TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN GIỮA KỲ MÔN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

MÔ HÌNH HIDDEN MARKOV VỚI THUẬT TOÁN VITERBI

Người hướng dẫn: **PGS TS. LÊ ANH CƯỜNG**

Người thực hiện:

PHẠM DƯƠNG THÀNH LONG – *51603190*

Lớp : **16050304**

Khoá : **20**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020

MỤC LỤC

DANH MỤC CÁC HÌNH VẼ VÀ BẢNG BIỂU

PHẦN 1 – GIỚI THIỆU, CƠ SỞ LÝ THUYẾT

1.1 Tổng quan:

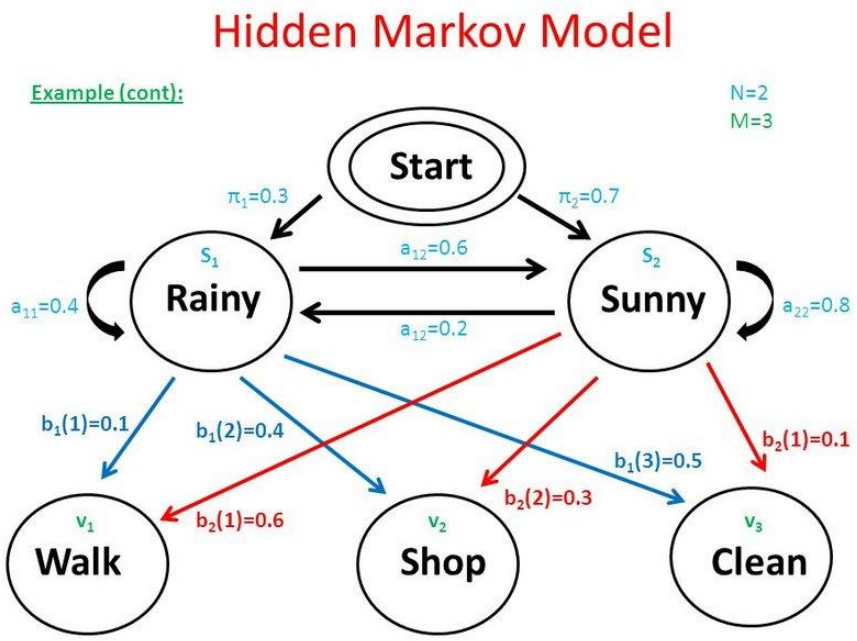
1.1.1 Nguồn gốc:

Học thuyết markov được hình thành và phát triển từ những *năm 1900* và đến những năm *1960 – 1970* mới được phổ biến một cách rộng rãi, đặc biệt là trong lĩnh vực nhận dạng tiếng nói. *Đến năm 1989*, thì nó được chuyển khai sang lĩnh vực học máy.

1.1.2 Định nghĩa:

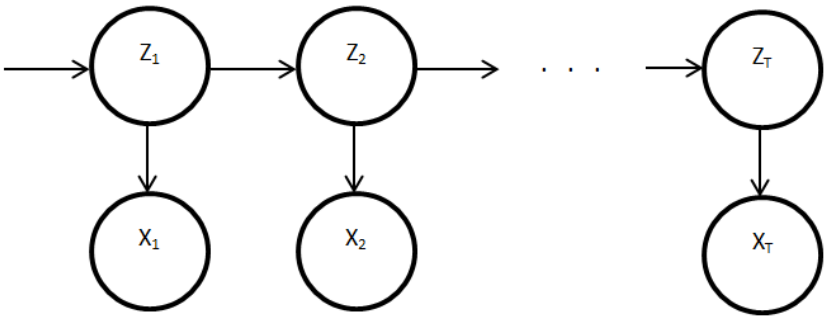
1.1.2.1 Hidden markov model

Hidden markov model *(HMM)* hay *(tiếng việt là Mô hình Markov ẩn)* là một ***mô hình thống kê.*** Trong đó, hệ thống được mô hình hóa được cho là một ***quá trình Markov*** với các tham số không biết trước . Nhiệm vụ của mô hình là ***xác định các tham số ẩn từ các tham số quan sát được,*** dựa trên sự thừa nhận này. Các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp.



Hình 1.1.2.1: Một ví dụ về Hidden markov model.

Mô hình Markov ẩn được biểu diễn dưới dạng đồ thị chuyển trạng thái (Hình 1.1.2.2). Các nút là các trạng thái, $\boldsymbol{Z_i}$ là các trạng thái ẩn, $\boldsymbol{X_i}$ là các trạng thái quan sát được. Các đường mũi tên là các lần chuyển trạng thái có gán xác suất.



Hình 1.1.2.1: Mô hình Hidden markov model (minh hoạ).

Việc xác định các tham số ẩn dựa vào các tham số đã biết thông qua các ***xác suất chuyển trạng thái*** (từ trạng thái Z_{t-1} sang trạng thái Z_t) và ***xác suất nhả trạng thái*** (trạng thái quan sát được X_t nhận trạng thái ẩn Z_t).

Trong mô hình Markov ẩn tổng quát bậc n, trạng thái Z_t phụ thuộc vào ***n*** trạng thái đứng trước nó.

Ví dụ: xét n = 1, tức trạng thái Z_t phụ thuộc vào trạng thái Z_{t-1} và độc lập với các trạng thái khác, tức là:

- $P(Z_t \mid Z_{t-1}, Z_{t-2}, \ldots) = P(Z_t \mid Z_{t-1})$ *(1)*
- $P(Z_t \mid Z_{t-1}, Z_{t-2}, \ldots) = P(Z_t \mid Z_{t-1})\ P(Z_{t-1} \mid Z_{t-2})$ *(2)*

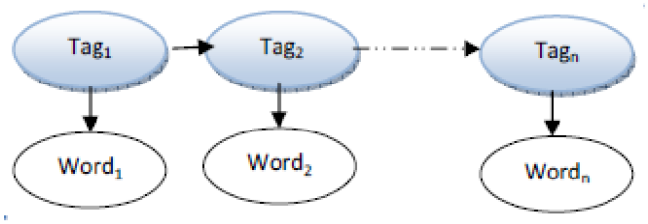
1.1.3 Ứng dụng

Ứng dụng của nó thể hiện ở nhiều lĩnh vực khác nhau như: Lĩnh vực thống kê, lĩnh vực tin sinh học *(bio-infomatics),* Xử lý tín hiệu, nhận dạng mẫu,. v.v.. Đặc biệt trong báo cáo lần này sẽ là ***part of speech.***

2. Xây dựng mô hình part of speech:

- **Input:** của bài toán là *một câu tiếng anh* ***đã được tách từ***.
- **Output:** Nhãn hay còn gọi ***postag*** cho từng từ trong câu.

Giả sử:



Cho một chuỗi các từ (word): $S = w_1, w_2 \dots w_n$

Tập hữu hạn các Pos Tag từ loại $T = \{ t_1, t_2, \dots, t_m \}$

Lúc này, các từ w_i là các ***đối tượng quan sát được*** và các nhãn T_i là các ***trạng thái ẩn***. Ta phải xác định nhãn từ loại t_i tương ứng cho mỗi từ w_i để thu được $T^* = t_1, t_2, \dots, t_n$ *(với* $t_i \in T$*)*. mục tiêu của ta là đi tìm ***P(T* / S) lớn nhất***. Ta sẽ minh hoạ công thức sau:

$$T^* = \text{argmax } P\left(\frac{T}{S}\right) \quad (3)$$

Áp dụng Bayes vào $P\left(\frac{T}{S}\right)$ ta được:

$$\boldsymbol{P(T \mid S) = \frac{P(S|T) * P(T)}{P(S)}} \quad (4)$$

Mục tiêu bài toán là tìm chuỗi nhãn phù hợp nhất hay ***làm cực đại công thức (3)*** nên lúc này, mẫu số trong tất cả các trường hợp là giống nhau, vì vậy ta có thể loại bỏ nó. Do đó, bài toán trở thành tìm chuỗi các nhãn thỏa mãn công thức (5):

$$T^* = \text{argmax } P(S \mid T)\, P(T) \quad (5)$$

Khai triển (5) theo luật chuỗi xác suất ta được:

$$T^* = \text{argmax } P(w_1, w_2, \dots, w_n \mid t_1, t_2, \dots, t_n) * P(t_1, t_2, \dots, t_n) \quad (6)$$

Từ (6) áp dung vào mô hình Markov ẩn bậc 1:

$$T^* = \text{argmax } \prod_{i=1}^{n} p(w_i|t_i) \prod_{i=1}^{n} p(t_i|t_{i-1}) \quad (7)$$

Tiếp theo ta áp dụng thuật toán Viterbi để tìm dãy trạng thái tối ưu, dựa trên công thức truy hồi ta có công thức dưới đây:

$$\sigma_{i+1}(t_j) = \max_{1 \leq k \leq T} [\, \sigma_i(t_k) * P(w_i|t_i) * P(t_i|t_{i-1}) \,] \quad (8)$$

Tương tự ta cũng có,

$$\psi_{i+1}(t_j) = \text{argmax}_{1 \leq k \leq T} [\, \sigma_i(t_k) * P(w_i|t_i) * P(t_i|t_{i-1}) \,] \quad (9)$$

Ví dụ 1 :

Giả sử ta có 4 Postag: t_0, t_1, t_2, t_3 *(trong đó* t_0 *là nhãn bắt đầu)* với một câu đã tách từ: S= [w_1, w_2, w_3]. *Sau đó ta có 2 bảng sau:*

Trước \ Sau	t_1	t_2	t_3
t_0	0.3	0.4	0.3
t_1	0.2	0.2	0.6
t_2	0.4	0.1	0.5
t_3	0.1	0.8	0.1

Bảng 2.1 xác suất chuyển từ loại

Pos Tag \ Word	w_1	w_2	w_3
t_1	0.01	0.02	0.02
t_2	0.8	0.01	0.5
t_3	0.19	0.97	0.48

Bảng 2.2 xác suất nhận pos_tag của từ loại.

Quá trình xác định từ loại cho mỗi từ được mô tả như sau:

Bước 1: $\varphi(t_0) = 1$

Xác suất Viterbi cho các thẻ từ đầu tiên :

$$\varphi(t_1) = \varphi(t_0)\, p(t_1|\, t_0)\, \mathrm{p}(w_1|\, t_1) = 0.003$$

Tương tự ta có,

- $\varphi(t_2) = \mathbf{0.32}$
- $\varphi(t_3) = 0.057$

Xác suất Viterbi cho các thẻ từ thứ 2:

$$\varphi_2(\boldsymbol{t_1}) = \max\{\, \varphi_1(t_1)\, p(t_1|\, t_1)\, \mathrm{p}(w_2|\, t_1),$$
$$\varphi_1(t_2)\, p(t_1|\, t_2)\, \mathrm{p}(w_2|\, t_1),$$
$$\varphi_1(t_3)\ p(t_1|\, t_3)\, \mathrm{p}(w_2|\, t_1)\,\}$$

$$\varphi_2(\mathrm{t}_1) = 0.00256 \ \text{->}\ \psi_2(t_1) = t_3$$

Tương tự, ta có:

- $\varphi_2(t_2)$ =0.000456 -> $\rightarrow$ $\psi_2(t_3) = t_2$
- $\varphi_2(\boldsymbol{t_3})$ =**0.1552** -> $\rightarrow$ $\psi_2(t_3) = t_3$

Xác suất Viterbi cho các thẻ từ thứ 3:

$$\varphi_3(t_1) = \max\{\varphi_1(t_1)\, p(t_1|\, t_1)\, \mathrm{p}(w_3|\, t_1),$$
$$\varphi_1(t_2)\, p(t_1|\, t_2)\, \mathrm{p}(w_3|\, t_1),$$
$$\varphi_1(t_3)\ \ p(t_1|\, t_3)\, \mathrm{p}(w_3|\, t_1)\}$$
$$= 0{,}0003104 \text{-> } \psi_3(t_1) = t_3$$

Tương tự, ta có:

- $\varphi_3(\boldsymbol{t_2})$ =**0,06208** -> $\rightarrow$ $\psi_2(t_3) = t_2$
- $\varphi_3(t_3)$ =0,0074496-> $\rightarrow$ $\psi_2(t_3) = t_2$

Từ các kết quả trên, qua quá trình suy lui ta có chuỗi tối ưu là:

$\boldsymbol{t_0 \rightarrow t_2 \rightarrow t_3 \rightarrow t_2}$

Tương ứng với :

w_1/t_2, w_2/t_3, $w_3/\ t_2$