Natural Language Processing
自然語言處理

黃瀚萱
Department of Computer Science
National Chengchi University
2020 Fall

Sequence Labeling II:
Information Extraction and Chinese Word Segmentation

Schedule

Date	Topic
9/16	Introduction
9/23	Linguistic Essentials
9/30	Collocation
10/7	Language Model
10/14	Performance Evaluation and Word Sense Disambiguation
10/21	Text Classification (HW1 will be assigned)
10/28	Invited Talk: NLP and Cybersecurity (Term Project)
11/4	**POS Tagging**
11/11	Midterm Exam

Schedule

Date	Topic
11/18	**Chinese Word Segmentation**
11/25	Word Embeddings
12/2	Neural Networks for NLP
12/9	Parsing
12/16	Discourse Analysis
12/23	Invited Talk
12/30	Final Project Presentation I
1/6	Final Project Presentation II
1/13	Final Exam

Agenda

- Sequence labeling
  - Conditional random fields (CRFs)
  - Deep neural networks
- Sequence labeling tasks
  - Named entity recognition
  - Relation extraction
  - Chinese word segmentation

Maximum-Entropy Markov Model

Limitation of HMMs

- The assumption of HMM is very limited
  - y_t is determined by only $y_{t\text{-}1}$ and x_t.
- Many kind of information are missing
  - $x_{t\text{-}2}$, $x_{t\text{-}1}$, x_{t+1}, x_{t+2}
  - Position of t (the first word, the second word, the last word, etc)
  - Subword information of x_t such as its suffix and prefix

Maximum-Entropy Markov Model

- Maximum-entropy model is actually the multi-class logistic regression model.

- Maximum-entropy Markov model

  - Markov model with logistic regression classifier

Information of entire ***x*** can be used at any time *t*

$$P_{MEMM}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} \underline{P(y_t|y_{t-1}, \mathbf{x})}$$

Logistic regression classifier

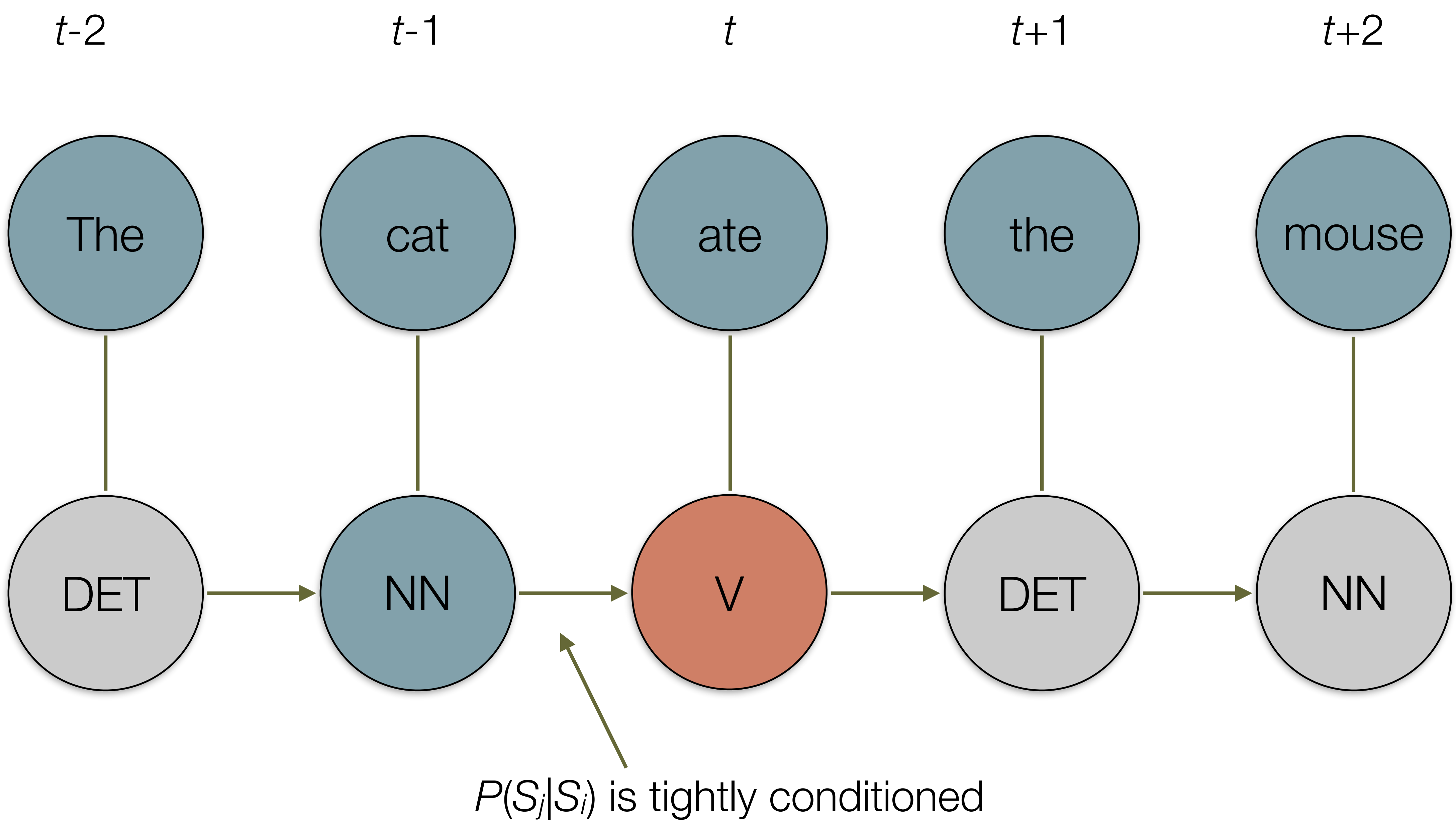
$$P(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z_t(y_{t-1}, \mathbf{x})} \exp\left[\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right]$$

Normalization term

Weight of the feature *k*

Occurrence of the feature *k*

Maximum-Entropy Markov Model



Future states cannot affect the posterior distribution over earlier states

Features of Markov Maximum Entropy Model

- Without the assumption of statistical independency
- Various (dependent) features for each position *t* can be considered.
  - $f_i(yt, y_{t\text{-}1}, x_t)$: 1 if x_t is in the uppercase and y_t is Proper Noun
  - $f_j(y_t, y_{t\text{-}1}, x_t)$: 1 if $x_{t\text{-}1}$ ends without 's', x_t ends with 's', $y_{t\text{-}1}$ is Noun, and y_t is Verb

$$P_{MEMM}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|y_{t-1}, \mathbf{x})$$

$$P(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z_t(y_{t-1}, \mathbf{x})} \exp\left[\sum_{k=1}^{K} \theta_k \underline{f_k(y_t, y_{t-1}, x_t)}\right]$$

Weight of each feature is obtain with MLE

Conditional Random Fields

- Most popular model for sequence labeling

Information of entire ***x*** can be used at any time *t*

$$P_{CRF}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} \underline{P(y_t|y_{t-1}, \mathbf{x})}$$

Logistic regression classifier

$$P(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left[\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right]$$

Normalization term sum over labels of an entire sequence

Weight of the feature *k*

Occurrence of the feature *k*

Key difference between MMEM and CRF

MEMM vs CRF

$$P_{MEMM}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|y_{t-1}, \mathbf{x})$$

$$P(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z_t(y_{t-1}, \mathbf{x})} \exp\left[\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right]$$

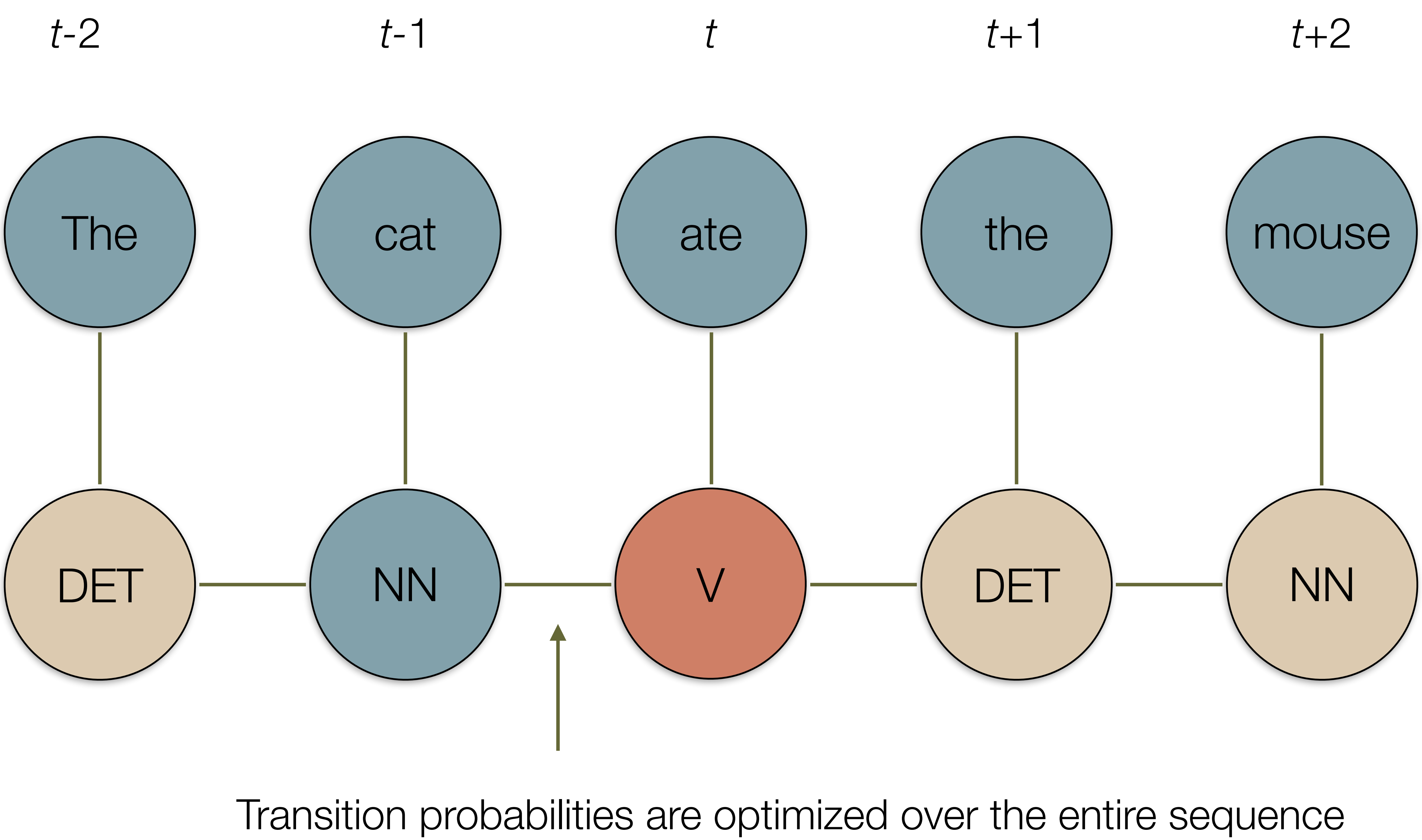
Normalization term of the sequence at ***t***

$$P_{CRF}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|y_{t-1}, \mathbf{x})$$

$$P(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left[\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right]$$

Normalization term sum over labels of an entire sequence ***x***

Conditional Random Fields



Decoding by the Viterbi Algorithm

$$P_{CRF}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|y_{t-1}, \mathbf{x})$$

$$P(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left[\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right]$$

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y}} P_{CRF}(\mathbf{y}|\mathbf{x})$$

max(P(IN | V, $\mathbf{x}$), P(IN | NN, $\mathbf{x}$))

P(DET | IN, $\mathbf{x}$)

	<s>	the	cat	runs	to	the	mouse
<s>	1	0	0	0	0	0	0
DET	0	1	0	0	0	0.098	0
IN	0	0	0	0	0.098	0	0
V	0	0	0	0.49	0	0	0
NN	0	0	1	0.03	0	0	0.098

Viterbi Algorithm for MEMM/CRF

$$\delta_t(i) = \max_{y_1, y_2, \ldots, y_{t-1}} P(y_1, y_2, ..., y_{t-1}, y_t = S_i, x_1, x_2, ..., x_t | \lambda)$$

1. Initialization:

$$\delta_1(i) = P(i | {<}\mathrm{s}{>}, \mathbf{x})$$

Base case for t = 1

$$\psi_1(i) = 0$$

No previous state

2. Recursion:

$$\delta_t(i) = \max_{j}^{M} \delta_{t-1}(j) P(i | j, \mathbf{x})$$

General case for t > 1

$$\psi_t(i) = \arg\max_{j}^{M} \delta_{t-1}(j) P(i | j, \mathbf{x})$$

Best i for $\delta_t(j)$ denoting the best previous state for $y_t = S_i$

3. Termination:

$$P(\mathbf{y}|\mathbf{x}) = \max_{i}^{M} \delta_T(i)$$

$$\bar{\mathbf{y}} = \arg\max_{i}^{M} \delta_T(i)$$

$$\bar{y}_t = \psi_{t+1}(\bar{y}_{t+1}) \quad \text{for } t = T-1, T-2, ..., 1$$

Feature Engineering for MEMM and CRFs

Sequence Labeling

x
My
dog
ate
my
cake
.

y
PRP
NN
VBD
PRP
NN
PU

Features for Sequence Labeling

x
My
dog
ate
my
cake
.

x_t	x_{t-1}	x_{t+1}	x_{t-1}, x_t	x_t, x_{t+1}	Capital Initial	Punct	Begin	End	Shape
my	<s>	dog	<s>, my	my dog	1	0	1	0	Xx
dog	my	ate	My dog	dog ate	0	0	0	0	xxx
ate	dog	my	dog ate	ate my	0	0	0	0	xxx
my	ate	cake	ate my	my cake	0	0	0	0	xx
cake	my	.	my cake	cake .	0	0	0	0	xxxx
.	cake	</s>	cake .	. </s>	0	1	0	1	.

y
PRP
NN
VBD
PRP
NN
PU

Feature Templates

x
My
dog
ate
my
cake
.

Templates	Patterns	Examples
Unigrams	x_t x_{t-1} x_{t-2} x_{t-3} ... x_{t+1} x_{t+2} x_{t+3} ...	ate dog My <s> my cake .
Bigrams	$x_t x_{t+1}$ $x_{t-1} x_t$ $x_{t-2} x_{t-1}$ $x_{t+1} x_{t+2}$ ...	ate/my dog/ate My/dog my/cake
Trigram	$x_{t-1} x_t x_{t+1}$ $x_t x_{t+1} x_{t+2}$ $x_{t-2} x_{t-1} x_t$ $x_{t-3} x_{t-2} x_{t-1}$ $x_{t+1} x_{t+2} x_{t+3}$	dog/ate/my ate/my/cake my/dog/ate <s>/my/dog my/cake/.
Variance	$x_{t-1} x_{t+1}$ $x_{t-2} x_{t+1}$ $x_{t-1} x_{t+2}$	dog/my My/my dog/cake

In addition to the character level, more information could be added as features.

Feature Extraction for CRFs

- Extracting features for a sequence (an instance)

```
def extract_sent_features(x):
    sent_features = []
    for i in range(len(x)):
        sent_features.append(extract_char_features(x, i))
    return sent_features
```

- Extracting features for each unit
  - Word for POS tagging or NER recognition
  - Character for Chinese word segmentation

```
def extract_char_features(sent, position):
    char_features = {}
    for i in range(-3, 4):
        if len(sent) > position + i >= 0:
            char_features['char_at_%d' % i] = sent[position + i]
    return char_features
```

Only unigram features x_{t-3}, x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, x_{t+3}

Feature Extraction for CRFs

- Extracting features for a sequence (an instance)

```
def extract_sent_features(x):
    sent_features = []
    for i in range(len(x)):
        sent_features.append(extract_char_features(x, i))
    return sent_features
```

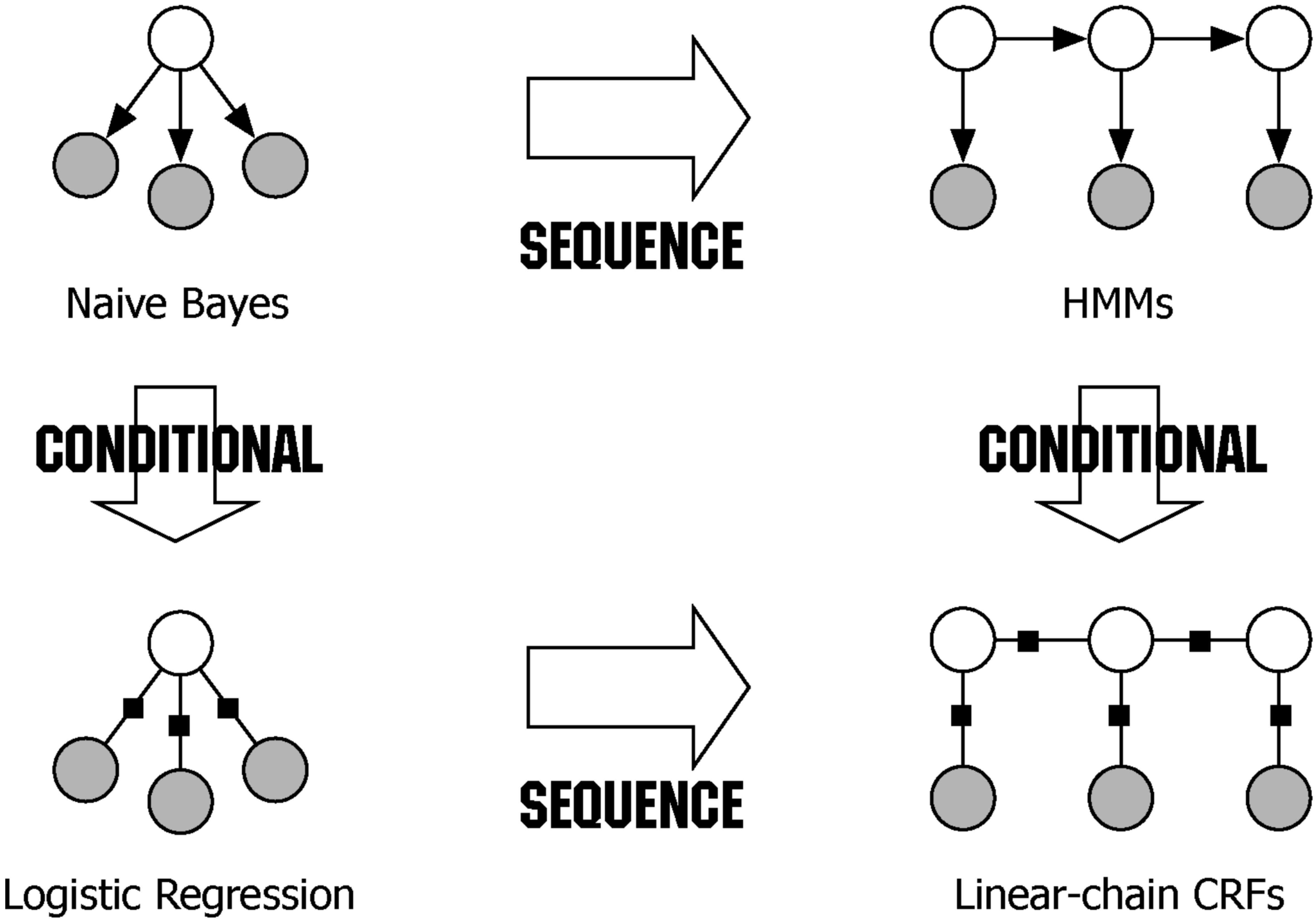
- Extracting features for each unit
    - Word for POS tagging or NER recognition
    - Character for Chinese word segmentation

```
def extract_char_features(sent, position):
    char_features = {}
    for i in range(-3, 4):
        if len(sent) > position + i >= 0:
            char_features['char_at_%d' % i] = sent[position + i]
    return char_features
```

Only unigram features x_{t-3}, x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, x_{t+3}

Summary

Model	Pros	Cons
HMM	• Simple to train • Friendly for unsupervised learning	y_t only depends on $y_{t\text{-}1}$ and x_t
MEMM	• Information of entire $\boldsymbol{x}$ can be used to decide y_t • Local optimal • Better performances	High complexity of training
CRF	• Information of entire $\boldsymbol{x}$ can be used to decide y_t • Global optimal • Even better than MEMM	Higher complexity of training



Sutton & McCallum, 2011

Information Extraction and Named Entity Recognition

Information Extraction

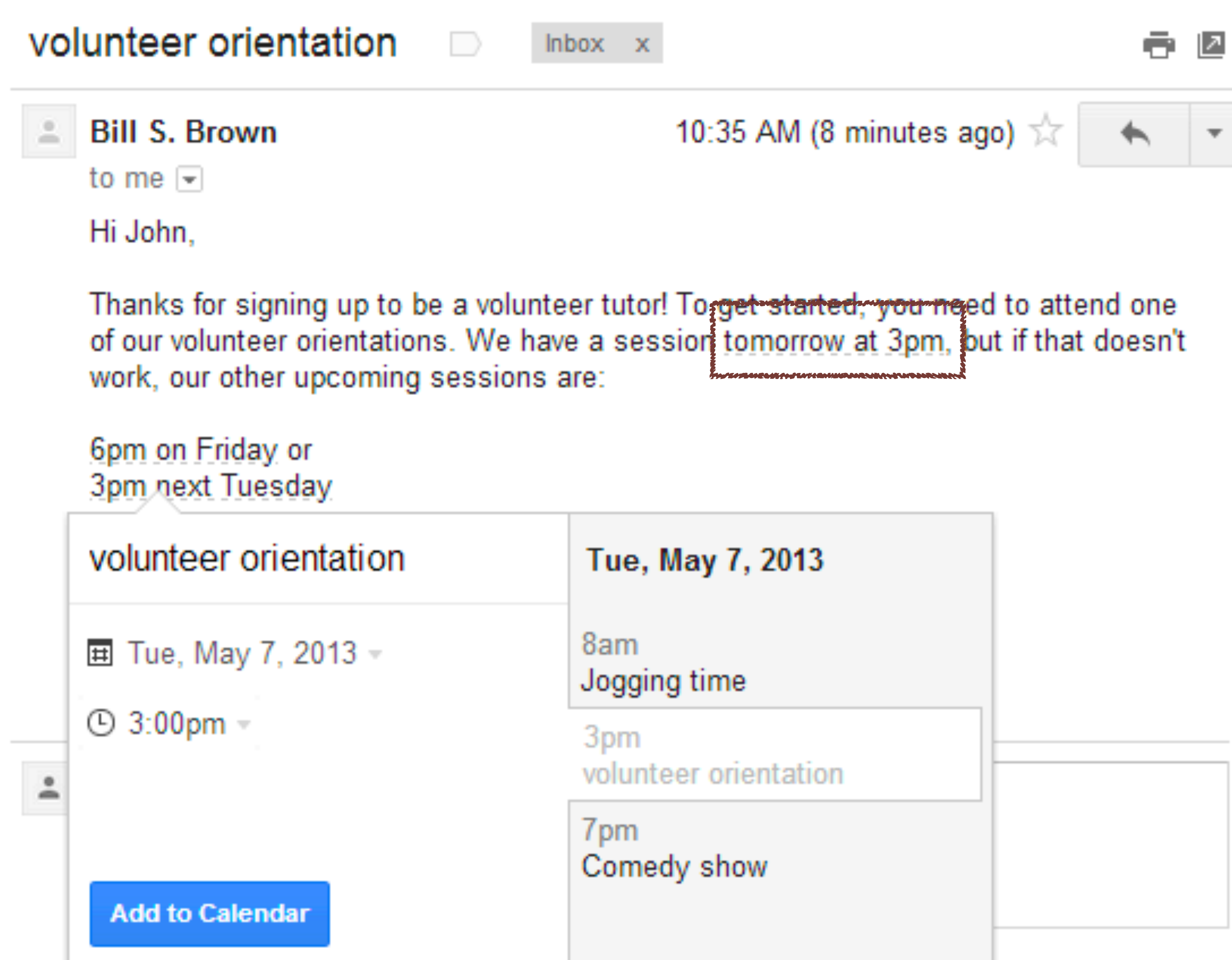
- Information extraction (IE) systems are aimed at understanding relevant parts of a long textual piece.
  - Relation extraction

... Bill Clinton's wife is Hillary ... → *Spouse*(Bill Clinton, Hillary Clinton)

- Automatically organize information for human to digest or for subsequent applications to utilize
  - KB construction for question answering
  - Text mining

Low Level Information Extraction

- Pattern matching
  - Dictionary and rules
  - Regular expression



Flight Ticket Information Extraction

Huang Hen Hsen 先生: 您在07/19/2019旅行的机票和信息 Inbox ×

Air France Flight 552
Landed - Confirmation #N8NUHI

巴黎 CDG	Terminal		台北市 TPE	Terminal	Gate
12:50 AM	2E		6:59 PM	2	D8

Air France 557
TPE to CDG Jul 19, 10:22 AM

Air France 552
CDG to TPE Aug 9, 12:50 AM

Air France <admin@ticket-airfrance.com> Unsubscribe
to me

Wed, Jul 10, 11:02 AM

Chinese > English Translate message Turn off for: Chinese ×

Jul
19
Fri

航班 台北 (Taipei)-巴黎 (Paris)

When Fri Jul 19 10:25 – Sat Jul 20, 2019 00:25 (CST)
Where Taiwan Taoyuan Intl Airport (TPE) - 航站楼 2
Who Unknown Organizer*

Add to calendar »

Agenda
Fri Jul 19, 2019

All day Stay at Hôtel L'Antoine
10:25 Flight to 巴黎 (AF 557)
10:25 航班 台北 (Taipei)-巴黎 (Paris)
No later events

Named Entity Recognition (NER)

- To identify names in text
  - An important subtask of information extraction

People's names

Roles' names

Crosby won an **Oscar for Best Actor** for his role as **Father Chuck O'Malley** in the **1944** motion picture **Going My Way** and was nominated for his reprise of the role in **The Bells of St. Mary's** opposite **Ingrid Bergman** the next year, becoming the first of six actors to be nominated twice for playing the same character. In **1963**, **Crosby** received the first **Grammy Global Achievement Award**. He is one of 33 people to have three stars on the **Hollywood Walk of Fame**, in the categories of motion pictures, radio, and audio

Date

Picture titles

Awards

Locations

Usages of Results from NER

- Named entities can indexed and linked

  Crosby won an Oscar for Best Actor for his role as Father Chuck O'Malley in the 1944 motion picture *Going My Way* and was nominated for his reprise of the role in *The Bells of St. Mary's* opposite Ingrid Bergman the next year, becoming the first of six actors to be nominated twice for playing the same character. In 1963, Crosby

- Aspect-based sentiment analysis
  - Attaching sentiment polarities to named entities
- Relation extraction
  - Further task that is aimed at finding relationship between named entities.

Sequence Labeling for NER

- Formulate the task of NER as sequence labeling

Crosby was born on **May 3, 1903** in **Tacoma, Washington.** In **1906**, his family moved to **Spokane** in eastern **Washington** state, where he was raised. In 1913, his father built a house at **508 E. Sharp Avenue**. The house sits on the campus of his alma mater, **Gonzaga University**. It functions today as a museum housing over 200 artifacts from his life and career, including his **Oscar**.

Word	Crosby	was	born	on	May	3	,	1903	in	Tacoma	,	Washington
Label	PER	O	O	O	TIME	TIME	TIME	TIME	O	LOC	LOC	LOC

BIO Scheme

- To distinguish the position of a word in a phrase
  - **B**egin, **I**nside, **O**utside

Crosby was born on **May 3, 1903** in **Tacoma, Washington.** In **1906**, his family moved to **Spokane** in eastern **Washington** state, where he was raised. In 1913, his father built a house at **508 E. Sharp Avenue**. The house sits on the campus of his alma mater, **Gonzaga University**. It functions today as a museum housing over 200 artifacts from his life and career, including his **Oscar**.

Word	Crosby	was	born	on	May	3	,	1903	in	Tacoma	,	Washington
Label	PER-B	O	O	O	TIME-B	TIME-I	TIME-I	TIME-I	O	LOC-B	LOC-I	LOC-I

Features for NER

- Words
  - Current words
  - Previous/next words (contextual information)
- POS
- Previous labels
- **Word shapes**

Word Shapes

- Mapping words to simplified surface form that captures the length, capitalization, numerals, internal punctuation marks, and so on.
  - Most proper nouns begin with a capital letter.
  - Money expressions, dates, and percentages are in numerals.

Word	Crosby	was	born	on	May	3	,	1903	in	Tacoma	,	Washington
Shapes	**Xxxxx**	**xxx**	**xxxx**	**xx**	**Xxx**	**#**	**,**	**####**	**xx**	**Xxxxxx**	**,**	**Xxxxxxxxxx**
Label	PER-B	O	O	O	TIME-B	TIME-I	TIME-I	TIME-I	O	LOC-B	LOC-I	LOC-I

Chinese Word Segmentation

Chinese Word Segmentation

- No explicit boundaries of Chinese words

The dog ate my cake

那 / 隻 / 狗 / 吃 / 了 / 我 / 的 / 蛋糕

- No Clear Definition of Chinese Words

那 / 隻 / 狗 / 吃 / 了 / 我的 / 蛋糕

國立 / 政治/ 大學 vs 國立政治大學

Ambiguity of Chinese Word Segmentation

- The determination of Chinese word boundaries is inherent ambiguity and should be resolved semantically.

- All 日 日文 文章 章魚 魚 are valid and frequent words in Chinese.

日文 / 章魚 / 怎麼 / 說

日 / 文章 / 魚 / 怎麼 / 說

Segmentation as Labeling

- We know some models for labeling each word in a sentence, but how to perform the word segmentation with the models?

- Text segmentation as sequence labeling

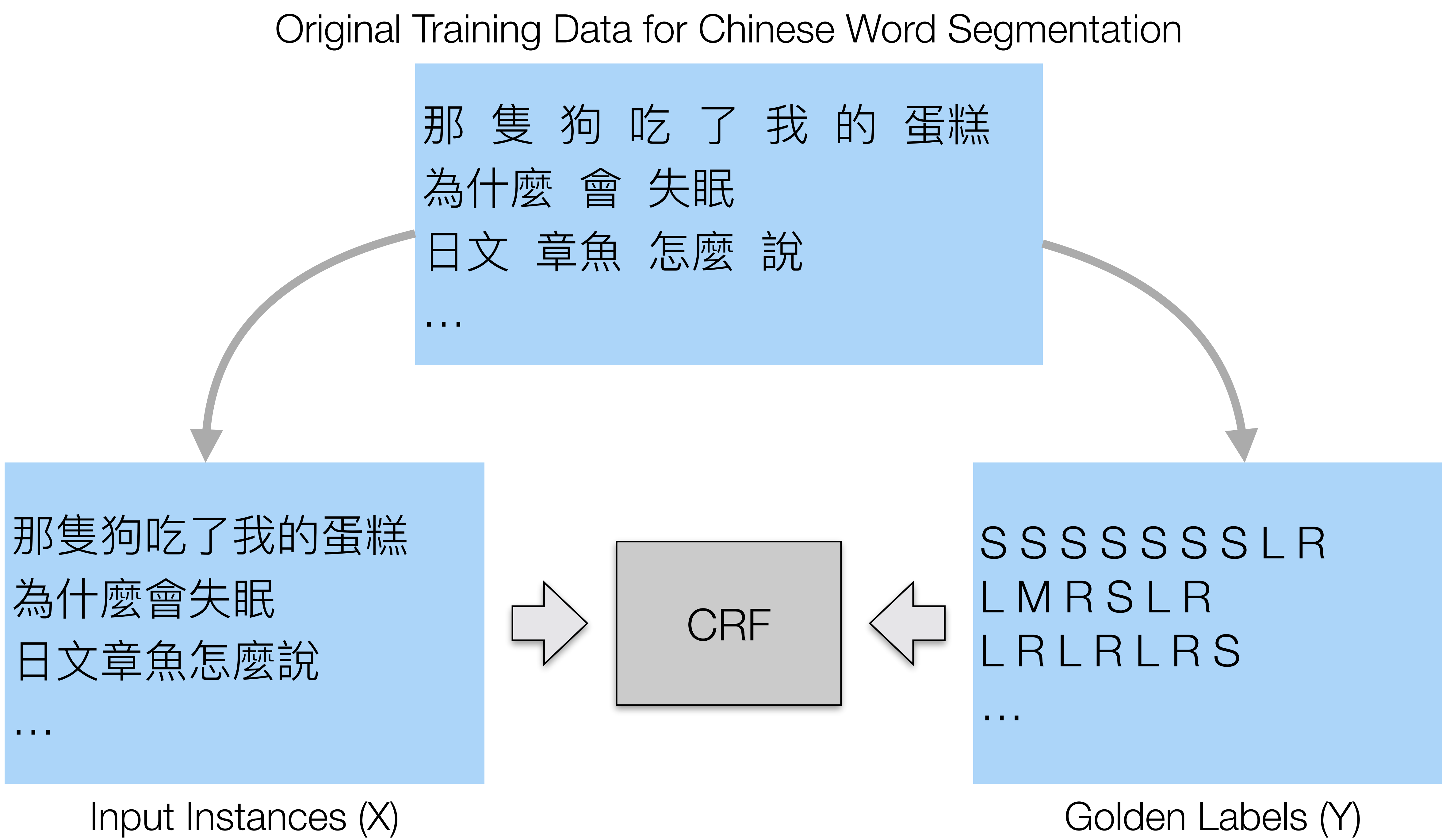
日	文	章	魚	怎	麼	說
Begin	Inside	Begin	Inside	Begin	Inside	Begin

Tagging Scheme

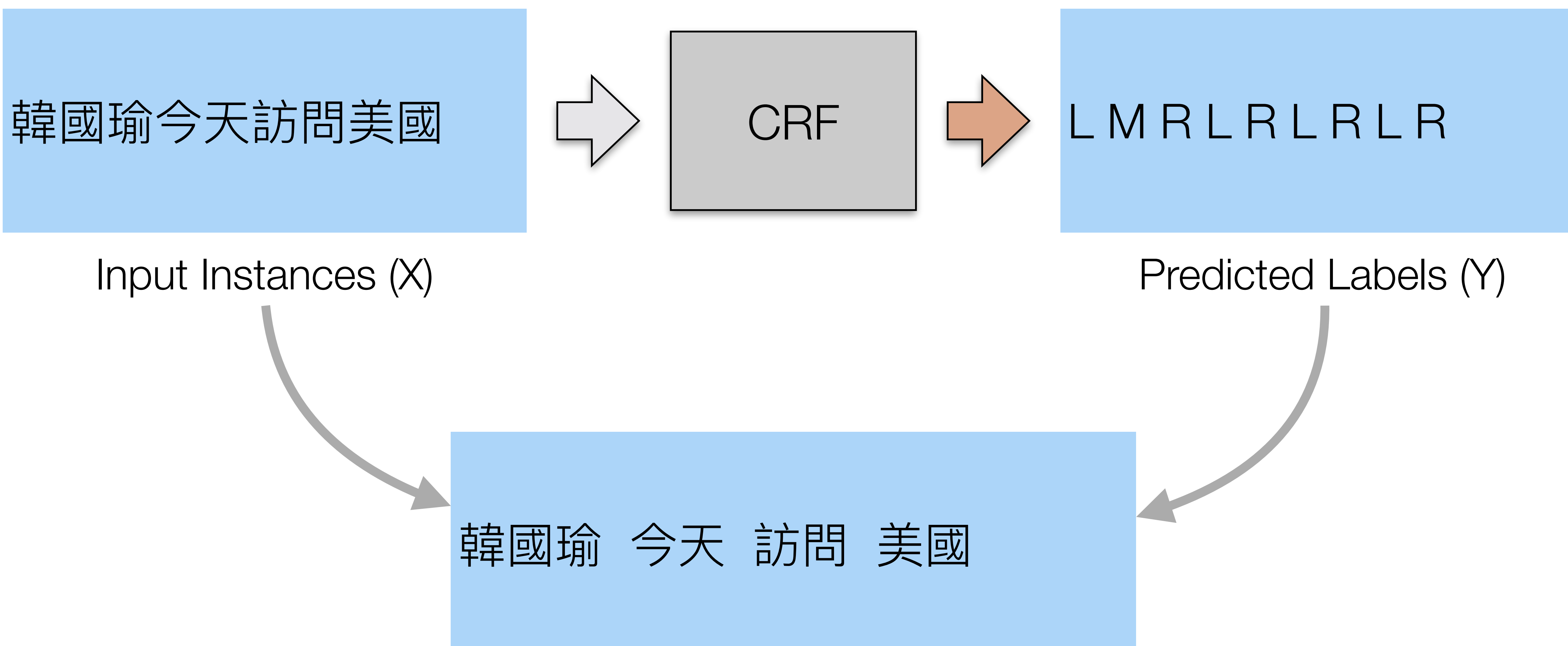
- Begin/Inside/Outside
- Begin/Middle/End
- Left/Right/Middle/Single

Scheme	為	什	麼	會	失	眠
BIO	Begin	Inside	Inside	Begin	Begin	Inside
B/M/E	Begin	Middle	End	Begin	Begin	End
L/M/R/S	Left	Middle	Right	Single	Left	Right

Training Stage



Prediction Stage



Generate the segmented tokens by adding space proceeding each L and S

Linguistic Features

- Characters
  - Unigram
  - Bigram
  - Trigram
- Other features
  - Phonetic information
  - Radical
  - Character Type

Character Type

```
import unicodedata

unicodedata.category(“資")

unicodedata.name(“訊")
```

Chr	name()	category()
5	DIGIT FIVE	Nd
b	LATIN SMALL LETTER B	Ll
Q	LATIN CAPITAL LETTER Q	Lu
æ	LATIN SMALL LETTER AE	Ll
資	CJK UNIFIED IDEOGRAPH-8CC7	Lo
한	HANGUL SYLLABLE HAN	Lo
ㄉ	BOPOMOFO LETTER D	Lo
.	FULL STOP	Po
,	COMMA	Po
。	FULLWIDTH COMMA	Po
，	IDEOGRAPHIC FULL STOP	Po
"	QUOTATION MARK	Po
「	LEFT CORNER BRACKET	Ps
😄	SMILING FACE WITH OPEN	So

Dictionary Features

- If an n-gram of characters is listed in a dictionary as a Chinese word.

t	0	1	2	3	4	5	6	7
x	<s>	日	文	章	魚	怎	麼	說
y	<s>	L	R	L	R	L	R	S
Radical	N/A	日	文	音	魚	心	广	言
Dictionary L		**T (日文）**	**T （文章）**	**T （章魚)**	**F**	**T (怎麼)**	**F**	**F**
Dictionary R		**F**	**T (日文)**	**T (文章)**	**T (章魚)**	**F**	**T (怎麼)**	**F**
Dictionary M		**F**	**F**	**F**	**F**	**F**	**F**	**F**
Dictionary S		**T**	**T**	**T**	**T**	**T**	**F**	**T**

Training Corpora for Chinese Word Segmentation

- Penn Chinese Treebank
  - With POS tagging
- 2005 Chinese Word Segmentation Bakeoff
  - Benchmark publicly used
  - http://sighan.cs.uchicago.edu/bakeoff2005/

Corpus	Training	Test	Language
Academia Sinica	708,953	14,432	Traditional
Hong Kong City University	53,019	1,493	Traditional
Microsoft Research	86,924	3,985	Simplfied
Peking University	19,056	1,945	Simplified

Evaluation for Segmentation Tasks

- score provided by Bakeoff 2005 is widely used.

$$TP : \text{Number of words correctly segmented}$$

$$FP : \text{Number of words incorrectly segmented}$$

$$N : \text{Number of words in ground-truth}$$

$$Precision = \frac{TP}{TP + FP}$$

$$\boxed{Recall = \frac{TP}{N}}$$

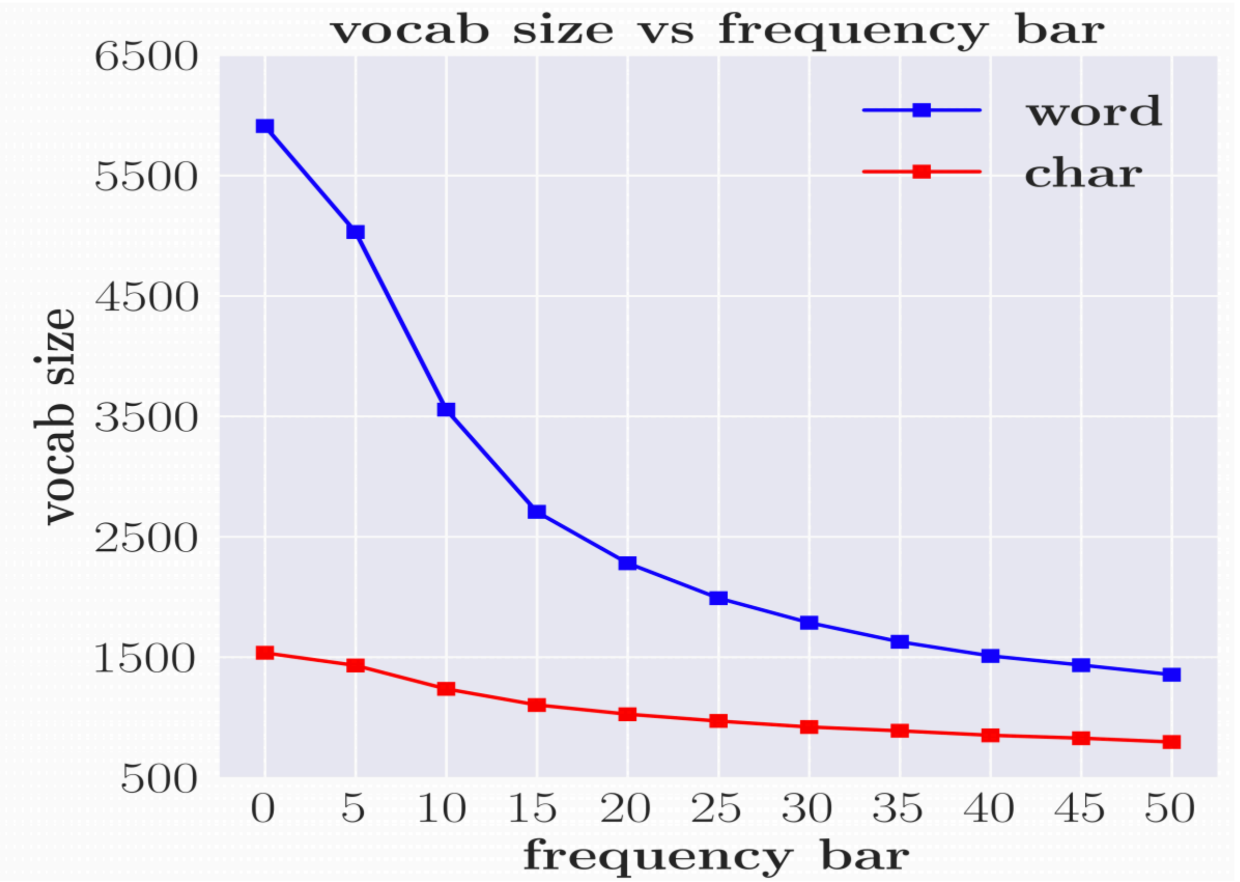
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Chinese Word Segmentation as Preprocessing

- Most NLP models for English are **word-based models**
  - We perform tokenization as the first step
  - In English, word is regarded as the basic unit of concepts and meaning.
- Until recent years, most NLP models for Chinese are also **word-based models**
  - Chinese word segmentation is one of the most wide-used preprocessing in Chinese tasks.

Disadvantages of Word-based Models for Chinese

- Sparsity of words
  - Leading to overfitting
  - Unable to deal with OOVs



bar	# distinct	prop of vocab	prop of corpus
∞	50,266	100%	100%
4	38,889	77.4%	10.1%
1	24,458	48.7%	4.0%

Disadvantages of Word-based Models for Chinese

- Errors of Chinese word segmentation will bias downstream NLP tasks.
  - If a proper name is incorrectly segmented to two tokens, and the NER model will fail to identify the proper name.
- The definition of a word may vary from humans
  - Resulting inconsistent segmentation according to different training data

<table>
<tr><th>Characters</th><th>姚</th><th>明</th><th>進</th><th>入</th><th>總</th><th>決</th><th>賽</th></tr>
<tr><td>CTB</td><td colspan="2">姚明</td><td colspan="2">進入</td><td colspan="3">總決賽</td></tr>
<tr><td>PKU</td><td>姚</td><td>明</td><td colspan="2">進入</td><td>總</td><td colspan="2">決賽</td></tr>
</table>

Disadvantages of Word-based Models for Chinese

- How much benefit Chinese word segmentation will provide it all about how much additional semantic information is present in a label Chinese word segmentation corpus.

- Today, the character-based datasets are usually much larger than the word-based datasets.

  - Millions of training pairs of English/Chinese translation

  - Chinese word segmentation datasets contain less than 100K sentences.

Word-based Models vs Character-based Models

- With deep learning, character-based models generally outperform word-based ones.

TestSet	Seq2Seq +Attn (word)	Seq2Seq +Attn (char)	Seq2Seq +Attn+BOW	Seq2Seq (char) +Attn+BOW
MT-02	42.57	44.09 (+1.52)	43.42	46.78 (+3.36)
MT-03	40.88	44.57 (+3.69)	43.92	47.44 (+3.52)
MT-04	40.98	44.73 (+3.75)	43.35	47.29 (+3.94)
MT-05	40.87	42.50 (+1.63)	42.63	44.73 (+2.10)
MT-06	39.33	42.88 (+3.55)	43.31	46.66 (+3.35)
MT-08	33.52	35.36 (+1.84)	35.65	38.12 (+2.47)
Average	39.69	42.36 (+2.67)	42.04	45.17 (+3.13)

Dataset	**description**	**char valid**	**word valid**	**char test**	**word test**
chinanews	1260K/140K/112K	91.81	91.82	91.80	**91.85** (+0.05)
dianping	1800K/200K/500K	78.80	78.47	**78.76** (+0.36)	78.40
ifeng	720K/80K/50K	86.04	84.89	**85.95** (+1.09)	84.86
jd_binary	3600K/400K/360K	92.07	91.82	**92.05** (+0.16)	91.89
jd_full	2700K/300K/250K	54.29	53.60	**54.18** (+0.81)	53.37

Remarks

- In 2019, Chinese word segmentation is no longer needed for many Chinese processing tasks.
  - The powerful pre-trained sentence encoders like BERT dominate most NLP tasks.
  - And BERT is character-based.
- However, Chinese word segmentation is still useful for simple model such as logistic regression with the bag-of-word features.
  - Finding keywords in the classification.

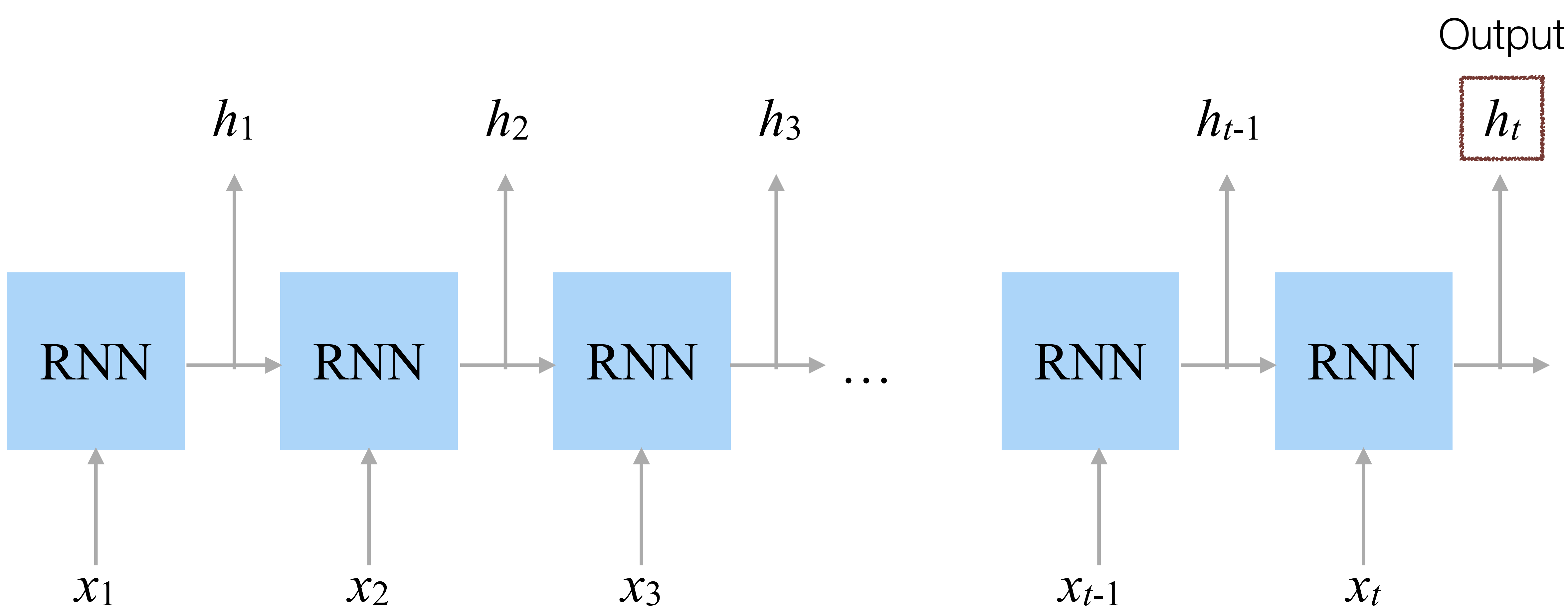
Toolkits for Chinese Word Segmentation

- jieba
  - Based on Simplified Chinese with Traditional Chinese supporting
  - Simple, fast, mediocre performance
- Stanford CoreNLP
  - Based on Simplified Chinese
  - Powerful, high performance
- CKIP
  - Focused on Traditional Chinese
  - API-based

Sequence Labeling with Deep Neural Networks

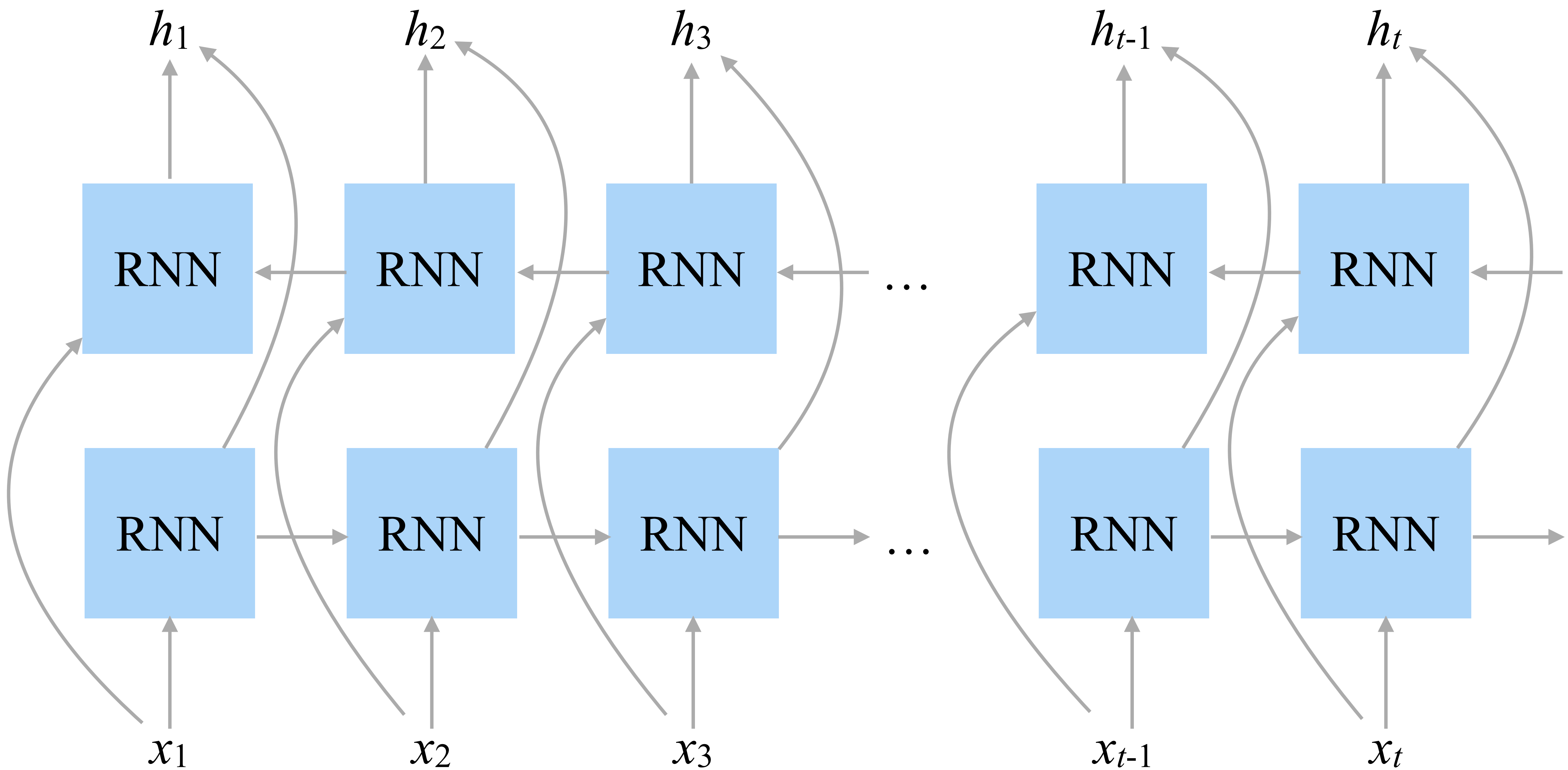
Recurrent Neural Network

- The output of step t-1 is passed to next step
- Unlike HMM, the output of step t is determined by not only the information of x_t and y_{t-1}, but also the information from 1, 2, 3, …, t-2
- No Markov assumption

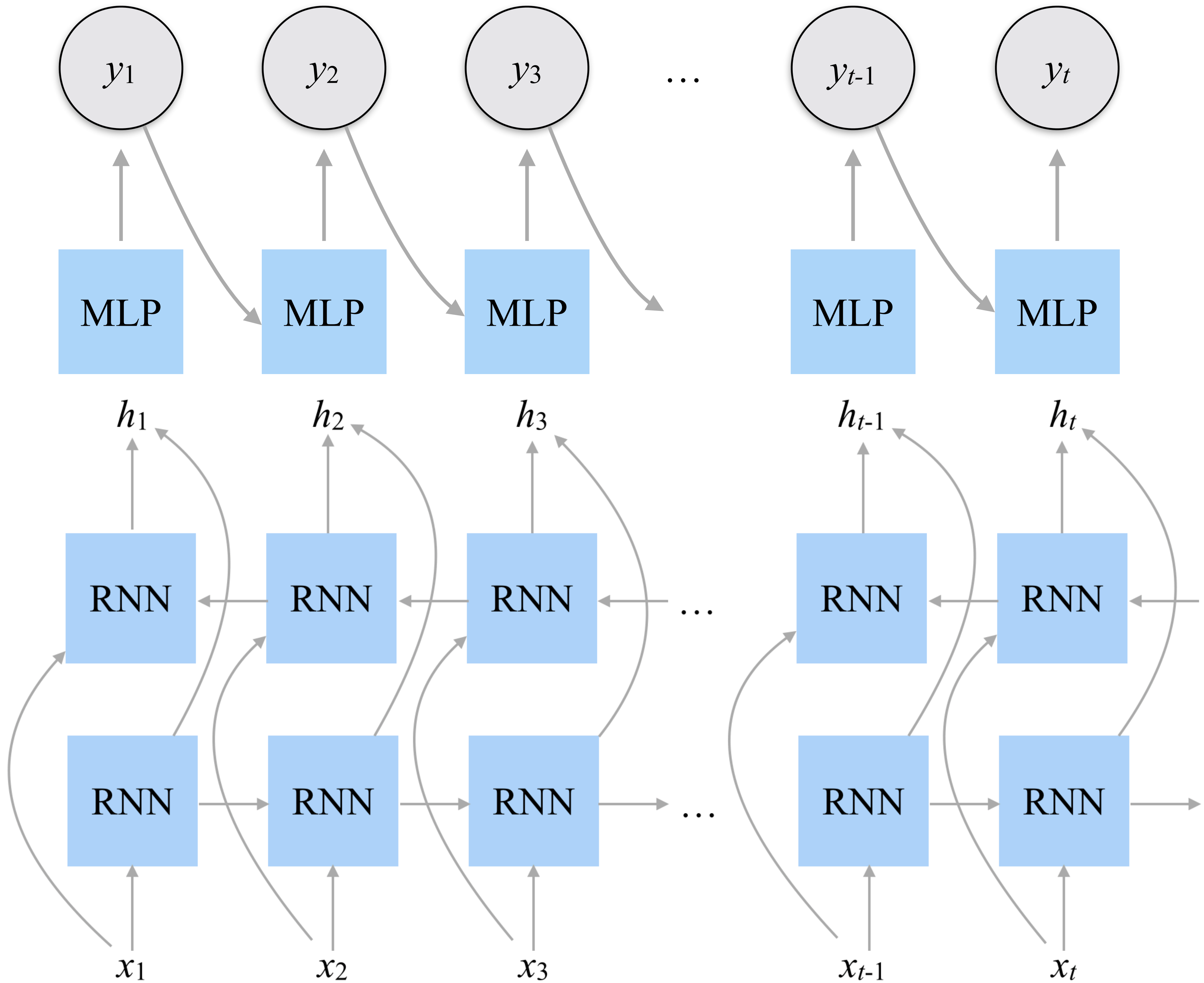


Bi-Directional RNN

- Adding additional RNN in the opposite direction.
- Take both forward/backward contextual information at the same time.



Bi-Directional RNN with CRFs



Sequence Labeling with BERT

- Given a sentence as a sequence of tokens, predict the label for each tokens
  - Token is a word (most languages) or a character (Chinese)
- POS tagging
- Chinese word segmentation
- Information extraction