BÀI TOÁN GÁN NHÃN TỪ LOẠI

Lê Duy Thức (thuc.leduy.int@gmail.com)

Phần 1 :Ví dụ HMM cho bài toán gán nhãn:

Cho câu -> **Tôi đi học**

Đây là câu trên gồm các từ đã dán nhãn:

Tôi/S đi/V học/N

Trong đó: S, V, N là các **nhãn** tương ứng

+ Ps: đầu câu

+ S : chủ từ (tôi, tao, mẹ, mày)

+ N: danh từ chung (cơm, cây, nước,…)

+ V động từ

(có cả đống nhãn nhưng chỉ vd 3 nhãn cơ bảng thôi cho dễ hiểu)

Phần 2:

- Cho tập dữ liệu dc huấn luyện (lấy vd 5 dòng trong 15k dòng)

start/Ps Lại/R đi/V trên/E đường/N Nguyễn_Ái_Quốc/Np ./CH end/P_s

start/Ps Anh/N ta/P tưởng/V bà/Nc mẹ/N "/CH bôn/X "/CH lắm/R ./CH end/P_s

start/Ps Trái_tim/N ơi/T ,/CH vỡ/V ra/R và/C đừng/R bao_giờ/P rung/V nữa/R ./CH end/P_s

start/Ps N.Anh/X bé_nhỏ/A yêu_dấu/V đêm/N nay/P ở/V đâu/P .../X end/P_s

start/Ps thế/T rồi/C không/R còn/R mơ/V được/V gì/P nữa/R ,/CH cứ/R trượt/V theo/E những/L đường_cong/N mềm_mại/A .../X end/P_s

start/Ps Mẹ/N mình/P chưa/R khóc/V bao_giờ/P trong/E những/L lần/N tiễn/V con/N đi/V ./CH end/P_s

start/Ps Thanh/Np bảo/V không/R dám/V kể/V tình_hình/N thật/A của/E đơn_vị/N cho/E gia_đình/N vì/E nó/P sợ/V gia_đình/N không/R yên_tâm/A ./CH end/P_s

start/Ps Càng/R nghĩ/V càng/R buồn/V ,/CH nhất_là/X trong/E những/L ngày/N này/P ./CH end/P_s

start/Ps Thế/T đấy/P ,/CH mẹ/N mình/P chắc/V là/C không/R thế/P ./CH end/P_s

Mục tiêu của mô hình HMM cho bài toán này là từ **một tập dữ liệu đã được gán nhãn** cho sẵn (cái đống trên), người dùng nhập vô 1 câu mới, **chương trình tự gán nhãn lại**

Ví dụ input: "***Tôi ăn cơm"*** từ tập trên suy ra "***start/Ps Tôi/S ăn/V cơm/N"***

Phần 3: mô tả chạy tay trên giấy

Input : *Tôi ăn cơm*

Từ tập dữ liệu cho sẵn(đống ở trên), suy ra dc 2 bảng dưới đây, làm sao suy ra – nói sau

	S	*V*	*N*
Start	*0.3*	*0.2*	*0.5*
N	*0.2*	*0.7*	*0.1*
V	*0.5*	*0.3*	*0.2*
S	*0.1*	*0.85*	*0.05*

Bảng Nhãn-nhãn

	Tôi	*ăn*	*cơm*
N	*0.4*	*0.2*	*0.5*
V	*0.1*	*0.6*	*0.05*
S	*0.5*	*0.1*	*0.05*

Bảng Từ-Nhãn

Bảng Từ-Nhãn: là xác xuất của từ đó là loại từ gì- nhãn gì.

VD: + nhìn vô thấy Tôi – S : 0.5 tức là "Tôi" thường xuất hiện trong câu với nhãn là S(chủ từ) chiếm 50%– như *Tôi ăn bánh* chả hạn. Còn "Tôi" – V : 0.1 thì tôi ở đây là V(động từ) --> rất hiếm có 10%,…

+ lưu ý 1 trong 1 cột vd Tôi xuất hiện là Tôi-N:0.4, Tôi-V:0.1, Tôi-S:0.5 . Tôi có thể là N V S nhưng cộng hết lại phải là 100% (0.4 + 0.1 + 0.5)

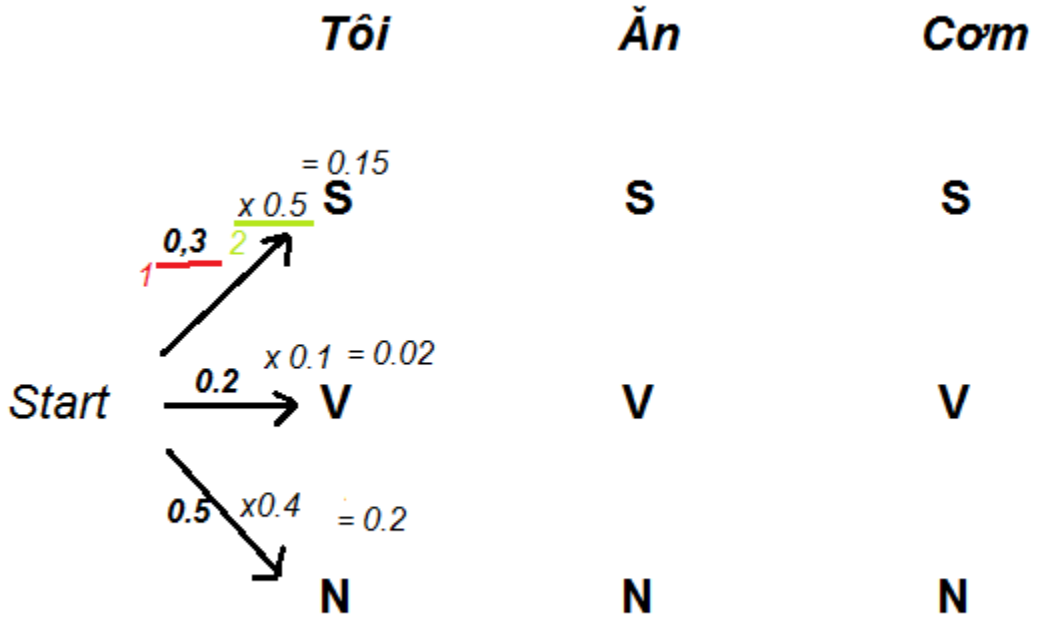
Bảng Nhãn-Nhãn:

VD + nhìn vô thấy S-V : 0.5 . Tức là: sau một S(chủ từ) thì khả năng xuất hiện V(danh từ) là 50% vd: *"Tao/S chạy/V"*, *"Mày/S ngủ/V"*. Hay V-N : 0.7 thì sau V(động từ) có tới 70% thường là danh từ(N) vd : *"đem/V cơm/N"*, *"Tao/S thích/V bánh/N"*

+ *Start* có nghĩa là bắt đầu câu: N-Start : 0.5 có nghĩa 1 câu thường bắt đầu với danh từ(N)

Phần 3 Cách chạy

B1:



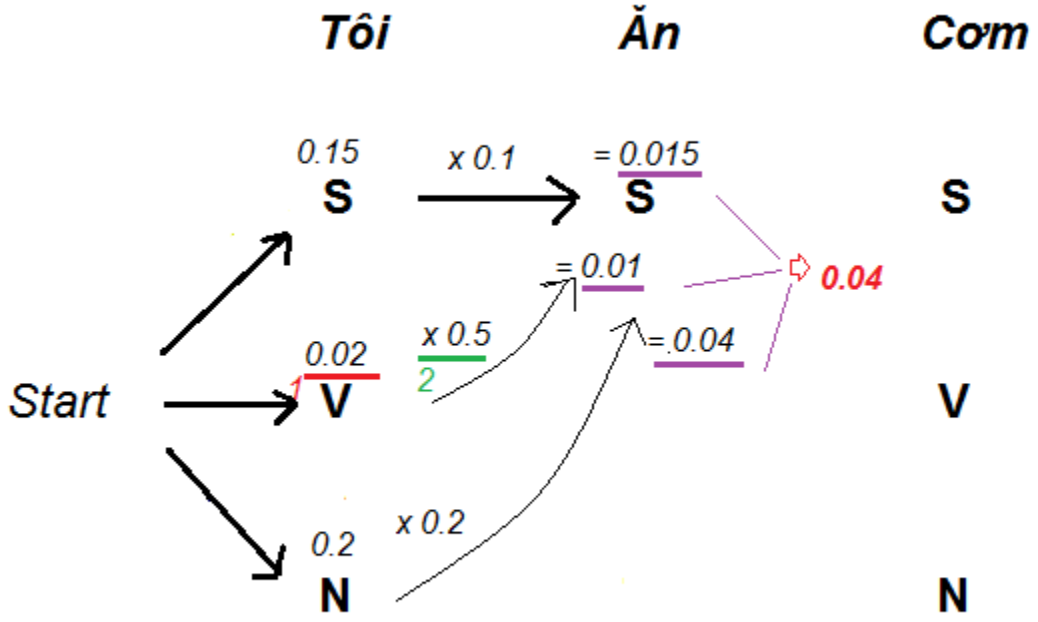
1: nhìn bảng Nhãn-nhãn: S-Start: 0.3

2 : Nhìn bảng Từ nhãn: Tôi – S: 0.5

⇨ Nhân ra vậy node S = 0.3 x 0.5 = 0.15

Tương tự 2 nhãn V và N nhân ra 0.02 và 0.2

B2:



(Khác bước 1)

1 : Số vừa tính ở node V khi nãy 0.02

2 mũi tên từ V -> S thì nhìn bảng Nhãn nhãn: ***Cột S hàng V*** *S-V:0.5*

⇨ ***0.02 * 0.5 = 0.01***

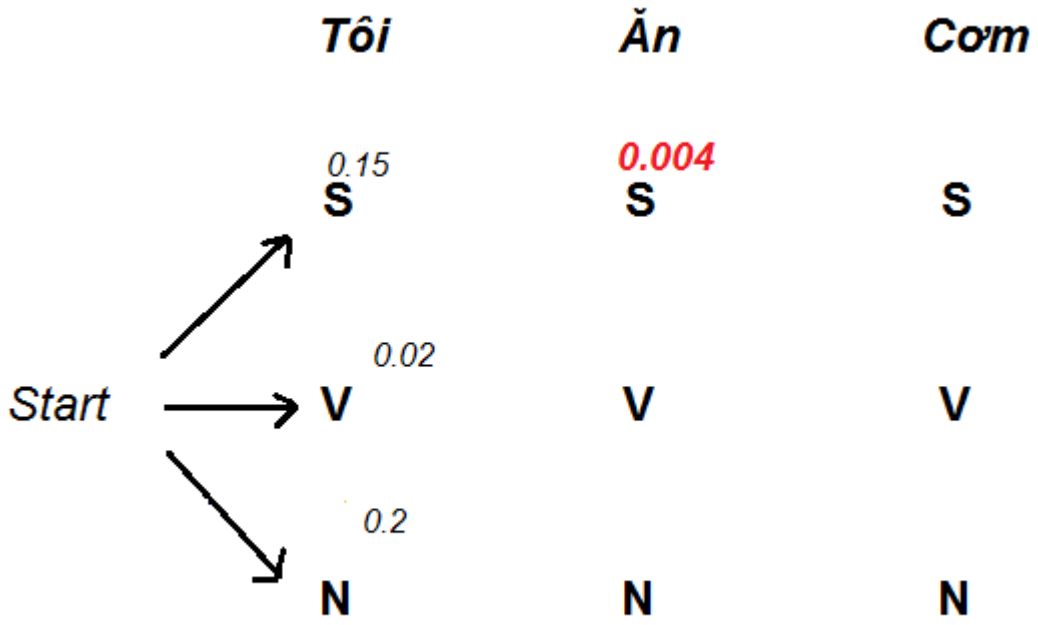
*Tính tương tự được S->S=****0.015***

*V->S=****0.01***

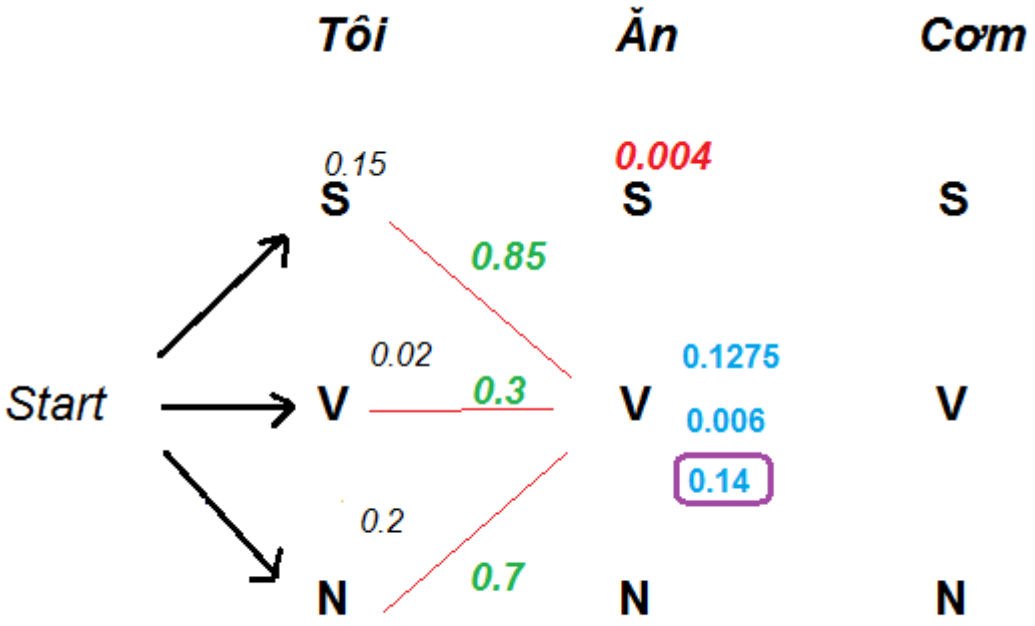
*N->S=**0,04***

- ⇨ *0.04 lớn nhất. Lấy 0.04*
- ⇨ *Ta đang tính cột "Ăn" => nhìn bảng Ăn-S : 0.1*
- ⇨ *0.04 * 0.1 = 0.004*

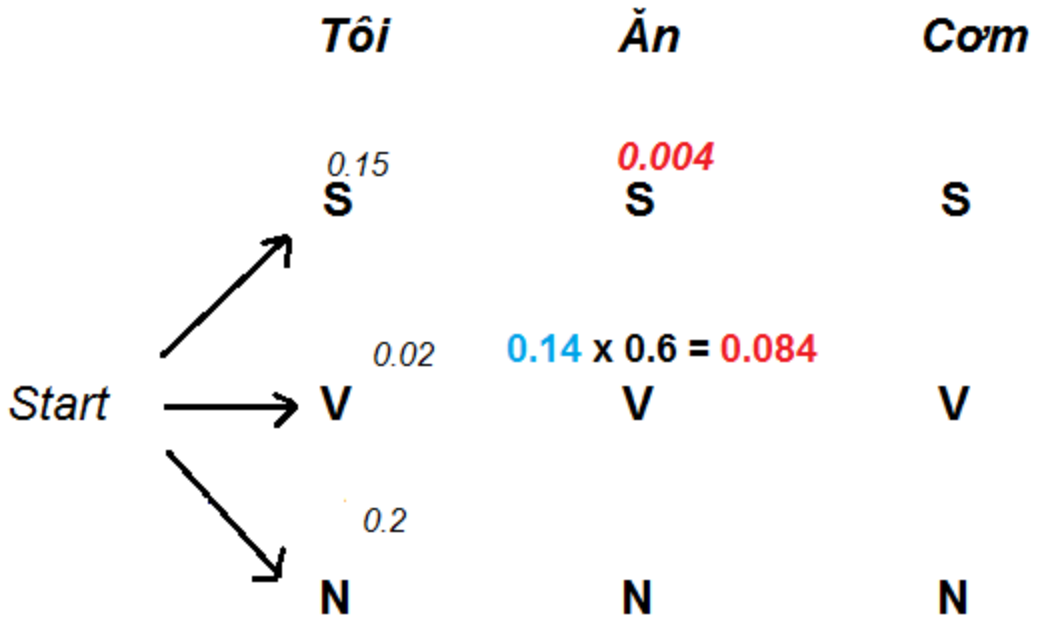
Nó sẽ ra như vầy



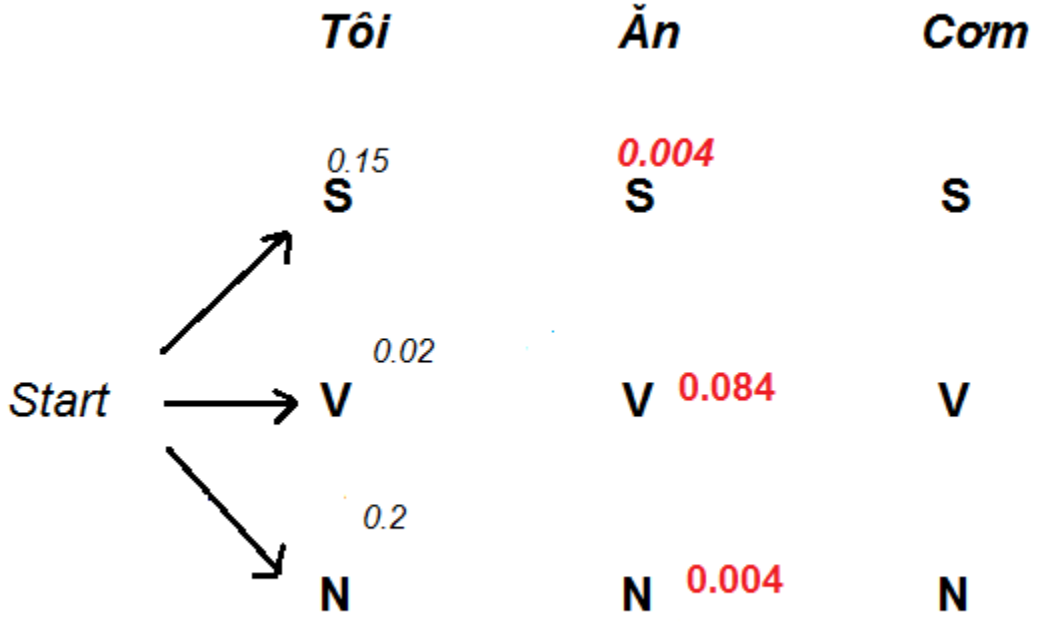
Tiếp tục y vậy làm với 2 node Ăn-V



0.14 lớn nhất lấy 0.14 ra và nhân với Ăn-V:0.6 = 0.084

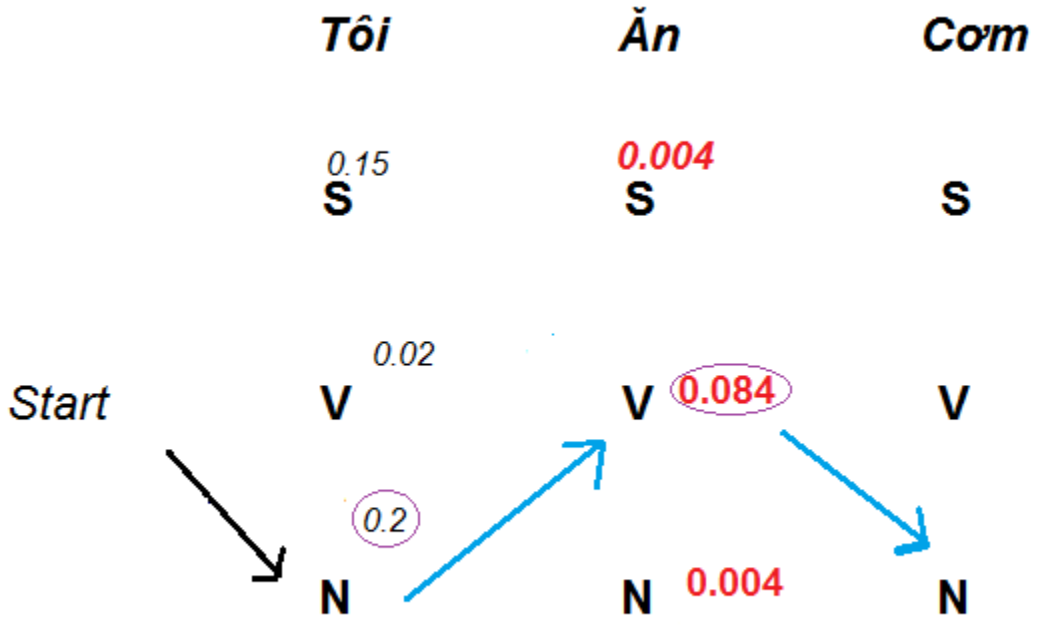


Làm tiếp cho ra như vẦY



Thức mệt quá bạn làm tương tự cái cột "Cơm" luôn. Tương tự như trên

Sau đó cuối cùng vẻ đường đi từ Start qua các nhãn. Đi theo số lớn nhất như vầy nè



- ***Rồi kết luận input "Tôi ăn cơm" đc gán nhãn Tôi/N ăn/V cơm/N dùng HMM như thế đấy***
- ***Thức cho xác xuất ko chuẩn đáng lẽ kết quả ra Tôi phải là S chứ ko phải N***