

Copyright Notice

These slides are distributed under the Creative Commons License.

[DeepLearning.AI](#) makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite [DeepLearning.AI](#) as the source of the slides.

For the rest of the details of the license, see <https://creativecommons.org/licenses/by-sa/2.0/legalcode>



deeplearning.ai

Part of Speech Tagging

Outline

- What is part of speech tagging?
- Markov chains
- Hidden Markov models
- Viterbi algorithm
- Example
- Coding assignment!

What is part of speech?

Why not learn something ?

adverb

adverb

verb

noun

punctuation
mark,
sentence
closer

Part of speech (POS) tagging

Part of speech tags:

lexical term	tag	example
noun	NN	something, nothing
verb	VB	learn, study
determiner	DT	the, a
w-adverb	WRB	why, where
...	...	

Why not learn something ?

WRB **RB** **VB** **NN** .

Applications of POS tagging



Named entities



Co-reference resolution



Speech recognition



deeplearning.ai

Markov Chains

Example

Why not learn  ...

verb **verb?**
 noun?
 ... ?

Part of Speech Dependencies

Why not learn  ...

verb verb?
└─> **noun?**
...?

The Most Likely Next Word

Why not learnswimming?

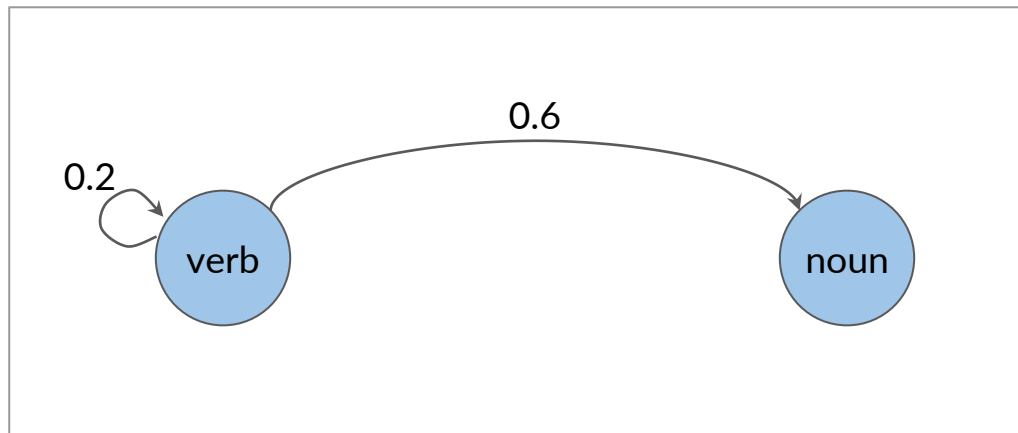
verb **noun**

Less Likely Words

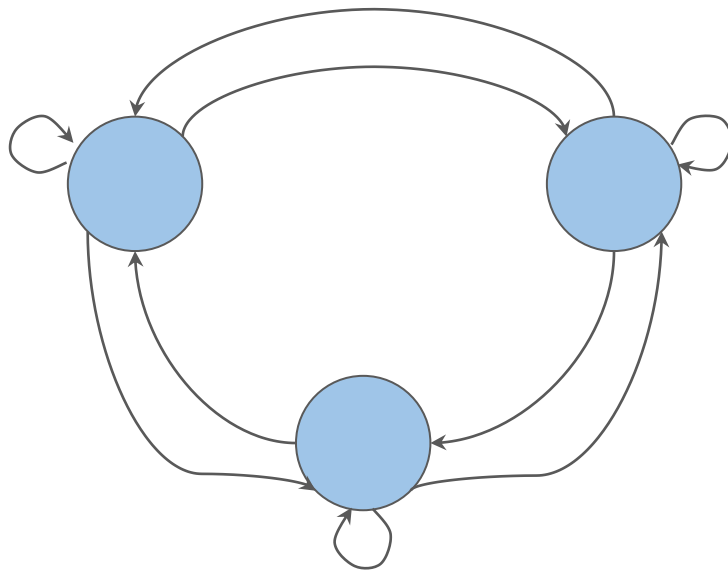
Why not learnswim?

verb verb

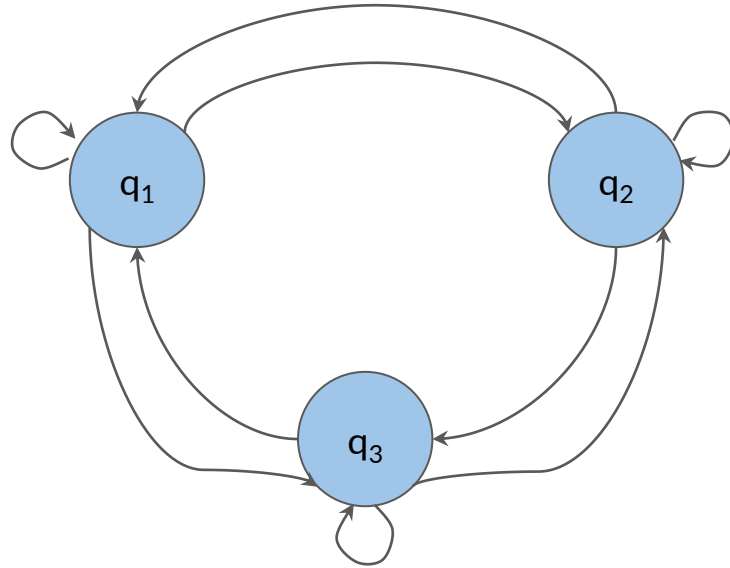
Visual Representation



What are Markov chains?



States



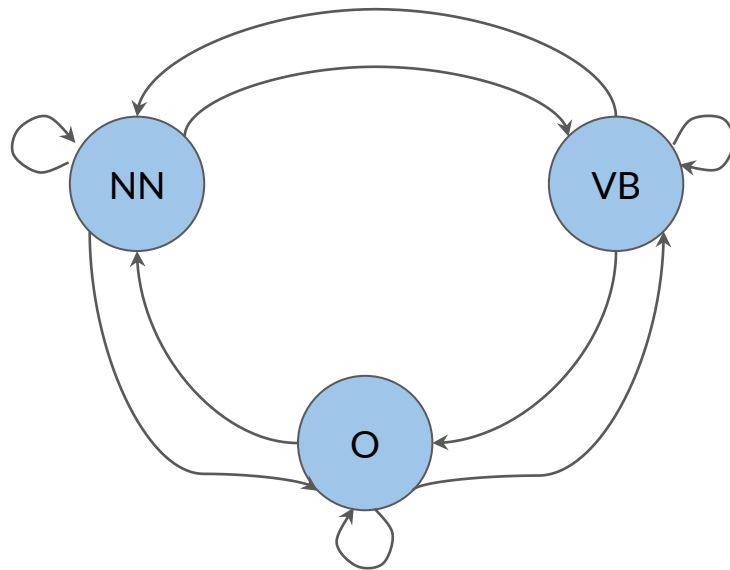
$$Q = \{q_1, q_2, q_3\}$$



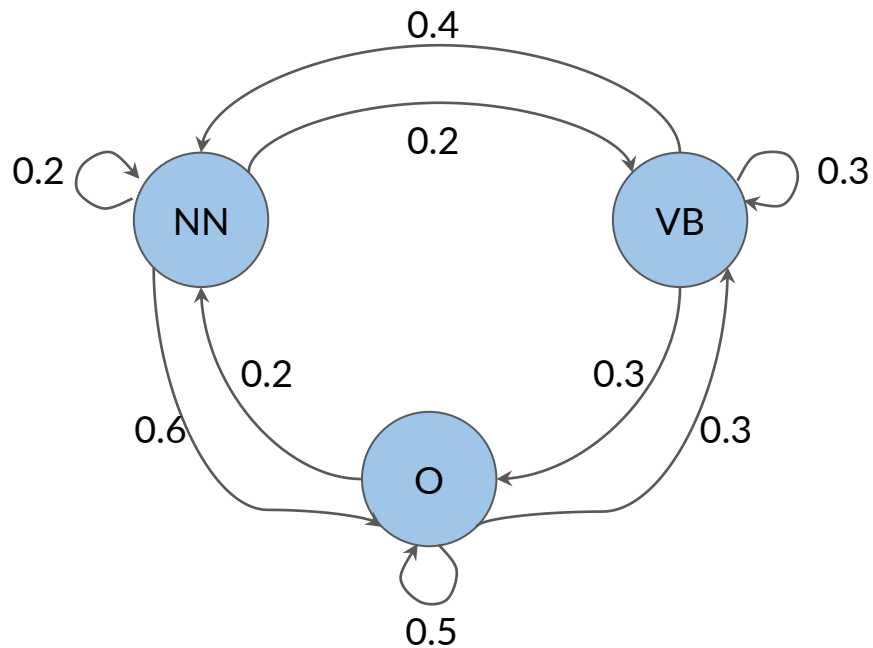
deeplearning.ai

Markov Chains and POS Tags

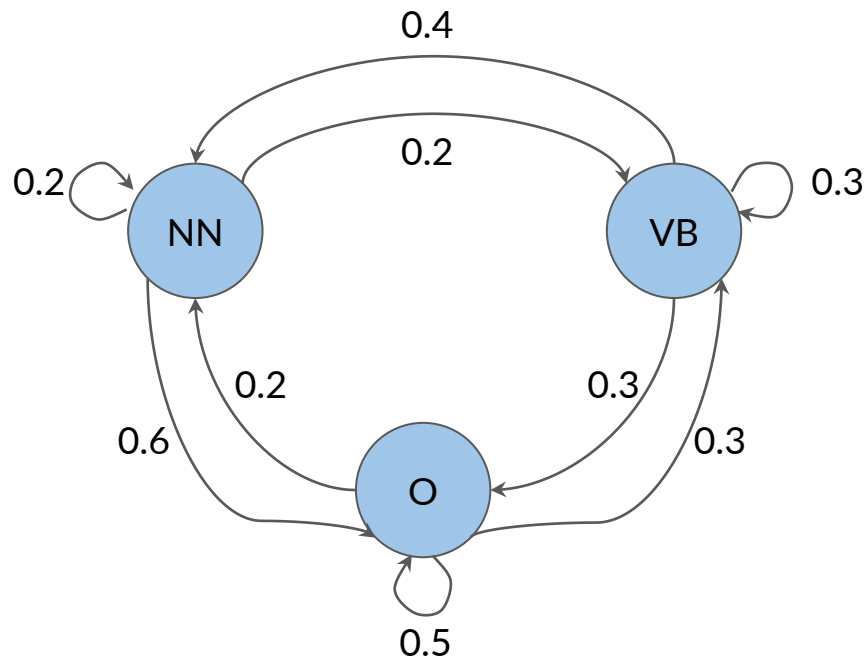
POS tags as States



Transition probabilities

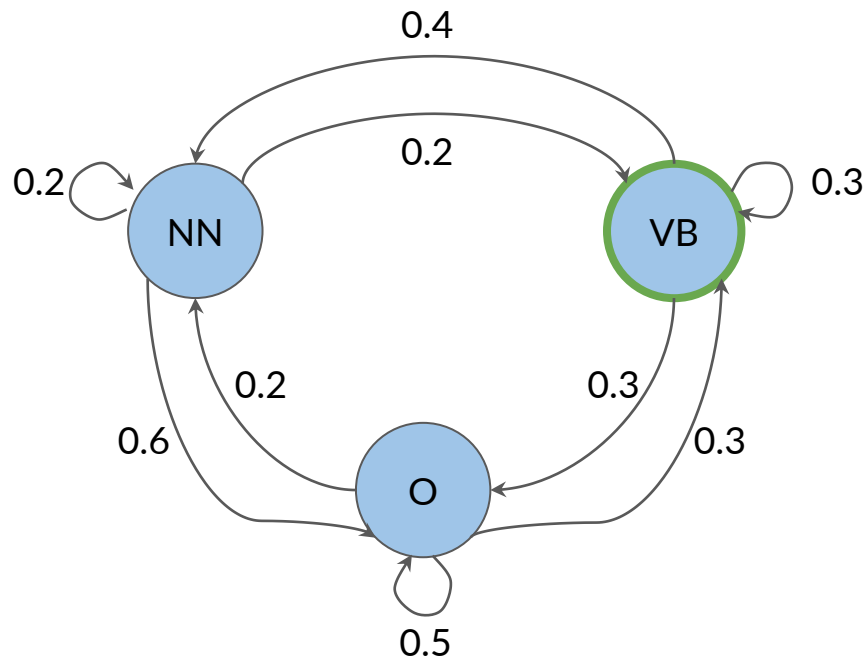


Transition probabilities



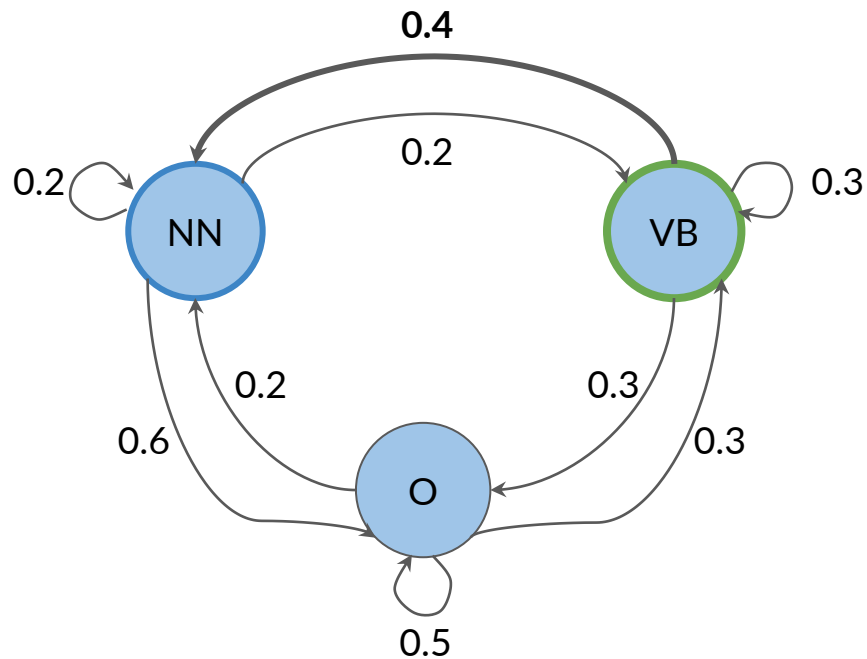
Why not **learn** something ?

Transition probabilities



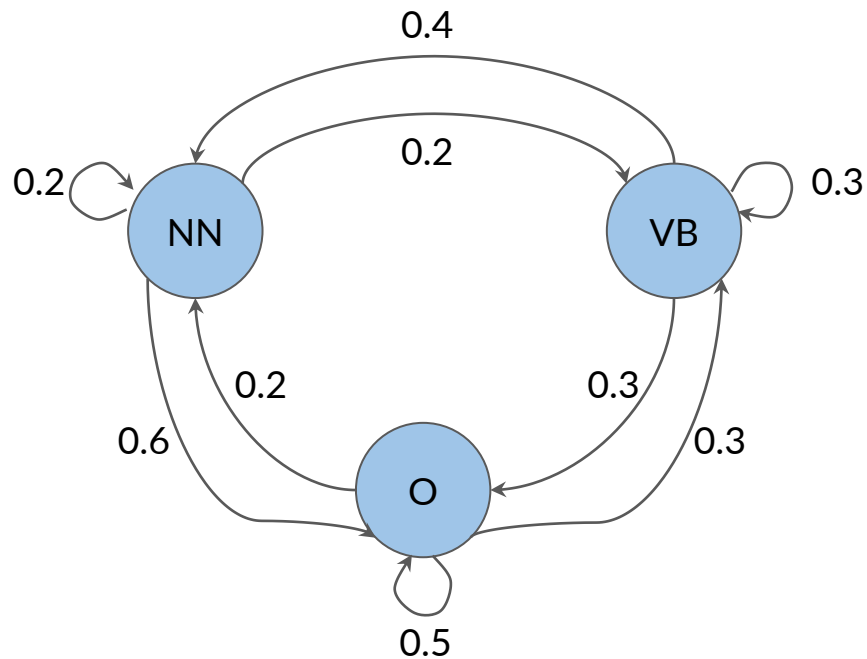
Why not **learn** something ?

Transition probabilities



Why not learn something ?

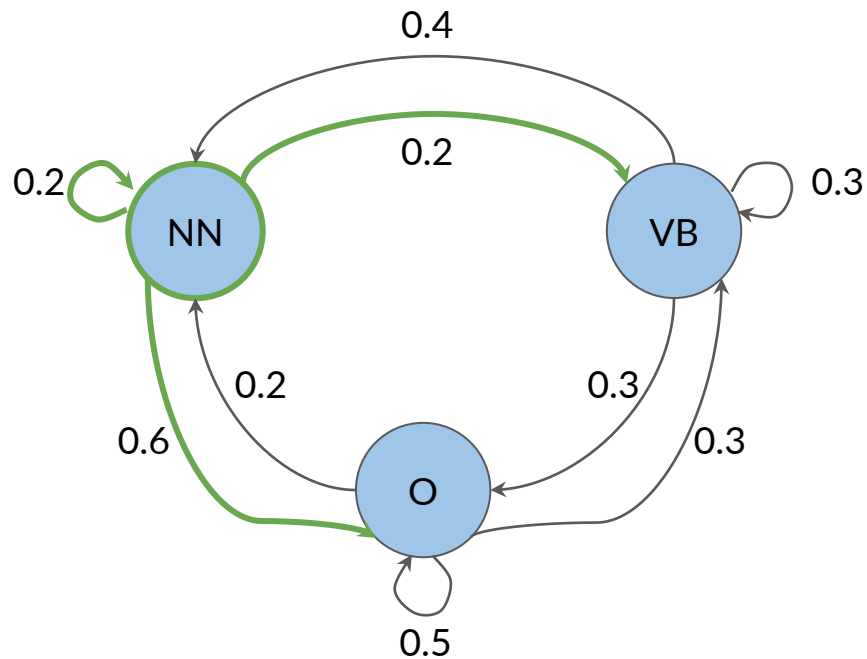
The transition matrix



$A =$

	NN	VB	O
NN (noun)	0.2	0.2	0.6
VB (verb)	0.4	0.3	0.3
O (other)	0.2	0.3	0.5

The transition matrix

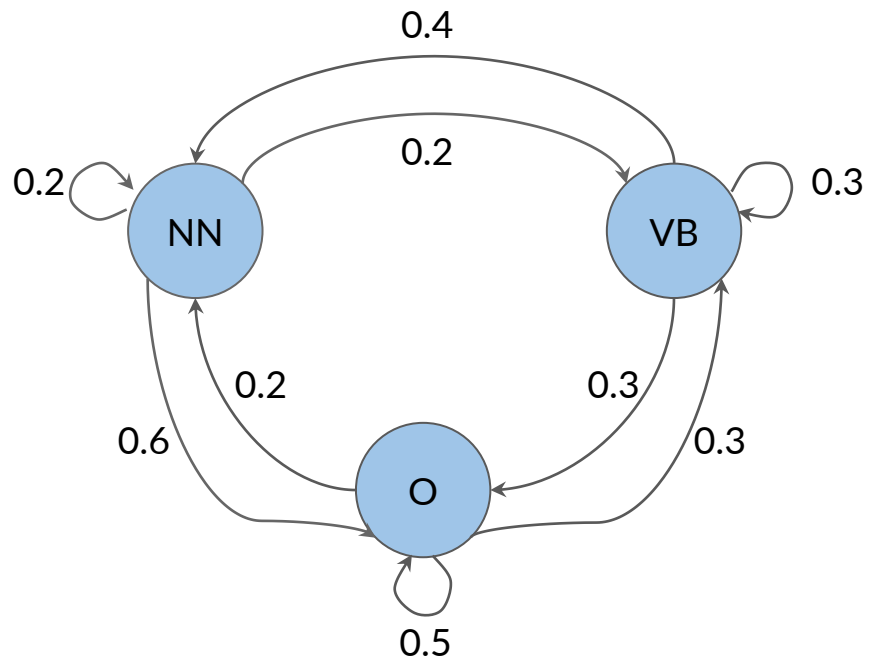


$A =$

	NN	VB	O
NN (noun)	0.2	0.2	0.6
VB (verb)	0.4	0.3	0.3
O (other)	0.2	0.3	0.5

$$\sum_{j=1}^N a_{ij} = 1$$

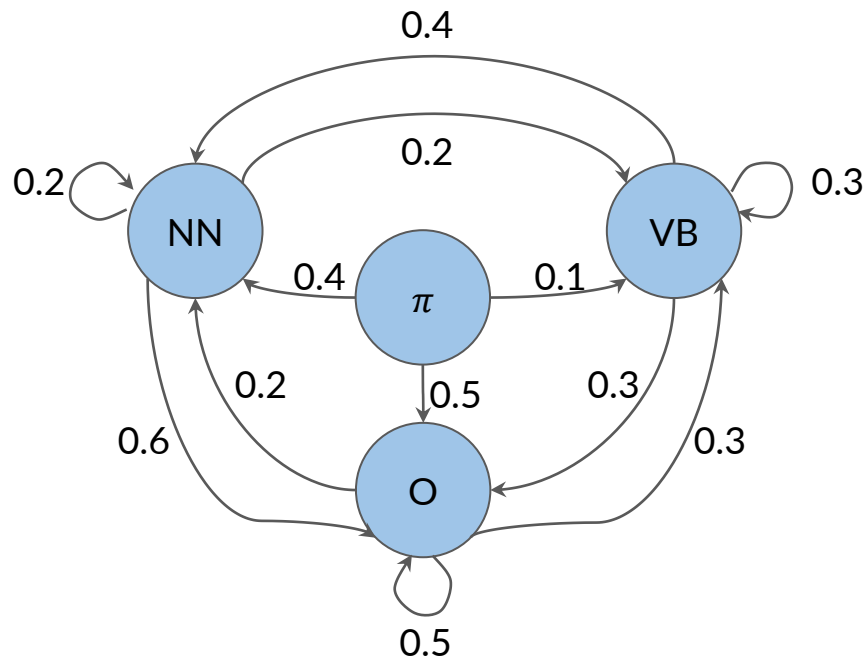
The first word



Why not learn something ?

NN?
VB?
O?

Initial probabilities



$A =$

	NN	VB	O
π (initial)	0.4	0.1	0.5
NN (noun)	0.2	0.2	0.6
VB (verb)	0.4	0.3	0.3
O (other)	0.2	0.3	0.5

Transition table and matrix

$A =$

	NN	VB	O
π (initial)	0.4	0.1	0.5
NN (noun)	0.2	0.2	0.6
VB (verb)	0.4	0.3	0.3
O (other)	0.2	0.3	0.5

$$A = \begin{pmatrix} 0.4 & 0.1 & 0.5 \\ 0.2 & 0.2 & 0.6 \\ 0.4 & 0.3 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

Summary

States

$$Q = \{q_1, \dots, q_N\}$$

Transition matrix

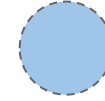
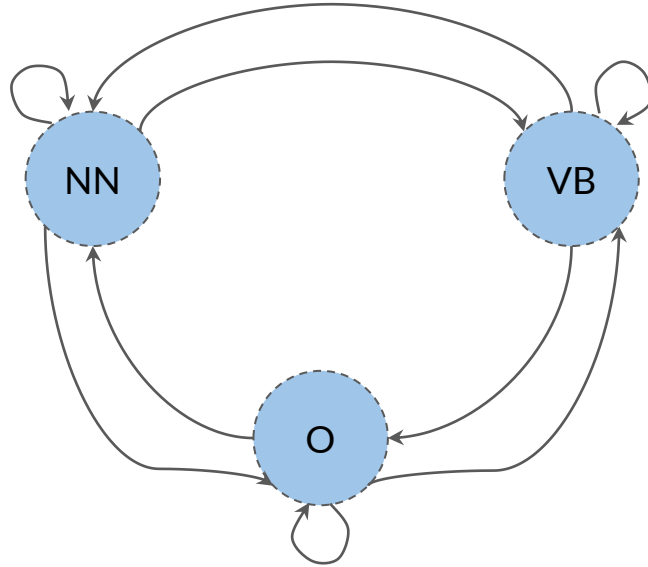
$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{N+1,1} & \dots & a_{N+1,N} \end{pmatrix}$$



deeplearning.ai

Hidden Markov Models

Hidden Markov Model



hidden states

you



jump = verb

machine



jump = ?

you



jump =
verb
run = verb
fly = verb

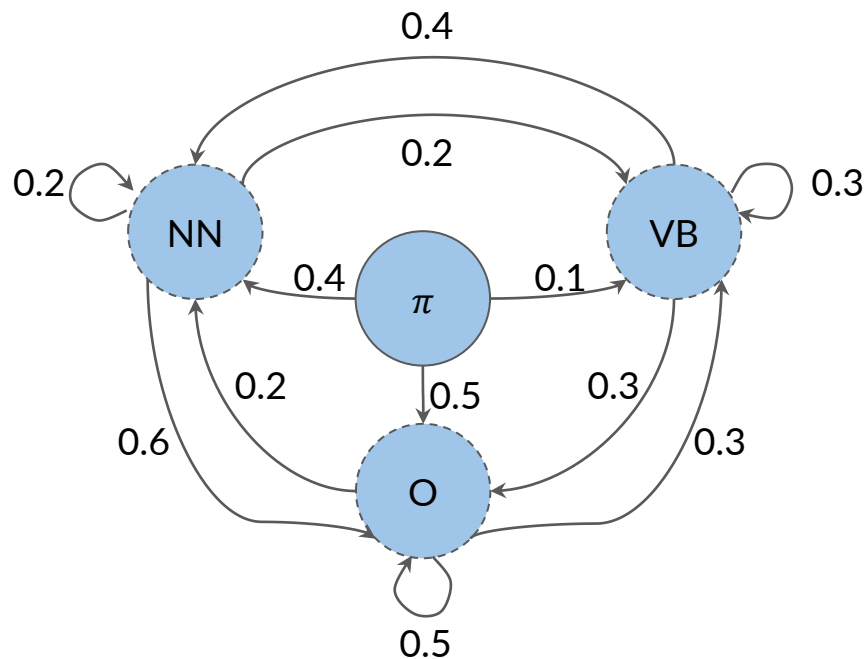
machine



jump*
run
fly

*observable

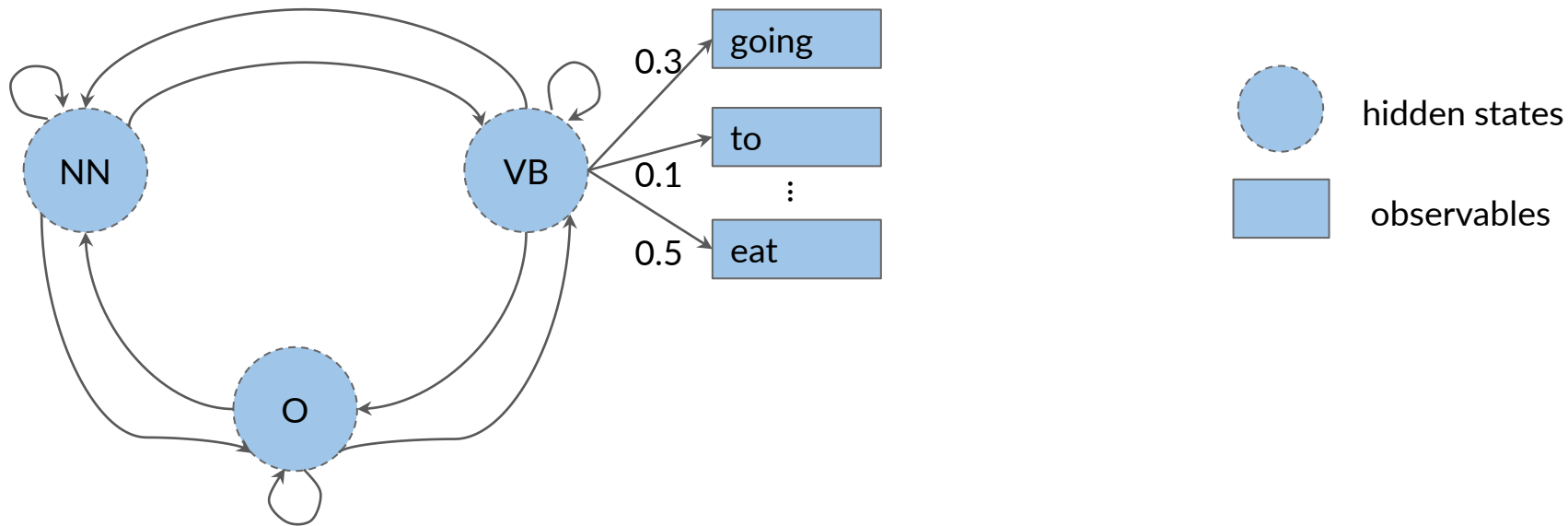
Transition probabilities



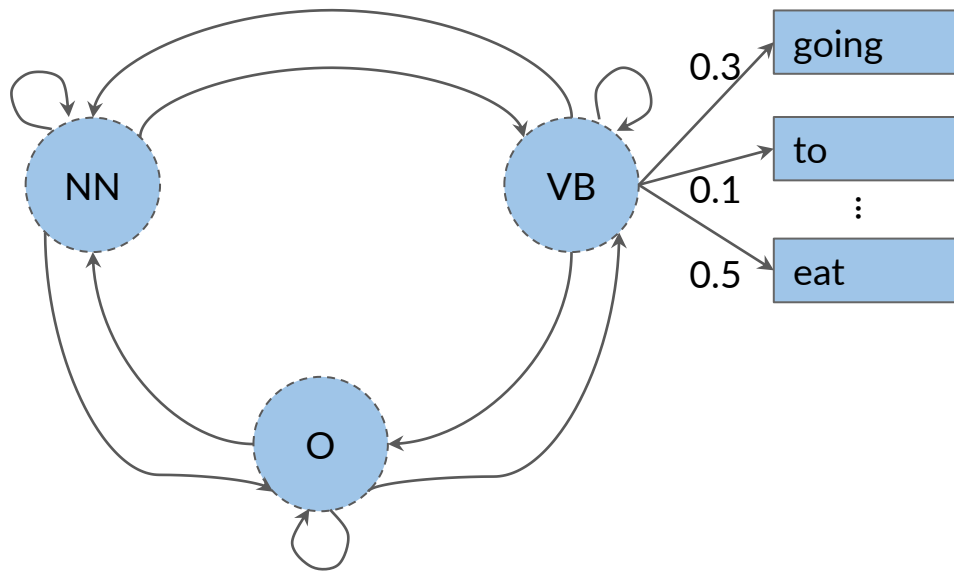
$A =$

	NN	VB	O
π (initial)	0.4	0.1	0.5
NN (noun)	0.2	0.2	0.6
VB (verb)	0.4	0.3	0.3
O (other)	0.2	0.3	0.5

Emission probabilities



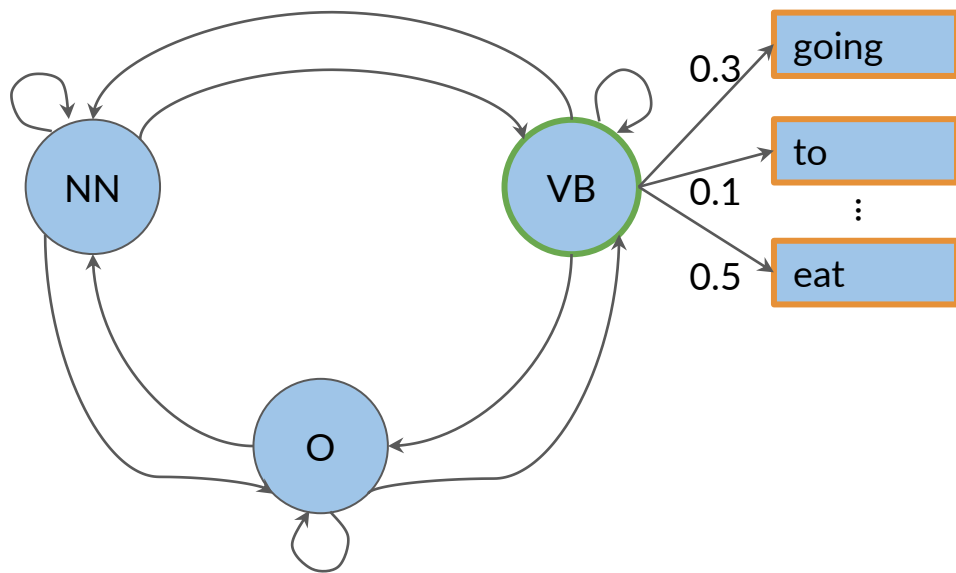
Emission probabilities



$B =$

	going	to	eat	...
NN (noun)	0.5	0.1	0.02	
VB (verb)	0.3	0.1	0.5	
O (other)	0.3	0.5	0.68	

Emission probabilities



$B =$

	going	to	eat	...
NN (noun)	0.5	0.1	0.02	
VB (verb)	0.3	0.1	0.5	
O (other)	0.3	0.5	0.68	

The emission matrix

$B =$

	going	to	eat	...
NN (noun)	0.5	0.1	0.02	
VB (verb)	0.3	0.1	0.5	
O (other)	0.3	0.5	0.68	

$$\sum_{j=1}^V b_{ij} = 1$$

He lay on his back.

I'll be back.

Summary

States

$$Q = \{q_1, \dots, q_N\}$$

Transition matrix

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{N+1,1} & \dots & a_{N+1,N} \end{pmatrix}$$

Emission matrix

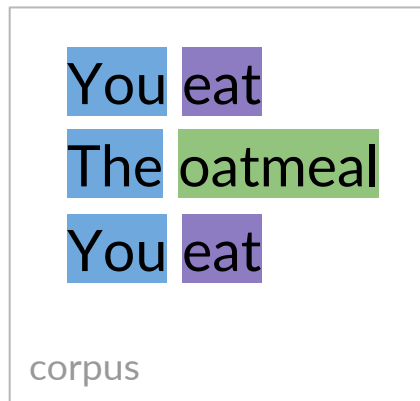
$$B = \begin{pmatrix} b_{11} & \dots & b_{1V} \\ \vdots & \ddots & \vdots \\ b_{N1} & \dots & b_{NV} \end{pmatrix}$$



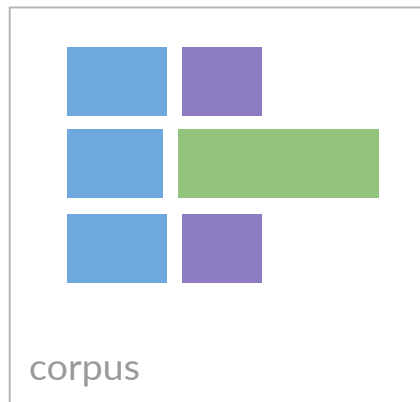
deeplearning.ai

Calculating Probabilities

Transition probabilities



Transition probabilities

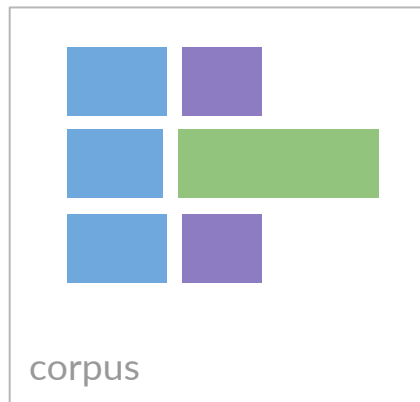


Count: 2



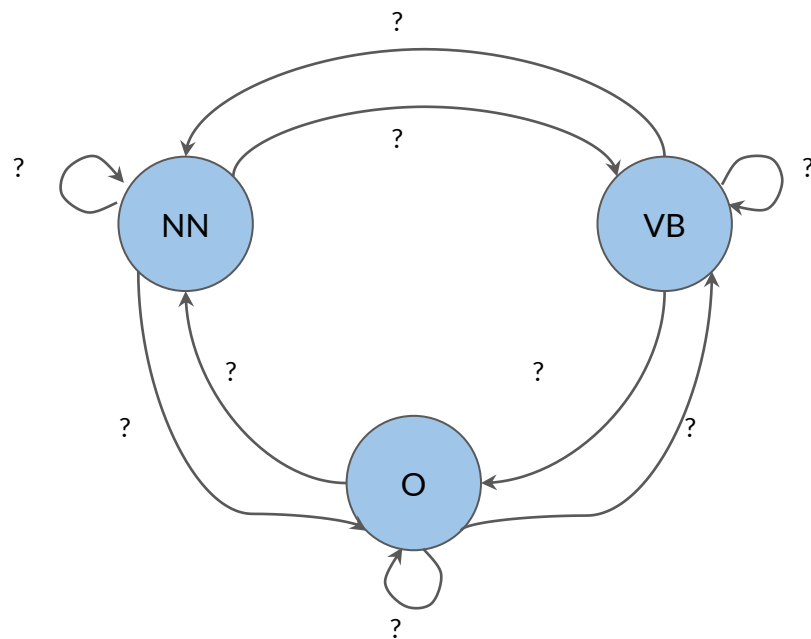
Count: 3

Transition probabilities



transition probability:  +  = $\frac{2}{3}$

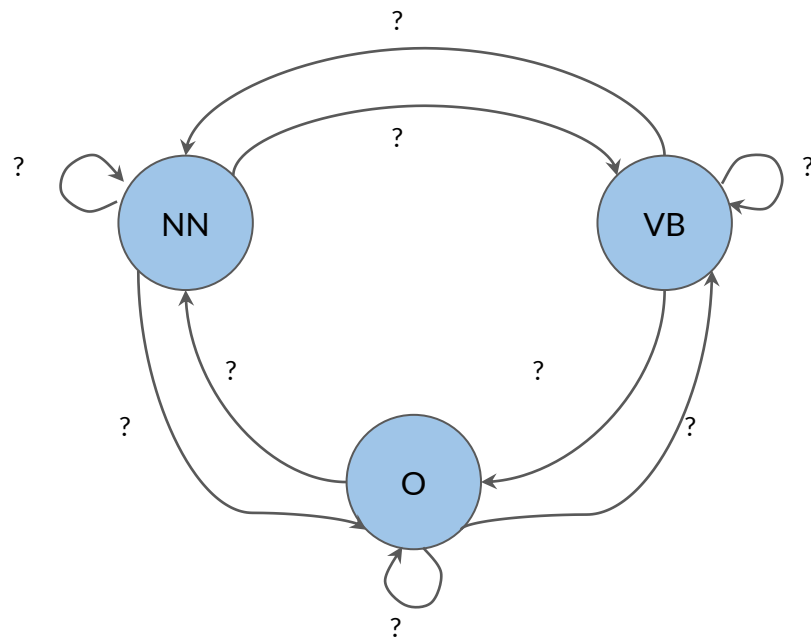
Transition probabilities



1. Count occurrences of tag pairs

$$C(t_{i-1}, t_i)$$

Transition probabilities



1. Count occurrences of tag pairs

$$C(t_{i-1}, t_i)$$

1. Calculate probabilities using the counts

$$P(t_i | t_{i-1}) = \frac{C(t_{i-1}, t_i)}{\sum_{j=1}^N C(t_{i-1}, t_j)}$$

The corpus

In a Station of the Metro
The apparition of these faces in the crowd :
Petals on a wet , black bough .

Ezra Pound –
1913

Preparation of the corpus

<s> In a Station of the Metro

<s> The apparition of these faces in the crowd
:

<s> Petals on a wet , black bough .

Ezra Pound –
1913

Preparation of the corpus

<s> in a station of the metro
<s> the apparition of these faces in the crowd
:
<s> petals on a wet , black bough .

Ezra Pound –
1913



deeplearning.ai

Populating the Transition Matrix

Populating the transition matrix

$A =$

	NN	VB	O
π			
NN (noun)			
VB (verb)			
O (other)			

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π			
NN (noun)			
VB (verb)			
O (other)			

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	$C(\pi, \text{NN})$		
NN (noun)	$C(\text{NN}, \text{NN})$		
VB (verb)	$C(\text{VB}, \text{NN})$		
O (other)	$C(\text{O}, \text{NN})$		

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1		
NN (noun)	$C(\text{NN}, \text{NN})$		
VB (verb)	$C(\text{VB}, \text{NN})$		
O (other)	$C(\text{O}, \text{NN})$		

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1		
NN (noun)	0		
VB (verb)	$C(\text{VB}, \text{NN})$		
O (other)	$C(\text{O}, \text{NN})$		

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1		
NN (noun)	0		
VB (verb)	0		
O (other)	$C(O, NN)$		

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1		
NN (noun)	0		
VB (verb)	0		
O (other)	6		

<s> in a station of the metro
<s> the apparition of these faces in the crowd
:
<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1		
NN (noun)	0		
VB (verb)	0		
O (other)	6		

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1	0	
NN (noun)	0	0	
VB (verb)	0	0	0
O (other)	6	0	

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1	0	2
NN (noun)	0	0	
VB (verb)	0	0	0
O (other)	6	0	

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1	0	2
NN (noun)	0	0	6
VB (verb)	0	0	0
O (other)	6	0	

<s> in a station of the metro

<s> the apparition of these faces in the crowd

...

<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1	0	2
NN (noun)	0	0	6
VB (verb)	0	0	0
O (other)	6	0	8

<s> in a station of the metro
<s> the apparition of these faces in the crowd
:
<s> petals on a wet , black bough .

Ezra Pound –
1913

Populating the transition matrix

$A =$

	NN	VB	O
π	1	0	2
NN	0	0	6
VB	0	0	0
O	6	0	8

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{\sum_{j=1}^N C(t_{i-1}, t_j)}$$

Populating the transition matrix

$A =$

	NN	VB	O	
π	1	0	2	3
NN	0	0	6	6
VB	0	0	0	0
O	6	0	8	14

$$P(\text{NN}|\pi) = \frac{C(\pi, \text{NN})}{\sum_{j=1}^N C(\pi, t_j)} = \frac{1}{3}$$

Populating the transition matrix

$A =$

	NN	VB	O	
π	1	0	2	3
NN	0	0	6	6
VB	0	0	0	0
O	6	0	8	14

$$P(\text{NN}|\text{O}) = \frac{C(\text{O}, \text{NN})}{\sum_{j=1}^N C(\text{O}, t_j)} = \frac{6}{14}$$

Populating the transition matrix

$A =$

	NN	VB	O	
π	1	0	2	3
NN	0	0	6	6
VB	0	0	0	0
O	6	0	8	14

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{\sum_{j=1}^N C(t_{i-1}, t_j)}$$

Smoothing

$A =$

	NN	VB	O	
π	$1+\epsilon$	$0+\epsilon$	$2+\epsilon$	$3+3^*\epsilon$
NN	$0+\epsilon$	$0+\epsilon$	$6+\epsilon$	$6+3^*\epsilon$
VB	$0+\epsilon$	$0+\epsilon$	$0+\epsilon$	$0+3^*\epsilon$
O	$6+\epsilon$	$0+\epsilon$	$8+\epsilon$	$14+3^*\epsilon$

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i) + \boxed{\epsilon}}{\sum_{j=1}^N C(t_{i-1}, t_j) + \boxed{N} * \boxed{\epsilon}}$$

Smoothing

$A =$

	NN	VB	O
π	0.3333	0.0003	0.6663
NN	0.0001	0.0001	0.9996
VB	0.3333	0.3333	0.3333
O	0.4285	0.0000	0.5713

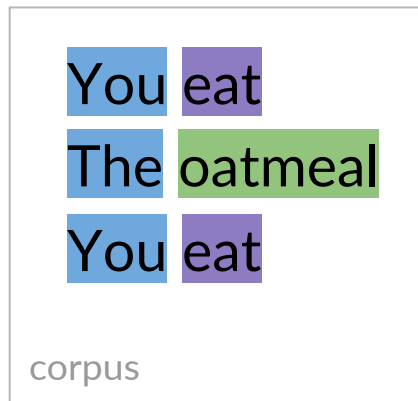
$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i) + \epsilon}{\sum_{j=1}^N C(t_{i-1}, t_j) + N * \epsilon}$$



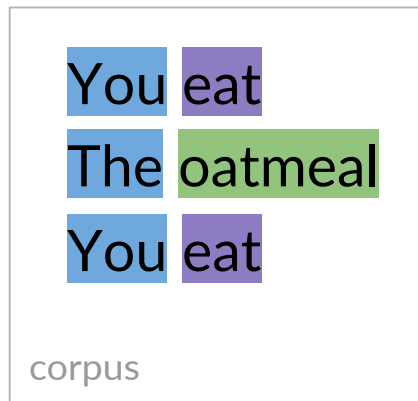
deeplearning.ai

Populating the Emission Matrix

Emission probabilities



Transition probabilities

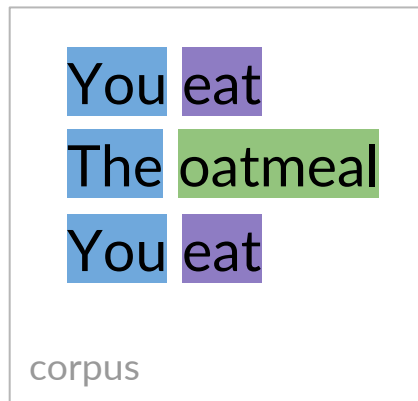


You

Count: 2

Count: 3

Transition probabilities



emission probability: You = $\frac{2}{3}$

The emission matrix

$B =$

	in	a	...
NN (noun)			
VB (verb)			
O (other)			

<s> in a station of the metro

<s> the apparition of these faces in the crowd
:

<s> petals on a wet , black bough .

Ezra Pound –
1913

The emission matrix

$B =$

	in	a	...
NN (noun)	$C(\text{NN}, \text{in})$		
VB (verb)	$C(\text{VB}, \text{in})$		
O (other)	$C(\text{O}, \text{in})$		

<s> in a station of the metro

<s> the apparition of these faces in the crowd
:

<s> petals on a wet , black bough .

Ezra Pound –
1913

The emission matrix

$B =$

	in	a	...
NN (noun)	0		
VB (verb)	$C(\text{VB}, \text{in})$		
O (other)	$C(\text{O}, \text{in})$		

<s> in a station of the metro

<s> the apparition of these faces in the crowd

:

<s> petals on a wet , black bough .

Ezra Pound –
1913

The emission matrix

$B =$

	in	a	...
NN (noun)	0		
VB (verb)	0		
O (other)	$C(O, in)$		

<s> in a station of the metro

<s> the apparition of these faces in the crowd
:

<s> petals on a wet , black bough .

Ezra Pound –
1913

The emission matrix

$B =$

	in	a	...
NN (noun)	0		
VB (verb)	0		
O (other)	2		

<s> in a station of the metro

<s> the apparition of these faces in the crowd
:

<s> petals on a wet , black bough .

Ezra Pound –
1913

The emission matrix

$B =$

	in	a	...
NN (noun)	0
VB (verb)	0
O (other)	2

$$\begin{aligned} P(w_i|t_i) &= \frac{C(t_i, w_i) + \epsilon}{\sum_{j=1}^V C(t_i, w_j) + N * \epsilon} \\ &= \frac{C(t_i, w_i) + \epsilon}{C(t_i) + N * \epsilon} \end{aligned}$$

Summary

1. Calculate transition and emission matrix
1. How to apply smoothing



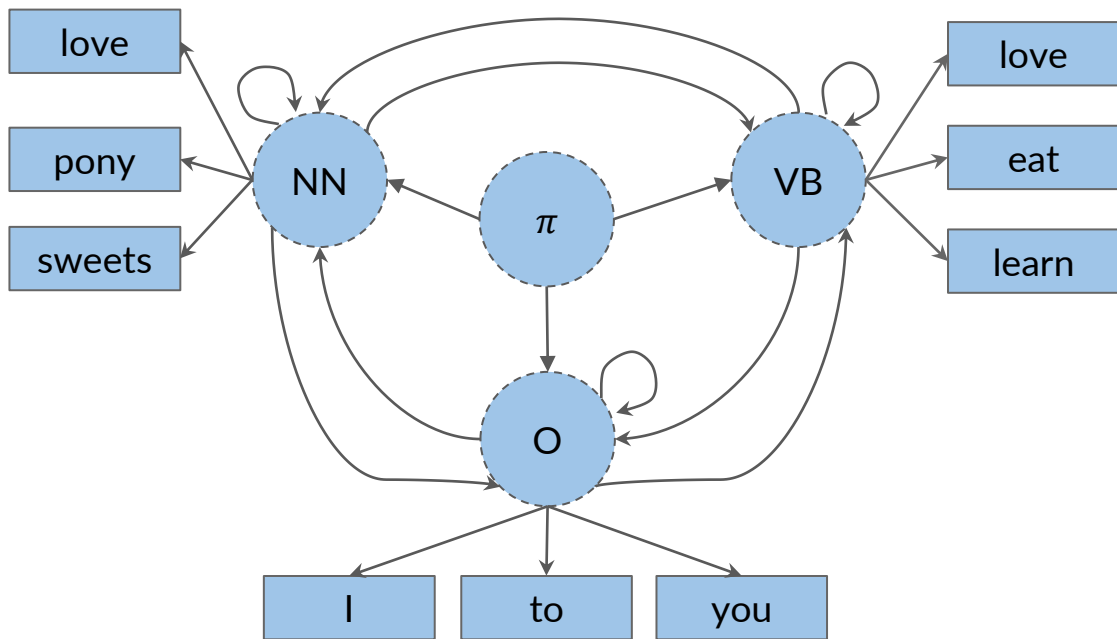
deeplearning.ai

The Viterbi Algorithm

Why not learn something ?

? ? ? ? ?

Viterbi algorithm – a graph algorithm

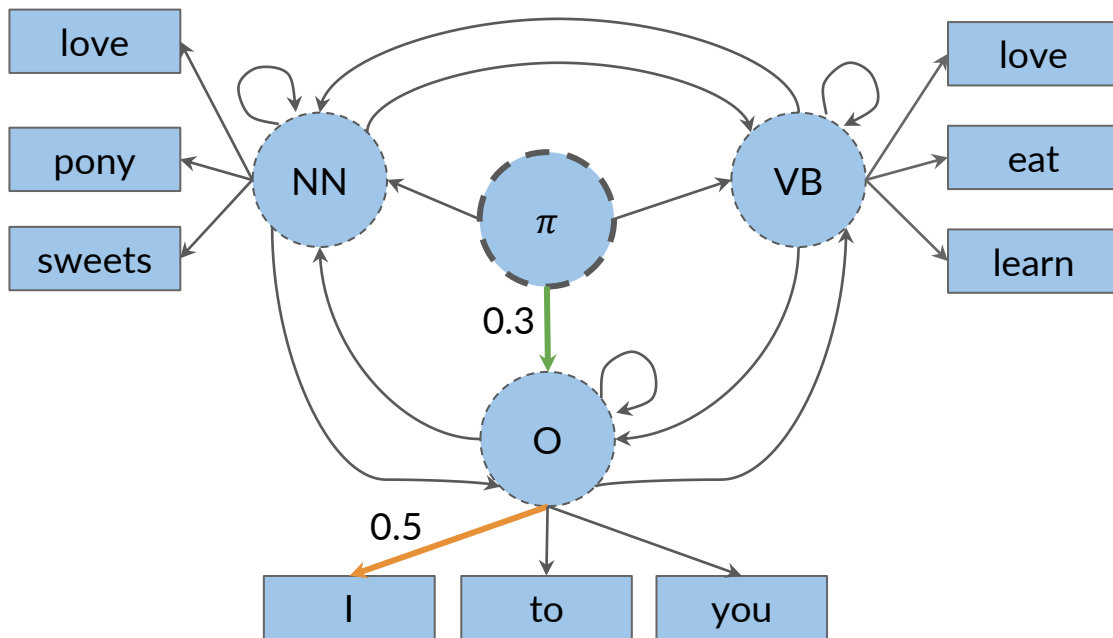


<s>

I
to

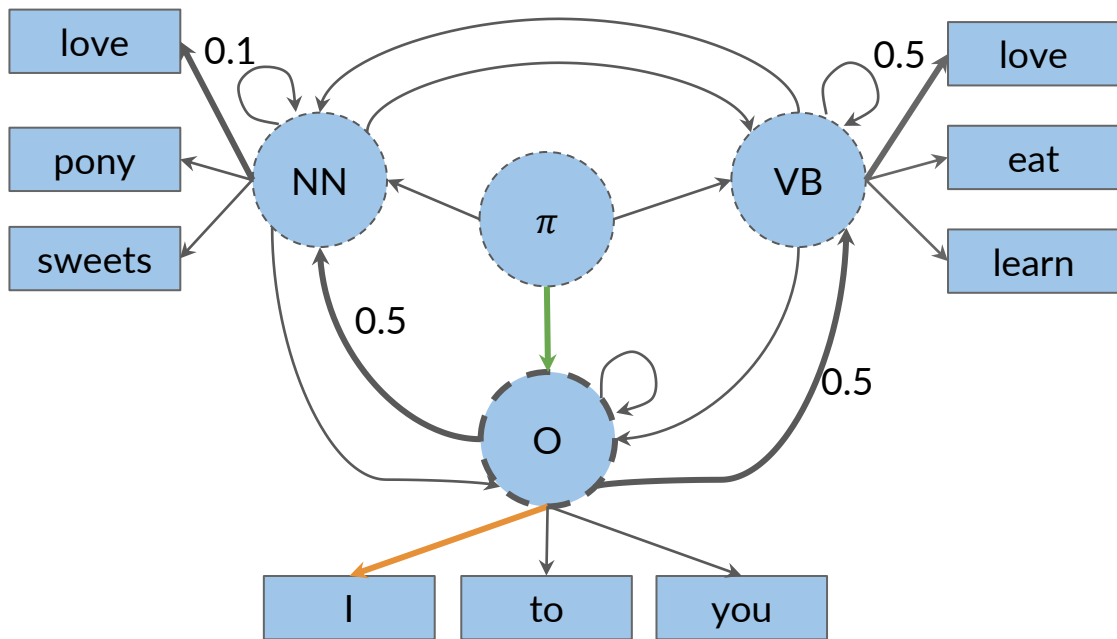
love
learn

Viterbi algorithm – a graph algorithm



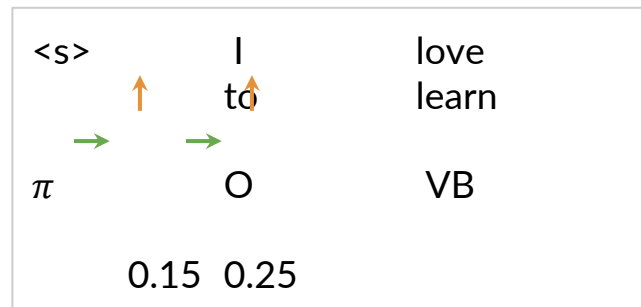
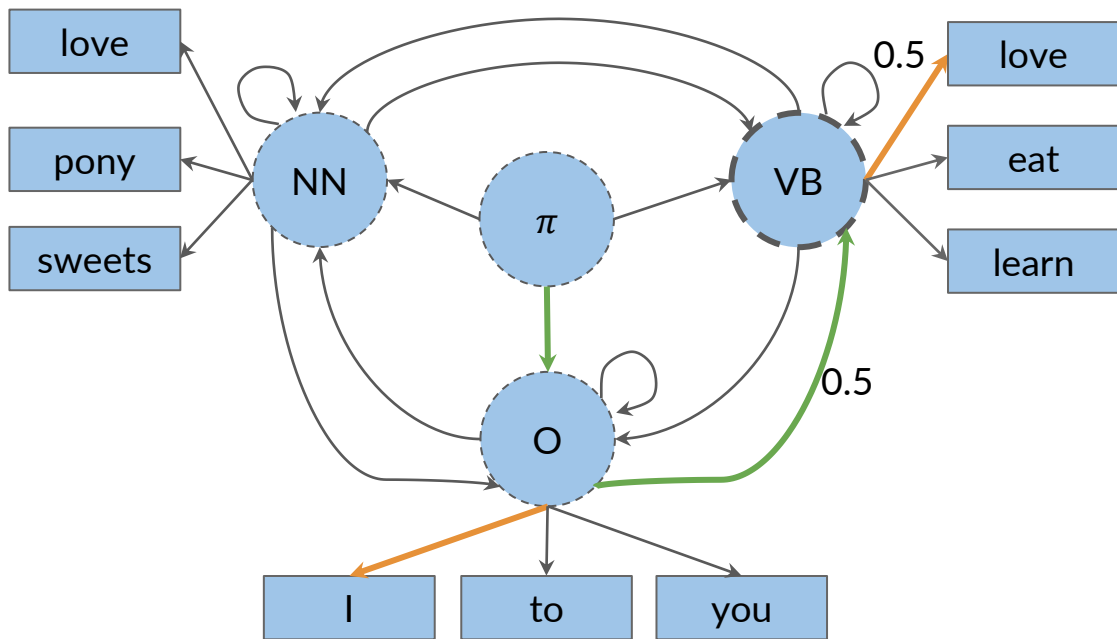
<s>	I	love
	to	learn
π	O	
	0.15	

Viterbi algorithm – a graph algorithm

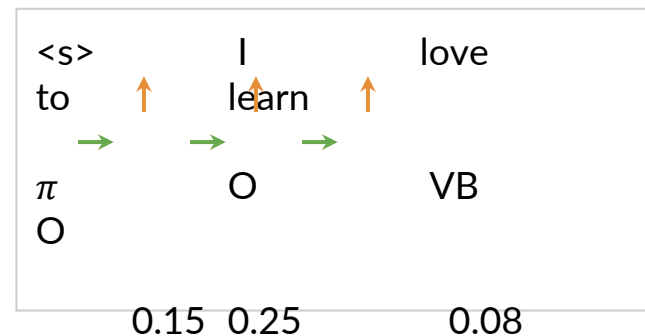
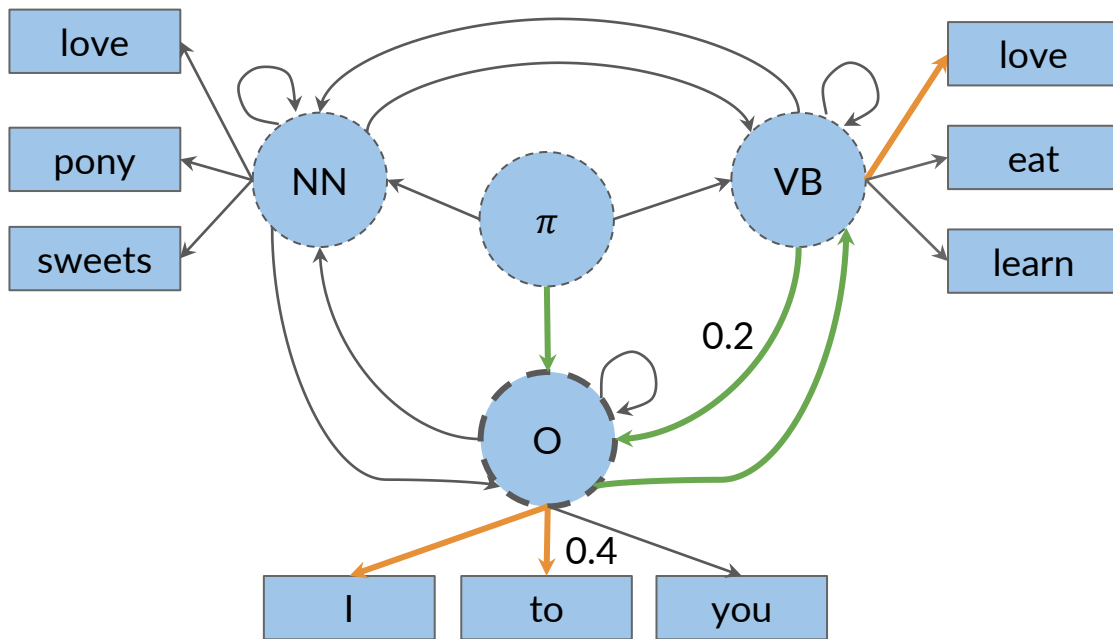


<s>	I	love
	to	learn
π	O	
	0.15	

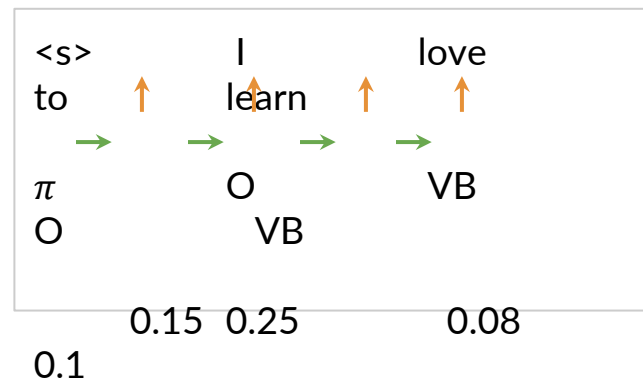
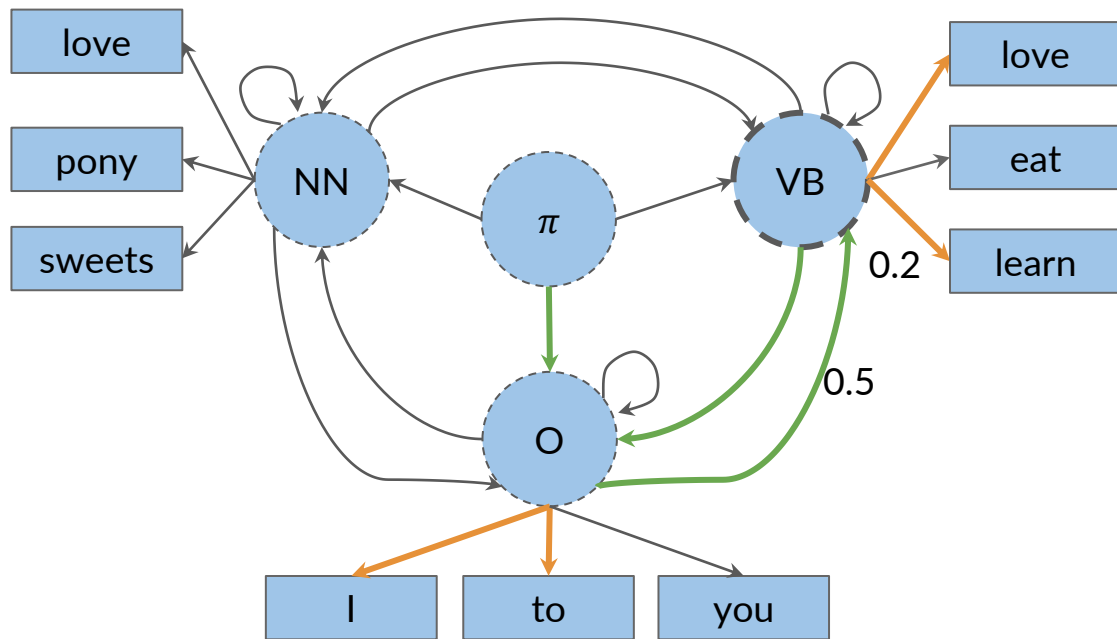
Viterbi algorithm – a graph algorithm



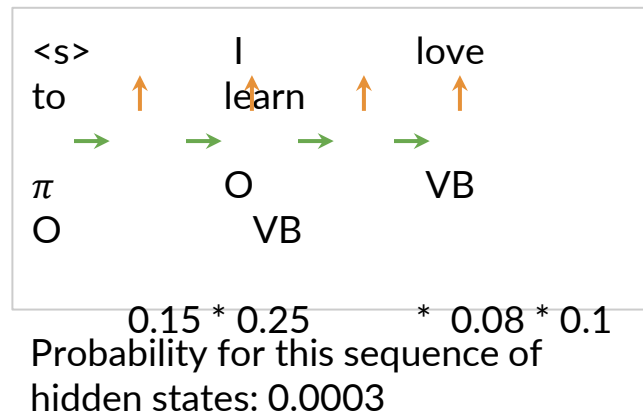
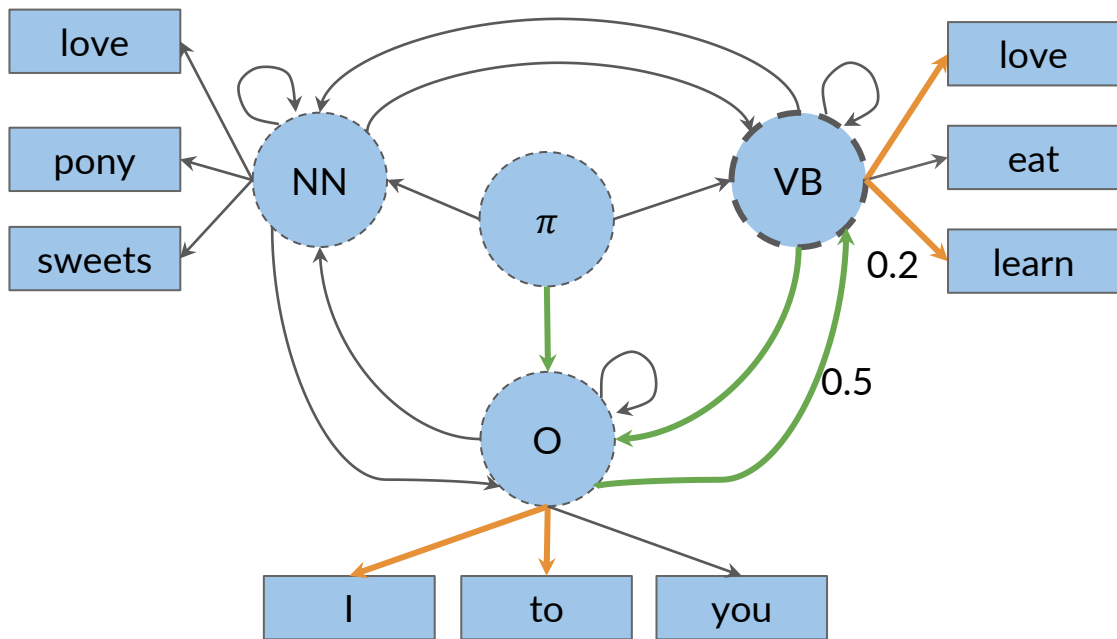
Viterbi algorithm – a graph algorithm



Viterbi algorithm – a graph algorithm



Viterbi algorithm – a graph algorithm



Viterbi algorithm – Steps

1. Initialization step
2. Forward pass
3. Backward pass

$C =$

	w_1	w_2	...	w_K
t_1				
...				
t_N				

$D =$

	w_1	w_2	...	w_K
t_1				
...				
t_N				



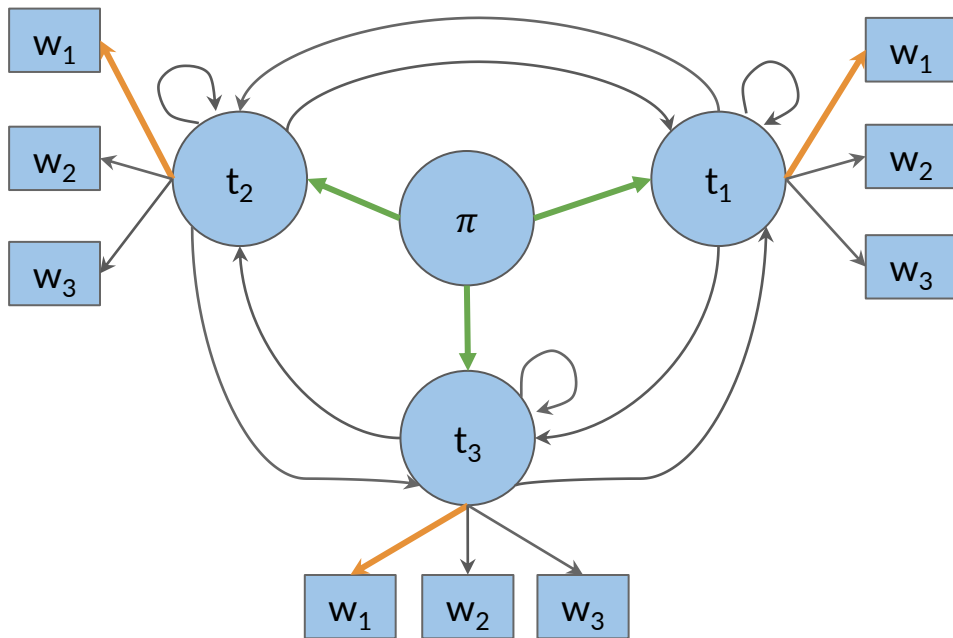
deeplearning.ai

Viterbi: Initialization

Viterbi algorithm – Steps

1. Initialization step

Initialization step



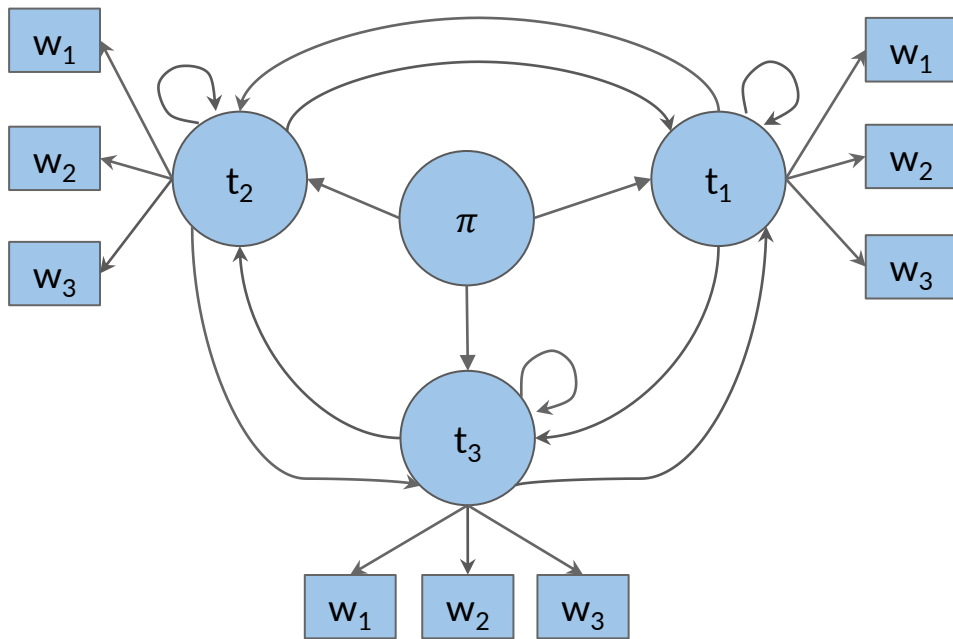
$C =$

	w_1	w_2	...	w_K
t_1	$c_{1,1}$			
...				
t_N	$c_{N,1}$			

$$c_{i,1} = \pi_i * b_{i, \text{index}(w_1)}$$

$$= a_{1,i} * b_{i, \text{index}(w_1)}$$

Initialization step



$D =$

	w_1	w_2	...	w_K
t_1	$d_{1,1}$			
...				
t_N	$d_{N,1}$			

$$d_{i,1} = 0$$



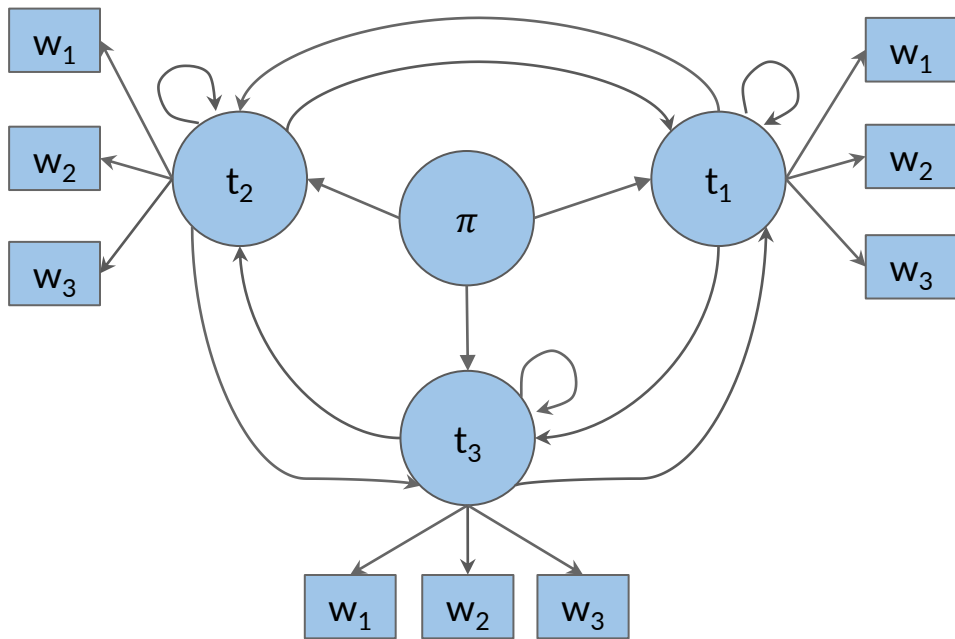
deeplearning.ai

Viterbi: Forward Pass

Viterbi algorithm – Steps

2. Forward pass

Forward pass

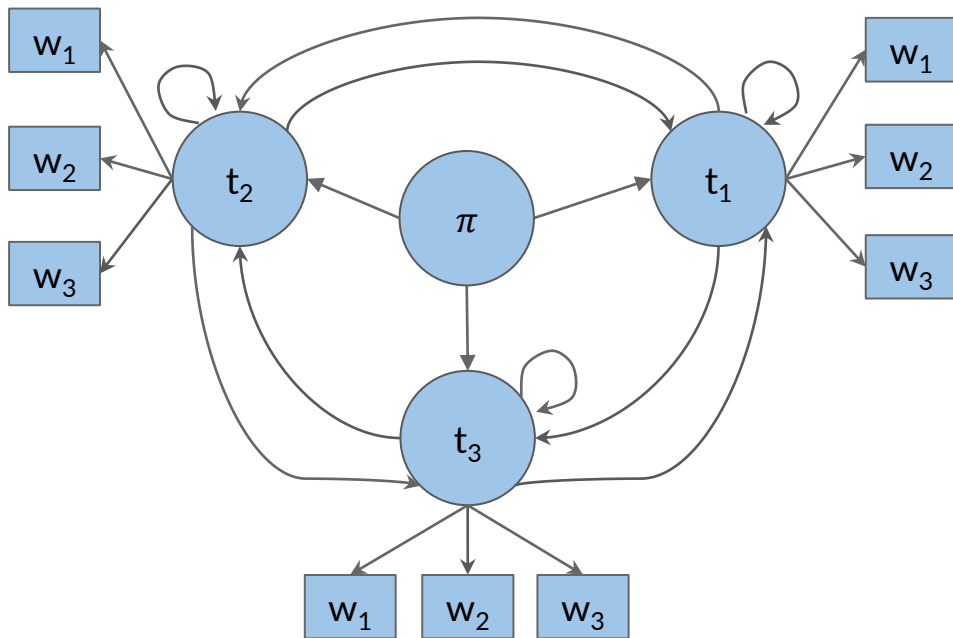


$C =$

	w_1	w_2	...	w_K
t_1	$c_{1,1}$	$c_{1,2}$		$c_{1,K}$
...				
t_N	$c_{N,1}$	$c_{N,2}$		$c_{N,K}$

$$c_{i,j} = \max_k c_{k,j-1} * a_{k,i} * b_{i, \text{index}(w_j)}$$

Forward pass

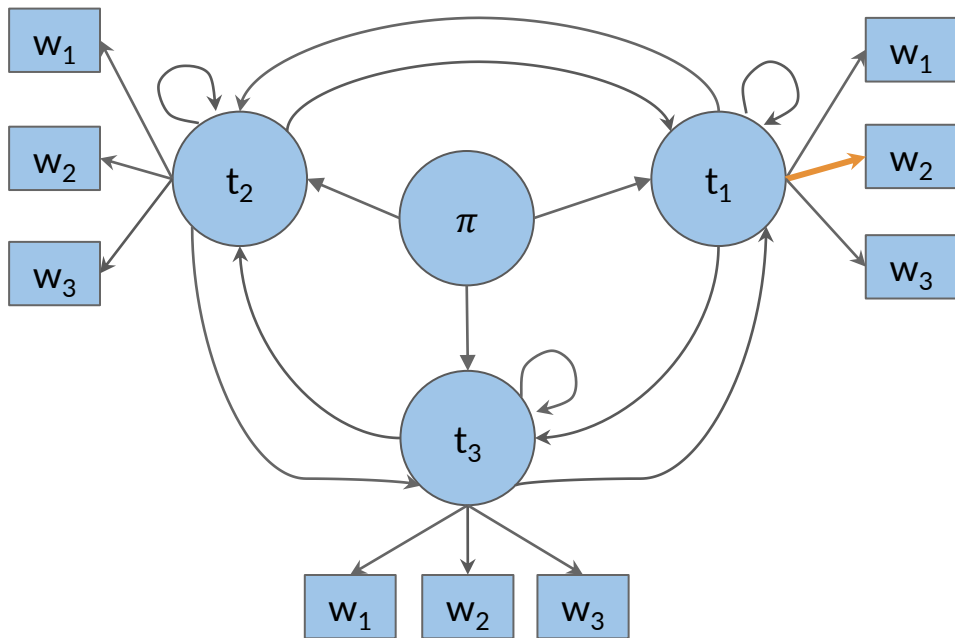


$C =$

	w_1	w_2	...	w_K
t_1	$c_{1,1}$	$c_{1,2}$		$c_{1,K}$
...				
t_N	$c_{N,1}$	$c_{N,2}$		$c_{N,K}$

$$c_{1,2} = \max_k c_{k,1} * a_{k,1} * b_{1, \text{index}(w_2)}$$

Forward pass

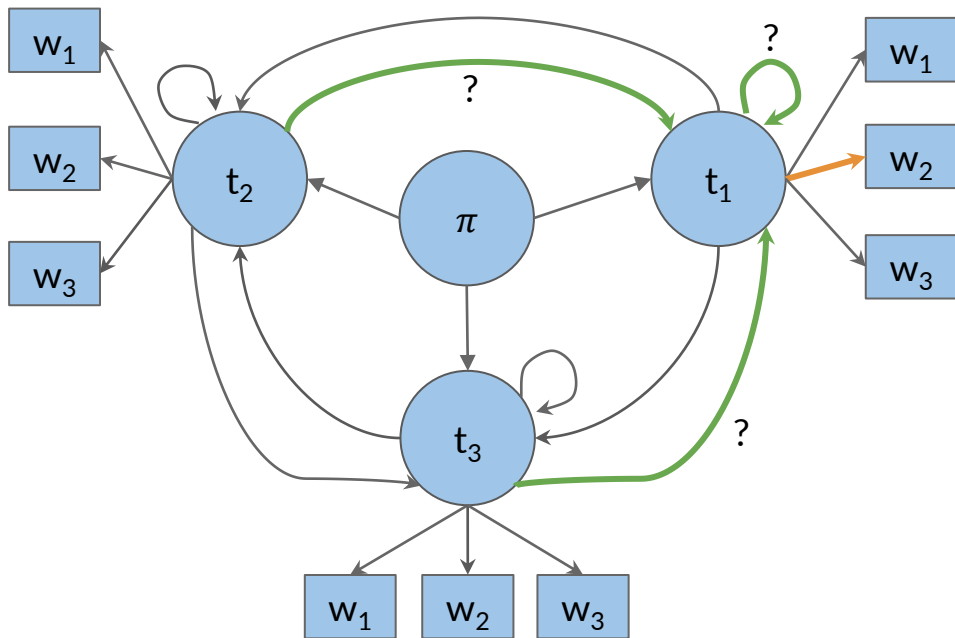


$C =$

	w_1	w_2	...	w_K
t_1	$c_{1,1}$	$c_{1,2}$		$c_{1,K}$
...				
t_N	$c_{N,1}$	$c_{N,2}$		$c_{N,K}$

$$c_{1,2} = \max_k c_{k,1} * a_{k,1} * b_{1, \text{index}(w_2)}$$

Forward pass

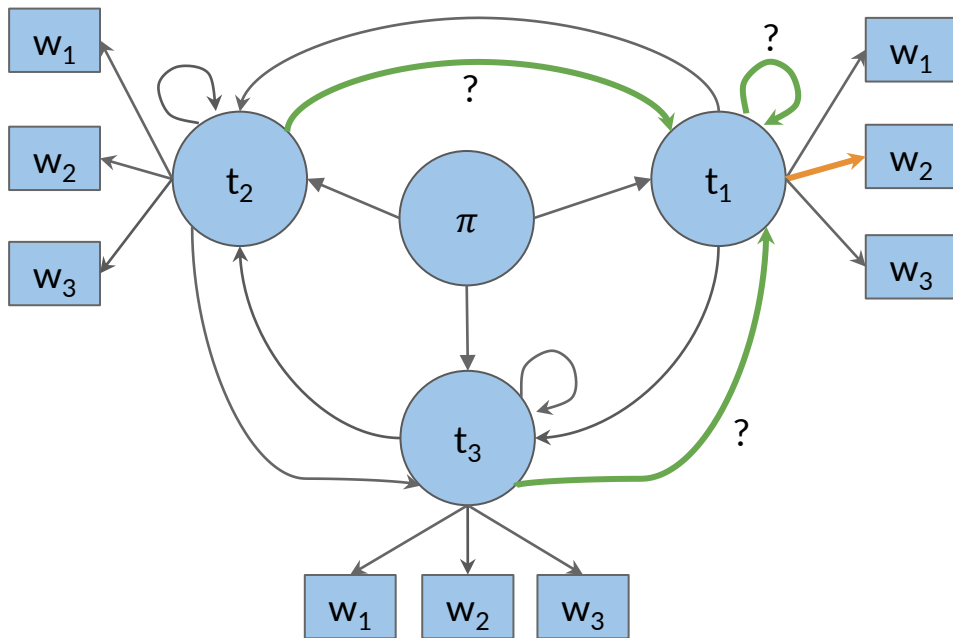


$C =$

	w_1	w_2	...	w_K
t_1	$c_{1,1}$	$c_{1,2}$		$c_{1,K}$
...				
t_N	$c_{N,1}$	$c_{N,2}$		$c_{N,K}$

$$c_{1,2} = \max_k c_{k,1} * a_{k,1} * b_{1, \text{index}(w_2)}$$

Forward pass

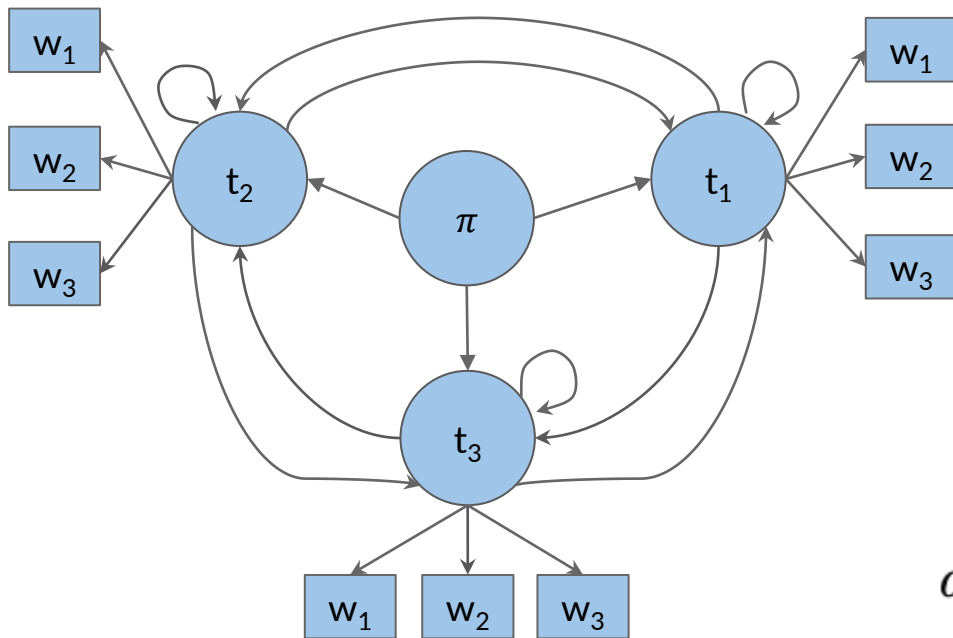


$C =$

	w_1	w_2	...	w_K
t_1	$c_{1,1}$	$c_{1,2}$		$c_{1,K}$
...				
t_N	$c_{N,1}$	$c_{N,2}$		$c_{N,K}$

$$c_{1,2} = \max_k c_{k,1} * a_{k,1} * b_{1, \text{index}(w_2)}$$

Forward pass



$D =$

	w_1	w_2	...	w_K
t_1	$d_{1,1}$	$d_{1,2}$		$d_{1,K}$
...				
t_N	$d_{N,1}$	$d_{N,2}$		$d_{N,K}$

$$c_{i,j} = \max_k c_{k,j-1} * a_{k,i} * b_{i, \text{index}(w_j)}$$

$$d_{i,j} = \operatorname{argmax}_k c_{k,j-1} * a_{k,i} * b_{i, \text{index}(w_j)}$$



deeplearning.ai

Viterbi: Backward Pass

Viterbi algorithm – Steps

3. Backward pass

Backward pass

$C =$

	w_1	w_2	...	w_K
t_1	$c_{1,1}$	$c_{1,2}$		$c_{1,K}$
...				
t_N	$c_{N,1}$	$c_{N,2}$		$c_{N,K}$

$$s = \operatorname{argmax}_i c_{i,K}$$

$D =$

	w_1	w_2	...	w_K
t_1	$d_{1,1}$	$d_{1,2}$		$d_{1,K}$
...				
t_N	$d_{N,1}$	$d_{N,2}$		$d_{N,K}$

Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1

$\langle s \rangle$	w_1	w_2
	w_3	w_4
	w_5	

Backward pass

$C =$

	w_1	w_2	w_3	w_4	w_5
t_1	0.25	0.125	0.025	0.0125	0.01
t_2	0.1	0.025	0.05	0.01	0.003
t_3	0.3	0.05	0.025	0.02	0.0000
t_4	0.2	0.1	0.000	0.0025	0.0003

$$s = \underset{i}{\operatorname{argmax}} c_{i,K} = 1$$

Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1

$s = \operatorname{argmax}_i c_{i,K} = 1$

$\langle s \rangle$	w_1	w_2
	w_3	w_4
	w_5	

Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1

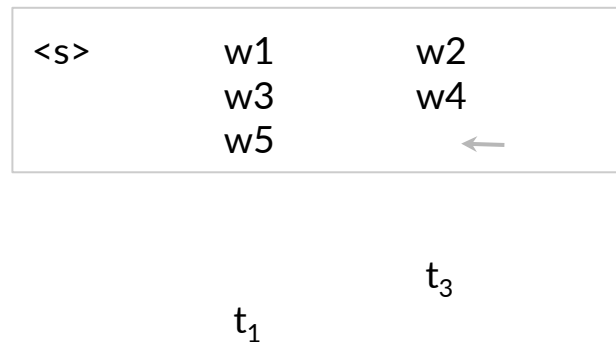
$\langle s \rangle$	w_1	w_2
	w_3	w_4
	w_5	

t_1

Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1



Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1

<s>

w_1
 w_3
 w_5

w_2
 w_4



t_1
 t_1

t_3

Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1

<s>

w_1
 w_3
 w_5

w_2
 w_4



t_1
 t_1

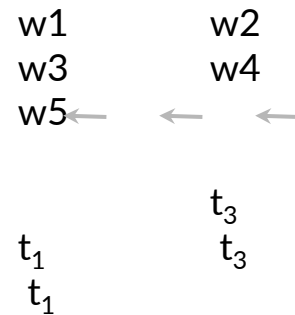
t_3

Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1

<s>

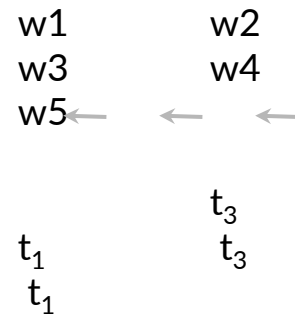


Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1

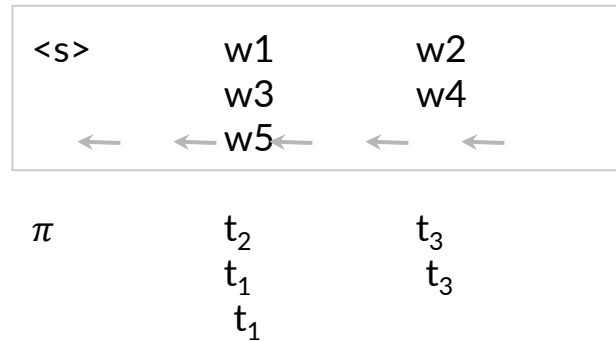
<s>



Backward pass

$D =$

	w_1	w_2	w_3	w_4	w_5
t_1	0	1	3	2	3
t_2	0	2	4	1	3
t_3	0	2	4	1	4
t_4	0	4	4	3	1



Implementation notes

1. In Python index starts with 0!
2. Use log probabilities

$$c_{i,j} = \max_k c_{k,j-1} * a_{k,i} * b_{i, \text{index}(w_j)}$$



$$\log(c_{i,j}) = \max_k \log(c_{k,j-1}) + \log(a_{k,i}) + \log(b_{i, \text{index}(w_j)})$$

Summary

1. From word sequence to POS tag sequence
2. Viterbi algorithm
3. Log probabilities