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:

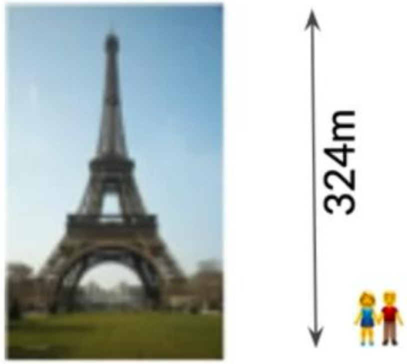
lexical term	**tag**	**example**
noun	NN	something, nothing
verb	VB	learn, study
determiner	DT	the, a
w-adverb	WRB	why, where
...	...	

Why not learn something ?

WRB RB VB NN .

Applications of POS tagging

Named entities



Co-reference resolution

Speech recognition

Example

Why not learn ...

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

Part of Speech Dependencies

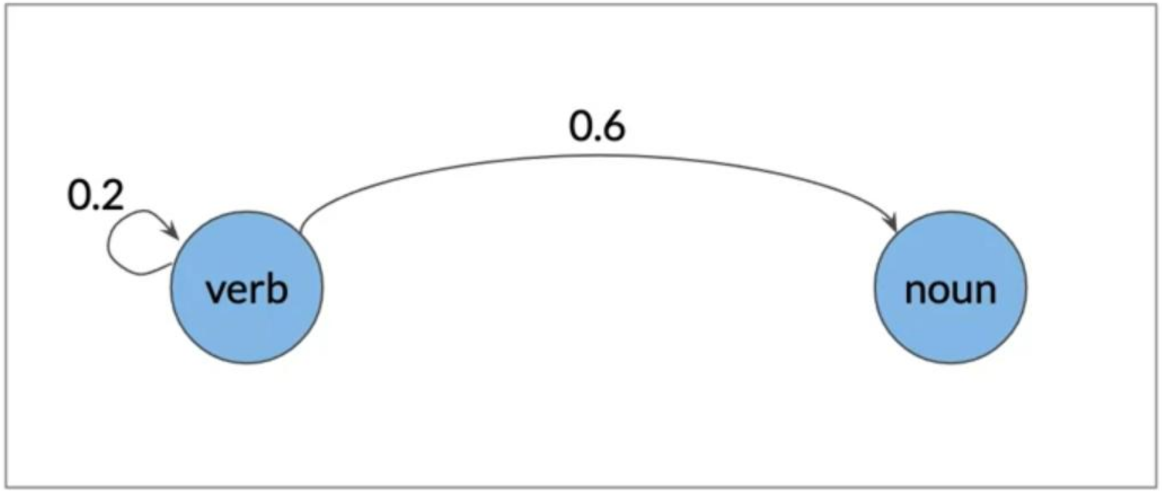
Why not learn ...

verb verb?

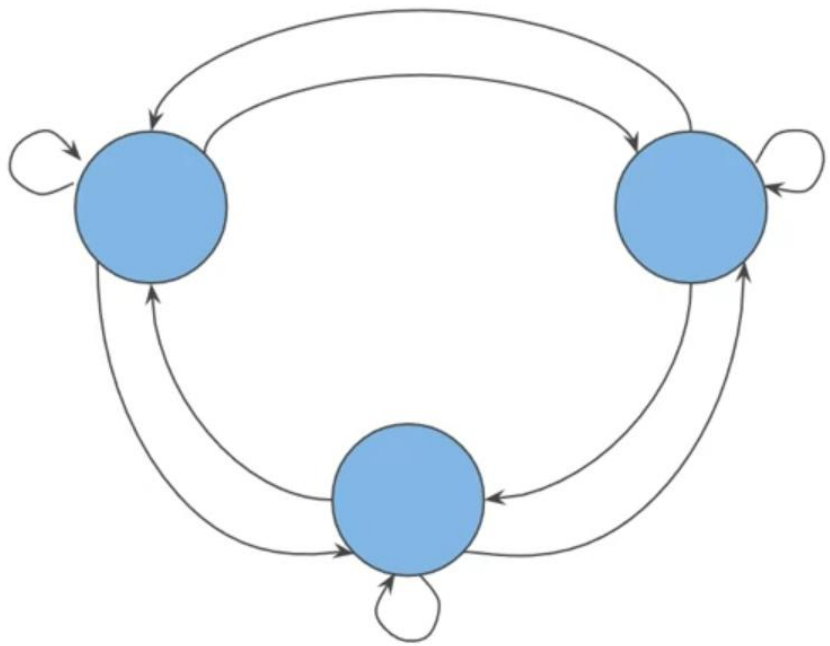
noun?

...?

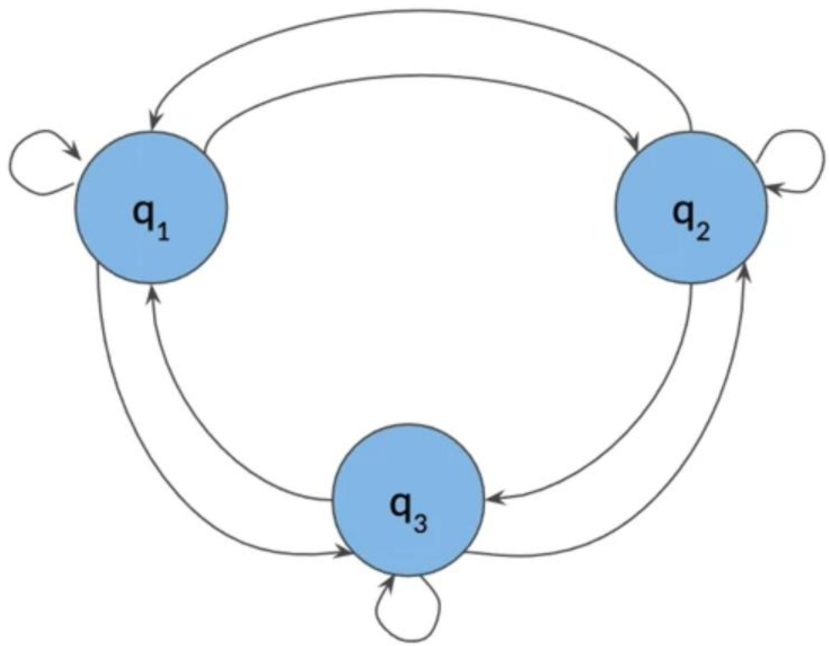
Visual Representation



What are Markov chains?

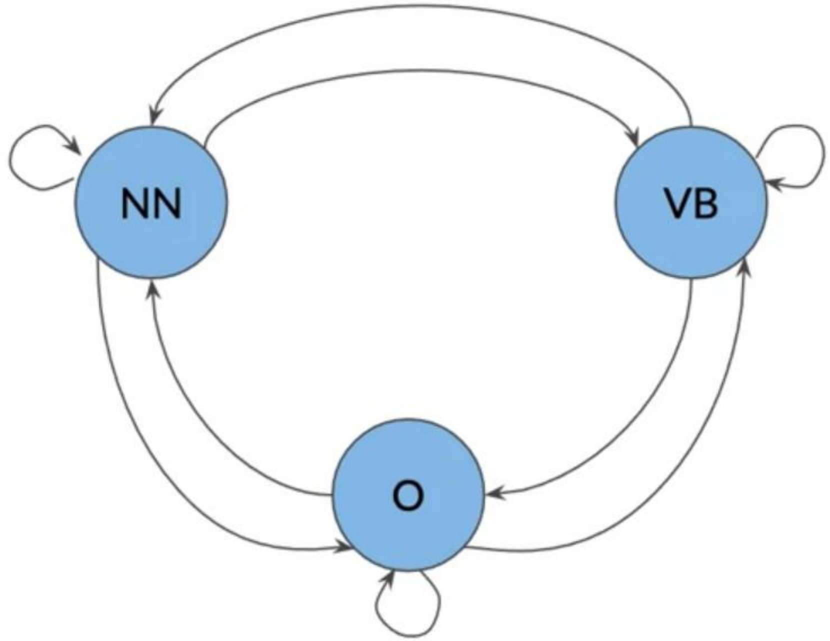


States

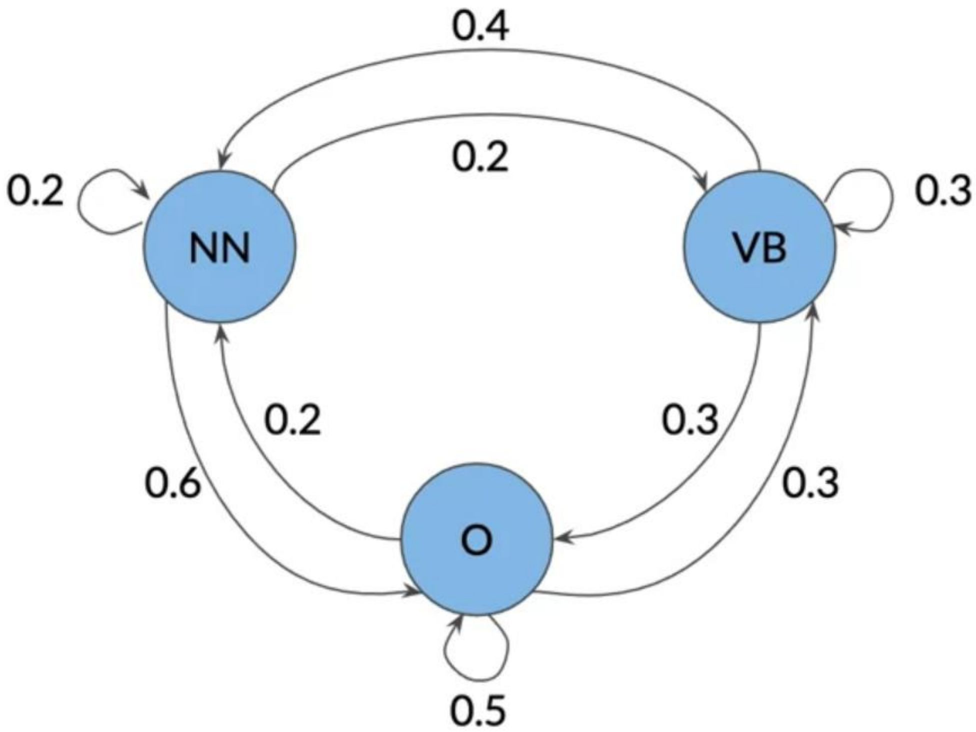


$$Q = \{q_1, q_2, q_3\}$$

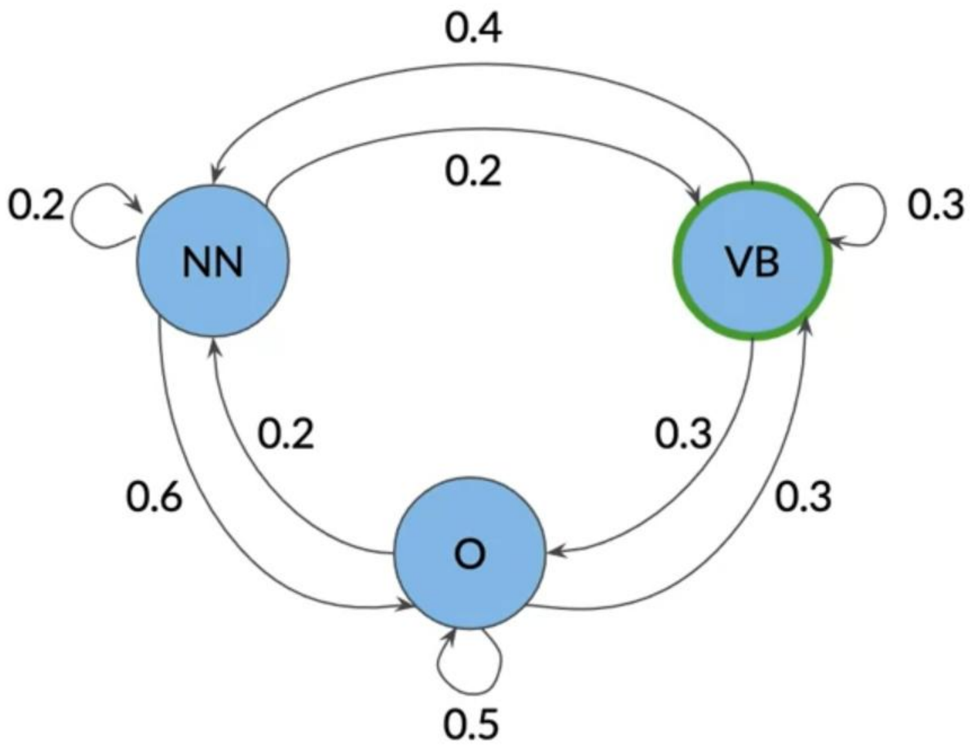
POS tags as States



Transition probabilities

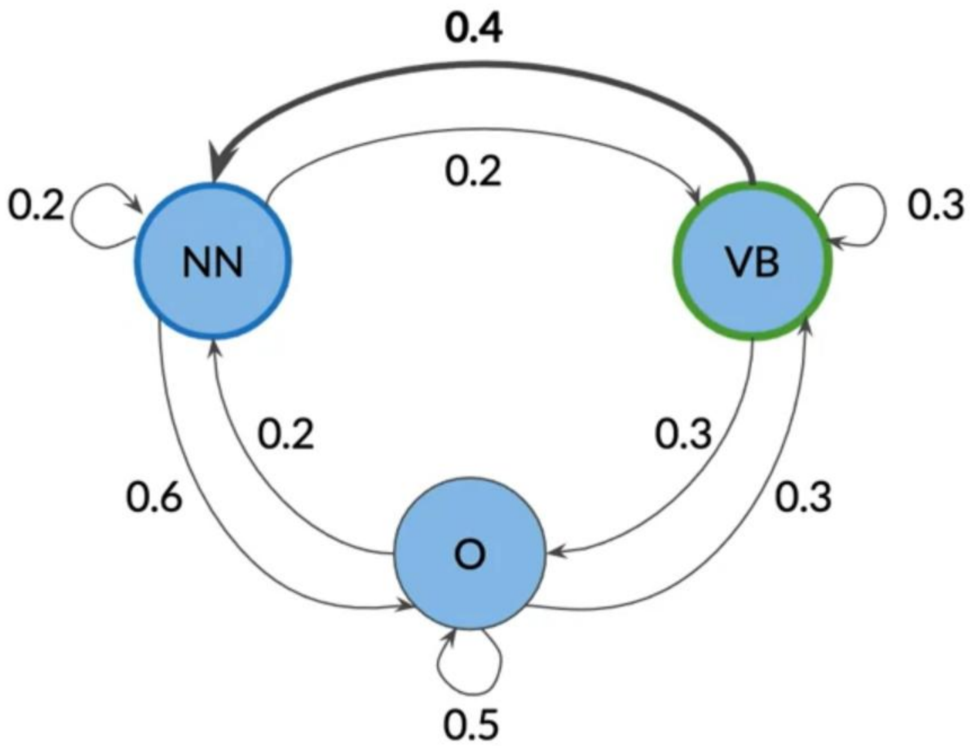


Transition probabilities



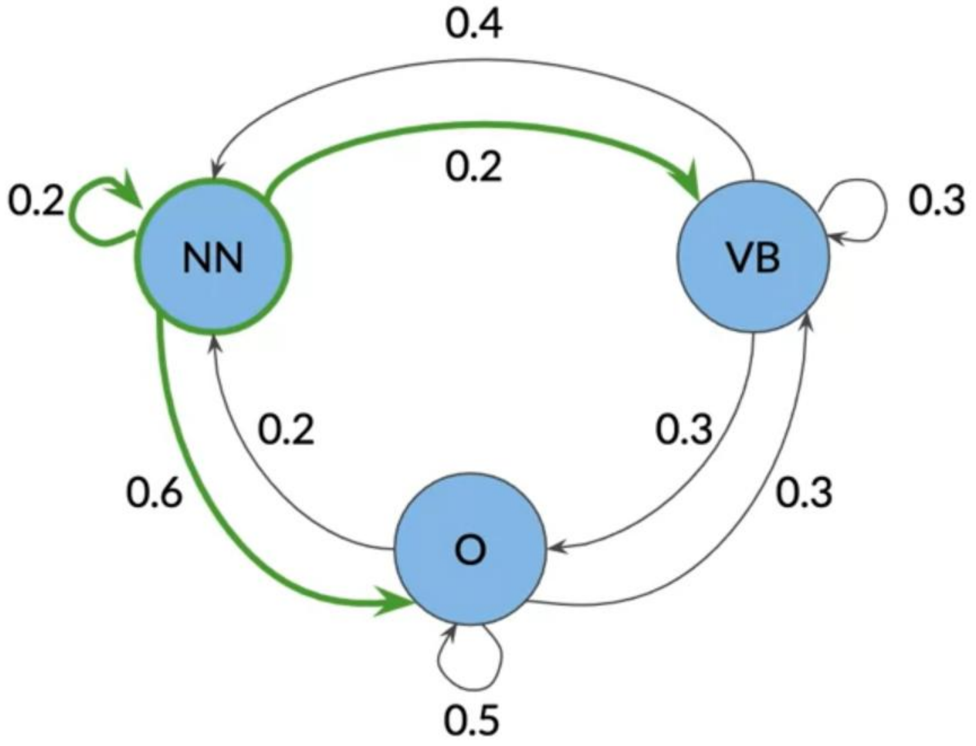
Why not learn something ?

Transition probabilities



Why not learn something ?

The transition matrix

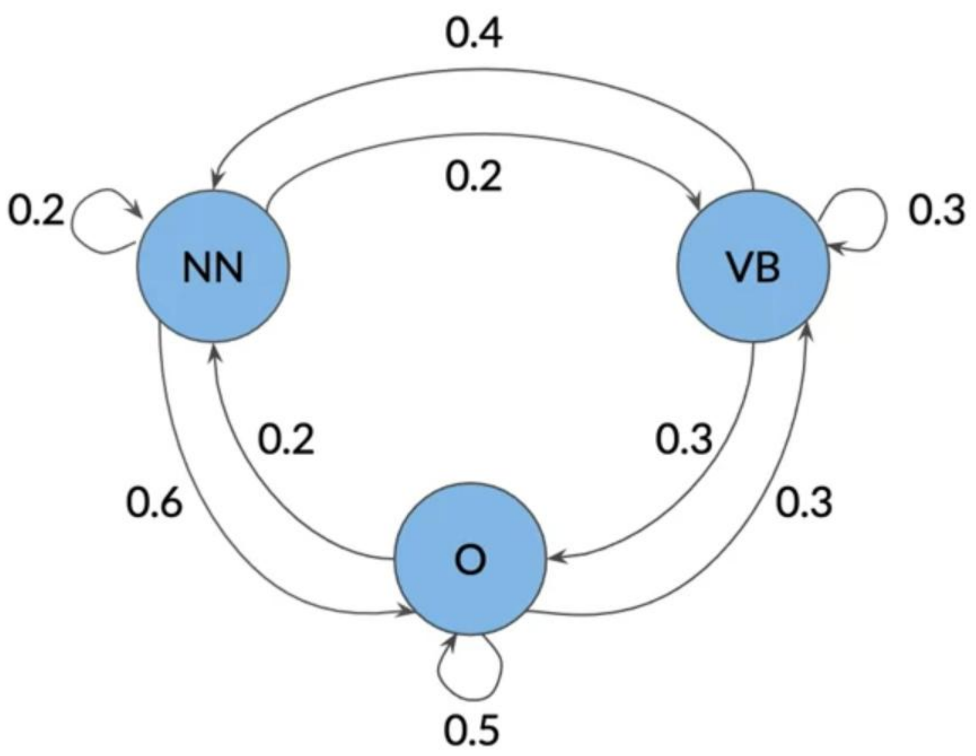


$A =$

	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

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

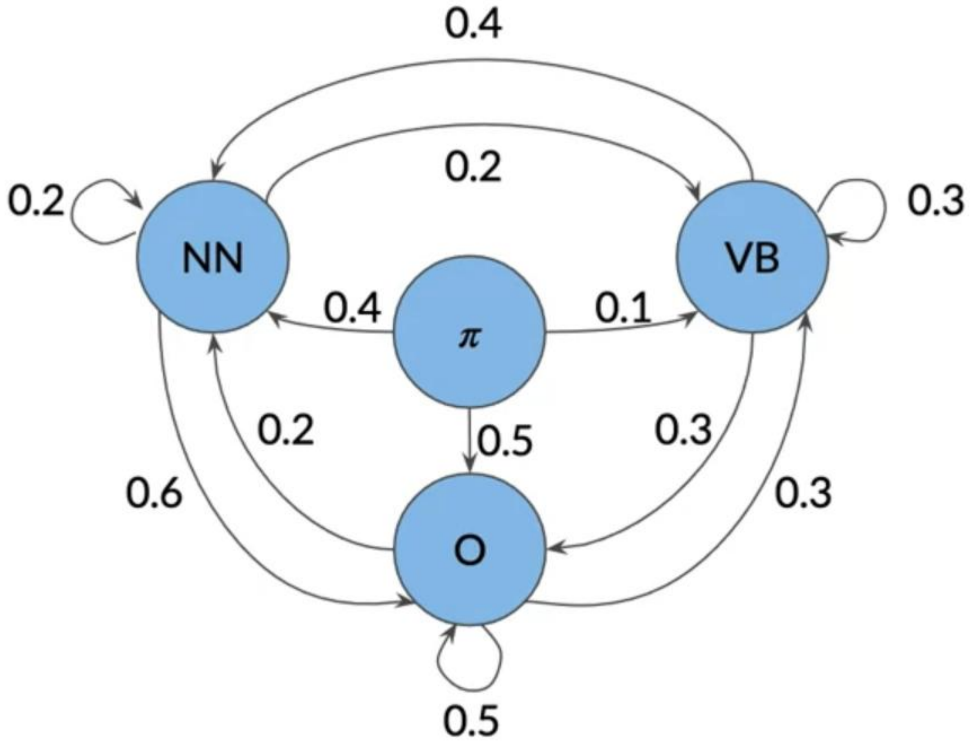
The first word



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

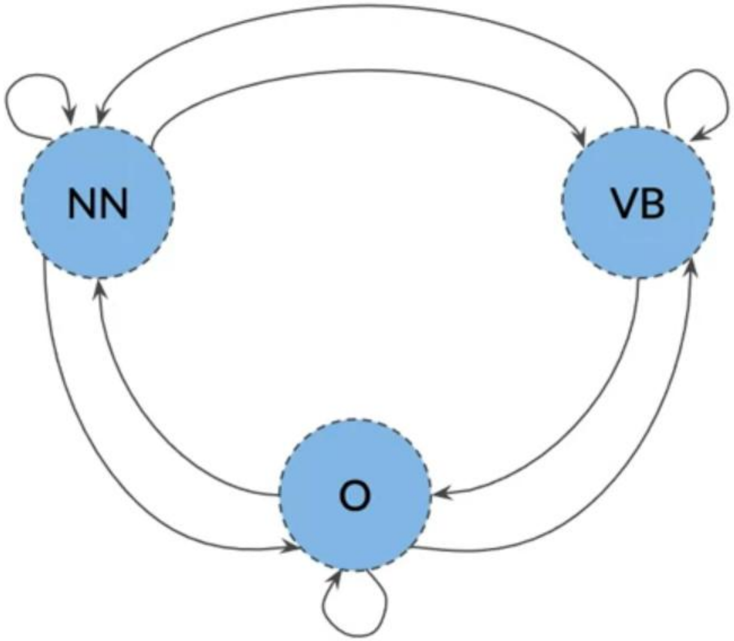
States

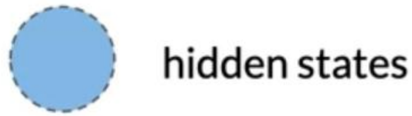
$$Q = \{q_1, \dots, q_N\}$$

Transition matrix

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

Hidden Markov Model





you	machine
😮	👾
jump = verb	jump = ?

you

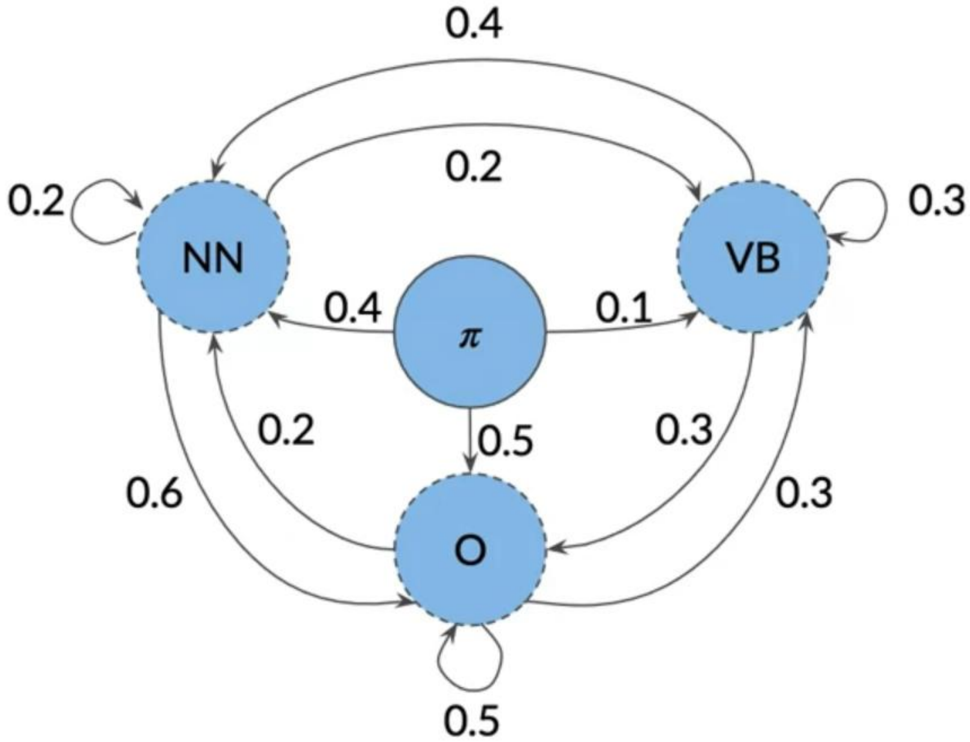
jump = verb
run = verb
fly = verb

machine

jump
run
fly

*observable

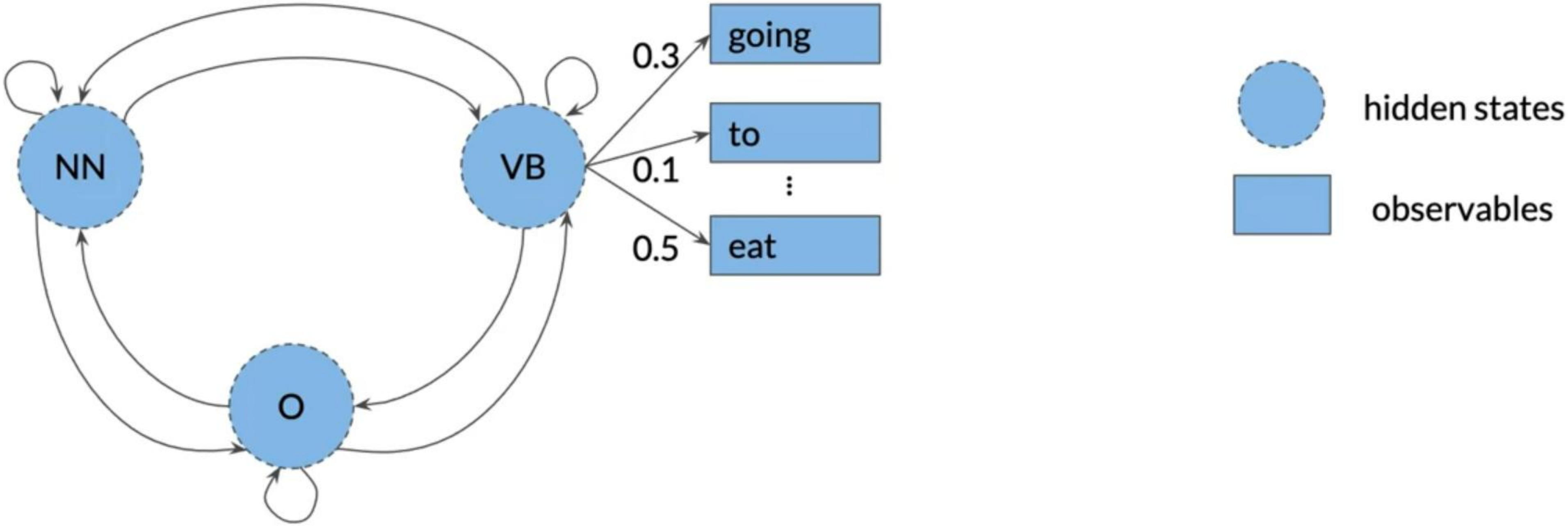
Transition probabilities



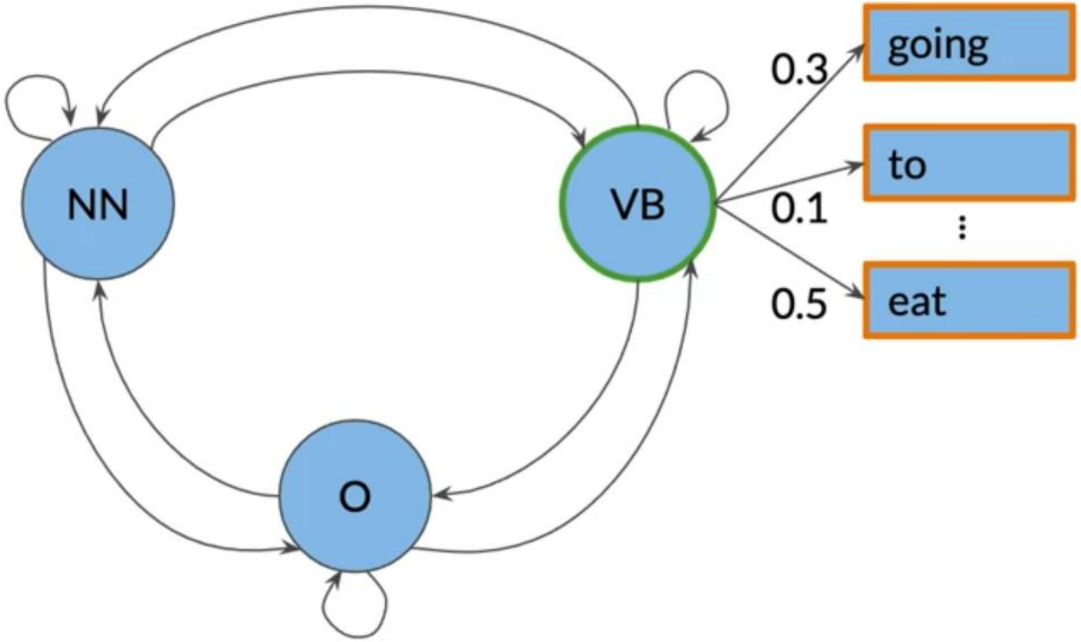
$A =$

	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	

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

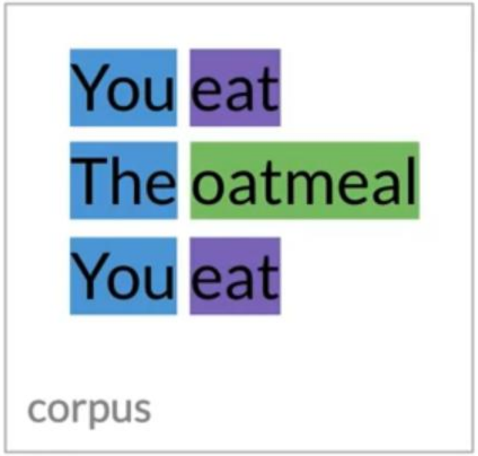
States

Transition matrix

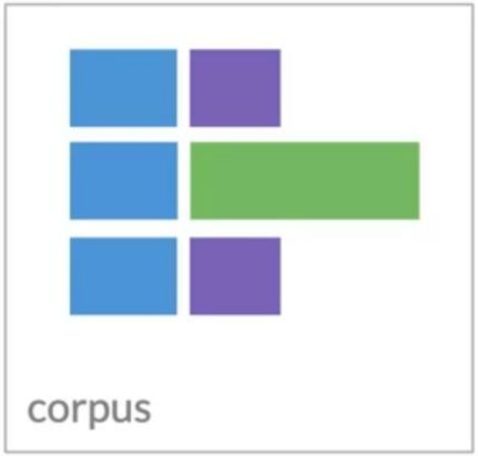
Emission matrix

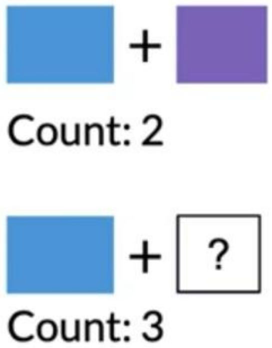
$$Q = \{q_1, \ldots, q_N\} \quad A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{N+1,1} & \cdots & a_{N+1,N} \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & \cdots & b_{1V} \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{NV} \end{pmatrix}$$

Transition probabilities

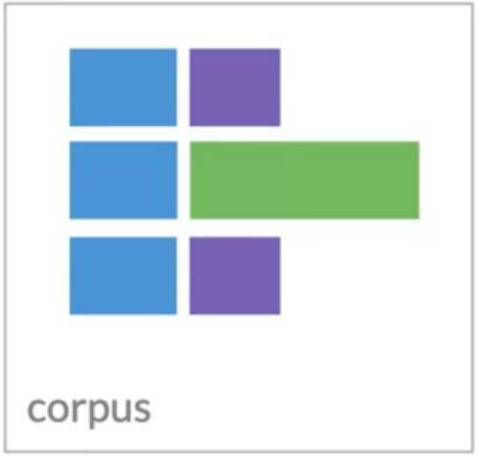


Transition probabilities



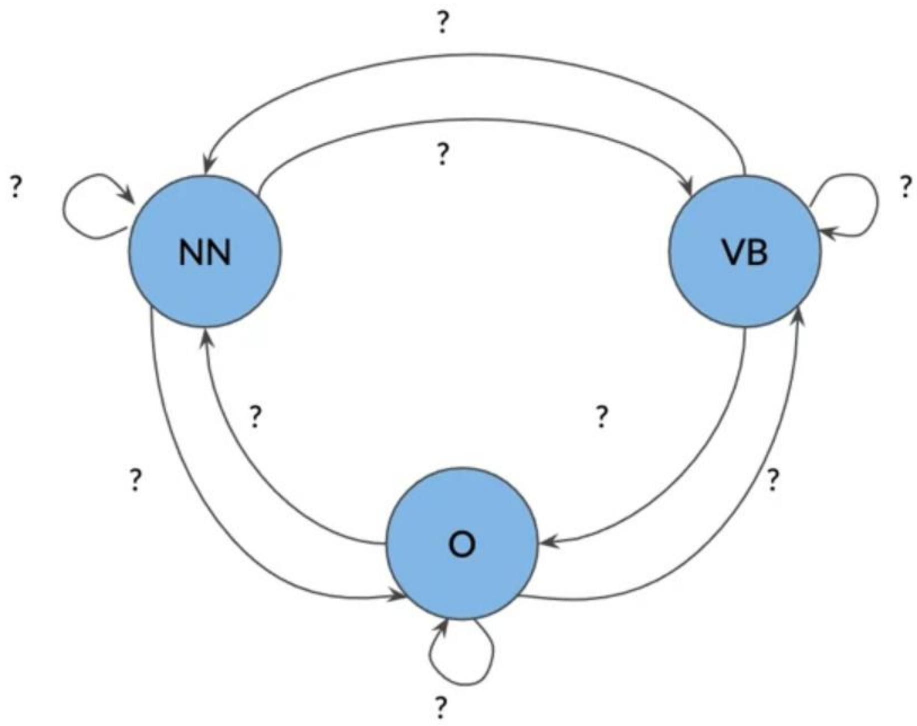


Transition probabilities



transition probability:  +  = ⅔

Transition probabilities



1. Count occurrences of tag pairs

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

2. Calculate probabilities using the counts

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

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

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

<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)	$\boxed{0}$		
VB (verb)	$C(\text{VB,NN})$		
O (other)	$C(\text{O,NN})$		

<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	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	$\boxed{8}$

<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

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

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

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

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

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

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+ε	0+ε	2+ε	3+3*ε
NN	0+ε	0+ε	6+ε	6+3*ε
VB	0+ε	0+ε	0+ε	0+3*ε
O	6+ε	0+ε	8+ε	14+3*ε

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

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

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

Emission probabilities

You eat
The oatmeal
You eat

corpus

You

Count: 2

Count: 3

Emission probabilities

You eat
The oatmeal
You eat

corpus

emission probability: You = ⅔

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)	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	...	...

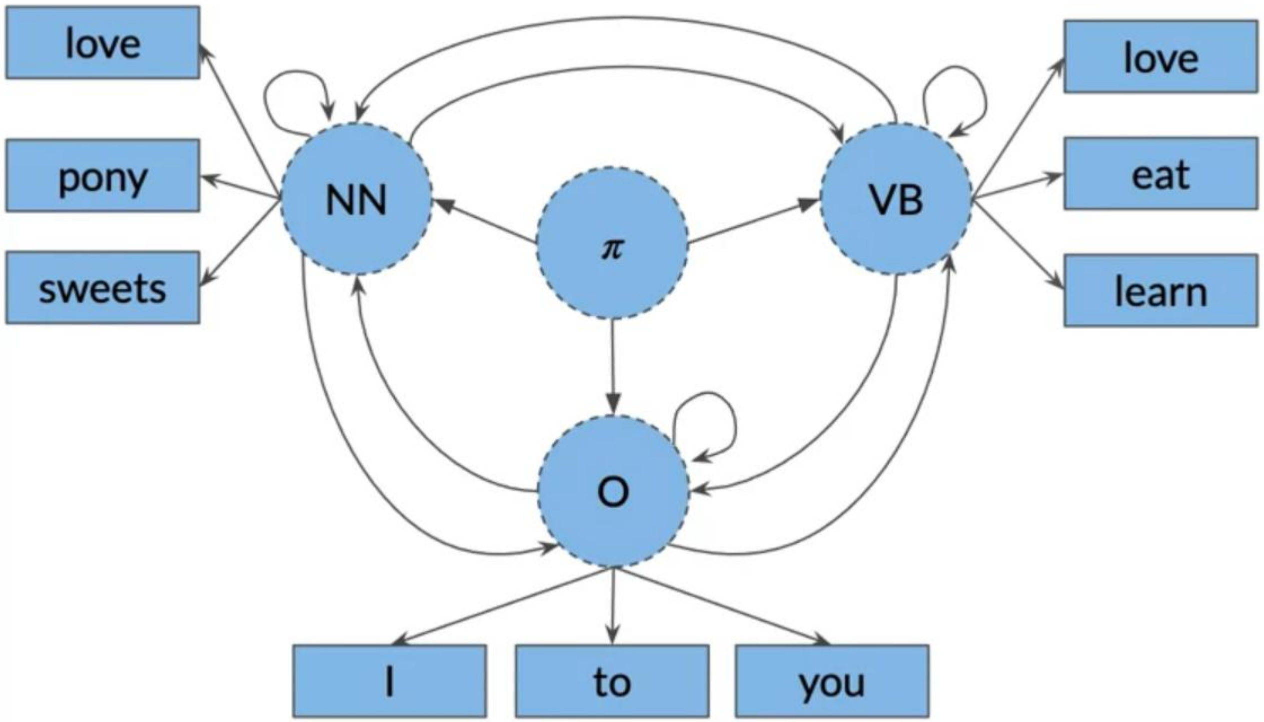
$$P(w_i|t_i) = \frac{C(t_i, w_i) + \epsilon}{\sum_{j=1}^{V} C(t_i, w_j) + N * \epsilon}$$
$$= \frac{C(t_i, w_i) + \epsilon}{C(t_i) + N * \epsilon}$$

Summary

1. Calculate transition and emission matrix

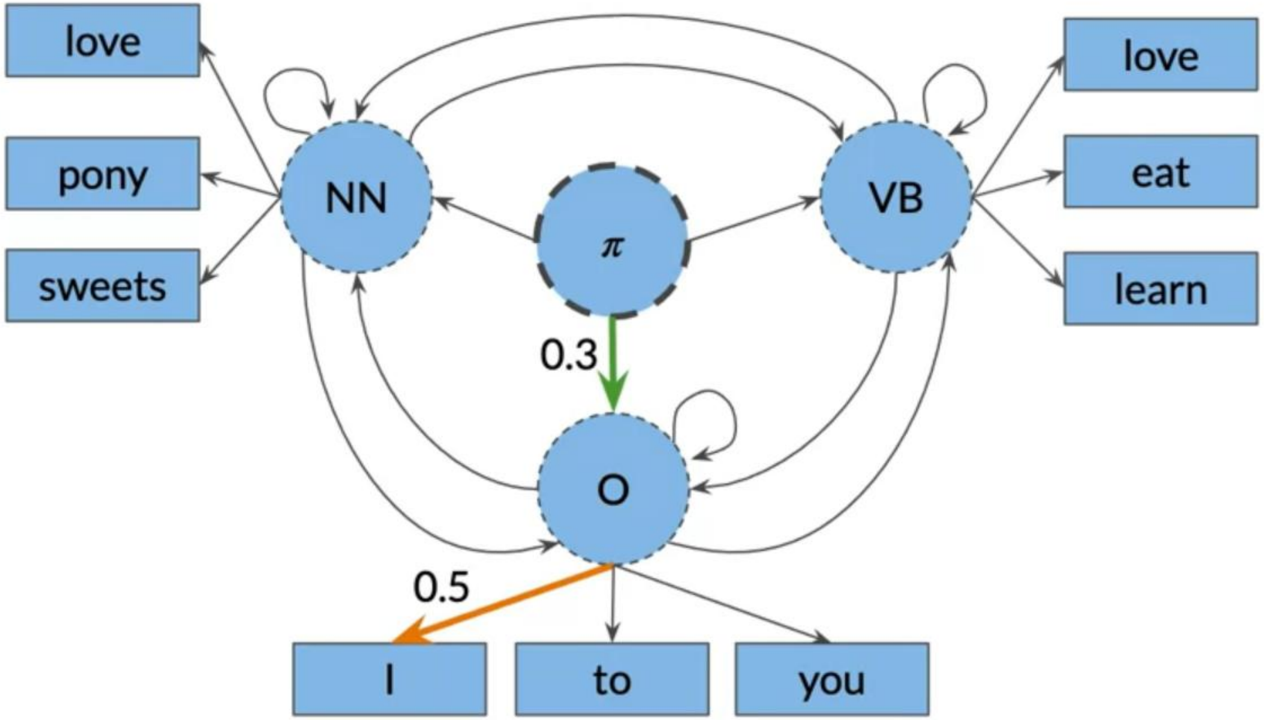
2. How to apply smoothing

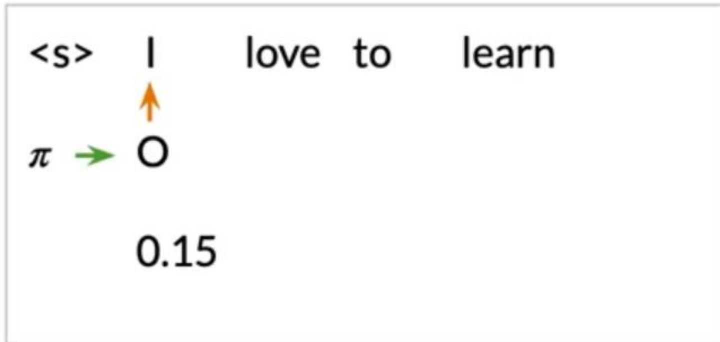
Viterbi algorithm – a graph algorithm



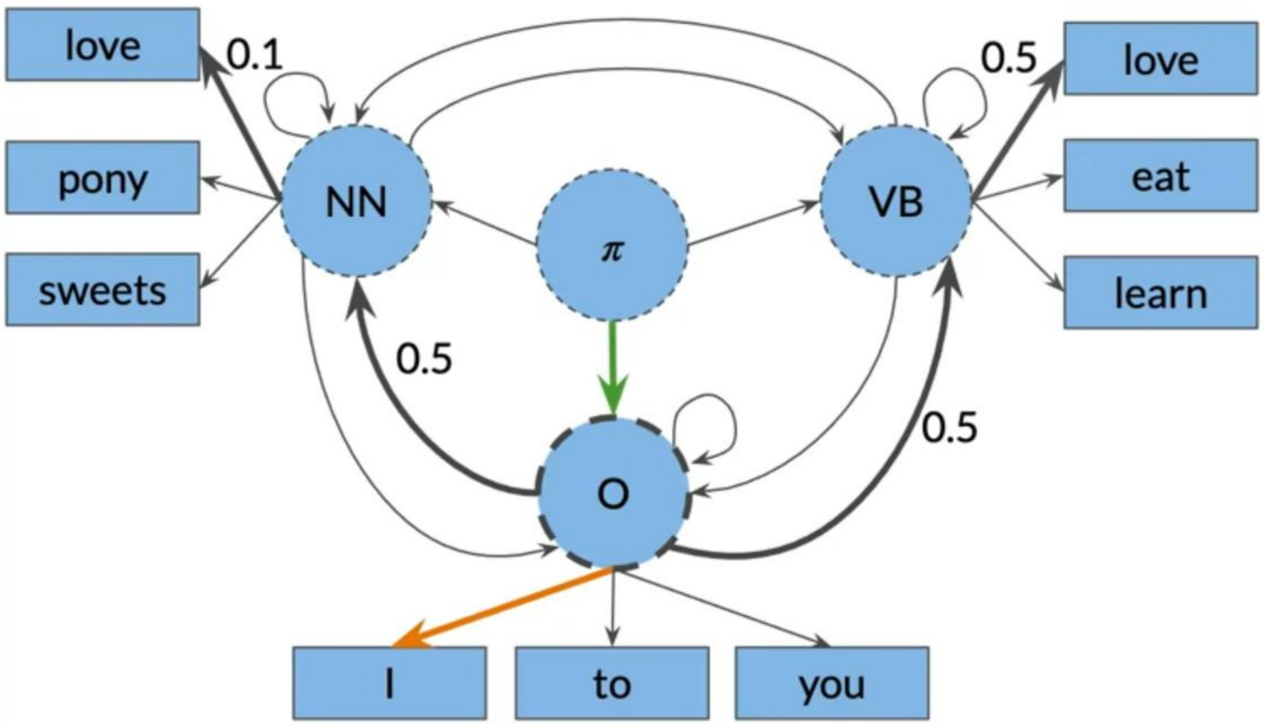
<s> I love to learn

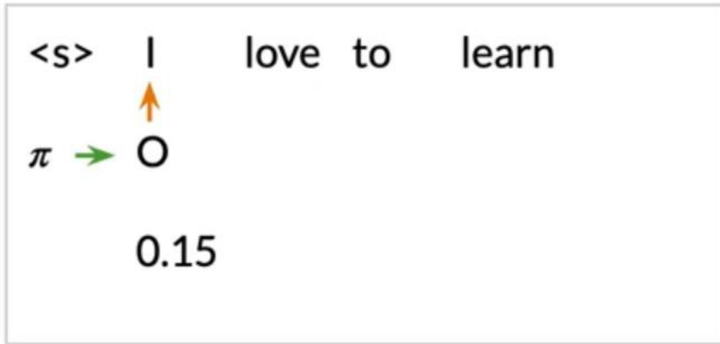
Viterbi algorithm – a graph algorithm



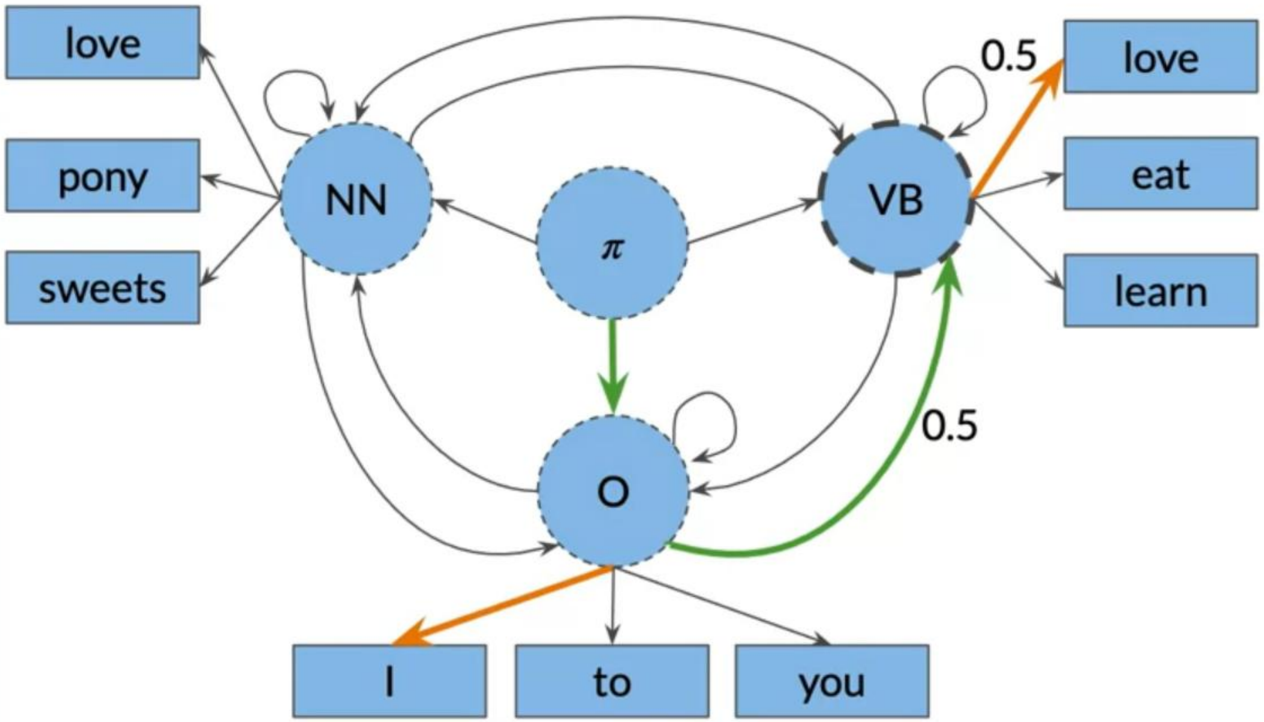


Viterbi algorithm – a graph algorithm





Viterbi algorithm – a graph algorithm

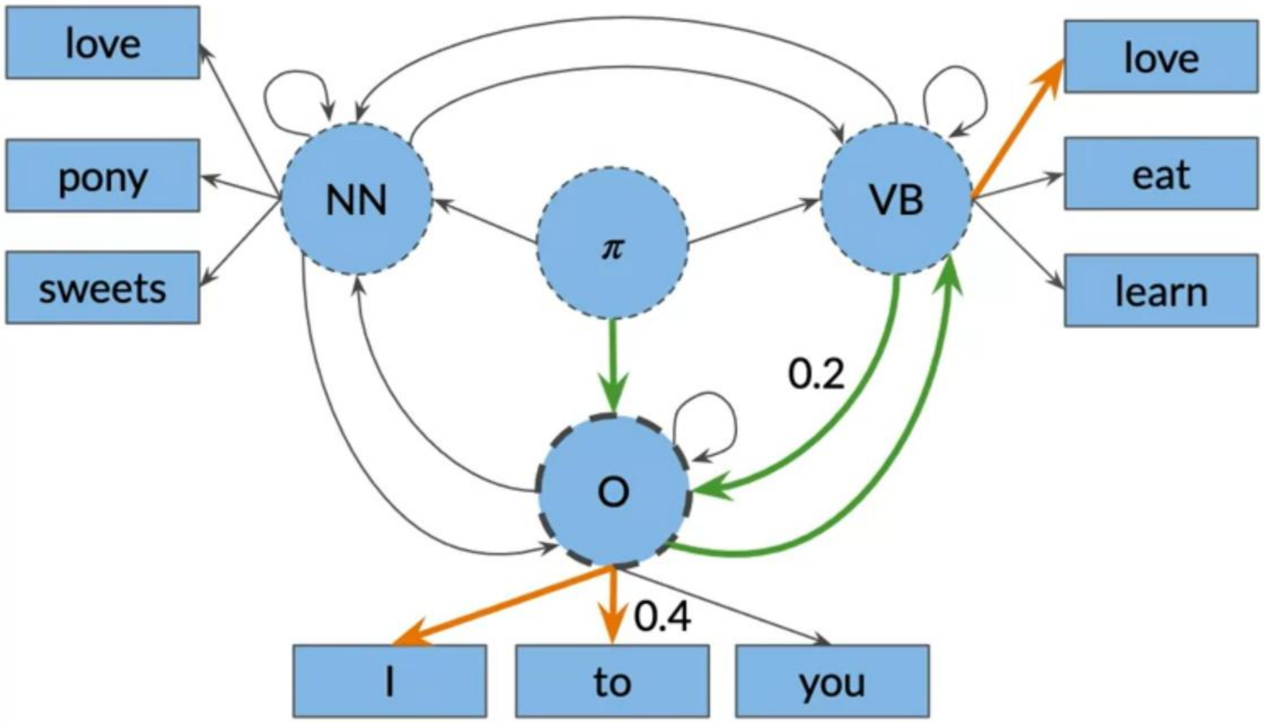


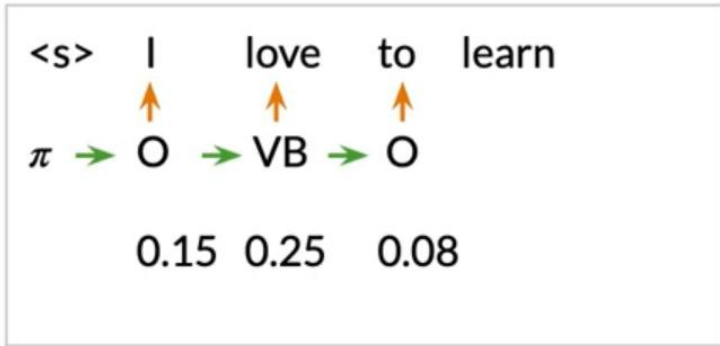
<s> I love to learn

$\pi \rightarrow O \rightarrow VB$

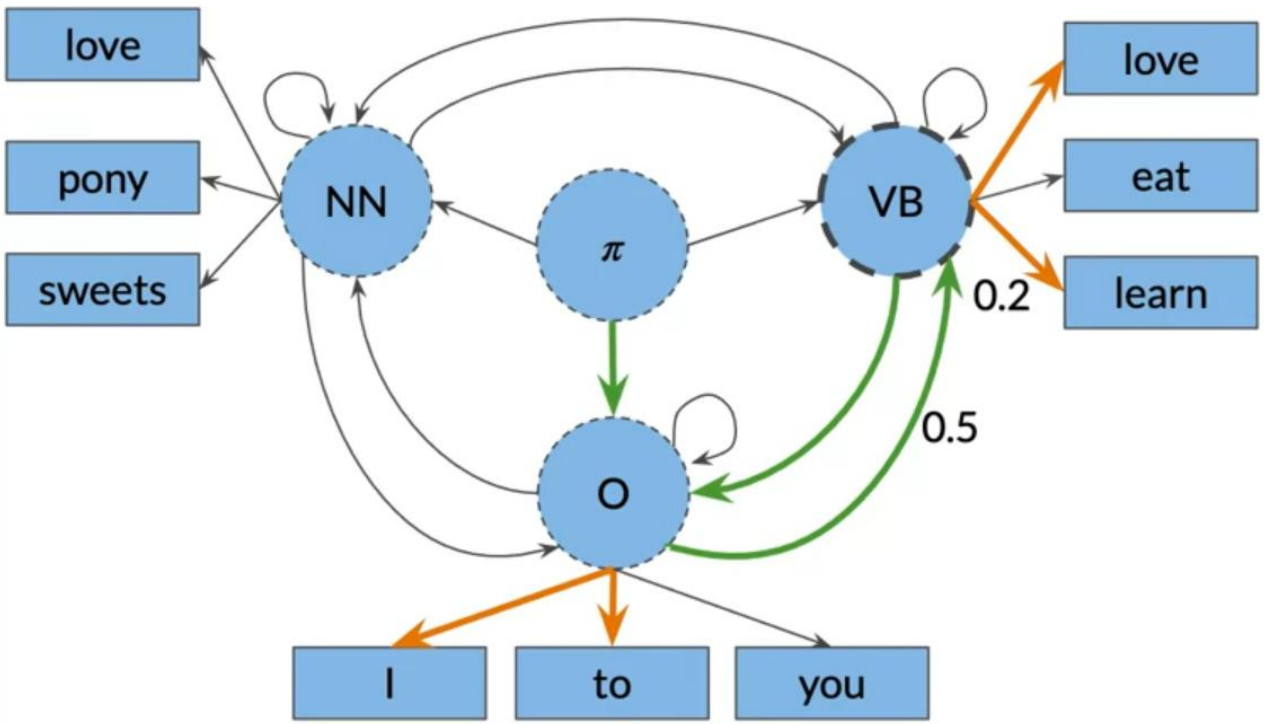
0.15 0.25

Viterbi algorithm – a graph algorithm



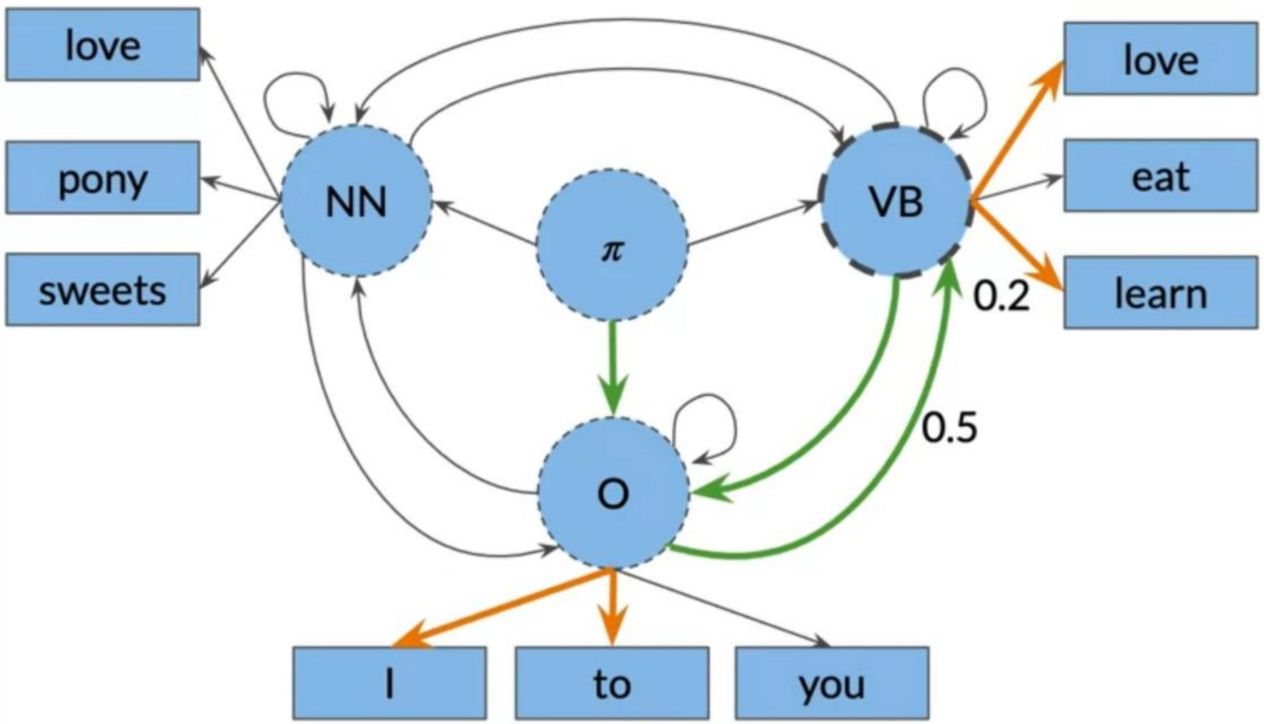


Viterbi algorithm – a graph algorithm



<s>	I	love	to	learn
π →	O →	VB →	O →	VB
	0.15	0.25	0.08	0.1

Viterbi algorithm – a graph algorithm



$$
\begin{array}{ccccc}
\text{<s>} & \text{I} & \text{love} & \text{to} & \text{learn} \\
 & \uparrow & \uparrow & \uparrow & \uparrow \\
\pi \rightarrow & O & \rightarrow VB & \rightarrow O & \rightarrow VB
\end{array}
$$

$0.15 * 0.25 * 0.08 * 0.1$

Probability for this sequence of hidden states: 0.0003

Viterbi algorithm – Steps

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

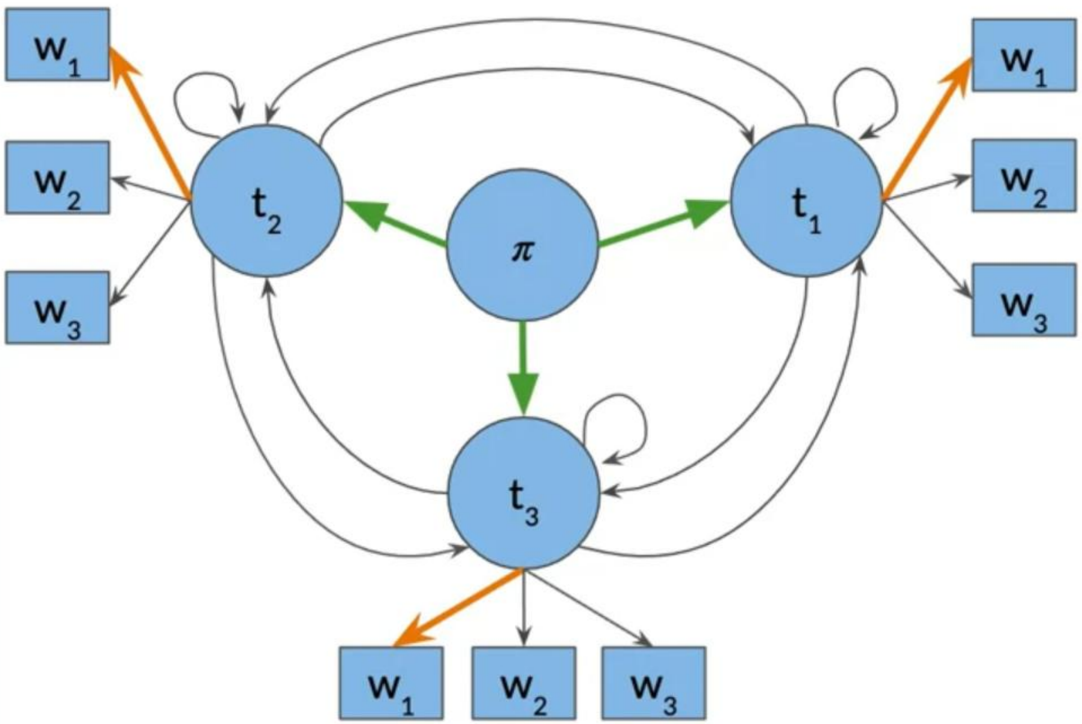
$C =$

	w_1	w_2	...	w_K
t_1				
...				
t_N				

$D =$

	w_1	w_2	...	w_K
t_1				
...				
t_N				

Initialization step

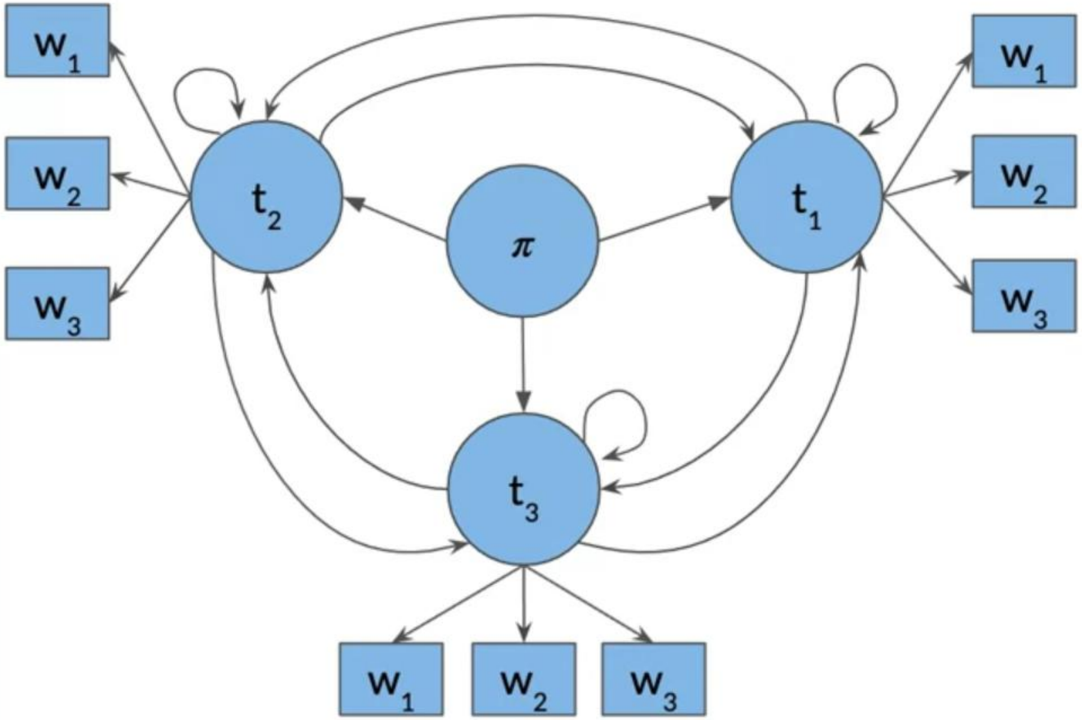


$C =$

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

$$c_{i,1} = \boxed{\pi_i} * \boxed{b_{i,cindex(w_1)}}$$
$$= a_{1,i} * b_{i,cindex(w_1)}$$

Initialization step

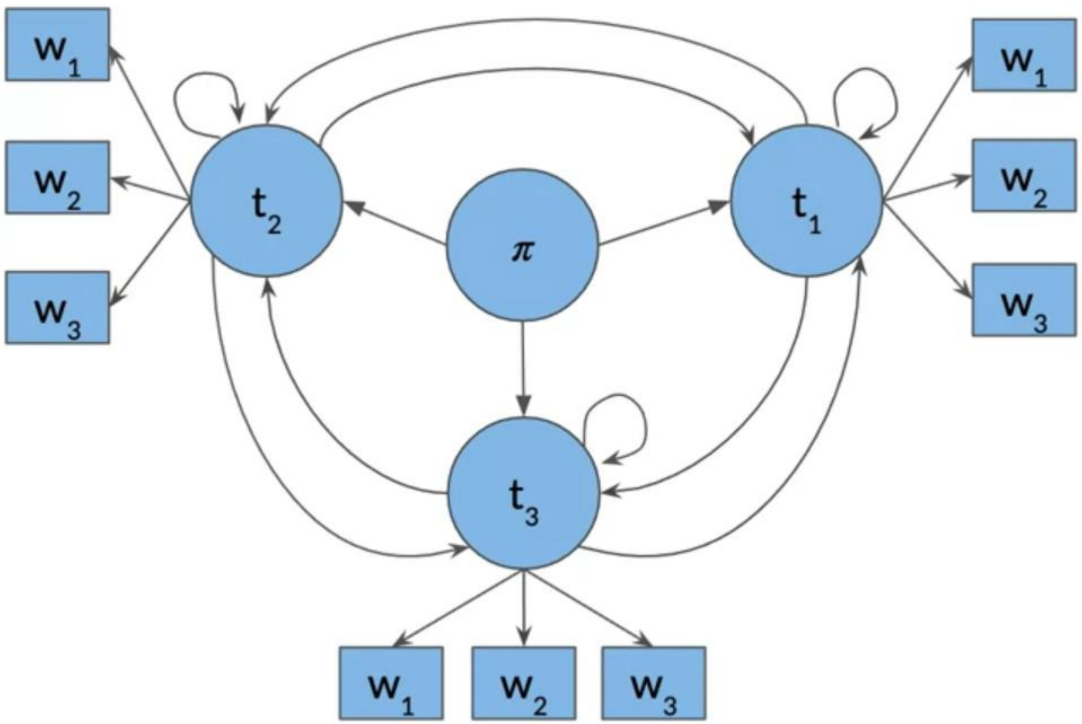


$D =$

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

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

Forward pass

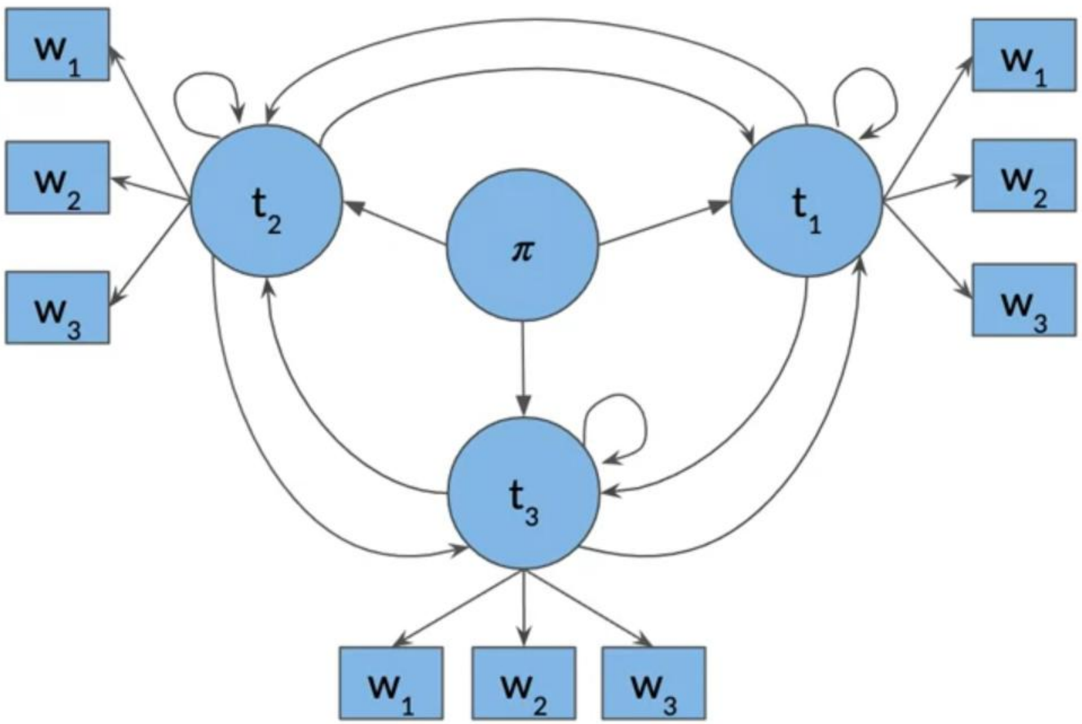


$C =$

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

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

Forward pass

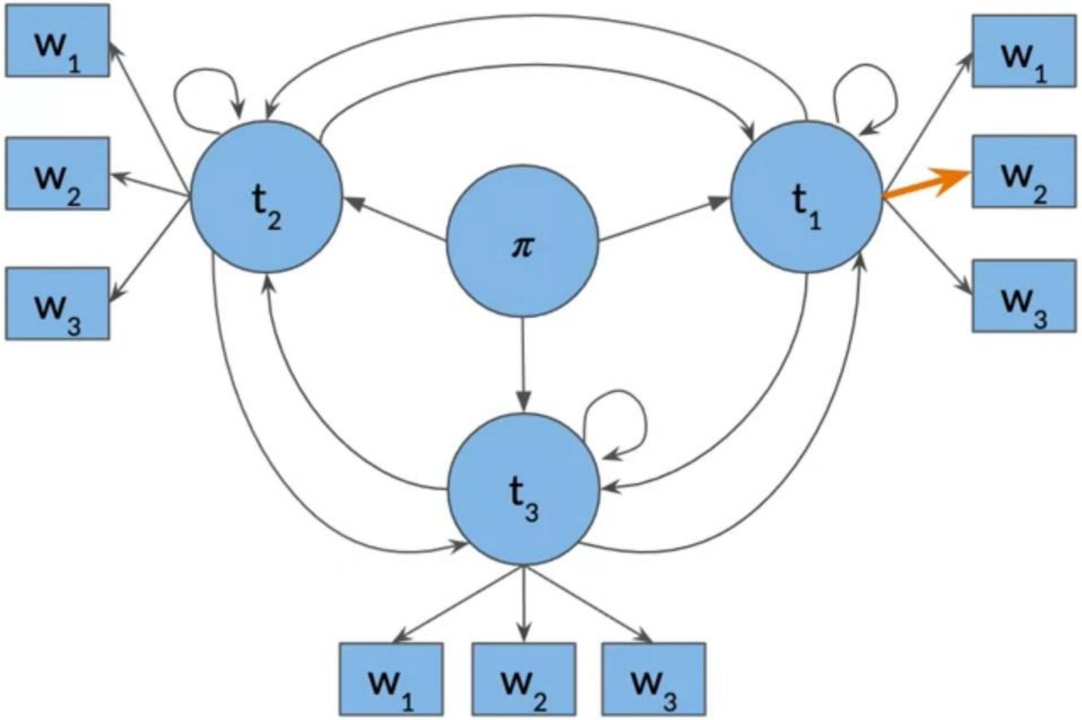


$C =$

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

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

Forward pass

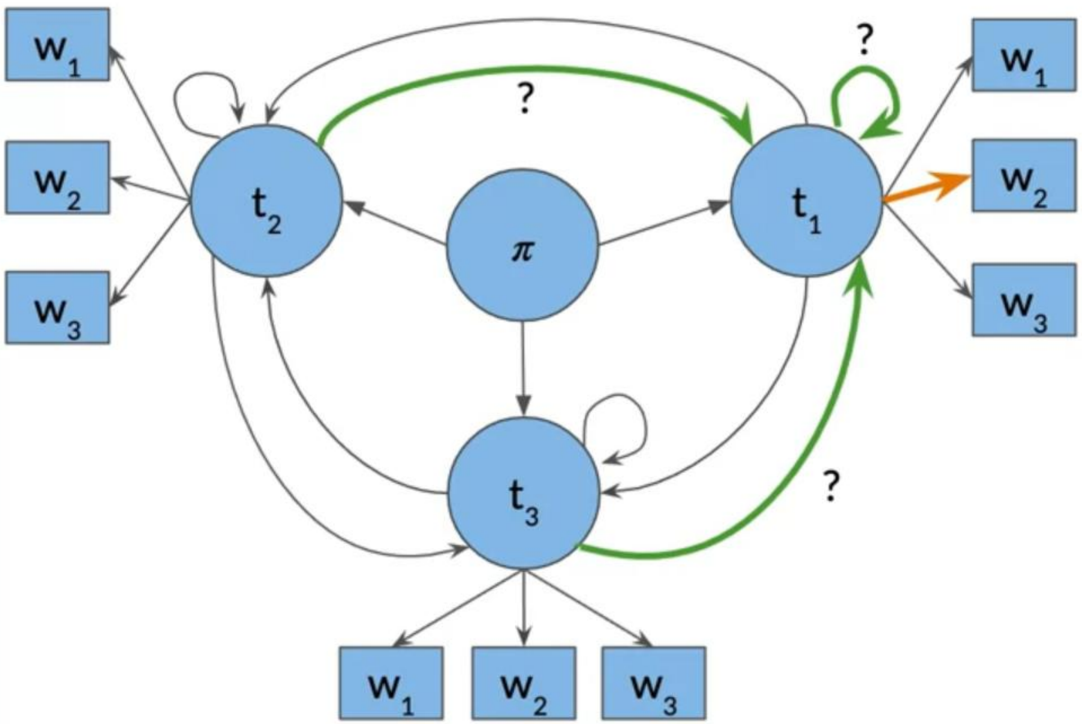


$C =$

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

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

Forward pass

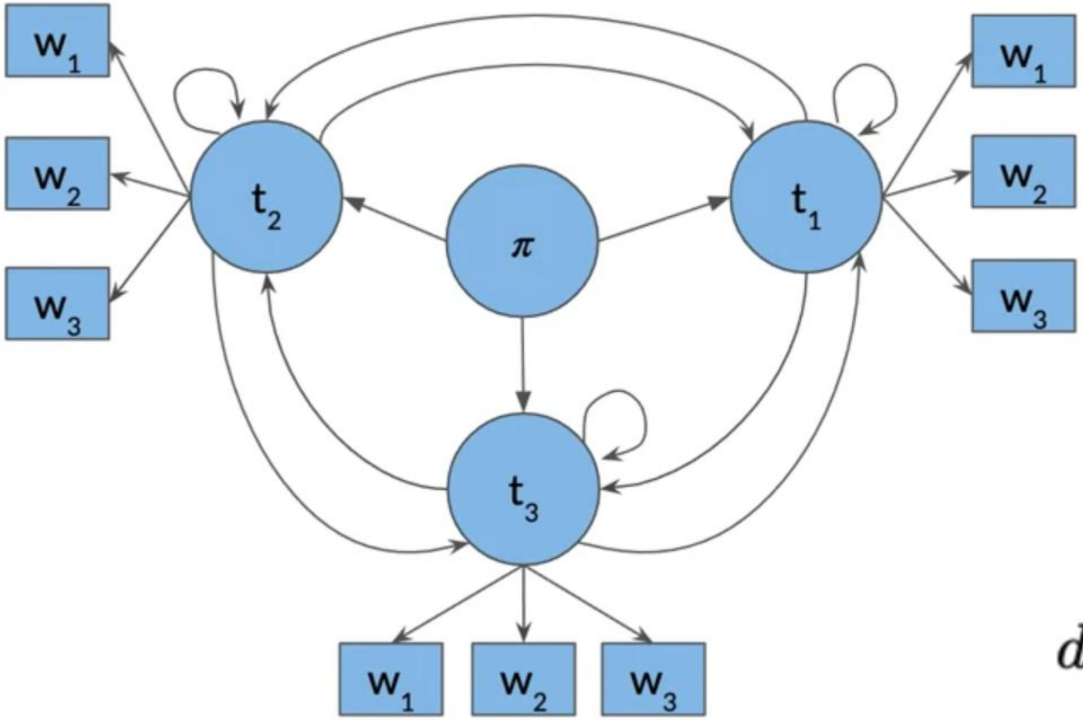


$C =$

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

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

Forward pass



$D =$

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

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

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

Backward pass

$C =$

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

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

$D =$

	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

<s> w1 w2 w3 w4 w5

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

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

<s> w1 w2 w3 w4 w5

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>	w1	w2	w3	w4	w5
					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

<s>	w1	w2	w3	w4	w5
				t_3 ←	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

<s> w1 w2 w3 w4 w5

$t_1 \leftarrow t_3 \leftarrow t_1$

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

<s> w1 w2 w3 w4 w5

$t_3 \leftarrow t_1 \leftarrow t_3 \leftarrow t_1$

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

<s> w1 w2 w3 w4 w5

$\pi \leftarrow t_2 \leftarrow t_3 \leftarrow t_1 \leftarrow t_3 \leftarrow t_1$

Implementation notes

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

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

$$\downarrow$$

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