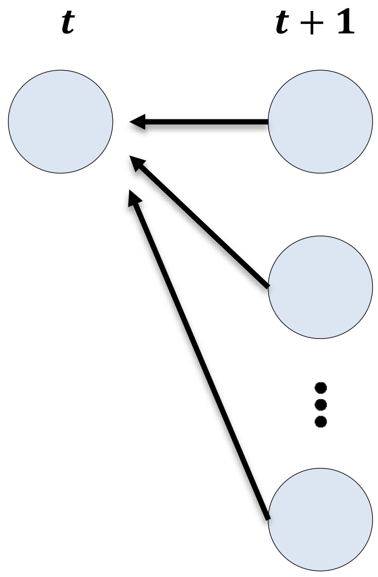
- **Hidden Markov Model (HMM)**
  - evaluation of HMM (forward algorithm, backward algorithm)
  - given O($\vec{o}$) and HMM(λ), find probability O($\vec{o}$)

Forward Algorithm



- $\alpha_1(i) = \pi_i b_i(o_1)$
- $\alpha_t(i) = \left[\sum_j \alpha_{t-1}(j) a_{ji}\right] b_i(o_t)$

Backward Algorithm



- $\beta_{t=T}(i) = 1$
- $\beta_t(i) = \left[\sum_j \beta_{t+1}(j) a_{ji} b_j(o_{t+1})\right]$

Conditional Random Field

- **Hidden Markov Model (HMM)**
  - Decoding with **Viterbi algorithm**
  - Given model **HMM(λ), observation($\vec{o}$)** , find sequence of **hidden state($\vec{s}$)**

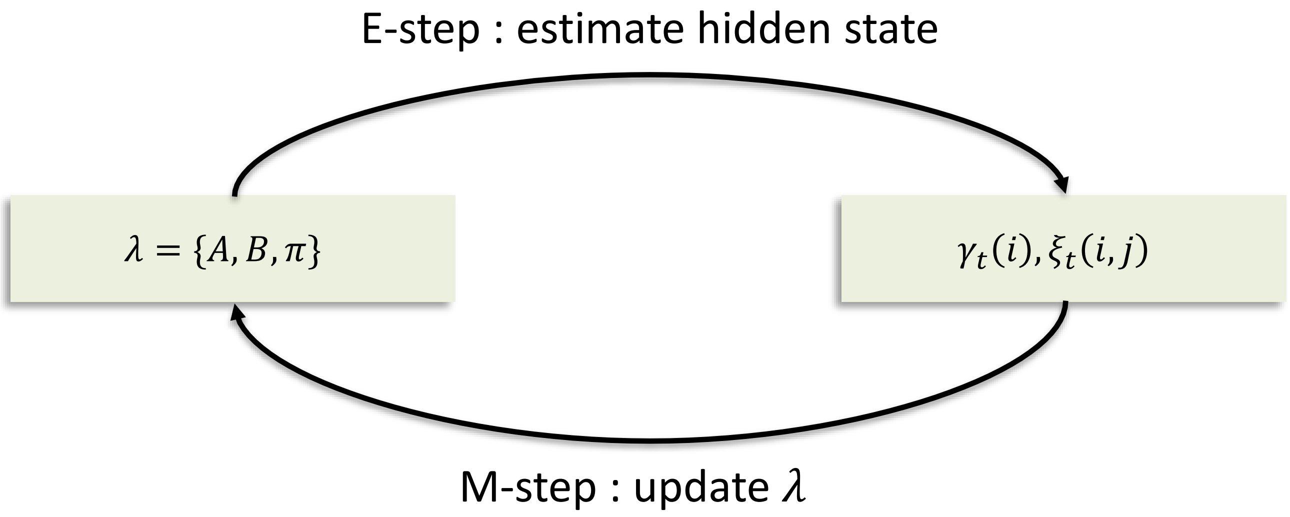
Viterbi probability $v_t(i)$: probability of s_i at t

- $v_1(i) = \pi_i b_i(o_1)$
- $v_t(i) = \max_{j}\left(\mathrm{v}_{\mathrm{t}-1}(j) \cdot a_{ji}\right) \cdot b_i(o_t)$
- $\tau_t(i) = argmax_j\left(v_{t-1}(j) \cdot a_{ji}\right)$
- $\hat{q}_T = argmax_j\left(v_T(j)\right)$
- $\hat{q}_t = \tau_{t+1}(\hat{q}_{t+1})$
- $\hat{Q}_t = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_t)$

Conditional Random Field

- **Hidden Markov Model (HMM)**
  - Learning with **E-M step** (Baum-Welch Algorithm)
  - Given model **observation($\vec{o}$)** , find HMM(λ^*)
  - maximum likelihood estimation of parameter HMM(λ^*)
  - use γ and ξ with forward probability(α), backward probability(β)

$$\boldsymbol{HMM(\lambda^*) = argmax_{\lambda} P(O|\lambda)}$$



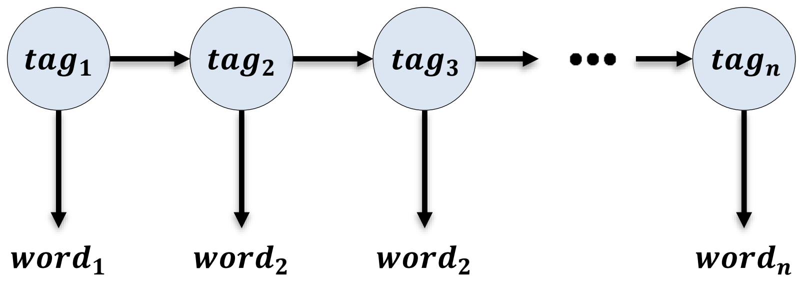
$\gamma_t(i)$: $prob. of\ s_i\ at\ t$
$\xi_t(i,j)$: $prob. of\ s_i\ at\ t\ and\ s_j\ at\ t+1$

Conditional Random Field

- **Maximum Entropy Markov Model (MEMM)**

  - HMM model

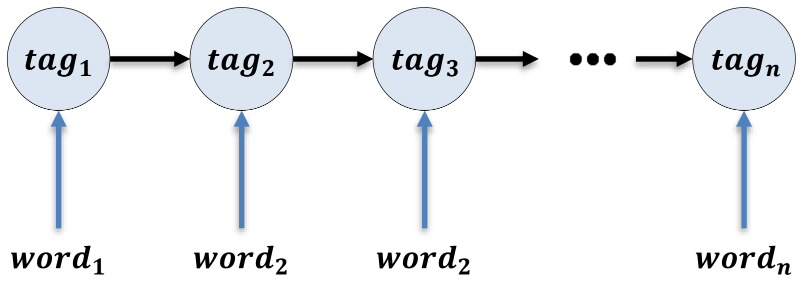
    - $P(tag \mid tag)$, state transition probability
    - $P(word \mid tag)$, emission probability



$$HMM = \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1})\, P(o_t \mid s_t)$$

  - MEMM model

    - $P(tag \mid tag, word)$
    - $P(s_t \mid s_{t-1},\ o_t)$



$$MEMM = \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}, o_t)$$

Conditional Random Field

- **Maximum Entropy Markov Model (MEMM)**
  - HMM's transition and observation functions are replaced by $P(s \mid s', o)$
    - $\alpha_t(i) = \sum_{s'} \alpha_{t-1}(s') \boxed{P(s|s')P(o_t|s)}$ → $\alpha_t(i) = \sum_{s'} \alpha_{t-1}(s') \boxed{P(s|o_t)}$

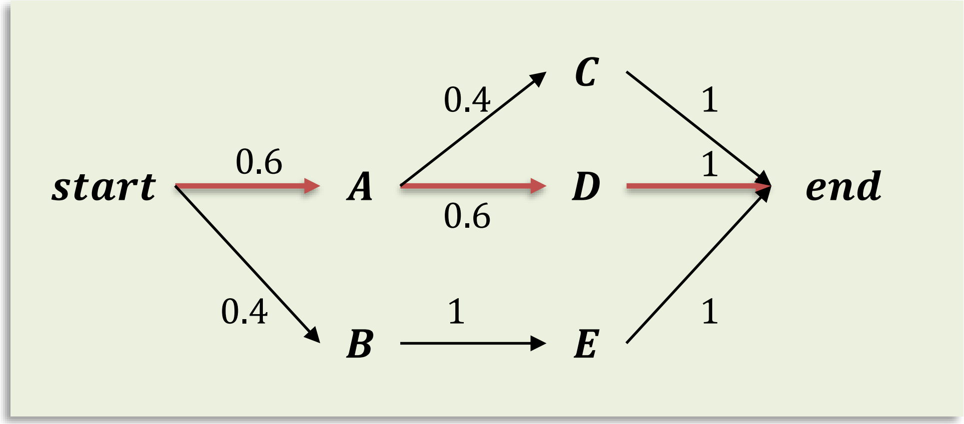
$$P(s \mid s', o) = \frac{1}{Z(o,\ s')} \exp\left(\sum_{i=1}^{n} w_i \cdot f_i(o, s')\right)$$

normalizing factor sum to one

$$f(o_t, s_t) = \begin{cases} 1, & if\ b(o_t)\ is\ true\ and\ s = s_t \\ 0, & otherwise. \end{cases}$$

Conditional Random Field

- **Maximum Entropy Markov Model (MEMM)**
  - Label bias problem
    - state A is frequent, transition C or D
    - answer is < A, D >, but predict < B, E>



$$P(A, D \mid o) = 0.6 \times 0.6 \times 1.0 = 0.36$$
$$P(B, E \mid o) = 0.4 \times 1.0 \times 1.0 = 0.4$$

$$P(A, D \mid o) < P(B, E \mid o)$$

Conditional Random Field

- **Conditional Random Field (CRF)**
  - Local normalization to global normalization

MEMM's

$$\boxed{P(s \mid s', o)} = \frac{1}{\boxed{Z}(o,\ s')} \exp\left(\sum_{i=1}^{n} w_i \cdot f_i(o, s')\right)$$

locality of normalization

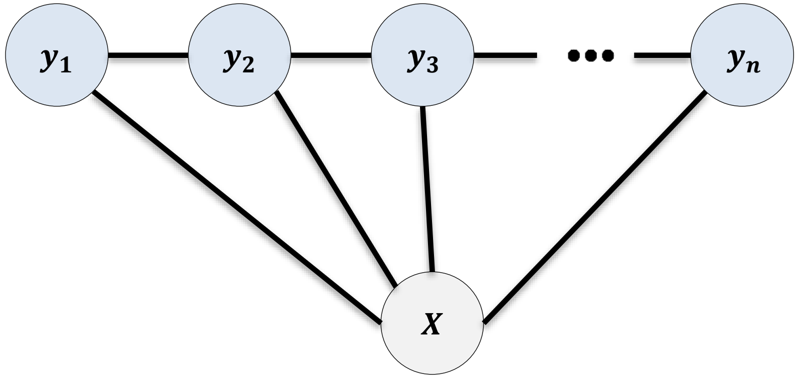
weights

$$P(\bar{y} \mid \bar{x}, w) = \frac{\exp\left(\sum_i \sum_j w_j \cdot f_j(y_{i-1}, y_i, \bar{x}, i)\right)}{\sum_{y'} \exp\left(\sum_i \sum_j w_j \cdot f_j\left(y'_{i-1}, y'_i, \bar{x}, i\right)\right)}$$

all possible label sequence

length of sequence X

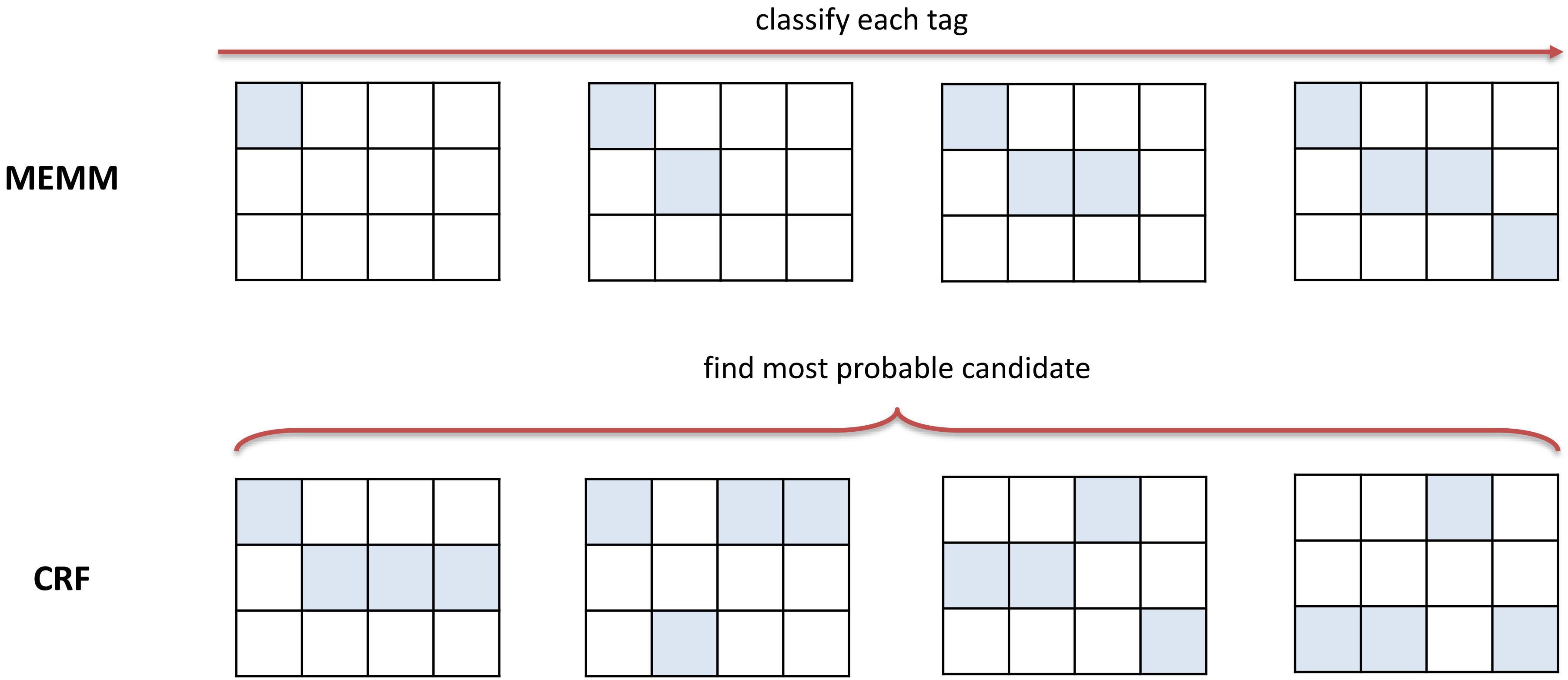
feature function



$X = x_1, x_2, \ldots, x_n$

Conditional Random Field

- **Conditional Random Field (CRF)**
  - Local normalization to global normalization

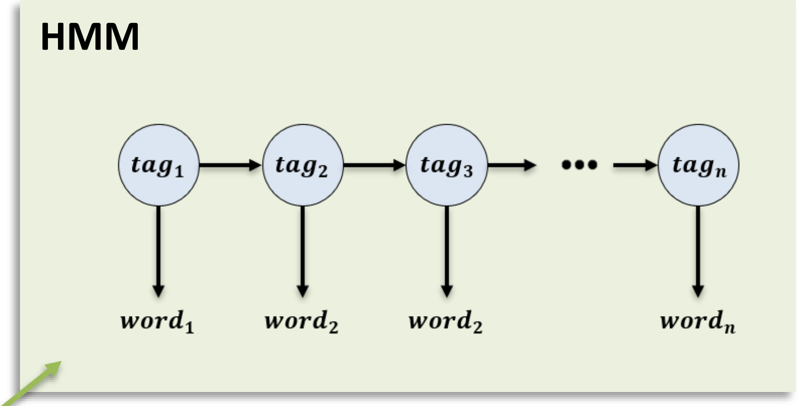


Conditional Random Field

- **Conditional Random Field (CRF)**
  - Feature function : produce a real value

$$P(\bar{y} \mid \bar{x}, w) = \frac{\exp\left(\sum_i \sum_j w_j \cdot \boxed{f_j(y_{i-1}, y_i, \bar{x}, i)}\right)}{\sum_{y'} \exp\left(\sum_i \sum_j w_j \cdot f_j(y'_{i-1}, y'_i, \bar{x}, i)\right)}$$

    - y_{i-1} : previous state
    - y_i : current state
    - $\bar{x}$: entire sentence $\{x_1, x_2, \ldots, x_n\}$
    - i : position of word



looks like state transition probability in HMM

$$f_1(y_{i-1}, y_i, \bar{x}, i) \begin{cases} 1 & if\ y_{i-1} = PERSON\ and\ y_i = OTHER \\ 0 & otherwise \end{cases}$$

✓ depends on its corresponding weight w_1

$w_1 > 0$: f_1 active (increase probability of $\bar{y}$)
$w_1 < 0$: f_1 deactive (decrease probability of $\bar{y}$)

Conditional Random Field

- **Conditional Random Field (CRF)**

  - Feature function : produce a **real value**

"John said so"

$$f_1(y_{i-1}, y_i, \bar{x}, i) \begin{cases} 1 \;\; if \; y_i = PERSON \; and \; x_i = John \\ 0 \;\; otherwise \end{cases}$$

$$f_2(y_{i-1}, y_i, \bar{x}, i) \begin{cases} 1 \;\; if \; y_i = PERSON \; and \; \boxed{x_{i+1} = said} \\ 0 \;\; otherwise \end{cases}$$

overlapping features

HMM cannot use next word

$$\boldsymbol{y_1 = PERSON \; to \; w_1 + w_2}$$

Conditional Random Field

- **Conditional Random Field (CRF)**

  - Training : parameter estimation

    use **maximum likelihood estimation**(gradient ascent)

    $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$, $x^{(1)} = x_{1:N}^{(1)}$: use dataset

    learning rate

    $$\bar{w}^* = \bar{w} + \boxed{\alpha} \times \frac{\partial}{\partial w_k} \boxed{L(\bar{w})}$$

    regularization

    $$L(\bar{w}) = \sum_{i=1}^{n} \log p(\bar{y}|\bar{x}, \bar{w}) - \boxed{\frac{\lambda}{2}||\bar{w}||^2}$$

  - Decoding

    use **Viterbi Algorithm**

    $$\underset{\bar{y}\in Y}{\operatorname{argmax}}\big(p(\bar{y}|\bar{x}, \bar{w})\big) = \underset{\bar{y}\in Y}{\operatorname{argmax}} \sum_{j=1}^{m} \bar{w} \cdot \bar{f}(\bar{x}, j, y_j, y_{j-1})$$

    $$\delta[i, y] = \max_{y'\in Y}\big[\delta[j-1, y'] \times \bar{w} \cdot \bar{f}(\bar{x}, j, y', y)\big]$$

Conditional Random Field

- **Conditional Random Field (CRF)**
  - Decoding

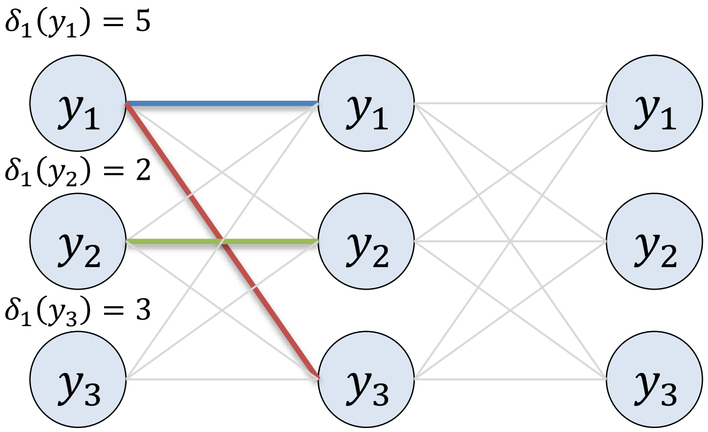
$$\psi_t(y_t) = \exp\left(\sum_i w_i f_i(y_t, y_{t-1}, x_t)\right)$$

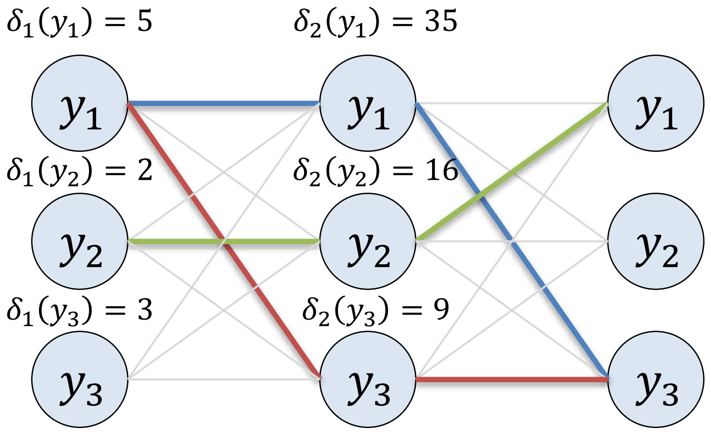
ψ_2	y_1	y_2	y_3
y_1	7	2	1
y_2	1	8	1
y_3	3	2	1

$\delta_2(y_1) = \delta_1(y_1) \times \psi_2(y_1)$ = 5x7
$\delta_2(y_2) = \delta_1(y_2) \times \psi_2(y_2)$ = 2x8
$\delta_2(y_3) = \delta_1(y_3) \times \psi_2(y_3)$ = 3x3

ψ_3	y_1	y_2	y_3
y_1	1	3	4
y_2	3	1	2
y_3	1	1	2

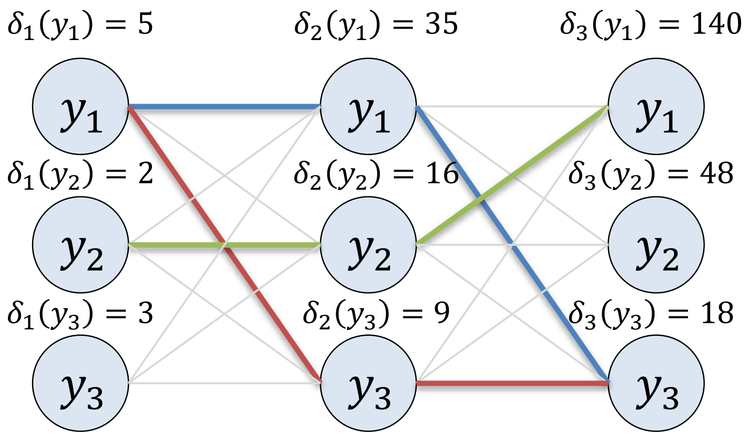
$\delta_3(y_1) = \delta_2(y_1) \times \psi_3(y_1)$ = 35x4
$\delta_3(y_2) = \delta_2(y_2) \times \psi_3(y_2)$ = 16x3
$\delta_3(y_3) = \delta_2(y_3) \times \psi_3(y_3)$ = 9x2

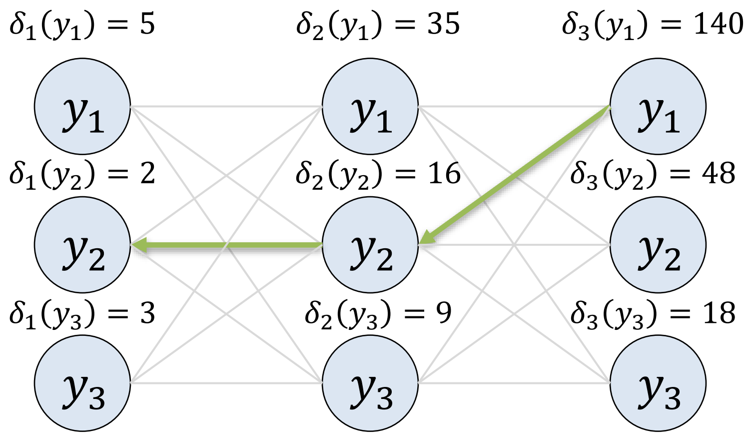




Conditional Random Field

- **Conditional Random Field (CRF)**
  - Decoding

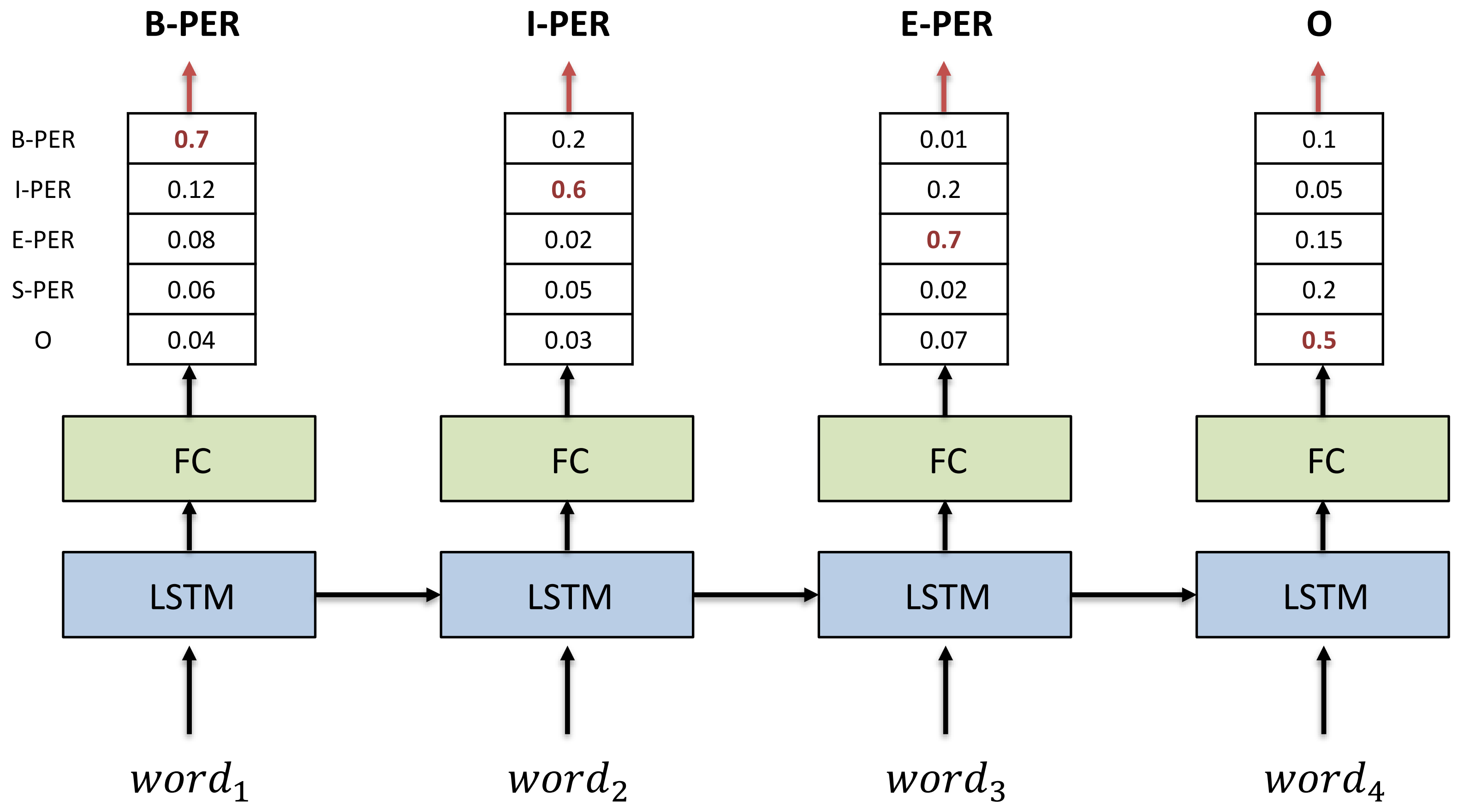




$$\bar{y} = \{y_2, y_2, y_1\}$$

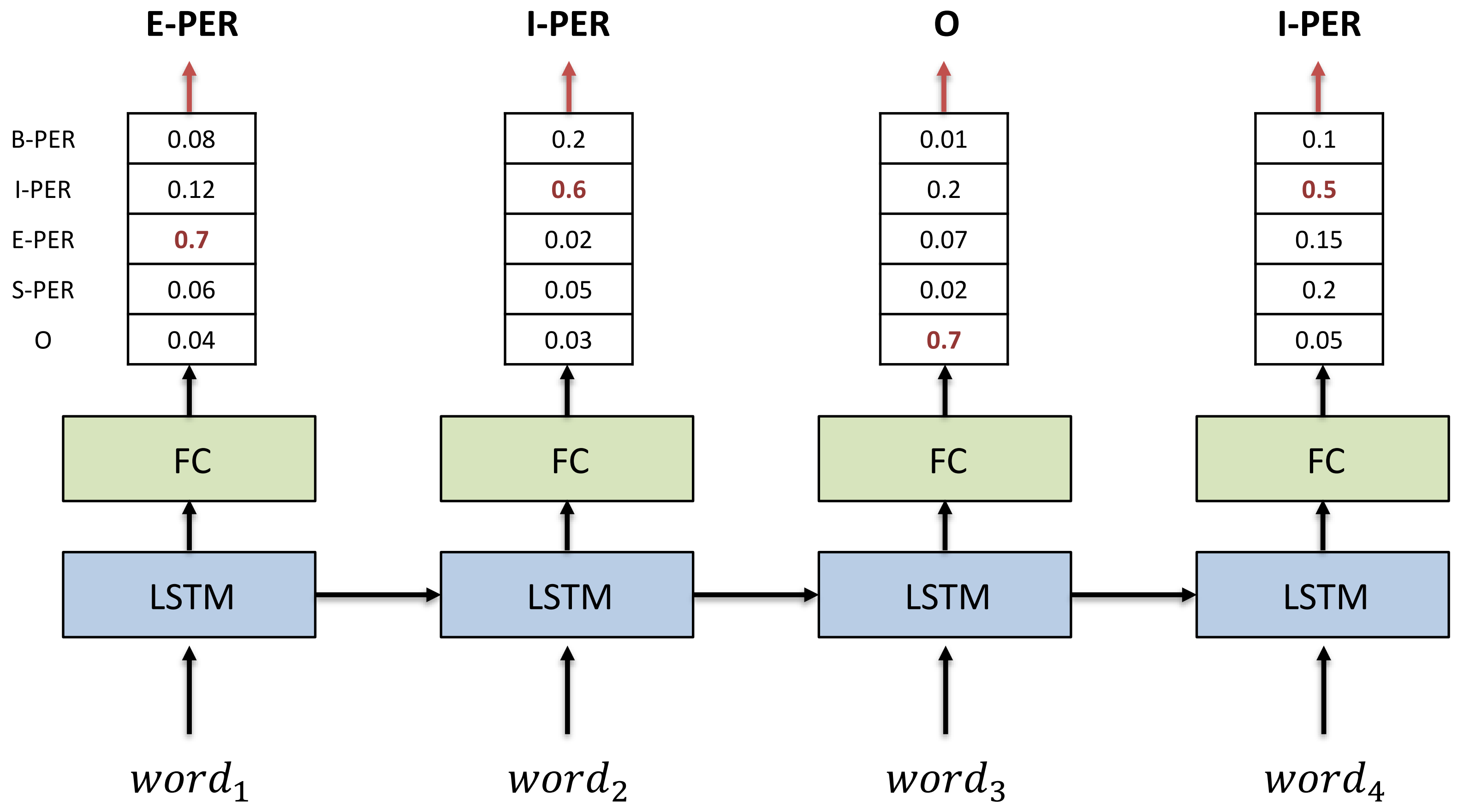
Conditional Random Field

- **Tagging task (BIOES) with LSTM**
  - **word 1, 2, 3 represent PERSON**
  - **word 4 is outside span**



Conditional Random Field

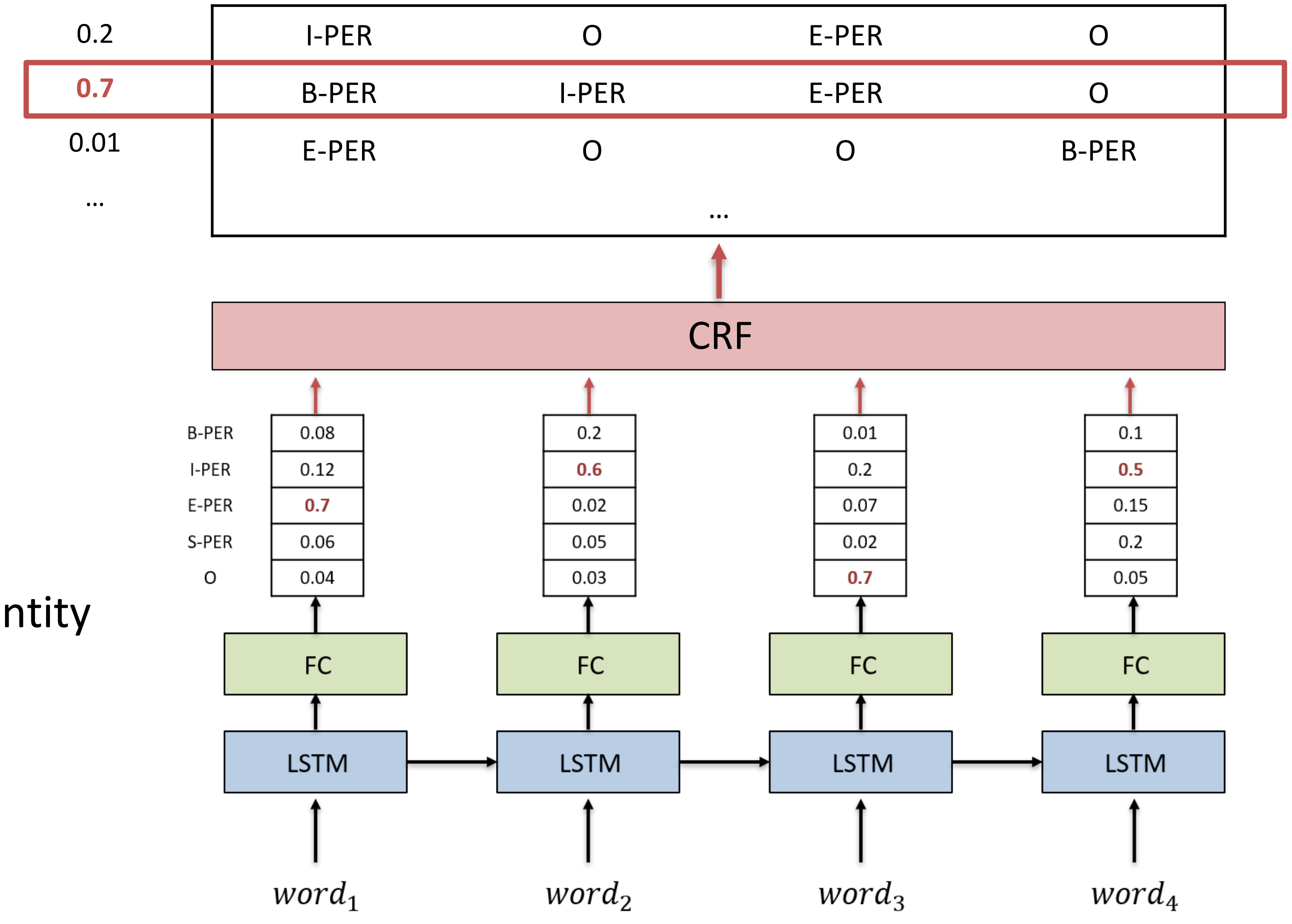
- **Tagging task (BIOES) with LSTM (Problem)**
  - **Cannot start with E**
  - **Cannot start with I**



Conditional Random Field

- **Tagging task (BIOES) with LSTM + CRF**
  - **CRF learn BIOES features**
  - **CRF learn other features of tagging**

1. Not start with I, E
2. E comes with end of 2 words entity
3. O-E, O-I patterns not exist



Thank you