Học viện Công nghệ Bưu chính Viễn thông
Khoa Công nghệ thông tin 1

Xử lý ngôn ngữ tự nhiên
(Natural Language Processing-NLP)

Tách từ & Gán nhãn từ loại
(Word Segmentation & Part-of-Speech Tagging)

Ngô Xuân Bách

Nội dung

- Tách từ tiếng Việt
- Gán nhãn từ loại (tiếng Anh)

Tách từ tiếng Việt

- Tiếng Việt vs. Tiếng Anh
  - Tiếng Anh:
    - Từ được ngăn cách bởi khoảng trắng (space)
  - Tiếng Việt:
    - Từ có thể gồm một hoặc nhiều âm tiết, các âm tiết được ngăn cách bởi khoảng trắng

- Tách từ tiếng Việt
  - Là bước xử lý cơ bản trong xử lý văn bản tiếng Việt
  - Chất lượng của hệ tách từ ảnh hưởng trực tiếp tới các bước xử lý tiếp sau như gán nhãn từ loại, phân tích cú pháp, phân tích ngữ nghĩa, cũng như các ứng dụng xử lý ngôn ngữ tự nhiên

Bài toán tách từ tiếng Việt

- Input
  - Là một câu tiếng Việt (gồm một chuỗi âm tiết được cách nhau bởi khoảng trắng)
- Output
  - Là câu mà đã xác định ranh giới giữa các từ (câu gồm một chuỗi các từ)
- Ví dụ
  - Input: Để đóng vai quần chúng chúng tôi phải mặc quần áo theo lối phong kiến
  - Output: Để đóng vai quần_chúng chúng_tôi phải mặc quần_áo theo lối phong_kiến

Cách tiếp cận dựa trên từ điển

- Phương pháp Longest matching
  - Duyệt câu đầu vào tuần tự từ trái qua phải và chọn từ dài nhất nếu từ đó có trong từ điển
  - Ví dụ
    - Input: Đó là cách để truyền thông tin
    - Output: Đó là cách để truyền_thông tin

Cách tiếp cận dựa trên từ điển

- Phương pháp Maximal matching
  - Sử dụng từ điển, tạo ra tất cả các cách tách từ có thể cho một câu bất kỳ, sau đó câu được tách từ đúng được chọn là câu chứa ít từ nhất
  - Ví dụ
    - Input: Đó là cách để tuyên truyền thông tin
    - Output: Đó là cách để tuyên_truyền thông_tin

Đặc điểm của cách tiếp cận dựa trên từ điển

- Ưu điểm
  - Đơn giản, dễ cài đặt
- Nhược điểm
  - Phụ thuộc vào từ điển
  - Độ chính xác không cao
  - Không giải quyết được trường hợp nhập nhằng

Cách tiếp cận dựa trên thống kê

- Coi bài toán tách từ là bài toán gán nhãn chuỗi
- Input
  - Là một câu, được biểu diễn dưới dạng một chuỗi các âm tiết
- Output
  - Là một xâu nhãn, mỗi nhãn tương ứng với một âm tiết
  - Nhãn B_W: gán cho các âm tiết bắt đầu một từ
  - Nhãn I_W: gán cho các âm tiết còn lại
  - Nhãn O: gán cho các ký tự đặc biệt
- Ví dụ

Đó	là	cách	để	truyền	thông	tin	.
B_W	B_W	B_W	B_W	B_W	B_W	I_W	O

Cách tiếp cận dựa trên thống kê

- Sử dụng các thuật toán học máy thống kê trên tập dữ liệu có gán nhãn (kho ngữ liệu – corpus) để xây dựng mô hình tách từ

- Một số thuật toán học máy
  - Mô hình Markov ẩn (Hidden Markov Model)
  - Mô hình cực đại Entropy (Maximum Entropy Markov Model)
  - Trường ngẫu nhiên (Conditional Random Field)

Đánh giá kết quả tách từ

- Precision (độ chính xác)

$$Precision = \frac{Số\ lượng\ từ\ tách\ đúng}{Số\ lượng\ từ\ được\ tách}$$

- Recall (độ bao phủ)

$$Recall = \frac{Số\ lượng\ từ\ tách\ đúng}{Số\ lượng\ từ\ thực\ tế}$$

- F1

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Gán nhãn từ loại (tiếng Anh)

Phân tích hình vị tiếng Anh

- Chia một từ thành các hình vị
  - unkindly = un + kind + ly

- Xác định nhãn từ loại của các từ
  - Nhiều từ tiếng Anh có thể có nhiều từ loại
    - Phần lớn các danh từ có thể được sử dụng như là động từ
    - Book, play, work, …
  - Trong một câu (ngữ cảnh) cụ thể, mỗi từ có một nhãn từ loại xác định
  - Xác định đúng nhãn từ loại của từ trong câu đóng vai trò quan trọng trong xử lý ngôn ngữ tự nhiên

Gán nhãn từ loại

- Đầu tiên xác định các nhãn từ loại ứng viên cho các từ bằng cách sử dụng từ điển

- Sau đó xếp hạng (rank) các ứng viên
  - Dựa vào sự xuất hiện của các nhãn từ loại
    - Ví dụ: breakfast: danh từ hoặc động từ
    - `breakfast' thường được sử dụng là danh từ hơn
  - Dựa vào các nhãn từ loại xung quanh
    - Ví dụ: the (từ hạn đinh) breakfast (danh từ hoặc động từ)
    - Danh từ thường xuất hiện sau một từ hạn định

Làm sao để xác định nhãn đúng?

- Bằng tay (người làm)

- Tự động
  - Với bài toán gãn nhãn từ loại tiếng Anh, các phương pháp dựa trên mô hình xác suất đã được sử dụng thành công từ vài thập kỷ nay

Mô hình xác suất cho gán nhãn từ loại

- $P(C_1 \dots C_n | w_1 \dots w_n)$
  - C_i là nhãn từ loại, w_i là từ
  - Ta cần tìm $C_1 \dots C_n$ sao cho xác suất trên là cực đại

- Sử dụng quy tắc Bayes

$$P(C_1 \dots C_n | w_1 \dots w_n) = \frac{P(C_1 \dots C_n) P(w_1 \dots w_n | C_1 \dots C_n)}{P(w_1 \dots w_n)}$$

  - Do mẫu số không phụ thuộc vào nhãn, ta cần tìm $C_1 \dots C_n$ sao cho tử số là cực đại

Ví dụ

- Chuỗi từ: time flies like an arrow

- Chuỗi nhãn 1: $n\ v\ prep\ \det\ n$

$$P(n\ v\ prep\ det\ n \mid time\ flies\ like\ an\ arrow) = \frac{P(n\ v\ prep\ det\ n)P(time\ flies\ like\ an\ arrow \mid n\ v\ prep\ det\ n)}{\cancel{P(time\ flies\ like\ an\ arrow)}}$$

- Chuỗi nhãn 2: $n\ n\ v\ \det\ n$

$$P(n\ n\ v\ det\ n \mid time\ flies\ like\ an\ arrow) = \frac{P(n\ n\ v\ det\ n)P(time\ flies\ like\ an\ arrow \mid n\ n\ v\ det\ n)}{\cancel{P(time\ flies\ like\ an\ arrow)}}$$

Xấp xỉ mô hình xác suất

- $P(C_1 \dots C_n)P(w_1 \dots w_n|C_1 \dots C_n)$

- Hạng thức thứ nhất

$$P(C_1 \dots C_n) = P(C_1)P(C_2|C_1)P(C_3|C_2C_1) \dots P(C_n|C_1 \dots C_{n-1})$$
$$\approx P(C_1)\prod_{i=2}^{n} P(C_i|C_{i-1})$$

- Hạng thức thứ hai

$$P(w_1 \dots w_n|C_1 \dots C_n) \approx \prod_{i=1}^{n} P(w_i|C_i)$$

Xấp xỉ mô hình xác suất

- Tổng hợp lại

$$P(C_1 \dots C_n)P(w_1 \dots w_n|C_1 \dots C_n) = \prod_{i=1}^{n} P(C_i|C_{i-1}) \prod_{i=1}^{n} P(w_i|C_i)$$

- Ở đây $P(C_1) = P(C_1|C_0) = P(C_1|\emptyset)$
- C_0 (hay $\emptyset$) là ký tự đặc biệt, ký hiệu đầu câu

Ví dụ: Time flies like an arrow

- Chuỗi nhãn 1: $n\ v\ prep\ \det n$

$$P(n\ v\ prep\ \det n)$$
$$\approx P(n|\emptyset)P(v|n)P(prep|v)P(\det|prep)P(n|\det)$$
$$P(time\ flies\ like\ an\ arrow|n\ v\ prep\ \det n)$$
$$\approx P(time|n)P(flies|v)P(like|prep)P(an|\det)P(arrow|n)$$

- Chuỗi nhãn 2: $n\ n\ v\ \det n$

$$P(n\ n\ v\ \det n) \approx P(n|\emptyset)P(n|n)P(v|n)P(\det|v)P(n|\det)$$
$$P(time\ flies\ like\ an\ arrow|n\ n\ v\ \det n)$$
$$\approx P(time|n)P(flies|n)P(like|v)P(an|\det)P(arrow|n)$$

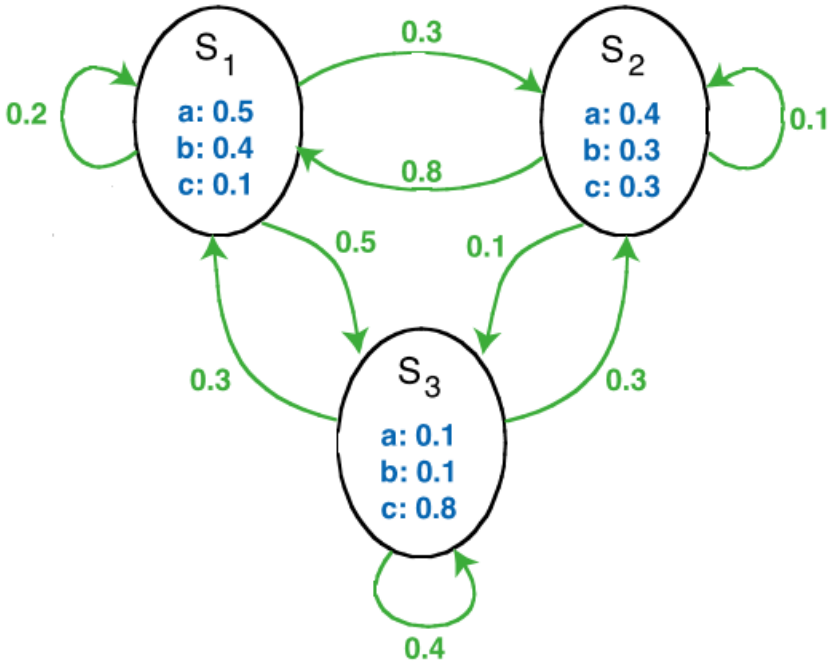
Hidden Markov Model

- Mô hình xác suất cho gán nhãn từ loại $\prod_{i=1}^{n} P(C_i|C_{i-1}) \prod_{i=1}^{n} P(w_i|C_i)$ là mô hình Markov ẩn (Hidden Markov Model - HMM)

- Mô hình Markov ẩn (HMM) là gì ?
  - Là mô hình xác suất cho một chuỗi các trạng thái (states) và các ký hiệu đầu ra (output symbols)
    - Xem xét phép chuyển trạng thái trong đó một ký hiệu đầu ra được sinh ra tại mỗi bước
  - Được áp dụng thành công cho bài toán gán nhãn từ loại, nhận dạng tiếng nói, và nhiều bài toán khác

Hidden Markov Model

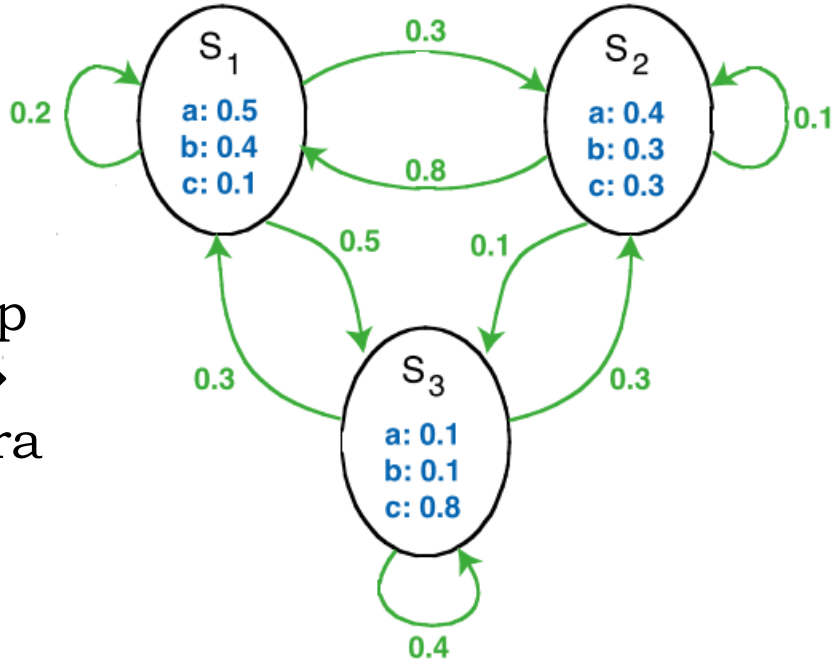
- Sơ đồ chuyển trạng thái của HMM
  - Đường màu xanh lá cây: xác suất chuyển trạng thái
  - Chữ màu xanh da trời: xác suất sinh ký hiệu đầu ra tại mỗi trạng thái



Hidden Markov Model

- Sơ đồ chuyển trạng thái của HMM
  - Đường màu xanh lá cây: xác suất chuyển trạng thái
  - Chữ màu xanh da trời: xác suất sinh ký hiệu đầu ra tại mỗi trạng thái

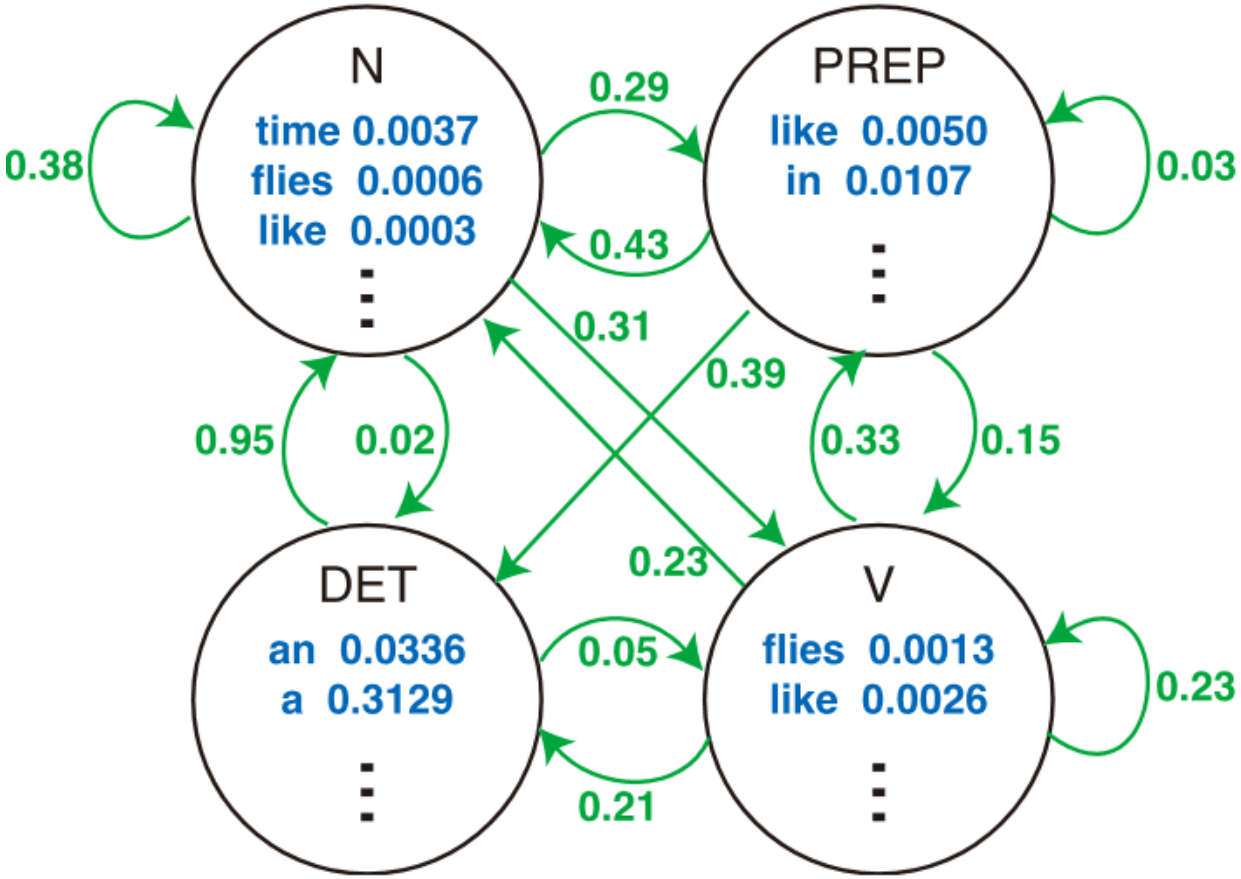
Giả sử:
$P(S_1|\emptyset) = 0.7;$
$P(S_2|\emptyset) = 0.2;$
$P(S_3|\emptyset) = 0.1.$
Tính xác suất trong trường hợp phép chuyển trạng thái là $S_1 \rightarrow S_2 \rightarrow S_3$ và chuỗi ký hiệu đầu ra được sinh ra là "abc" ?



Định nghĩa hình thức của HMM

- $< S, K, \Pi, A, B >$
  - S: tập các trạng thái
  - K: tập các ký hiệu đầu ra
  - Π: tập các xác suất của trạng thái khởi tạo $P(S_i)$
  - A: tập các xác suất chuyển trạng thái $P(S_j|S_i)$
  - B: tập các xác suất sinh ký hiệu đầu ra $P(a_i|S_i)$

HMM cho gán nhãn từ loại



$P(N|\phi)=0.57$
$P(DET|\phi)=0.28$
$P(PREP|\phi)=0.11$
$P(V|\phi)=0.04$
ϕ=beginning of a sentence

HMM cho gán nhãn từ loại

- $< S, K, \Pi, A, B >$
  - S: tập các nhãn từ loại
  - K: tập các từ
  - Π: $P(C_i|\emptyset)$
  - A: $P(C_j|C_i)$
  - B: $P(w_i|C_i)$

Huấn luyện HMM

- Tính các xác suất của HMM thế nào?
- Phương pháp sử dụng kho ngữ liệu đã được gán nhãn từ loại
  - Kho ngữ liệu gán nhãn từ loại
    - Tập hợp các câu trong đó mỗi một từ đã được gán nhãn từ loại đúng
  - $P(C_j|C_i)$ có thể được ước lượng từ tần suất xuất hiện đồng thời của các cặp nhãn từ loại
  - $P(w_i|C_i)$ có thể được ước lượng từ tần suất của từ
  - Là phương pháp học có giám sát
  - Tương đối khó khăn trong việc xây dựng (thu thập) dữ liệu huấn luyện

Bảng xác suất

$P(C_j|C_i)$

C_j \ C_i	ϕ	N	V	DET	PREP
N	392	1111	326	1050	605
	0.57	0.38	0.23	0.95	0.43
V	28	918	313	52	204
	0.04	0.31	0.23	0.05	0.15
DET	194	78	289	0	541
	0.28	0.03	0.21	0.00	0.39
PREP	71	840	456	0	38
	0.11	0.28	0.33	0.00	0.03
Total	685	2947	1384	1102	1388
	1.00	1.00	1.00	1.00	1.00

Bảng xác suất

$P(w_i|C_i)$

w_i \ C_i	N	V	DET	PREP
time	13	0	0	0
	0.0044	0.0000	0.0000	0.0000
flies	2	2	0	0
	0.0007	0.0014	0.0000	0.0000
like	1	4	0	7
	0.0003	0.0029	0.0000	0.0050
...	...	...	...	...
Total	2947	1384	1102	1388
	1.0000	1.0000	1.0000	1.0000

Huấn luyện HMM

- Phương pháp sử dụng dữ liệu chưa gán nhãn (plain text)
  - Dữ liệu chưa gán nhãn (Plain text)
    - Tập hợp các câu không có bất kỳ thông tin nào thêm ví dụ như nhãn từ loại
  - HMM có thể được huấn luyện sử dụng thuật toán forward-backward
  - Là phương pháp học không giám sát
  - Mô hình xây dựng được sau khi huấn luyện có thể không thật sự tốt

Gán nhãn từ loại với HMM

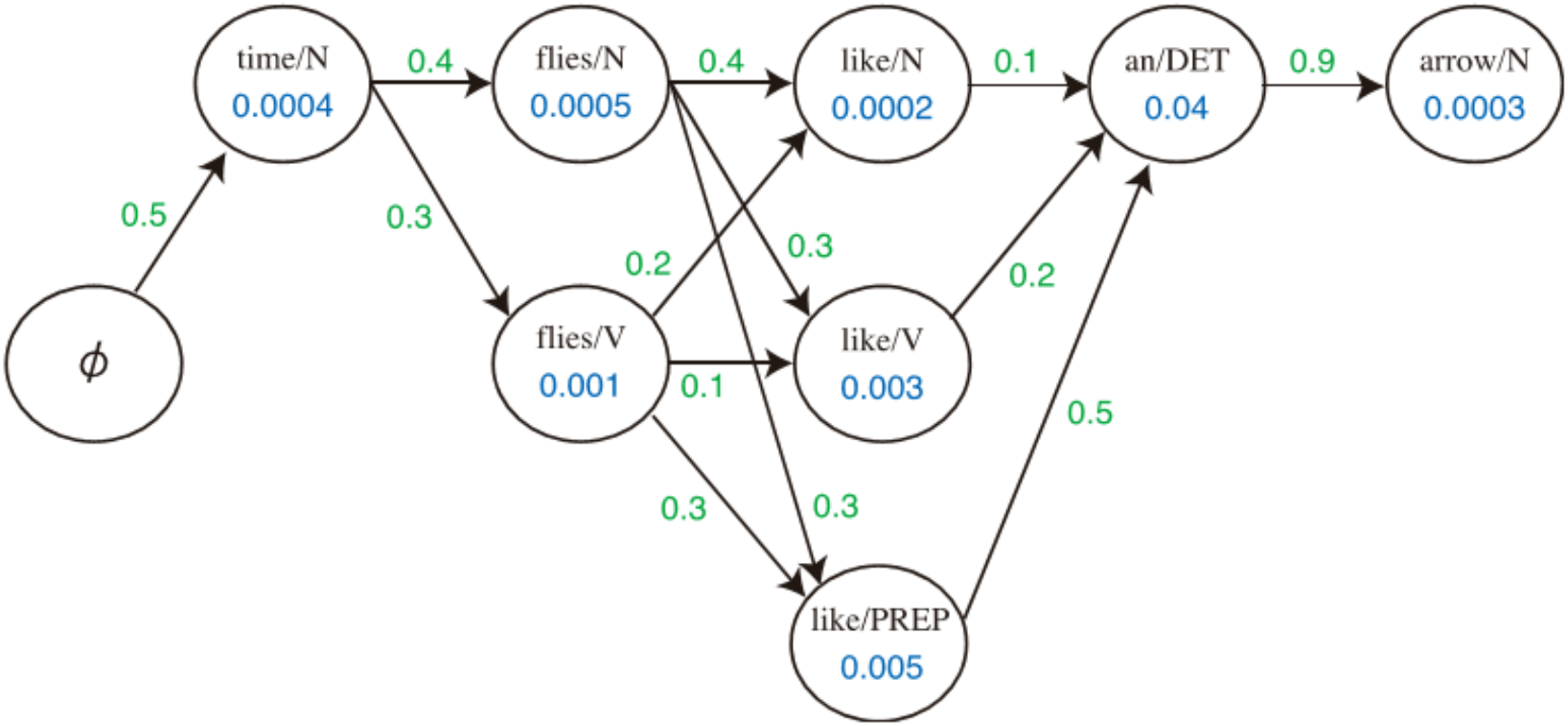
- Xây dựng lưới từ
  - Sử dụng từ điển, lấy các nhãn từ loại có thể cho một từ, sau đó thêm vào các nút tương ứng
  - Tạo đường liên kết giữa các nút
  - Một đường đi (trong lưới từ) là một cách gán nhãn từ loại

word dictionary

word	*POS*
an	DET
arrow	N
flies	N,V
like	N,V,P
time	N

- Chọn cách gán nhãn đúng
  - Đưa xác suất $P(w_i|C_i)$ vào các nút, $P(C_j|C_i)$ vào các liên kết
  - Tìm đường đi khả dĩ nhất (xác suất cao nhất)
    - Sử dụng thuật toán Viterbi

Xây dựng lưới từ

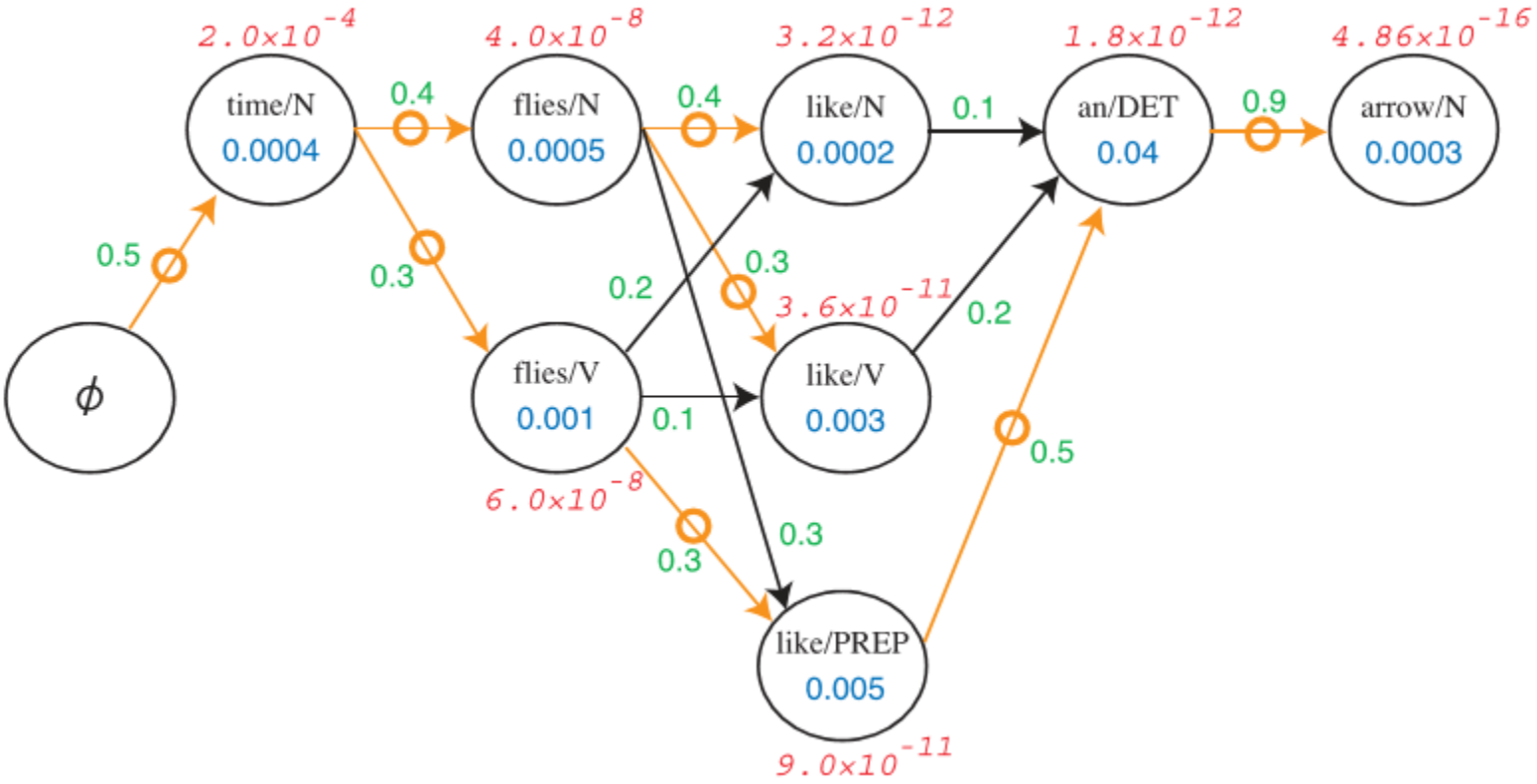


Thuật toán Viterbi

- **for** $i = 1$ **to** n
  - Ghi lại xác suất cục bộ (local) tốt nhất cho mỗi nút X tại vị trí i
    - Xác suất lớn nhất trong số các xác suất của tất cả các đường đi tới nút X
    - Đồng thời ghi lại đường đi tương ứng với xác suất lớn nhất này
  - Tính xác suất cục bộ tốt nhất
    - (xác suất cục bộ tốt nhất của nút Y tại vị trí $i-1$) × (xác suất chuyển trạng thái từ Y sang X) × (xác suất sinh ký hiệu đầu ra tại nút X)
    - Chọn xác suất cao nhất cho mỗi nút Y tại vị trí $i-1$

- Để tìm đường đi khả dĩ nhất cho một câu đầu vào, chọn nút có xác suất cục bộ tốt nhất là lớn nhất trong số các nút tại vị trí n

Thuật toán Viterbi



- Local best probabilities of nodes are written in italic
- ○ stands for the best path

Độ phức tạp tính toán

- Khi tính xác suất của tất cả các đường đi
  - $O(c^n)$
    - c: số lượng nhãn từ loại
    - n: số lượng từ trong câu đầu vào
- Khi sử dụng thuật toán Viterbi
  - $O(c^2 n)$
    - Nhanh hơn rất nhiều

Đánh giá kết quả gán nhãn từ loại

- Độ chính xác (accuracy)

$$Accuracy = \frac{Số\ lượng\ từ\ gán\ nhãn\ đúng}{Số\ lượng\ từ\ được\ gán\ nhãn}$$

- Có thể đánh giá riêng cho từng loại nhãn từ loại
  - Sử dụng Precision, Recall, và F_1

Tóm tắt

- Tách từ là bài toán cơ bản trong xử lý tiếng Việt
  - Các phương pháp dựa trên từ điển đơn giản, cho độ chính xác hạn chế
  - Các phương pháp dựa trên thống kê yêu cầu dữ liệu huấn luyện, cho độ chính xác cao hơn
- Gán nhãn từ loại là vấn đề quan trọng trong phân tích hình vị tiếng Anh
  - Mô hình HMM đã được áp dụng rất thành công
  - HMM có thể được huấn luyện tự động từ dữ liệu
  - Thuật toán Viterbi cho phép tìm chuỗi nhãn từ loại tốt nhất một cách hiệu quả