

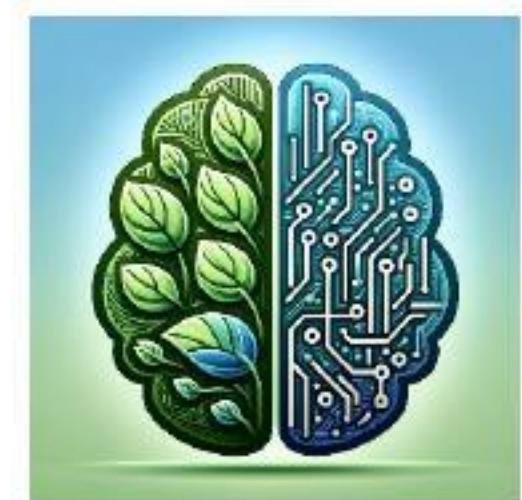


Natural Language Processing with Deep Learning

Bài 7: PART-OF-SPEECH TAGGING



https://csc.edu.vn/data-science-machine-learning/natural-language-processing-with-deep-learning_293





PART-OF-SPEECH TAGGING

I. Tổng quan về Part-Of-Speech Tagging

II. Markov Chains

III. Markov Models

IV. Giải thuật Viterbi

Tổng quan Part-of-speech Tagging

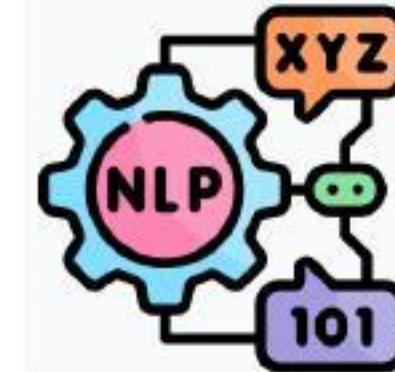


PART-OF-SPEECH TAGGING



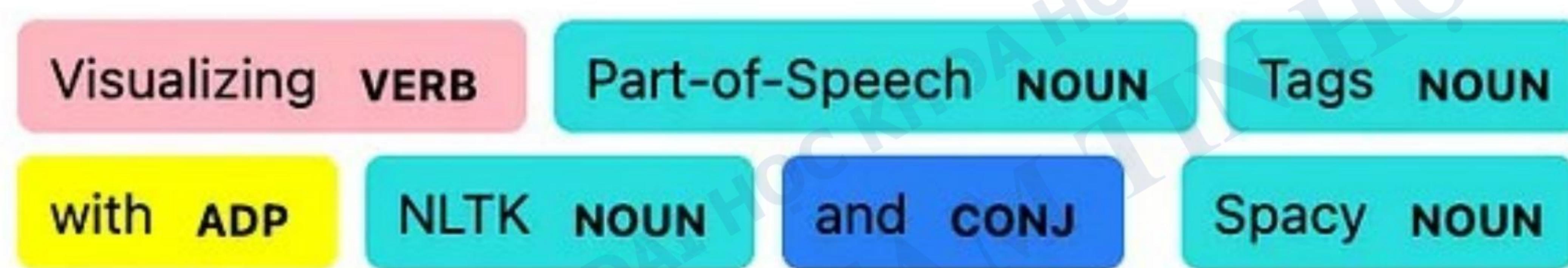
Là các nhãn được gán cho các từ trong câu, chỉ ra loại từ ngũ pháp hoặc chức năng cú pháp của chúng.

→ Đóng vai trò quan trọng trong các tác vụ xử lý ngôn ngữ tự nhiên như phân tích văn bản, dịch máy và truy xuất thông tin.



Tổng quan Part-of-speech Tagging

POS Tags cung cấp thông tin quan trọng về cấu trúc và ý nghĩa của một câu.



- Giúp làm rõ ý nghĩa của từ và giải quyết các mâu thuẫn cú pháp.
- Phân loại từ loại cho phép xác định danh từ, động từ, tính từ, trạng từ, đại từ, giới từ và các loại từ loại khác.



Tổng quan Part-of-speech Tagging

Why not learn something new?

WRB RB VB NN ADJ

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



PART-OF-SPEECH TAGGING

I. Tổng quan về Part-Of-Speech Tagging

II. Markov Chains

III. Markov Models

IV. Giải thuật Viterbi

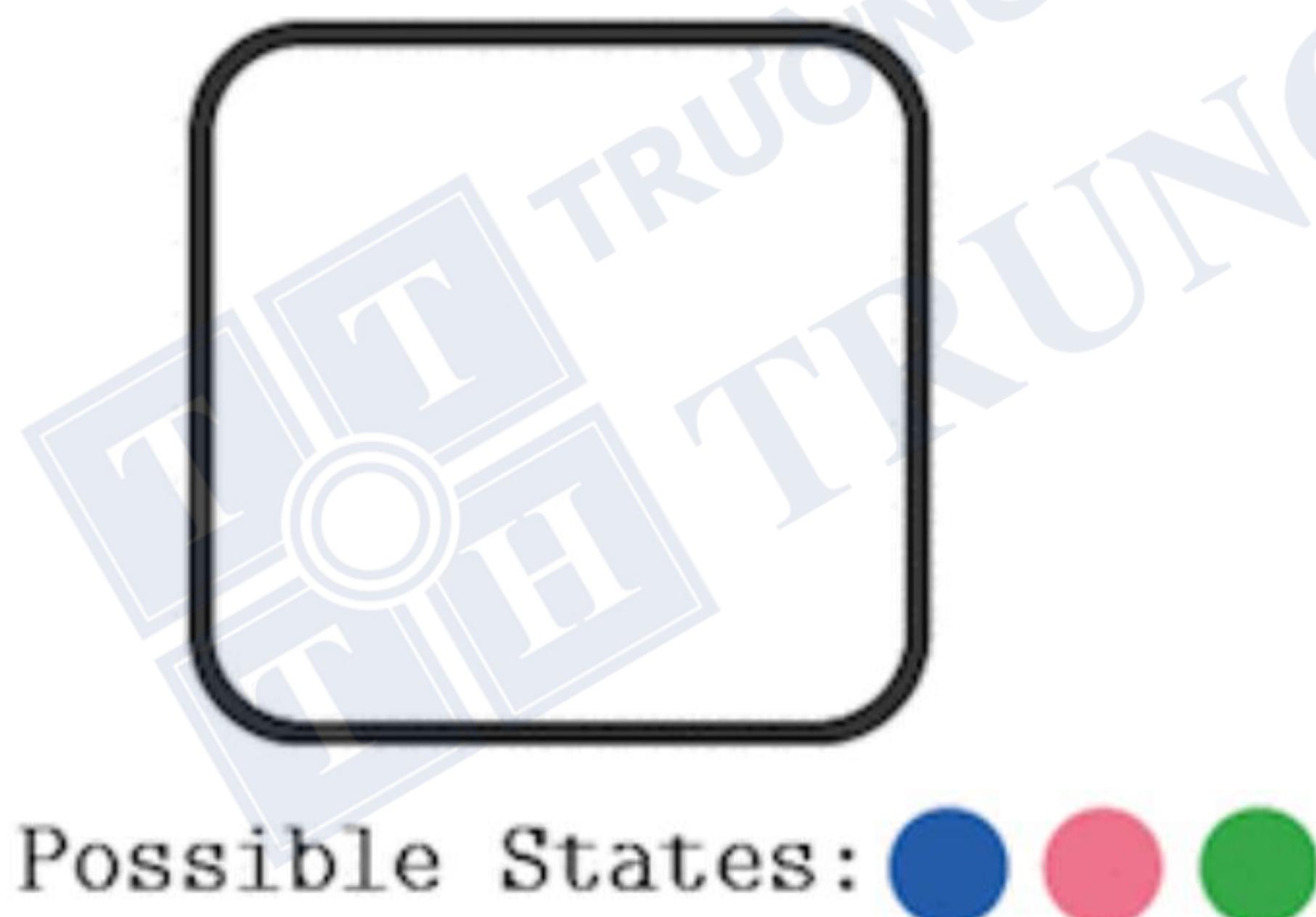


Tổng quan Markov Chains

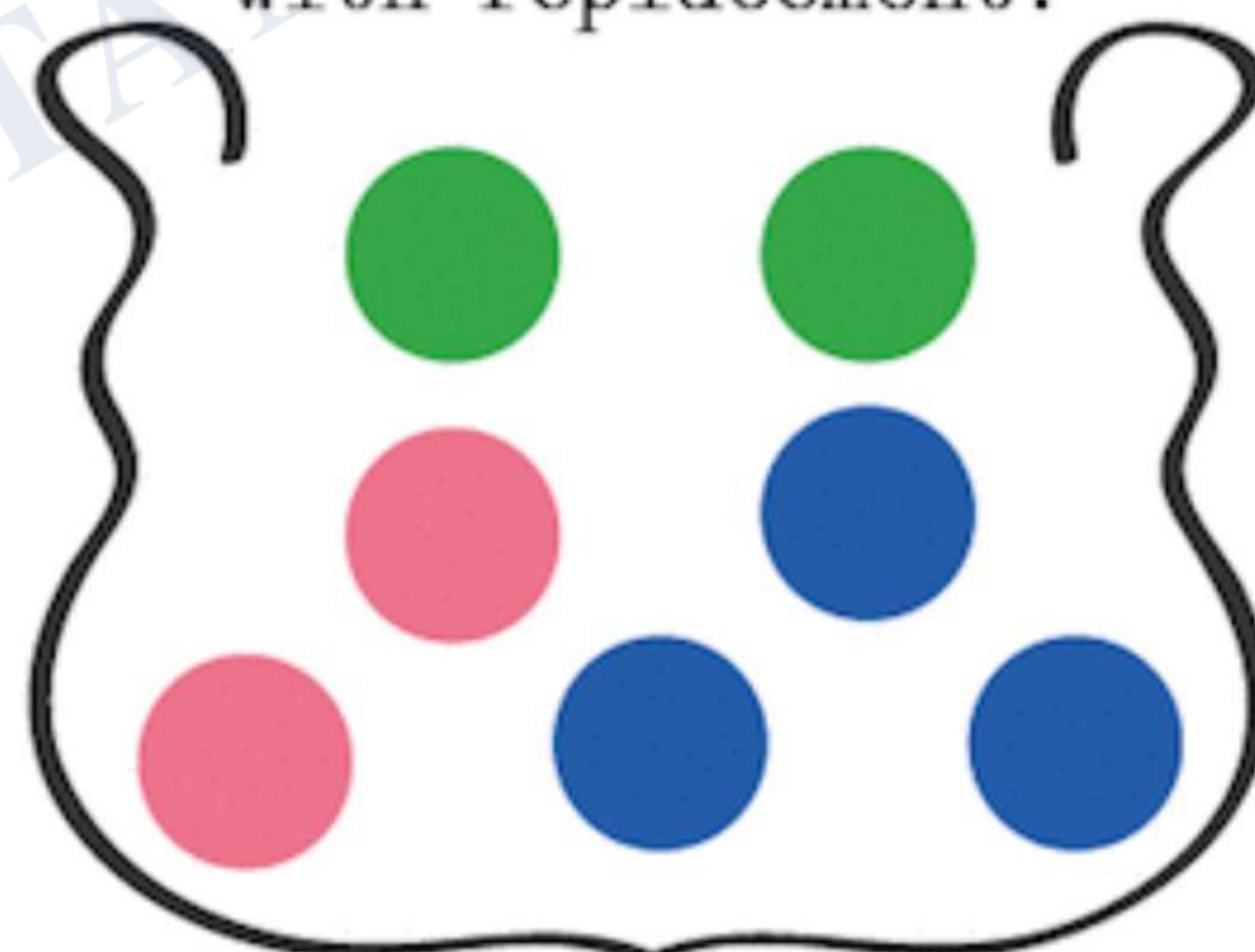
Markov Chains là mô hình hóa một hệ thống ngẫu nhiên qua việc xác định xác suất chuyển đổi từ trạng thái này sang các trạng thái khác nhau theo thời gian.

Markov Chain

Random Variable



Bag of Balls
With replacement!





Tổng quan Markov Chains

MARKOV PROPERTY

là xác suất chuyển đổi chỉ phụ thuộc vào trạng thái hiện tại mà không phụ thuộc vào quá khứ.



Điều này giúp đơn giản hóa quy trình tính toán:

- Không dùng cho các quá trình có yêu cầu quá khứ hay thông tin nằm ngoài state hiện tại.
- Dùng mô hình phức tạp hơn: **Hidden Markov Model, ...**



Tổng quan Markov Chains

Markov Chains được sử dụng để xác định xác suất của từ tiếp theo trong một chuỗi
→ Dự đoán từ có xác suất cao nhất.

Why not learn something ?

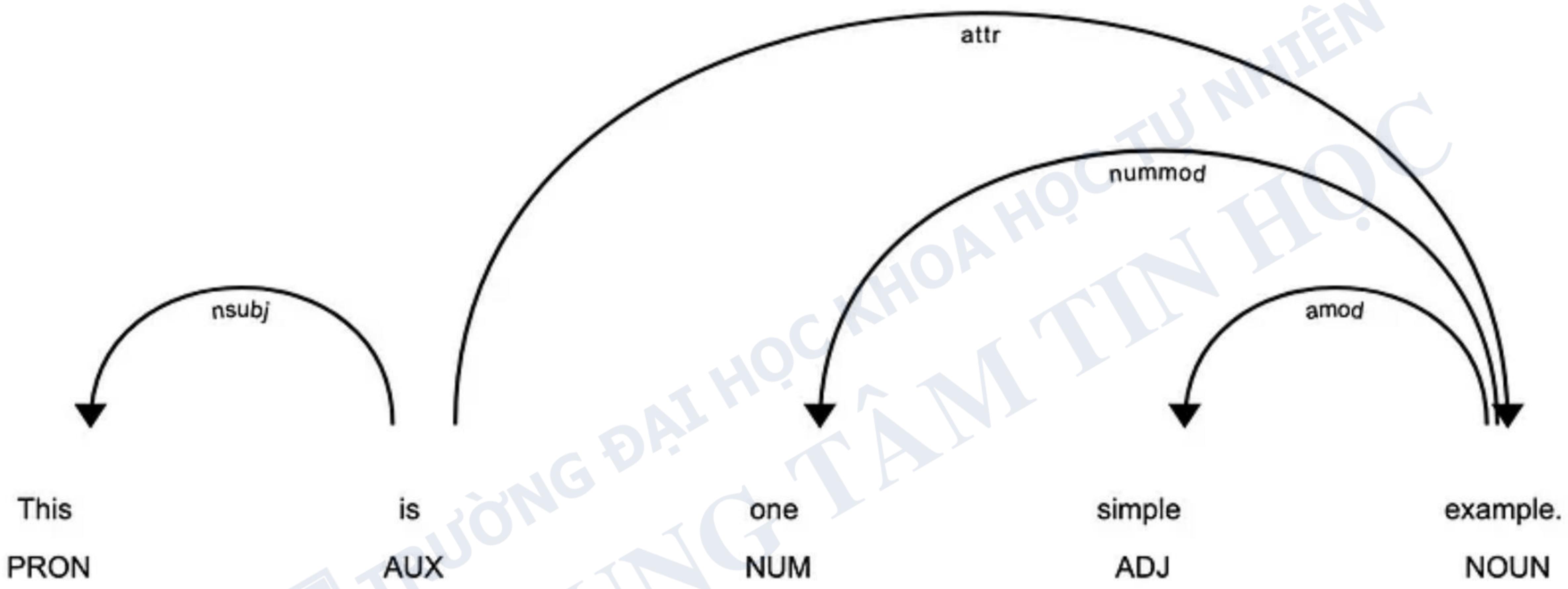
VB NN

→ Xác suất 1 NN
theo sau chữ
learn cao hơn

Why not learn swimming ?

VB VB

Tổng quan Markov Chains



Markov chains thể hiện mối quan hệ giữa **POS tags** của các từ trong chuỗi.

→ POS tags của từ tiếp theo.



PART-OF-SPEECH TAGGING

I. Tổng quan về Part-Of-Speech Tagging

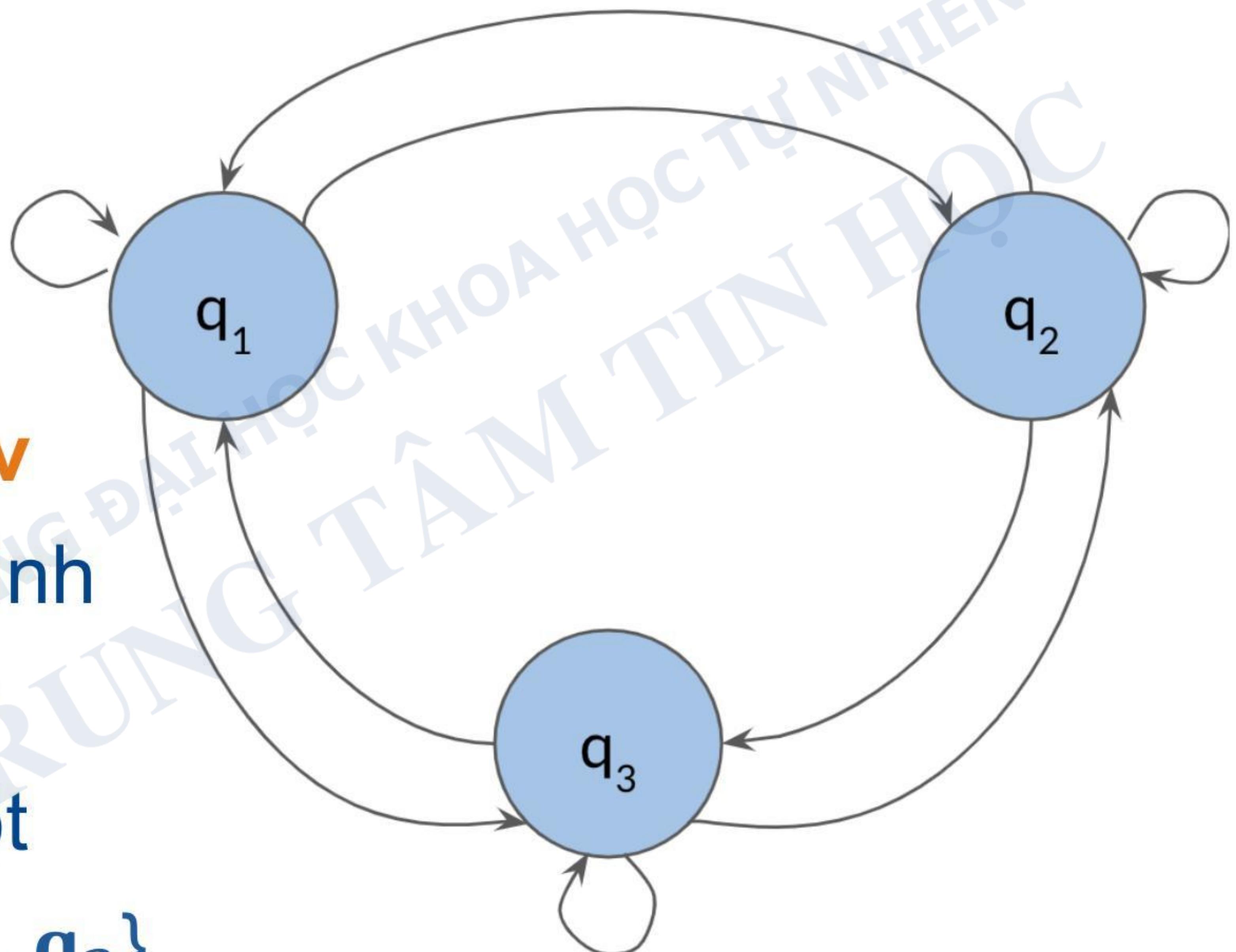
II. Markov Chains

III. Markov Models

IV. Giải thuật Viterbi

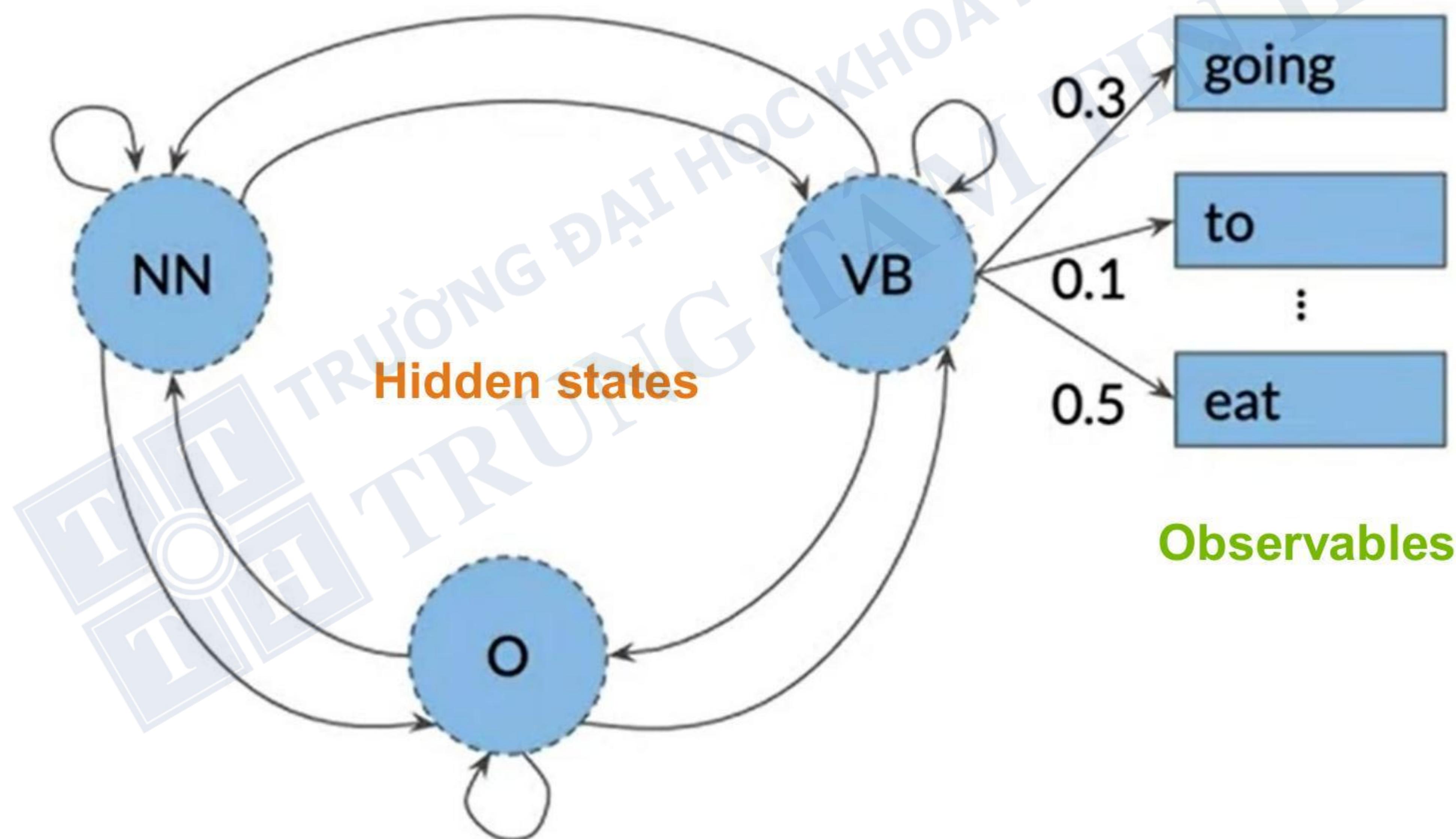
Mô hình Simple Markov Model

Simple Markov Model là mô hình có một Markov chains, hay một state $Q=\{q_1, q_2, q_3\}$



Mô hình Simple Markov Model

Hidden Markov Model là mô hình gồm các Markov chains (state) quan sát được và Markov chains ẩn.



Mô hình Hidden Markov Model

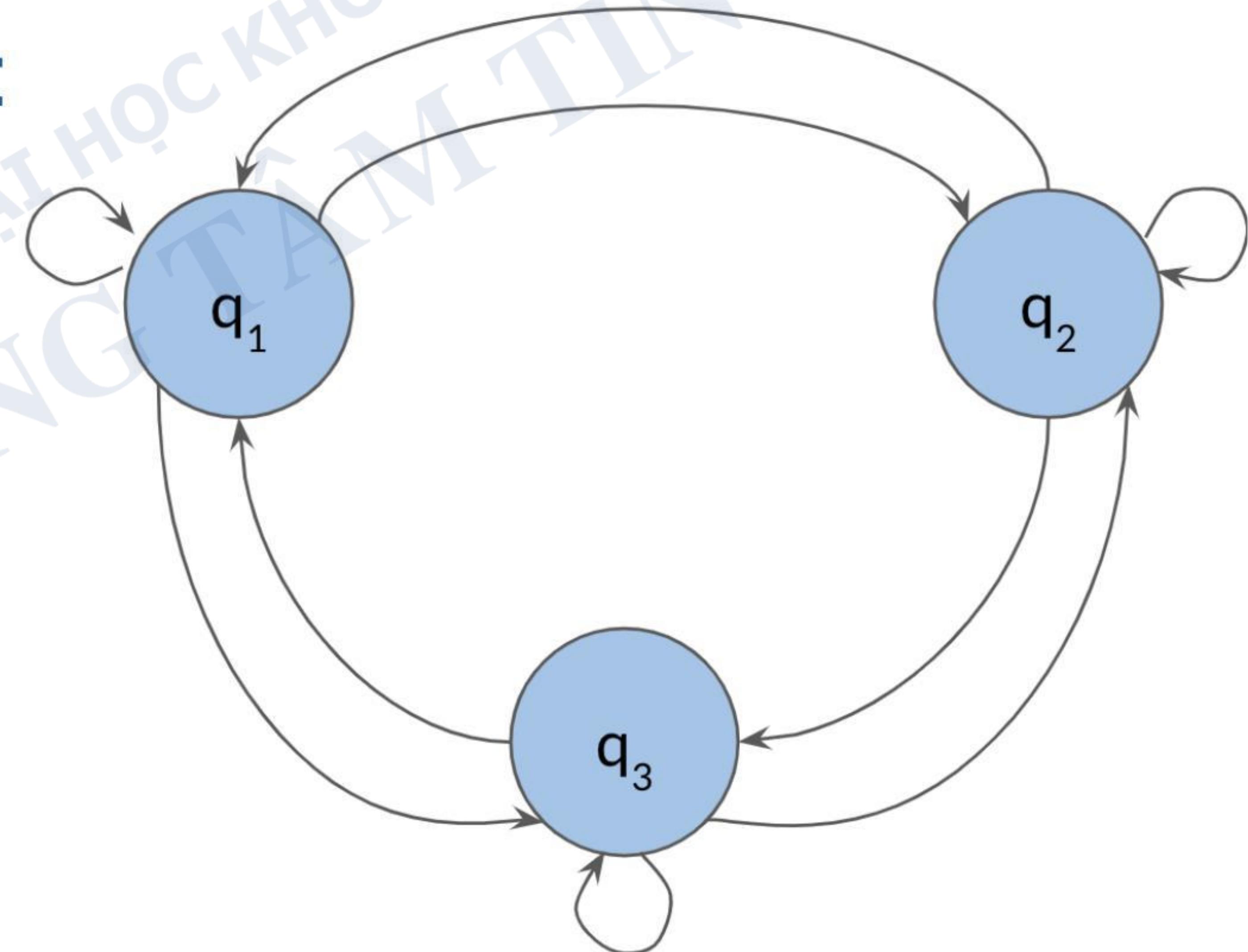
Lập mô hình Hidden Markov:

1. $Q=\{q_1, q_2, q_3\}$ là các state trong mô hình.
2. **Transition matrix:** xác suất của các POS tags.
3. **Emission matrix:**

xác suất của các từ.

State q_1 có:

- $P(q_1, q_1)$
- $P(q_1, q_2)$
- $P(q_1, q_3)$





Tổng quan Hidden Markov Model

Bước 1:
Xác định các state
(POS tags/ từ)

Bước 2:
Tính ma trận xác
suất của **POS**
tags

Bước 3:
Tính ma trận xác
suất của **từ**
(word)

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

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

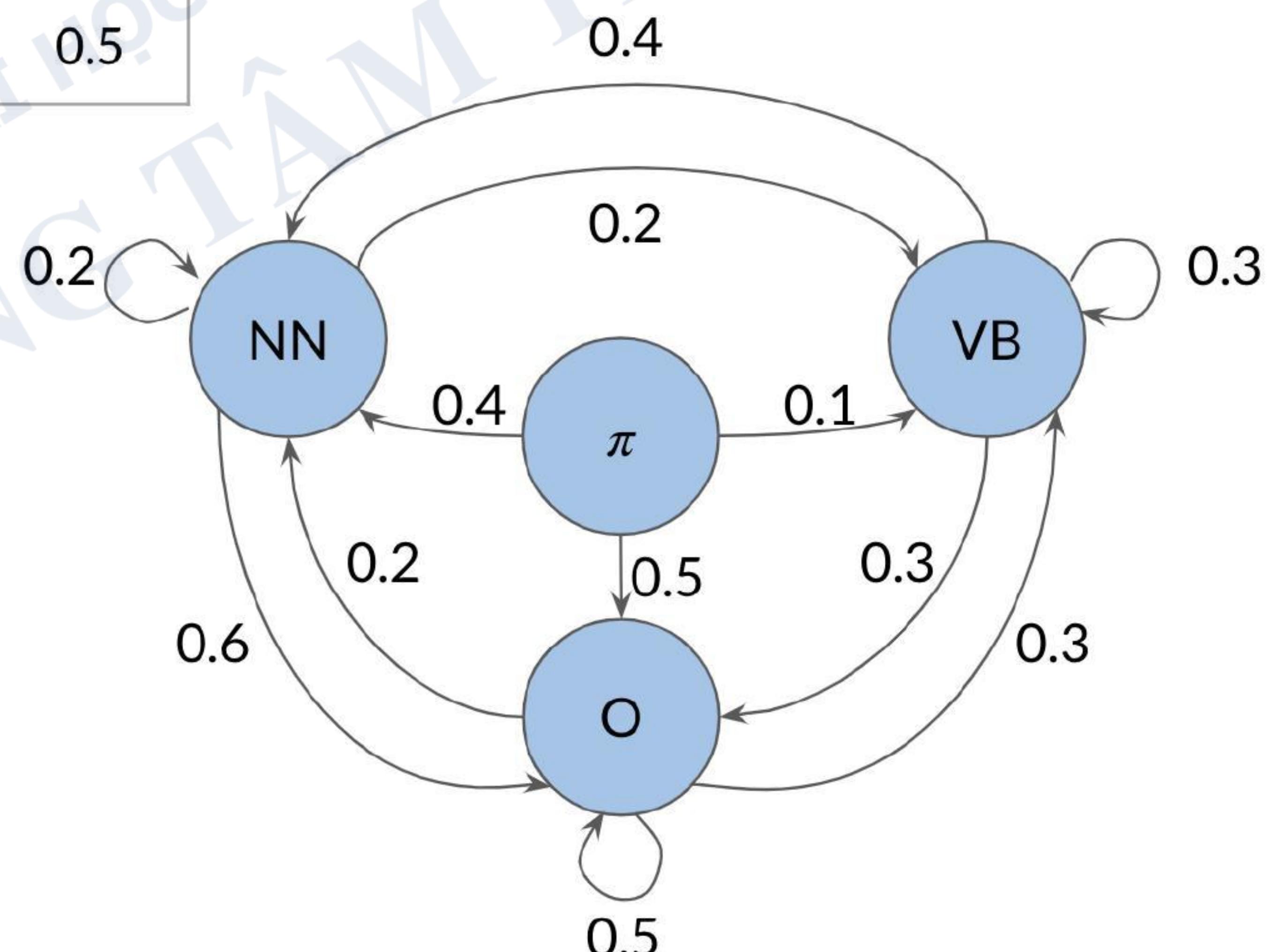
Mô hình Hidden Markov Model

	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

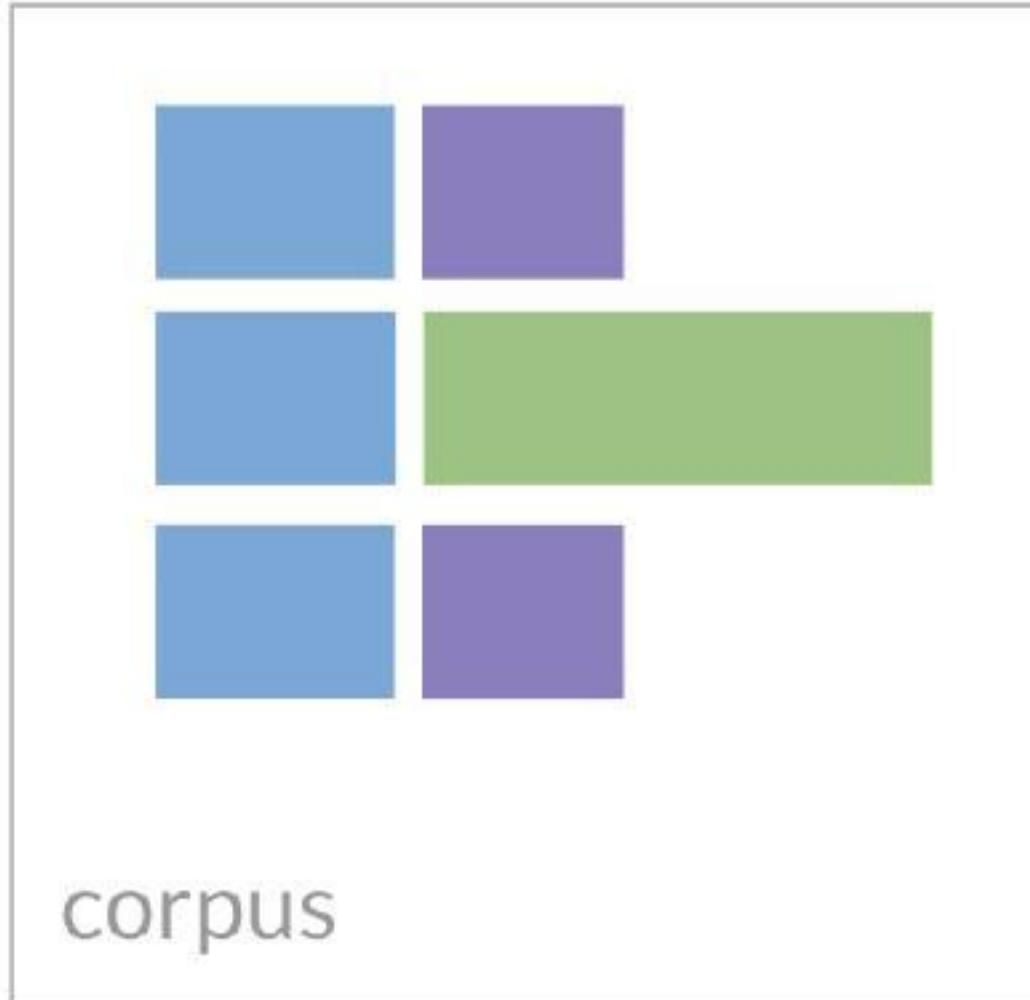
$A =$

= 1

Tính transition
matrix



Mô hình Hidden Markov Model



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

- Đếm số lần 1 cặp POS tags đi cùng nhau.

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

- Tính xác suất của cặp POS tags đó.

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



Mô hình Hidden Markov Model

<S> in a station of the metro

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

<S> petals on a wet, black bough.

$$A = \begin{array}{c|ccc} & \text{NN} & \text{VB} & \text{O} \\ \hline \pi & 1 & 0 & 2 \\ \text{NN (noun)} & 0 & 0 & 6 \\ \text{VB (verb)} & 0 & 0 & 0 \\ \text{O (other)} & 6 & 0 & 8 \end{array} C(t_{i-1}, t_i)$$



Mô hình Hidden Markov Model

<S> in a station of the metro

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

<S> petals on a wet, black bough.

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

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



Smoothing

$$P(NN|NN) = 0$$

→ Dùng Smoothing để tránh mô hình bị triệt tiêu

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

$A =$

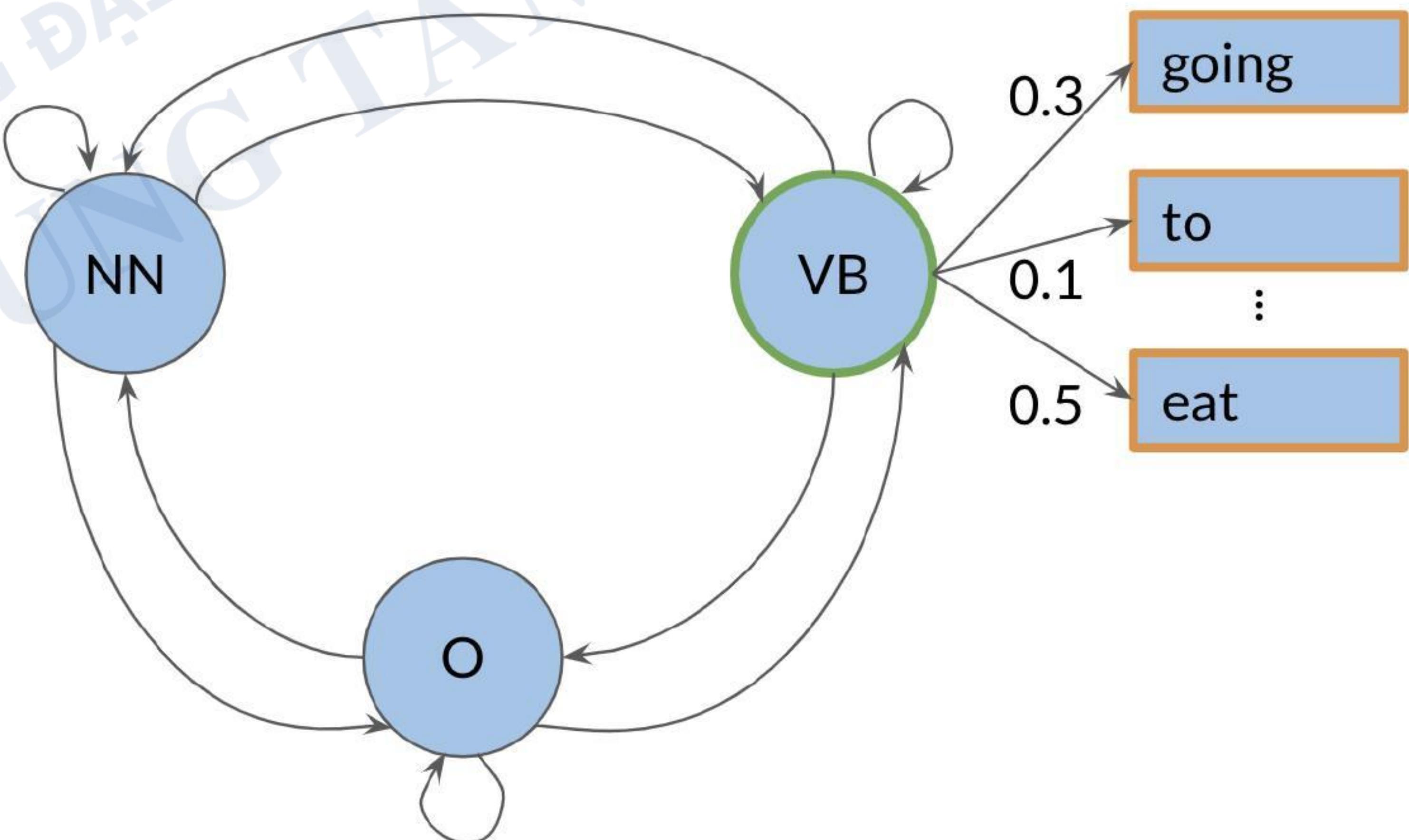
	NN	VB	O	
π	$1+\epsilon$	$0+\epsilon$	$2+\epsilon$	$3+3^*\epsilon$
NN	$0+\epsilon$	$0+\epsilon$	$6+\epsilon$	$6+3^*\epsilon$
VB	$0+\epsilon$	$0+\epsilon$	$0+\epsilon$	$0+3^*\epsilon$
O	$6+\epsilon$	$0+\epsilon$	$8+\epsilon$	$14+3^*\epsilon$

Mô hình Simple Markov Model

$B =$

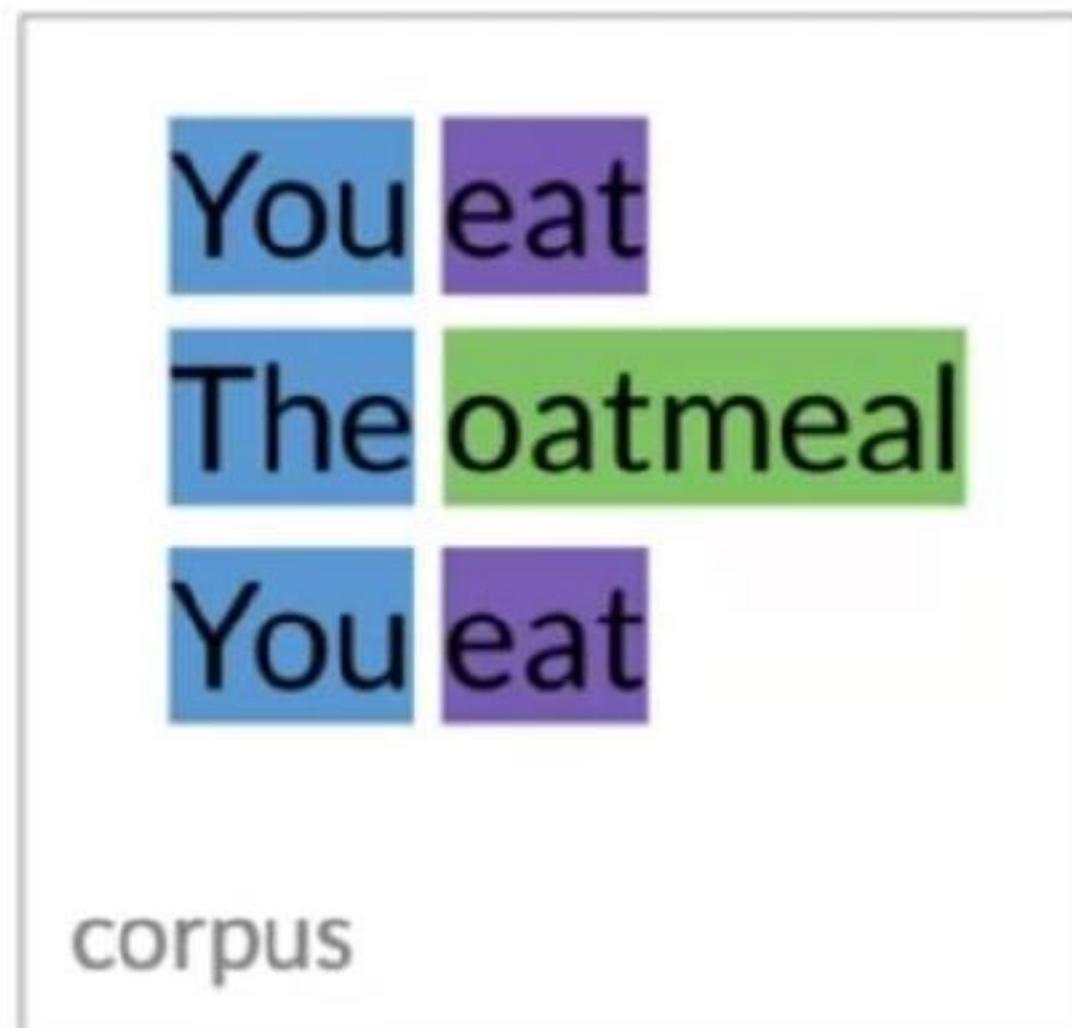
	going	to	eat	...
NN (noun)	0.5	0.1	0.02	
VB (verb)	0.3	0.1	0.5	
O (other)	0.3	0.5	0.68	

Tính emission
matrix





Mô hình Simple Markov Model



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

1. Đếm số lần từ w thuộc POS tag t

$$C(t_i, w_i)$$

2. Tính xác suất của từ w là tag t (đã smooth)

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



Mô hình Hidden Markov Model

<S> in a station of the metro

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

<S> petals on a wet, black bough.

$$B = \begin{array}{c|cccc} & \text{in} & \text{a} & \dots & \\ \hline \text{NN (noun)} & C(\text{NN}, \text{in}) & & & \\ \text{VB (verb)} & C(\text{VB}, \text{in}) & & & \\ \text{O (other)} & C(\text{O}, \text{in}) & & & \end{array} \quad C(t_i, w_i)$$



Mô hình Hidden Markov Model

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

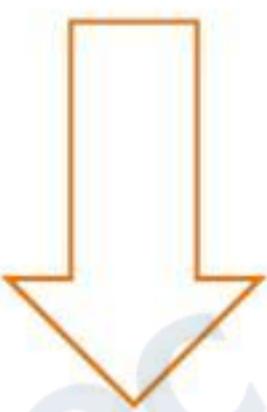
$$B = \begin{matrix} & \text{in} & \text{a} & \dots \\ \text{NN (noun)} & 0 & & \\ \text{VB (verb)} & 0 & & \\ \text{O (other)} & 2 & & \end{matrix} \quad C(t_i, w_i)$$

Tính emission matrix theo công thức $P(w_i | t_i)$



Ví dụ Giải thuật Viterbi

One fish two fish red fish blue fish

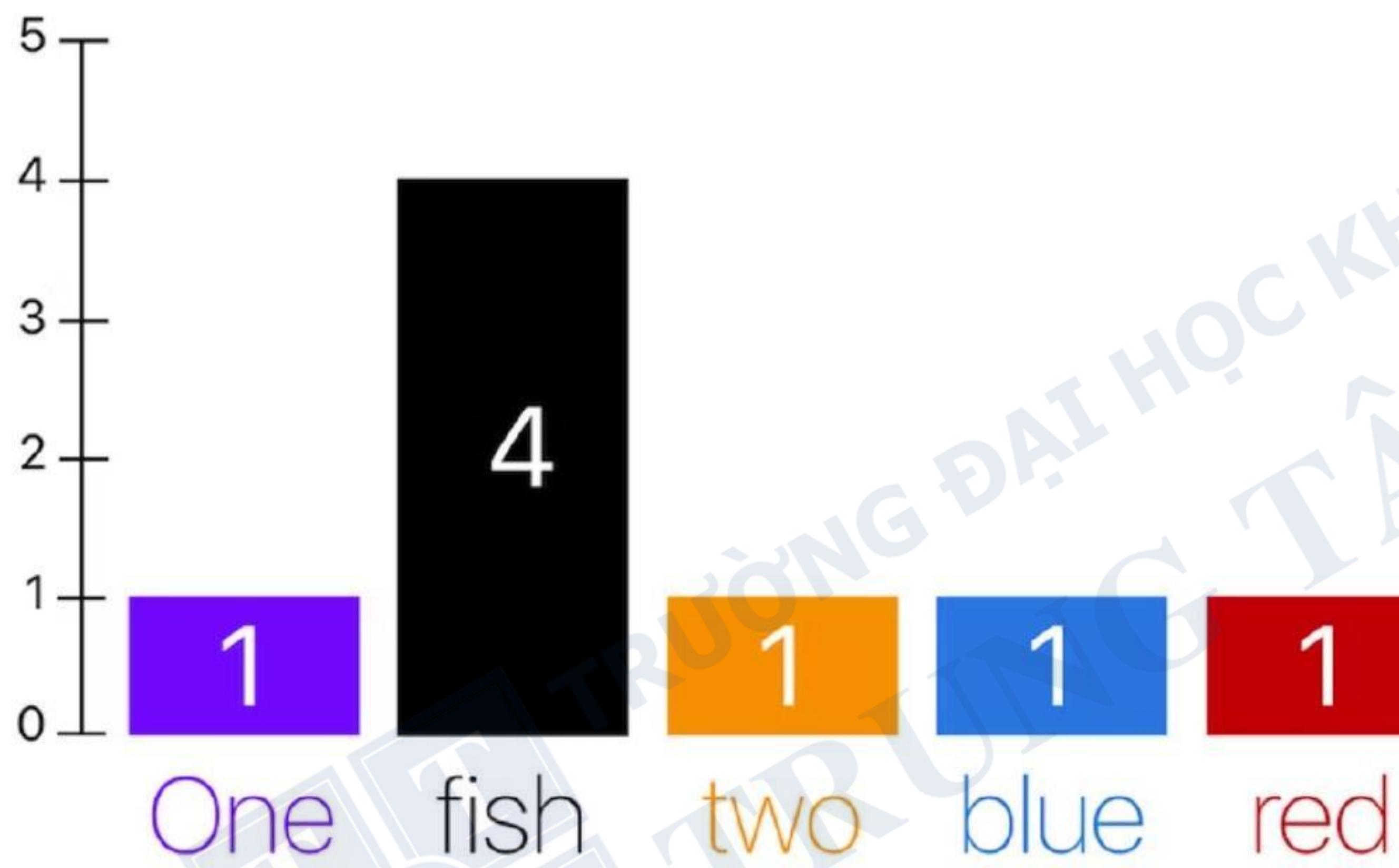


One fish **two** fish **red** fish **blue** fish





Giải thuật Viterbi



One	:	1
fish	:	4
two	:	1
red	:	1
blue	:	1

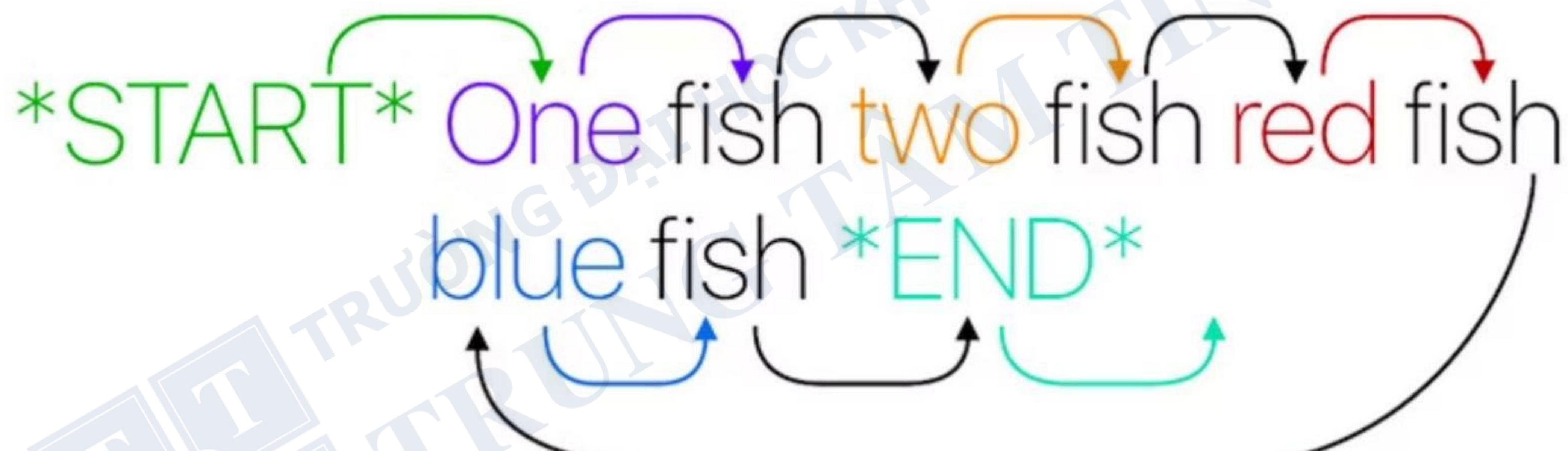


Ví dụ Giải thuật Viterbi

START One fish two fish red fish
blue fish *END*

START	:	1
One	:	1
fish	:	4
two	:	1
red	:	1
blue	:	1
END	:	1

Neural Network & Linear Regression





Ví dụ Neural Network & Linear Regression

(*Start*, One)

(One, fish)

(fish, two)

(two, fish)

(fish, red)

(red, fish)

(fish, blue)

(blue, fish)

(fish, *END*)

(*END*, none)

(*Start*, One)

(One, fish)

(fish, two) (fish, red) (fish, blue) (fish, *END*)

(two, fish)

(red, fish)

(blue, fish)

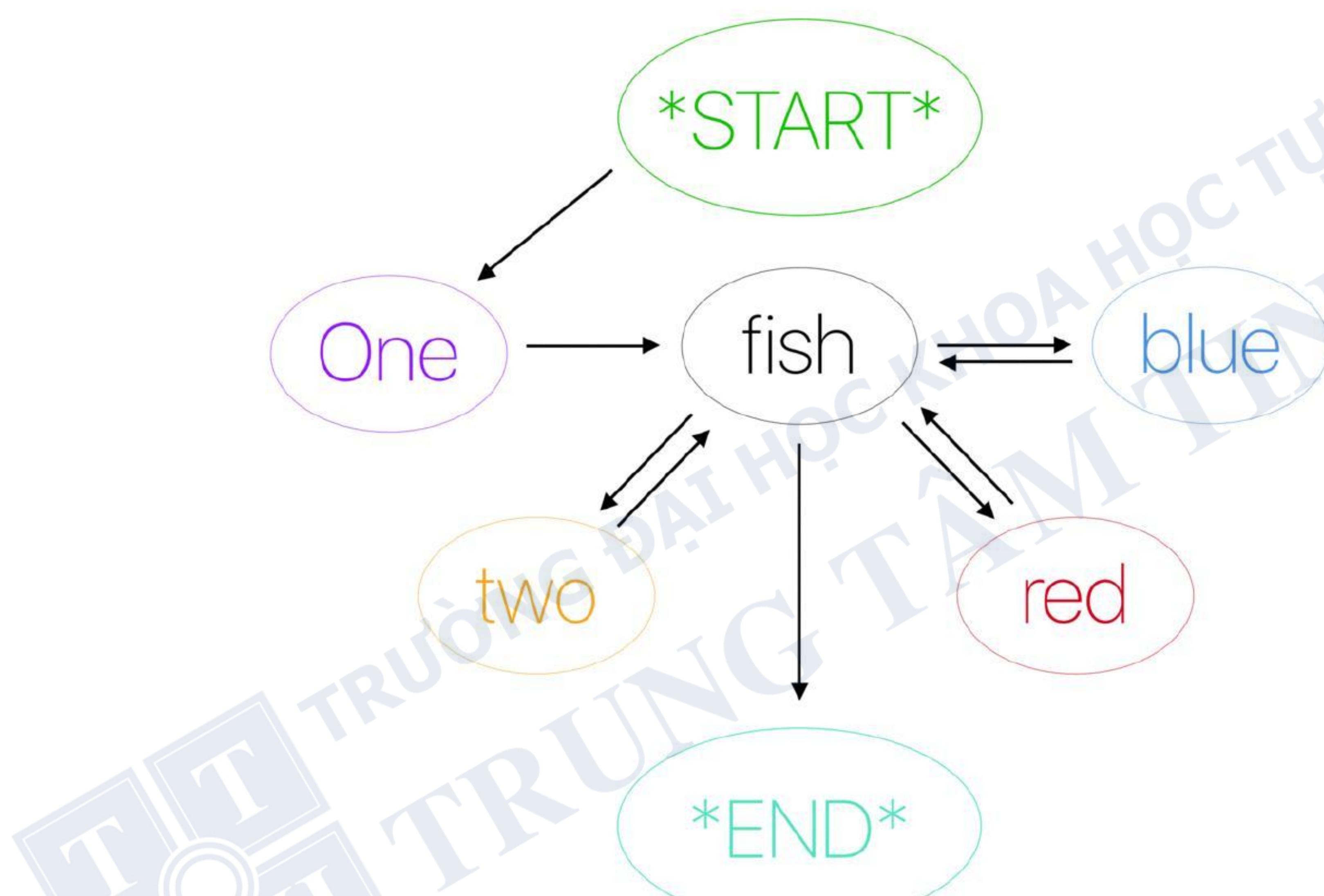
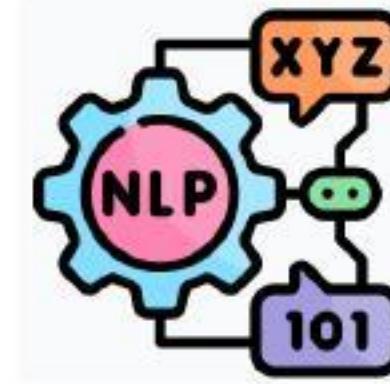
(*END*, none)



Bài toán Non-linear Regression

Start	:	[One]
One	:	[fish]
fish	:	[two, red, blue, *END*]
two	:	[fish]
red	:	[fish]
blue	:	[fish]
END	:	[none]

Bài toán Non-linear Regression





PART-OF-SPEECH TAGGING

I. Tổng quan về Part-Of-Speech Tagging

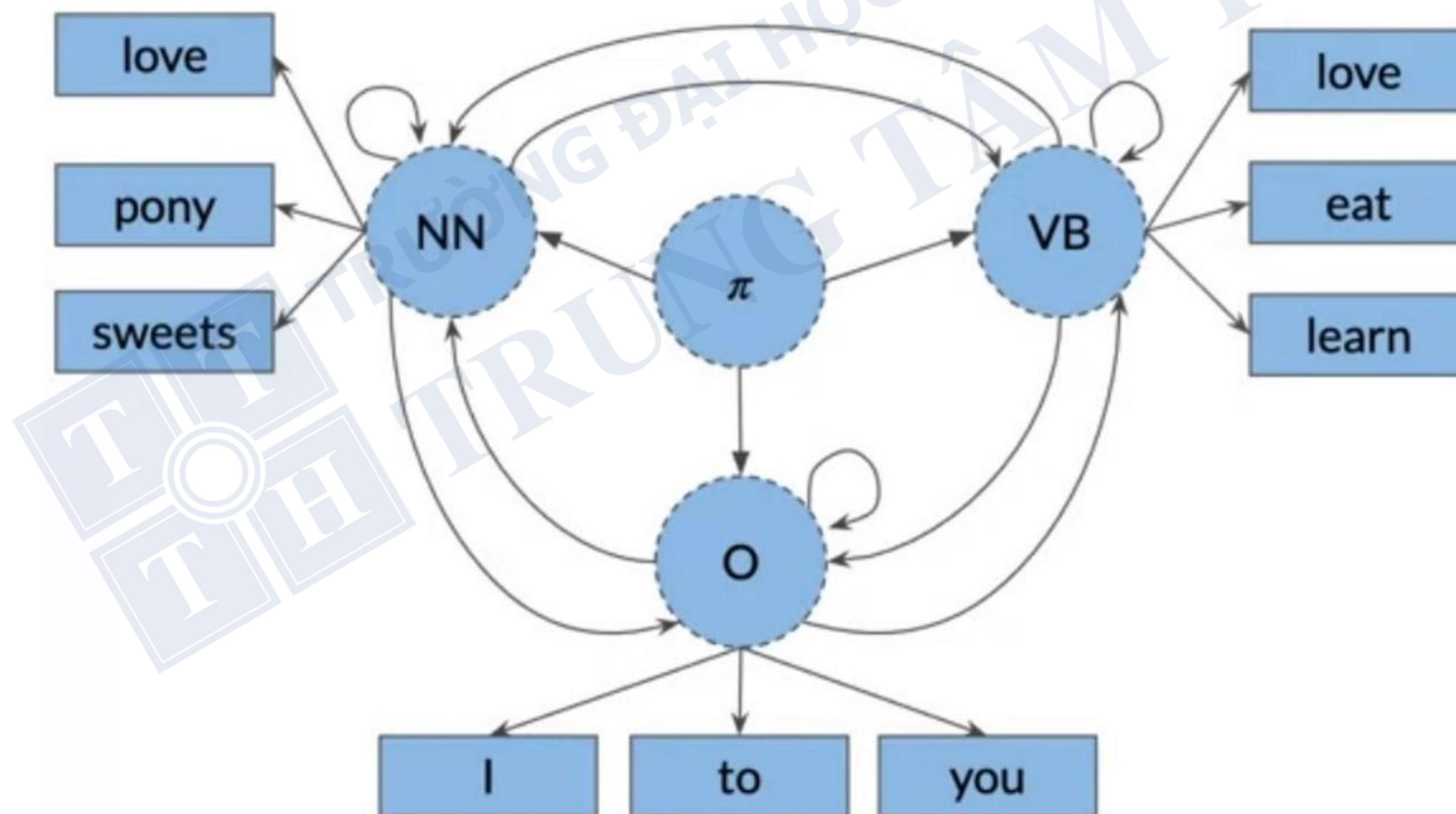
II. Markov Chains

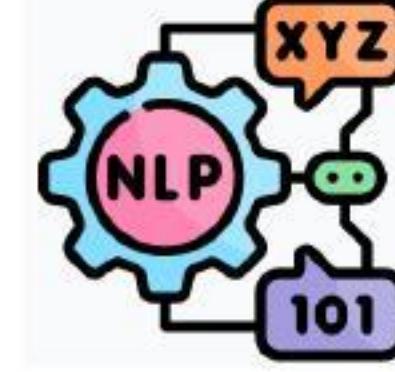
III. Markov Models

IV. Giải thuật Viterbi

Giải thuật Viterbi

Giải thuật Viterbi là giải thuật để tìm ra chuỗi trạng thái có xác suất cao nhất, cho trước một Hidden Markov Model.





Giải thuật Viterbi

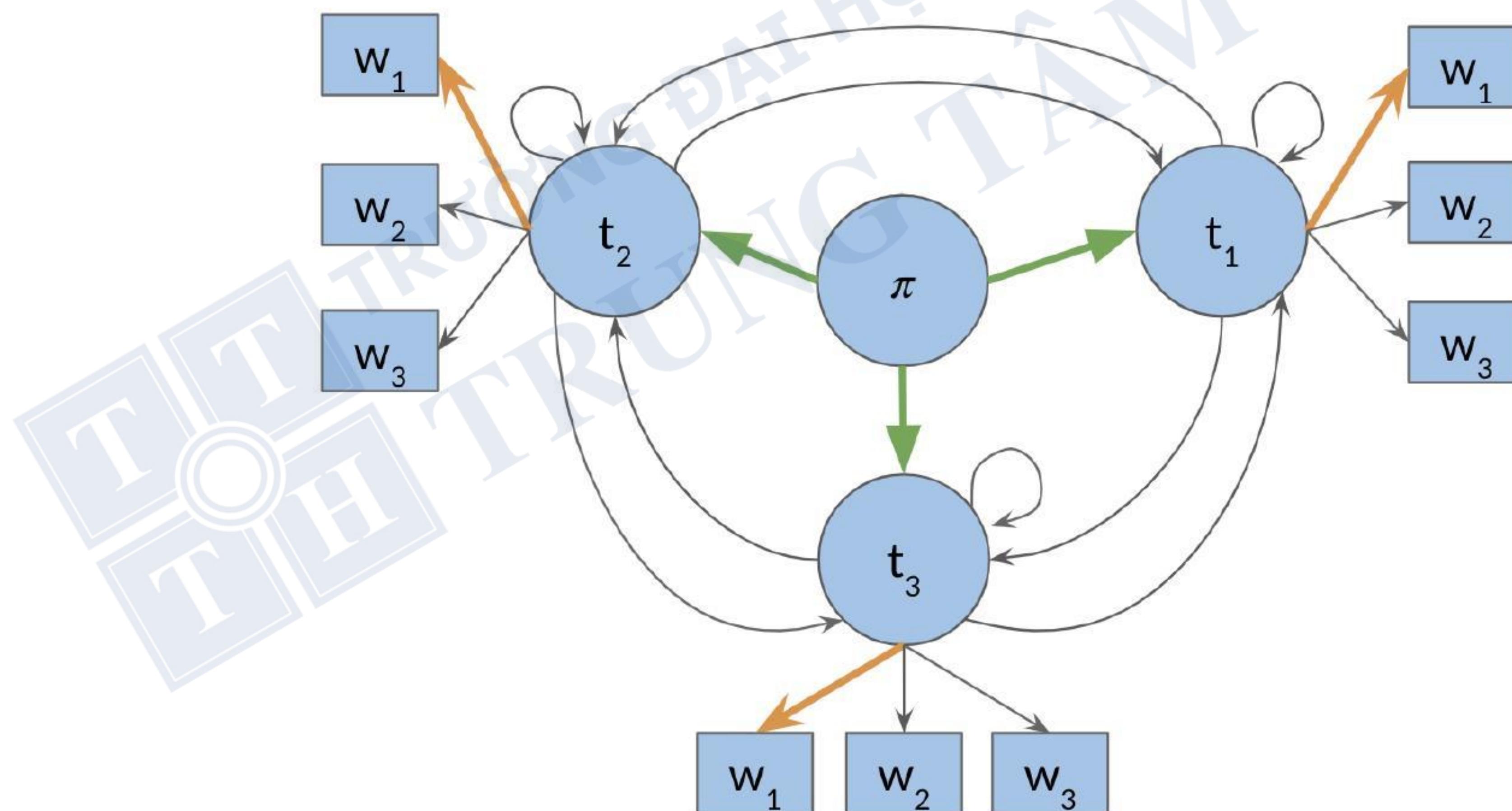
Chi tiết giải thuật Viterbi:

1. $Q=\{q_1, q_2, q_3\}$ là các state trong mô hình.
2. **Transition matrix A:** xác suất của các POS tags.
3. **Emission matrix B:** xác suất của các từ.
4. **Word | POS tag matrix C:** xác suất từ thuộc các POS tag.
5. **Index matrix D:** bảng index các từ và các POS tag tương ứng với từ.

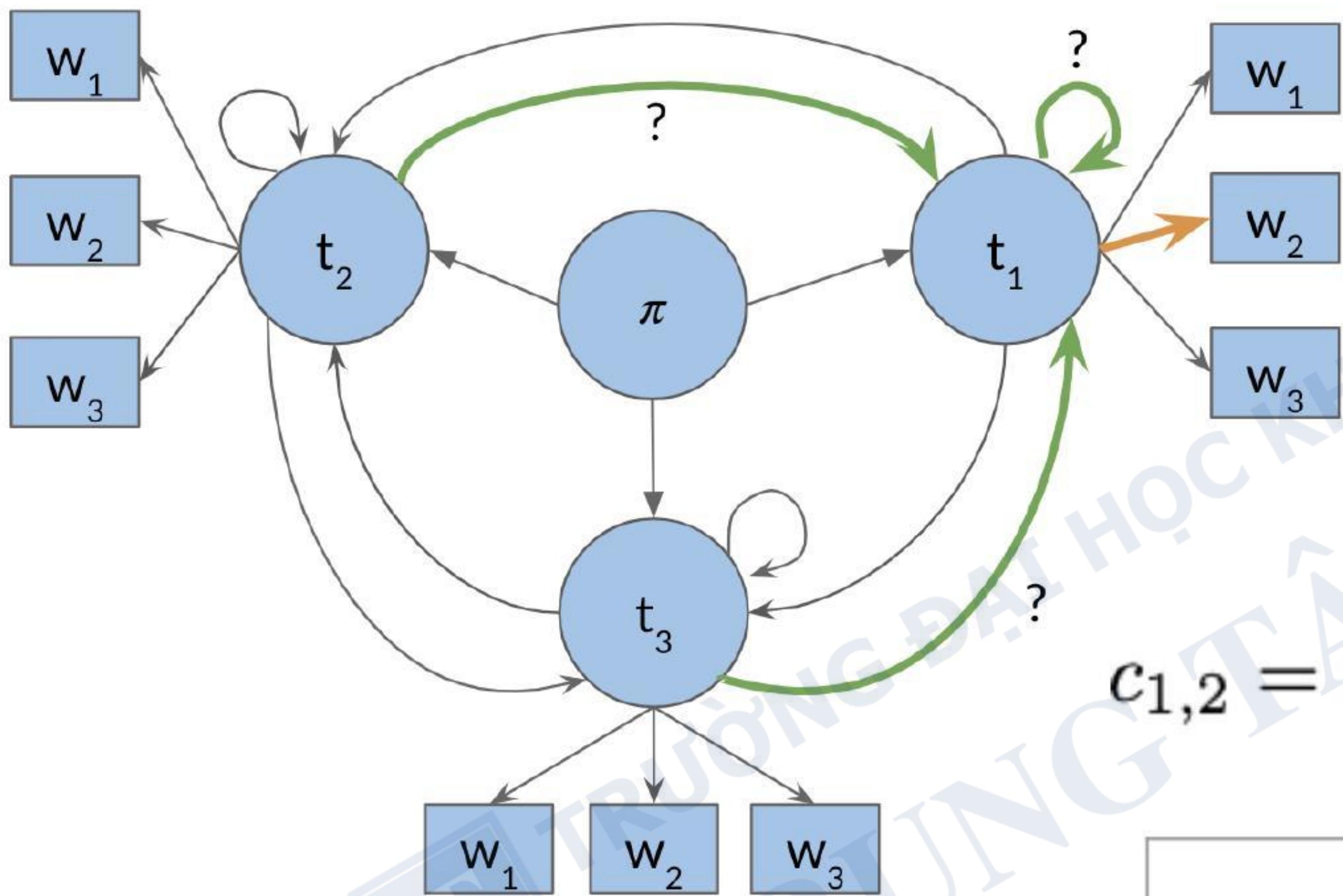
Giải thuật Viterbi - Tạo Matrix C

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

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



Matrix C – Forward Pass

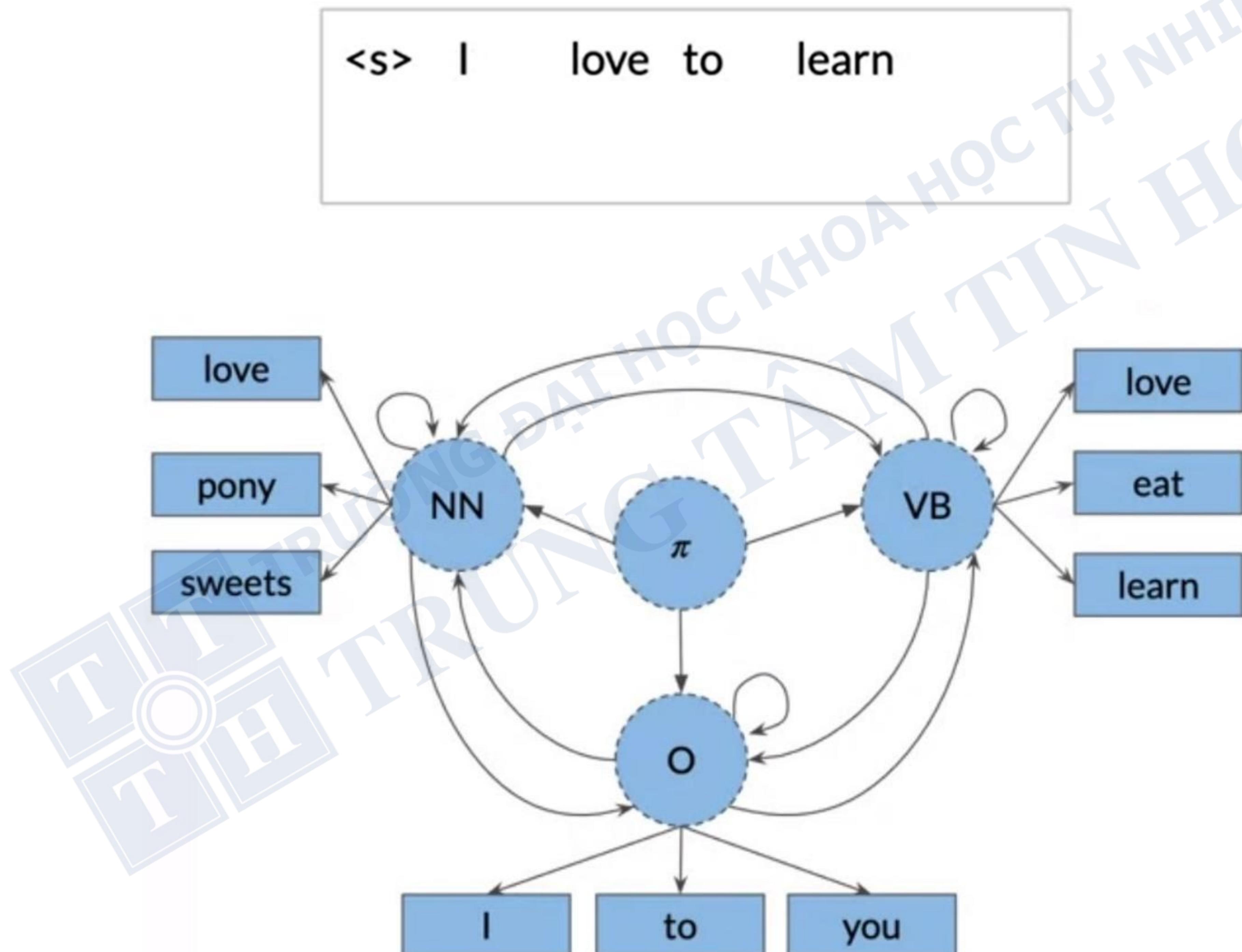


$$c_{1,2} = \max_k c_{k,1} * a_{k,1} * b_{1,cindex(w_2)}$$

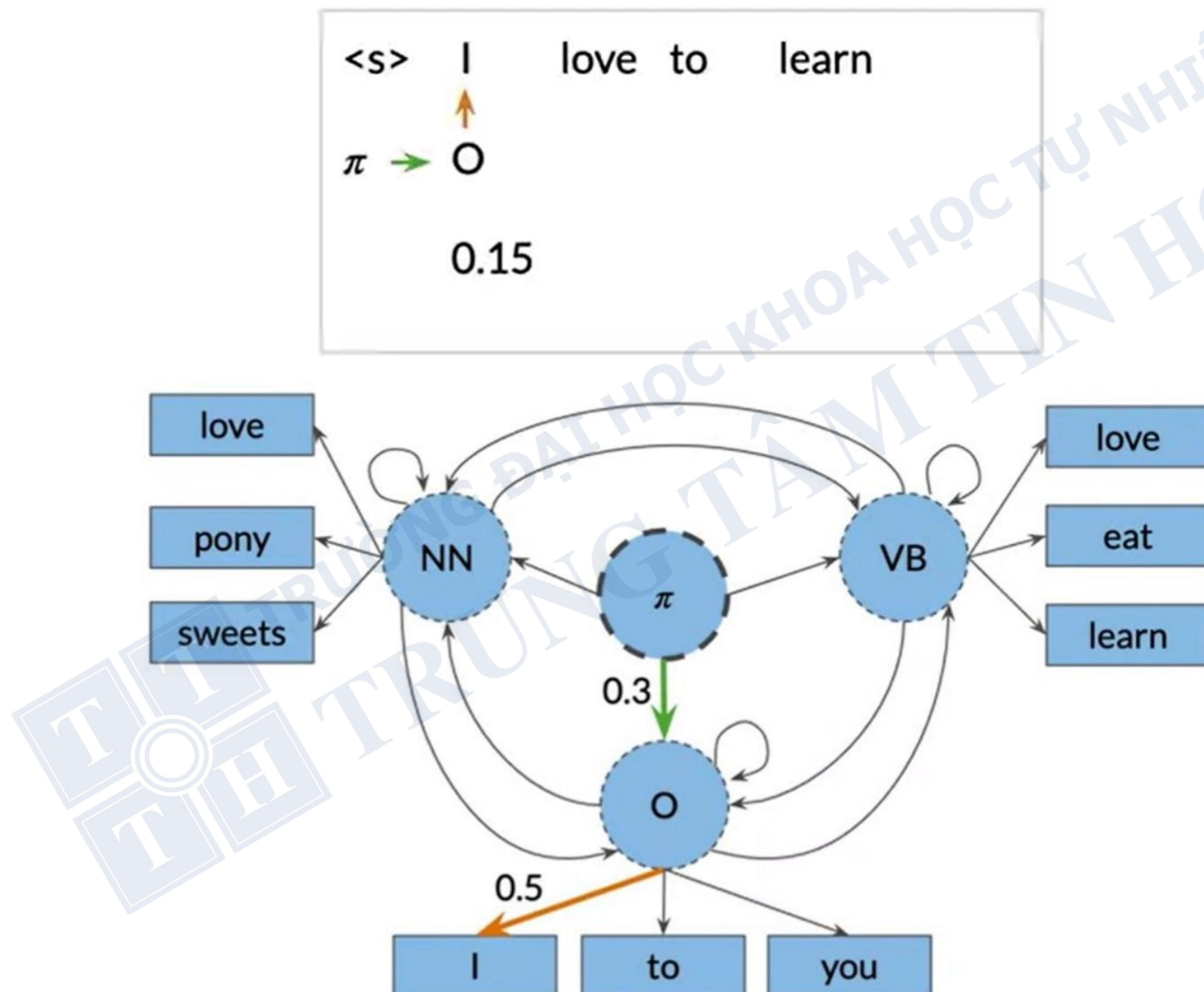
$C =$

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

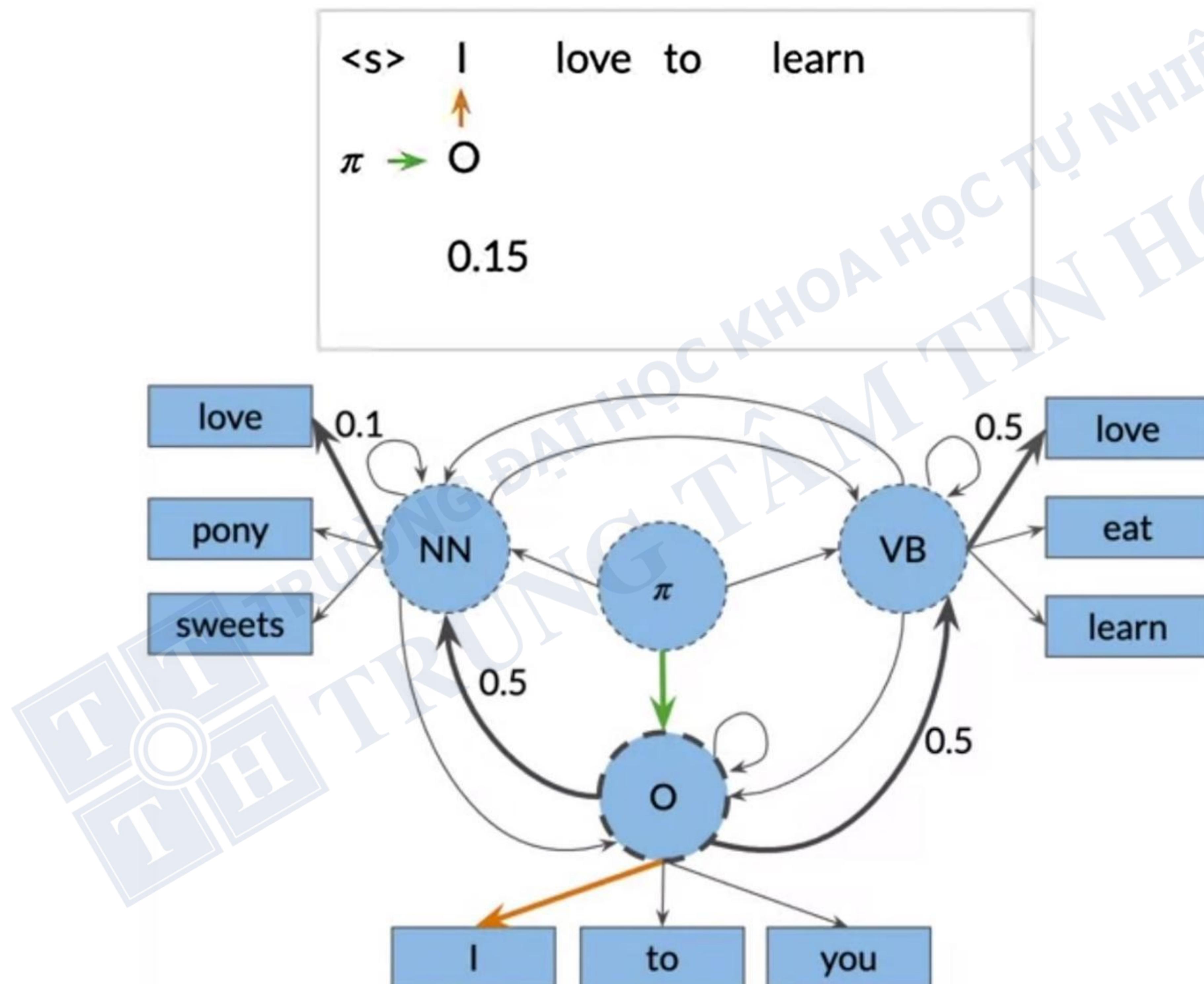
Giải thuật Viterbi - Tạo Matrix C



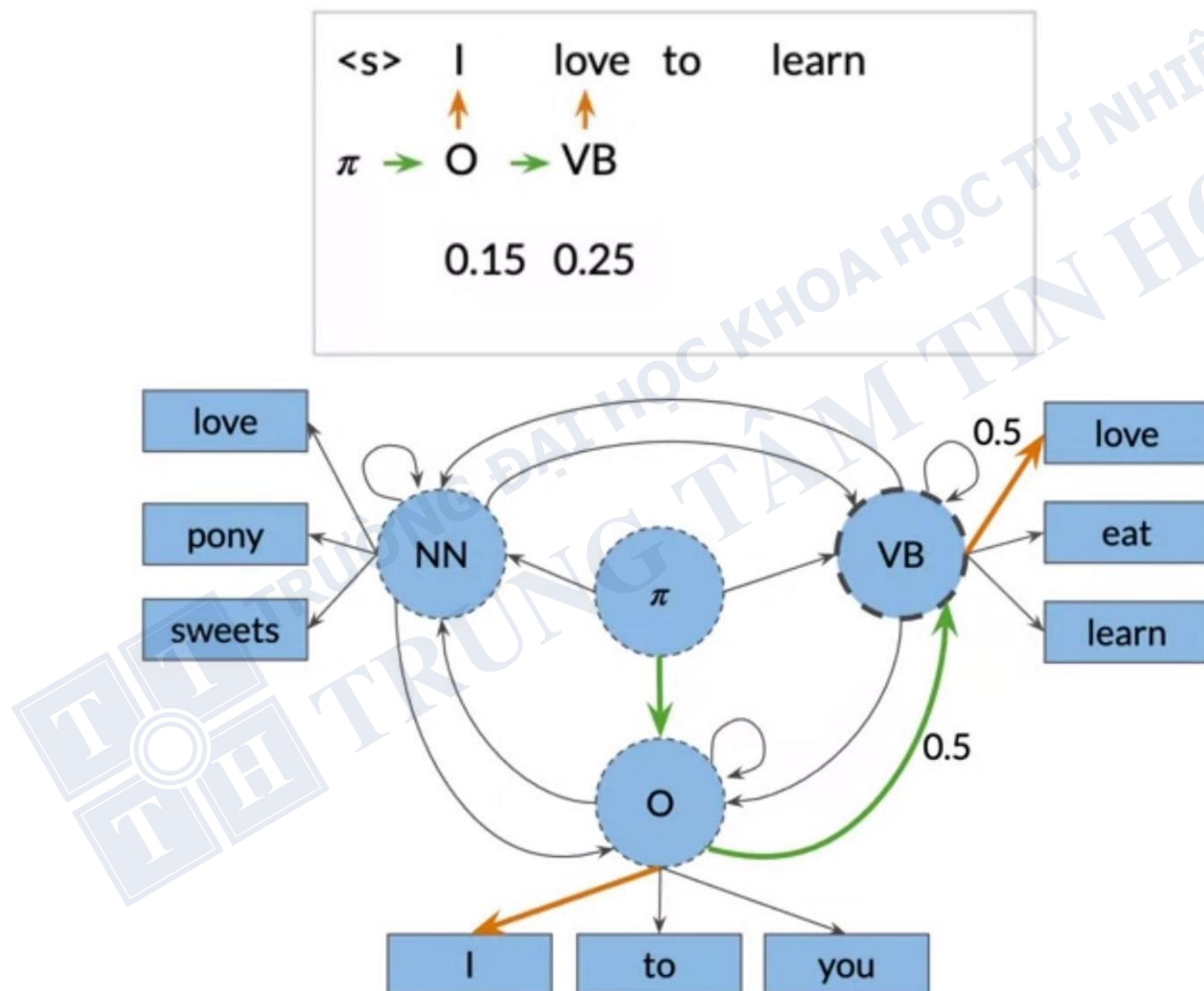
Giải thuật Viterbi - Tạo Matrix C



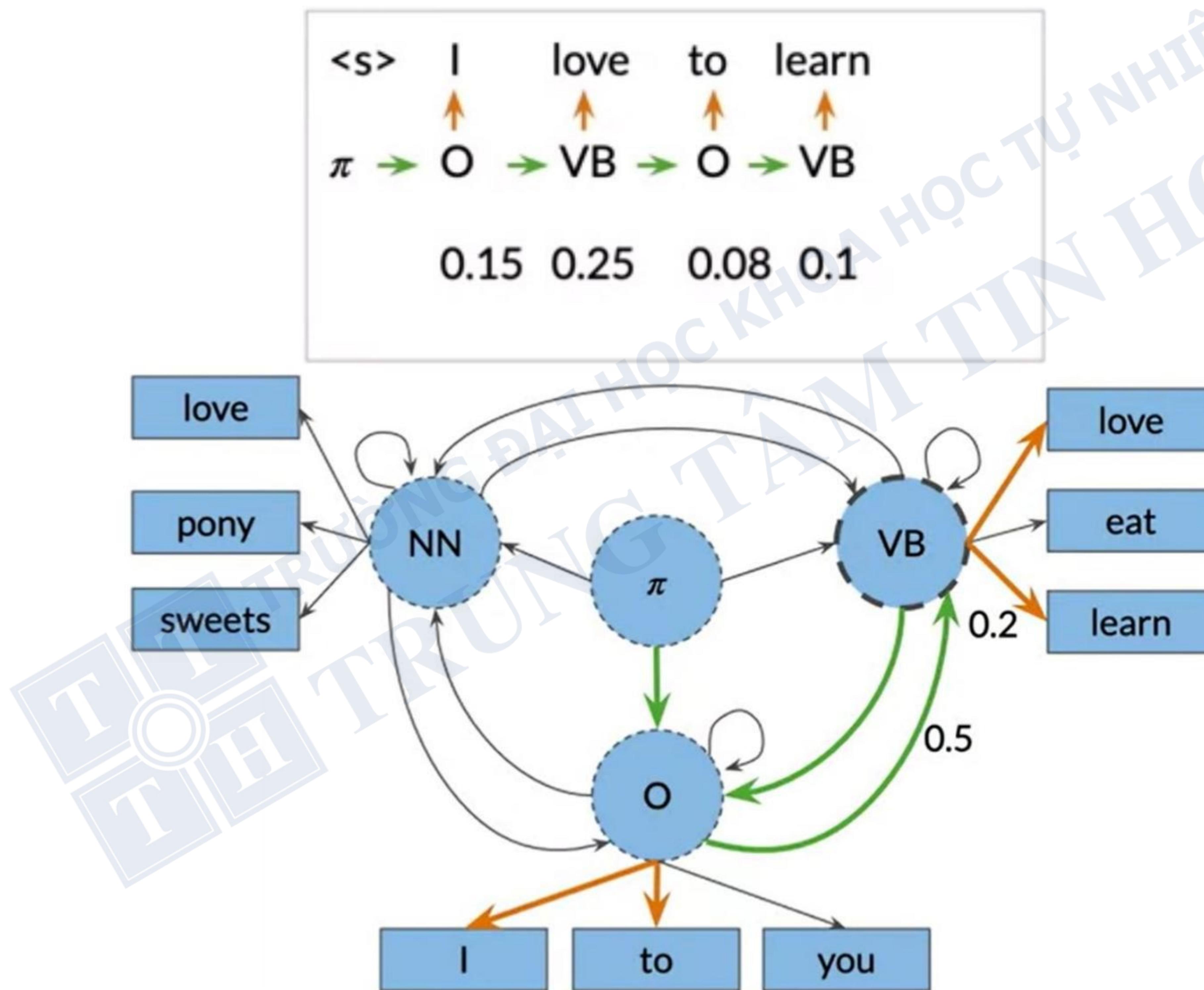
Giải thuật Viterbi - Tạo Matrix C



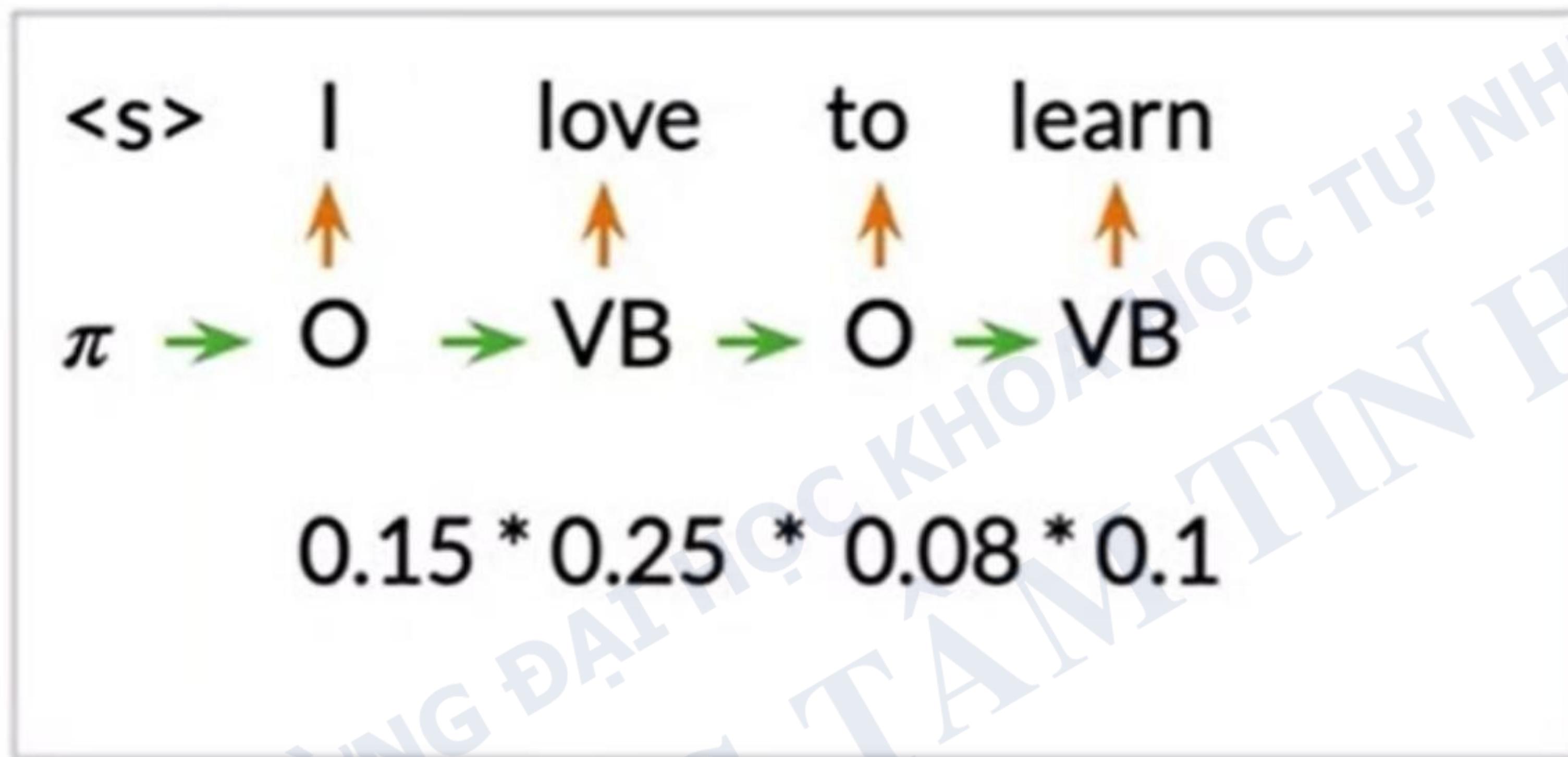
Giải thuật Viterbi - Tạo Matrix C



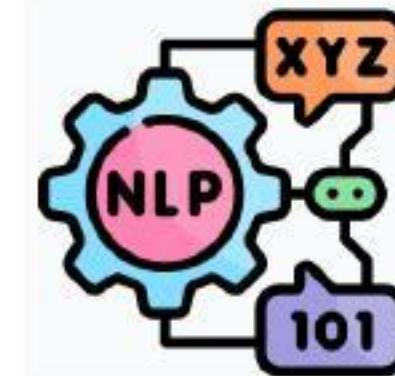
Giải thuật Viterbi - Tạo Matrix C



Mô hình Viberti - Tạo Matrix C



0.0003 là xác xuất của chuỗi trên.



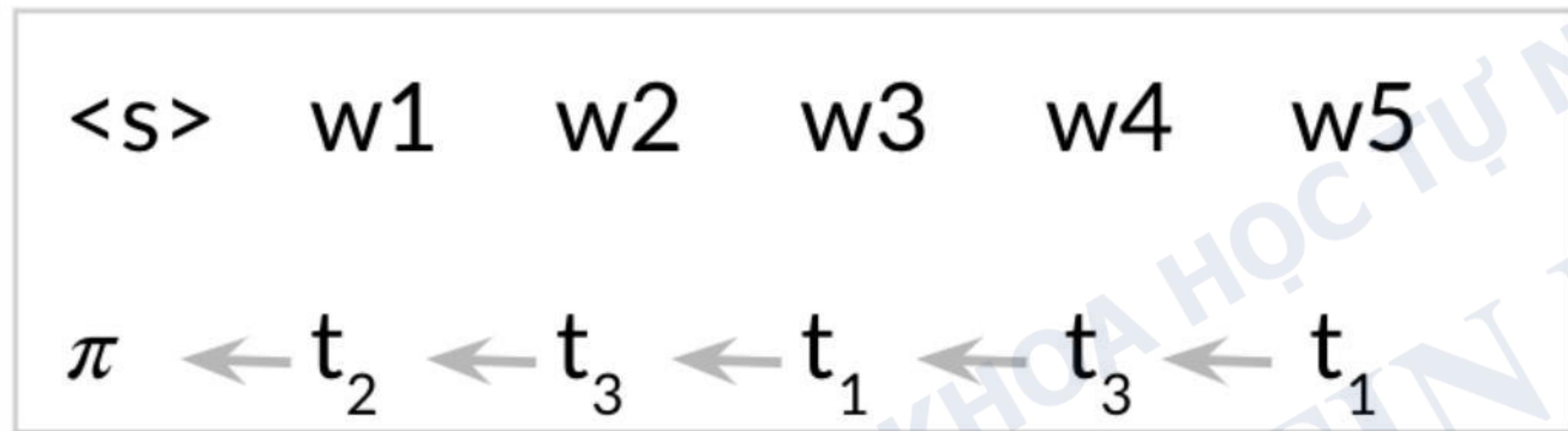
Matrix C – Backward Pass

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

$C =$

	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

Matrix D – Backward Pass



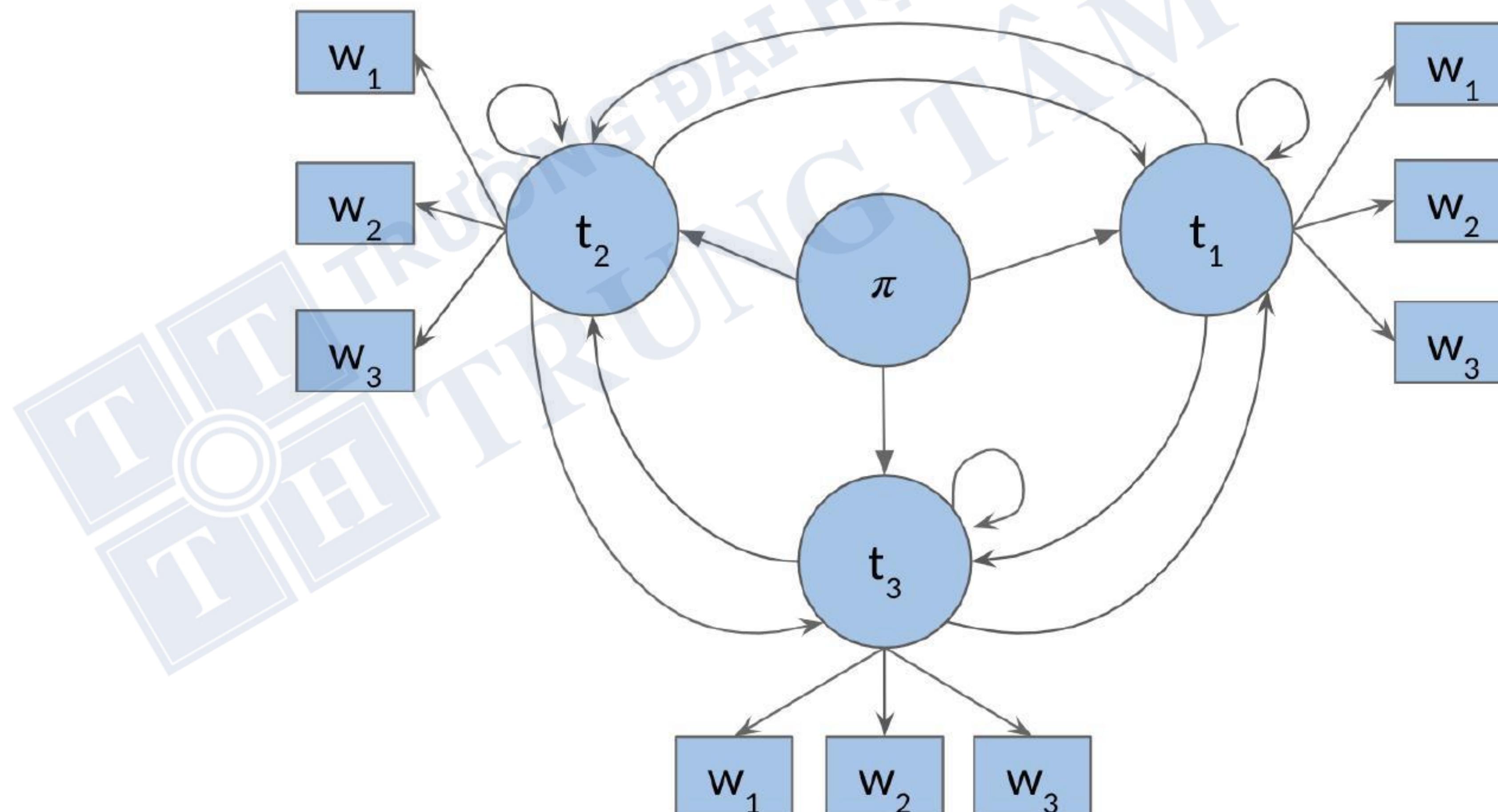
$D =$

	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

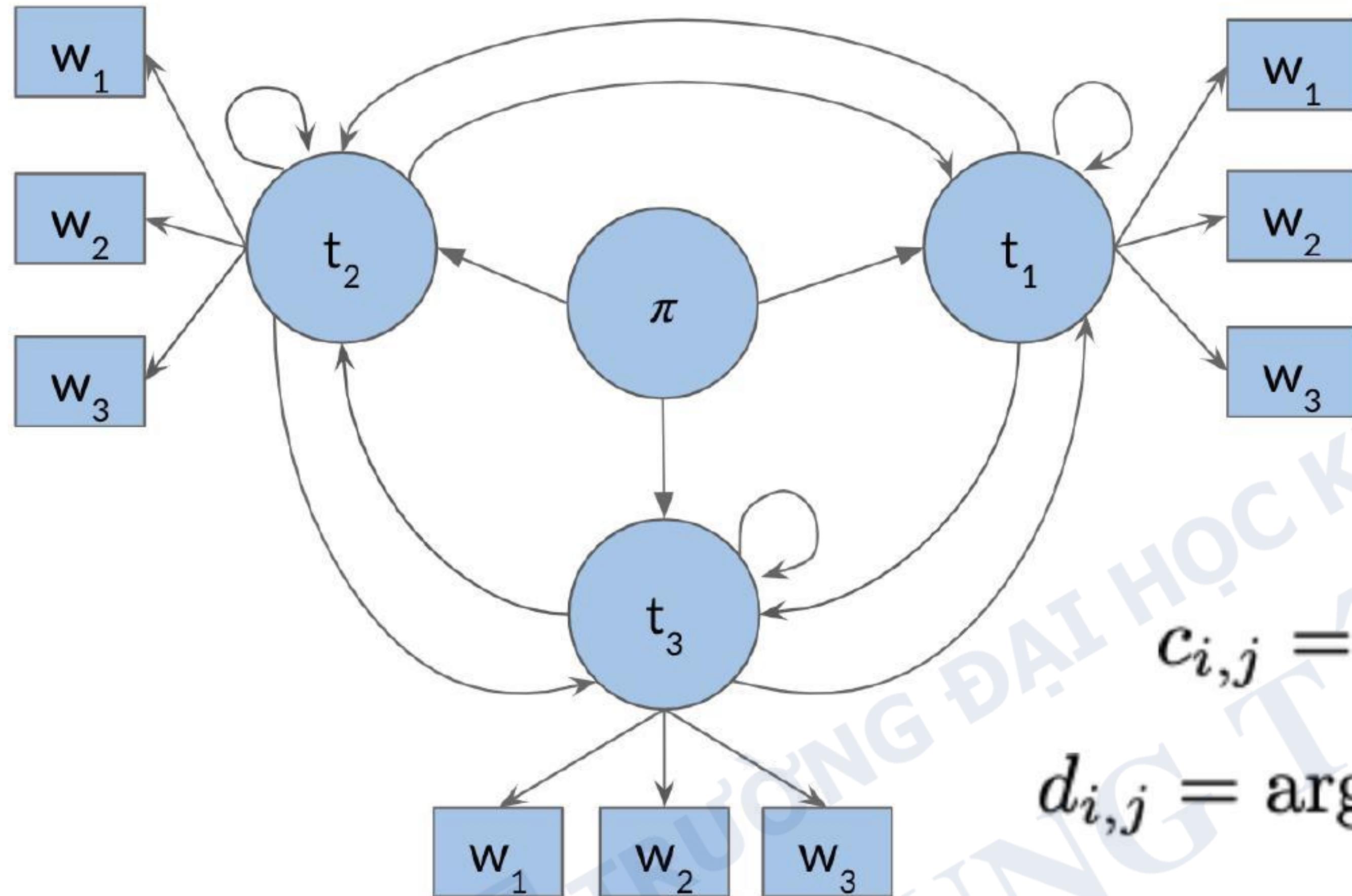
Giải thuật Viterbi - Tạo Matrix D

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

$d_{i,1} = 0$



Matrix D – Forward Pass



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

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

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

$D =$

Code Demo



DEMO



Q&A

