



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



4. Markov Chains



Markov chains (Chuỗi Markov)

- Chuỗi Markov (Markov chain) là một quá trình ngẫu nhiên mô tả một dãy các biến cố khả dĩ trong đó xác suất của mỗi biến cố chỉ phụ thuộc vào trạng thái của biến cố trước đó.
- Tính chất quan trọng: khi dự đoán sự xuất hiện của trạng thái kế tiếp, thì **trạng thái hiện tại (current state)** đóng vai trò quyết định.



Andrei Andreyevich Markov
(1856 - 1922)

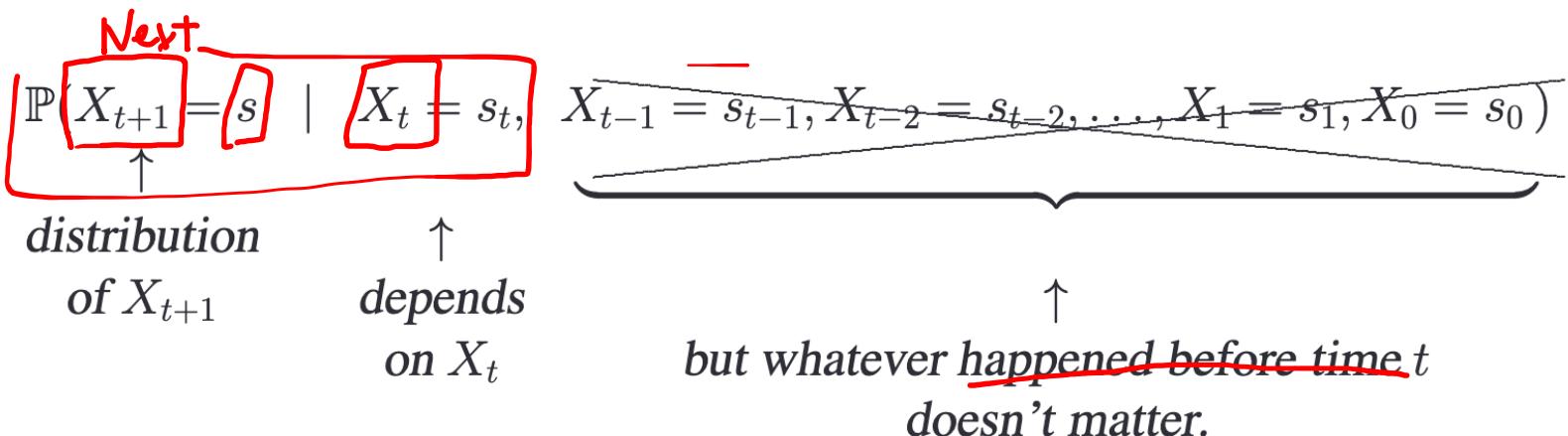


Markov assumption

- Dự đoán tương lai dựa vào hiện tại, không dựa vào quá khứ.

Markov Assumption: $P(q_i = a | q_1 \dots q_{i-1}) = P(\underline{q_i} = a | \underline{q_{i-1}})$

VD: dự báo thời tiết của ngày mai thì dựa vào trạng thái thời tiết của ngày hiện tại, không dựa vào thời tiết của các ngày trước đó





Các thành phần của mô hình Markov

- Q: Tập các trạng thái (states)
- A: ma trận chuyển xác suất. Mỗi phần tử trong ma trận a_{ij} cho biết xác suất chuyển từ trạng thái i sang trạng thái j.
- 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

$$A = a_{11} a_{12} \dots a_{N1} \dots a_{NN}$$

a **transition probability matrix** A , each a_{ij} representing the probability of moving from state i to state j , s.t.
 $\sum_{j=1}^n a_{ij} = 1 \quad \forall i$

$$\pi = \pi_1, \pi_2, \dots, \pi_N$$

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, $\sum_{i=1}^n \pi_i = 1$

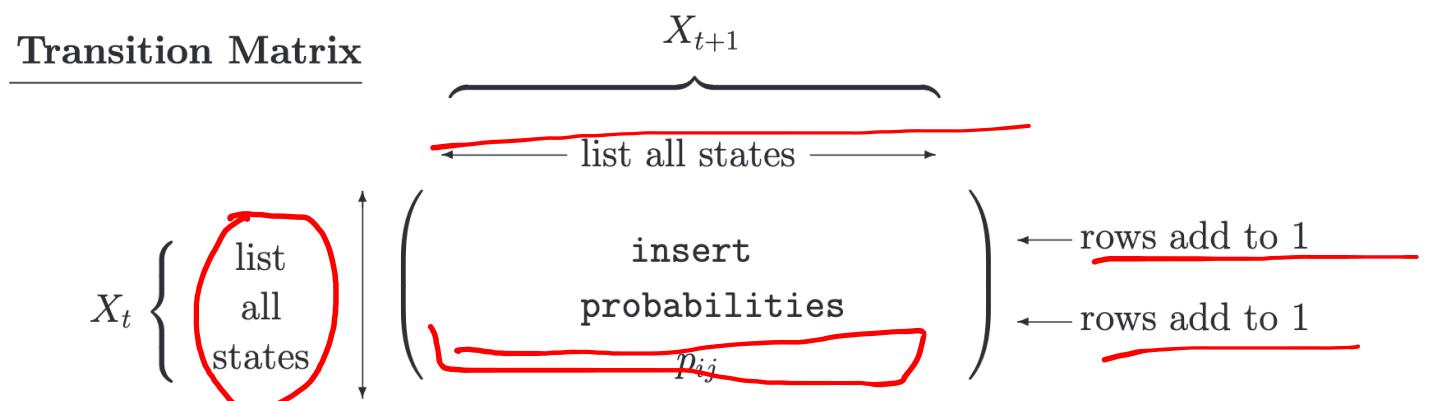


Ma trận chuyển trạng thái (transition matrix)

- Mỗi dòng biểu diễn trạng thái Hiện tại (NOW).
- Mỗi cột biểu diễn trạng thái Tiếp theo (NEXT).
- Mỗi ô $p(i, j)$ biểu diễn xác suất có điều kiện để chuyển từ trạng thái NOW i sang NEXT t.

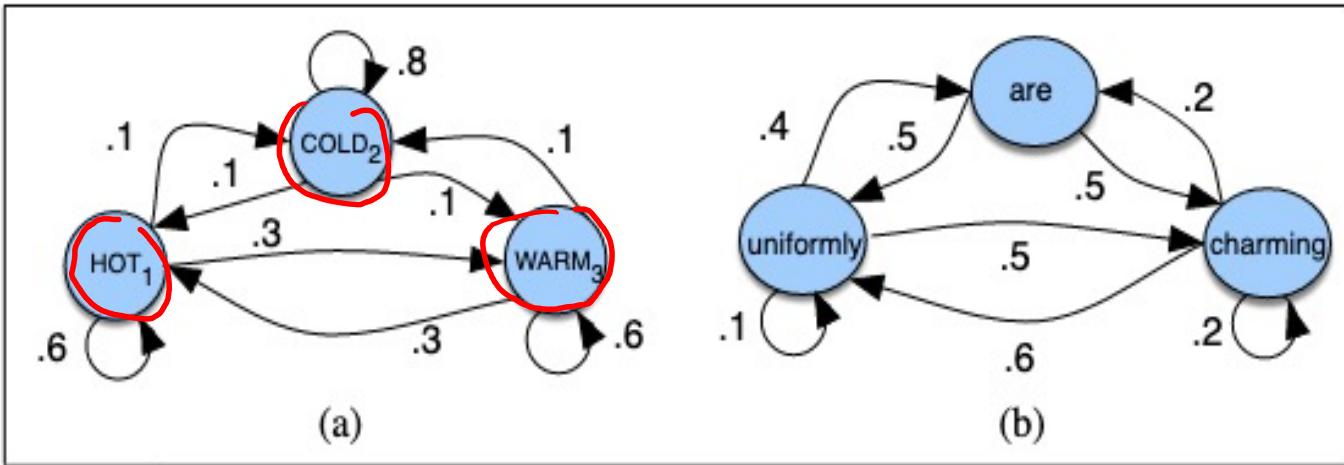
$$p_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i).$$

- Giá trị tổng mỗi dòng là 1.





Ví dụ



Xét hình (a):

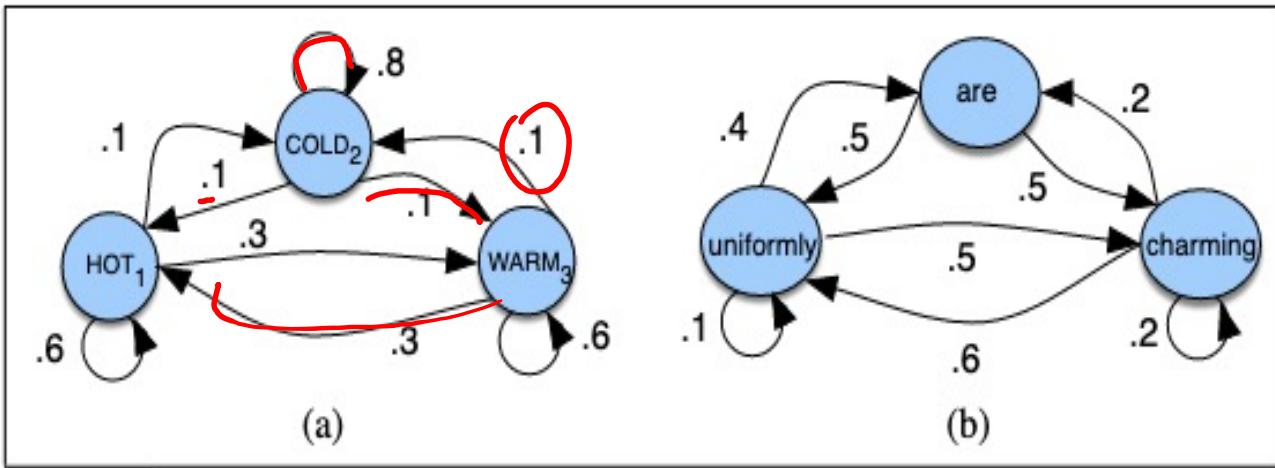
$$Q = \{\text{HOT}, \text{COLD}, \text{WARM}\}$$

$$\pi = [0.7, 0.1, 0.2]$$

Xây dựng ma trận chuyển xác suất A.



Ví dụ 1



Xét hình (a):

$$Q = \{\text{HOT}, \text{COLD}, \text{WARM}\}$$

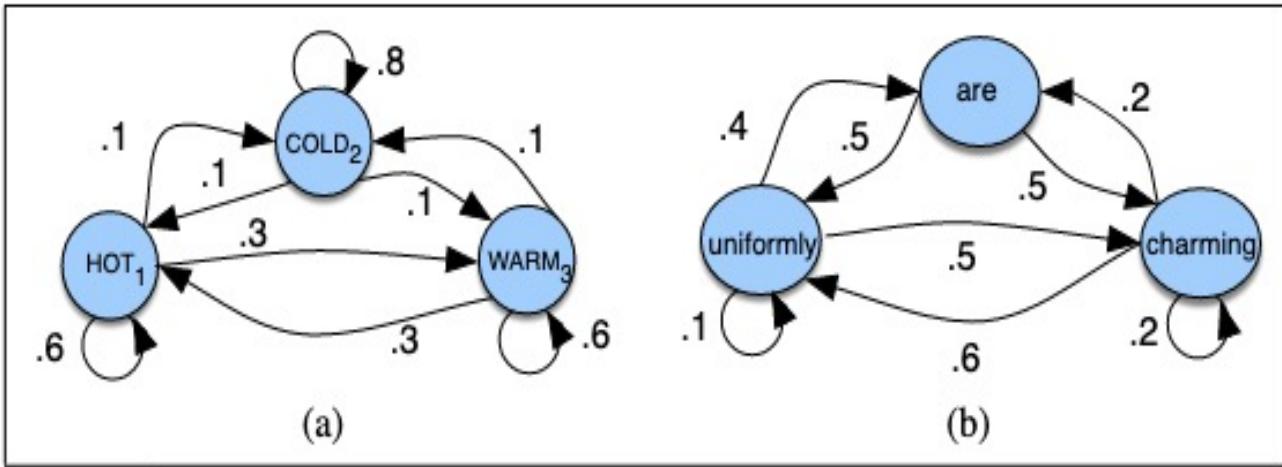
$$\pi = [0.7, 0.1, 0.2]$$

Xây dựng ma trận chuyển xác suất A.

	HOT	COLD	WARM
π	0.7	0.1	0.2
HOT	0.6	0.1	0.3
COLD	0.1	0.8	0.1
WARM	0.3	0.1	0.6



Ví dụ 2



Xét hình (b):

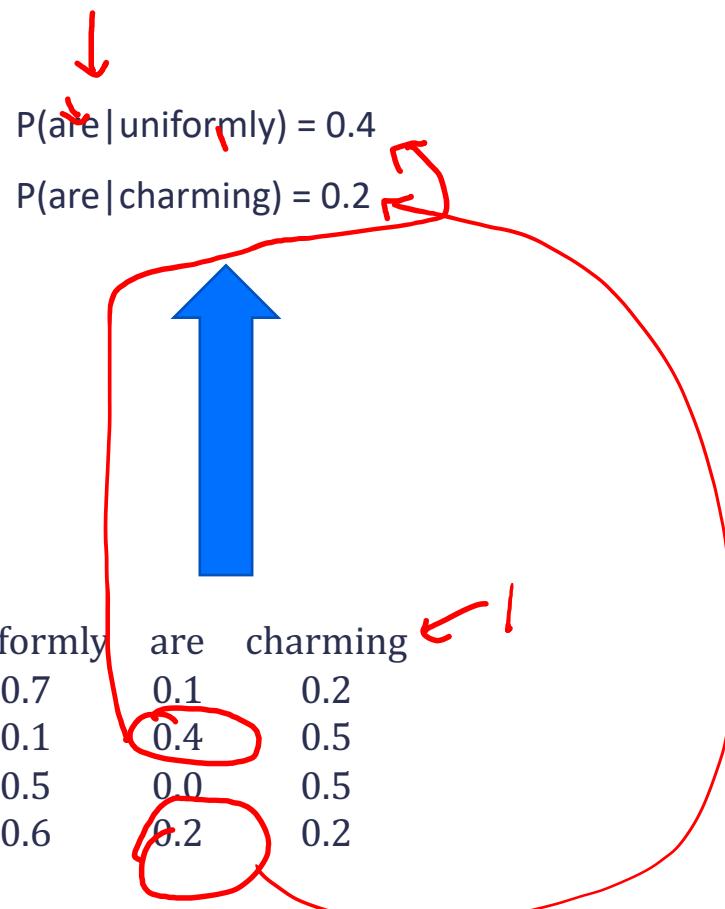
$$Q = \{\text{uniformly, are, charming}\}$$

$$\pi = [0.7, 0.1, 0.2]$$

Xây dựng ma trận chuyển xác suất A.

π
uniformly
are
charming

	uniformly	are	charming
uniformly	0.7	0.1	0.2
are	0.1	0.4	0.5
charming	0.5	0.0	0.5
	0.6	0.2	0.2





Bài tập

- Dựa vào ma trận xác suất chuyển ở Hình (a), tính xác suất chuyển cho các chuỗi sau:

(1) ~~hot~~¹-hot²-hot³-hot⁴

(2) cold hot cold hot

	HOT	COLD	WARM
π	0.7	0.1	0.2
HOT	0.6	0.1	0.3
COLD	0.1	0.8	0.1
WARM	0.3	0.1	0.6

→ $P(\text{hot}|\pi) * P(\text{hot}|\text{hot}) * P(\text{hot}|\text{hot}) * P(\text{hot}|\text{hot})$
= 0.7 * 0.6 * 0.6 * 0.6 =