BÀI 6: TRÍCH CHỌN THÔNG TIN

AI Academy Vietnam

Thông tin giảng viên

- TS. Nguyễn Minh Tiến
- Email: minhtienhy@gmail.com
- Mobile: 0983 860 318

- Research interest: natural language processing, information extraction.

Nội dung buổi học

- Giới thiệu về trích chọn thông tin (Information Extraction - IE)
- Gán nhãn chuỗi (Sequence labeling)
  - Gán nhãn từ loại (POS tagging)
  - Nhận dạng thực thể (Named Entity Recognition - NER)
- Mô hình Markove ẩn (Hidden Markov Models - HMMs)
- Mô hình CRFs (Conditional Random Fields - CRFs)
- Sử dụng công cụ cho IE

Giới thiệu về trích rút thông tin

Trích chọn thông tin

- **Information Extraction**
- Trích chọn thông tin là bài toán tự động trích chọn các thông tin có cấu trúc từ các văn bản không cấu trúc.
- Tổ chức lại thông tin một cách có hệ thống, các thông tin trích chọn được có thể đưa vào database làm đầu vào cho các thuật toán khác (data mining).

Firm XYZ is a full service advertising agency specializing in direct and interactive marketing. Located in Bigtown CA, Firm XYZ is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automative account. Experience in online marketing, automative and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables . . . Compensation: $50,000-$80,000 Hiring Organization: Firm XYZ

INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Bigtown, CA.
COMPANY	Firm XYZ
SALARY	$50,000-$80,000

Trích chọn thông tin

Cập nhật dữ liệu vào CSDL thông qua trích chọn thông tin từ các đoạn văn bản

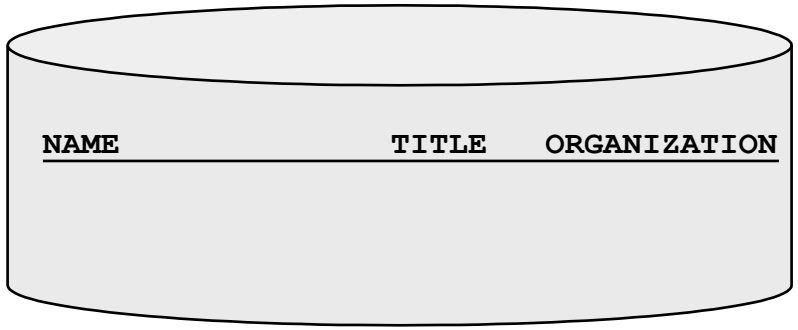
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



Trích chọn thông tin

Cập nhật dữ liệu vào CSDL thông qua trích chọn thông tin từ các đoạn văn bản

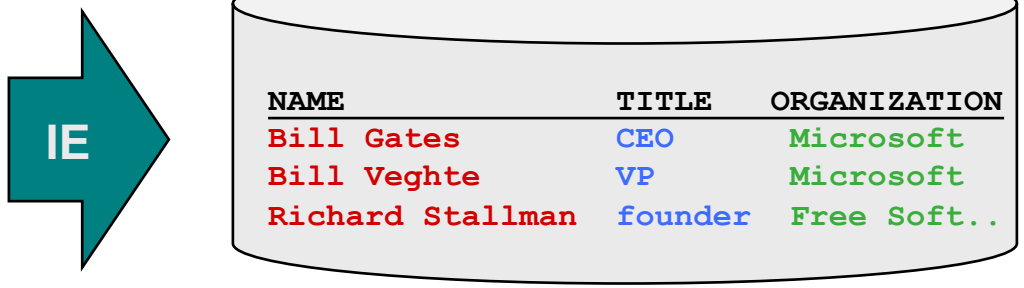
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage imp rovement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



Độ phức tạp của bài toán

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office is 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Trích xuất các mối quan hệ

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

"Named entity" extraction

Các bài toán con

- Named Entity Recognition (NER)
- Coreference Resolution
- Entity Linking
- Relation Extraction
- Event Extraction

Trích rút quan hệ: Dữ liệu dịch bệnh

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire, is finding itself hard pressed to cope with the crisis…

Information Extraction System

Date	*Disease Name*	*Location*
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Trích rút quan hệ: tương tác proteins

"We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."

CBF-A $\xleftarrow[\text{complex}]{\text{interact}}$ CBF-C

CBF-B $\xrightarrow[\text{associates}]{}$ CBF-A-CBF-C complex

Phân giải đồng tham chiếu

John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tue
29, 1917, at 3:00 pm,[7] the second son of Joseph P. Kennedy, Sr., and Rose Fitzgerald; R
turn, was the eldest child of John "Honey Fitz" Fitzgerald, a prominent Boston political fi
was the city's mayor and a three-term member of Congress. Kennedy lived in Brookline
years and attended Edward Devotion School, Noble and Greenough Lower School, and the Dexter
School, through 4th grade. In 1927, the family moved to 5040 Independence Avenue in
Bronx, New York City; two years later, they moved to 294 Pondfield Road in Bronxville, N
where Kennedy was a member of Scout Troop 2 (and was the first Boy Scout to become
President).[8] Kennedy spent summers with his family at their home in Hyannisport,
Massachusetts, and Christmas and Easter holidays with his family at their winter home i
Beach, Florida. For the 5th through 7th grade, Kennedy attended Riverdale Country School, a
private school for boys. For 8th grade in September 1930, the 13-year old Kennedy attended
Canterbury School in New Milford, Connecticut.

Gán nhãn chuỗi

Bài toán gán nhãn chuỗi

- Sequence labeling
- Nhiều bài toán NLP có thể đưa về bài toán gán nhãn chuỗi
- **Đầu vào:** một chuỗi các từ
- **Đầu ra:** chuỗi các từ đã được gán nhãn

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

POS tagging

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

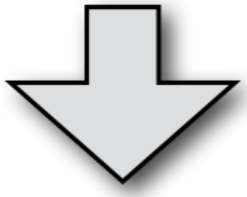
Named entity recognition

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Word segmentation

Gán nhãn chuỗi

	She	promised	to	back	the	bill
$\mathbf{w} =$	$w^{(1)}$	$w^{(2)}$	$w^{(3)}$	$w^{(4)}$	$w^{(5)}$	$w^{(6)}$



$\mathbf{t} =$	$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	$t^{(4)}$	$t^{(5)}$	$t^{(6)}$
	PRP	**VBD**	**TO**	**VB**	**DT**	**NN**

- Cho một chuỗi các từ $\mathbf{w}=w^{(1)}\ldots w^{(n)}$, tìm chuỗi các nhãn (tag) có khả năng xảy ra cao nhất $\mathbf{t}=t^{(1)}\ldots t^{(n)}$

$$\mathbf{t}^* = \text{argmax}_t P(\mathbf{t} \mid \mathbf{w})$$

Gán nhãn từ loại

- Part of Speech tagging – POS tagging
- Mỗi từ trong câu được gán nhãn thẻ từ loại tương ứng của nó
- **Đầu vào:** 1 đoạn văn bản đã tách từ + tập nhãn
- **Đầu ra:** cách gán nhãn chính xác nhất

Gán nhãn từ loại

- Các ứng dụng:
  - Tổng hợp tiếng nói: record - N: [‘reko:d], V: [ri’ko:d];
  - Tiền xử lý cho phân tích cú pháp.
  - Nhận dạng tiếng nói, tìm kiếm, dịch máy, v.v…

Tập nhãn cho tiếng Anh

- Tập ngữ liệu Brown: 87 nhãn
- 3 tập thường được sử dụng:
  - Nhỏ: 45 nhãn - Penn treebank
  - Trung bình: 61 nhãn, British national corpus
  - Lớn: 146 nhãn, C7

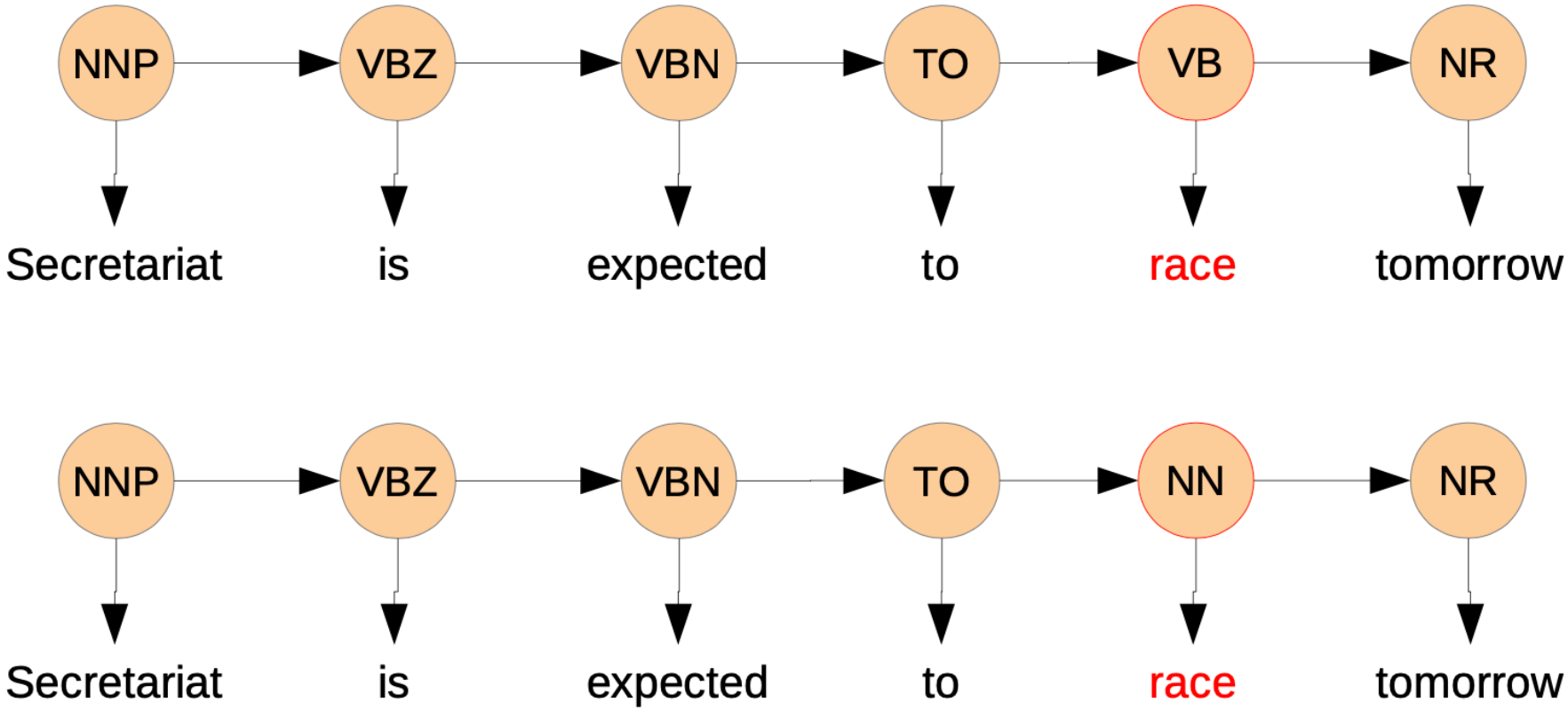
Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	' or "
POS	Possessive ending	*'s*	"	Right quote	' or "
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	[, (, {, <
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	],), }, >
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	. ! ?
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	: ; ... – -
RP	Particle	*up, off*			

Ví dụ

- **There/EX** are/VBP 70/CD children/NNS **there/RB**
- EX: từ chỉ sự tồn tại there
- RB: phó từ
- →Khó khăn trong gán nhãn từ loại: nhập nhằng.

Ví dụ

Secretariat[NNP] is[VBZ] expected[VBN] to[TO] race[?] tomorrow[NR] .



Gán nhãn từ loại tiếng Việt

Câu tiếng Việt đã tách từ	Qua những lần từ Sài_Gòn về Quảng_Ngãi kiểm_tra công_việc , Sophie và Jane thường trò_chuyện với Mai , cảm_nhận ngọn_lửa_sống và niềm_tin mãnh_liệt từ người phụ_nữ VN này .
Câu tiếng Việt đã được gán nhãn từ loại	Qua những lần từ Sài_Gòn về Quảng_Ngãi kiểm_tra công_việc , Sophie và Jane thường trò_chuyện với Mai , cảm_nhận ngọn_lửa_sống và niềm_tin mãnh_liệt từ người phụ_nữ VN này .
Chú thích từ loại	DANH TỪ ĐỘNG TỪ TÍNH TỪ ĐẠI TỪ ĐỊNH TỪ SỐ TỪ PHỤ TỪ GIỚI TỪ CẢM TỪ LIÊN TỪ THÁN TỪ TRỢ TỪ TỪ ĐƠN LẺ TỪ VIẾT TẮT KHÔNG XÁC ĐỊNH

Nhận dạng thực thể

- Named-entity recognition (NER)
- Bài toán con quan trọng của trích rút thông tin
- Thực thể (entity) là đối tượng hoặc tập hợp các đối tượng trong thế giới tự nhiên được mô tả bằng ngôn ngữ
- Phân loại:
  - Tên người
  - Tên địa điểm
  - Tên tổ chức
  - Giá trị số
  - Thời gian

Nhận dạng thực thể

- Nhận dạng trong văn bản các nhóm thực thể có tên đã được định trước như tên người, tổ chức, địa điểm, thời gian, …
- Các nhãn (tag)
  - PERS
  - ORG
  - LOC
  - DATE

Nhận dạng thực thể

Pierre Vinken , 61 years old , will join IBM ‘s board as a nonexecutive director Nov. 29 .

[PERS Pierre Vinken] , 61 years old , will join [ORG IBM] ‘s board as a nonexecutive director [DATE Nov. 2] .

Nhãn BIO

- Định nghĩa các nhãn (tag) mới:
  - **B-PERS, B-DATE, …:** Đánh dấu bắt đầu thực thể có tên (Begin)
  - **I-PERS, I-DATE, …:** Đánh dấu các từ tiếp theo của thực thể có tên (Inside)
  - **O:** Đánh dấu các từ không tuộc thực thể có tên (Outside)

Nhãn BIO

```
[PERS Pierre Vinken] , 61 years old , will join
[ORG IBM] ‘s board as a nonexecutive director
[DATE Nov. 2] .
```

```
Pierre_B-PERS Vinken_I-PERS ,_O 61_O years_O old_O ,_O
will_O join_O IBM_B-ORG ‘s_O board_O as_O a_O
nonexecutive_O director_O Nov._B-DATE 29_I-DATE ._O
```

VLSP 2016

	POS tag	Chunking tag	NE	Nested NE
Anh	N	B-NP	O	O
Thanh	Np	I-NP	I-PER	O
là	V	B-VP	O	O
cán_bộ	N	B-NP	O	O
Uỷ ban	N	B-NP	B-ORG	O
nhân_dân	N	I-NP	I-ORG	O
Thành_phố	N	I-NP	I-ORG	B-LOC
Hà_Nội	Np	I-NP	I-ORG	I-LOC
.	.	O	O	O

Các phương pháp

- Các phương pháp rule-based
  - Email, Thời gian, Số điện thoại, URL, Số lượng tiền
- Học máy thống kê
  - Mô hình Markov ẩn (Hidden Markov Model - HMM)
  - Maximum Entropy Markov Model (MEMM)
  - Conditional Random Field (CRF)
- Học sâu
  - RNN/LSTM
  - BERT

Các thư viện

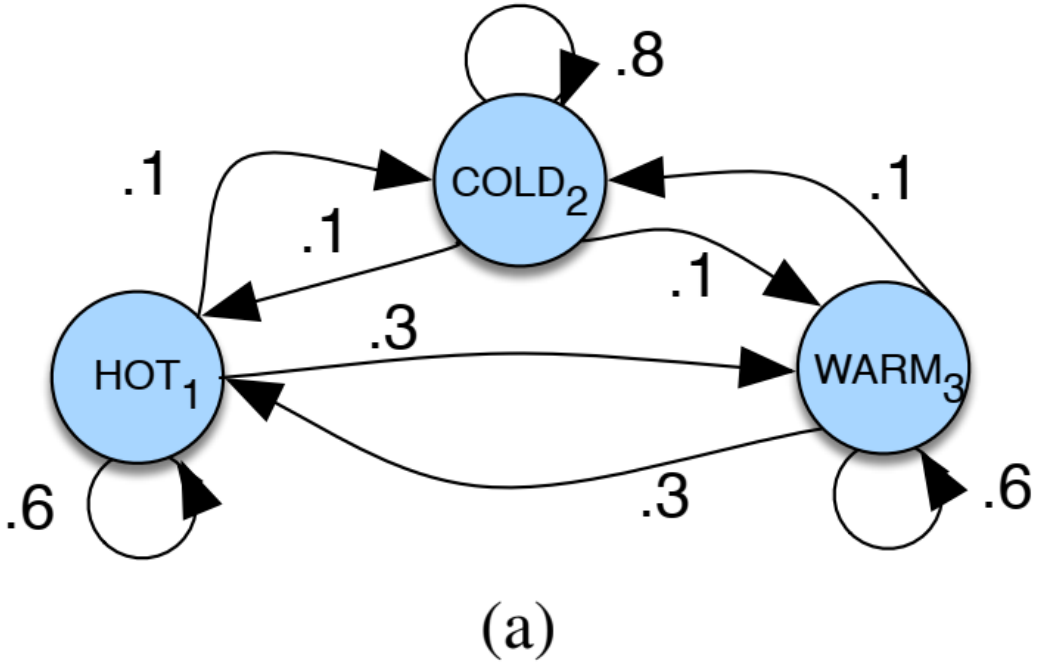
- NLTK
- Spacy
- Standford Core NLP
- Allen NLP
- Flair

Mô hình Markov ẩn

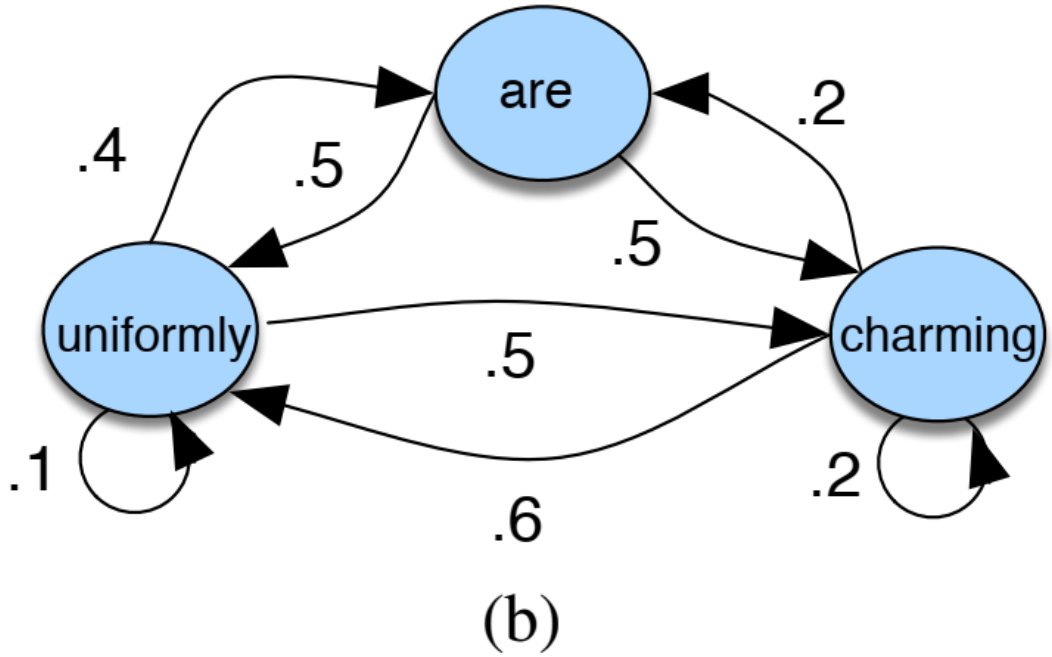
Các mô hình Markov

- Là một trong những mô hình học máy quan trọng
- Các mô hình Markov cơ bản
  - Mô hình chuỗi Markov
  - Mô hình Markov ẩn
- Mô hình chuỗi Markov (Markov chain), hay còn được gọi là mô hình Markov có thể quan sát được, là mô hình Markov đơn giản nhất
- Mô hình chuỗi Markov và Markov ẩn đều được mở rộng từ Automat hữu hạn
  - Một automat hữu hạn có trọng số có các cạnh được gắn với các xác suất, biểu diễn xác suất đi vào cạnh đó. Tổng tất cả các xác suất của các cạnh đi ra từ một đỉnh phải bằng 1.
  - Chuỗi Markov là trường hợp đặc biệt của automat hữu hạn có trọng số, khi đó chuỗi đầu vào sẽ quyết định các trạng thái mà automat sẽ đi qua.

Chuỗi Markov



(a)



(b)

Phân bố ban đầu $\pi = [0.1, 0.7, 0.2]$

Chuỗi Markov

- Các thành phần:

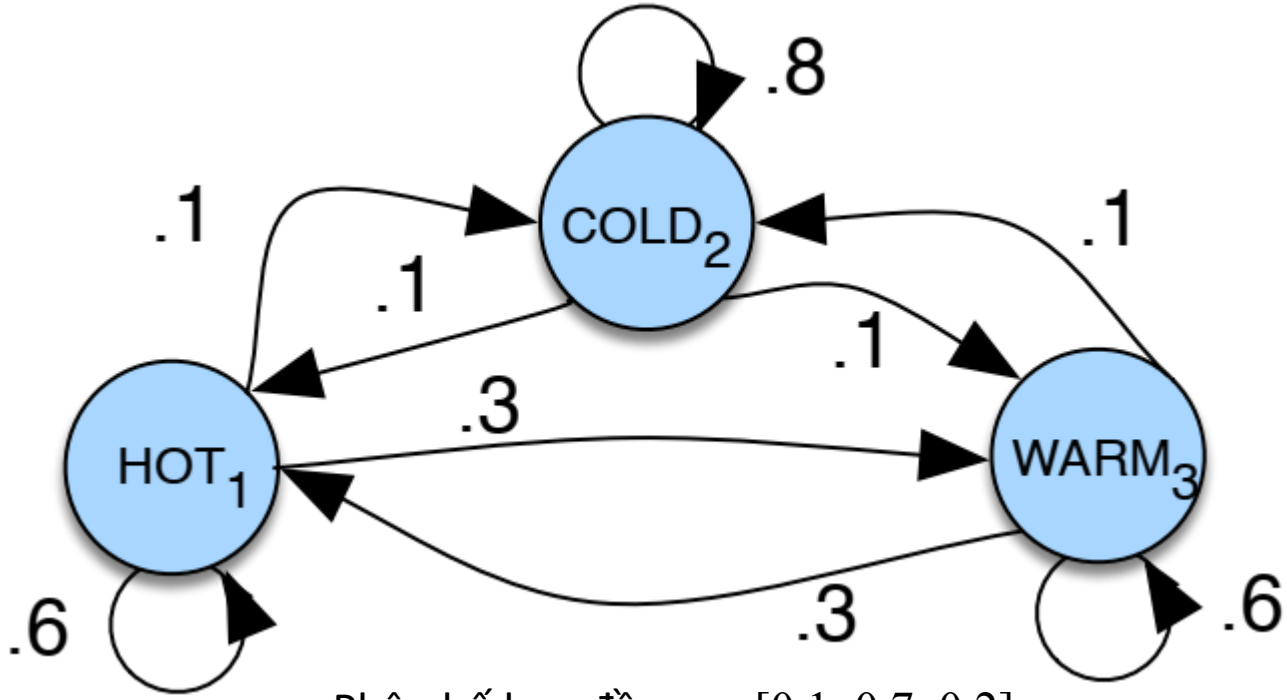
$Q = q_1 q_2 \ldots q_N$	tập hợp N trạng thái
$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$	ma trận xác suất chuyển trạng thái A, a_{ij} biểu diễn xác suất chuyển từ trạng thái i sang trạng thái j $\sum_{j=1}^{N} a_{ij} = 1 \; \forall i$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	phân bố xác suất ban đầu của các trạng thái $\sum_{i=1}^{N} \pi_i = 1$

- Giả định Markov: xác suất của một trạng thái chỉ phụ thuộc vào trạng thái trước nó

$$P(q_i = a | q_1 \ldots q_{i-1}) = P(q_i = a | q_{i-1})$$

Ví dụ

- Hãy tính xác suất của các chuỗi sau:
  - hot hot hot hot
  - cold hot cold hot



Phân bố ban đầu $\pi = [0.1, 0.7, 0.2]$

Mô hình Markov ẩn

- Mô hình chuỗi Markov dùng để tính xác suất của một chuỗi sự kiện mà chúng ta có thể quan sát được
- Tuy nhiên, trong nhiều trường hợp thì có những sự kiện chúng ta quan tâm có thể không quan sát trực tiếp được
- Mô hình Markov ẩn cho phép chúng ta xem xét cả các sự kiện quan sát được và các sự kiện ẩn.

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

POS tagging

Các thành phần

$Q = q_1 q_2 \ldots q_N$	tập hợp N trạng thái
$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$	ma trận xác suất chuyển trạng thái A, a_{ij} biểu diễn xác suất chuyển từ trạng thái i sang trạng thái j $\sum_{j=1}^{N} a_{ij} = 1 \; \forall i$
$O = o_1 o_2 \ldots o_T$	chuỗi sự kiện quan sát được
$B = b_i(o_t)$	emission probabilities: xác suất sự kiện o_t được sinh ra từ trạng thái q_i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	phân bố xác suất ban đầu của các trạng thái $\sum_{i=1}^{N} \pi_i = 1$

Xác suất chuyển trạng thái

- Xác suất chuyển trạng thái được tính bằng cách đếm số lần xuất hiện của các nhãn trên kho ngữ liệu

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

- Ví dụ, MD xuất hiện 13124 lần, trong đó có 10471 lần VB xuất hiện ngay sau nó

$$P(VB|MD) = \frac{C(MD, VB)}{C(MD)} = \frac{10471}{13124} = .80$$

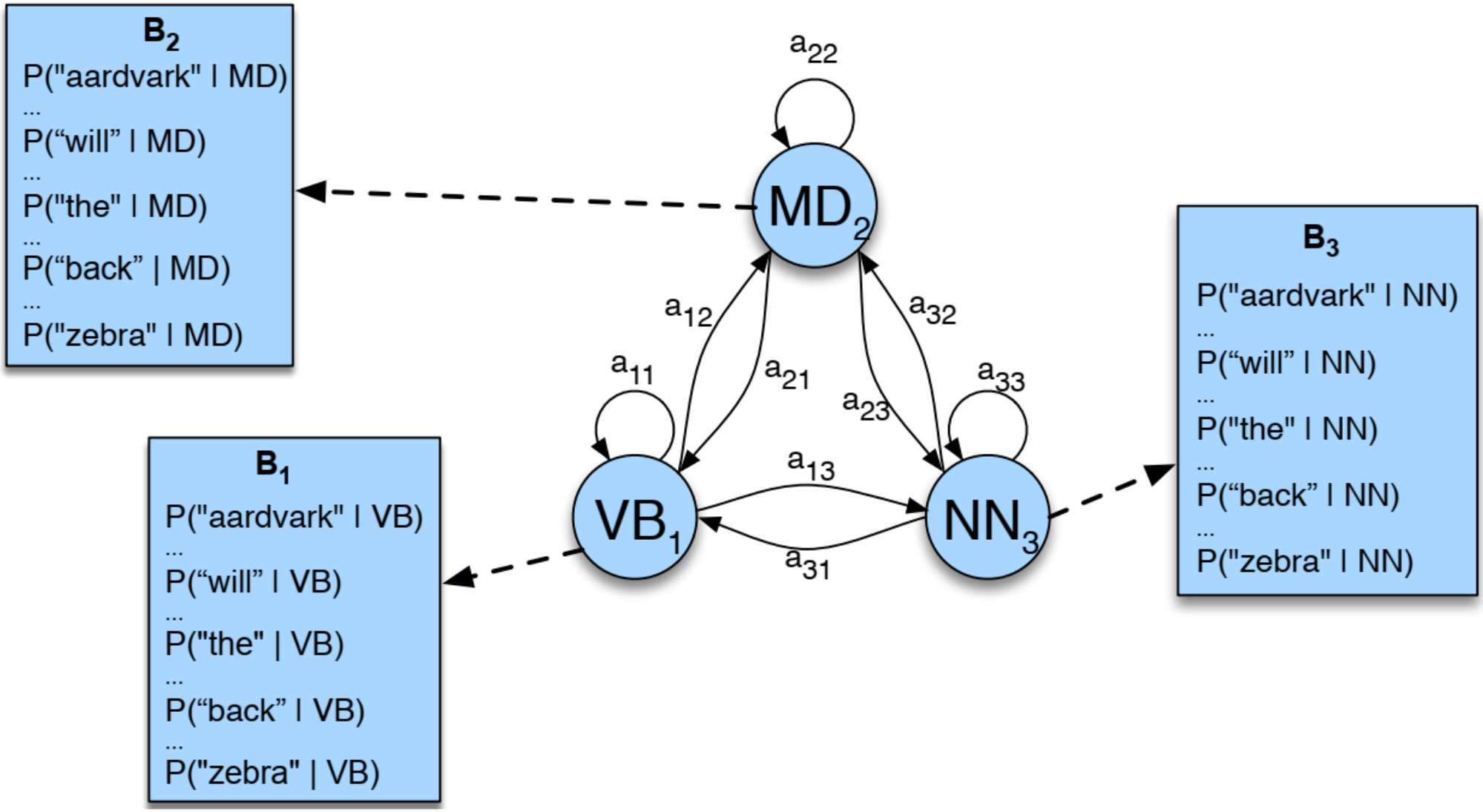
Xác suất chuyển

- Xác suất sự kiện o_t được sinh ra từ trạng thái q_i
- Xác suất một nhãn đi với một từ

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

$$P(will|MD) = \frac{C(MD, will)}{C(MD)} = \frac{4046}{13124} = .31$$

Mô hình Markov ẩn



HMM cho gán nhãn chuỗi

- Xác định chuỗi trạng thái ẩn tương ứng với chuỗi quan sát được, được gọi là decoding
- Cho đầu vào là một HMM $\lambda = (A, B)$ và một chuỗi quan sát được $O = o_1 o_2 \ldots o_T$, tìm chuỗi các trạng thái $Q = q_1 q_2 \ldots q_N$

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n | w_1^n)$$

- Định lý Bayes

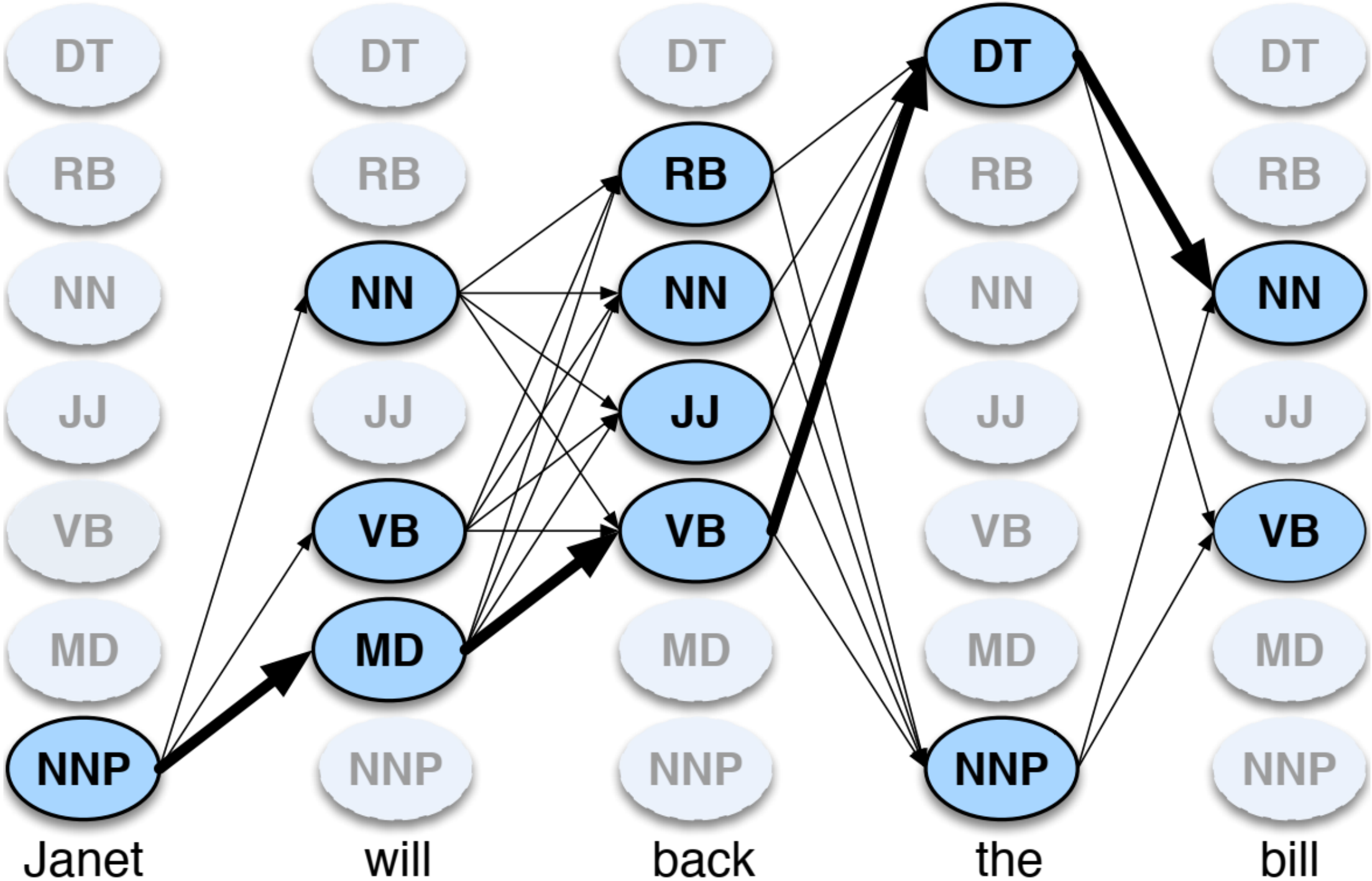
$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)}$$

HMM cho gán nhãn chuỗi

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(w_1^n | t_1^n) P(t_1^n)$$

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n | w_1^n) \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i | t_i)}^{\text{emission}} \overbrace{P(t_i | t_{i-1})}^{\text{transition}}$$

- Giải quyết bằng thuật toán Viterbi (quy hoạch động)


DT
RB
NN
JJ
VB
MD
NNP
Janet
will
back
the
bill

Mô hình CRFs

Conditional Random Fields

- Lafferty et al. 2001
- Được áp dụng rộng rãi trong nhiều lĩnh vực từ XLNNTN đến thị giác máy, phân tích chuỗi trong sinh học
- CRF là phương pháp thống kê nhưng thường vẫn được sử dụng kết hợp với các mô hình học sâu

https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers

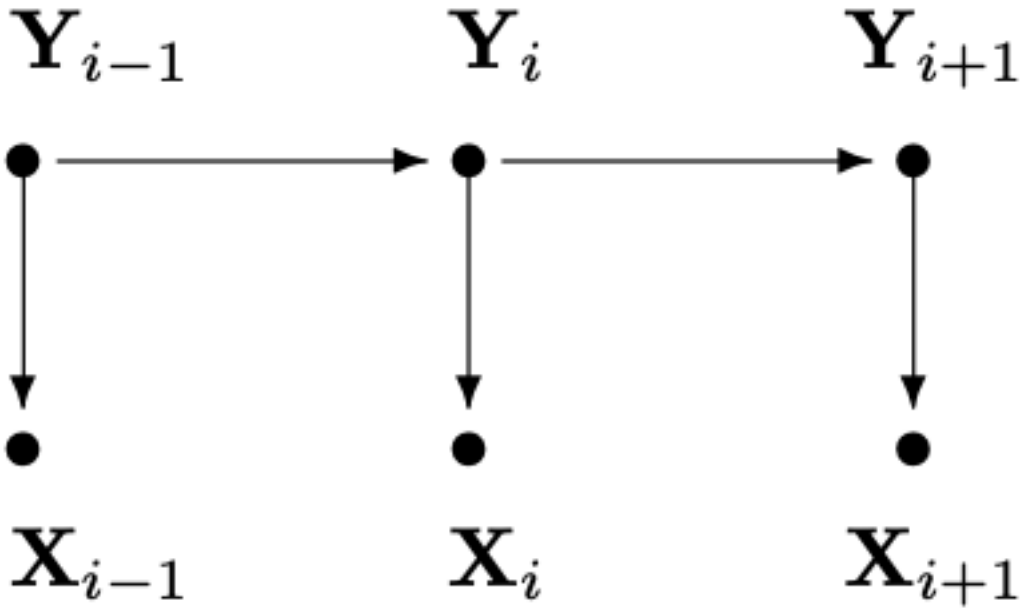
Generative và discriminative

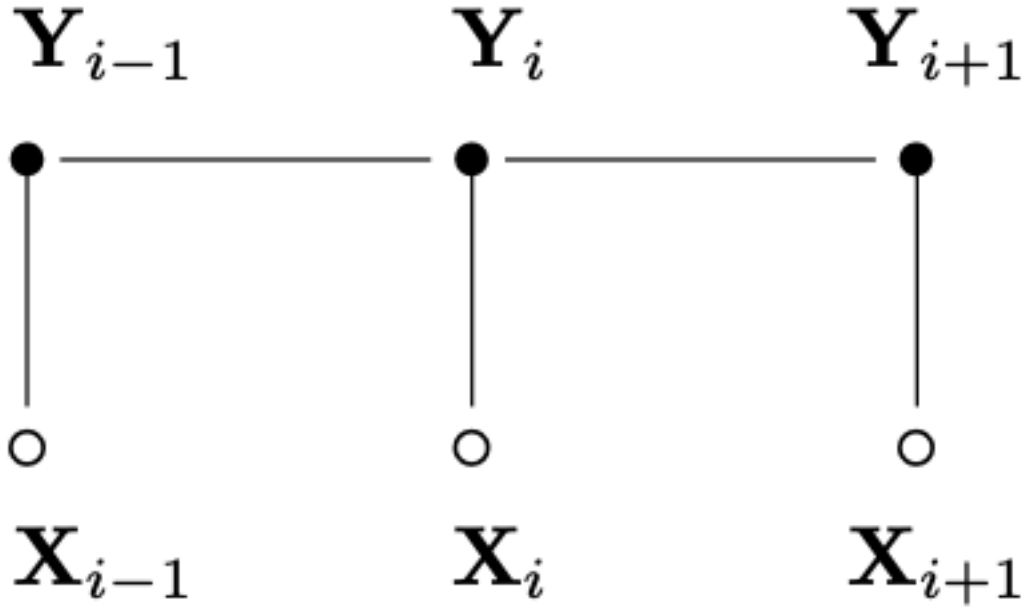
- Đều cố gắng mô hình hóa phân bố xác suất trên (y, x)
- HMM: mô hình generative của chuỗi đầu vào x, **mô tả phân bố "sinh" ra x khi đã biết nhãn y** (áp dụng định lý Bayes)

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|x) = \underset{y}{\operatorname{argmax}}\, P(x|y)P(y)$$

- Các mô hình discriminative (CRF) **trực tiếp** mô hình hóa P(y|x) bằng các hàm đặc trưng

HMMs và CRFs





- Generative
- Trạng thái quan sát X được sinh ra từ mô hình

- Discriminative
- Trạng thái quan sát X không được sinh ra từ mô hình

Conditional Random Fields

- Phân bố P(y|x) trong CRF được định nghĩa như sau

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

- Vector tham số $\theta = \{\theta_k\} \in \Re^K$
- Hàm chuẩn hóa

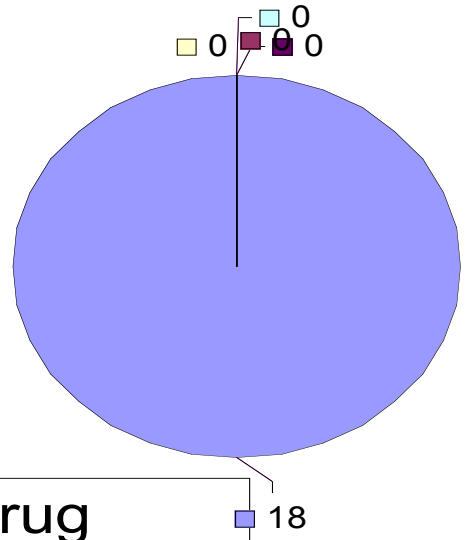
$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

Các đặc trưng

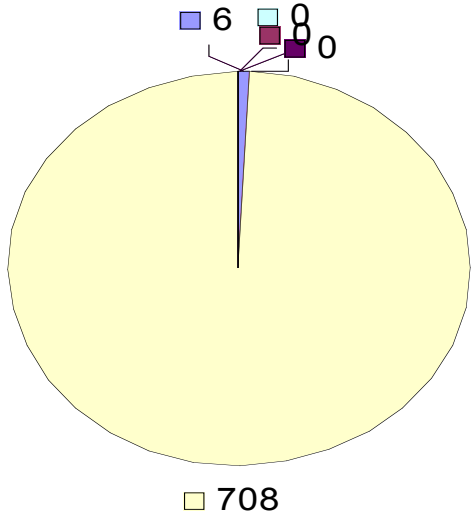
- Các đặc trưng của từ hiện tại, các từ trước và sau nó, nhãn trước nó
  - Chứa tiền/hậu tố đặc thù
  - Chứa số, chữ viết hoa, gạch ngang
  - Viết hoa toàn bộ
  - Word shape
  - Nhãn từ loại
- Sử dụng đặc trưng nào phụ thuộc từng bài toán và bộ dữ liệu huấn luyện

Các đặc trưng

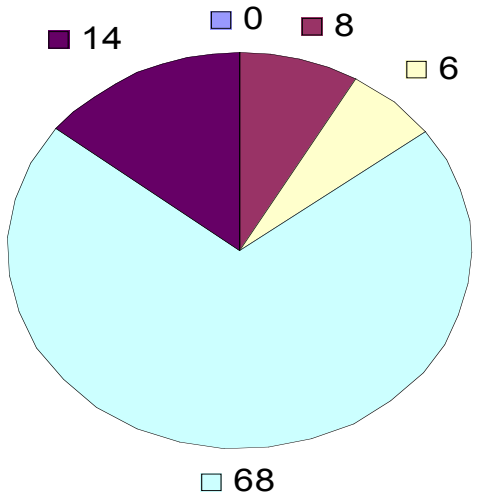
oxa



:



field



Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

Word shape

- Biểu diễn từ một cách đơn giản bằng việc mã hóa các ký tự viết thường thành ‘x’, viết hoa thành ‘X’, số thành ‘d’

I.M.F	X.X.X	X.X.X
DC10-30	XXdd-dd	Xd-d
well-dressed	xxxx-xxxxxxx	x-x

Huấn luyện mô hình

- Xác định các tham số của mô hình $\theta = \{\theta_k\} \in \Re^K$
- Maximum likelihood: dữ liệu huấn luyện đạt xác xuất lớn nhất với các tham số được chọn (tương tự logistic regression)

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_t^{(i)}, y_{t-1}^{(i)}, \mathbf{x}_t^{(i)}) - \sum_{i=1}^{N} \log Z(\mathbf{x}^{(i)}) - \sum_{k=1}^{K} \frac{\theta_k^2}{2\sigma^2}$$

- Có thể áp dụng các thuật toán tối ưu như gradient descent

Dự đoán

- Sau khi đã huấn luyện mô hình, với mỗi đầu vào x cần dự đoán nhãn tương ứng sao cho

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|x)$$

- Có thể áp dụng thuật toán Viterbi để tìm chuỗi trạng thái (chuỗi nhãn) để xác suất P(y|x) đạt giá trị lớn nhất.

Thực nhiệm so sánh CRF và HMM

- Lafferty et al. 2001
- Penn treebank POS tagging (45 tags)
- Sử dụng các đặc trưng chính tả:
  - có bắt đầu bằng số hoặc chữ viết hoa không,
  - chứa dấu gạch ngang không,
  - có chứa các hậu tố sau không: -ing, -ogy, -ed, -s, -ly, -ion, -tion, -ity, -ies
- oov = out-of-vocabulary (not observed in the training set)

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM$^+$	4.81%	26.99%
CRF$^+$	4.27%	23.76%

$^+$Using spelling features

Sử dụng công cụ cho trích rút thông tin

Bài thực hành

- Sử dụng thư viện sklearn_crfsuite để huấn luyện mô hình CRF cho bài toán NER
- Sử dụng mô hình CRF để trích chọn thông tin từ văn bản

Q&A

Thank you!