Hybrid semi-Markov CRF for Neural Sequence Labeling

Zhi-Xiu Ye, Zhen-Hua Ling
University of Science and Technology of China

Background

Sequence labeling is a type of pattern recognition task that involves the algorithmic assignment of a categorical label to each member of a sequence of observed values.
Take named entity recognition as an example:
sentence:
Barack Obama was born in Hawaii.

CRF-style(word-level) label:
B-PER I-PER O O O B-LOC

HSCRF-style(segment-level) label:
(1,2,PER) (3,3,O) (4,4,O) (5,5,O) (6,6,LOC)

Contributions

★ Propose the **Hybrid semi-Markov CRF** (HSCRF) architecture which employs both word-level and segment-level labels for segment score calculation.

★ Propose a **joint CRF-HSCRF training framework** and a naive joint decoding algorithm for neural sequence labeling.

★ The proposed model achieves **state-of-the-art** performance in CoNLL 2003 NER shared task **without** external knowledge.

Source code available!!!

`https://github.com/ZhixiuYe/HSCRF-pytorch`

Our implementation is based on python and the **PyTorch** library.

A comparison between CRFs and HSCRFs

1. Input data

a Input sentence $\mathbf{x} = \{x_1, ..., x_n\}$
b Word-level label: $\mathbf{y} = \{y_1, ..., y_n\}$
c Segment-level label: $\mathbf{s} = \{s_1, s_2, ..., s_p\}$

a, b for CRFs and **a, b, c** for HSCRFs.

2. Word-level representations

CRFs and HSCRFs share the **same** word representations

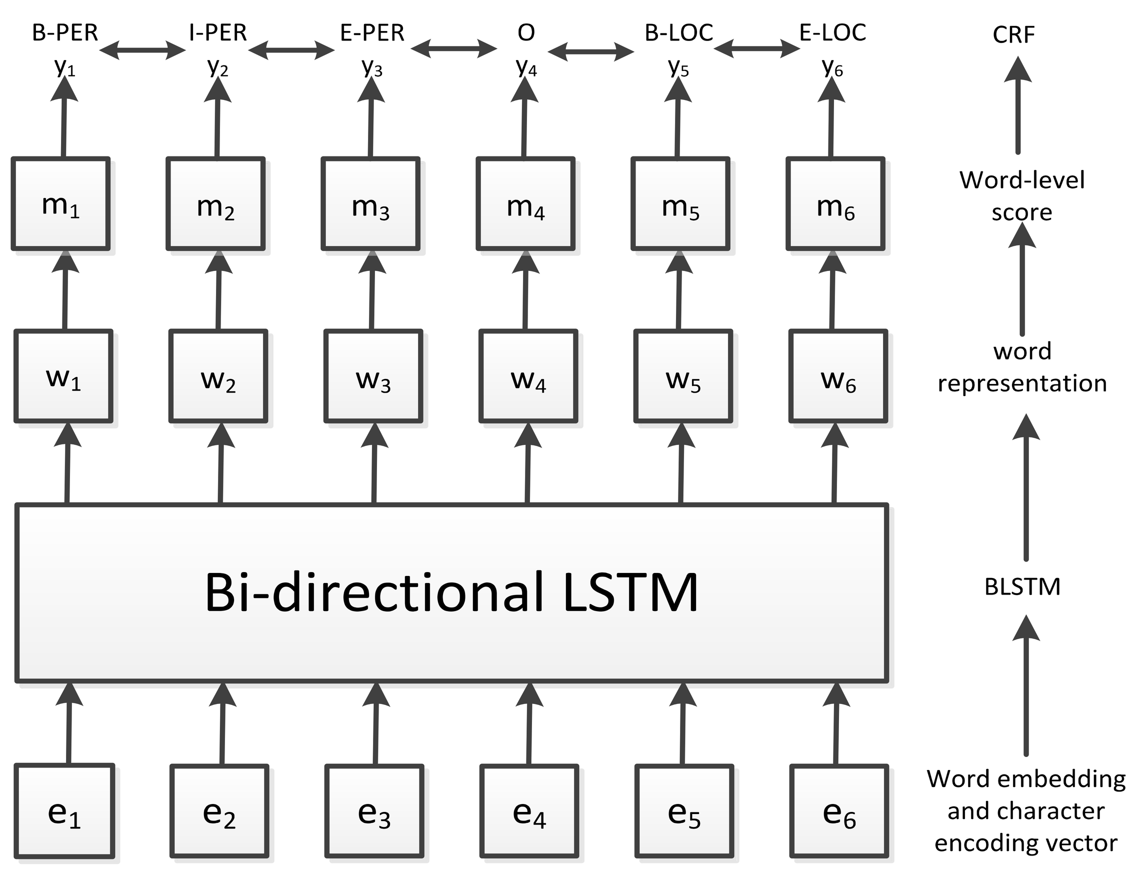
$$\boldsymbol{w}_i = \mathrm{BLSTM}(\boldsymbol{e}_i),$$

where $\boldsymbol{e}_i$ is the word embedding of x_i.

3. Score computation

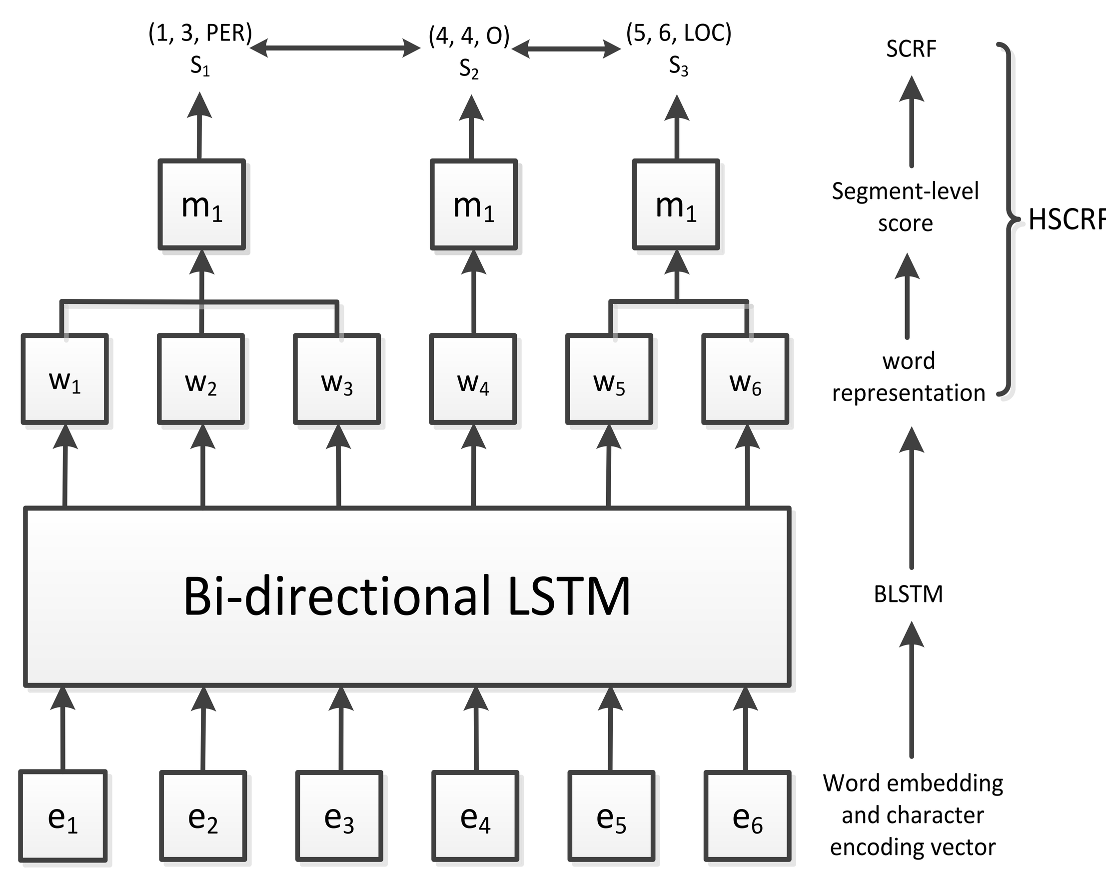
- In **CRFs**, we compute the score of **word**-level label $\boldsymbol{m}_i$ via the representation of i-th word $\boldsymbol{w}_i$.
- In **HSCRFs**, the score of **segment**-level label $\boldsymbol{m}_i$ is computed by the summation of the scores of the **word**-level label.

Figure 1: CRFs with neural networks



$$m_{CRF_i} = \varphi_c(y_k, \boldsymbol{w}_k) = \boldsymbol{a}_{y_k}^{\top}\boldsymbol{w}_k$$

Figure 2: HSCRFs with neural networks



$$m_{HSCRF_i} = \sum_{k=\mathbf{b}_i}^{\mathbf{e}_i} \varphi_c(y_k, \boldsymbol{w}_k) = \sum_{k=b_i}^{e_i} \boldsymbol{a}_{y_k}^{\top}\boldsymbol{w}_k$$

Joint training and decoding

1. Training

- A CRF output layer and a HSCRF output layer are **integrated** into an unified neural network.
- The model parameters are **shared** and optimized by minimizing the **summation** of the loss functions of the CRF layer and the HSCRF layer with equal weights as follows:

$$loss = loss_{CRF} + loss_{HSCRF}$$

2. Decoding

- **Two** label sequences, $\mathbf{s}_c$ and $\mathbf{s}_h$, for an input sentence can be obtained using the CRF output layer and the HSCRF output layer respectively.
- Choose the one between $\mathbf{s}_c$ and $\mathbf{s}_h$ with **lower** *loss* as the final result.

Experiments

Dataset: CoNLL 2003 shared task: English named entity recognition.

Table 1: Model performance (F1 score) on CoNLL 2003 NER task for entities with different lengths, where LM for language model[1], GSCRF for grSemi-CRF[2], JNT for our proposed joint model.

Model	Entity Length						
	1	2	3	4	5	≥ 6	all
LM-BLSTM-CRF	91.68	91.88	82.64	75.81	73.68	72.73	91.17
LM-BLSTM-GSCRF	91.57	91.68	83.61	74.32	76.64	73.64	91.06
LM-BLSTM-HSCRF	91.65	91.84	82.97	76.20	78.95	74.55	91.27
LM-BLSTM-JNT(JNT)	**91.73**	**92.03**	**83.78**	**77.27**	**79.66**	**76.55**	**91.38**

Table 2: Comparison with existing works

Model	Test Set F1 Score	
	Type	Value ($\pm$std)
Zhuo et al.(2016)	reported	88.12
Lample et al.(2016)	reported	90.94
Ma and Hovy(2016)	reported	91.21
Rei(2017)	reported	86.26
Liu et al.(2018)	mean	91.24 $\pm$ 0.12
	max	91.35
CNN-BLSTM-JNT(JNT)	mean	91.26 $\pm$ 0.10
	max	91.41
LM-BLSTM-JNT(JNT)	mean	**91.38$\pm$ 0.10**
	max	**91.53**

- **Word**-level labels may supervise models to learn word-level descriptions which tend to benefit the recognition of **short** entities.
- **Segment**-level labels may guide models to capture the descriptions of combining words for whole entities which help to recognize **long** entities.
- By utilizing **both** labels, the proposed joint model can achieve better overall performance of recognizing entities with different lengths.

[1] Empower Sequence Labeling with Task-Aware Neural Language Model.
[2] Segment-level sequence modeling using gated recursive semi-Markov conditional random fields.