### 基于神经网络的中文分词技术

Neural Models for Chinese Word Segmentation

Xinchi Chen (Fudan University)

**Advisors: Prof. Xuanjing Huang** 

**Prof. Xipeng Qiu** 

**Direction: Natural Language Processing** 

### What is Chinese word segmentation (CWS)?



中国官员应邀到美国开会中国/官员/应邀/到/美国/开会

大学生活好,还是中学生活好 大学/生活/好,还是/中学/生活/好

如果你饿了,我就下面给你吃 如果/你/饿/了,我/就/下/面/给/你/吃

### Why CWS is important?

- ➤ Parsing
- > Text classification
- > IR
- > Translation
- > QA
- **>**.....











### Why CWS is hard?



- > Low out-of-vocabulary (OOV) word recall rate
- Disambiguate
- Evaluation
  - Criteria & granularity
  - ➤ New Evaluation Metric¹

1. A New Psychometric-inspired Evaluation Metric for Chinese Word Segmentation. ACL2016. Peng Qian et al.



### 人名、地名、机构名:

刘德华长坂坡 耀华路

#### 网名:

你是我的谁旺仔小馒头

### 公司名、产品名

摩托罗拉 谷歌 爱国者 腾讯 网易 新浪诺基亚C5 尼康D700

斯坦福张首晟旗下的丹华资本暑假招收intern



斯坦福张首晟旗下的丹华资本暑假招收intern

斯坦福|张首晟|旗下|的|**丹华资|本**暑假|招收|intern

斯坦福|张首晟|旗下|的|**丹华资本**|暑假|招收|intern

### Disambiguate



这样/的/人/才能/经受住/考验这样/的/人才/能/经受住/考验

学生会/写/文章 学生/会/写/文章

乒乓球**拍/卖**/完了 乒乓球**/拍卖**/完了

#### **Evaluation**



$$R = \frac{c}{N}$$

$$P = \frac{c}{c+e}$$

$$F = \frac{2 \times P \times R}{P+R}$$

N: 黄金标准分割的单词数

e: 分词器错误标注的单词数

c: 分词器正确标注的单词数

### Granularity



上海市 上海/市

江泽民 江/泽民

### **Prevalent Approaches**



- ▶ 基于规则的分词
- > 基于统计的分词

#### 基于统计的分词



- > 基于词的方法
- ▶ 基于字的序列标注方法¹
- ▶基于词和基于字的方法的结合

1. N. Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing* 8(1):29–48.

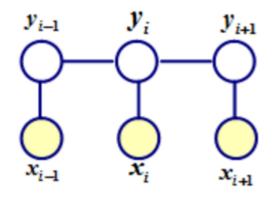
### 基于字的序列标注方法

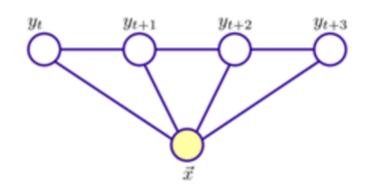


- ➤ HMM (隐马模型)
- > MEMM (最大熵隐马模型)
- > CRF (条件随机场模型)

### Conditional Random Fields (Linear-chain CRFs)

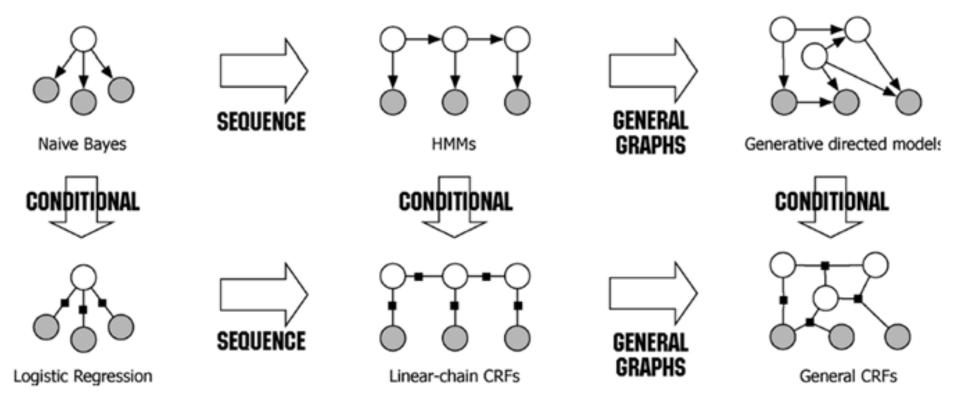






### Conditional Random Fields (Linear-chain CRFs)





#### CRF++实现中文分词



```
毎
     k
         В
日
     k
         Ι
新
         I <== 扫描到这一行,代表当前位置
聞
                   1
                      # Unigram
     k
         Ι
                      U00:%x[-2,0] ==> 毎
                   2
社
     k
         Ι
                      U01:%x[-1,0] ==> \Box
                   3
特
     k
         В
                      U02:%x[0,0] ==> 新
                   4
別
     k
         Ι
                      U03:%x[1,0] ==> 間
                   5
顧
     k
         В
                      U04:%x[2,0] ==> 社
                   6
問
                                                     ==> 毎/日/新
     k
         Τ
                   7
                      U05: %x[-2,0]/%x[-1,0]/%x[0,0]
                                                     ==> 日/新/聞
                      U06: x[-1,0]/x[0,0]/x[1,0]
                   8
4
         В
     n
                                                     ==> 新/聞/社
                   9
                      U07: %x[0,0]/%x[1,0]/%x[2,0]
                                                     ==> 日/新
                  10
                      U08: %x[-1,0]/%x[0,0]
                                                     ==> 新/聞
                  11
                      U09: %x[0,0]/%x[1,0]
                  12
                  13
                      # Bigram
                  14
                      В
```

### Neural Chinese Word Segmentation



- ▶建模复杂的特征
  - ➤ Gated Recursive Neural Network
- > 建模长距离的依赖
  - ➤ Long Short-Term Memory Neural Network
- ▶利用不同标准语料
  - Adversarial Multi-Criteria Learning for CWS

### Neural network based CWS



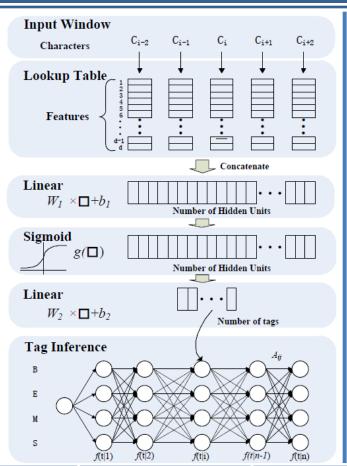
Feature Template
$C_i \ (i = -22)$
$C_i C_{i+1} \ (i = -21)$
$C_{-1}C_{1}$
$Pu(C_0)$
$T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

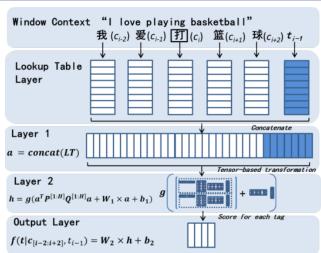
Unigram Bigram

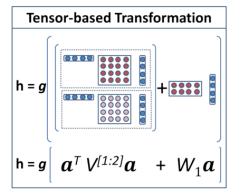
A\*A\* (一心一意)

A\*B\* (朝三暮四)

\*A\*B (朝三暮四)







Models	PKU	MSRA	СТВ6
Conventional	96.1	97.4	95.7
Zheng et al. (EMNLP 2013)	92.8	93.9	93.7
Pei et al. (ACL 2014)	95.2	97.2	-

## Optimization



- ➤ Max Margin
- > CRFs

### Max Margin



For a given training instance (xi, yi), we search for the tag sequence with the highest score:

$$y^* = \operatorname*{arg\,max}_{\hat{y} \in Y(x)} s(x_i, \hat{y}, \theta)$$

➤ The object of Max-Margin training is that the highest scoring tag sequence is the correct one: y \* = yi and its score will be larger up to a margin to other possible tag sequences

$$s(x, y_i, \theta) \ge s(x, \hat{y}, \theta) + \triangle(y_i, \hat{y})$$
$$\triangle(y_i, \hat{y}) = \sum_{i=1}^{n} \kappa \mathbf{1} \{ y_{i,j} \ne \hat{y}_j \}$$

### Max Margin



This leads to the regularized objective function for m training examples:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} l_i(\theta) + \frac{\lambda}{2} ||\theta||^2$$
  
$$l_i(\theta) = \max_{\hat{y} \in Y(x_i)} (s(x_i, \hat{y}, \theta) + \triangle(y_i, \hat{y}))$$
  
$$-s(x_i, y_i, \theta))$$

The subgradient of equation is:

$$\frac{\partial J}{\partial \theta} = \frac{1}{m} \sum_{i} \left( \frac{\partial s(x_i, \hat{y}_{max}, \theta)}{\partial \theta} - \frac{\partial s(x_i, y_i, \theta)}{\partial \theta} \right) + \lambda \theta$$

### **CRFs**



Specifically, given a sequence with n characters  $X = \{x_1, \ldots, x_n\}$ , the aim of CWS task is to figure out the ground truth of labels  $Y^* = \{y_1^*, \ldots, y_n^*\}$ :

$$Y^* = \underset{Y \in \mathcal{L}^n}{\arg \max} \, p(Y|X), \tag{1}$$

where  $\mathcal{L} = \{B, M, E, S\}.$ 

### **CRFs**



$$p(Y|X) = \frac{\Psi(Y|X)}{\sum_{Y' \in \mathcal{L}^n} \Psi(Y'|X)}.$$

Here,  $\Psi(Y|X)$  is the potential function, and we only consider interactions between two successive labels (first order linear chain CRFs):

$$\Psi(Y|X) = \prod_{i=2}^{n} \psi(X, i, y_{i-1}, y_i),$$

$$\psi(\mathbf{x}, i, y', y) = \exp(s(X, i)_y + \mathbf{b}_{y'y}),$$

$$s(X, i) = \mathbf{W}_s^{\mathsf{T}} \mathbf{h}_i + \mathbf{b}_s,$$

# Feed-forward Neural Network for CWS [Xiaoqing Zheng, Hanyang Chen, Tianyu Xu; EMNLP15]



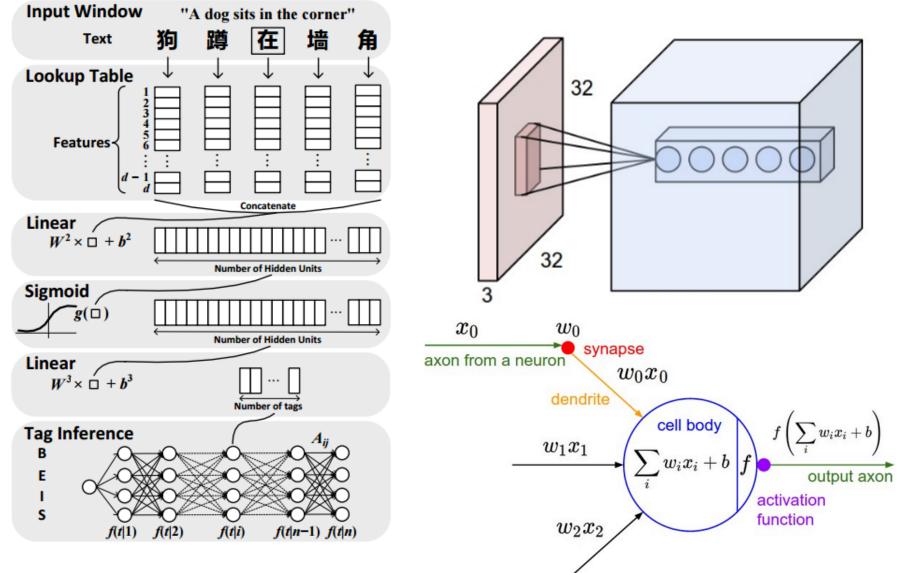
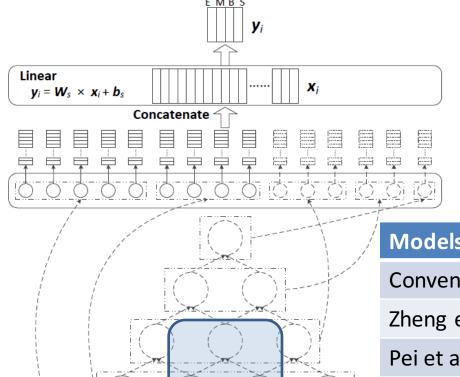
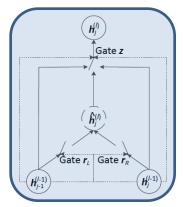


Figure 1: The neural network architecture.





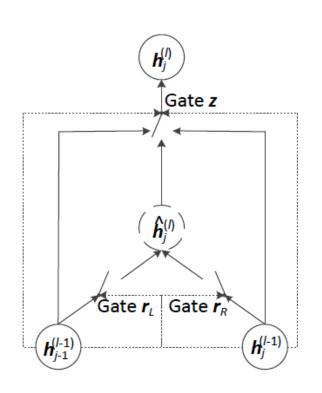


/	Models	PKU	MSRA	СТВ6
	Conventional	96.1	97.4	95.7
	Zheng et al. (EMNLP 2013)	92.8	93.9	93.7
	Pei et al. (ACL 2014)	95.2	97.2	-
	This work (ACL 2015)	96.4	97.6	95.8



### Gate mechanism



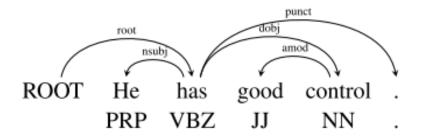


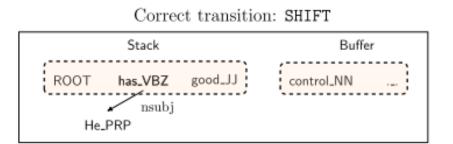
$$\begin{split} \mathbf{h}_{j}^{(l)} &= \begin{cases} \mathbf{z}_{N} \odot \hat{\mathbf{h}}_{j}^{l} + \mathbf{z}_{L} \odot \mathbf{h}_{j-1}^{l-1} + \mathbf{z}_{R} \odot \mathbf{h}_{j}^{l-1}, & l > 1, \\ \mathbf{c}_{j}, & l = 1, \end{cases} \\ \hat{\mathbf{h}}_{j}^{l} &= \tanh(\mathbf{W}_{\hat{\mathbf{h}}} \begin{bmatrix} \mathbf{r}_{L} \odot \mathbf{h}_{j-1}^{l-1} \\ \mathbf{r}_{R} \odot \mathbf{h}_{j}^{l-1} \end{bmatrix}) \\ & \begin{bmatrix} \mathbf{r}_{L} \\ \mathbf{r}_{R} \end{bmatrix} = \sigma(\mathbf{G} \begin{bmatrix} \mathbf{h}_{j-1}^{l-1} \\ \mathbf{h}_{j}^{l-1} \end{bmatrix}), \\ \mathbf{z} &= \begin{bmatrix} \mathbf{z}_{N} \\ \mathbf{z}_{L} \\ \mathbf{z}_{R} \end{bmatrix} = \begin{bmatrix} 1/Z \\ 1/Z \\ 1/Z \end{bmatrix} \odot \exp(\mathbf{U} \begin{bmatrix} \hat{\mathbf{h}}_{j}^{l} \\ \mathbf{h}_{j-1}^{l-1} \\ \mathbf{h}_{j}^{l-1} \end{bmatrix}) \end{split}$$

- Transition-based Dependency Parsing [EMNLP 2015]
- Sentence Modeling [EMNLP 2015]

### Transition based Dependency Parsing



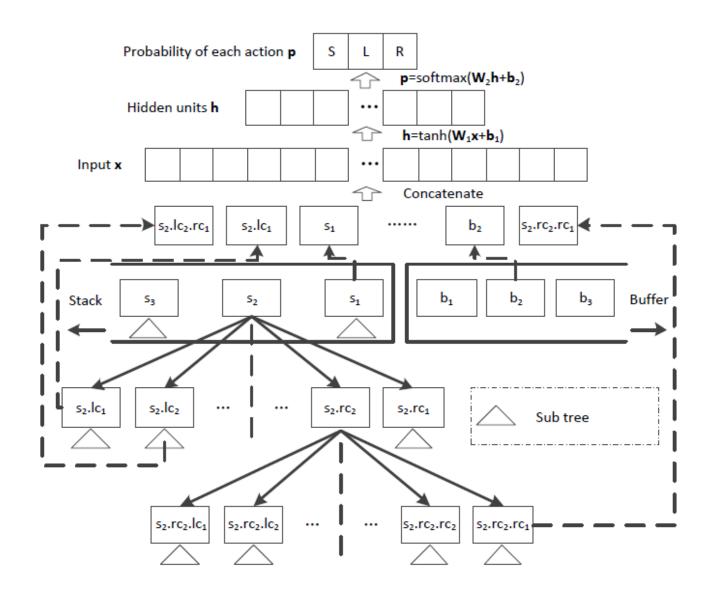




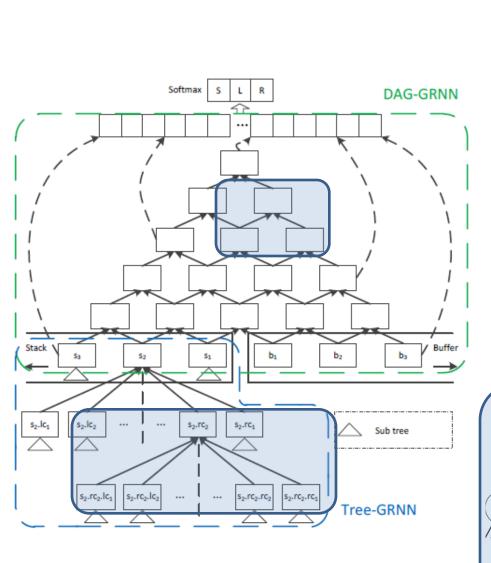
Transition	Stack	Buffer	A
	[ROOT]	[He has good control .]	0
SHIFT	[ROOT He]	[has good control .]	
SHIFT	[ROOT He has]	[good control .]	
LEFT-ARC(nsubj)	[ROOT has]	[good control .]	A∪ nsubj(has,He)
SHIFT	[ROOT has good]	[control .]	
SHIFT	[ROOT has good control]	[.]	
LEFT-ARC(amod)	[ROOT has control]	[.]	$A \cup amod(control,good)$
RIGHT-ARC(dobj)	[ROOT has]	[.]	A∪ dobj(has,control)
RIGHT-ARC(root)	[ROOT]		$A \cup \text{root}(\text{ROOT},\text{has})$

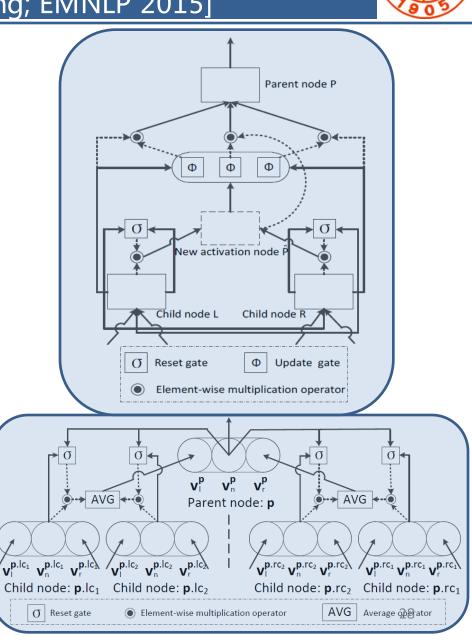
### 基于神经网络的依存句法模型





#### Transition-based Dependency Parsing Using Two Heterogeneous G Recursive Neural Networks [X Chen, Y Zhou, C Zhu, X Qiu, X Huang; EMNLP 2015]





### Experiments



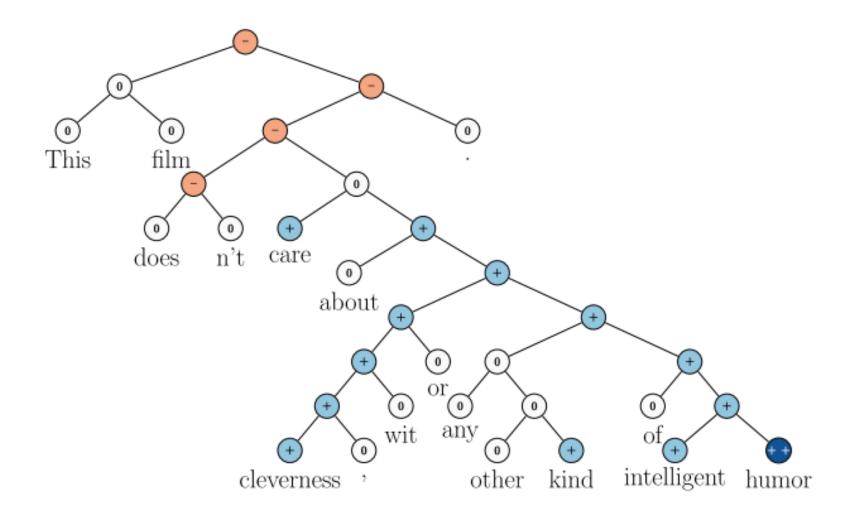
models	Dev		Test	
models	UAS	LAS	UAS	LAS
Malt:standard	90.0	88.8	89.9	88.5
Malt:eager	90.1	88.9	90.1	88.7
MSTParser	92.1	90.8	92.0	90.5
Chen's Parser	92.2	91.0	92.0	90.7
Plain	91.1	90.0	91.2	89.7
Tree-RNN	92.4	91.0	92.1	90.8
Tree-GRNN	92.6	91.1	92.4	91.0
Tree-RNN+DAG-GRNN	92.8	91.9	92.4	91.5
Tree-GRNN+DAG-GRNN	92.6	91.9	92.6	91.6

Table 2:	Performance of different models on	PTB3
dataset.	UAS: unlabeled attachment score.	LAS:
labeled a	attachment score.	

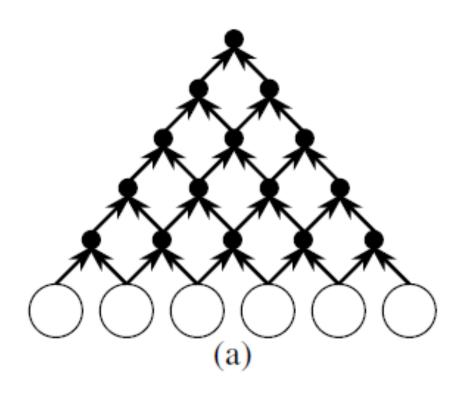
models	Dev		Test	
models	UAS	LAS	UAS	LAS
Malt:standard	82.4	80.5	82.4	80.6
Malt:eager	91.2	79.3	80.2	78.4
MSTParser	84.0	82.1	83.0	81.2
Chen's Parser	84.0	82.4	83.9	82.4
Plain	81.6	79.3	81.1	78.8
Tree-RNN	83.5	82.5	83.8	82.7
Tree-GRNN	84.2	82.5	84.3	83.1
Tree-RNN+DAG-GRNN	84.5	83.3	84.5	83.1
Tree-GRNN+DAG-GRNN	84.6	83.6	84.7	83.7

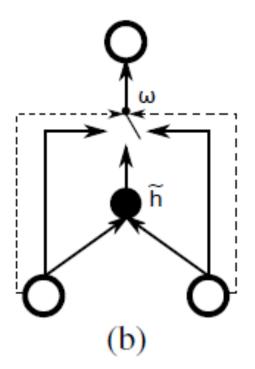
Table 3: Performance of different models on CTB5 dataset. UAS: unlabeled attachment score. LAS: labeled attachment score.



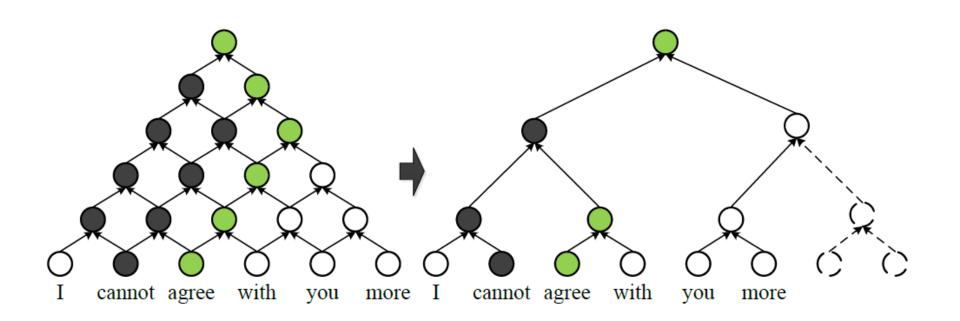












Methods	SST-1	SST-2	QC
NBoW (Kalchbrenner et al., 2014)	42.4	80.5	88.2
PV (Le and Mikolov, 2014)	44.6*	82.7*	91.8*
CNN-non-static (Kim, 2014)	48.0	87.2	93.6
CNN-multichannel (Kim, 2014)	47.4	88.1	92.2
MaxTDNN (Collobert and Weston, 2008)	37.4	77.1	84.4
DCNN (Kalchbrenner et al., 2014)	48.5	86.8	93.0
RecNTN (Socher et al., 2013b)	45.7	85.4	-
RAE (Socher et al., 2011)	43.2	82.4	-
MV-RecNN (Socher et al., 2012)	44.4	82.9	-
AdaSent (Zhao et al., 2015)	-	-	92.4
GRNN (our approach)	47.5	85.5	93.8

# Long Short-term Neural Network based CWS [X Chen, X Qiu, C Zhu, P Liu, X Huang; EMNLP 2015]

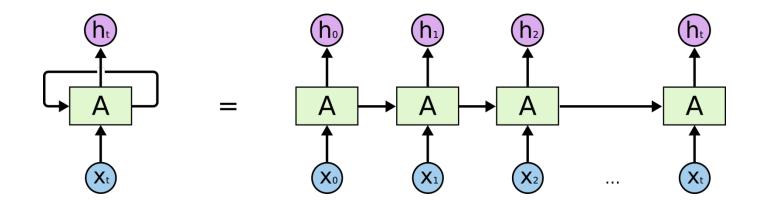


冬天(winter),能(can)穿(wear)多少(amount)穿(wear)多少(amount);

夏天(summer),能(can) 穿(wear) 多(more) 少(little) 穿(wear) 多(more) 少(little)。

#### Recurrent Neural Networks

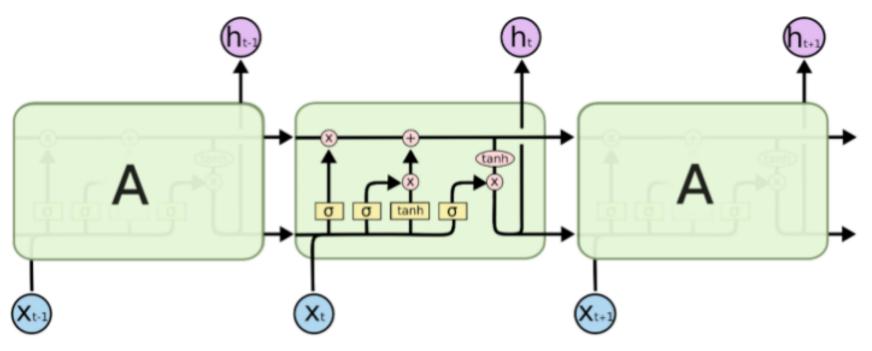




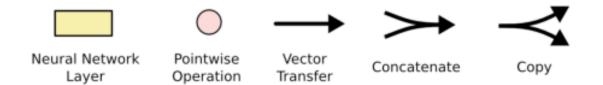
$$\mathbf{h}^{(t)} = g(\mathbf{U}\mathbf{h}^{(t-1)} + \mathbf{W}\mathbf{x}^{(t)} + \mathbf{b}),$$

# Long Short-Term Memory Neural Networks [J Schmidhuber; 1997]





The repeating module in an LSTM contains four interacting layers.



# Long Short-Term Memory Neural Networks [J Schmidhuber ; 1997]



$$\begin{bmatrix} \mathbf{i}_{i} \\ \mathbf{o}_{i} \\ \mathbf{f}_{i} \\ \tilde{\mathbf{c}}_{i} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \phi \end{bmatrix} \left( \mathbf{W}_{g}^{\mathsf{T}} \begin{bmatrix} \mathbf{e}_{x_{i}} \\ \mathbf{h}_{i-1} \end{bmatrix} + \mathbf{b}_{g} \right),$$

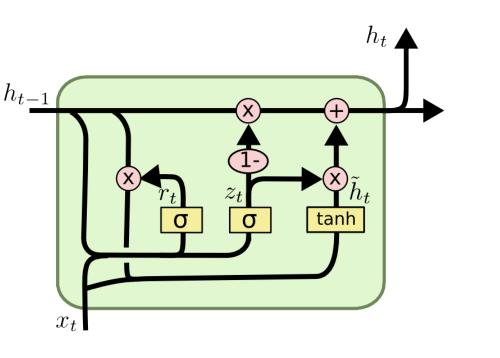
$$\mathbf{c}_{i} = \mathbf{c}_{i-1} \odot \mathbf{f}_{i} + \tilde{\mathbf{c}}_{i} \odot \mathbf{i}_{i},$$

$$\mathbf{h}_{i} = \mathbf{o}_{i} \odot \phi(\mathbf{c}_{i}),$$

#### Gated Recurrent Neural Networks



Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling



$$z_{t} = \sigma (W_{z} \cdot [h_{t-1}, x_{t}])$$

$$r_{t} = \sigma (W_{r} \cdot [h_{t-1}, x_{t}])$$

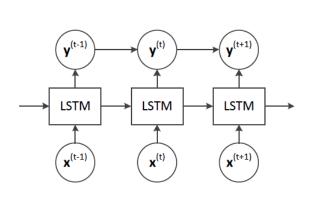
$$\tilde{h}_{t} = \tanh (W \cdot [r_{t} * h_{t-1}, x_{t}])$$

$$h_{t} = (1 - z_{t}) * h_{t-1} + z_{t} * \tilde{h}_{t}$$

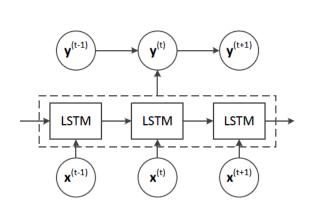
## Long Short-term Neural Network based CWS

[X Chen, X Qiu, C Zhu, P Liu, X Huang; EMNLP 2015]

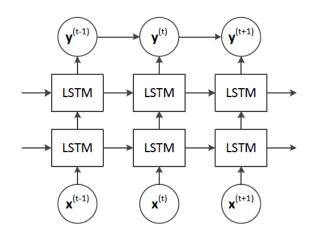




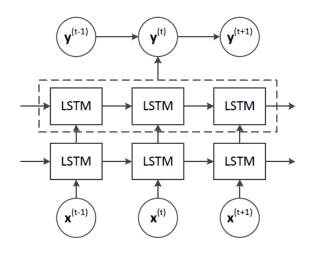
(a) LSTM-1



(c) LSTM-3



(b) LSTM-2



(d) LSTM-4

# Long Short-term Neural Network based CWS [X Chen, X Qiu, C Zhu, P Liu, X Huang; EMNLP 2015]



Models	PKU	MSRA	СТВ6
Conventional	96.1	97.4	95.7
Zheng et al. (EMNLP 2013)	92.8	93.9	93.7
Pei et al. (ACL 2014)	95.2	97.2	-
GRNN(ACL 2015)	96.4	97.6	95.8
This work (EMNLP 2015)	96.5	97.4	96.0

## Adversarial Multi-Criteria Learning for Chinese Word Segmentation [X Chen, Zhan Shi, X Qiu, X Huang; ACL 2017]

Corpora	Yao	Ming	reaches	the	final
CTB	姚明		进入	总决赛	
PKU	姚	明	进入	总	决赛

Table 1: Illustration of the different segmentation criteria.

## Adversarial Multi-Criteria Learning for Chinese Word Segmentation [X Chen, Zhan Shi, X Qiu, X Huang; ACL 2017]



Formally, assume that there are M corpora with heterogeneous segmentation criteria. We refer  $\mathcal{D}_m$  as corpus m with  $N_m$  samples:

$$\mathcal{D}_m = \{ (X_i^{(m)}, Y_i^{(m)}) \}_{i=1}^{N_m}, \tag{2}$$

where  $X_i^m$  and  $Y_i^m$  denote the *i*-th sentence and the corresponding label in corpus *m*.

## Adversarial Multi-Criteria Learning for Chinese Word Segmentation [X Chen, Zhan Shi, X Qiu, X Huang; ACL 2017]



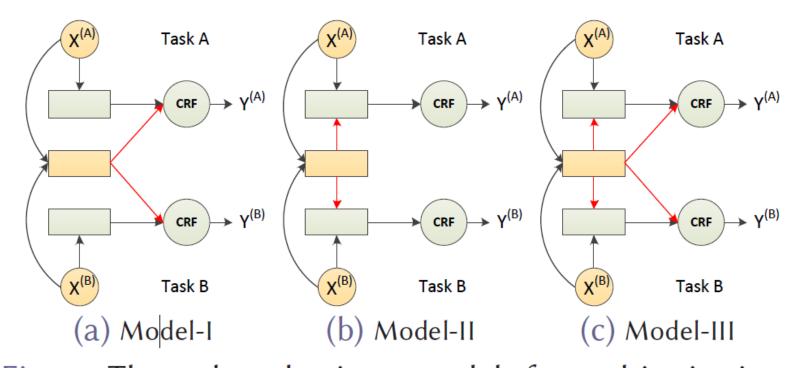


Figure: Three shared-private models for multi-criteria learning. The yellow blocks are the shared Bi-LSTM layer, while the gray block are the private Bi-LSTM layer. The yellow circles denote the shared embedding layer. The red information flow indicates the difference between three models.

#### Objective function



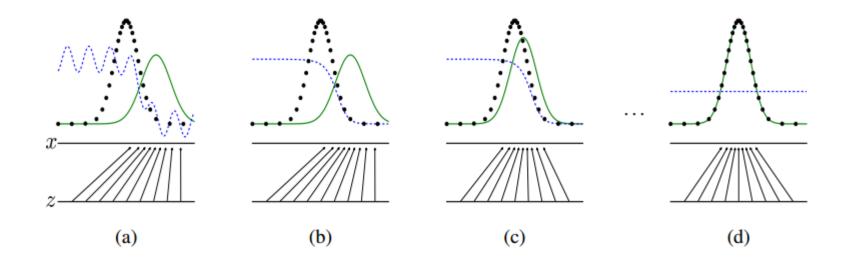
The parameters of the network are trained to maximize the log conditional likelihood of true labels on all the corpora. The objective function  $\mathcal{J}_{seg}$  can be computed as:

$$\mathcal{J}_{seg}(\Theta^{m}, \Theta^{s}) = \sum_{m=1}^{M} \sum_{i=1}^{N_{m}} \log p(Y_{i}^{(m)} | X_{i}^{(m)}; \Theta^{m}, \Theta^{s}),$$
(18)

where  $\Theta^m$  and  $\Theta^s$  denote all the parameters in private and shared layers respectively.

## Generative Adversarial Networks [IJ Goodfellow - 2014]

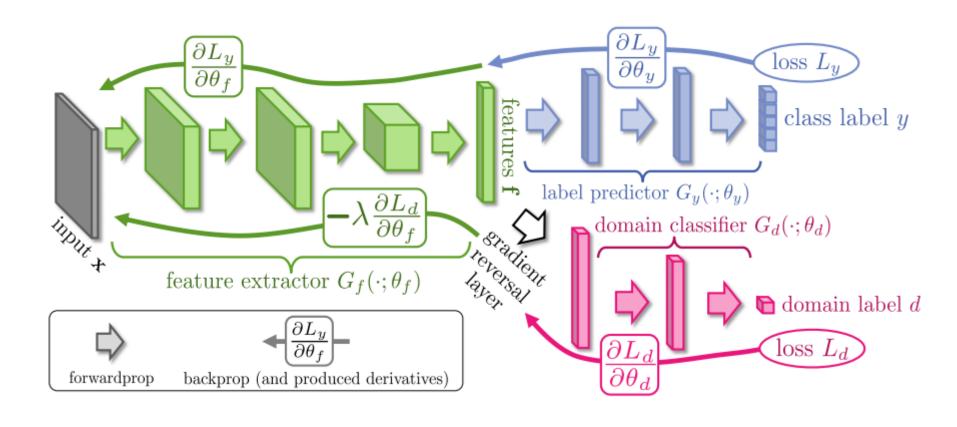




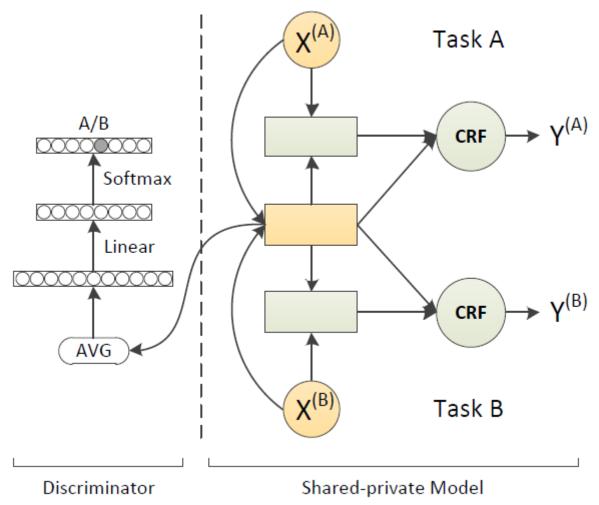
$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log (1 - D(G(\boldsymbol{z})))].$$

# Unsupervised Domain Adaptation by Backpropagation [Yaroslav Ganin, et al.]









#### Adversarial loss function



The criterion discriminator maximizes the cross-entropy of predicted criterion distribution  $p(\mid X)$  and true criterion.

$$\max_{\Theta^d} \mathcal{J}_{adv}^1(\Theta^d) = \sum_{m=1}^M \sum_{i=1}^{N_m} \log p(m|X_i^{(m)}; \Theta^d, \Theta^s). \quad (20)$$

An adversarial loss aims to produce shared features, such that a criterion discriminator cannot reliably predict the criterion by using these shared features. Therefore, we maximize the entropy of predicted criterion distribution when training shared parameters.

$$\max_{\Theta^s} \mathcal{J}_{adv}^2(\Theta^s) = \sum_{m=1}^M \sum_{i=1}^{N_m} \mathbf{H}\left(p(m|X_i^{(m)}; \Theta^d, \Theta^s)\right),\,$$

### **Training**



$$\mathcal{J}(\Theta; \mathcal{D}) = \mathcal{J}_{seg}(\Theta^m, \Theta^s) + \lambda \mathcal{J}_{adv}^1(\Theta^d) + \lambda \mathcal{J}_{adv}^2(\Theta^s),$$

## Algorithm 1 Adversarial multi-criteria learning for CWS task.

```
1: for i = 1; i \le n epoch; i + + do
        # Train tag predictor for CWS
        for m = 1; m <= M; m + + do
 3.
            # Randomly pick data from corpus m
 4:
 5: \mathcal{B} = \{X, Y\}_1^{b_m} \in \mathcal{D}^m
     \Theta^s += \alpha \nabla_{\Theta^s} \mathcal{J}(\Theta; \mathcal{B})
 6.
            \Theta^m += \alpha \nabla_{\Theta^m} \mathcal{J}(\Theta; \mathcal{B})
 7:
       end for
 8.
 9:
        # Train criterion discriminator
        for m = 1; m <= M; m + + do
10.
            \mathcal{B} = \{X, Y\}_1^{b_m} \in \mathcal{D}^m
11:
            \Theta^d = \alpha \nabla_{\Theta^d} \mathcal{J}(\Theta; \mathcal{B})
12:
13.
        end for
14: end for
```

### Experiments



	Dataset	S	$N_{ m w}$	$N_{ m c}$	$ \mathcal{D}_w $	$ \mathcal{D}_c $	$N_s$
05	MSRA	Train	2.4M	4.1M	88.1K	5.2K	86.9K
Sighan05	WISICA	Test	0.1M	0.2M	12.9K	2.8K	4.0K
lgh	AS	Train	5.4M	8.4M	141.3K	6.1K	709.0K
S	AS	Test	0.1M	0.2M	18.8K	3.7K	14.4K
	PKU	Train	1.1M	1.8M	55.2K	4.7K	47.3K
	TKU	Test	0.2M	0.3M	17.6K	3.4K	6.4K
	СТВ	Train	0.6M	1.1M	42.2K	4.2K	23.4K
	CID	Test	0.1M	0.1M	9.8K	2.6K	2.1K
80	CKIP	Train	0.7M	1.1M	48.1K	4.7K	94.2K
Sighan08	CKII	Test	0.1M	0.1M	15.3K	3.5K	10.9K
lgh	CITYU	Train	1.1M	1.8M	43.6K	4.4K	36.2K
S	CITIO	Test	0.2M	0.3M	17.8K	3.4K	6.7K
	NCC	Train	0.5M	0.8M	45.2K	5.0K	18.9K
	NCC	Test	0.1M	0.2M	17.5K	3.6K	3.6K
	SXU	Train	0.5M	0.9M	32.5K	4.2K	17.1K
	DAU	Test	0.1M	0.2M	12.4K	2.8K	3.7K

Models		MSRA	AS	PKU	CTB	CKIP	CITYU	NCC	SXU	Avg.
	P	95.70	93.64	93.67	95.19	92.44	94.00	91.86	95.11	93.95
Bi-LSTM	R	95.99	94.77	92.93	95.42	93.69	94.15	92.47	95.23	94.33
D1-L311VI	F	95.84	94.20	93.30	95.30	93.06	94.07	92.17	95.17	94.14
	OOV	66.28	70.07	66.09	76.47	72.12	65.79	59.11	71.27	68.40
	P	95.69	93.89	94.10	95.20	92.40	94.13	91.81	94.99	94.03
Stacked Bi-LSTM	R	95.81	94.54	92.66	95.40	93.39	93.99	92.62	95.37	94.22
Stacked DI-LSTM	F	95.75	94.22	93.37	95.30	92.89	94.06	92.21	95.18	94.12
	OOV	65.55	71.50	67.92	75.44	70.50	66.35	57.39	69.69	68.04
Multi-Criteria Learr										
	P	95.67	94.44	94.93	95.95	93.99	95.10	92.54	96.07	94.84
Model-I	R	95.82	95.09	93.73	96.00	94.52	95.60	92.69	96.08	94.94
Wiodel-1	F	95.74	94.76	94.33	95.97	94.26	95.35	92.61	96.07	94.89
	OOV	69.89	74.13	72.96	81.12	77.58	80.00	64.14	77.05	74.61
	P	95.74	94.60	94.82	95.90	93.51	95.30	92.26	96.17	94.79
Model-II	R	95.74	95.20	93.76	95.94	94.56	95.50	92.84	95.95	94.94
Model-11	F	95.74	94.90	94.28	95.92	94.03	95.40	92.55	96.06	94.86
	OOV	69.67	74.87	72.28	79.94	76.67	81.05	61.51	77.96	74.24
	P	95.76	93.99	94.95	95.85	93.50	95.56	92.17	96.10	94.74
Model-III	R	95.89	95.07	93.48	96.11	94.58	95.62	92.96	96.13	94.98
Wiodel-III	F	95.82	94.53	94.21	95.98	94.04	95.59	92.57	96.12	94.86
	OOV	70.72	72.59	73.12	81.21	76.56	82.14	60.83	77.56	74.34
Adversarial Multi-C	riteria L	earning								
	P	95.95	94.17	94.86	96.02	93.82	95.39	92.46	96.07	94.84
Model-I+ADV	R	96.14	95.11	93.78	96.33	94.70	95.70	93.19	96.01	95.12
Wiodel-1+ADV	F	96.04	94.64	94.32	96.18	94.26	95.55	92.83	96.04	94.98
	OOV	71.60	73.50	72.67	82.48	77.59	81.40	63.31	77.10	74.96
	P	96.02	94.52	94.65	96.09	93.80	95.37	92.42	95.85	94.84
Model-II+ADV	R	95.86	94.98	93.61	95.90	94.69	95.63	93.20	96.07	94.99
	F	95.94	94.75	94.13	96.00	94.24	95.50	92.81	95.96	94.92
	OOV	72.76	75.37	73.13	82.19	77.71	81.05	62.16	76.88	75.16
	P	95.92	94.25	94.68	95.86	93.67	95.24	92.47	96.24	94.79
Model-III+ADV	R	95.83	95.11	93.82	96.10	94.48	95.60	92.73	96.04	94.96
Model-III+AD (	F	95.87	94.68	94.25	95.98	94.07	95.42	92.60	96.14	94.88
	OOV	70.86	72.89	72.20	81.65	76.13	80.71	63.22	77.88	74.44



### Traditional & Simplified Chinese



	AS	CKIP	CITYU	Avg.
Base line	94.20	93.06	94.07	93.78
This work	94.12	93.24	95.20	94.19

Table 5: Performance on 3 traditional Chinese datasets with shared parameters fixed shared. The shared parameters are trained on 5 simplified Chinese datasets. Here, we conduct Model-I without incorporating adversarial training strategy.

#### **Error Analysis**



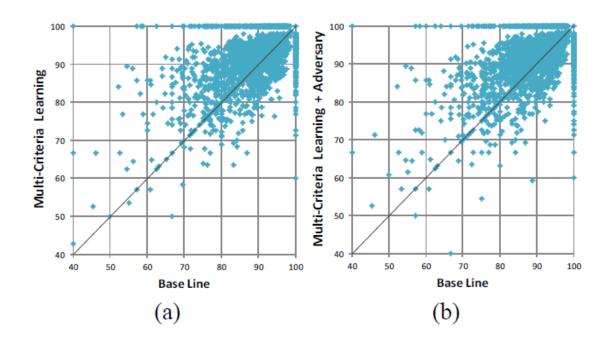


Figure 5: F-measure scores on test set of CITYU dataset. Each point denotes a sentence, with the (x, y) values of each point denoting the F-measure scores of the two models, respectively. (a) is comparison between Bi-LSTM and Model-I. (b) is comparison between Bi-LSTM and Model-I with adversarial training.

### Case Study



Models	PKU-2333		MSRA-89		
Golds	Lu Wu Xuan 卢 武铉		Mu Ling Ying 穆玲英		
Base Line	卢武铉		穆	玲英	
Model-I	卢武铉		穆 玲英		
Modell-I+ADV	卢 武铉		穆玲英		

Table 6: Segmentation cases of personal name.

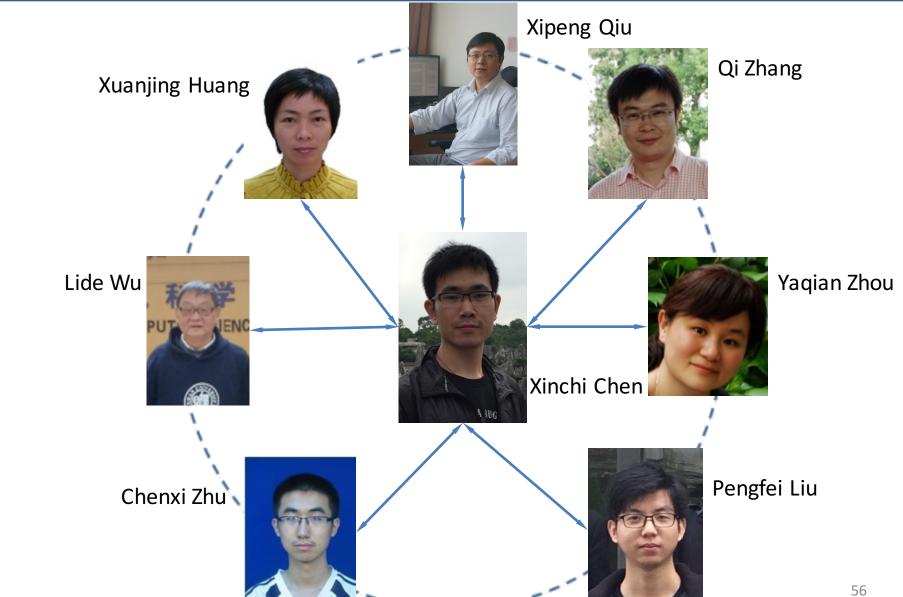
### Other Neural Models



- 1. Jianqiang Ma and Erhard W Hinrichs. 2015. **Accurate linear-time chinese word segmentation via embedding matching**. In *ACL* (1). pages 1733–1743.
- 2. Deng Cai and Hai Zhao. 2016. **Neural word segmentation learning for chinese**. *arXiv preprint arXiv:1606.04300*.
- 3. Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2009. A discriminative latent variable chinese segmenter with hybrid word/character information. NAACL. pages 56–64.
- 4. Yijia Liu, Wanxiang Che, Jiang Guo, Bing Qin, and Ting Liu. 2016. **Exploring segment representations for neural segmentation models**. *arXiv preprint arXiv:1604.05499*.
- 5. ......

## Team work





### **Publications**



- Xinchi Chen, Xipeng Qiu, Chenxi Zhu & Xuanjing Huang, Gated Recursive Neural Network For Chinese Word Segmentation, In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2015.
- ◆ Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu & Xuanjing Huang, Long Short-Term Memory Neural Networks for Chinese Word Segmentation, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.
- Xinchi Chen, Yaqian Zhou, Chenxi Zhu, Xipeng Qiu & Xuanjing Huang, Transition-based Dependency Parsing Using Two Heterogeneous Gated Recursive Neural Networks, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.
- ◆ Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Shiyu Wu & Xuanjing Huang, Sentence Modeling with Gated Recursive Neural Network, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015 [short].
- Chenxi Zhu, Xipeng Qiu, Xinchi Chen & Xuanjing Huang, A Re-Ranking Model For Dependency Parser With Recursive Convolutional Neural Network, In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2015.
- PengFei Liu, Xipeng Qiu, Xinchi Chen, Shiyu Wu & Xuanjing Huang, Multi-Timescale Long Short-Term Memory Neural Network for Modelling Sentences and Documents, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.
- ◆ Xinchi Chen, Zhan Shi, Xipeng Qiu & Xuanjing Huang, Adversarial Multi-Criteria Learning for Chinese Word Segmentation, In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2017.

#### Thank you for your attention!

## Thank you for your attention!

**Xinchi Chen (Fudan University)** 

Advisors: Prof. Xuanjing Huang

**Prof. Xipeng Qiu** 

**Direction: Natural Language Processing**