

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

HMM, MEMM, CRF

Geonil Yun,

rjsdlf45@gmail.com

2021. 08. 10.



Introduction

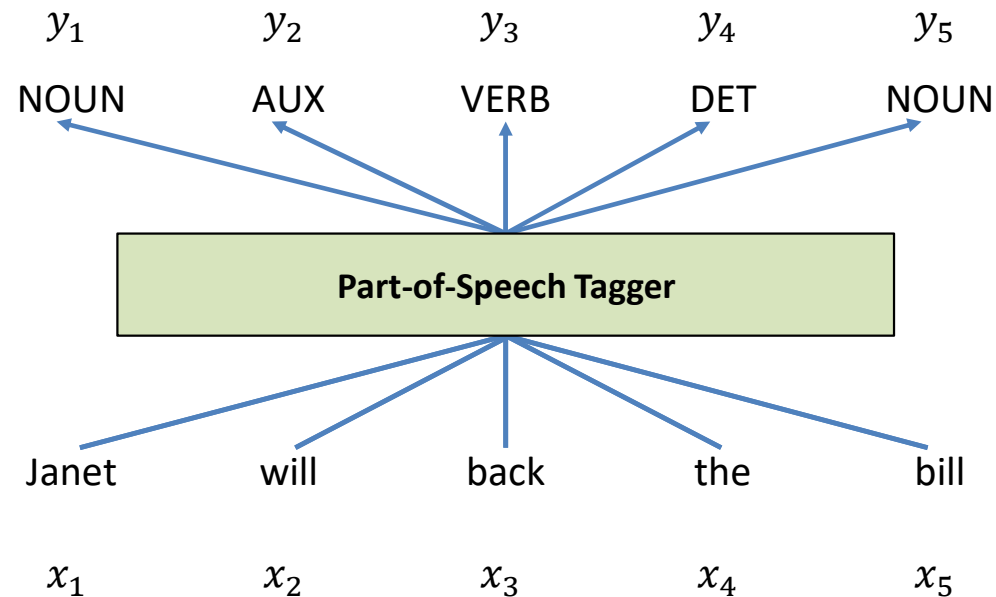
- Tagging

- Part-of-Speech (POS) tagging

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

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

words $\nearrow$ POS $\nearrow$



Introduction

- Tagging

- Named Entity Recognition (NER) tagging

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

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

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

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	
.	



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

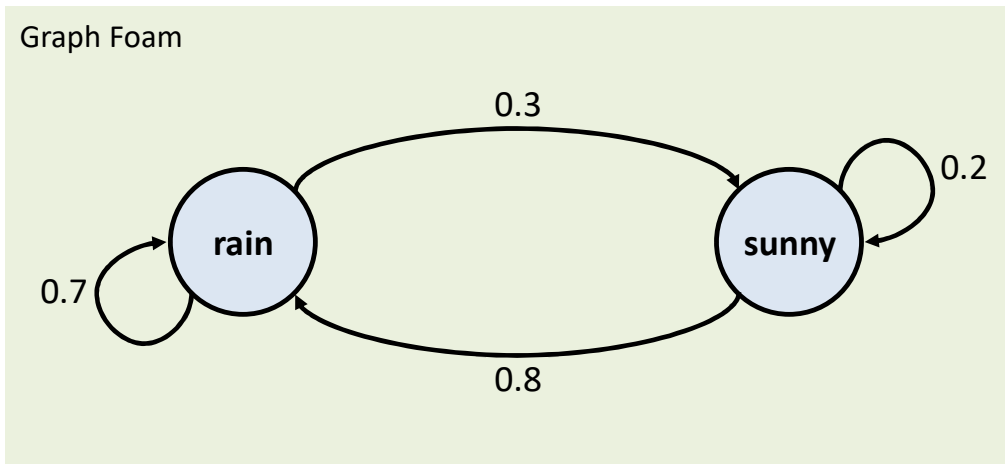
Words	Tag	BIO Tag	BIOES Tag
Jane	PER	B-PER	B-PER
Villanueva	PER	I-PER	E-PER
of		O	O
United	ORG	B-ORG	B-ORG
Airlines	ORG	I-ORG	I-ORG
Holding	ORG	I-ORG	E-ORG
discussed		O	O
the		O	O
Chicago	LOC	B-LOC	S-LOC
route		O	O
.		O	O

Conditional Random Field

- Hidden Markov Model (HMM)

- Based on Markov chain that embodies Markov Assumption

- Markov Assumption : $P(q_t | q_{t-1}, q_{t-2}, \dots, q_1) = P(q_t | q_{t-1})$



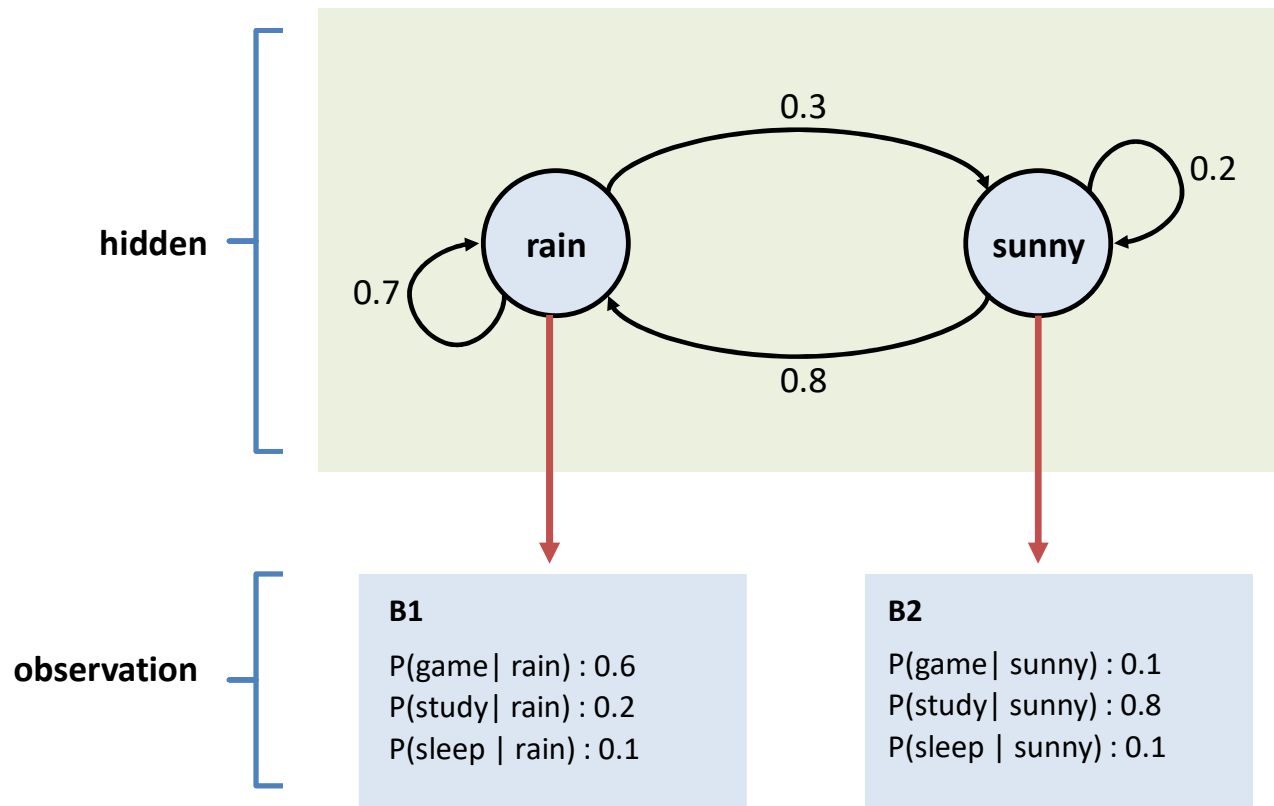
state transition probability matrix(A)

		t	
$t - 1$		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)

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

emission probability matrix (B)

$\backslash$	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

- $a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i)$, 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:
- $\sum_j b_j(v_k) = 1$

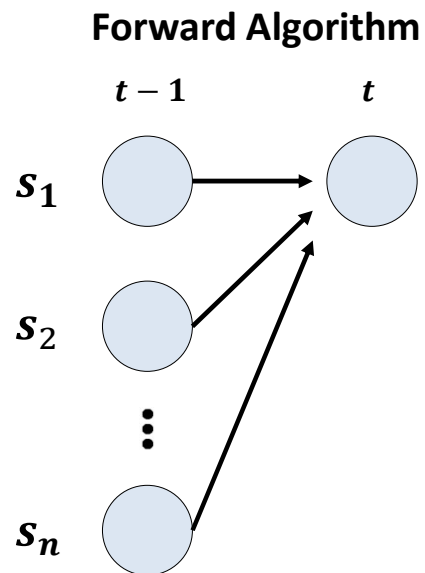
π : initial state probability

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

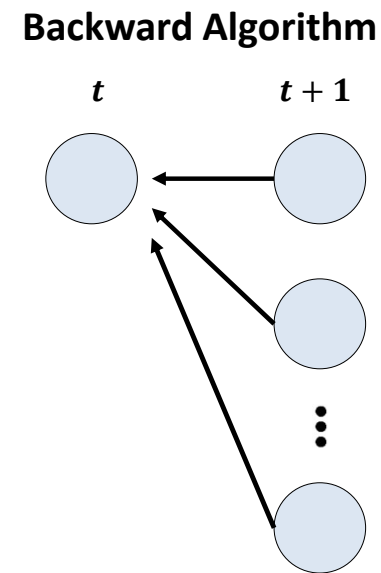
Conditional Random Field

- **Hidden Markov Model (HMM)**

- evaluation of HMM (forward algorithm, backward algorithm)
- given $O(\vec{o})$ and $HMM(\lambda)$, find probability $O(\vec{o})$



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



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

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

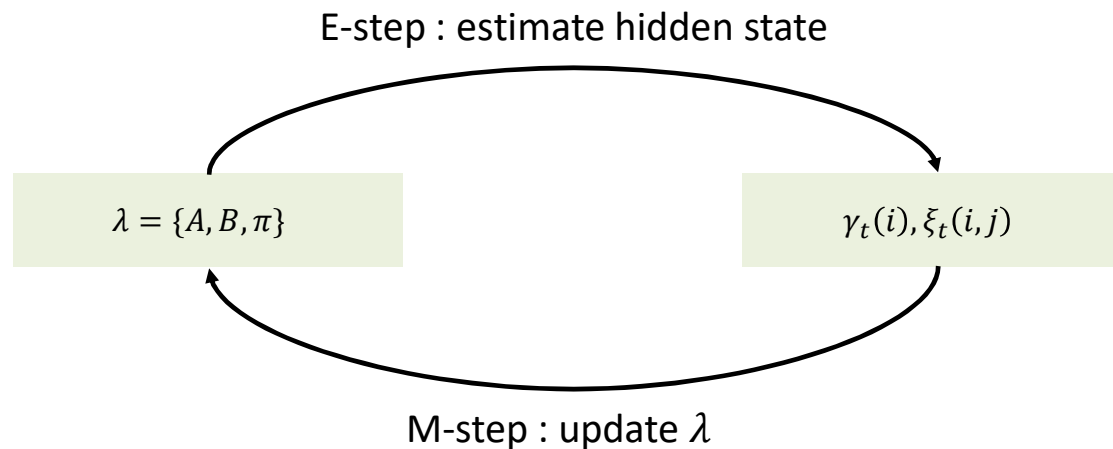
- $v_1(i) = \pi_i b_i(o_1)$
- $v_t(i) = \max_j (v_{t-1}(j) \cdot a_{ji}) \cdot b_i(o_t)$
- $\tau_t(i) = \operatorname{argmax}_j (v_{t-1}(j) \cdot a_{ji})$
- $\hat{q}_T = \operatorname{argmax}_j (v_T(j))$
- $\hat{q}_t = \tau_{t+1}(\hat{q}_{t+1})$
- $\hat{Q}_t = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_t)$

Conditional Random Field

- Hidden Markov Model (HMM)

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

$$HMM(\lambda^*) = \operatorname{argmax}_{\lambda} P(O|\lambda)$$



$\gamma_t(i)$: 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

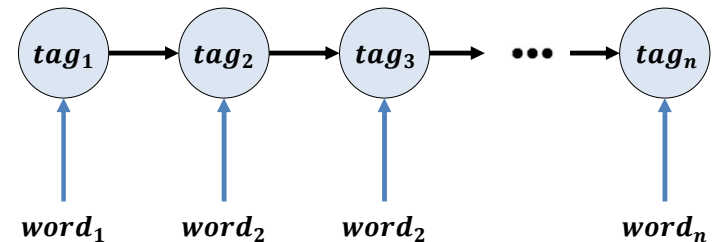
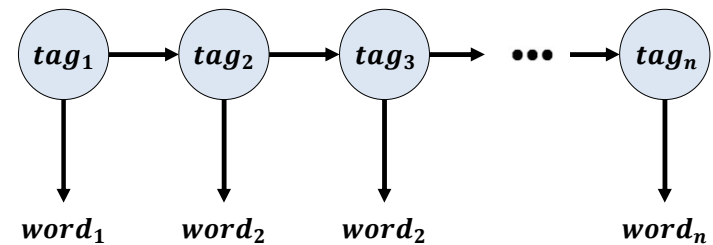
- $P(tag | tag)$, state transition probability
- $P(word | tag)$, emission probability

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

- MEMM model

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

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



Conditional Random Field

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

- HMM's transition and observation functions are replaced by $P(s | s', o)$

$$\alpha_t(i) = \sum_{s'} \alpha_{t-1}(s') \boxed{P(s|s')P(o_t|s)} \quad \longrightarrow \quad \alpha_t(i) = \sum_{s'} \alpha_{t-1}(s') \boxed{P(s|o_t)}$$

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

normalizing factor sum to one

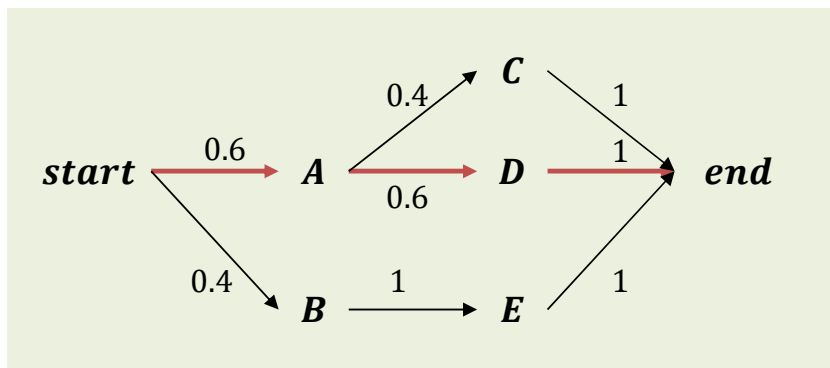
$$f(o_t, s_t) = \begin{cases} 1, & \text{if } b(o_t) \text{ is true and } s = s_t \\ 0, & \text{otherwise.} \end{cases}$$

Conditional Random Field

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

- Label bias problem

- state A is frequent, transition C or D
- answer is $\langle A, D \rangle$, but predict $\langle B, E \rangle$



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

$$P(B, E \mid o) = 0.4 \times 1.0 \times 1.0 = 0.4$$

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

Conditional Random Field

- Conditional Random Field (CRF)

- Local normalization to global normalization

MEMM's

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

locality of normalization

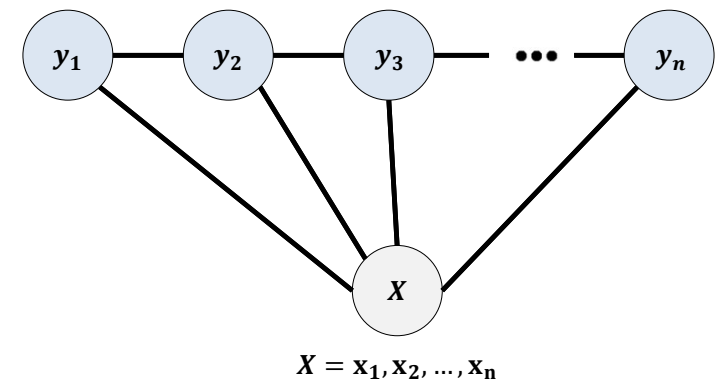
$$P(\bar{y} | \bar{x}, w) = \frac{\exp(\sum_i \sum_j w_j \cdot f_j(y_{i-1}, y_i, \bar{x}, i))}{\sum_{y'} \exp(\sum_i \sum_j w_j \cdot f_j(y'_{i-1}, y'_i, \bar{x}, i))}$$

weights

feature function

all possible label sequence

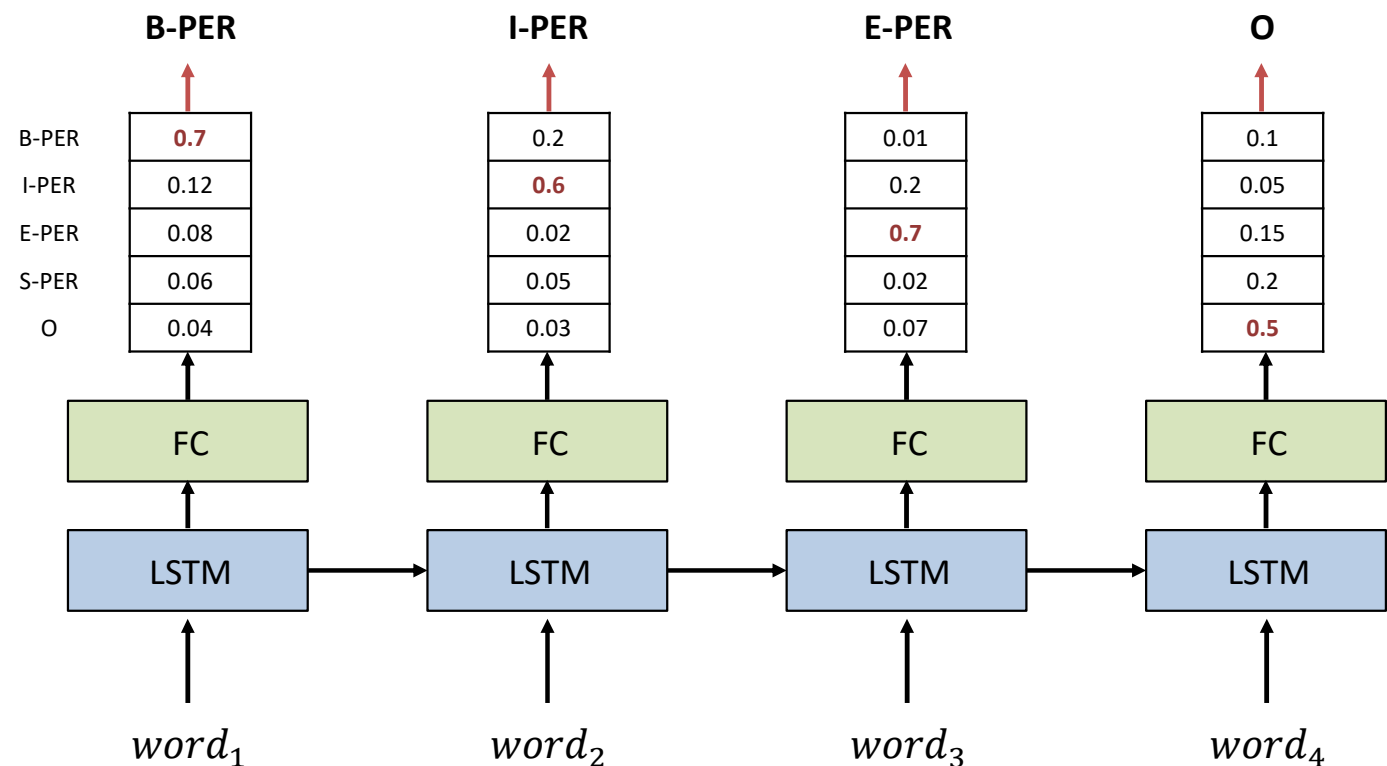
length of sequence X



Conditional Random Field

- Tagging task (BIOES) with LSTM

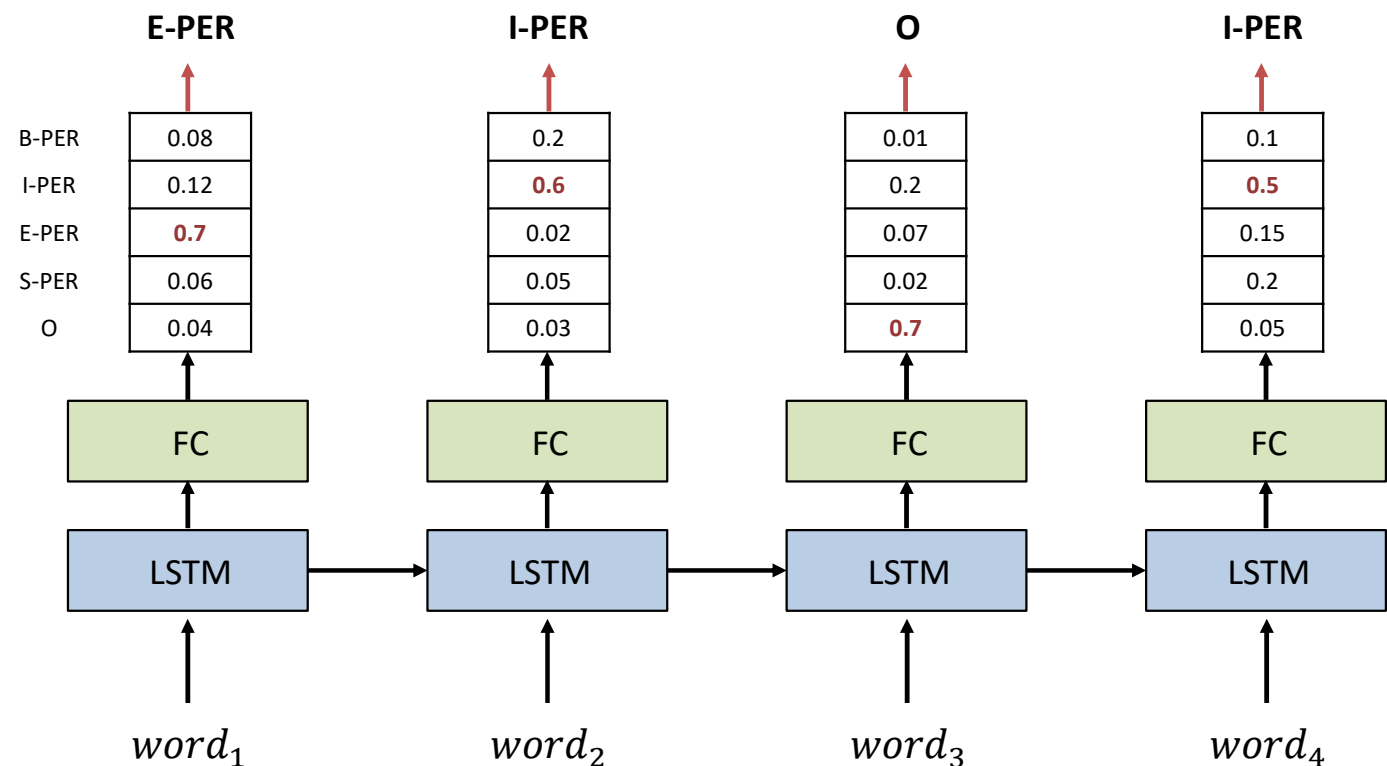
- word 1, 2, 3 represent PERSON
- word 4 is outside span



Conditional Random Field

- Tagging task (BIOES) with LSTM (Problem)

- Cannot start with E
- Cannot start with I

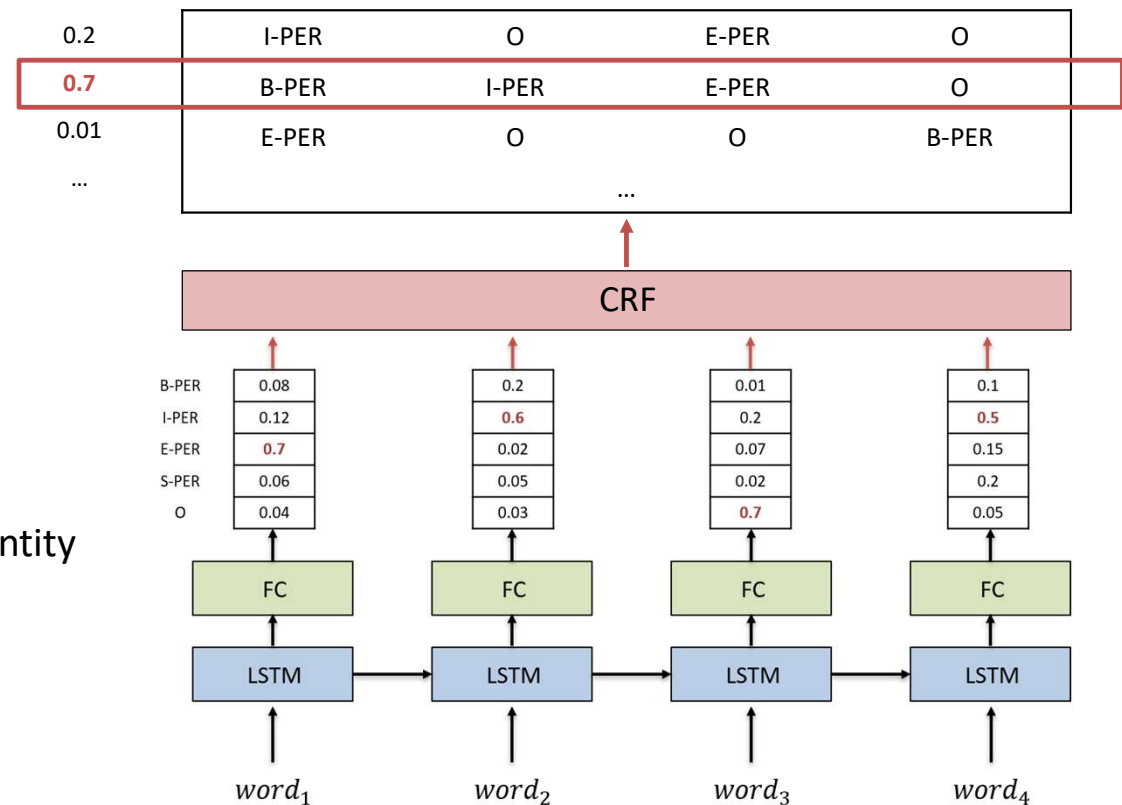


Conditional Random Field

- 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