



XỬ LÝ NGÔN NGỮ TỰ NHIÊN

CHƯƠNG 5: Sequential labeling

ThS. Lưu Thanh Sơn



NỘI DUNG

1. Định nghĩa bài toán.
2. Part of speech tagging
3. Named entity recognition
4. Markov chains
- 5. HMM Tagger**
6. CRF
7. Độ đo đánh giá



5. Hidden Markov Model (HMM)



Hidden Markov model (HMM)

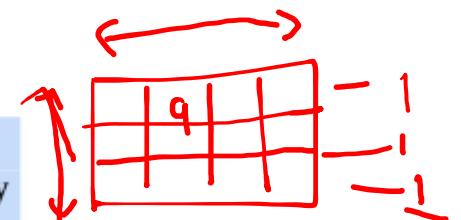
- Hidden Markov Model (HMM) là một mô hình dựa trên chuỗi Markov cho phép biểu diễn cho các sự kiện không thể quan sát trực tiếp.
 $w_1 w_2 w_3$
- Các sự kiện không thể biểu diễn trực tiếp:
 - Part of speech trong văn bản.
 - Mọi quan hệ giữa các thực thể trong một văn bản.



Các thành phần của HMM Model

- Q: tập trạng thái.
- A: ma trận chuyển xác suất.
- O: tập các sự kiện quan sát được (observation)
- B: xác suất khả năng xuất hiện của một sự kiện quan sát o trong tập O dựa trên trạng thái q trước đó (**emission probability**)
- π : xác suất khởi tạo. π_i là xác suất bắt đầu trạng thái thứ i.

$Q = q_1 q_2 \dots q_N$	a set of N states
$A = a_{11} \dots a_{ij} \dots a_{NN}$	a transition probability matrix A , each a_{ij} representing the probability of moving from state i to state j , s.t. $\sum_{j=1}^N a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \dots o_T$	a sequence of T observations , each one drawn from a vocabulary $V = v_1, v_2, \dots, v_V$
$B = b_i(o_t)$	a sequence of observation likelihoods , also called emission probabilities , each expressing the probability of an observation o_t being generated from a state q_i
$\pi = \pi_1, \pi_2, \dots, \pi_N$	an initial probability distribution over states. π_i is the probability that the Markov chain will start in state i . Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^n \pi_i = 1$





HMM Tagger

Tagging

- Là một mô hình dựa trên HMM, dùng để đánh nhãn cho các phần tử trong chuỗi. $\langle X_1 X_2 \dots X_n \rangle$
- HMM Tagger gồm 2 thành phần chính:
 - A: xác suất của một nhãn, thể hiện khả năng xuất hiện của một nhãn dựa trên nhãn trước nó.
 - B: xác suất xuất hiện của một từ w trong tập từ vựng khi biết nhãn t.



Xác suất xuất hiện của 1 nhãn

$$P(t_i|t_{i-1}) = \frac{\text{count}(t_{i-1}, t_i)}{\text{count}(t_{i-1})}$$

Ví dụ: Trong bộ dữ liệu WSJ:

- Nhãn MD xuất hiện 13124 lần
- Nhãn MD xuất hiện cùng với nhãn VB là 10471 lần
- Xác suất chuyển từ nhãn MD sang nhãn VB là:

MD VB

$$P(\text{VB}|MD) = \frac{\text{count}(MD, VB)}{\text{count}(MD)} = \frac{10471}{13124} = 0.8$$





Xác suất xuất hiện của một từ theo một nhãn

$$P(w_i|t_i) = \frac{\text{count}(t_i, w_i)}{\text{count}(t_i)}$$

Ví dụ: Trong bộ dữ liệu WSJ:



từ “will” xuất hiện với vai trò modal verb (WD) là 4046 lần.

Nhãn MD xuất hiện cùng với nhãn VB là 10471 lần

$$P(\overset{\leftarrow}{will}|MD) = \frac{\text{count}(MD, will)}{\text{count}(MD)} = \frac{4046}{10471} = 0.31\overset{\leftarrow}{}$$



HMM Decoding

- Input:

HMM: $\lambda = (\underset{\text{↑}}{A}, \underset{\text{↓}}{B})$

$O = o_1, o_2, \dots, o_n$

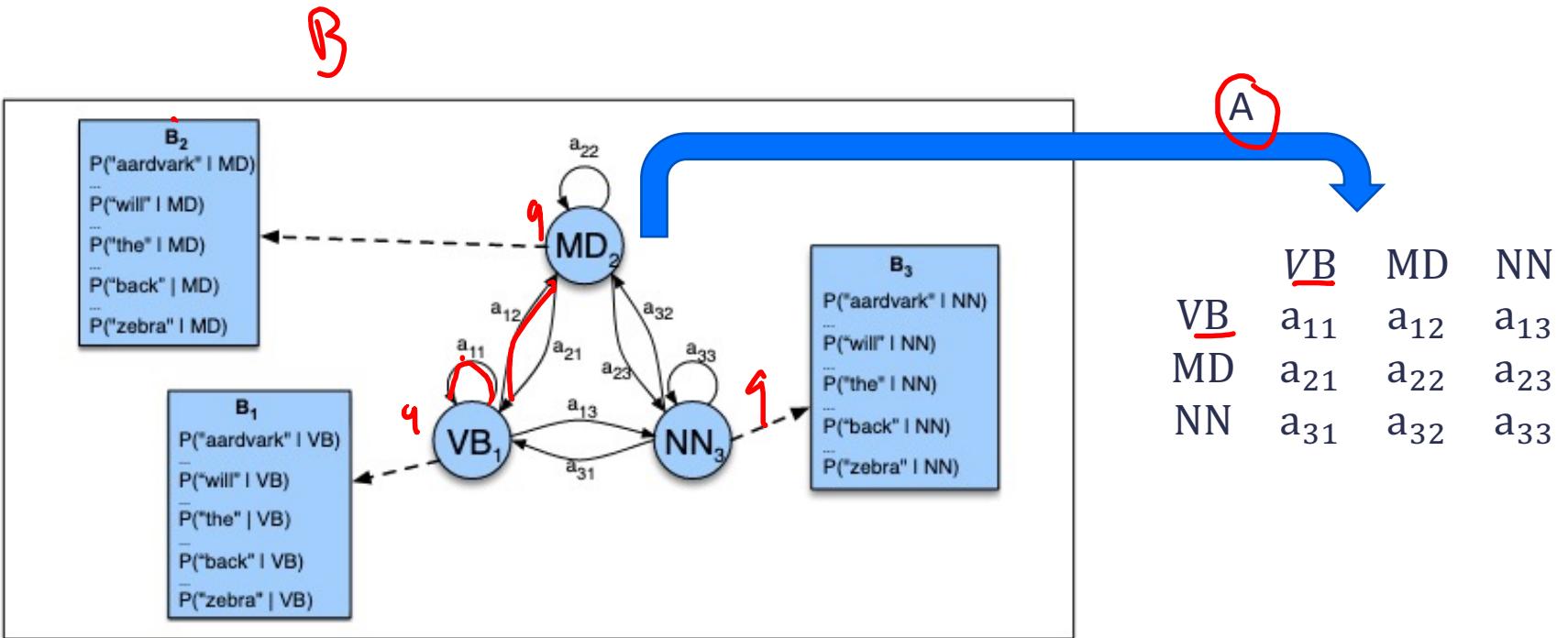
- Output:

Chuỗi có khả năng xuất hiện cao nhất $Q = \underset{\text{↑}}{\pi} q_1, q_2, \dots, q_n$





Ví dụ





HMM Decoding method

1. Tìm xác suất xuất hiện của một từ khi biết được loại từ của nó. Giả định rằng từ này độc lập với những từ lân cận và nhãn của các từ lân cận.

$$P(w_1 \dots w_n | t_1 \dots t_n) \approx \prod_{i=1}^n P(w_i | t_i)$$

2. Tìm xác suất xuất hiện của một nhãn khi biết nhãn trước nó. Giả định rằng nhãn này chỉ phụ thuộc vào một nhãn trước nó, và độc lập với các nhãn còn lại.

$$P(t_1 \dots t_n) \approx \prod_{i=1}^n P(t_i | t_{i-1})$$

3. Kết hợp 2 xác suất trên lại với nhau để tìm ra nhãn khả thi nhất cho chuỗi.

$$\rightarrow \hat{t}_{1:n} = \underset{t_1 \dots t_n}{\operatorname{argmax}} P(t_1 \dots t_n | w_1 \dots w_n) \approx \underset{t_1 \dots t_n}{\operatorname{argmax}} \prod_{i=1}^n \overbrace{P(w_i | t_i)}^{\text{emission transition}} \overbrace{P(t_i | t_{i-1})}^{\text{A}}$$



Thuật toán Viterbi

- Mục tiêu: ước tính xác suất tối đa của chuỗi trạng thái ẩn có khả năng xảy ra nhất, được gọi là đường dẫn Viterbi, dẫn đến một chuỗi các sự kiện được quan sát. *(Path n)*
- Công cụ sử dụng: **minimum edit distance.**



Thuật toán Viterbi (mã giả)

```
function VITERBI(observations of len  $T$ ,state-graph of len  $N$ ) returns best-path, path-prob
    create a path probability matrix viterbi[ $N,T$ ]
    for each state  $s$  from 1 to  $N$  do ; initialization step
        viterbi[ $s,1$ ]  $\leftarrow \pi_s * b_s(o_1)$ 
        backpointer[ $s,1$ ]  $\leftarrow 0$ 
    for each time step  $t$  from 2 to  $T$  do ; recursion step
        for each state  $s$  from 1 to  $N$  do
            viterbi[ $s,t$ ]  $\leftarrow \max_{s'=1}^N$  viterbi[ $s',t-1$ ] *  $a_{s',s}$  *  $b_s(o_t)$ 
            backpointer[ $s,t$ ]  $\leftarrow \operatorname{argmax}_{s'=1}^N$  viterbi[ $s',t-1$ ] *  $a_{s',s}$  *  $b_s(o_t)$ 
    bestpathprob  $\leftarrow \max_{s=1}^N$  viterbi[ $s,T$ ] ; termination step
    bestpathpointer  $\leftarrow \operatorname{argmax}_{s=1}^N$  viterbi[ $s,T$ ] ; termination step
    bestpath  $\leftarrow$  the path starting at state bestpathpointer, that follows backpointer[] to states back in time
    return bestpath, bestpathprob
```



Thuật toán Viterbi

- For a given state q_j at time t , the value $v_t(j)$ is computed as:

$$v_t(j) = \max_{i=1}^N v_{t-1}(i) a_{ij} b_j(o_t)$$

$v_{t-1}(i)$	the previous Viterbi path probability from the previous time step
a_{ij}	the transition probability from previous state q_i to current state q_j
$b_j(o_t)$	the state observation likelihood of the observation symbol o_t given the current state j



Thuật toán Viterbi

1. Initialization:

$$\rightarrow v_1(j) = \pi_j b_j(o_1) \quad 1 \leq j \leq N$$

$$\rightarrow bt_1(j) = 0 \quad 1 \leq j \leq N$$

2. Recursion

$$v_t(j) = \max_{i=1}^N v_{t-1}(i) a_{ij} b_j(o_t); \quad 1 \leq j \leq N, 1 < t \leq T$$

$$bt_t(j) = \operatorname{argmax}_{i=1}^N v_{t-1}(i) a_{ij} b_j(o_t); \quad 1 \leq j \leq N, 1 < t \leq T$$

3. Termination:

The best score: $P* = \max_{i=1}^N v_T(i)$

[- - - - -]

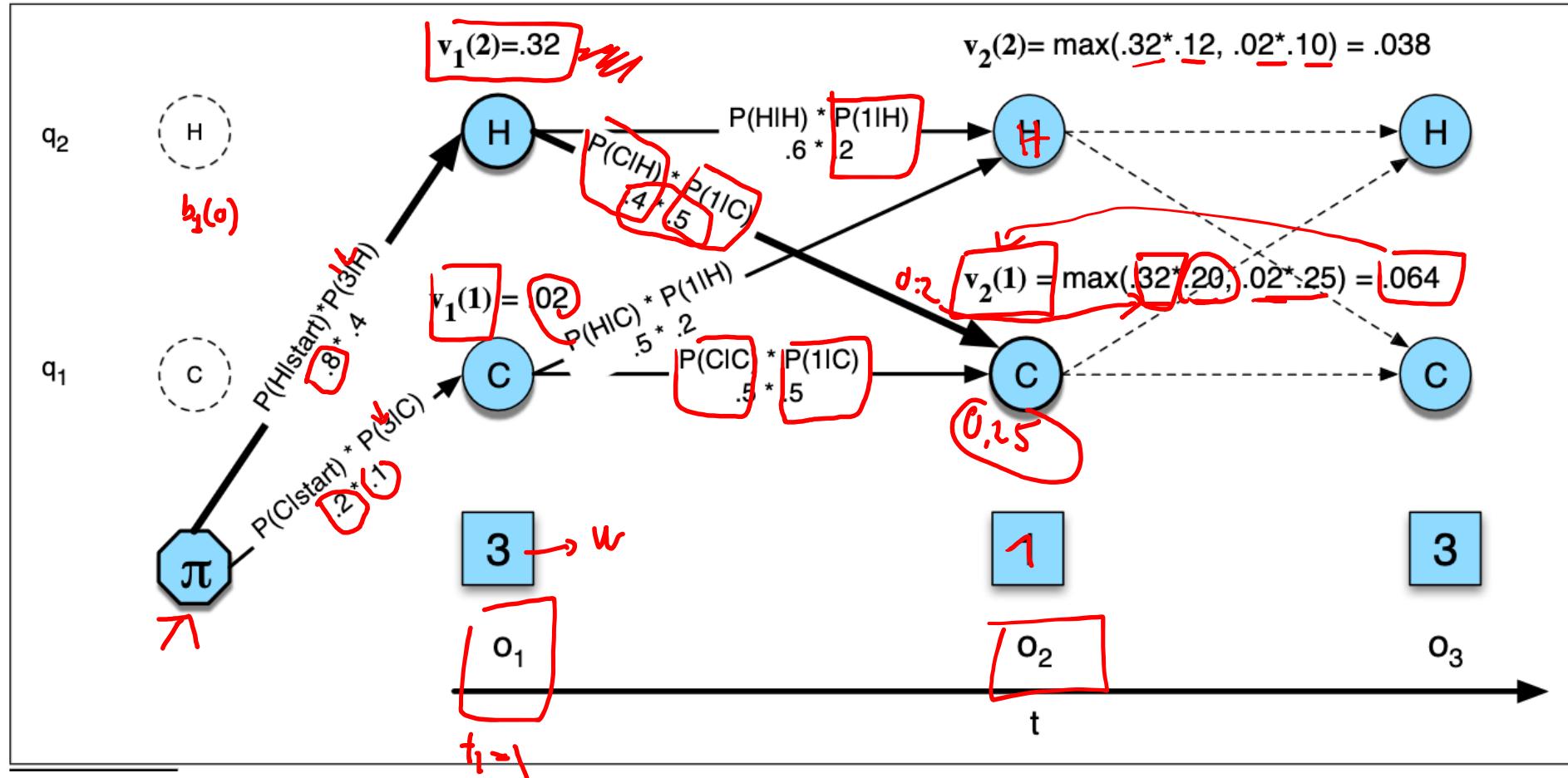
— — — — —

The start of backtrace: $q_T* = \operatorname{argmax}_{i=1}^N v_T(i)$

f



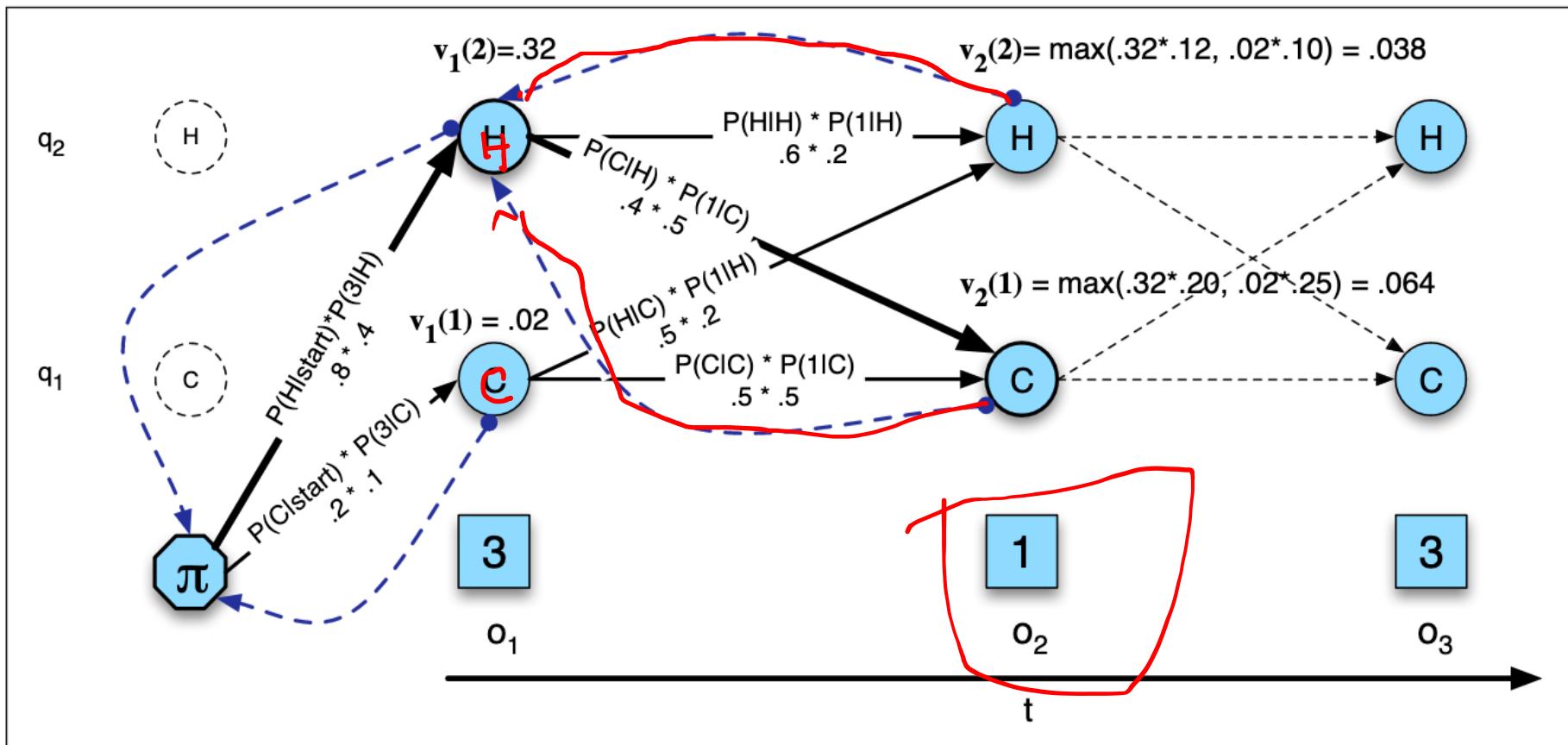
Minh họa Thuật toán Viterbi





Backtrace

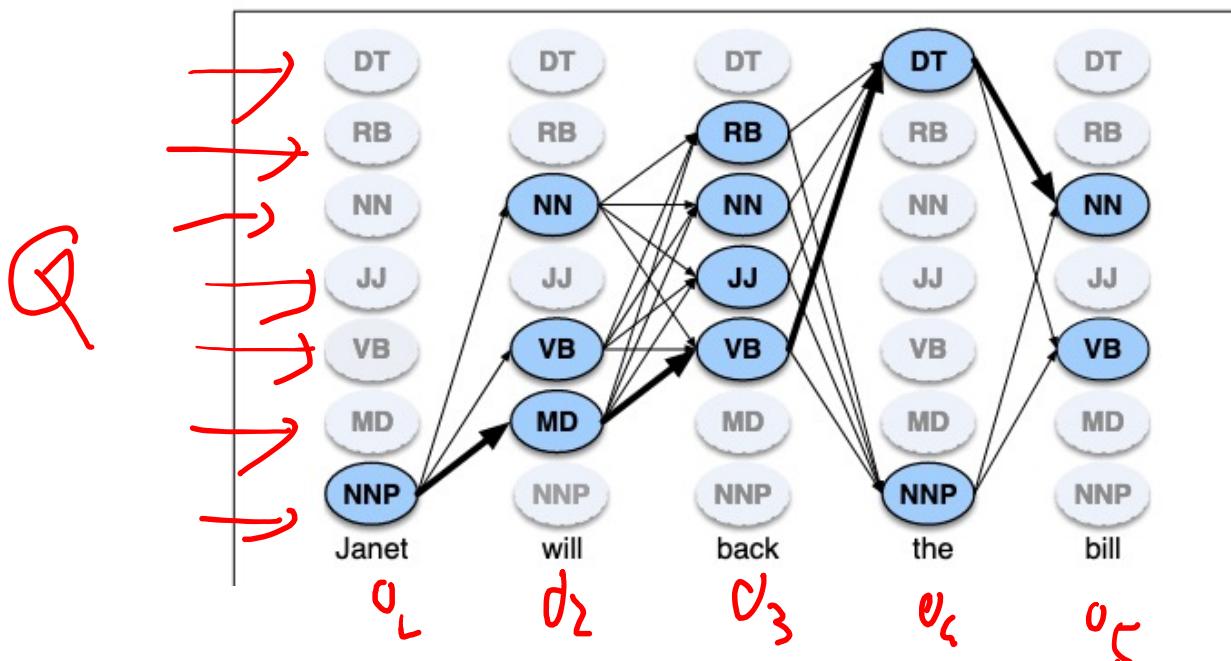
- Backpointers: Truy vết (backtrace) đường đi của từng hidden state khi chuyển trạng thái đến state hiện tại.





Biểu diễn cho thuật toán Viterbi

- Sử dụng một “lattice” để biểu diễn:
 - Mỗi cột biểu diễn cho một sự kiện quan sát được o .
 - Mỗi dòng biểu diễn cho một trạng thái s .





Ví dụ

- Văn bản: Janet will back the bill.
- Xác suất đầu ra mong muốn: NNP – MD – VB – DT – NN



Ví dụ

A

	NNP	MD	VB	JJ	NN	RB	DT
< s >	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

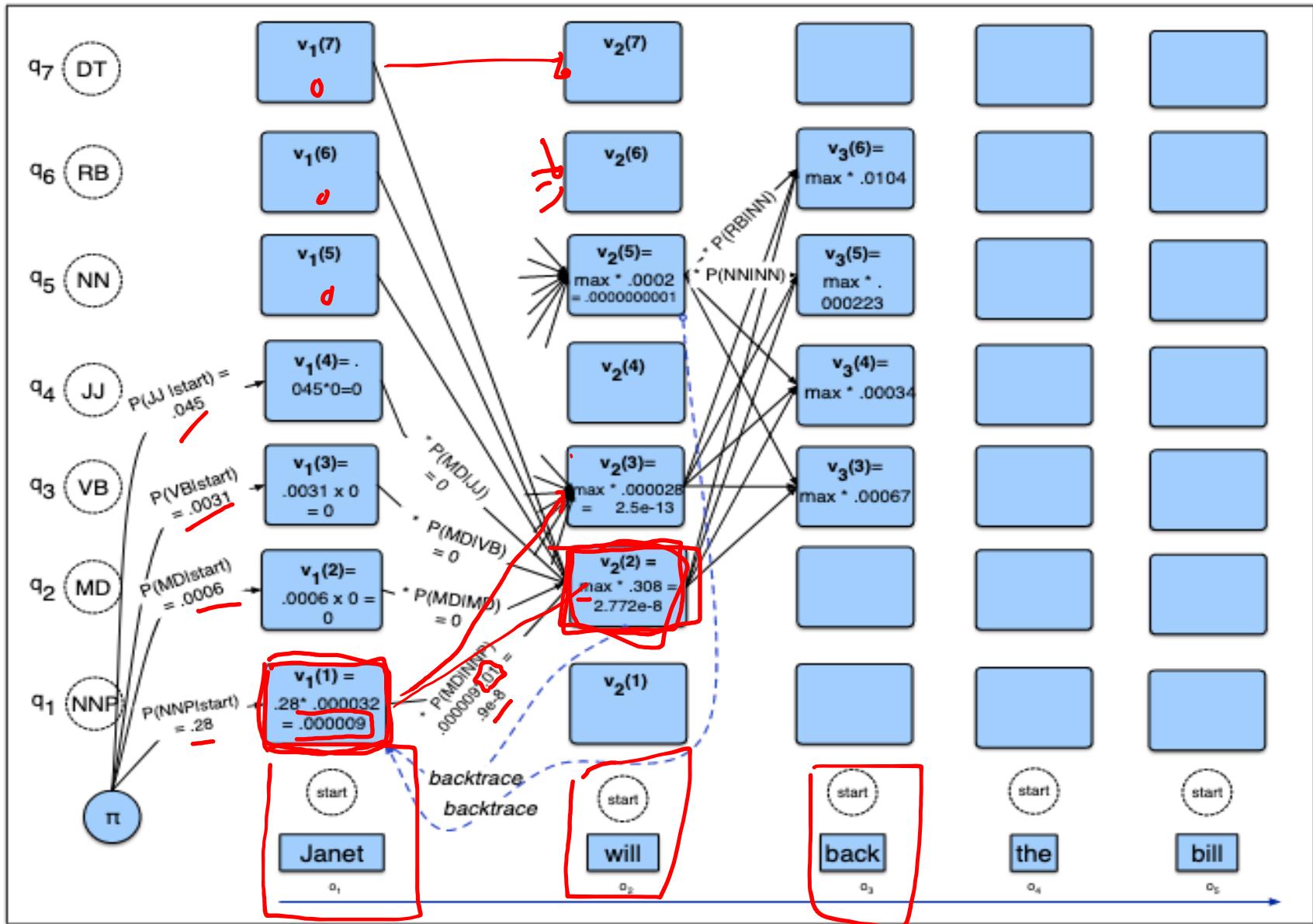
	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Xác suất chuyển ứng với từng nhãn
(tag)

Xác suất từ w xuất hiện khi biết
nhãn t

b(t)

Các xác suất trên được tính dựa trên bộ dữ liệu WSJ.



Tính tiếp cho các cột: "will", "back", "the", "bill".



Vấn đề với HMM

- Sự xuất hiện của các từ không biết nhãn (unknow words). Các từ này có thể là:

- Từ viết tắt (acronyms)
- Tên riêng (proper names)
- Các từ “mới”.

→ knowing the previous or following words might be a useful feature

F_k