Diving into SyntaxNet

Kim Ho Yeob

NLP Procedure

Tokenizing	형태소 분석
POS-Tagging	품사 태깅
Syntactic Parsing	문장 구조 분석
Semantic Parsing	문장 의미 분석

SyntaxNet ?

Tokenizing

POS-Tagging

Syntactic Parsing

Semantic Parsing

SyntaxNet

SyntaxNet Main Paper

Globally Normalized Transition-Based Neural Networks

Globally Normalized

Tools for solving problems of local normalization

Transition-Based

Method of dependency parsing

Neural Networks

SyntaxNet Main Paper

Globally Normalized Transition-Based Neural Networks

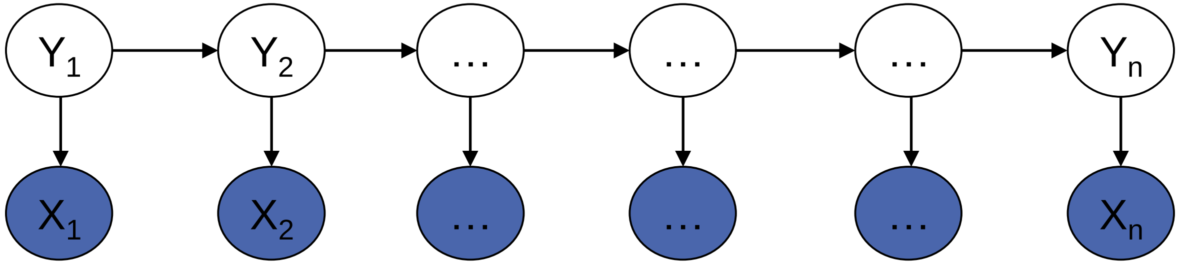
1. **Globally Normalized**
   Tools for solving problems of local normalization

2. **Transition-Based**
   Method of dependency parsing

Neural Networks

Globally Normalized?

- Globally Normalization starts from **Hidden Markov Model(HMM)**

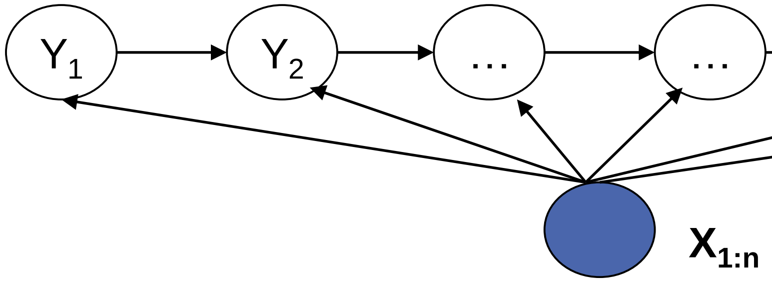


- **Shortcomings of Hidden Markov Model**
- HMM models direct dependence between each state and only its corresponding observation
  - NLP example: In a sentence segmentation task, segmentation may depend not just on a single word, but also on the features of the whole line such as line length, indentation, amount of white space, etc.

Referred From Label bias problem – Stanford Computer Science

Globally Normalized?

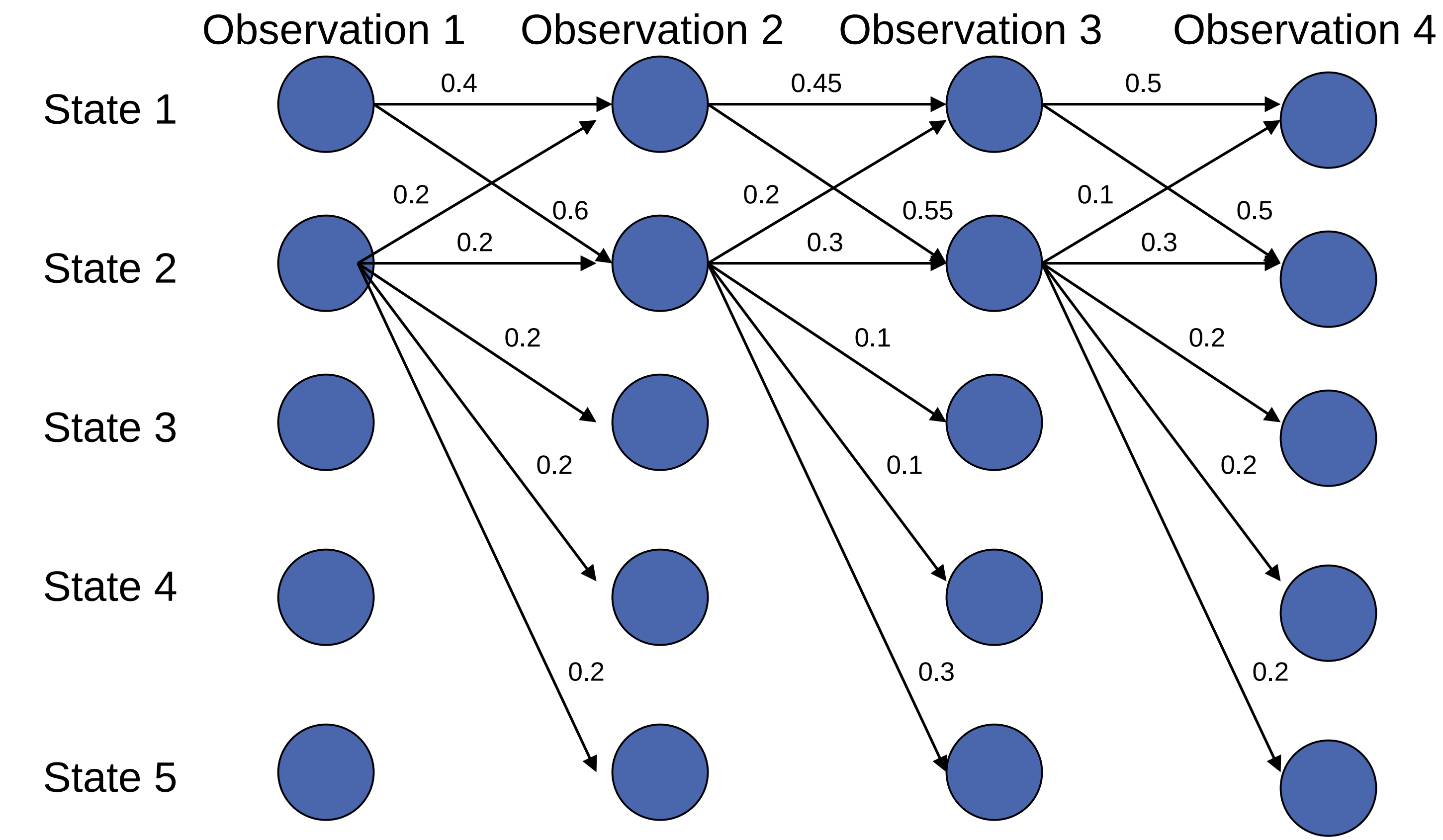
- **Maximum Entropy Markov Model(MEMM)**
  - Solution for shortcomings of HMM



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1}, \mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}_{1:n}))}{Z(y_{i-1}, \mathbf{x}_{1:n})}$$

- Models dependence between each state and the full observation sequence explicitly
  - More expressive than HMMs
- Discriminative model
  - Completely ignores modeling P(**X**): saves modeling effort
  - Learning objective function consistent with predictive function: P(**Y**|**X**)

MEMM: Label bias problem

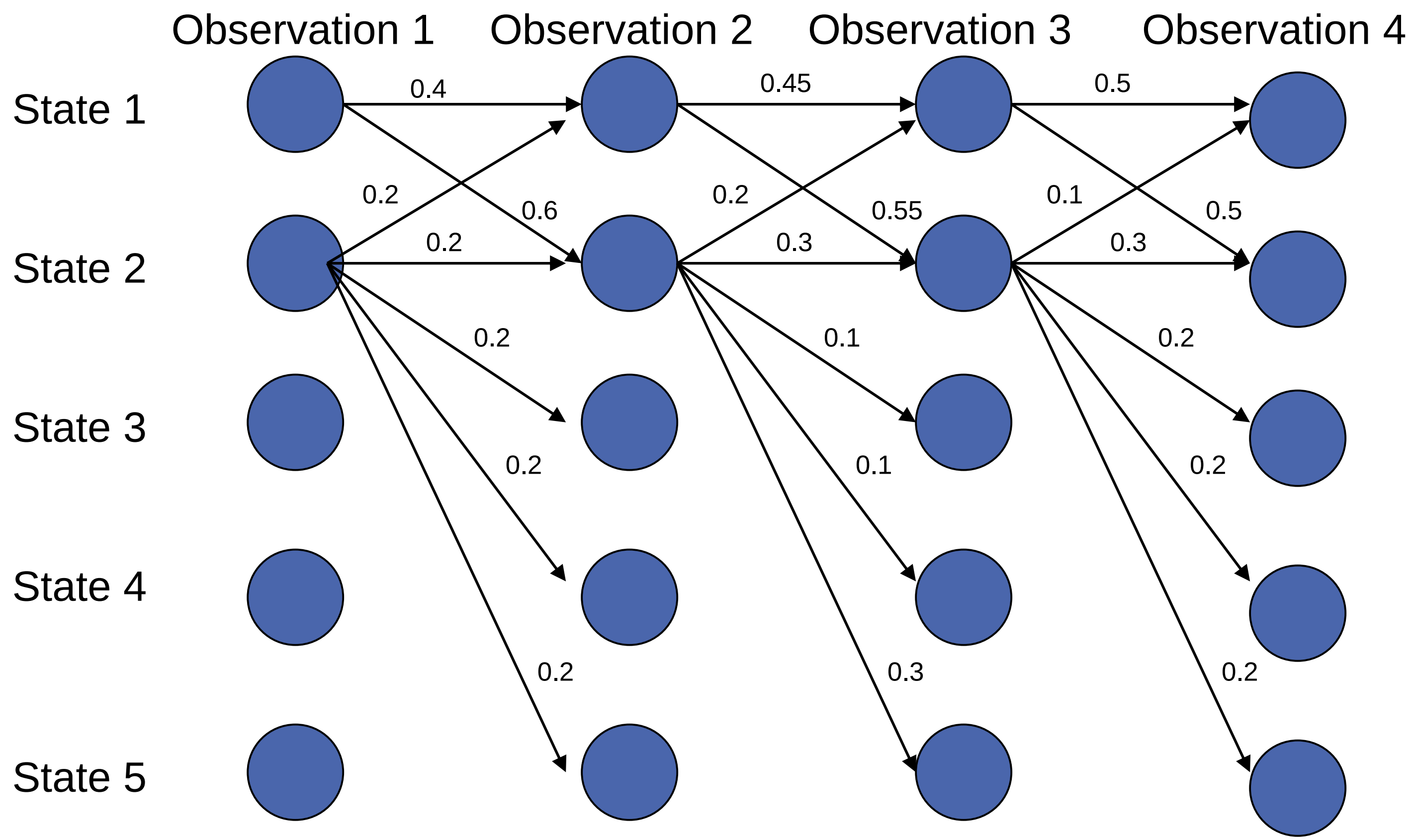


What the local transition probabilities say:

- State 1 almost always prefers to go to state 2
- State 2 almost always prefer to stay in state 2

[Referred From Label bias problem – Stanford Computer Science]()

MEMM: Label bias problem

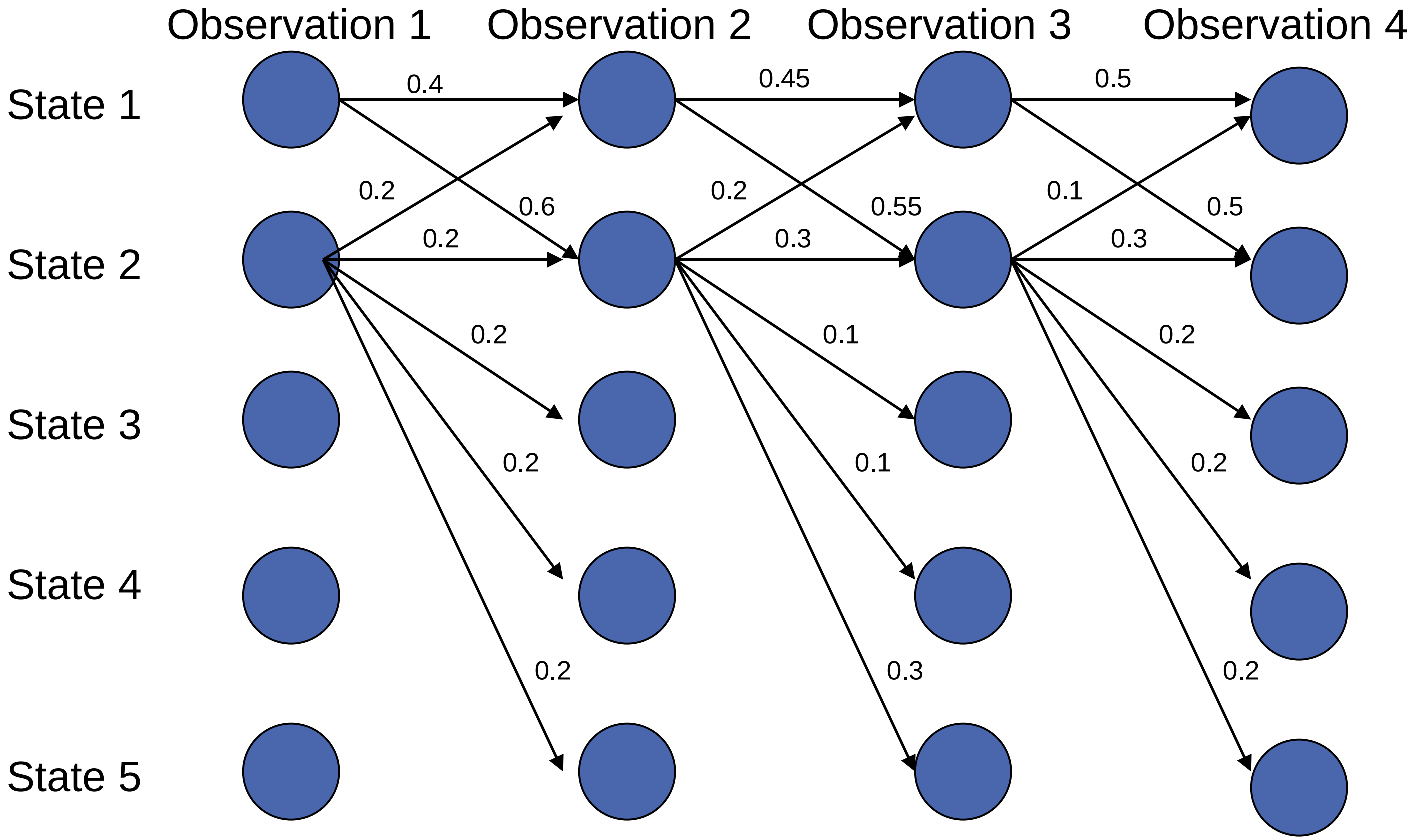


Probability of path 1-> 1-> 1-> 1:

- 0.4 x 0.45 x 0.5 = 0.09

[Referred From Label bias problem – Stanford Computer Science]

MEMM: Label bias problem



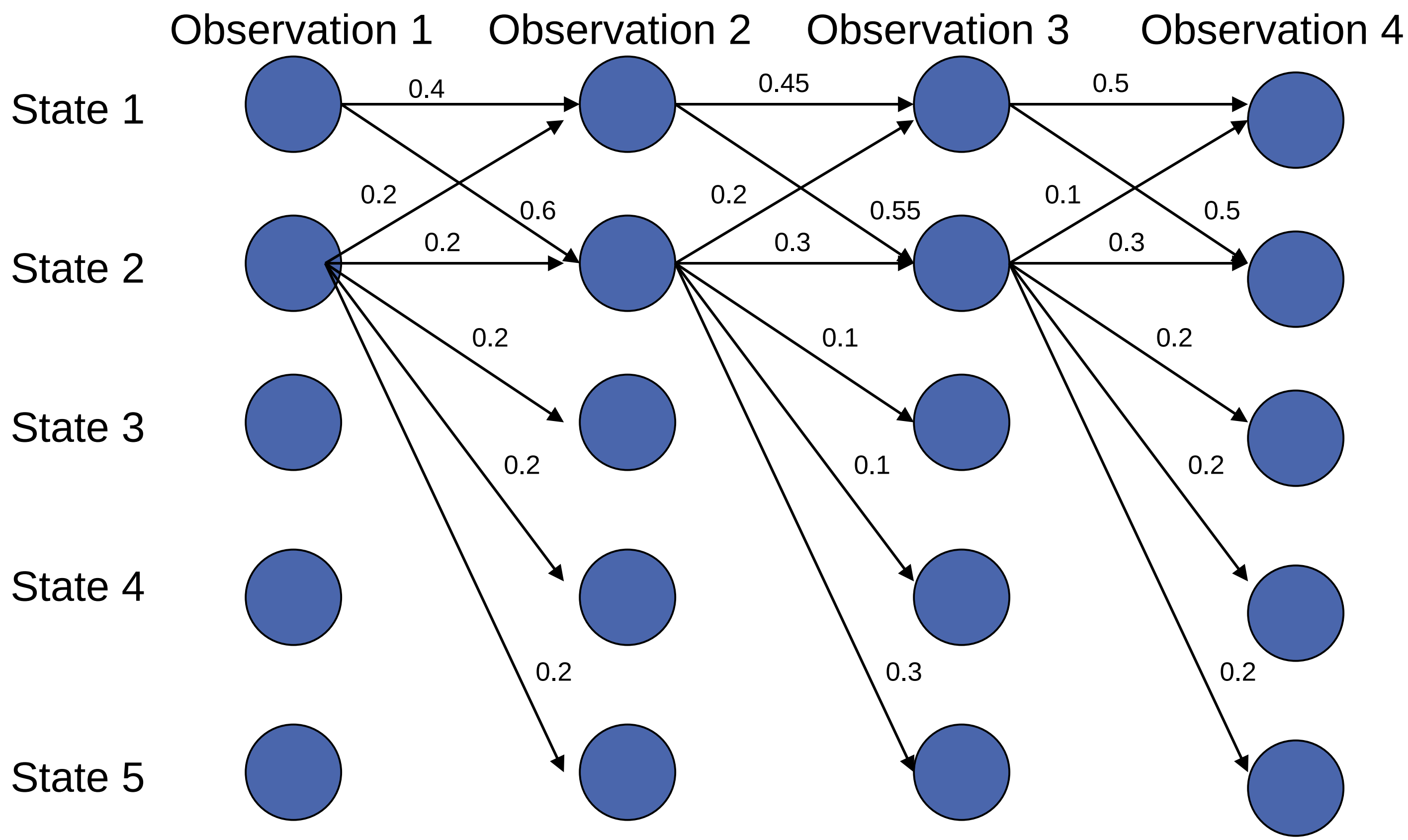
Probability of path 2->2->2->2 :

- 0.2 X 0.3 X 0.3 = 0.018

Other paths:
1-> 1-> 1-> 1: 0.09

[Referred From Label bias problem – Stanford Computer Science]

MEMM: Label bias problem



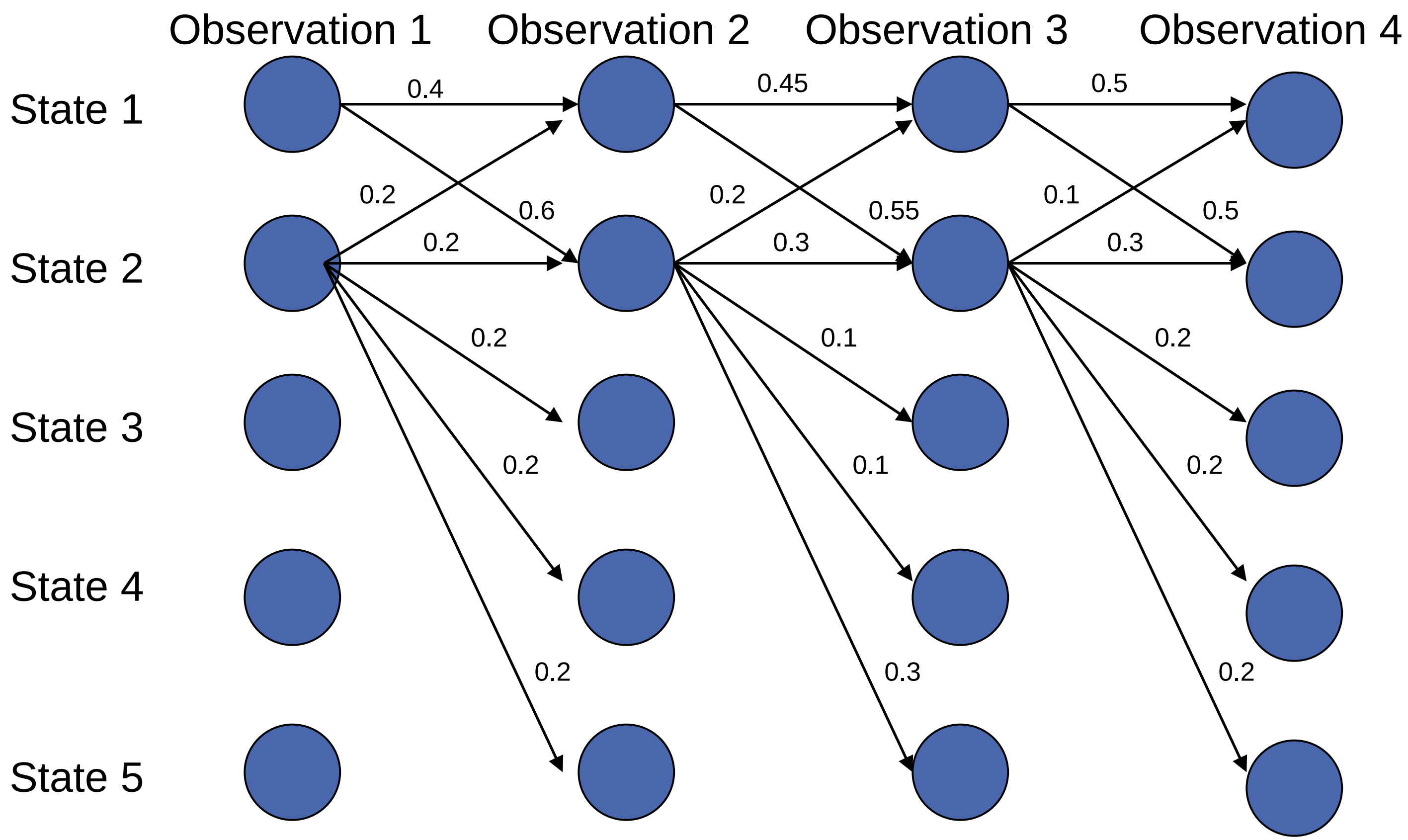
Probability of path 1->2->1->2:

- 0.6 X 0.2 X 0.5 = 0.06

Other paths:
1->1->1->1: 0.09
2->2->2->2: 0.018

[Referred From Label bias problem – Stanford Computer Science]

MEMM: Label bias problem



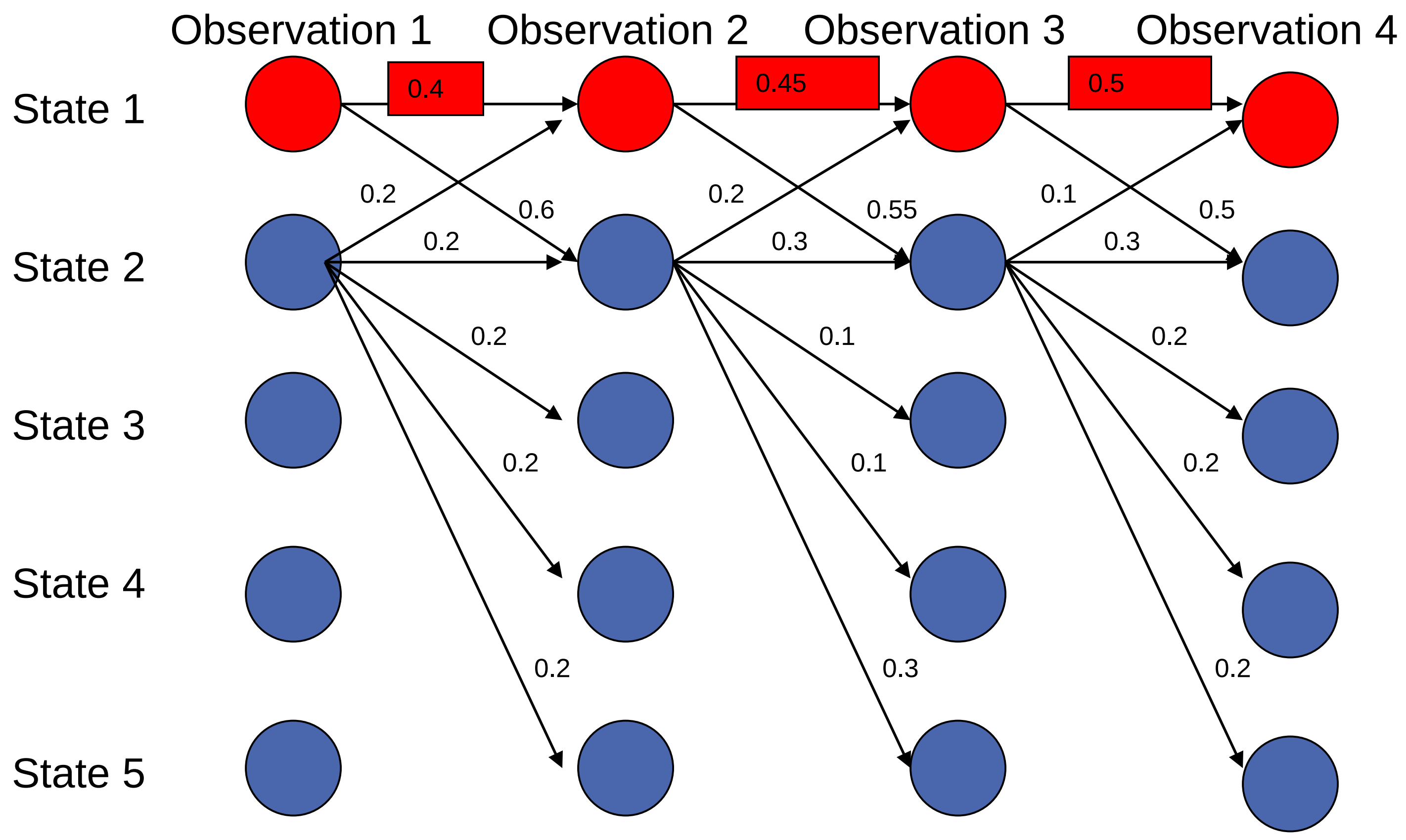
Probability of path 1->1->2->2:

- 0.4 X 0.55 X 0.3 = 0.066

Other paths:
1->1->1->1: 0.09
2->2->2->2: 0.018
1->2->1->2: 0.06

[Referred From Label bias problem – Stanford Computer Science]

MEMM: Label bias problem

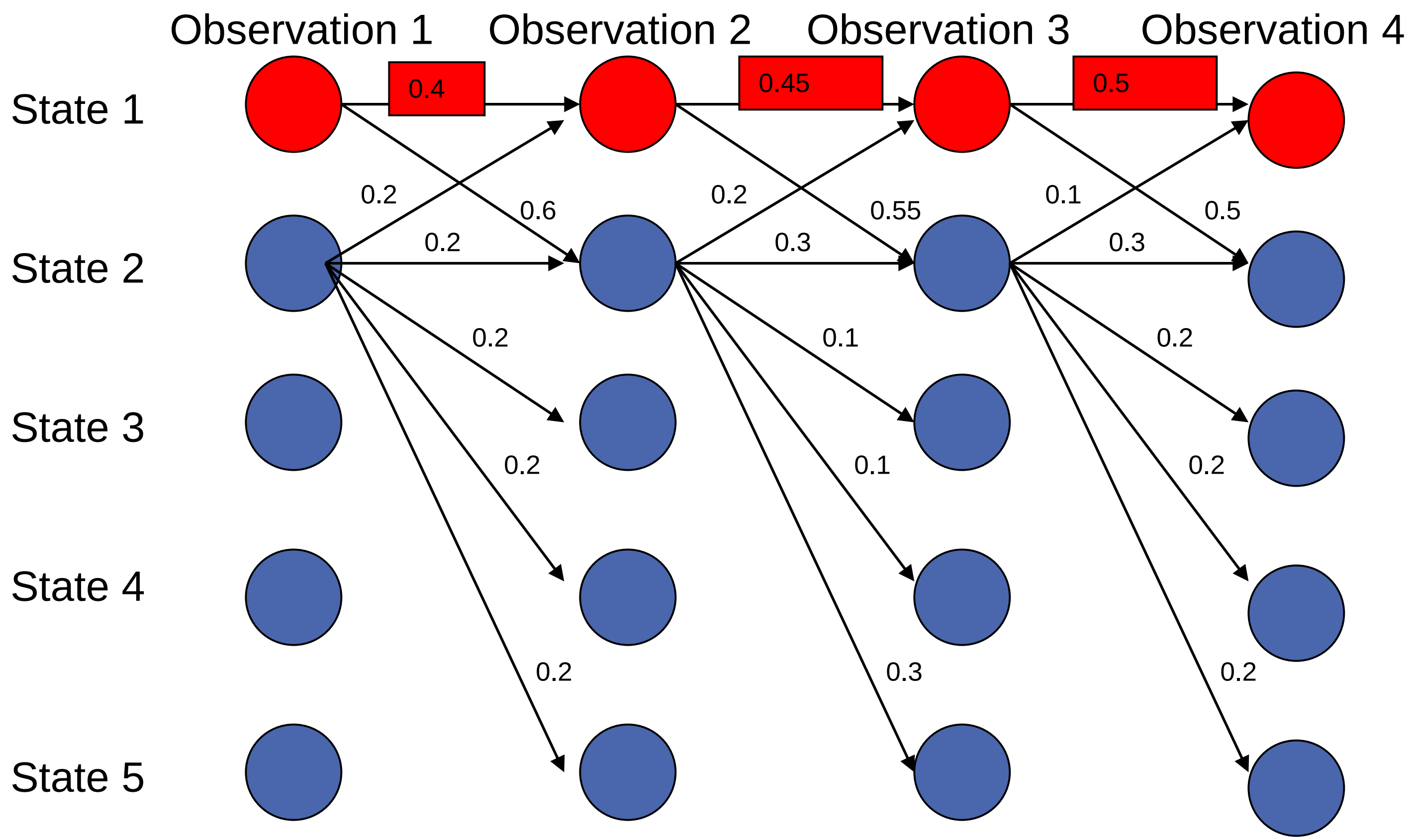


Most Likely Path: 1-> 1-> 1-> 1

- Although locally it seems state 1 wants to go to state 2 and state 2 wants to remain in state 2.
- **why?**

[Referred From Label bias problem – Stanford Computer Science]

MEMM: Label bias problem

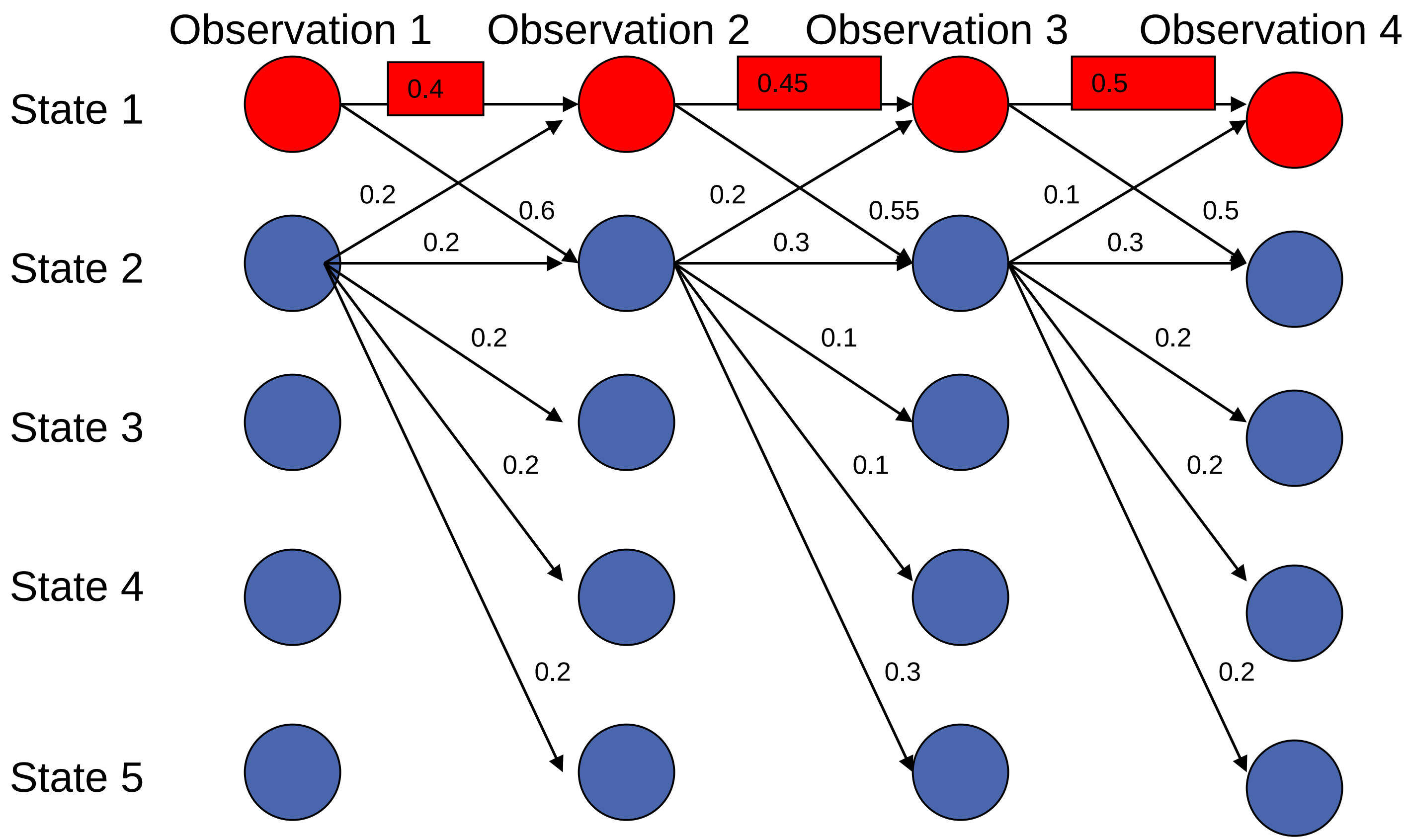


Most Likely Path: 1-> 1-> 1-> 1

- State 1 has only two transitions but state 2 has 5:
  - Average transition probability from state 2 is lower

Referred From Label bias problem – Stanford Computer Science

MEMM: Label bias problem

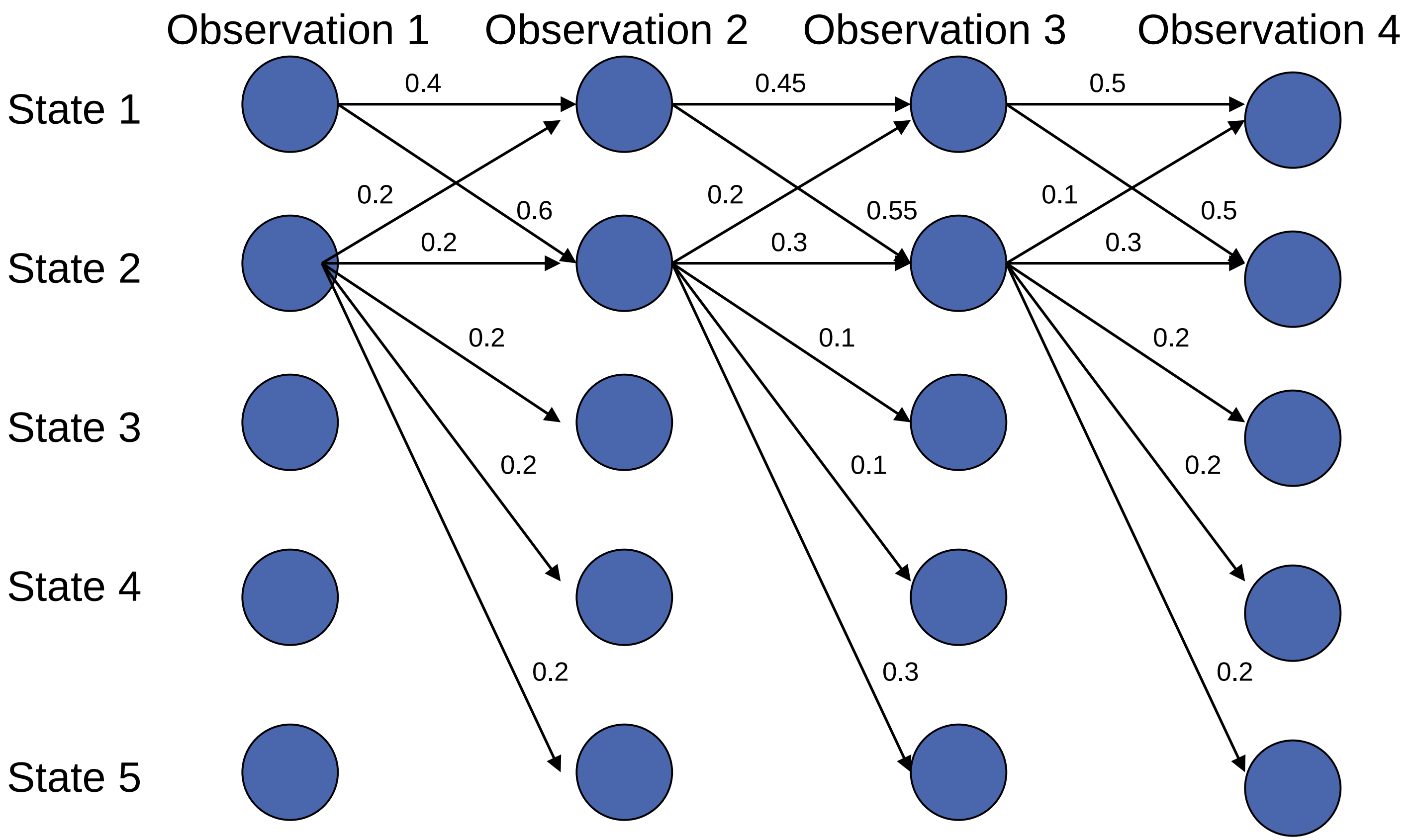


Label bias problem in MEMM:

- Preference of states with lower number of transitions over others

Referred From Label bias problem – Stanford Computer Science

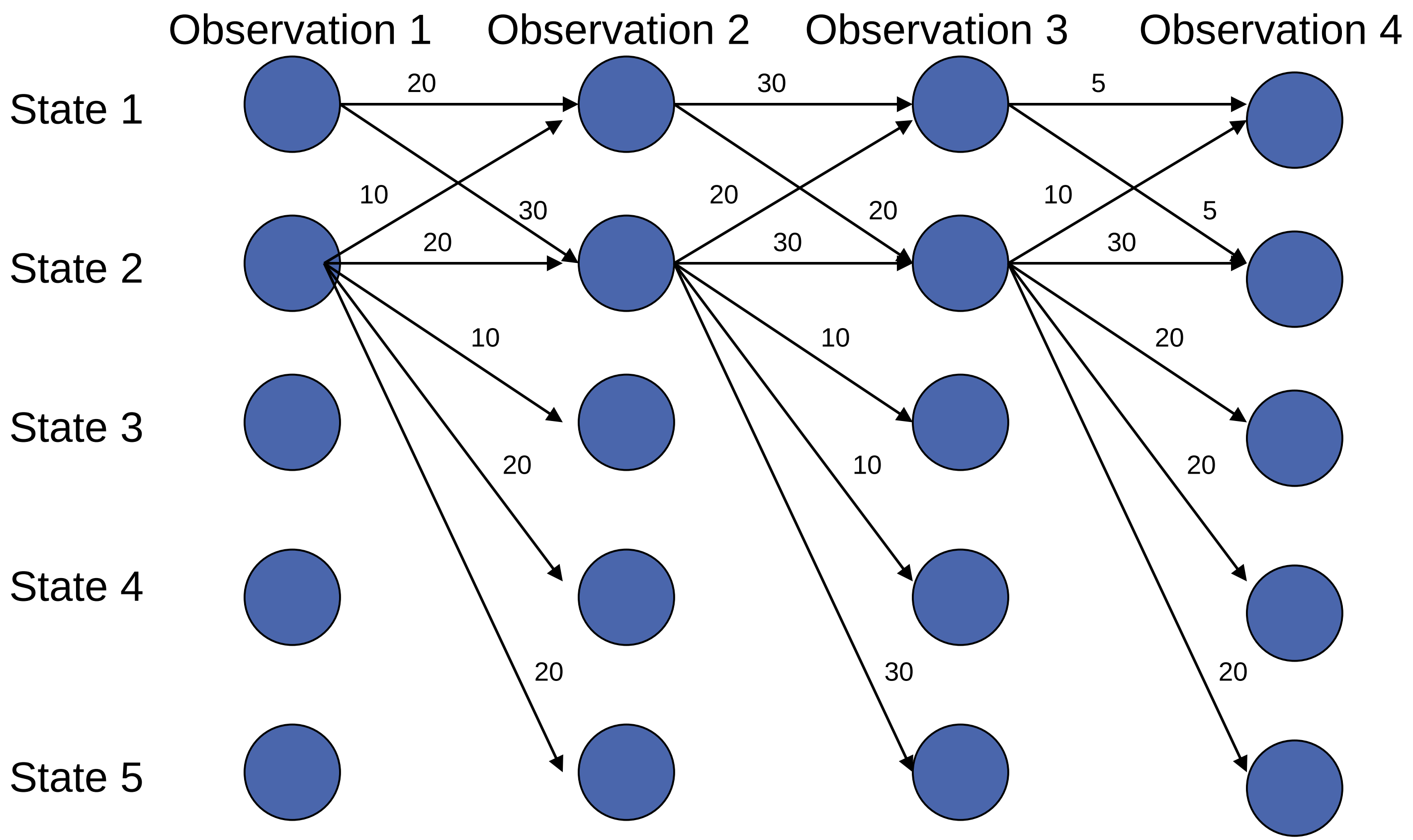
Solution: Do not normalize probabilities locally



From local probabilities ….

[Referred From Label bias problem – Stanford Computer Science]

Solution: Do not normalize probabilities locally

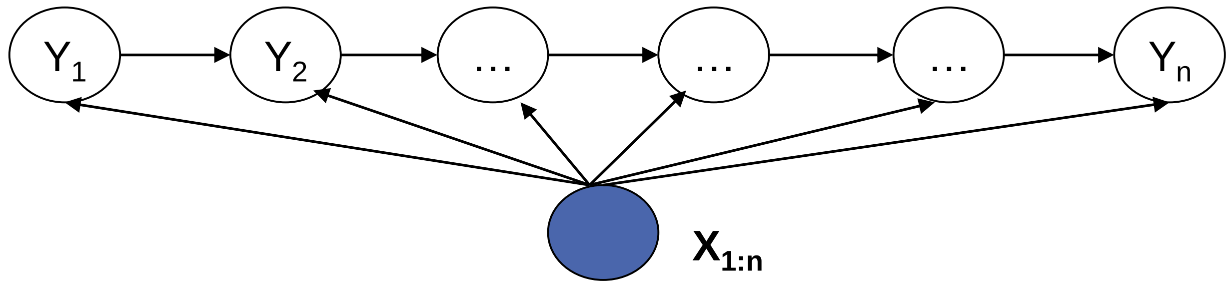


From local probabilities to local potentials

- States with lower transitions do not have an unfair advantage!

[Referred From Label bias problem – Stanford Computer Science]()

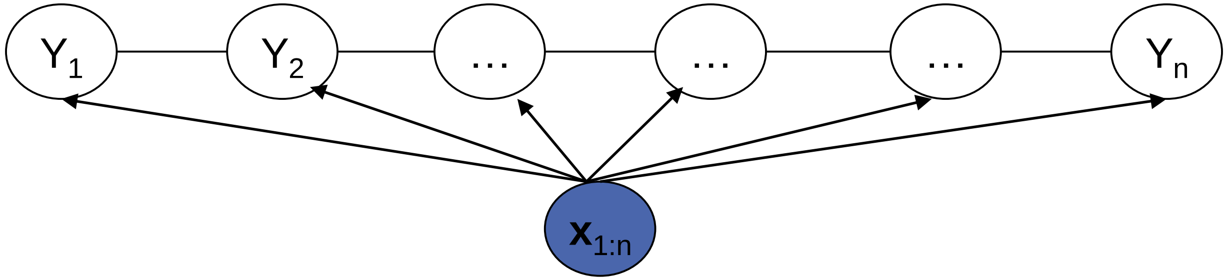
From MEMM ….



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) \quad = \quad \prod_{i=1}^{n} P(y_i|y_{i-1},\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}_{1:n}))}{Z(y_{i-1},\mathbf{x}_{1:n})}$$

[Referred From Label bias problem – Stanford Computer Science]

From MEMM to CRF



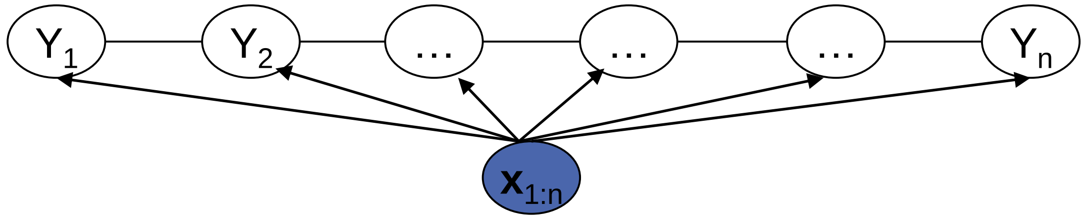
$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) \quad = \quad \frac{1}{Z(\mathbf{x}_{1:n})}\prod_{i=1}^{n}\phi(y_i, y_{i-1}, \mathbf{x}_{1:n}) = \frac{1}{Z(\mathbf{x}_{1:n})}\prod_{i=1}^{n}\exp(\mathbf{w}^T\mathbf{f}(y_i, y_{i-1}, \mathbf{x}_{1:n}))$$

CRF is a partially directed model

- Discriminative model like MEMM
- Usage of global normalizer Z(**x**) overcomes the label bias problem of MEMM
- Models the dependence between each state and the entire observation sequence (like MEMM)

Referred From Label bias problem – Stanford Computer Science

Conditional Random Fields

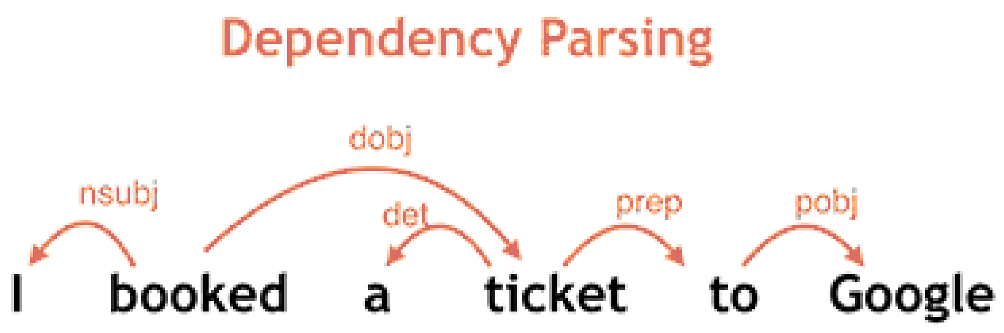


$$
\begin{aligned}
P(\mathbf{y}|\mathbf{x}) &= \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{i=1}^{n}\left(\sum_{k}\lambda_k f_k(y_i, y_{i-1}, \mathbf{x}) + \sum_{l}\mu_l g_l(y_i, \mathbf{x})\right)\right) \\
&= \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{i=1}^{n}(\lambda^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}) + \mu^T \mathbf{g}(y_i, \mathbf{x}))\right)
\end{aligned}
$$

$$
\text{where } Z(\mathbf{x}) = \sum_{\mathbf{y}}\exp\left(\sum_{i=1}^{n}(\lambda^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}) + \mu^T \mathbf{g}(y_i, \mathbf{x}))\right)
$$

[Referred From Label bias problem – Stanford Computer Science]

Dependency Parsing?



Transition Systems

$$S = (C,\ T,\ c_s,\ C_t)$$

- C: A set of configurations
- T: A set of Transitions
- c_s: An initialization function
- C_t: A set of terminal configurations

$$C = (\Sigma,\ B,\ A)$$

- Σ: STACK
- B: Buffer
- A: Arcs

Referred from Joakim Nivre's [class material](), and [presentation]()

Transition Systems

Initialization: $c_s(x = x_1, \ldots, x_n) = ([0], [1, \ldots, n], \emptyset)$

Terminal: $C_t = \{c \in C | c = ([0], [\,], A)\}$

Transitions:

$$(\sigma, [i|\beta], A) \Rightarrow ([\sigma|i], \beta, A) \quad (\textsc{Shift})$$

$$([\sigma|i|j], B, A) \Rightarrow ([\sigma|j], B, A \cup \{(j, l, i)\})^1 \quad (\textsc{Left-Arc}_l)$$

$$([\sigma|i|j], B, A) \Rightarrow ([\sigma|i], B, A \cup \{(i, l, j)\}) \quad (\textsc{Right-Arc}_l)$$

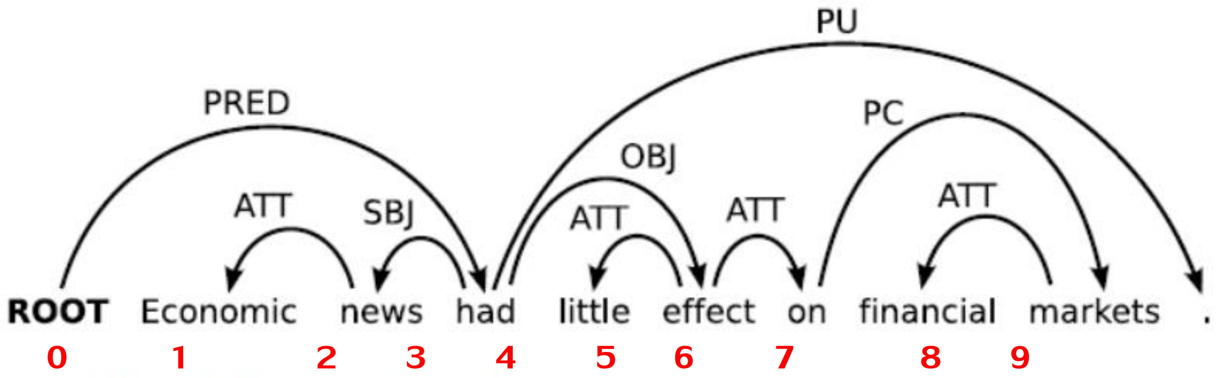
[1] Permitted only if $i \neq 0$.

The notation [σ|i] (for the stack) denotes a right-headed list with head i and tail σ.

The notation [j|β] (for the buffer) denotes a left-headed list with head j and tail β.

Referred from Joakim Nivre's class material, and presentation

Transition Sequence For Arc-Standard System



Transition	Configuration			
$c_s(x) =$	(	$[0]$,	$[1,\ldots,9]$,	$\emptyset$)
$\textsc{Shift} \Longrightarrow$	(	$[0,1]$,	$[2,\ldots,9]$,	$\emptyset$)
$\textsc{Shift} \Longrightarrow$	(	$[0,1,2]$,	$[3,\ldots,9]$,	$\emptyset$)
$\textsc{Left-Arc}_{\textsc{att}} \Longrightarrow$	(	$[0,2]$,	$[3,\ldots,9]$,	$A_1 = \{(2,\textsc{att},1)\}$)
$\textsc{Shift} \Longrightarrow$	(	$[0,2,3]$,	$[4,\ldots,9]$,	A_1)
$\textsc{Left-Arc}_{\textsc{sbj}} \Longrightarrow$	(	$[0,3]$,	$[4,\ldots,9]$,	$A_2 = A_1 \cup \{(3,\textsc{sbj},2)\}$)
$\textsc{Shift} \Longrightarrow$	(	$[0,3,4]$,	$[5,\ldots,9]$,	A_2)
$\textsc{Shift} \Longrightarrow$	(	$[0,\ldots,5]$,	$[6,\ldots,9]$,	A_2)
$\textsc{Left-Arc}_{\textsc{att}} \Longrightarrow$	(	$[0,3,5]$,	$[6,\ldots,9]$,	$A_3 = A_2 \cup \{(5,\textsc{att},4)\}$)
$\textsc{Shift} \Longrightarrow$	(	$[0,\ldots,6]$,	$[7,8,9]$,	A_3)
$\textsc{Shift} \Longrightarrow$	(	$[0,\ldots 7]$,	$[8,9]$,	A_3)
$\textsc{Shift} \Longrightarrow$	(	$[0,\ldots,8]$,	$[9]$,	A_3)
$\textsc{Left-Arc}_{\textsc{att}} \Longrightarrow$	(	$[0,\ldots 8]$,	$[9]$,	$A_4 = A_3 \cup \{(8,\textsc{att},7)\}$)
$\textsc{Right-Arc}_{\textsc{pc}} \Longrightarrow$	(	$[0,\ldots,6]$,	$[9]$,	$A_5 = A_4 \cup \{(6,\textsc{pc},8)\}$)
$\textsc{Right-Arc}_{\textsc{att}} \Longrightarrow$	(	$[0,3,5]$,	$[9]$,	$A_6 = A_5 \cup \{(5,\textsc{att},6)\}$)
$\textsc{Right-Arc}_{\textsc{obj}} \Longrightarrow$	(	$[0,3]$,	$[9]$,	$A_7 = A_6 \cup \{(3,\textsc{obj},5)\}$)
$\textsc{Shift} \Longrightarrow$	(	$[0,3,9]$,	$[\,]$,	A_7)
$\textsc{Right-Arc}_{\textsc{pu}} \Longrightarrow$	(	$[0,3]$,	$[\,]$,	$A_8 = A_7 \cup \{(3,\textsc{pu},9)\}$)
$\textsc{Right-Arc}_{\textsc{root}} \Longrightarrow$	(	$[0]$,	$[\,]$,	$A_9 = A_8 \cup \{(0,\textsc{root},3)\}$)
		Stack	Buffer	Arcs

Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

$[\text{ROOT}]_\Sigma$ $[\text{Economic, news, had, little, effect, on, financial, markets, .}]_B$

ROOT Economic news had little effect on financial markets .

Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

$[\text{ROOT}, \text{Economic}]_{\Sigma}$ $[\text{news}, \text{had}, \text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, .]_{B}$

ROOT Economic news had little effect on financial markets .

Referred from Joakim Nivre's [class material](), and [presentation]()

How Arc-Standard Transition works?

$[\text{ROOT}, \text{Economic}, \text{news}]_{\Sigma}$ $[\text{had}, \text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, .]_{B}$

ROOT **Economic** **news** **had** **little** **effect** **on** **financial** **markets** **.**

Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

$[\text{ROOT}, \text{news}]_{\Sigma}$ $[\text{had}, \text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, .]_{B}$

Referred from Joakim Nivre’s [class material](), and [presentation]()

How Arc-Standard Transition works?

$[\text{ROOT}, \text{news}, \text{had}]_\Sigma \quad [\text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, .]_B$

Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

$[\text{ROOT}, \text{had}]_{\Sigma}$ $[\text{little}, \text{effect}, \text{on}, \text{financial}, \text{markets}, .]_{B}$

Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

$[\text{ROOT}, \text{had}, \text{little}]_{\Sigma}$ $[\text{effect}, \text{on}, \text{financial}, \text{markets}, .]_{B}$

Referred from Joakim Nivre's [class material](), and [presentation]()

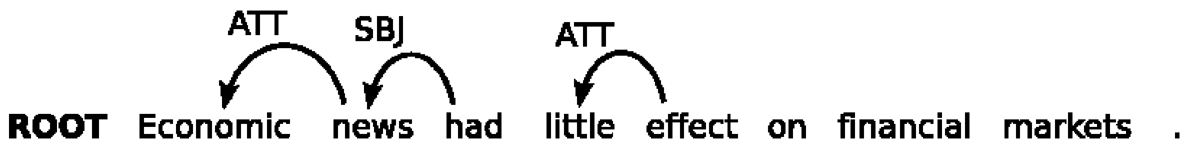
How Arc-Standard Transition works?

$[\text{ROOT}, \text{had}, \text{little}, \text{effect}]_{\Sigma}$ $[\text{on}, \text{financial}, \text{markets}, .]_{B}$

Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

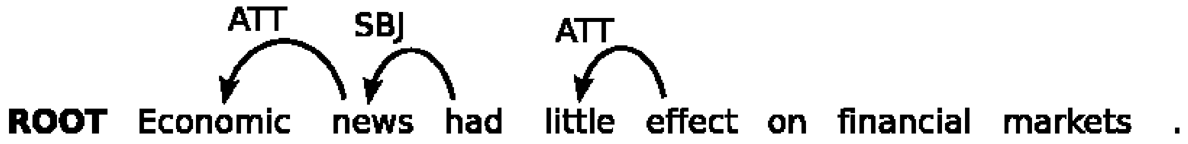
$[\text{ROOT}, \text{had}, \text{effect}]_{\Sigma}$ $[\text{on}, \text{financial}, \text{markets}, .]_{B}$



Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

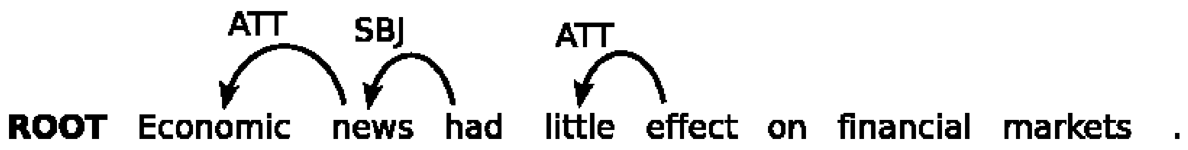
$[\text{ROOT}, \text{had}, \text{effect}, \text{on}]_{\Sigma}$ $[\text{financial}, \text{markets}, .]_{B}$



Referred from Joakim Nivre's [class material](), and [presentation]()

How Arc-Standard Transition works?

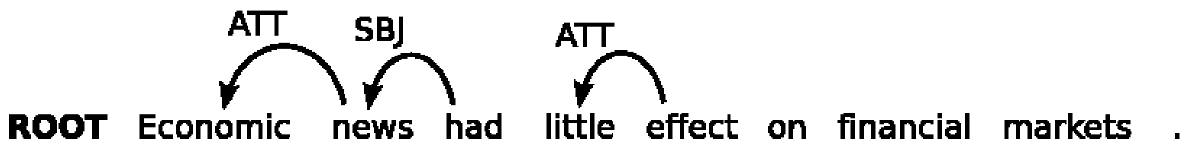
$[\text{ROOT}, \text{had}, \text{effect}, \text{on}, \text{financial}]_{\Sigma}$ $[\text{markets}, .]_{B}$



Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

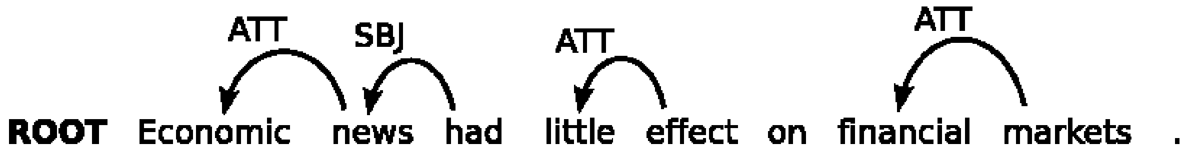
$[\text{ROOT}, \text{had}, \text{effect}, \text{on}, \text{financial}, \text{markets}]_{\Sigma} \quad [.]_{B}$



Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

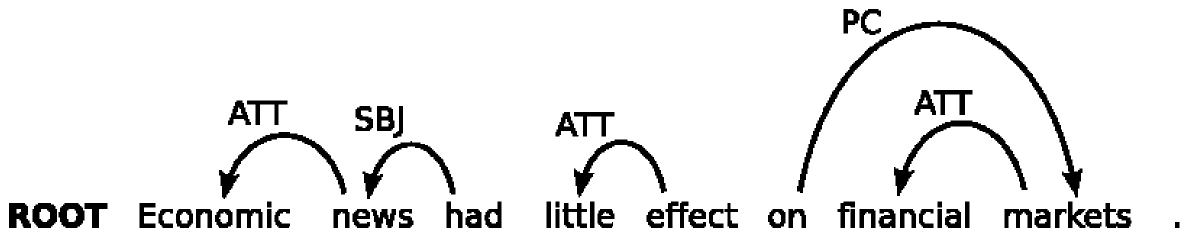
$[\text{ROOT}, \text{had}, \text{effect}, \text{on}, \text{markets}]_{\Sigma} \quad [.]_{B}$



Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

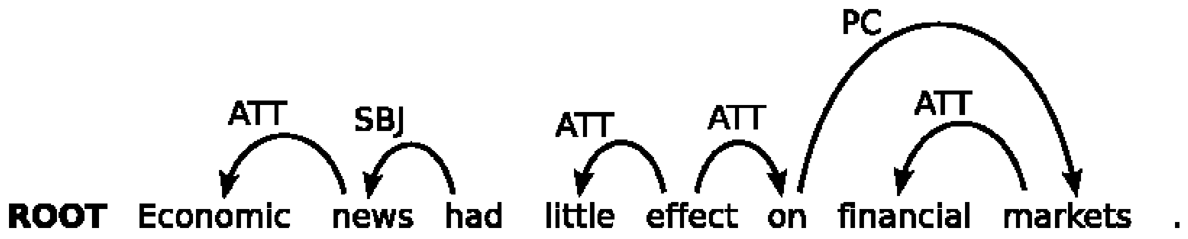
$[\text{ROOT}, \text{had}, \text{effect}, \text{on}]_{\Sigma} \quad [.]_{B}$



Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

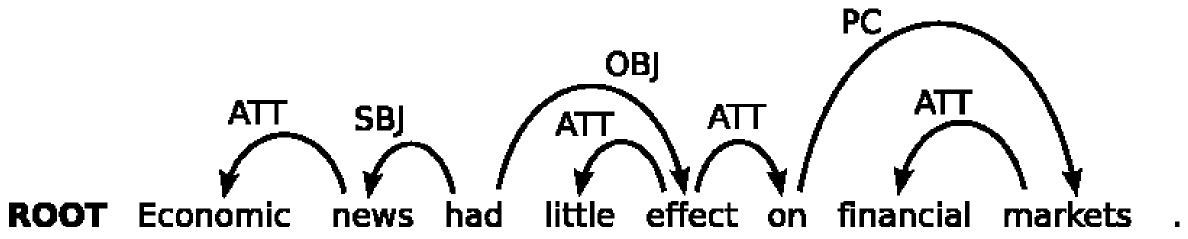
[ROOT, had, effect]$_\Sigma$ [.]$_B$



Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?

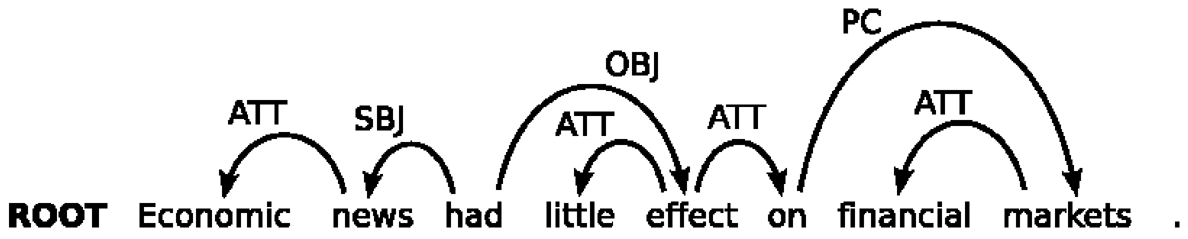
$[\text{ROOT}, \text{had}]_\Sigma \quad [.]_B$



Referred from Joakim Nivre's [class material](), and [presentation]()

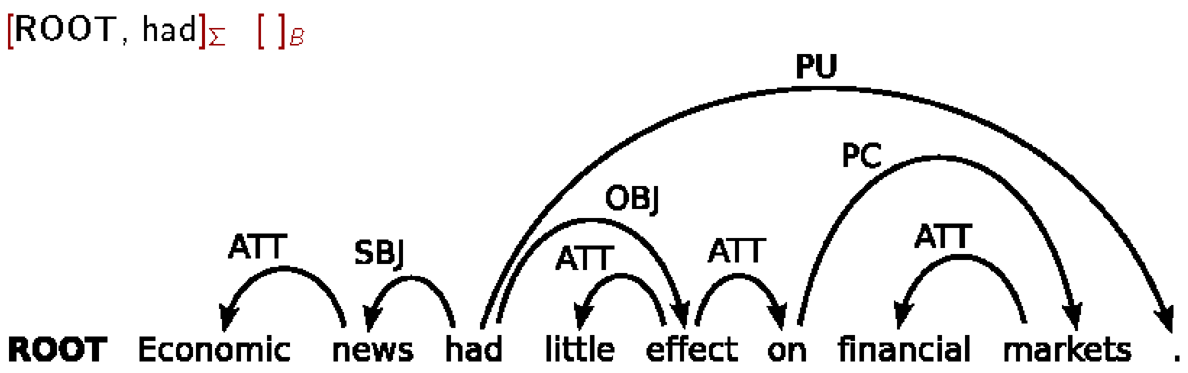
How Arc-Standard Transition works?

$[\text{ROOT}, \text{had}, .]_{\Sigma} \quad [\;]_{B}$



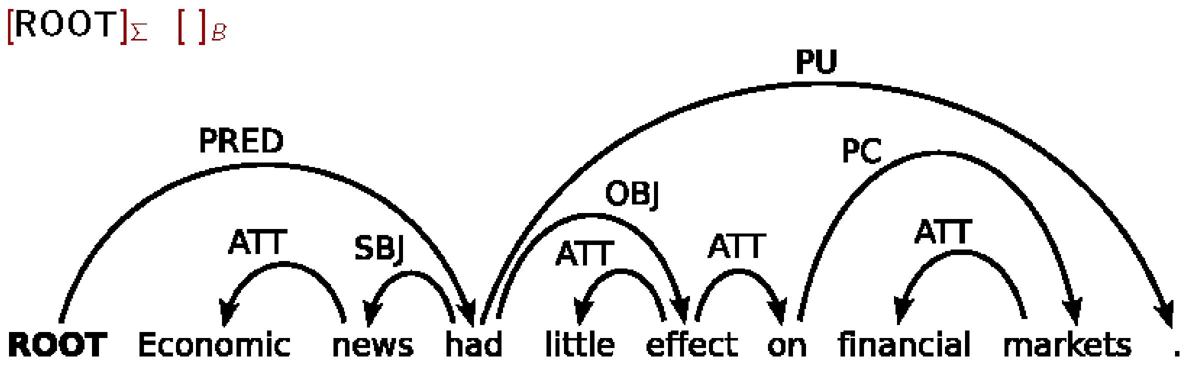
Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?



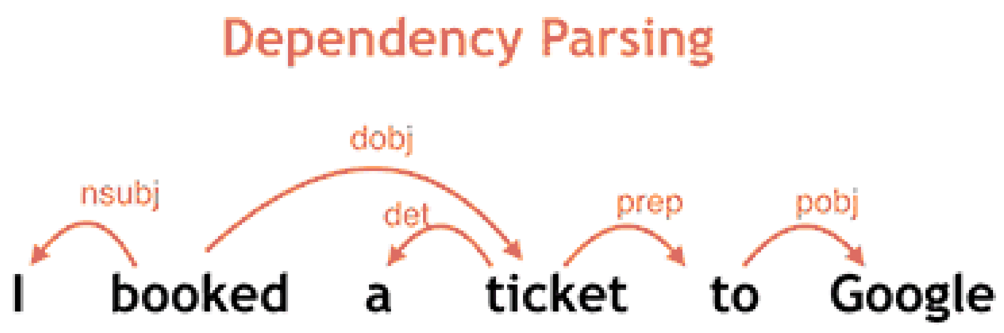
Referred from Joakim Nivre's class material, and presentation

How Arc-Standard Transition works?



Referred from Joakim Nivre's class material, and presentation

Dependency Parsing?



Question & Answer

Referrences

- A Fast and Accurate Dependency Parser using Neural Networks
  - By Danqi Chen, Christopher D. Manning (Stanford)
- Structured Training for Neural Network Transition-Based Parsing
  - By David Wiess, Chris Alberti, Michael Collins, Slav Petrov (Google)
- Globally Normalized Transition-Based Neural Networks
  - By Daniel Andor, Chris Alberti, David Wiess, … (Google)
- Transition-Based Parsing
  - class material by Joakim Nivre

Thank you