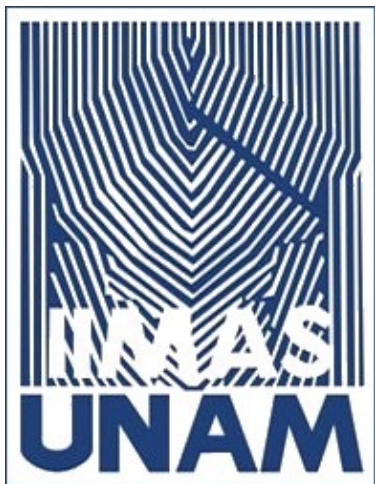


Procesamiento de Lenguaje Natural

Etiquetado gramatical (Partes de la oración)



Dra. Helena Gómez Adorno

helena.gomez@iimas.unam.mx

Dra. Gemma Bel

gbele@iingen.unam.mx

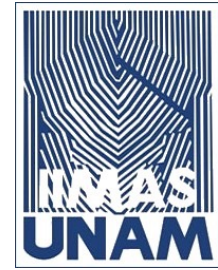


Correo del curso:

pln.cienciadedatos@gmail.com

Asistente:

Luis Ramon Casillas



Contenido

- Qué es el etiquetado gramatical?
- Cadenas de Markov
- Modelos ocultos de Markov
- Algoritmo de Viterbi

Why not learn something ?

adverb

adverb

verb

noun

punctuation
mark,
sentence
closer

Etiquetado gramatical (part of speech tagging)

- El *part of speech tagging* o *POS tagging* es el proceso de asignación de ciertas etiquetas o categorías a las palabras de un texto.

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 ?

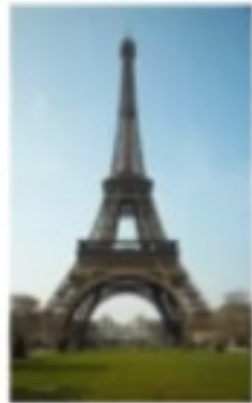
WRB **RB** **VB** **NN** .

Aplicaciones

- Ejemplo: La torre Eiffel está ubicada en París. Tiene 324 metros de alto.



Entidades nombradas



Resolución de co-referencia



Reconocimiento de voz

Ejemplo

Why not learn ...

verb

Ejemplo

Why not learn ...


verb

verb?


noun?

...?

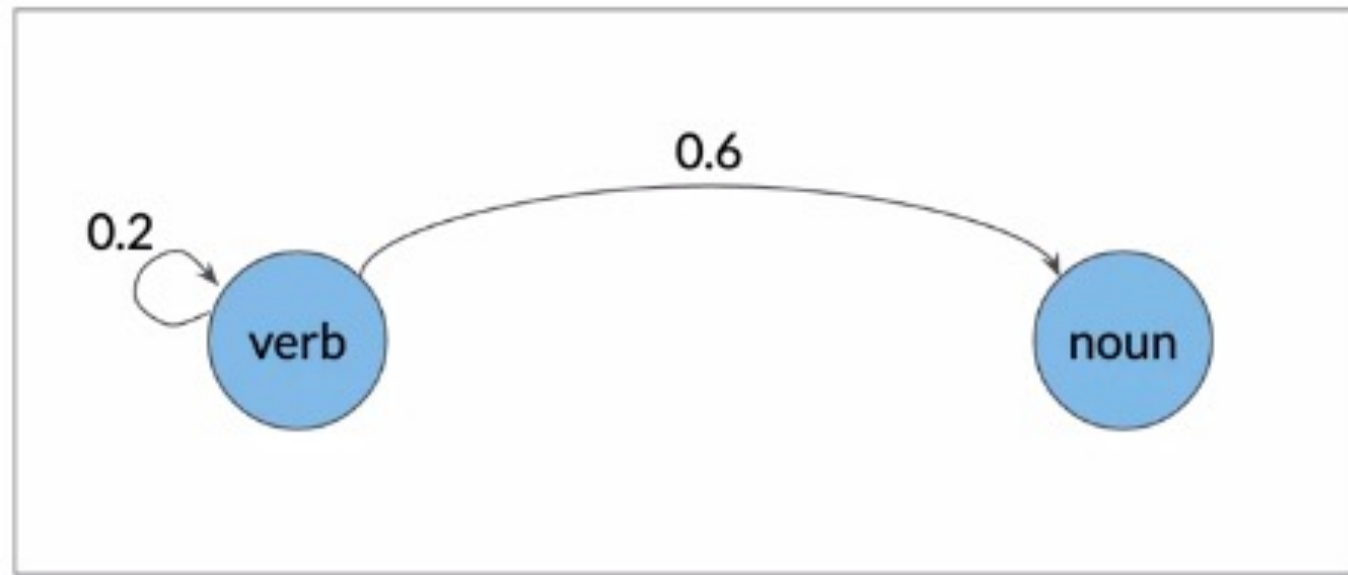
Ejemplo

Why not learn  ...
verb verb?
noun?
...?

Ejemplo

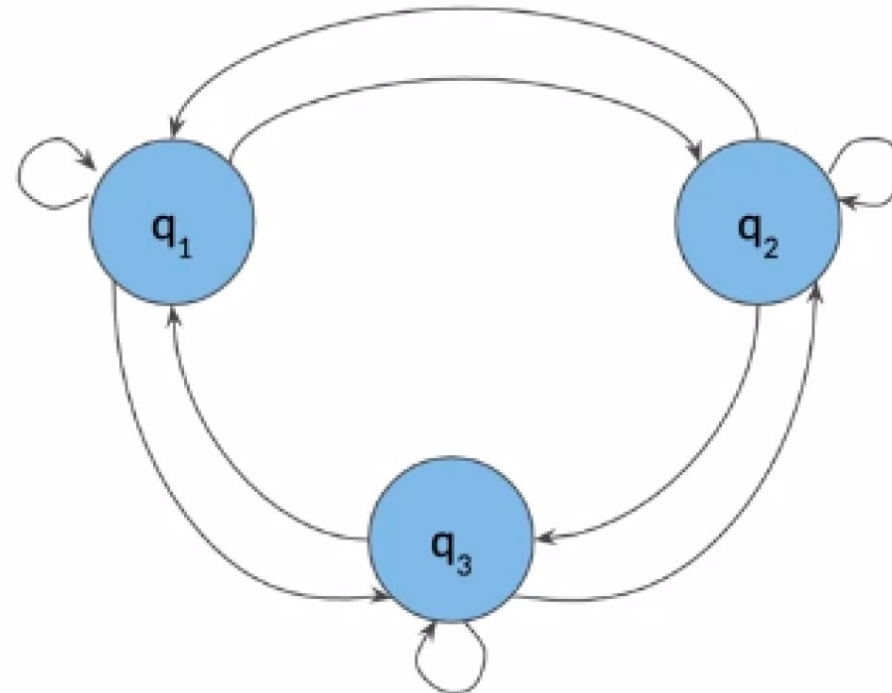
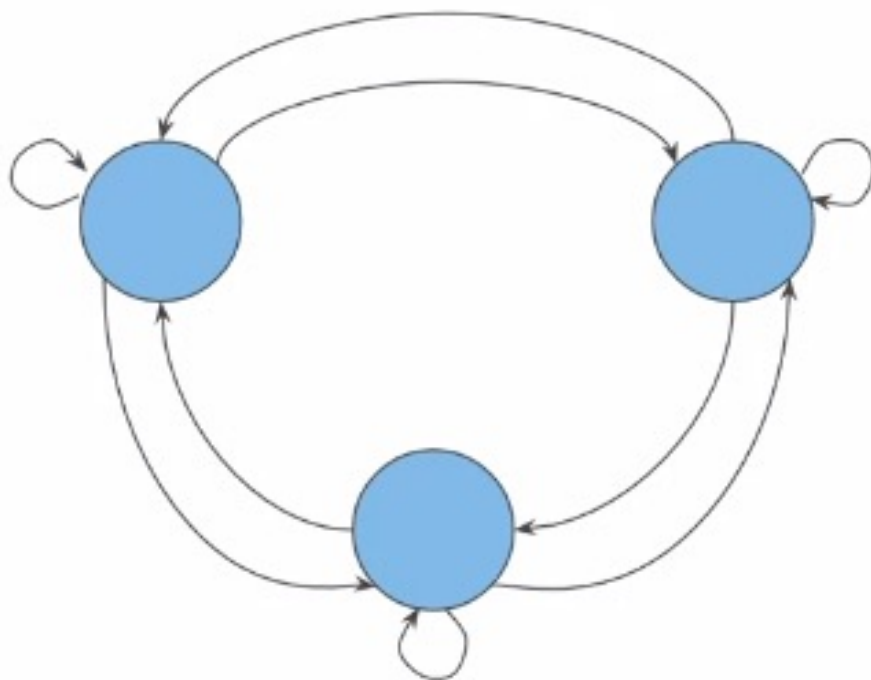
Why not learn  ...
verb verb?
 ↘ noun?
 ...?

Representación visual



Qué son las cadenas de Markov

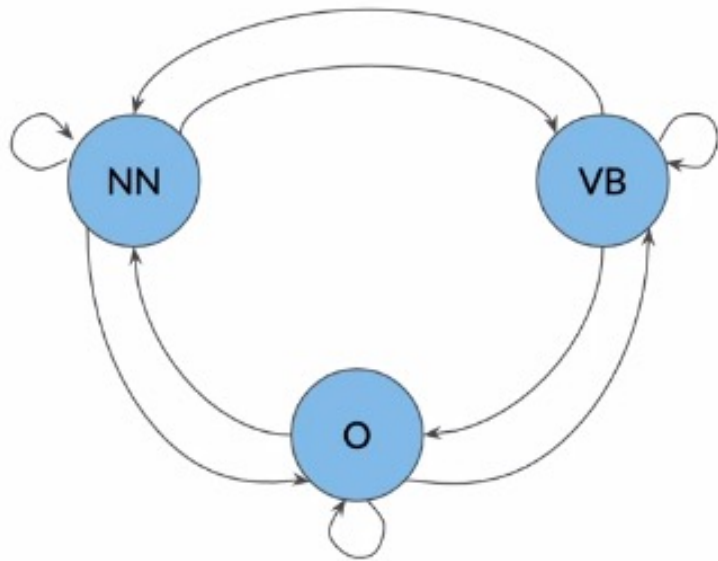
- Las cadenas de Markov son un tipo de proceso estocástico que describen una secuencia de eventos posibles. Para obtener la probabilidad del evento siguiente solo se necesita la probabilidad del evento previo.



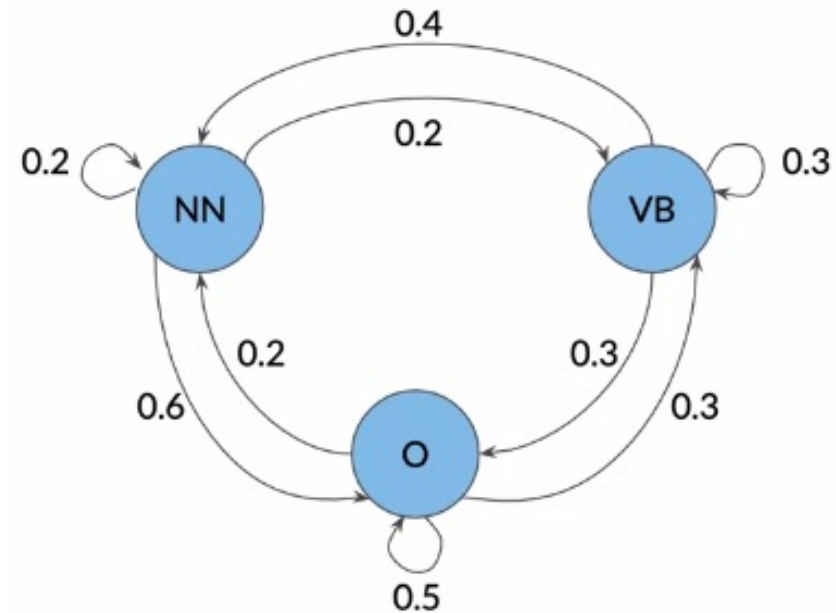
$$Q = \{q_1, q_2, q_3\}$$

Cadenas de Markov y etiquetas POS

- Las cadenas de Markov se utilizan en el etiquetado POS ya que solo necesita del estado actual para establecer probabilidades para el próximo estado.

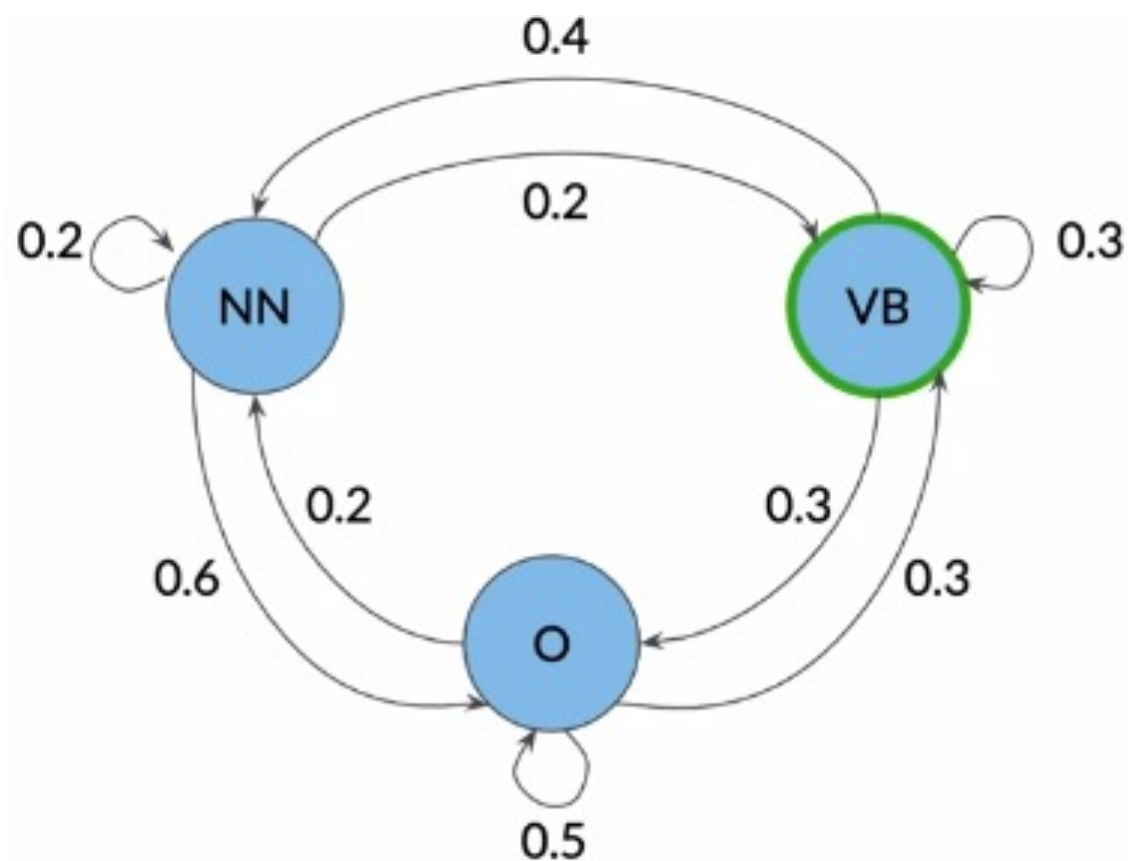


Etiquetas POS como estados



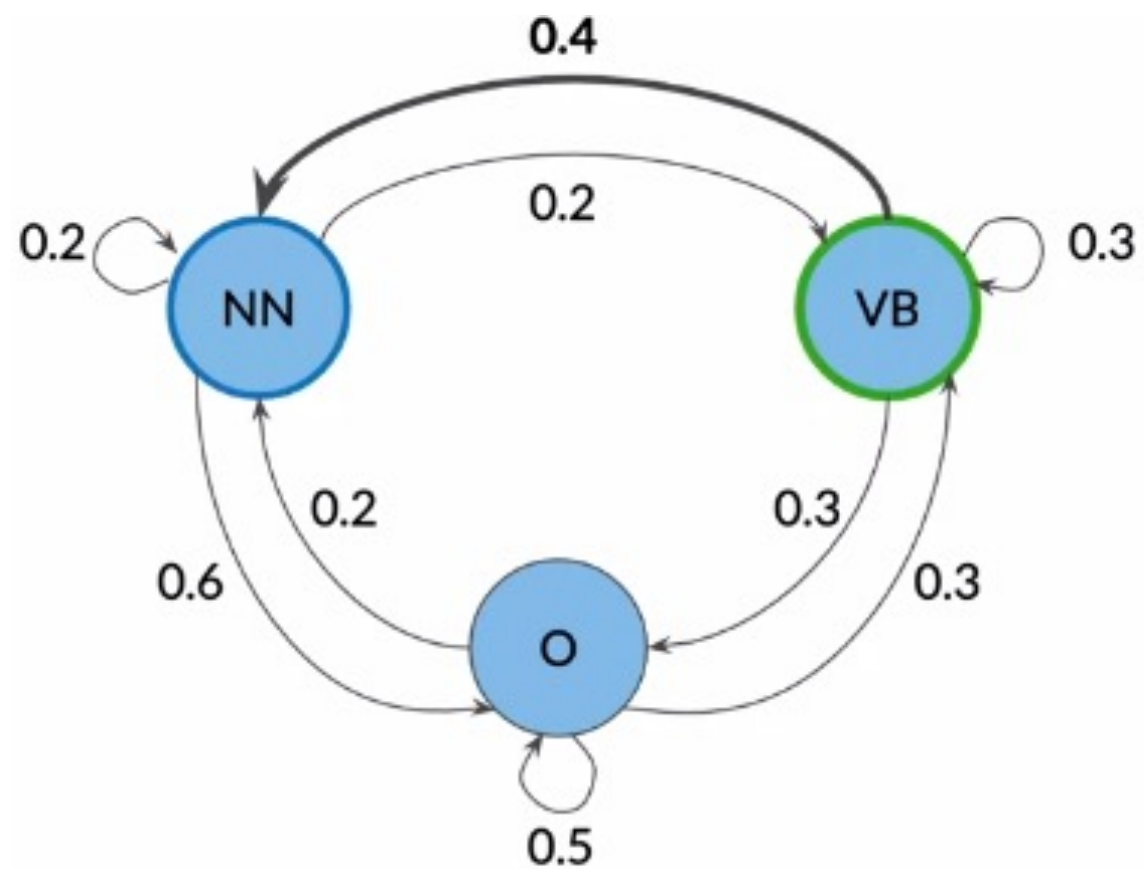
Probabilidades de transición

Probabilidades de transición



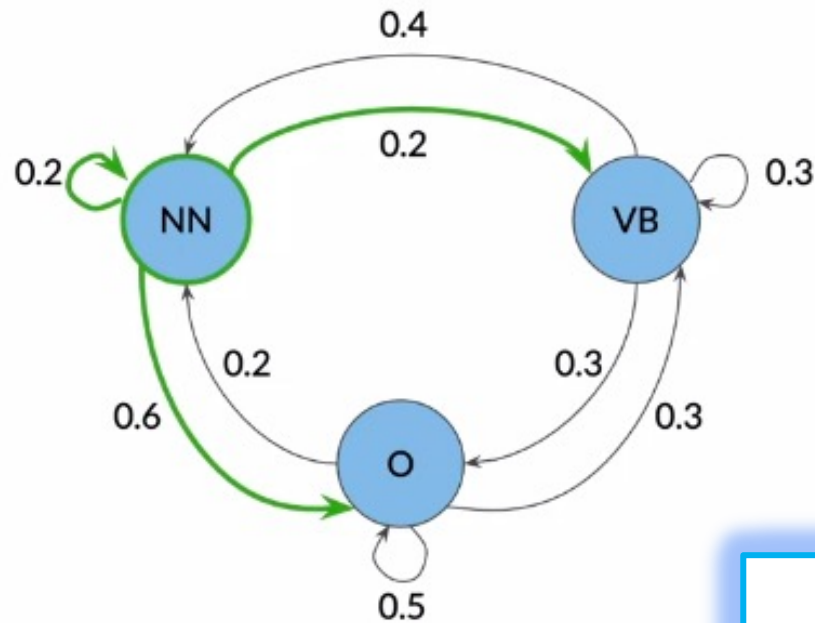
Why not **learn** something?

Probabilidades de transición



Why not learn something?

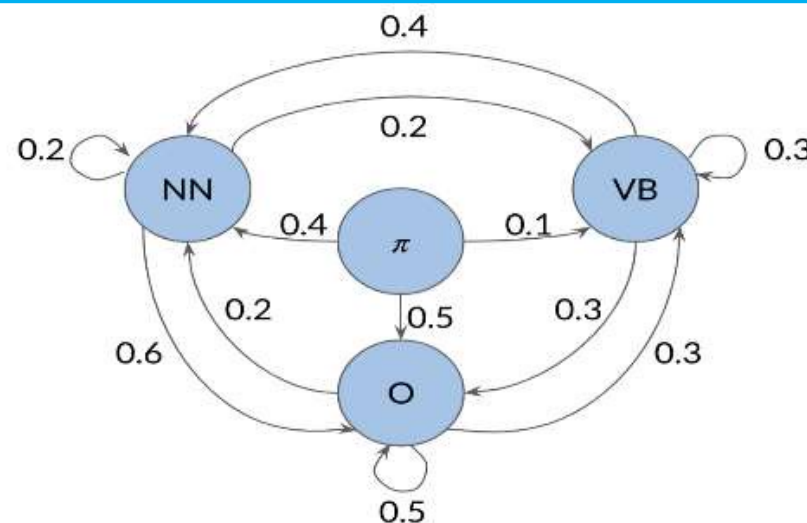
La matriz de transición



$A =$

	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

La primera palabra?



$A =$

	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

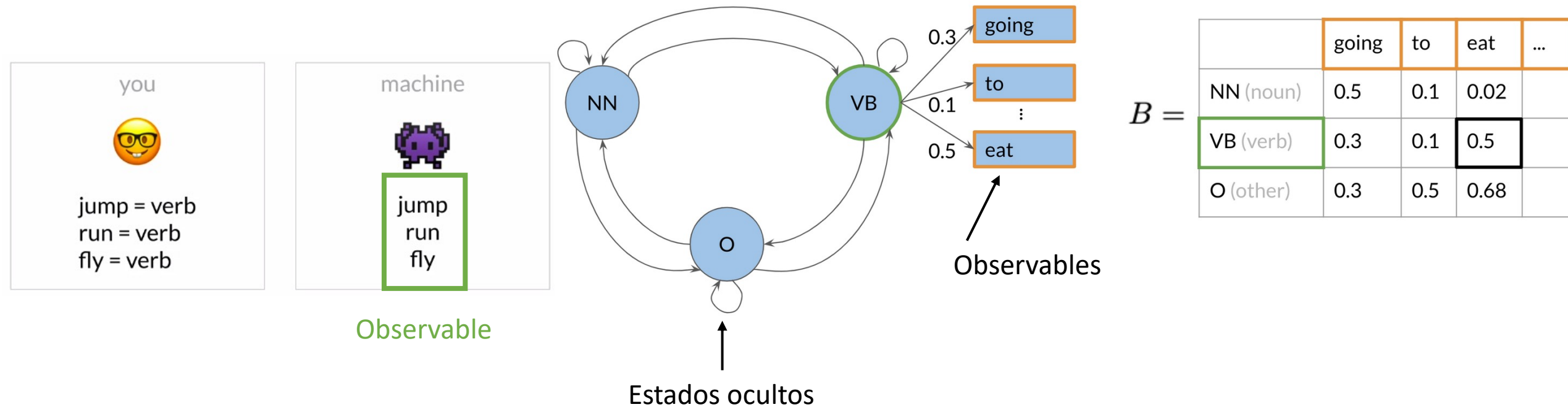
La matriz de transición

- En notación más general, puede escribir la matriz de transición A , dados algunos estados Q , de la siguiente manera:

States	Transition matrix
$Q = \{q_1, \dots, q_N\}$	$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{N+1,1} & \dots & a_{N+1,N} \end{pmatrix}$

Modelos ocultos de Markov

- En los modelos ocultos de Markov se usan probabilidades de emisión (probabilidad de pasar de una etiqueta a una palabra específica)



Modelos ocultos de Markov

- La matriz de emisión B, se utilizará con su matriz de transición A, para ayudar a identificar la parte gramatical de una palabra en una oración. Para calcular la matriz B, se utiliza un conjunto de datos etiquetado y calcular las probabilidades de pasar de un POS a cada palabra de su vocabulario.

States

Transition matrix

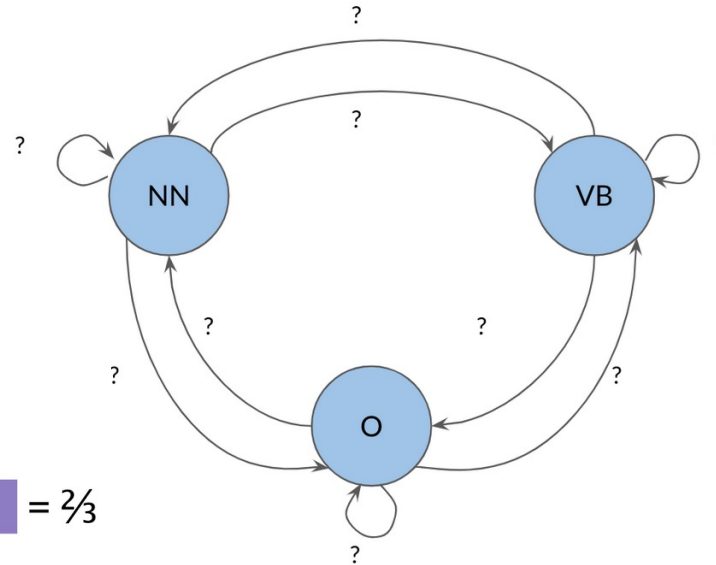
Emission matrix

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

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

Modelos ocultos de Markov

- El número de veces que el azul es seguido por el púrpura es 2 de 3. Usaremos la misma lógica para poblar nuestras matrices de transición y emisión.
- En la matriz de transición contaremos el número de veces que etiqueta $t_{(i-1)}$, $t_{(i)}$ aparecer cerca uno del otro y dividir por el número total de veces que aparece $t_{(i-1)}$ (que es el mismo que el número de veces que aparece seguido de cualquier otra cosa).



1. Count occurrences of tag pairs

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

2. Calculate probabilities using the counts

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

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



Calcular las probabilidades de transición

$A =$

	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

Calcular las probabilidades de transición

$A =$

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

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

Suavizado →

$A =$

	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	$6+\epsilon$	$0+\epsilon$	$8+\epsilon$	$14+3*\epsilon$

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

Calcular las probabilidades de emisión

- Igualmente, para completar la matriz de probabilidades de emisión se siguen los siguientes pasos:
- Identificar el número de veces que una palabra en específico aparece con cada etiqueta.
- Utilizar la misma formula de la matriz de transición.

$B =$

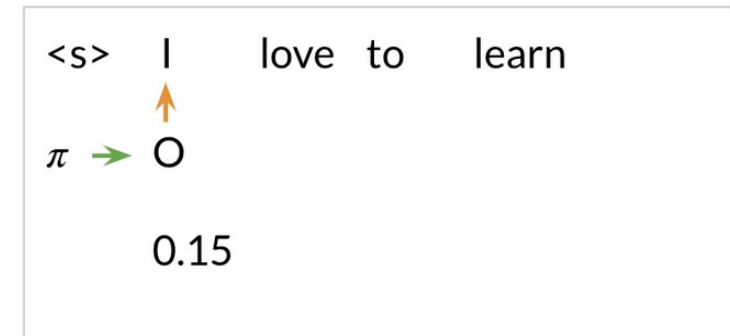
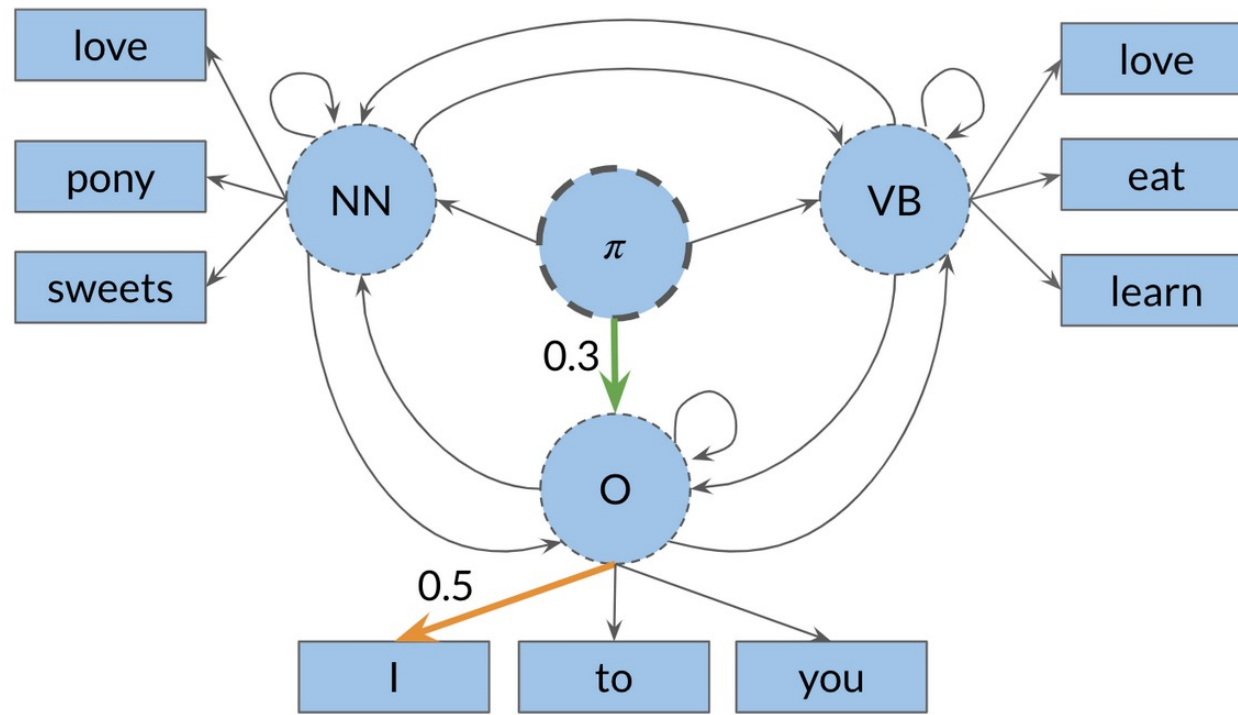
	in	a	...
NN (noun)	0		
VB (verb)	0		
O (other)	2		

<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

El Algoritmo de Viterbi

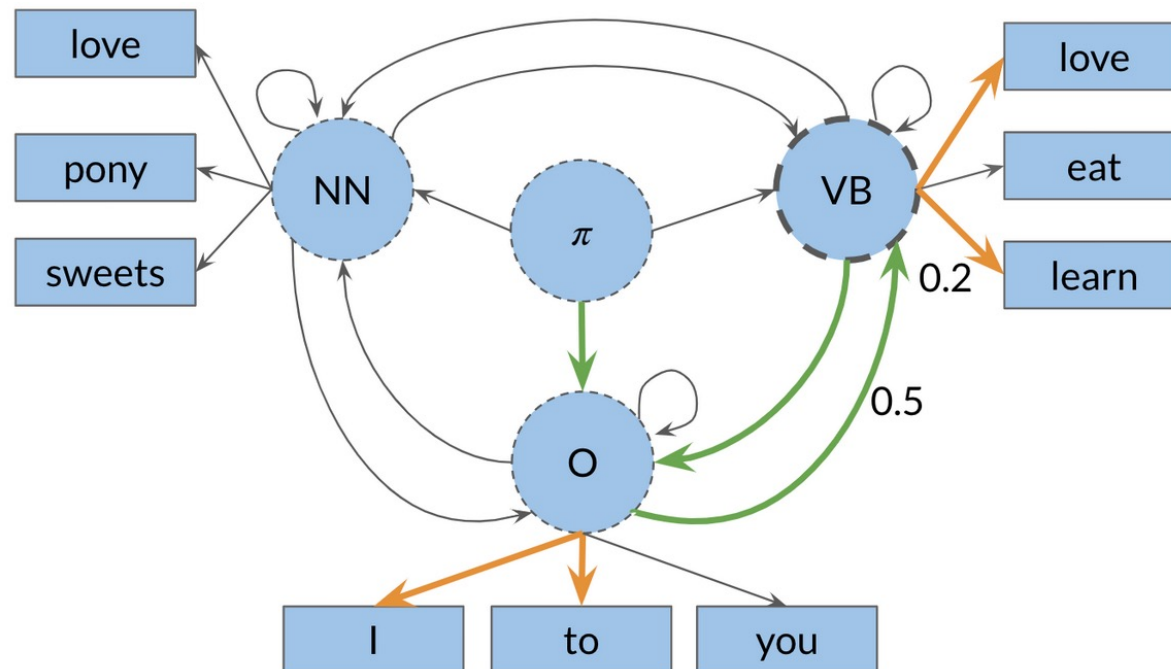
- El Algoritmo de Viterbi es un proceso que utiliza las probabilidades de transición y emisión para encontrar la probabilidad de una secuencia de palabras y la secuencia más probable.
- Para pasar de π a O, debes multiplicar la probabilidad de transición correspondiente (0.3) y la probabilidad de emisión correspondiente (0.5), lo que le da 0.15. Sigue haciendo eso para todas las palabras, hasta que obtenga la probabilidad de una secuencia completa.



El Algoritmo de Viterbi

Pasos del algoritmo:

- Inicialización
- Paso hacia delante
- Paso hacia atrás

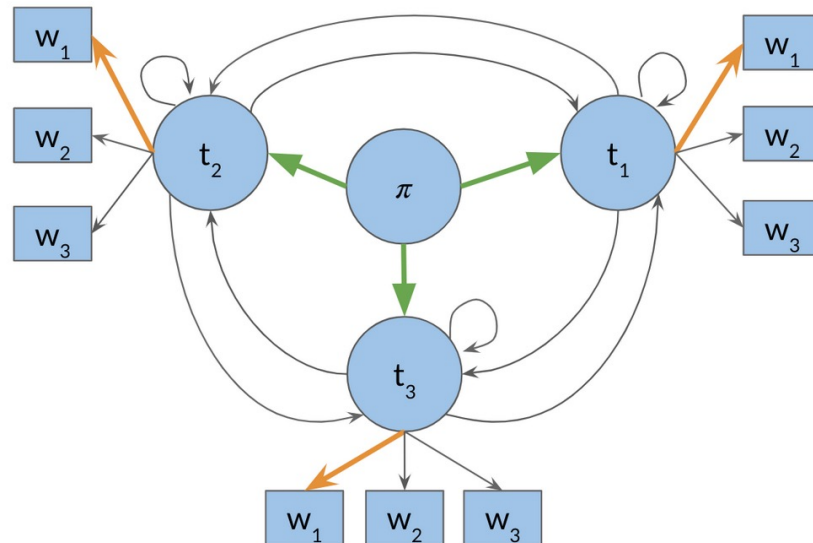


$\langle s \rangle$ I love to learn
↑ ↑ ↑ ↑
 $\pi \rightarrow O \rightarrow VB \rightarrow O \rightarrow VB$
 $0.15 * 0.25 * 0.08 * 0.1$

Probability for this sequence of hidden states: 0.0003

Inicialización de Viterbi

- Se calcula una matriz C de dimensión $(\text{num_etiquetas}, \text{num_palabras})$. Esta matriz tendrá las probabilidades que le dirán a qué categoría gramatical pertenece cada palabra.
- Para completar la primera columna, simplemente multiplique la distribución inicial π , para cada etiqueta, por $b_{i, \text{index}(w_1)}$. Donde i , corresponde a la etiqueta de la distribución inicial y $\text{index}(w_1)$, es el índice de la palabra 1 en la matriz de emisión.



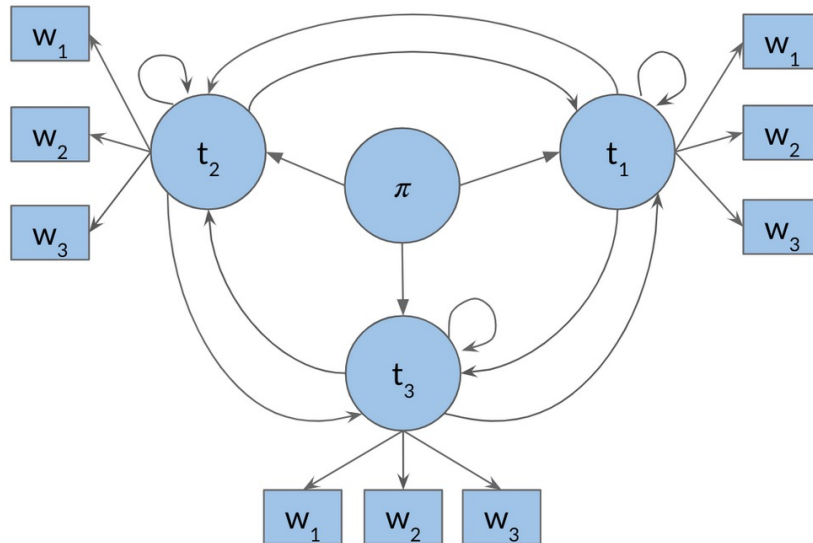
$C =$

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

$$\begin{aligned} c_{i,1} &= \pi_i * b_{i, \text{index}(w_1)} \\ &= a_{1,i} * b_{i, \text{index}(w_1)} \end{aligned}$$

Inicialización de Viterbi

- Ahora debemos realizar un seguimiento de la etiqueta gramatical de la que proviene. Por lo tanto, calculamos una matriz D , que nos permite almacenar las etiquetas que representan los diferentes estados por los que está pasando al encontrar la secuencia más probable de etiquetas gramaticales para la secuencia dada de palabras w_1, \dots, w_K . Al principio, establecemos la primera columna en 0, porque no proviene de ninguna etiqueta POS.



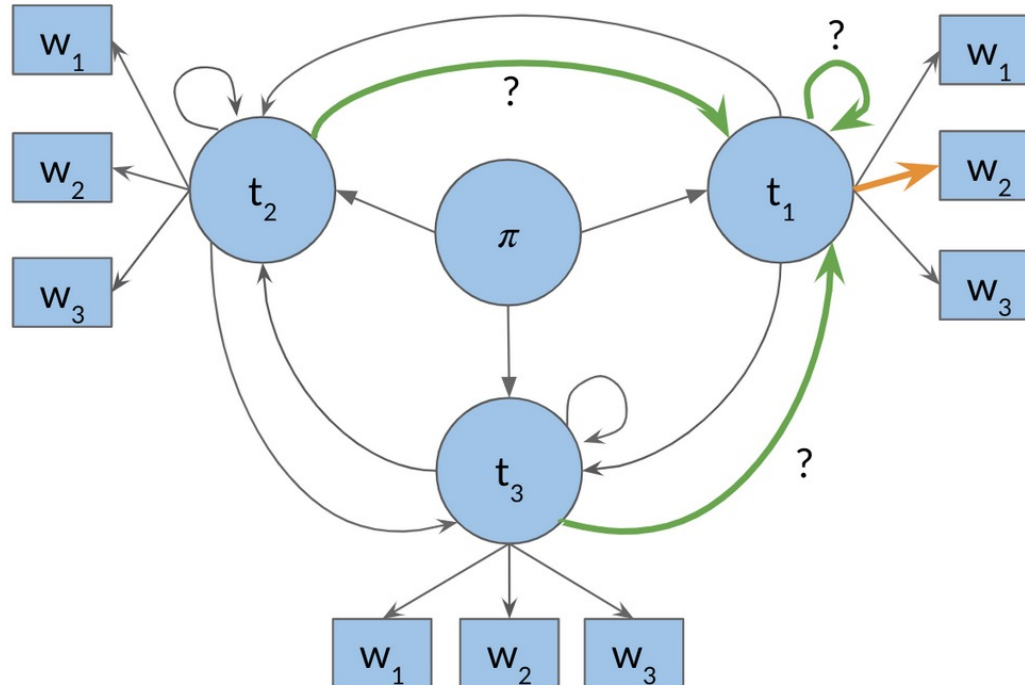
$D =$

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

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

Viterbi: Paso hacia adelante

- Para llenar una celda (por ejemplo, 1,2), debemos tomar el máximo de [K celdas en la columna anterior, multiplicado por la probabilidad de transición correspondiente del POS K al primer POS multiplicado por la probabilidad de emisión del primer POS y la palabra actual que está viendo]. Lo mismo para todas las celdas.



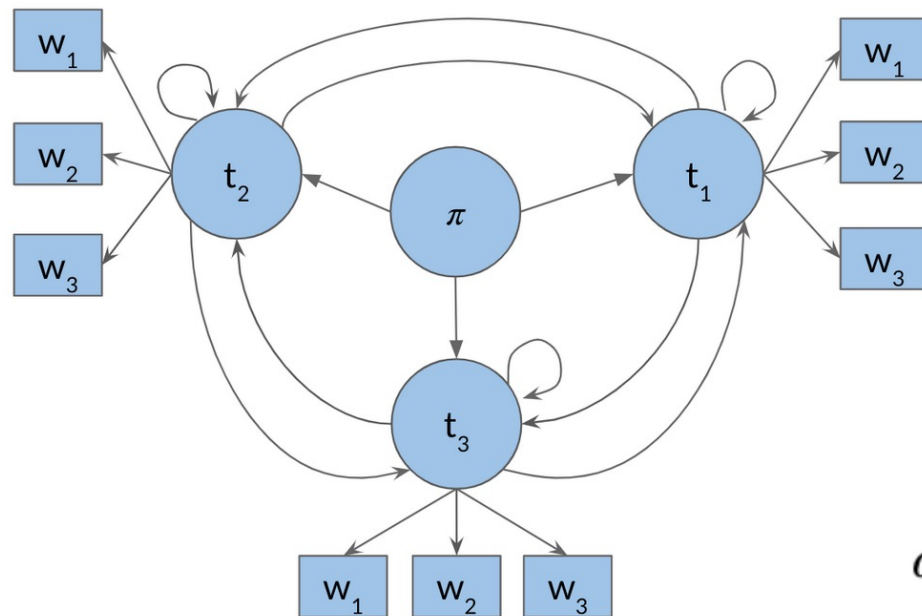
$C =$

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

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

Viterbi: Paso hacia adelante

- Ahora, para completar la matriz D, realizará un seguimiento del *argmax* de su procedencia de la siguiente manera:



$D =$

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

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

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

- La única diferencia entre $c_{i,j}$ y $d_{i,j}$, es que en el primero calcula la probabilidad y en el segundo realiza un seguimiento del índice de la fila de donde proviene esa probabilidad. Por lo tanto, realiza un seguimiento de qué k se usó para obtener esa probabilidad máxima.

Viterbi: Paso hacia atrás

$C =$

	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

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

$D =$

	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
π	$\leftarrow t_2$	$\leftarrow t_3$	$\leftarrow t_1$	$\leftarrow t_3$	$\leftarrow t_1$