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

Project 2

Gán nhãn từ loại sử dụng Hidden Markov Model

(Part-of-speech Tagging)

Giảng viên hướng dẫn : TS. Ngô Xuân Bách
Sinh viên thực hiện : Lê Thị Ngọc Châm
Lớp : D13CN6

Nội dung

- Tổng quan về dán nhãn từ loại.
- Hidden Markov Model (HMM).
- Áp dụng HMM vào bài toán gán nhãn từ loại.
- Tổng kết.

Nội dung

- Tổng quan về dán nhãn từ loại.
- Hidden Markov Model (HMM).
- Áp dụng HMM vào bài toán gán nhãn từ loại.
- Tổng kết.

Tổng quan về gán nhãn từ loại

- Gán nhãn từ loại là gì?
- Một số phương pháp dùng để gán nhãn từ loại Tiếng Anh.

Tổng quan về gán nhãn từ loại

❑ Gán nhãn từ loại là gì? [1]

- Gán nhãn từ loại là việc xác định các chức năng ngữ pháp của từ trong câu. Đây là bước cơ bản trước khi phân tích sâu văn phạm hay các vấn đề xử lý ngôn ngữ phức tạp khác.

- Thông thường, một từ có thể có nhiều chức năng ngữ pháp, ví dụ : "the book is interesting" và "We book a room"

  Từ "book" trong 2 câu trên về hình thức thì giống nhau nhưng chức năng ngữ pháp và ngữ nghĩa thì khác nhau.
  "the book is interesting" => "book" : quyển sách _ danh từ.
  "We book a room" => "book" : đặt trước _ động từ.

Tổng quan về gán nhãn từ loại

- Một số phương pháp dùng để gán nhãn tư loại Tiếng Anh.[1]
  - Gắn nhãn dựa trên mô hình Markov ẩn (MHH)
  - Các mô hình dựa trên bộ nhớ(Daelemans, 1996)
  - Mô hình dựa trên luật (Transformation Base Learning, Brill, 1995)
  - Maximum Entropy Markov Model (MEMM)
  - Cây quyết định (Schmid, 1994a)
  - Mạng nơ-ron(Schmid, 1994b).

Nội dung

- Tổng quan về dán nhãn từ loại.
- Hidden Markov Model (HMM).
- Áp dụng HMM vào bài toán gán nhãn từ loại.
- Tổng kết.

Hidden Markov Model (HMM)[2]

- Mô hình xác suất cho gán nhãn từ loại
- Xấp xỉ mô hình xác suất
- Hidden Markov Model (HMM).

Mô hình xác suất cho gán nhãn từ loại

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

- Sử dụng quy tắc Bayes

$$P(C_1 \ldots C_n | w_1 \ldots w_n) = \frac{P(C_1 \ldots Cn) P(w_1 \ldots w_2 | C_1 \ldots Cn)}{P(w_1 \ldots w_2)}$$

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

Xấp xỉ mô hình xác suất

- $P(C_1 \ldots C_n)P(w_1 \ldots w_n|C_1 \ldots C_n)$
- Hạng thức thứ nhất:

$$P\ (C_1 \ldots C_n) = P(C_1)\ P(C_2|C_1)\ P(C_3|C_2C_1) \ldots P(C_n|C_1 \ldots C_{n-1})$$

$$\approx P(C_1) \prod_{i=2}^{n} P(C_i \mid C_{i-1})$$

- Hạng thức thứ hai:

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

Xấp xỉ mô hình xác suất

- Tổng hợp lại:

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

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

Hidden Markov Model

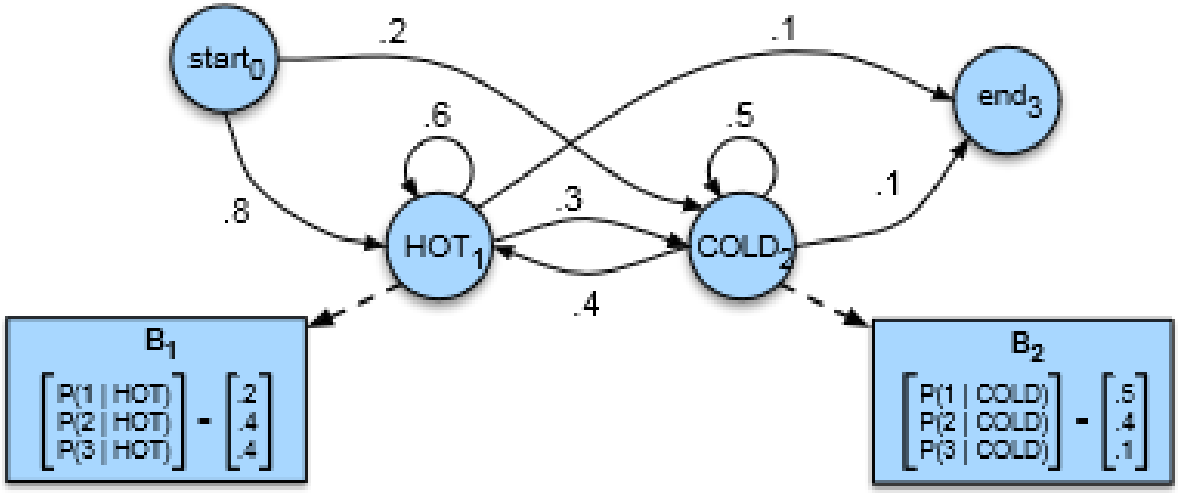
- Mô hình xác suất cho gán nhãn từ loại $\prod_{i=1}^{n} P(C_i|C_{i-1}) \prod_{i=1}^{n} P(w_i|C_i)$ là mô hình Markov ẩn.

- Mô hình Markov ẩn 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

- Các đường: xác suất chuyển trạng thái.
- Bảng : xác suất sinh ký hiệu đầu ra tại mỗi trạng thái.



Định nghĩa hình thức của HMM

- <S,K,Π,A,B>
  - S : tập các trạng thái
  - K : Tập các ký hiệu đầu ra
  - Π : Tập các xác suất của trạng thái khởi tạo
  - A : Tập các xác suất chuyển trạng thái
  - B : Tập các xác suất sinh ký hiệu đầu ra.

Nội dung

- Tổng quan về dán nhãn từ loại.
- Hidden Markov Model (HMM).
- Áp dụng HMM vào bài toán gán nhãn từ loại.
- Tổng kết.

Áp dụng HMM vào gán nhãn từ loại

- Huấn luyện HMM
- Gán nhãn từ loại với HMM
- Thuật toán Viterbi
- Thực nghiệm

Huấn luyện HMM

- Phương pháp sử dụng kho dữ liệu bao gồm 53020 câu Tiếng Anh và được gán nhãn từ loại đúng.

- $P(C_j|C_i)$ có thể ước lượng từ tần suất xuất hiện đồng thời của cặp nhãn từ loại.

$$P(C_j|C_i) = \frac{\text{Số cặp } C_iC_j}{\text{Tổng số nhãn } C_i}$$

- $P(w_i|C_i)$ có thể ước lượng từ tần suất của từ.

$$P(w_i|C_i) = \frac{\text{Số lần wi có nhãn} C_i}{\text{Tổng số từ } C_i}$$

Từ điển (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

Từ điển (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
...	...	...	...	...

Từ điển (bảng xác suất)

- Làm giảm độ lớn của từ điển bằng cách loại đi các số (1.222 hoặc 1,232) hoặc một số ký tự mà nhãn của của nó trong câu là chính nó. Với số thì nhãn của nói luôn luôn là CD.

Gán nhãn từ loại với HMM

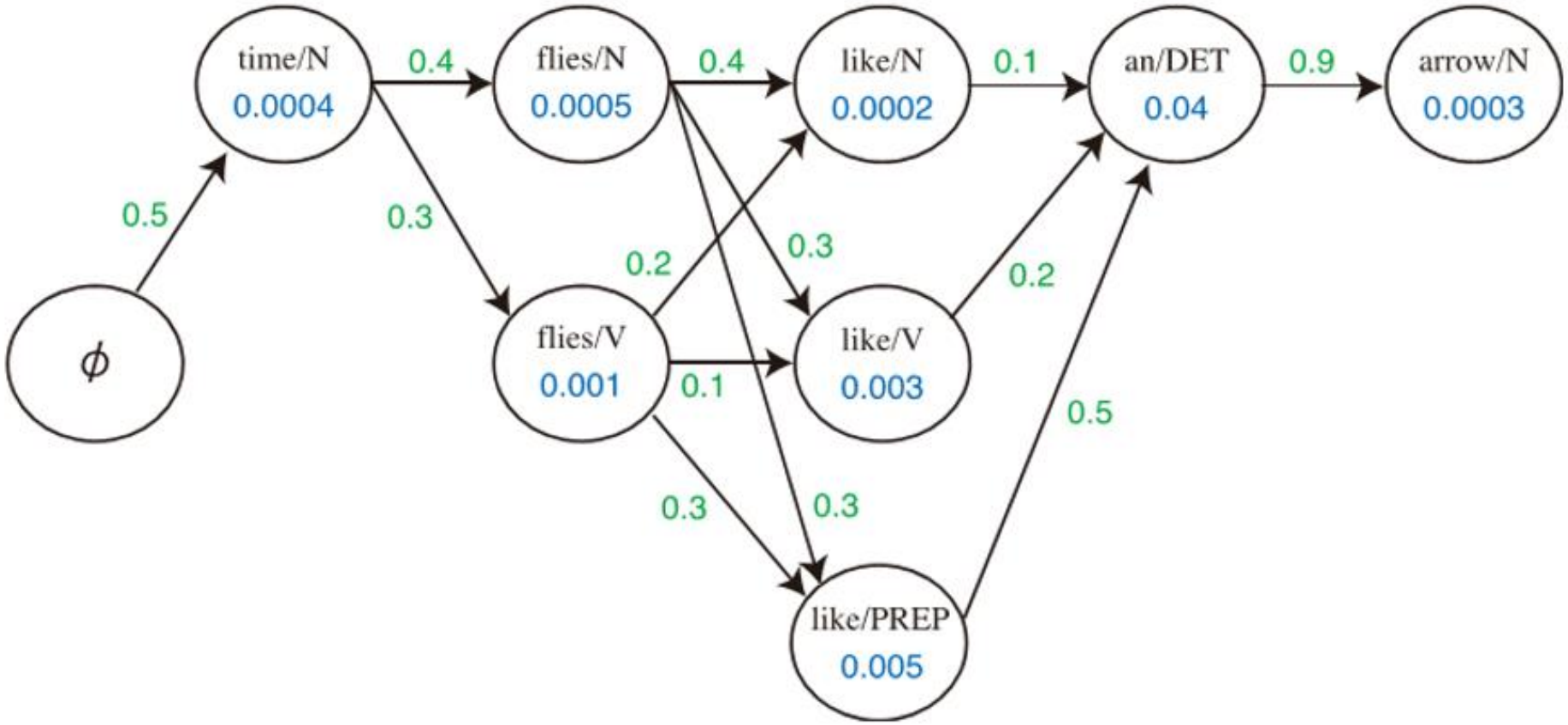
➢ Xây dựng lưới từ

- Sử dụng từ điển, lấy các nhãn từ loại có thể có 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

➢ Chọn cách gán nhãn đúng

- Đưa xác suất $P(w_i|C_i)$ vào 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

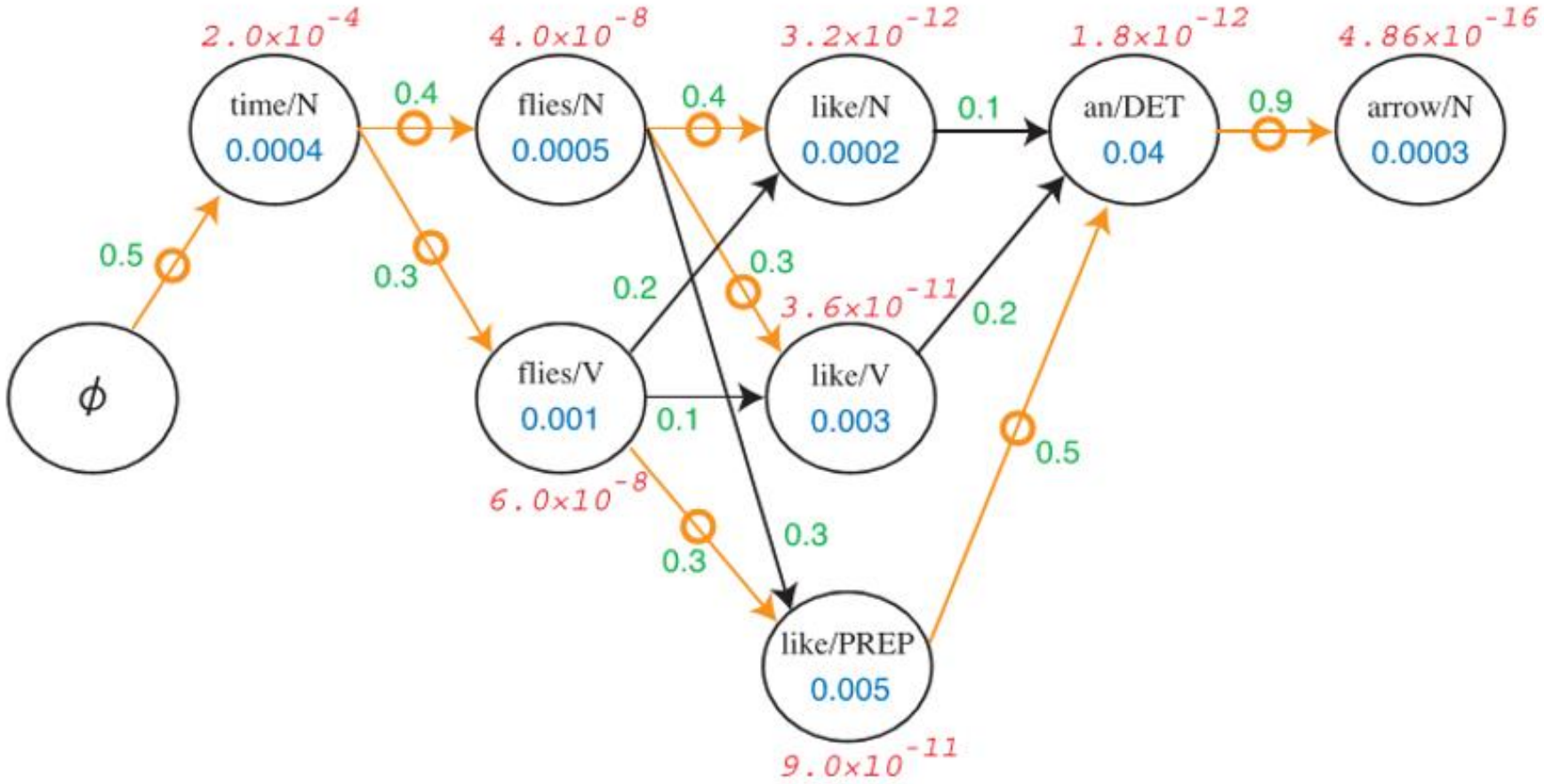
Xây dựng lưới từ



Thuật toán Viterbi

- **For i = 1 to n**
  - Ghi lại xác suất cục bộ 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) x (xác suất chuyển trạng thái từ Y sang X) x (xác suất sinh ký hiệu đầu ra tại nút X)
    - Chọn xác suất cao nhất cho nút Y tại vị trí i-1
- Để tìm đường đi khả dĩ nhất cho một câu nào đầu vào, chọn nút co xác suất cục bộ tốt nhất là lớn nhất trong số các nút tại vị trí n.

Thuật toán Viterbi



Thực nghiệm

Đánh giá riêng cho từng loại nhãn với các độ đo

- Precision (độ chính xác)

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

- Recall (độ bao phủ)

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

- F1

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Kết quả thực nghiệm

- Mean F1

JJ	90.15
NN	95.10
``	99.99
WRB	99.86
LS	2.86
PRP	98.83
DT	98.02
NNP	95.56
FW	49.48
NNS	96.47
JJS	94.31
JJR	86.03
UH	53.72
MD	99.60
VBD	93.36
WP	99.40
VBG	90.14
CC	99.44
''	95.77
CD	99.06
PDT	74.57
RBS	83.85
VBN	84.77

RBR	69.20
#	100.00
$	99.99
VBP	89.56
IN	97.06
WDT	93.77
SYM	88.85
NNPS	60.16
(	100.00
)	100.00
WP$	100.00
VB	93.62
,	99.99
.	100.00
VBZ	96.40
RB	89.46
PRP$	99.71
EX	97.38
POS	94.38
:	99.98
TO	99.88
RP	52.43

- Mean Accuracy

Mean Accuracy
95.56

Nhận xét kết quả

➢ Phân tích nhãn NNPS:

- Có 60.16% phân loại đúng vào nhãn NNPS, 22.15% phân loại sai vào nhãn NNP, 8.15 % là vào NNS và còn lại là các nhãn khác.
- Có 30 nhãn có khả năng đứng trước NNPS tất cả các nhãn này xác suất chuyển nhãn sang NNP đều cao hơn so với NNPS.
- Có 34.56 % từ có nhãn NNPS cũng có khả năng thuộc về nhãn NNP và 17.99% từ cũng có khả năng thuộc về nhãn NNS.
- Trong số các từ có thể là nhãn NNPS thì 24.12% từ xác suất là nhãn NNP lơn hơn hoặc bằng.

Từ những số liệu trên ta có thể thấy được sự nhập nhằng về ngôn ngữ đã dẫn đến kết quả thấp của việc gán nhãn.

Nhận xét kết quả

➢ Tương tự với nhãn RBR.

- Có 69.20 % phân loại đúng vào nhãn RBR, 24.21% phân loại sai vào nhãn JJR và 5.82% vào RB còn lại là các nhãn khác.
- Xác suất chuyển nhãn từ các nhãn khác sáng RBR đều có sự chênh lệch rất nhỏ so với chuyển sang nhãn JJR.
- Có 71.17% từ có nhãn RBR có thể là nhãn JJR và có 49.82% từ có khả năng là RBR thì xác suất là JJR lớn hơn hoặc bằng.

Những số liệu trên chỉ ra sự tương đương về vị trí ngữ pháp trong câu cũng như là ý nghĩa của các từ có nhãn RBR và JJR.

Danh mục tài liệu tham khảo

Silde được hoàn thành dựa vào các tài liệu tham khảo

[2] Slide "Xử lý ngôn ngữ tự nhiên – Tách từ và gán nhãn từ loại" _ TS. Ngô Xuân Bách. Học viện Công Nghệ bưu chính viễn thông.

- Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright 2014. Allrights reserved. Draft of September 1, 2014.

[1] Gán nhãn từ loại tiếng Việt dựa trên các phương pháp học máy thống kê. Phạm Xuân Hiếu(Trường Khoa học thông tin, Đại học Tohoku, Nhật Bản) , Lê Minh Hoàng (Đại học Sư Phạm Hà Nội), Nguyễn Cẩm Tú(Đại học Công nghệ, Đại học Quốc gia Hà Nội)