Neural Chinese Word Segmentation with Dictionary Knowledge

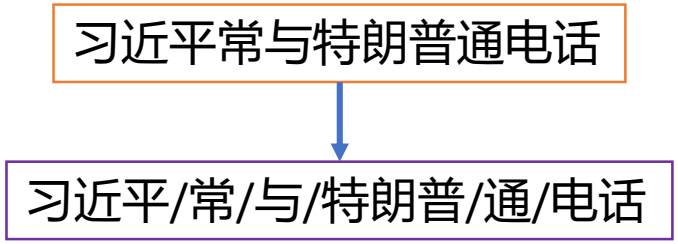
Junxin Liu[1], Fangzhao Wu[2], Chuhan Wu[1],
Yongfeng Huang[1], and Xing Xie[2]

Department of Electronic Engineering, Tsinghua University[1]
Microsoft Research Asia[2]

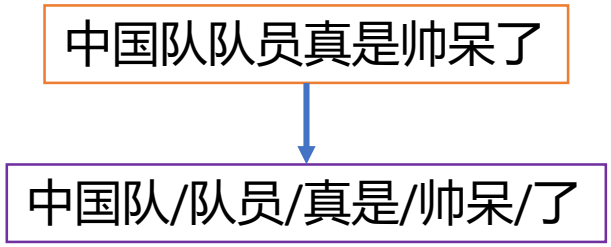
Chinese Word Segmentation

- Segment a Chinese sentence into a list of words
  - There is no natural delimiter such as whitespace to segment words in Chinese texts

Example 1



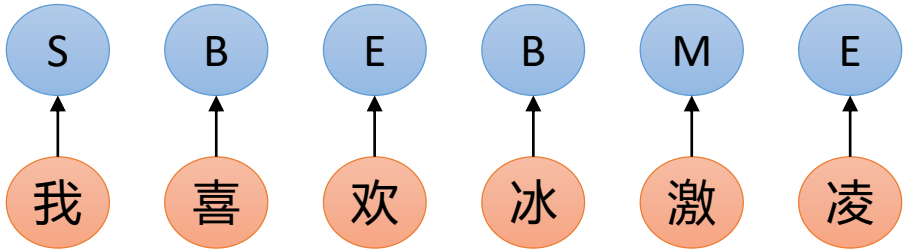
Example 2



- An important task in Chinese NLP field
- An essential step for many downstream tasks

Chinese Word Segmentation

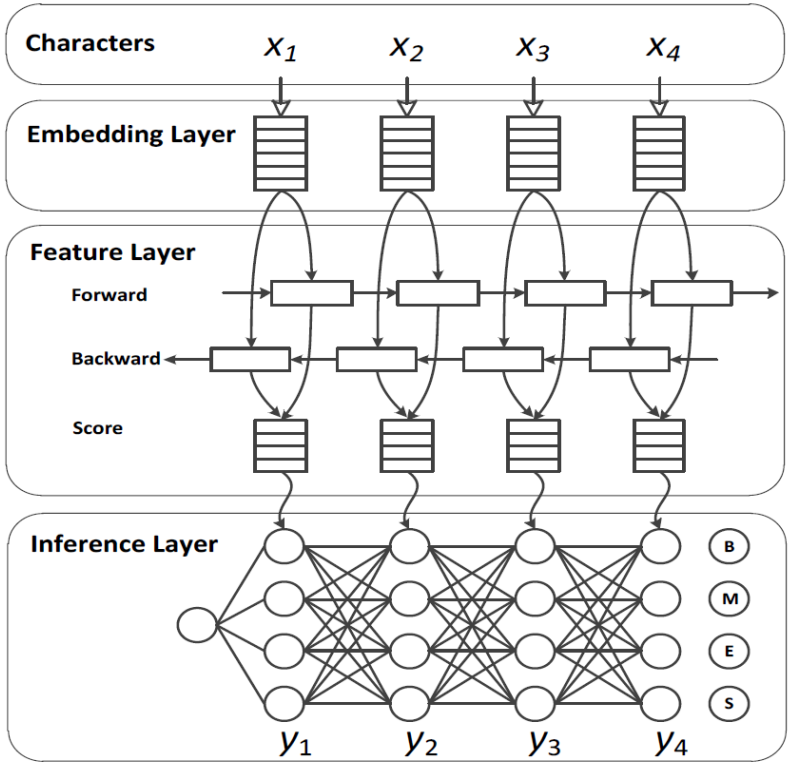
- Chinese word segmentation is usually modeled as a character-level sequence labeling task



- Deep learning methods have been widely used in Chinese word segmentation

Neural Chinese Word Segmentation

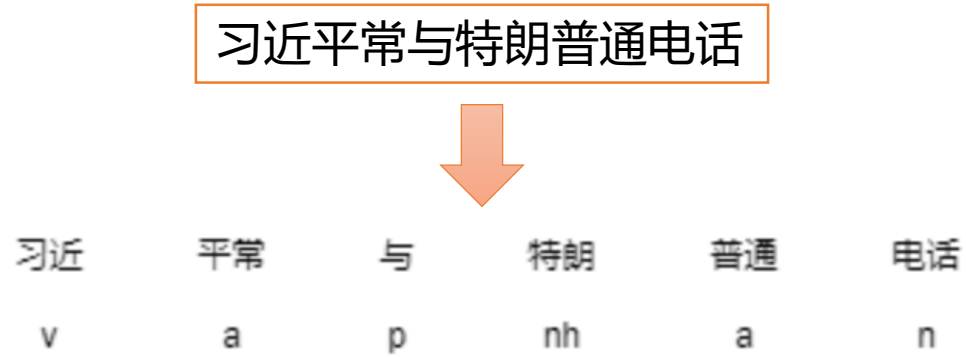
- LSTM-CRF architecture is very popular in neural Chinese word segmentation methods



Chen X, Shi Z, Qiu X, et al. Adversarial Multi-Criteria Learning for Chinese Word Segmentation, ACL. 2017, 1: 1193-1203.

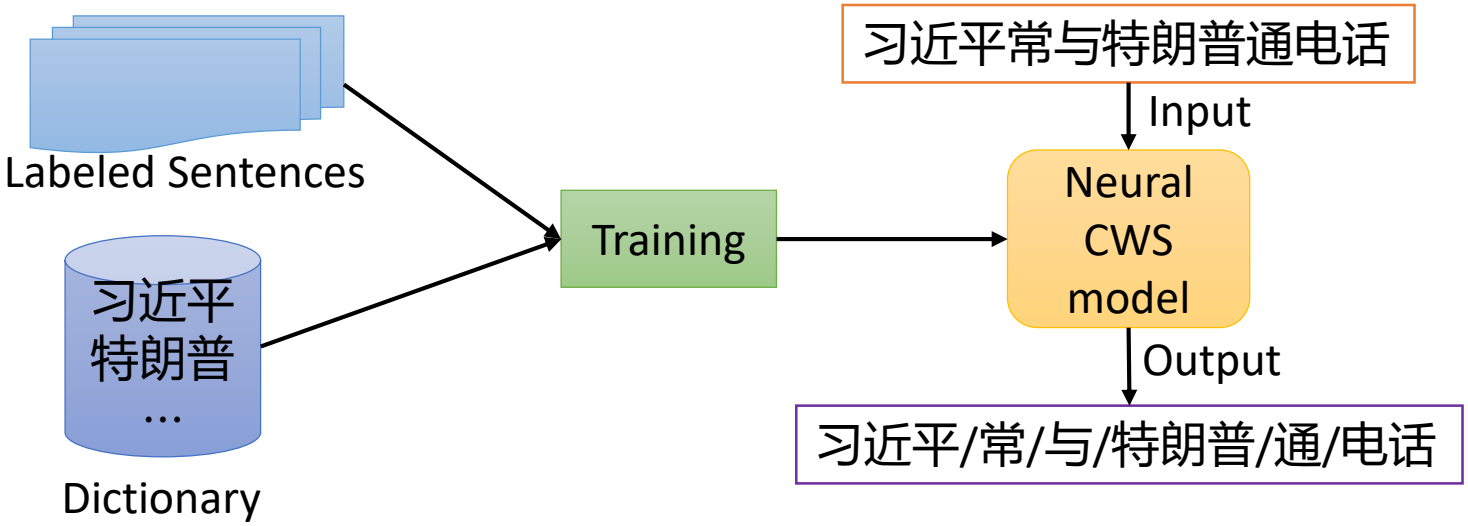
Neural Chinese Word Segmentation

- Advantages
  - End-to-end learning, free of feature engineering
  - Achieved promising results on benchmark datasets
- Drawbacks
  - Rely on a large number of labeled samples
  - Poor performance on OOV words and rare words
    - Very time-consuming and expensive to annotate sufficient sentences to cover every OOV and rare word for many times



Motivation

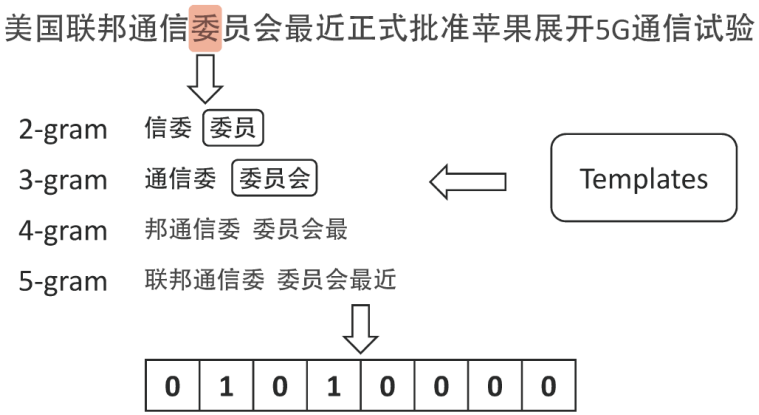
- Many OOV words and rare words are included and well-defined in many Chinese dictionaries
- Incorporate Chinese dictionaries into neural CWS
  - Improve the performance on OOV and rare words
  - Reduce the dependence on labeled data

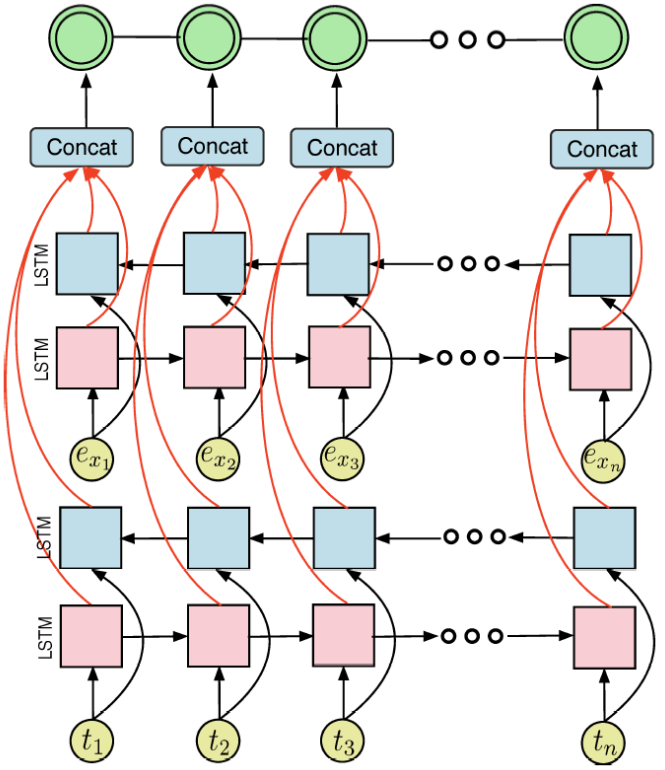


Existing Methods

- Existing method of neural CWS with dictionary
  - Feature engineering

Type	Template
2-gram	$x_{i-1}x_i,\ \ x_ix_{i+1}$
3-gram	$x_{i-2}x_{i-1}x_i,\ \ x_ix_{i+1}x_{i+2}$
4-gram	$x_{i-3}x_{i-2}x_{i-1}x_i,\ \ x_ix_{i+1}x_{i+2}x_{i+3}$
5-gram	$x_{i-4}x_{i-3},\ldots,x_i,\ \ x_ix_{i+1},\ldots,x_{i+4}$

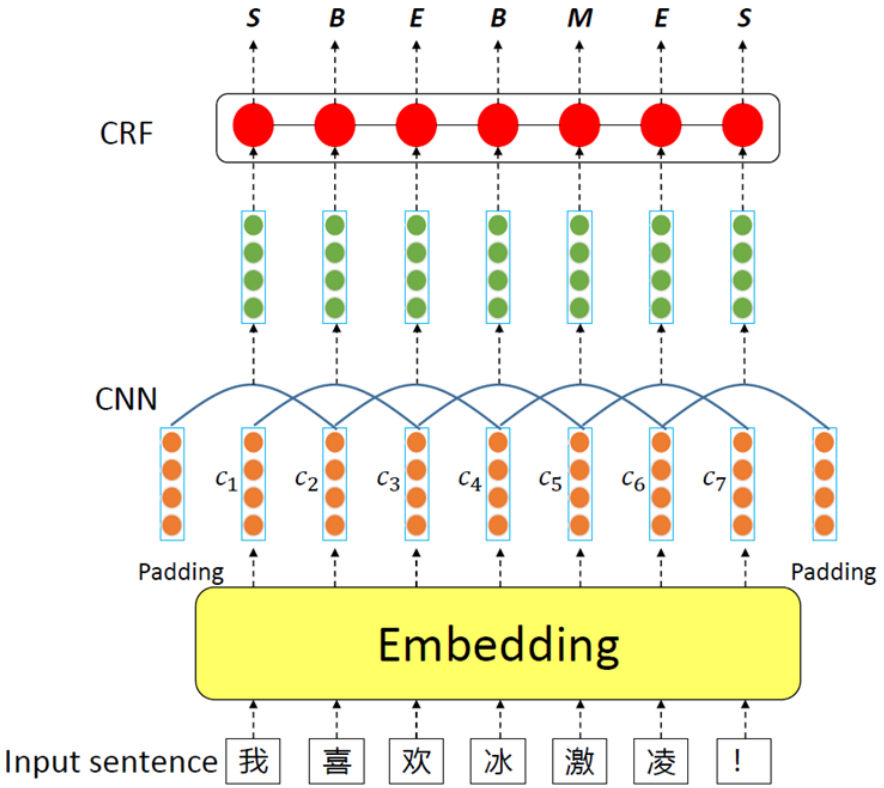




Zhang Q, Liu X, Fu J. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation. AAAI 2018.

Our Approach: Basic Model

- CNN-CRF architecture



$$\mathcal{L} = -\sum_{i=1}^{N} \log(p(\mathbf{y}_i|\mathbf{x}_i))$$

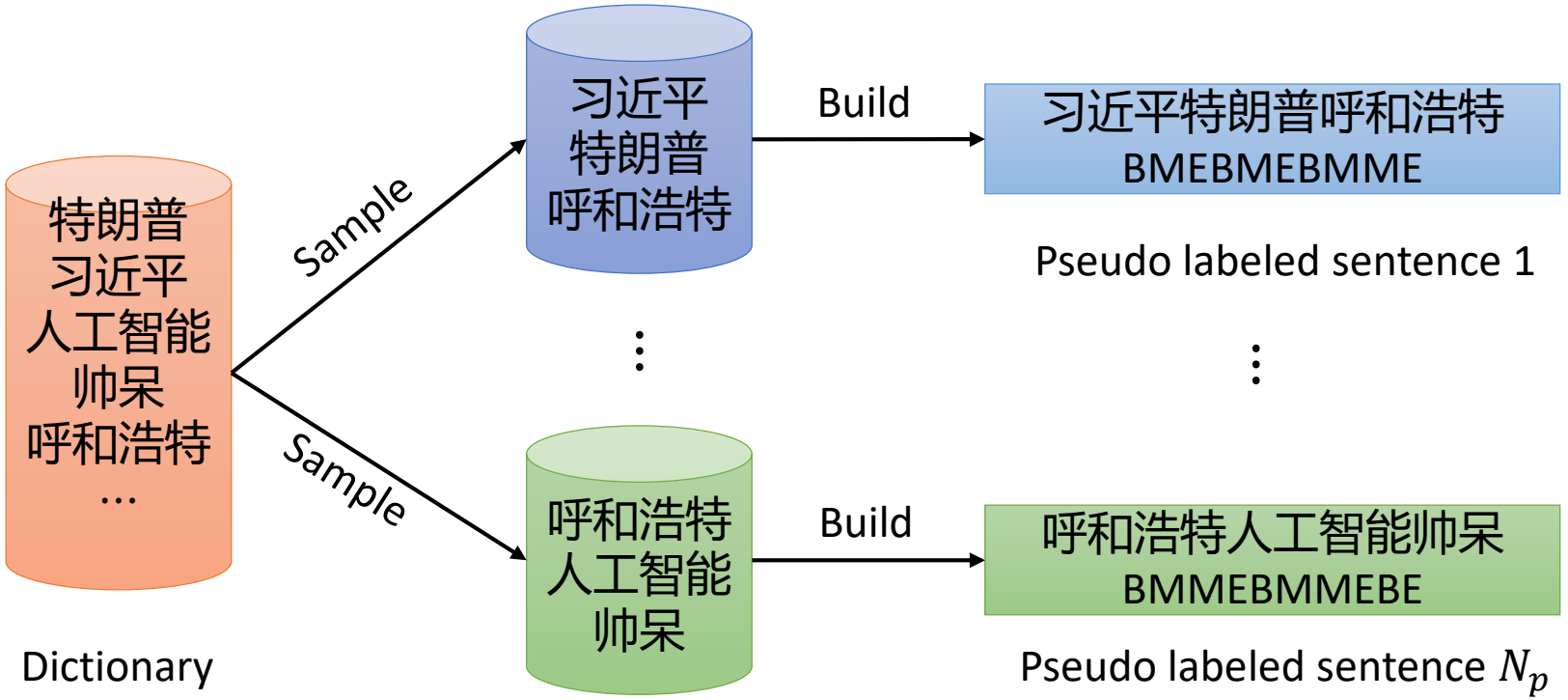
$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp(g(\mathbf{x},\mathbf{y}))}{\sum_{\mathbf{y}'\in\mathcal{Y}(\mathbf{x})}\exp(g(\mathbf{x},\mathbf{y}'))}$$

$$g(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M}(S_{i,y_i} + A_{y_{i-1},y_i}) \qquad S_i = \mathbf{W}^T\mathbf{h}_i + \mathbf{b},$$

$$h_i = f(\mathbf{w}^T \times \mathbf{c}_{i-\lceil\frac{k-1}{2}\rceil:i+\lfloor\frac{k-1}{2}\rfloor} + b)$$

Our Approach 1

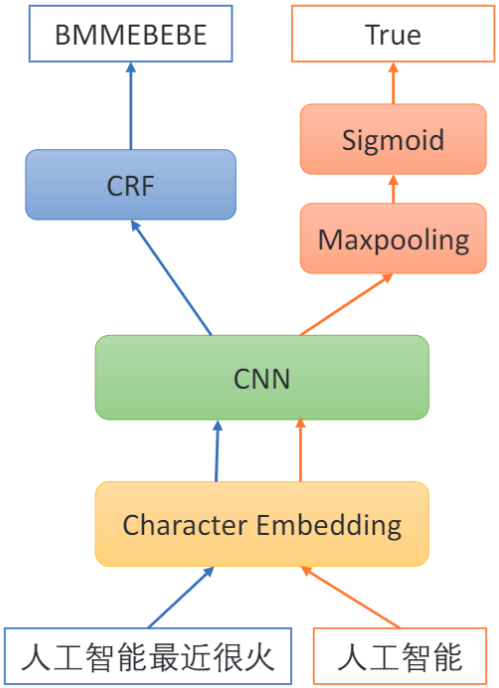
- Dictionary as pseudo labeled data
  - Use dictionary to build pseudo labeled sentences for CWS



$$\mathcal{L} = -\sum_{i=1}^{N} \log(p(\mathbf{y}_i|\mathbf{x}_i)) - \lambda_1 \sum_{i=1}^{N_p} \log(p(\mathbf{y}_i^s|\mathbf{x}_i^s))$$

Our Approach 2

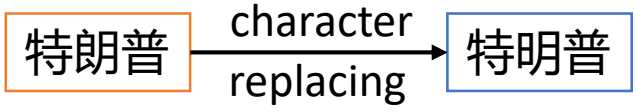
- Dictionary as word-level supervision
  - Word classification: classifying whether a character sequence is a Chinese word
  - Multi-task learning: jointly train CWS and word classification models



$$\mathcal{L} = -(1-\lambda_2)\sum_{i=1}^{N}\log(p(\mathbf{y}_i|\mathbf{x}_i)) + \lambda_2\sum_{i=1}^{N_w}\log(1+e^{-y_i s_i})$$

Labeled data for word classification:

- Positive sample: words sampled from dictionary
- Negative sample: negative sampling from real words with character replacing



Positive Negative

Experiments

- Datasets
  - Two benchmark datasets released by the third international Chinese language processing bakeoff

Dataset		#Sentence	#Word	#Character	OOV Rate
MSRA	Train	46.3K	1.27M	2.17M	-
	Test	4.4K	0.10M	0.17M	3.4%
UPUC	Train	18.8K	0.51M	0.83M	-
	Test	5.1K	0.15M	0.26M	8.8%

- Dictionary
  - Sogou Internet Chinese dictionary
- Our codes and data will be available at https://github.com/liujunxin/CWS_with_Dictionary

Experiments

- Performance evaluation

	1%			10%			100%		
	P	R	F	P	R	F	P	R	F
Chen et al. [3]	75.50	75.80	75.64	87.71	86.22	86.96	94.24	93.35	93.80
LSTM-CRF	75.88	74.86	75.36	85.52	84.81	85.16	94.26	93.29	93.78
CNN-CRF	75.59	74.43	75.00	89.72	89.14	89.43	95.03	94.53	94.78
Zhang et al. [17]	75.75	75.95	75.85	89.52	89.01	89.27	95.71	95.41	95.56
Ours_Pseudo	80.58	77.97	79.25	90.49	89.59	90.04	95.36	94.71	95.03
Ours_Multi	78.47	77.31	77.88	89.91	89.27	89.59	95.10	94.50	94.80

Results on the *MSRA* dataset

Experiments

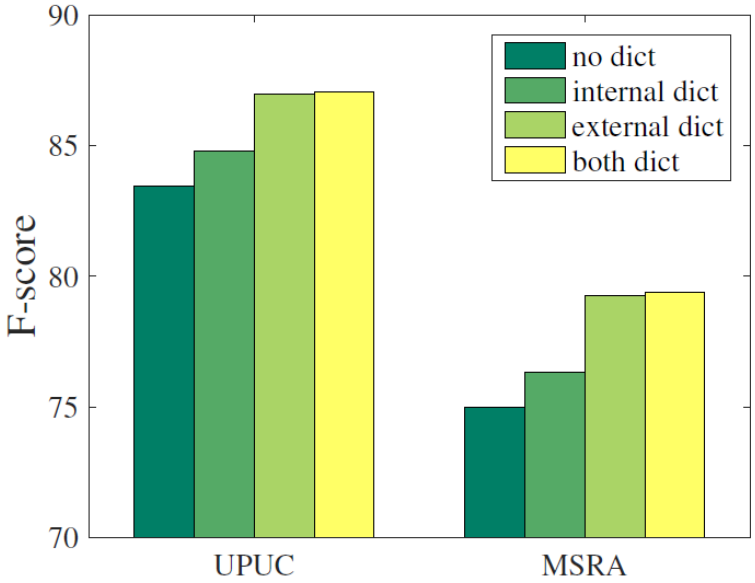
- Performance evaluation

	5%			25%			100%		
	P	R	F	P	R	F	P	R	F
Chen et al. [3]	82.31	82.60	82.44	88.00	89.90	88.94	90.79	92.92	91.84
LSTM-CRF	81.08	80.88	80.98	86.76	88.40	87.57	91.39	92.58	91.98
CNN-CRF	82.44	84.50	83.46	89.95	91.57	90.75	92.22	93.84	93.02
Zhang et al. [17]	83.38	84.98	84.17	89.93	91.41	90.66	92.60	93.89	93.24
Ours_Pseudo	87.37	86.56	86.97	90.97	92.04	91.50	92.77	94.09	93.43
Ours_Multi	84.59	86.22	85.40	90.43	91.68	91.05	92.35	93.93	93.13

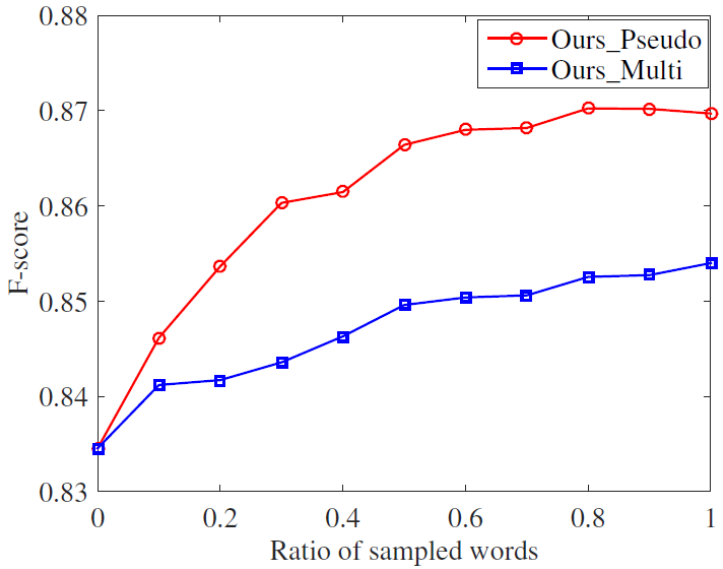
Results on the *UPUC* dataset

Experiments

- Influence of Dictionary



Influence of dictionary type

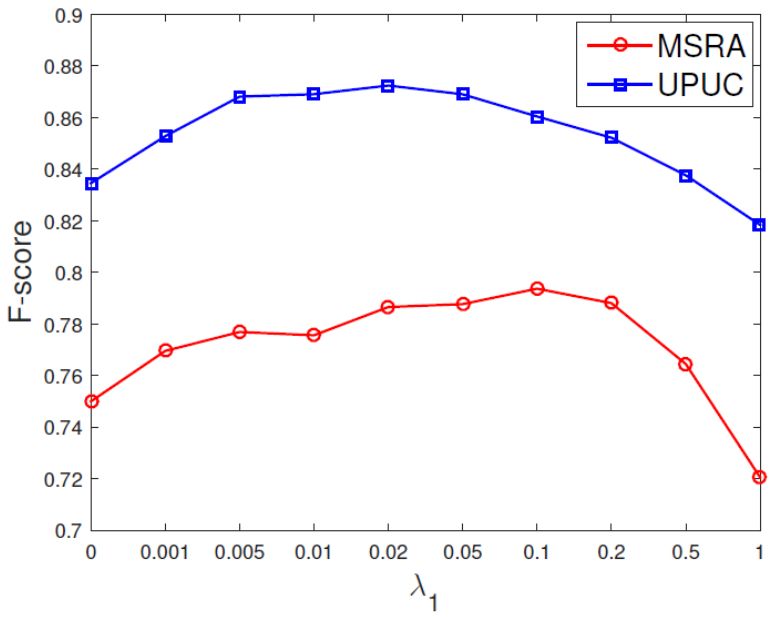


Influence of dictionary size

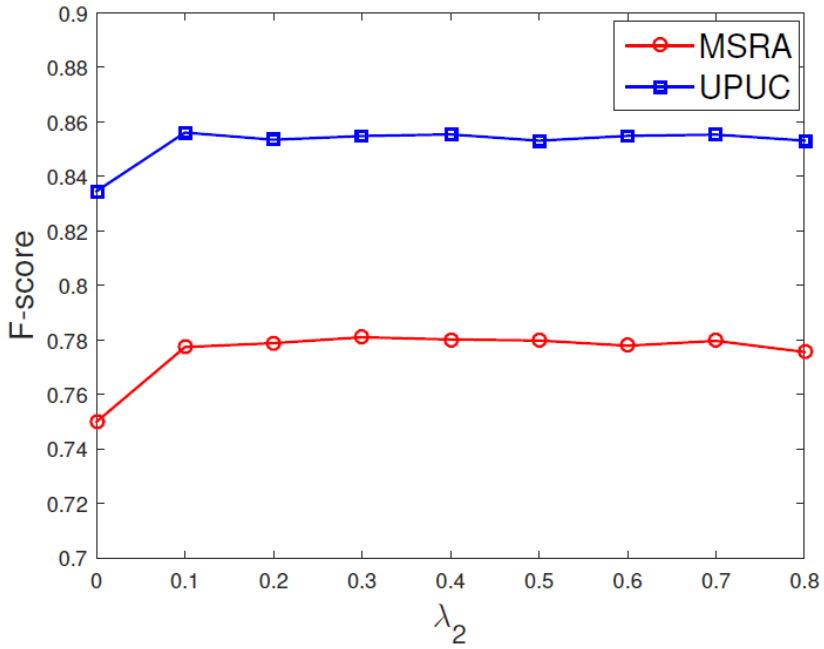
Internal dict: words from training data
External dict: Sogou Chinese dictionary

Experiments

- Influence of Parameters



Influence of λ_1



Influence of λ_2

Experiments

- Case Study

	Example 1	Example 2
Original	5 名男子和被害人有恩怨	警方一口气带回了５０多人
CNN-CRF	5 /名/男子/和/被/害/人/有/恩怨	警方/一/口/气/带回/了/５０多/人
+Internal dictionary	5 /名/男子/和/被/害/人/有/恩怨	警方/一口气/带回/了/５０多/人
+External dictionary	5 /名/男子/和/被害人/有/恩怨	警方/一口气/带回/了/５０多/人

“被害人” is an OOV word, and “一口气” is a rare word in training data.

Conclusion and Future Work

- Dictionary is useful for neural CWS
  - Improve the performance on OOV and rare words
  - Reduce the dependence on labeled data
- Proposed two simple but effective methods to incorporate dictionary knowledge into neural CWS
  - Dictionary as pseudo labeled data
  - Dictionary as word-level supervision and multi-task learning
- Future work:
  - Exploiting both dictionary and unlabeled data for CWS
  - Joint new word detection and Chinese word segmentation

Thank you