Improving Chinese Named Entity Recognition with Multi-grained Words and Part-of-Speech Tags via Joint Modeling

Chenhui Dou[1], Chen Gong[1*], Zhenghua Li[1], Zhefeng Wang[2], Baoxing Huai[2], Min Zhang[1]

[1]Institute of Artificial Intelligence, School of Computer Science and Technology, Soochow University, China, [2]Huawei Cloud, China
[1]20215227026@stu.suda.edu.cn [1]{gongchen18, zhli13, minzhang}@suda.edu.cn
[2]{wangzhefeng, huaibaoxing}@huawei.com

Task Definition

CNER (Chinese Named Entity Recognition):

Recognizing the entities with specific meanings in Chinese text.

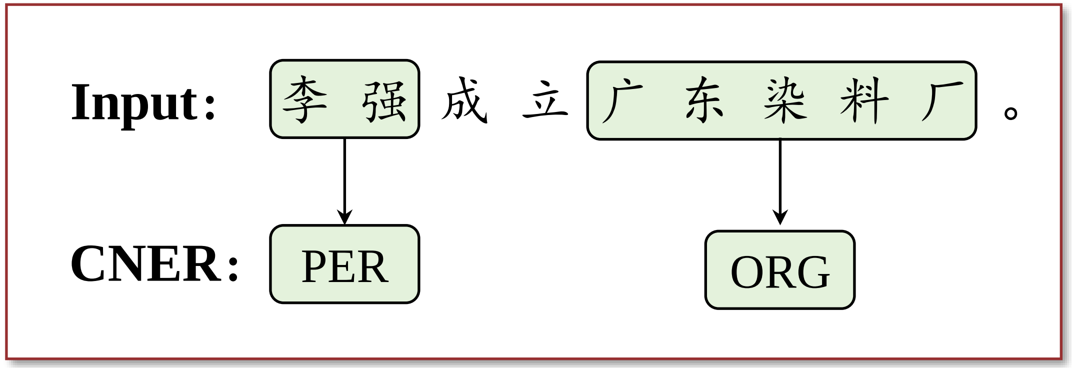

Figure 1: An example sentence with its CNER result: “李强(Li Qiang) 成立(sets up)广东(Kwangtung)染料(Dyestuff)厂(Plant).”

Motivation

- In Chinese, word information plays a very important role in NER. However, the integration of CNER and word information through previous methods is indirect and shallow.

- Existing methods usually only consider single-grained word segmentation.

Contributions

- Unified MWS-POS-NER representation and data
- Jointly modeling MWS-POS-NER with a two-stage parsing
- Extensive experiments and in-depth analysis

Data Representation

Representing MWS, POS, and NER in a unified manner by constructing the MWS-POS-NER tree structure.

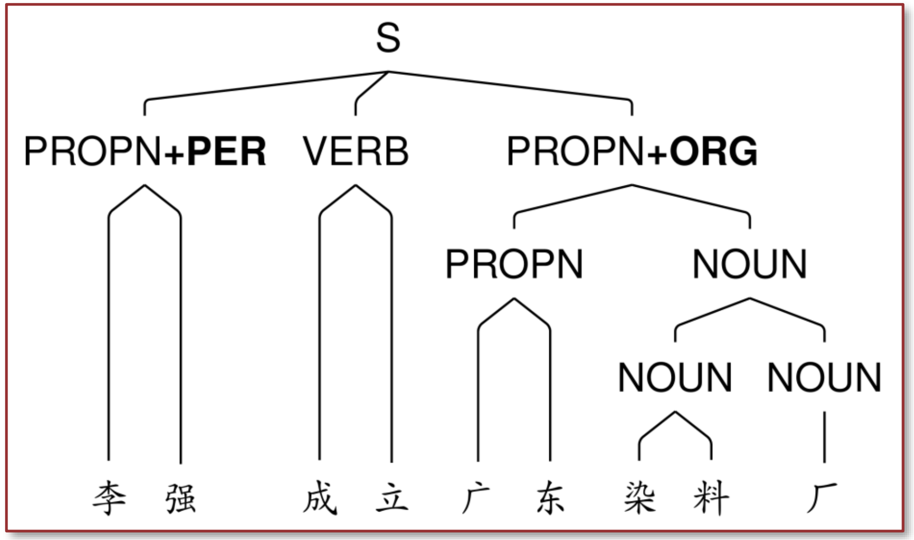

Figure 2: An example sentence of its Chinese MWS-POS-NER tree.

Data Construction

Step 1: Generating MWS tree with POS tags.

I. Training two conversion models.

II. Converting to PPD-side and MSR-side WS&POS results.

III. Representing the three different WS&POS results in the MWS-POS tree.

Step2: Attaching NE labels to MWS-POS tree.

- Attaching an extra NE label to its corresponding word non-terminal.
- Adding a new non-terminal node for the corresponding entity.

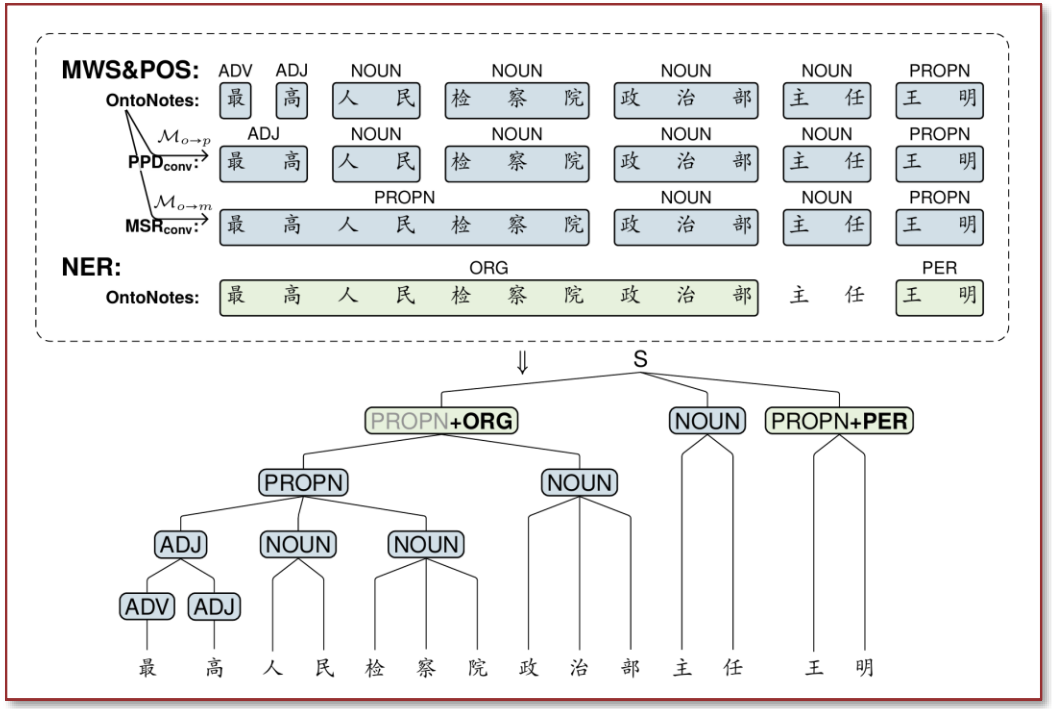

Figure 3: An example of how a MWS-POS-NER tree is generated.

Two-stage Parsing Framework

First-stage: Predicting MWS tree with POS tags.

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{(i,j,t)\in \boldsymbol{y}} s(i, j, t)$$

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} s(\boldsymbol{x}, \boldsymbol{y})$$

Second-stage: Recognizing named entities.

$$\hat{l} = \arg\max_{l\in\mathcal{N}} s(i, j, l)$$

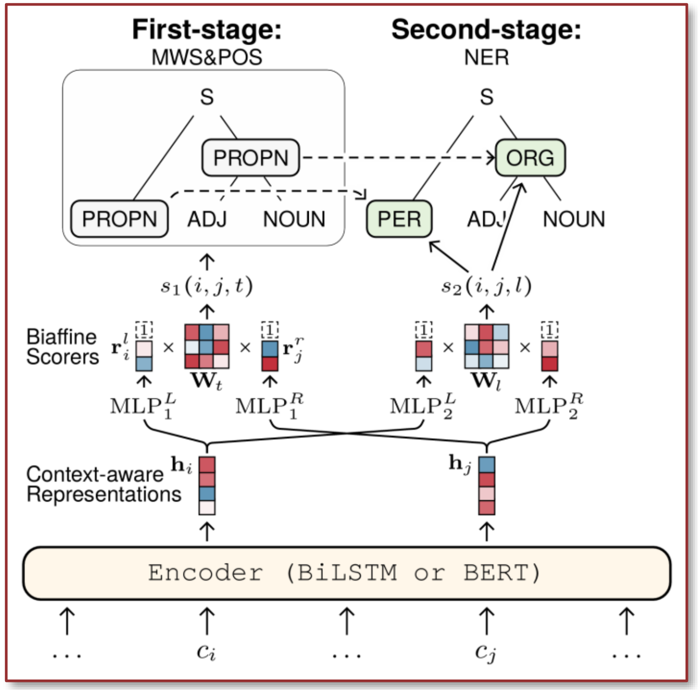

Figure 4: The architecture of the two-stage joint parsing framework.

Model Architecture

Inputs: Character embedding

$$\mathbf{e}_i = \mathbf{emb}(c_i)$$

Encoder: Three layers BiLSTM or BERT

Boundary representation: Two separate MLPs

$$\mathbf{r}_i^L ; \mathbf{r}_i^R = \mathbf{MLP}^L(\mathbf{h}_i) ; \mathbf{MLP}^R(\mathbf{h}_i)$$

Biaffine Scorer:

$$s(i,j,t) = \begin{bmatrix} \mathbf{r}_i^L \\ 1 \end{bmatrix}^{\mathrm{T}} \mathbf{W}_t \begin{bmatrix} \mathbf{r}_j^R \\ 1 \end{bmatrix}$$

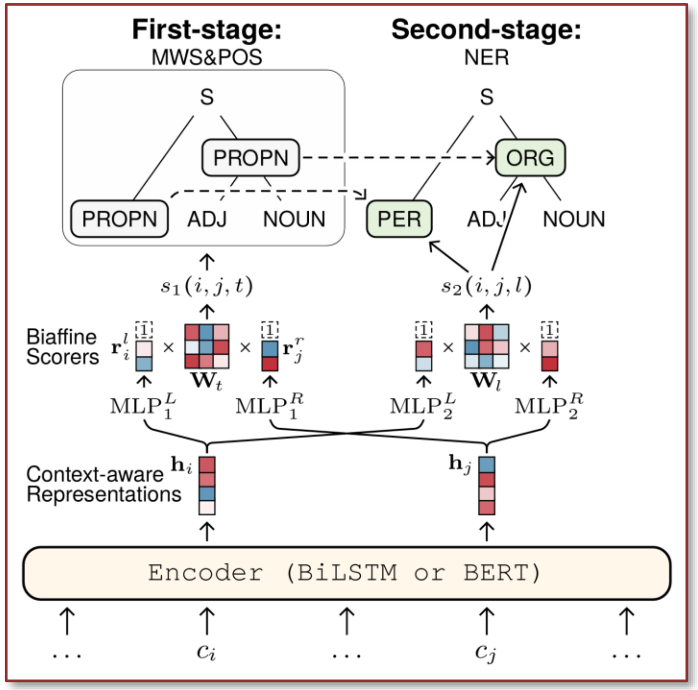

Figure 4: The architecture of the two-stage joint parsing framework.

Training Loss

First-stage: TreeCRF Loss

$$\mathcal{L}^{1st}(\boldsymbol{x}, \boldsymbol{y}^*) = -\log p(\boldsymbol{y}^* | \boldsymbol{x})$$

$$p(\boldsymbol{y}^* | \boldsymbol{x}) = \frac{e^{s(\boldsymbol{x}, \boldsymbol{y}^*)}}{\boldsymbol{Z}(\boldsymbol{x}) \equiv \sum_{\boldsymbol{y}' \in \mathcal{T}(\boldsymbol{x})} e^{s(\boldsymbol{x}, \boldsymbol{y}')}}$$

Second-stage: Cross Entropy Loss

$$\mathcal{L}^{2nd}(\boldsymbol{x}, \boldsymbol{z}^*) = \sum_{(i,j,l) \in \boldsymbol{z}^*} -\log \frac{e^{s(i,j,l)}}{\sum_{l'} e^{s(i,j,l')}}$$

Overall training loss

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*, \boldsymbol{z}^*) = \mathcal{L}^{1st}(\boldsymbol{x}, \boldsymbol{y}^*) + \mathcal{L}^{2nd}(\boldsymbol{x}, \boldsymbol{z}^*)$$

Datasets

Ontonotes4 & Ontonotes5

Datasets	Type	Train	Dev	Test
OntoNotes4	#Sent.	15,724	4,301	4,346
	#Entity	13,372	6,950	7,684
OntoNotes5	#Sent.	36,487	6,083	4,472
	#Entity	62,543	9,104	7,494

Table 1: Numbers of sentences and entities in OntoNotes4 and OntoNotes5 datasets.

Experimental Results

Model	OntoNotes4-Dev			OntoNotes5-Dev			Sent/s
	P	R	F1	P	R	F1	
Char-based	72.77	64.64	$68.42_{\pm 0.25}$	72.90	69.53	$71.17_{\pm 0.11}$	393
Joint model	75.86	66.21	$\mathbf{70.70}_{\pm 0.29}$	77.50	71.22	$\mathbf{74.22}_{\pm 0.03}$	349
Char-based w/ lexicon	74.63	72.72	$73.65_{\pm 0.19}$	74.25	74.39	$74.32_{\pm 0.20}$	136
Joint model w/ lexicon	76.07	72.37	$\mathbf{74.18}_{\pm 0.12}$	78.83	73.59	$\mathbf{76.12}_{\pm 0.12}$	131
Char-based w/ BERT	78.96	80.18	$79.55_{\pm 0.11}$	75.86	78.19	$77.01_{\pm 0.07}$	204
Joint model w/ BERT	80.39	80.44	$\mathbf{80.41}_{\pm 0.21}$	78.83	77.41	$\mathbf{78.09}_{\pm 0.16}$	179

Table 2: Development results on OntoNotes4 and OntoNotes5 datasets.

After introducing lexicon information or BERT encoder, the performance of the Joint model is superior to the 'Char-based' Baseline on both datasets.

Experimental Results

Model	F1
OntoNotes4	
Lattice LSTM (Zhang and Yang, 2018)	73.88
LR-CNN (Gui et al., 2019)	74.45
WC-LSTM (Liu et al., 2019)	74.43
PLTE† (Xue et al., 2020)	80.60
FLAT† (Li et al., 2020)	81.82
SoftLexicon† (Ma et al., 2020)	82.81
LEBERT† (Liu et al., 2021)	82.08
MECT† (Wu et al., 2021)	82.57
ATSSA† (Hu et al., 2022a)	83.31
ACT-S† (Ning et al., 2022)	**83.91**
W²NER† (Li et al., 2022)	83.08
Joint model†	82.82
OntoNotes5	
WC-LSTM (Liu et al., 2019)	75.95
DGLSTM-CRF (Jie and Lu, 2019)	77.40
FLAT† (Li et al., 2020)	77.87
SoftLexicon† (Ma et al., 2020)	79.71
LEBERT† (Liu et al., 2021)	78.30
W²NER† (Li et al., 2022)	79.04
Joint model†	**79.87**

Table 3: Comparison with previous works.

- Joint model achieves comparable performance with other latest models on OntoNotes4.
- Joint model achieves state-of-the-art results on OntoNotes5.

Experimental Results

Model	OntoNotes4	OntoNotes5
NER as sequence labeling		
Char-based	$81.70_{\pm 0.28}$	$78.30_{\pm 0.16}$
Word-based (orig.)	$79.28_{\pm 0.17}$	$78.14_{\pm 0.11}$
Joint NER w/ WS as tree parsing		
+SWS (orig.)	$81.82_{\pm 0.17}$	$79.34_{\pm 0.39}$
+SWS (fine)	$81.96_{\pm 0.32}$	$79.29_{\pm 0.29}$
+SWS (coarse)	$82.04_{\pm 0.23}$	$79.50_{\pm 0.05}$
+MWS	$\mathbf{82.11}_{\pm 0.16}$	$\mathbf{79.58}_{\pm 0.20}$
Joint NER w/ WS&POS as tree parsing		
+SWS (orig.)&POS	$82.20_{\pm 0.05}$	$79.69_{\pm 0.14}$
+SWS (fine)&POS	$81.97_{\pm 0.19}$	$79.64_{\pm 0.25}$
+SWS (coarse)&POS	$82.43_{\pm 0.24}$	$79.84_{\pm 0.41}$
+MWS&POS	$\mathbf{82.82}_{\pm 0.07}$	$\mathbf{79.87}_{\pm 0.20}$
w/o PROPN constraint	$82.55_{\pm 0.06}$	$79.82_{\pm 0.12}$
merge POS&NE label	$81.91_{\pm 0.58}$	$79.52_{\pm 0.29}$

Table 4: Ablation studies on models with BERT.

- Joint framework is better than pipeline framework.
- MWS is better than SWS. Coarse SWS is the best among SWS.
- POS is further helpful for CNER.
- PROPN constraint and distinguishing label space are effective.

Experimental Results

Model	OntoNotes4	OntoNotes5
NER as sequence labeling		
Char-based	$81.70_{\pm 0.28}$	$78.30_{\pm 0.16}$
Word-based (orig.)	$79.28_{\pm 0.17}$	$78.14_{\pm 0.11}$
Joint NER w/ WS as tree parsing		
+SWS (orig.)	$81.82_{\pm 0.17}$	$79.34_{\pm 0.39}$
+SWS (fine)	$81.96_{\pm 0.32}$	$79.29_{\pm 0.29}$
+SWS (coarse)	$82.04_{\pm 0.23}$	$79.50_{\pm 0.05}$
+MWS	$\mathbf{82.11}_{\pm 0.16}$	$\mathbf{79.58}_{\pm 0.20}$
Joint NER w/ WS&POS as tree parsing		
+SWS (orig.)&POS	$82.20_{\pm 0.05}$	$79.69_{\pm 0.14}$
+SWS (fine)&POS	$81.97_{\pm 0.19}$	$79.64_{\pm 0.25}$
+SWS (coarse)&POS	$82.43_{\pm 0.24}$	$79.84_{\pm 0.41}$
+MWS&POS	$\mathbf{82.82}_{\pm 0.07}$	$\mathbf{79.87}_{\pm 0.20}$
w/o PROPN constraint	$82.55_{\pm 0.06}$	$79.82_{\pm 0.12}$
merge POS&NE label	$81.91_{\pm 0.58}$	$79.52_{\pm 0.29}$

Table 4: Ablation studies on models with BERT.

- Joint framework is better than pipeline framework.
- MWS is better than SWS. Coarse SWS is the best among SWS.
- POS is further helpful for CNER.
- PROPN constraint and distinguishing label space are effective.

Experimental Results

Model	OntoNotes4	OntoNotes5
NER as sequence labeling		
Char-based	$81.70_{\pm 0.28}$	$78.30_{\pm 0.16}$
Word-based (orig.)	$79.28_{\pm 0.17}$	$78.14_{\pm 0.11}$
Joint NER w/ WS as tree parsing		
+SWS (orig.)	$81.82_{\pm 0.17}$	$79.34_{\pm 0.39}$
+SWS (fine)	$81.96_{\pm 0.32}$	$79.29_{\pm 0.29}$
+SWS (coarse)	$82.04_{\pm 0.23}$	$79.50_{\pm 0.05}$
+MWS	$\mathbf{82.11}_{\pm 0.16}$	$\mathbf{79.58}_{\pm 0.20}$
Joint NER w/ WS&POS as tree parsing		
+SWS (orig.)&POS	$82.20_{\pm 0.05}$	$79.69_{\pm 0.14}$
+SWS (fine)&POS	$81.97_{\pm 0.19}$	$79.64_{\pm 0.25}$
+SWS (coarse)&POS	$82.43_{\pm 0.24}$	$79.84_{\pm 0.41}$
+MWS&POS	$\mathbf{82.82}_{\pm 0.07}$	$\mathbf{79.87}_{\pm 0.20}$
w/o PROPN constraint	$82.55_{\pm 0.06}$	$79.82_{\pm 0.12}$
merge POS&NE label	$81.91_{\pm 0.58}$	$79.52_{\pm 0.29}$

Table 4: Ablation studies on models with BERT.

- Joint framework is better than pipeline framework.
- MWS is better than SWS. Coarse SWS is the best among SWS.
- POS is further helpful for CNER.
- PROPN constraint and distinguishing label space are effective.

The End

Thanks for your time!

Questions?