Information Extraction

Dr. Cao Truong Tran,
Computer Science Department, Faculty of Information Technology, Le Quy Don Technical University, Hanoi, Vietnam
Email: truongct@lqdtu.edu.vn

Contents

- Information Extraction
- Sequence labeling
  - POS tagging
  - Named Entity Recognition
- Hidden Markov Models
- Conditional Random Fields (CRFs)
- Using CRFs tools for IE

Information Extraction

Information Extraction

- Information extraction is the problem of automatically extracting structured information from unstructured documents.
- Organize the information systematically, the extracted information can be put into the database as input for other algorithms (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

Information Extraction

Updating data into the database by extracting information from text fragments

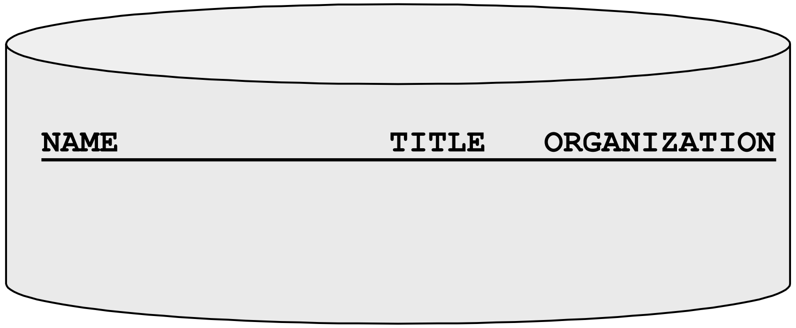
October 14, 2002, 4:00 a.m. PT

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…



Information Extraction

Updating data into the database by extracting information from text fragments

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…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

The complexity of the problem

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.

Relation Extraction

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

Sub Problems

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

Relation Extraction: Disease Outbreaks

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.

Relation Extraction: Protein Interactions

"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."

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

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

Resolving coreference (both within and across documents)

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; Ro
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 t
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 R
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.

Sequence labeling

Sequence labeling

- Many NLP problems can be reduced to sequence labeling
- Input: a string of words
- Output: string of labeled words

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

POS tagging

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

Word segmentation

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

Named entity recognition

Sequence labeling

	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**

- Given a sequence of words w=w(1)…w(n), find the sequence of tags with the highest probability t=t(1)…t(n)

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

Part of Speech tagging

- Part of Speech tagging – POS tagging
- Each word in the sentence is labeled with its corresponding word tag
- Input: 1 word delimited text + label set
- Output: the most accurate way of labeling

Part of Speech tagging

- Applications:
  - Speech synthesis: record - N: [‘reko:d], V: [ri’ko:d];
  - Preprocessor for parsing.
  - Speech recognition, search, etc.

Part of Speech tagging

- Brown corpus: 87 labels
- 3 commonly used sets:
  - Small: 45 longan - Penn treebank
  - Average: 61 labels, British national corpus
  - Large: 146 labels, 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*			

Example

- **There/EX** are/VBP 70/CD children/NNS **there/RB**
- EX: word for existence there
- RB: adverb
- Difficulty in labeling from categories: ambiguous.

Part of Speech tagging for Vietnamese

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 ■

Named-entity recognition

- Important subproblem of information extraction
- An entity is an object or collection of objects in the natural world described in language
- Classify:
  - Name
  - Place name
  - Organization Name
  - Numeric value
  - Time

Named-entity recognition

- Identify in text groups of entities with predefined names such as names of people, organizations, places, times, etc.
- tags
  - PERS
  - ORG
  - LOC
  - DATE

Named-entity recognition

```
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] .
```

BIO Labels

- Define new tags
  - B-PERS, B-DATE, …: Mark the start of the named entity (Begin)
  - I-PERS, I-DATE, …: Highlight the next words of the named entity (Inside)
  - O: Highlight words that do not have a named entity (Outside)

BIO Labels

[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

Approaches

- rule-based approaches
  - Email, Time, Phone Number, URL, Amount
- Statistical Machine Learning
  - Hidden Markov Model - HMM
  - Maximum Entropy Markov Model (MEMM)
  - Conditional Random Field (CRF)
- Deep learning
  - RNN/LSTM
  - BERT

Libraries

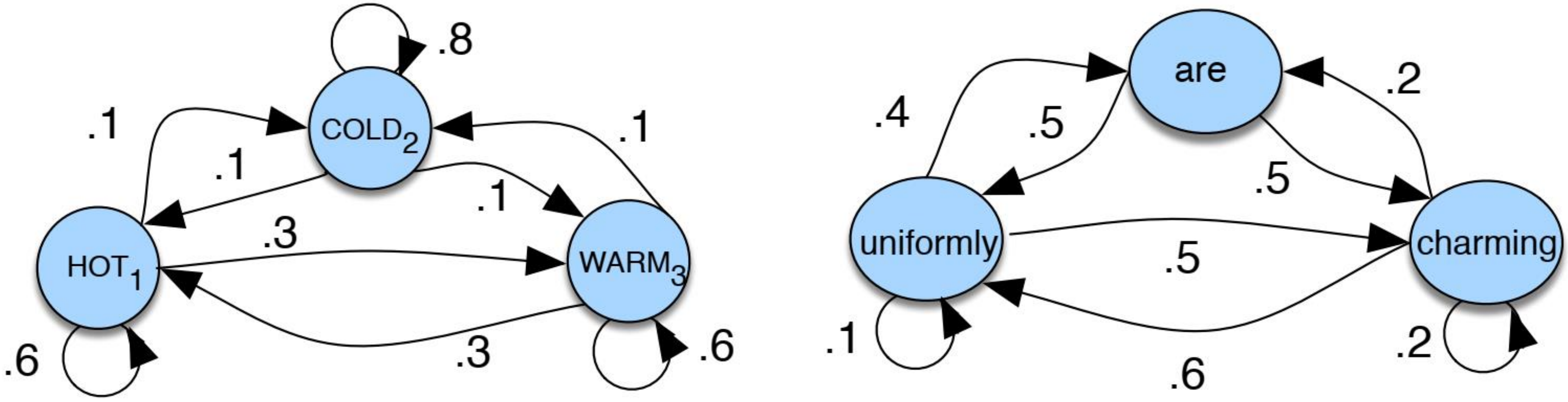
- NLTK
- Spacy
- Standford Core NLP
- Allen NLP
- Flair

Hidden Markov Models

Markov Models

- Hidden Markov model is one of the important machine learning models
- Basic Markov models: Markov chain model and hidden Markov model
- The Markov chain model, also known as the observable Markov model, is the simplest Markov model.
- Hidden Markov and Markov chain models are both extended from Finite Automat
- A weighted finite automat whose edges are attached to probabilities, representing the probability of entering that edge. The sum of all probabilities of the edges coming out of a vertex must be 1.
- Markov series is a special case of finite weighted automat, then the input sequence determines the states that the automat will go through.

Markov Chain



(a) (b)

Initial distribution π = [0.1, 0.7, 0.2]

Markov Chain

- Components:

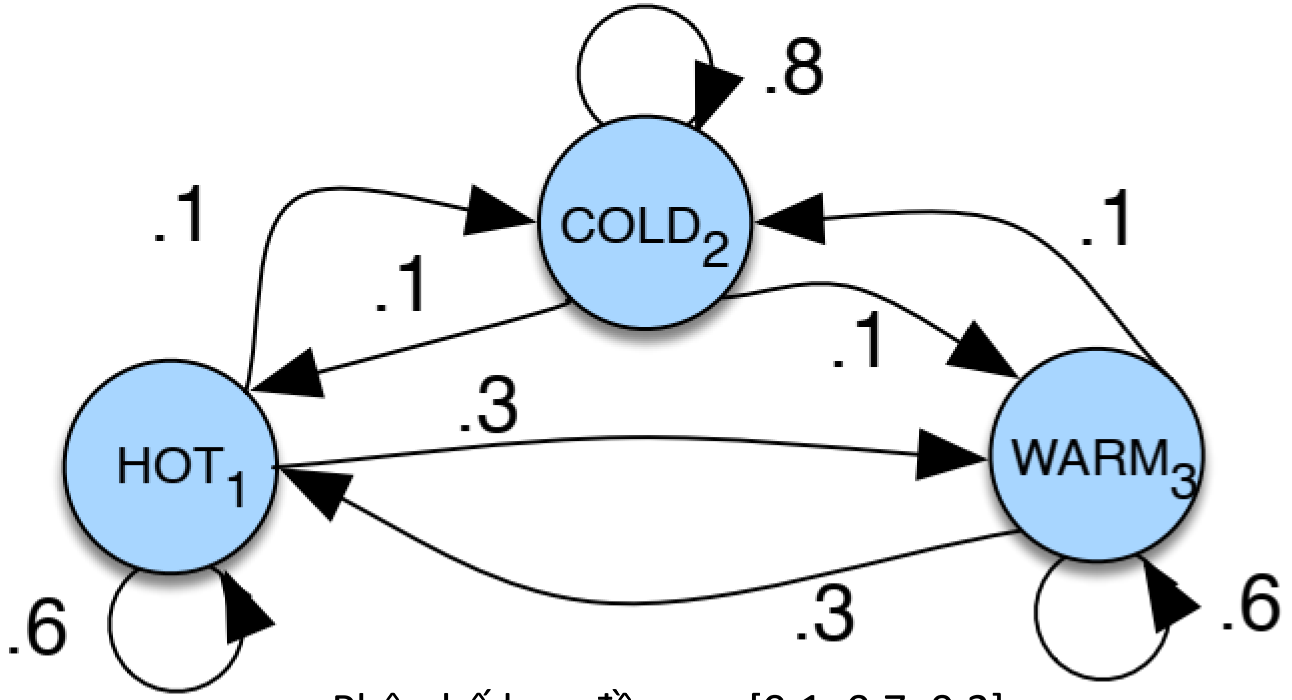
$Q = q_1 q_2 \ldots q_N$	set of N states
$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$	matrix transition probability state A, a_{ij} represents the probability of transition from state i to state j $\sum_{j=1}^{N} a_{ij} = 1 \ \forall i$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	initial probability distribution of states $\sum_{i=1}^{N} \pi_i = 1$

- Markov conjecture: the probability of a state depends only on the state before it

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

Example

- Calculate the probability of the following series:
  - hot hot hot hot
  - cold hot cold hot



Phân bố ban đầu π = [0.1, 0.7, 0.2]

+) P(H) * P(H | H) * P(H|H) * P(H|H)

= 0,1 . 0,6 . 0,6 . 0,6

= 0,0216

+) P(C) * P(H|C) * P(C|H) * P(H|C) = 0,7 . 0,1 . 0,1 . 0,1 = 0,0007

Hidden Markov Model

- The Markov series model is used to calculate the probability of an observable sequence of events
- However, in many cases there are events we are interested in that may not be directly observed
- The hidden Markov model allows us to consider both observable and hidden events.

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

POS tagging

Hidden Markov Model

- Components:

$Q = q_1 q_2 \ldots q_N$	set of N states
$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$	matrix transition probability state A, a_{ij} represents the probability of transition from state i to state j $\sum_{j=1}^{N} a_{ij} = 1 \ \forall i$
$O = o_1 o_2 \ldots o_T$	observable sequence of events
$B = b_i(o_t)$	emission probabilities: probability that event o_t is generated from state q_i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	initial probability distribution of states $\sum_{i=1}^{N} \pi_i = 1$

State transition probability

- The state transition probability is calculated by counting the number of occurrences of the labels in the corpus

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

- For example, MD appears 13124 times, of which 10471 times VB appears immediately after it

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

Emission probabilities

- Probability of event o_t generated from state q_i
- Probability that a label goes with a word

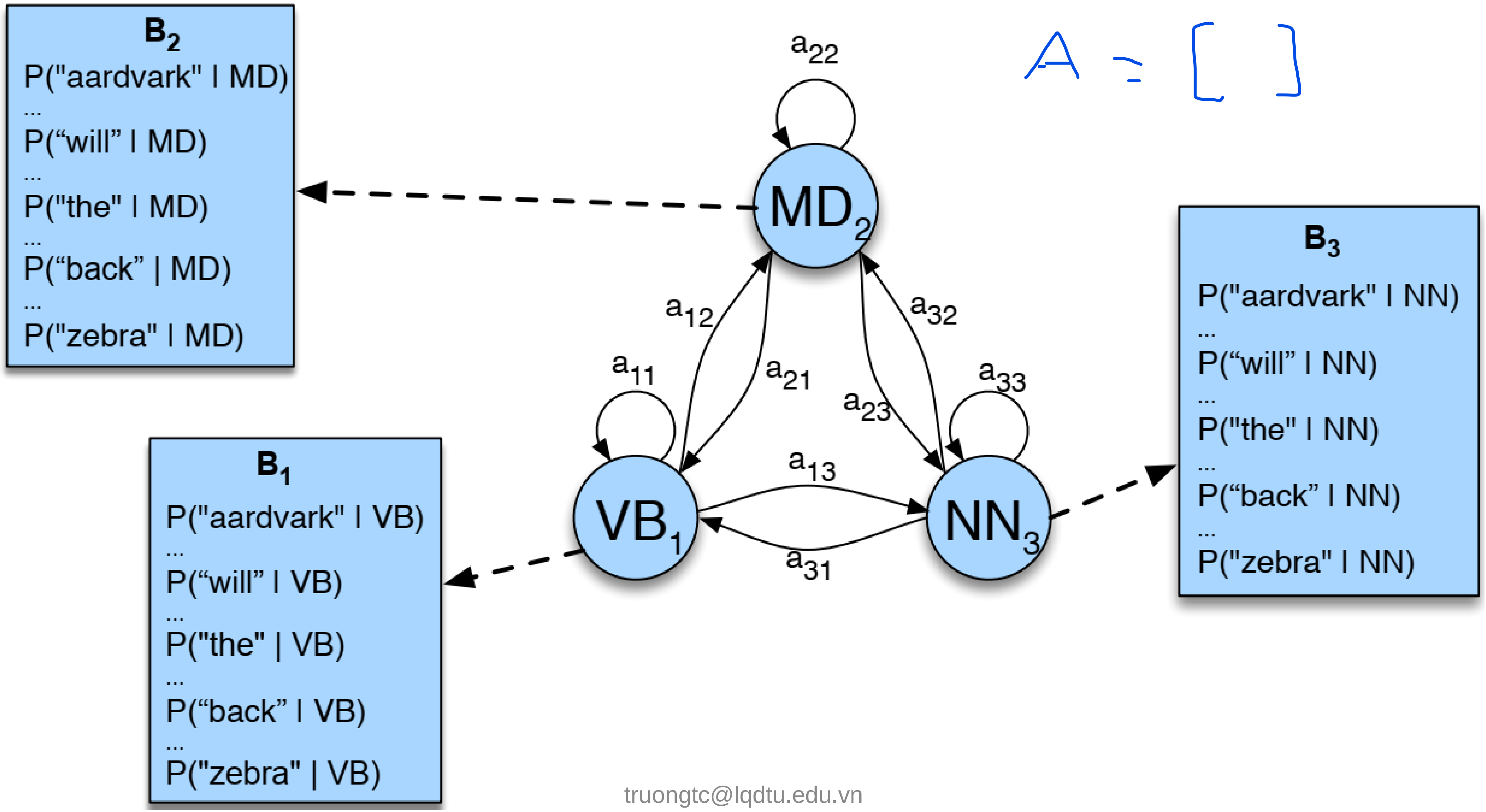
$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

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

Nếu là 1 modal verb, xác suất will là bn ?

Hidden Markov Model

$A = [\ ]$



HMM cho sequence labeling

- Determine the hidden state sequence corresponding to the observable sequence, called decoding
- Given as input an HMM $\lambda=(A,B)$ and an observable sequence O = $o_1o_2 \ldots o_T$, find the sequence of states Q = $q_1q_2 \ldots q_N$

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

- Bayes theory:

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

HMM cho sequence labeling

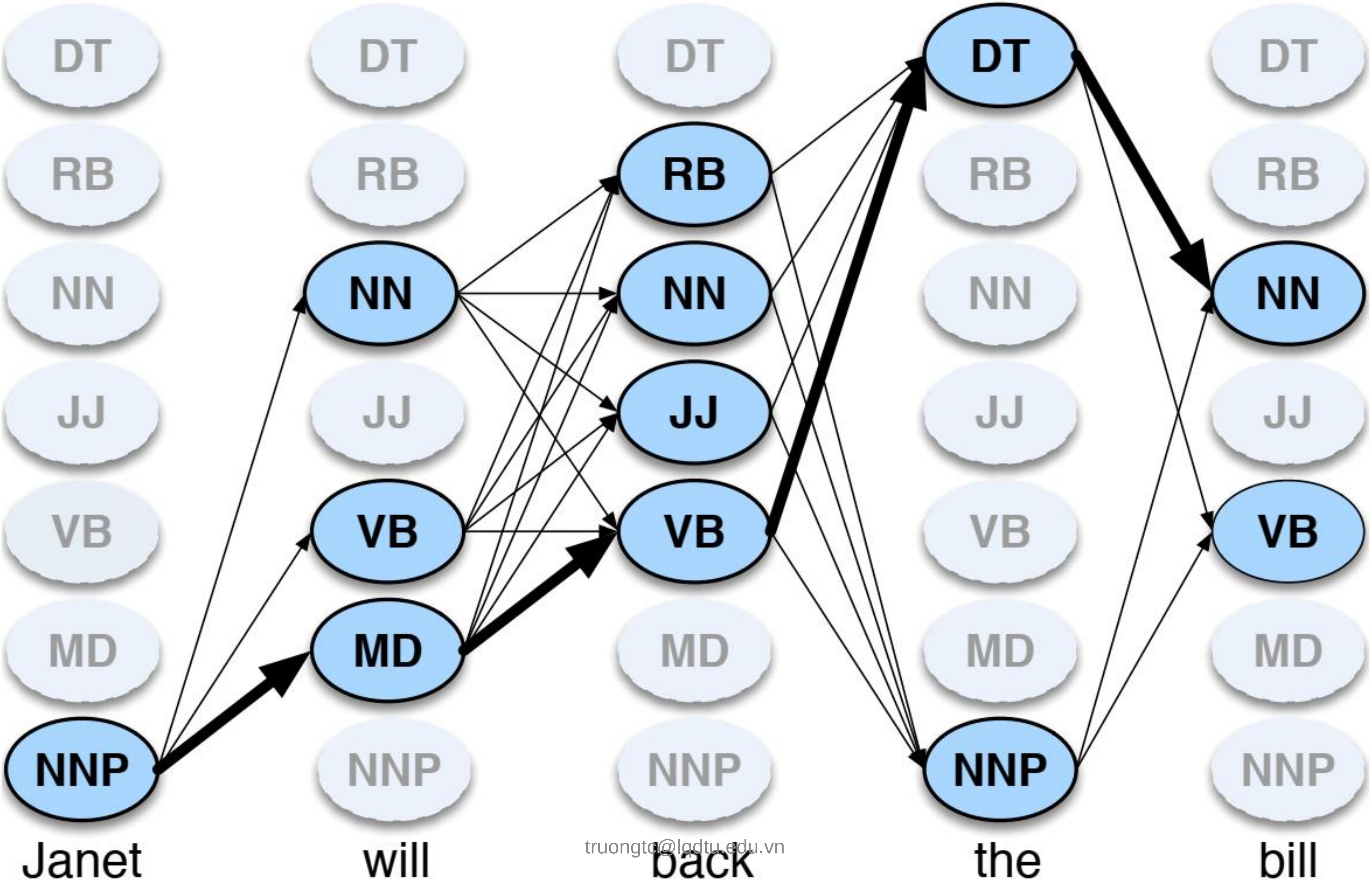
$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}}\, P(w_1^n|t_1^n)P(t_1^n)$$

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

- Solve with Viterbi algorithm (dynamic programming)

XS
xuất hiện của
1 từ chỉ phụ thuộc
vào thẻ của chính nó

phụ thuộc
vào thẻ
trước đó


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

Conditional Random Fields (CRFs)

Conditional Random Fields

- Lafferty et al. 2001
- Widely applied in many fields from NNTN to machine vision, sequence analysis in biology
- CRF is a statistical method but is often used in conjunction with deep learning models

Generative and discriminative

- All try to model the probability distribution on (y, x)
- HMM: generative model of the input sequence x, describing the distribution that "generates" x when the label y is known (using Bayes' theorem)

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

- Discriminative (CRF) models directly model P(y|x) using feature functions

Conditional Random Fields

- The distribution P(y|x) in the CRF is defined as follows

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

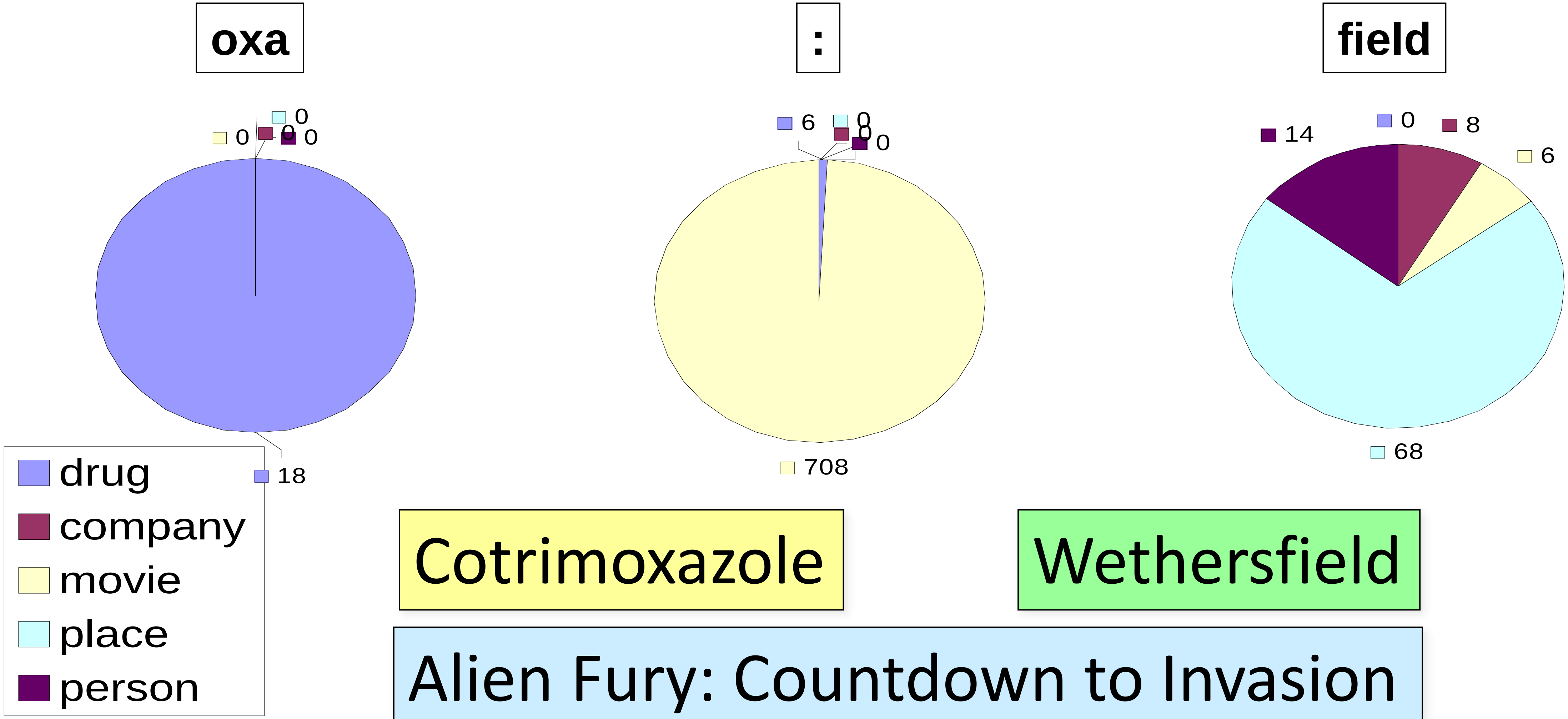
- Parameter vector $\theta = \{\theta_k\} \in \Re^K$
- Normalization function

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

Features

- Features of current word, words before and after it, label before it
  - Contains a specific prefix/suffix
  - Contains numbers, capital letters, dashes
  - All caps
  - Word shape
  - Label from category
- Which feature to use depends on the problem and the training data set

Features



Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

Word shape

- Simple word representation by encoding lowercase characters to 'x', uppercase to 'X', numbers to 'd'

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

Training models

- Determine the parameters of the model $\theta = \{\theta_k\} \in \Re^K$
- Maximum likelihood: the training data has the greatest probability for the selected parameters (similar to logistic regression)

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

- It is possible to apply optimization algorithms such as gradient descent

Predictions

- After training the model, for each input x it is necessary to predict the corresponding label such that

$$\hat{y} = \operatorname*{argmax}_{y} P(y|x)$$

- It is possible to apply the Viterbi algorithm to find the state series (label series) so that the probability P(y|x) reaches the maximum value.

Practice comparing CRF and HMM

- Lafferty et al. 2001
- Penn treebank POS tagging (45 tags)
- Use spelling features:
  - does it start with a number or a capital letter,
  - contains no dash,
  - contain the following suffixes: -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

Practice

- Using the sklearn_crfsuite library to train the CRF model for the NER . problem
- Using CRF model to extract information from text

Q&A

Thank you!