Does Higher Order LSTM Have Better Accuracy for Segmenting and Labeling Sequence Data?

Yi Zhang, Xu Sun, Shuming Ma, Yang Yang, Xuancheng Ren

MOE Key Laboratory of Computational Linguistics,

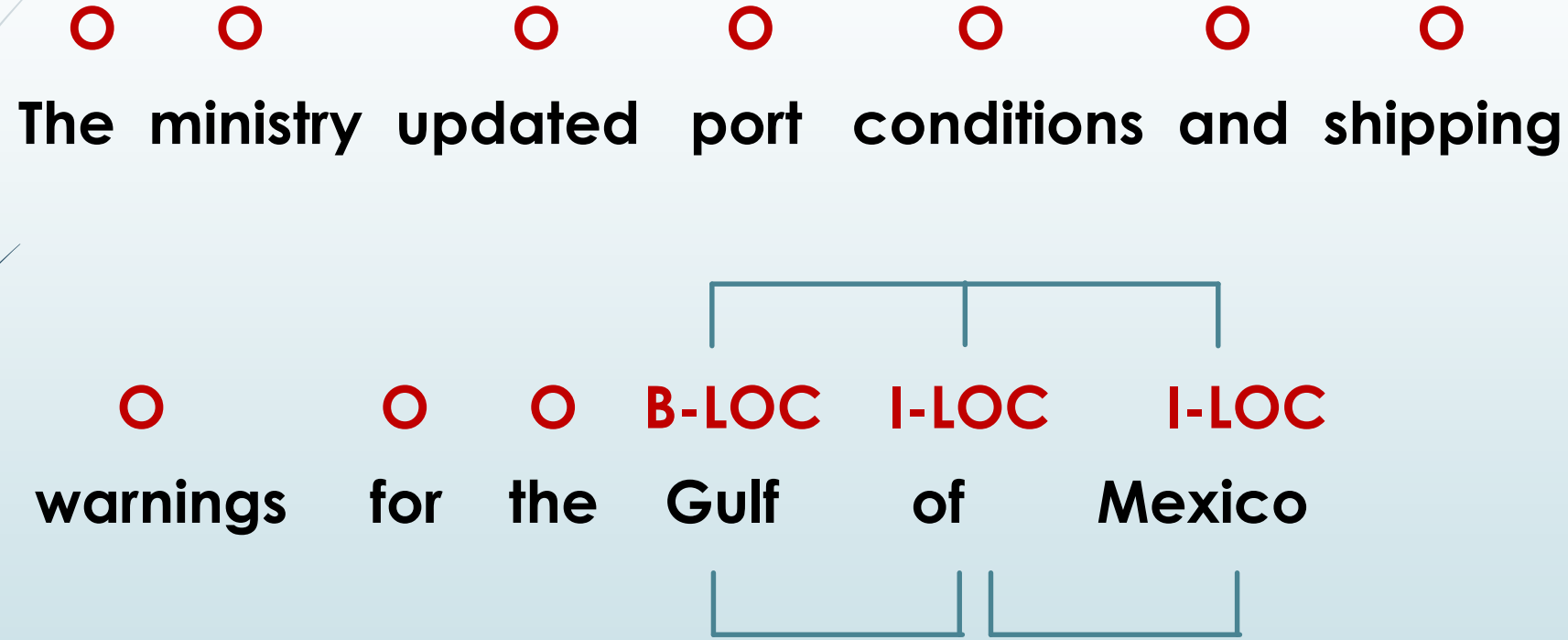
EECS, Peking University

Overview

- Motivation
- Models
- Experiments
- Conclusions

Overview

- Motivation
- Models
- Experiments
- Conclusions


O O O O O O O
The ministry updated port conditions and shipping
O O O B-LOC I-LOC I-LOC
warnings for the Gulf of Mexico

Current Methods

- **LSTM**

  makes prediction on single point

- **LSTM-CRF**

  usually considers two adjacent labels

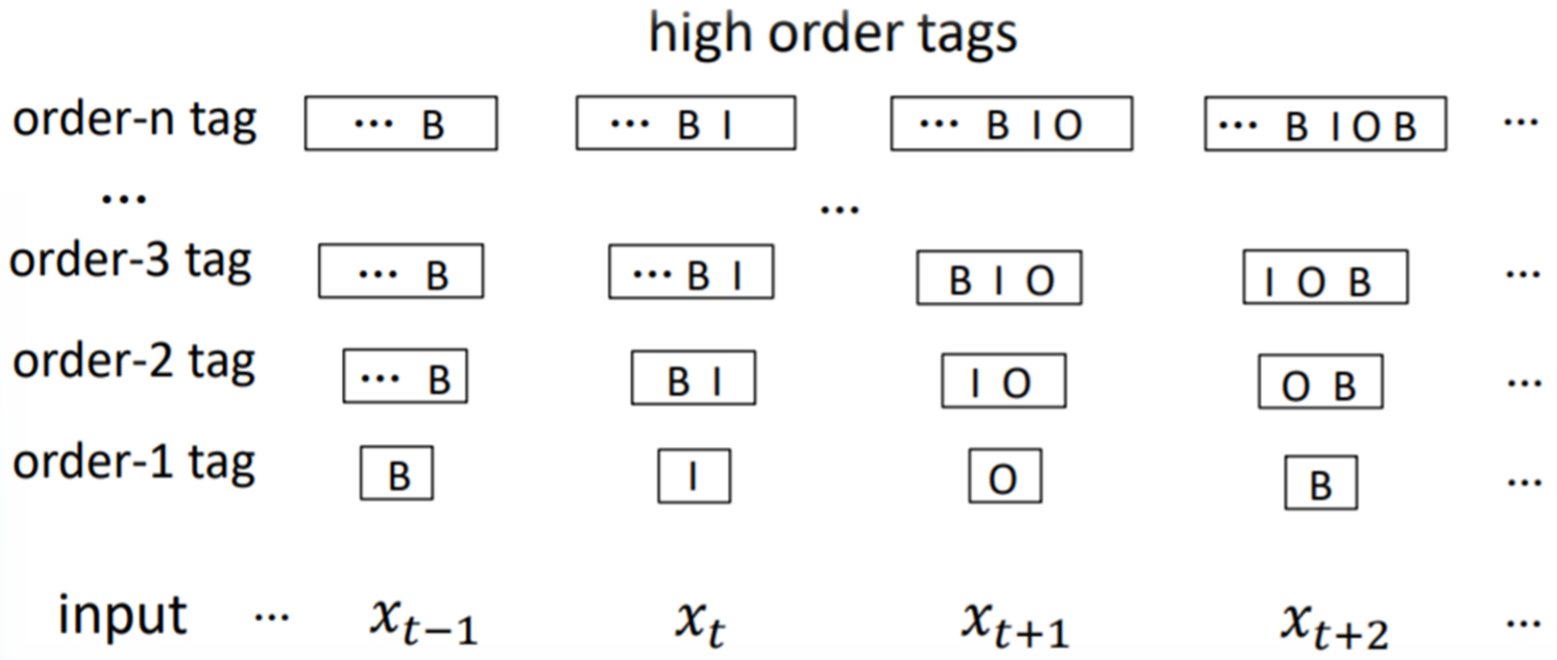
ignored
long tag dependencies

Overview

- Motivation
- Models
- Experiments
- Conclusions

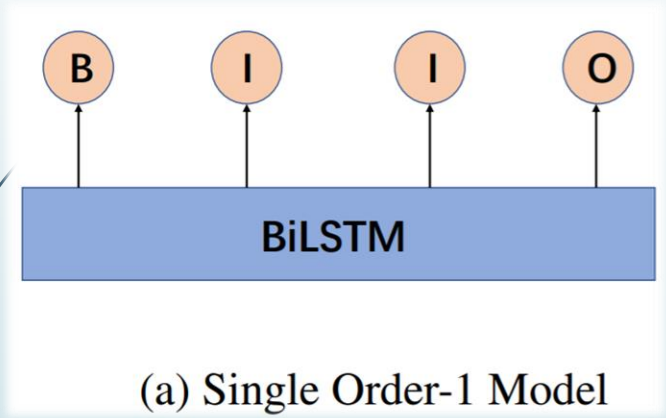
Single Order LSTM

- **High order tags**



Single Order LSTM

Order-1 Model (LSTM)



(a) Single Order-1 Model

$$s_1(y_1, y_2, \cdots, y_T|\boldsymbol{x};\theta) = \prod_{t=1}^{T} s(y_t|\boldsymbol{x};\theta)$$

For training:

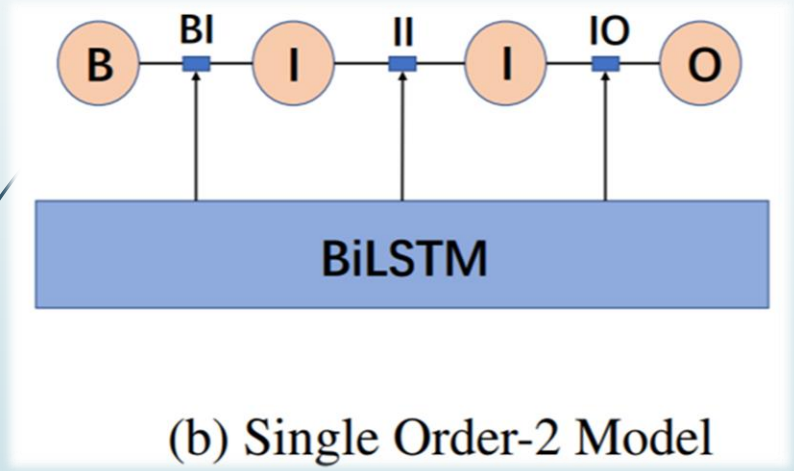
train with original tag set {B, I , O}

For testing:

choose the tag with max probability at each time step

Single Order LSTM

Order-2 Model



(b) Single Order-2 Model

$$s_2(y_1, y_2, \cdots, y_T|\boldsymbol{x};\theta) = \prod_{t=1}^{T} s(y_{t-1}y_t|\boldsymbol{x};\theta)$$

For training:

train with order-2 tag set {BB, BI , BO, …, OI, OO}

For testing:

1. choose the 2-order tag with max probability at each time step.

2. take the second tag of the 2-order tag. (BO ⟶ O)

Single Order LSTM

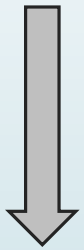
Order-n Model

The model can be further extended to order-n:

$$s_n(y_1, y_2, \cdots, y_T|\boldsymbol{x};\theta) = \prod_{t=1}^{T} s(y_{t-n+1} \cdots y_t|\boldsymbol{x};\theta)$$

Single Order LSTM

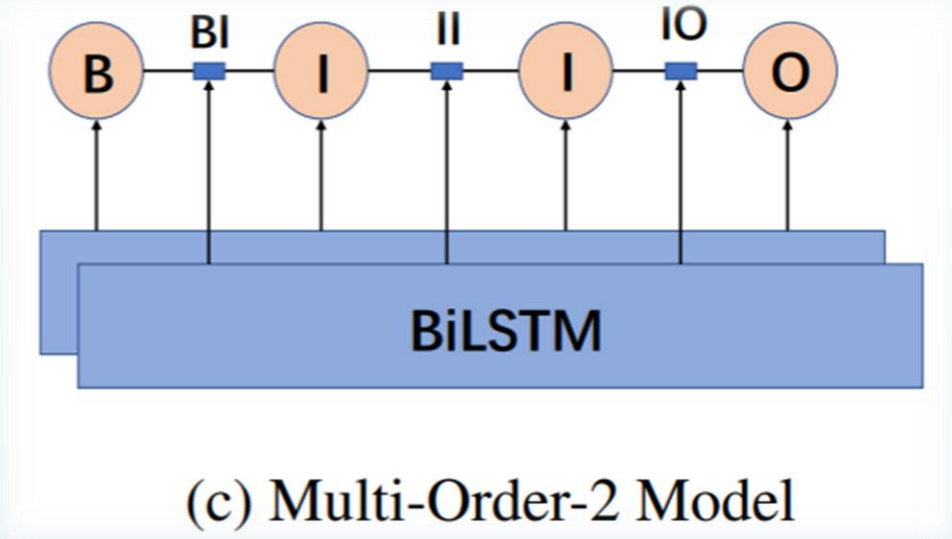
Single order LSTM



order ↑ **performance** ↓

- Try to capture long tag dependencies
- Training and testing are similar to LSTM
- Easy to implement
- Large tag set arises the difficulty of prediction
- Complex structure may lead to overfitting

Multi-Order BiLSTM



(c) Multi-Order-2 Model

$$y_1^*, y_2^*, \cdots, y_T^* = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, s_1(y_1, y_2, \cdots, y_T|\boldsymbol{x}; \theta_1) \times s_2(y_1, y_2, \cdots, y_T|\boldsymbol{x}; \theta_2)$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{t=1}^{T} s(y_t|\boldsymbol{x}; \theta_1) \times s(y_{t-1}, y_t|\boldsymbol{x}; \theta_2)$$

order-1 score order-2 score

Multi-Order BiLSTM

Extend to order-n case

$$y_1^*, y_2^*, \cdots, y_T^* = \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{i=1}^{k} s_{o_i}(y_1, y_2, \cdots, y_T | \boldsymbol{x}; \theta_i)$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{i=1}^{k} \prod_{t=1}^{T} s(y_{t-o_i+1} \cdots y_t | \boldsymbol{x}; \theta_i)$$

Multi-Order BiLSTM

- For training:

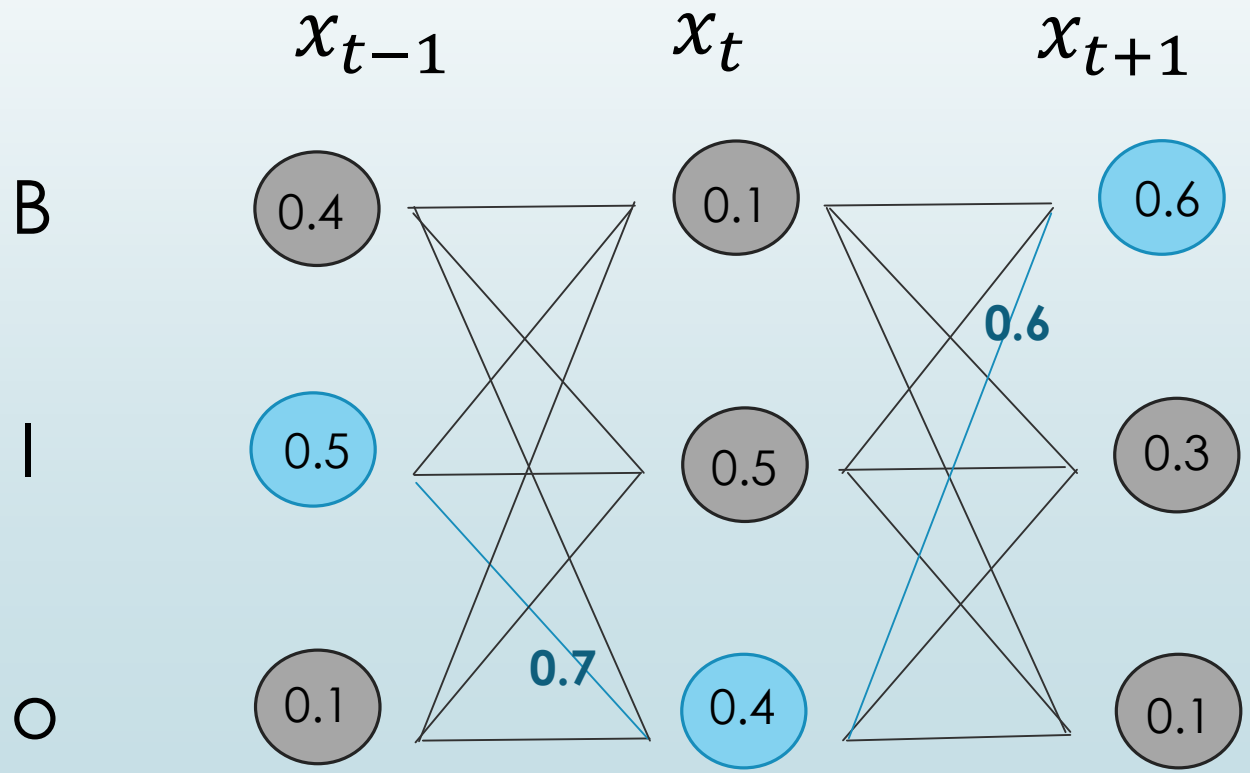
  we train *k* single order models separately

- For decoding

  we use a dynamic programming algorithm to search for the label sequence with the maximum score

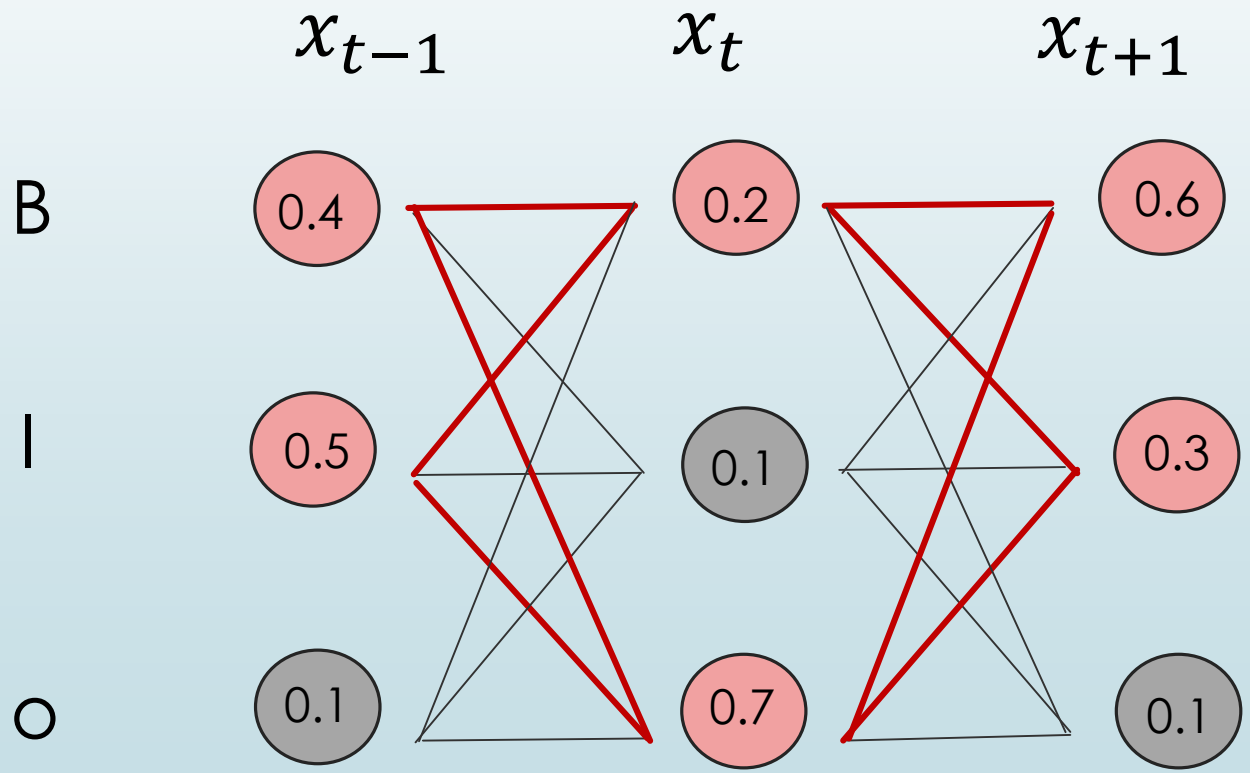
Scalable Decoding with Pruning

- original dynamic programing (order-2)



Scalable Decoding with Pruning

- with pruning (order-2)



Overview

- Motivation
- Models
- Experiments
- Conclusions

Effect of Multi-Order Setting

Model	All-Chunking	English-NER	Dutch-NER
Order-1	14	10	11
Order-2	154	39	44
Order-3	832	138	158

Effect of Pruning

Model	All-Chunking		English-NER		Dutch-NER	
	Time (s)	F1	Time (s)	F1	Time (s)	F1
Multi-Order-2 BiLSTM w/o pruning	31.59	94.93	19.23	90.23	26.60	80.95
Multi-Order-2 BiLSTM	13.64	94.93	13.13	90.23	18.42	80.95
Multi-Order-3 BiLSTM w/o pruning	215.21	95.01	51.78	90.70	69.79	81.76
Multi-Order-3 BiLSTM	44.81	95.01	20.43	90.70	28.66	81.76

Effect of Multi-Order BiLSTM

Model	All-Chunking	English-NER	Dutch-NER
Single Order-1 BiLSTM	93.89	88.23	77.20
Single Order-2 BiLSTM	93.71 (-0.18)	87.61 (-0.62)	76.61 (-0.59)
Single Order-3 BiLSTM	93.34 (-0.55)	87.47 (-0.76)	76.47 (-0.73)
Multi-Order-1 BiLSTM	93.89	88.23	77.20
Multi-Order-2 BiLSTM	94.93 (+1.04)	90.23 (+2.00)	80.95 (+3.75)
Multi-Order-3 BiLSTM	**95.01 (+1.12)**	**90.70 (+2.47)**	**81.76 (+4.56)**

Effect of Multi-Order BiLSTM

GOLD	The ministry updated port conditions and shipping warnings for the Gulf of Mexico (LOC), Caribbean and Pacific Coast
BiLSTM	The ministry updated port conditions and shipping warnings for the Gulf (LOC) of Mexico(LOC) , Caribbean and Pacific Coast
MO-BiLSTM	The ministry updated port conditions and shipping warnings for the Gulf of Mexico (LOC), Caribbean and Pacific Coast.
GOLD	About 200 Burmese students marched briefly from troubled Yangon Institute of Technology (ORG) in northern Rangoon on Friday.
BiLSTM	About 200 Burmese students marched briefly from troubled Yangon (LOC) Institute of Technology (ORG) in northern Rangoon on Friday.
MO-BiLSTM	About 200 Burmese students marched briefly from troubled Yangon Institute of Technology (ORG) in northern Rangoon on Friday.

Error Analysis

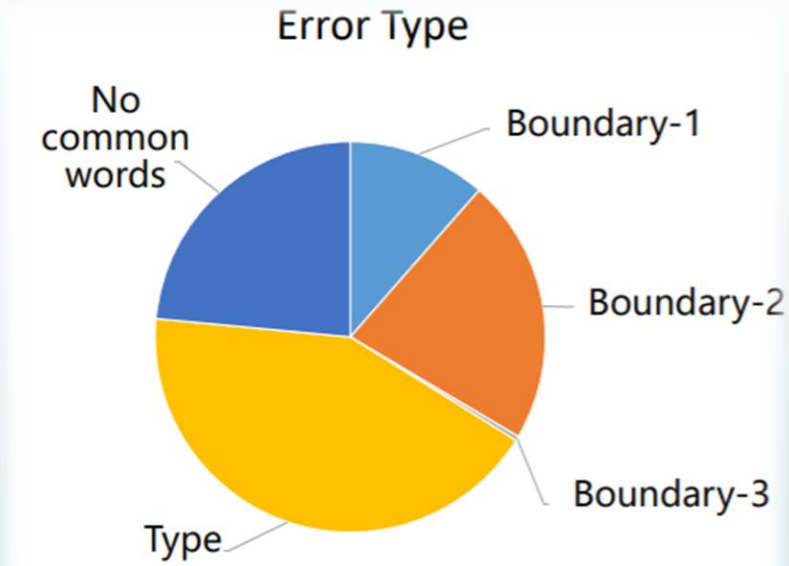
Boundary-1: the gold entity contains a predicted entity

Boundary-2: the gold entity is contained by a prediction

Boundary-3: the gold entity and prediction overlap

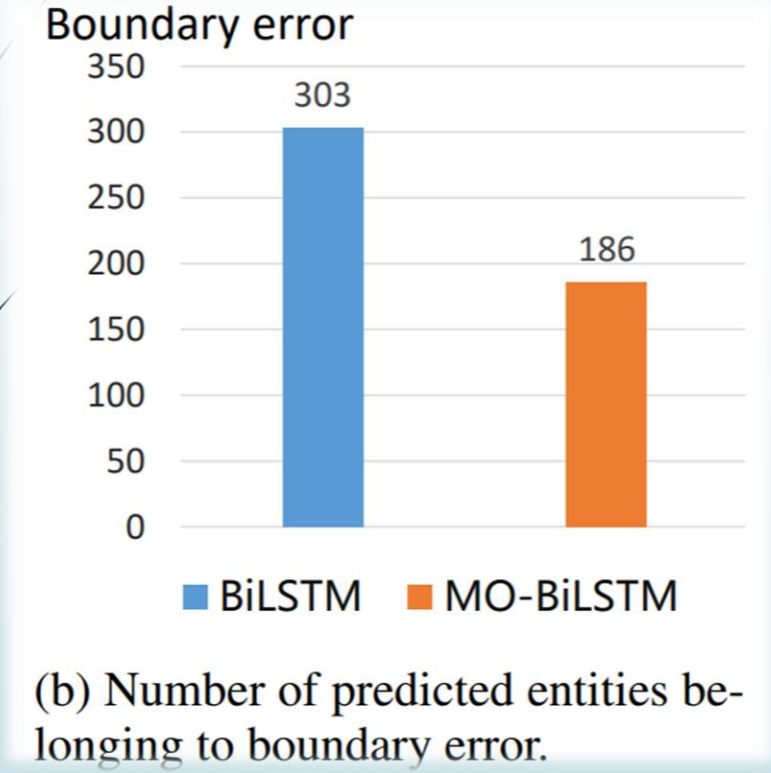
Type: correct boundaries and misclassified entity type

no common words: no common words between the predicted entity and any gold entity



(a) Error types of the predicted entities of MO-BiLSTM.

Error Analysis

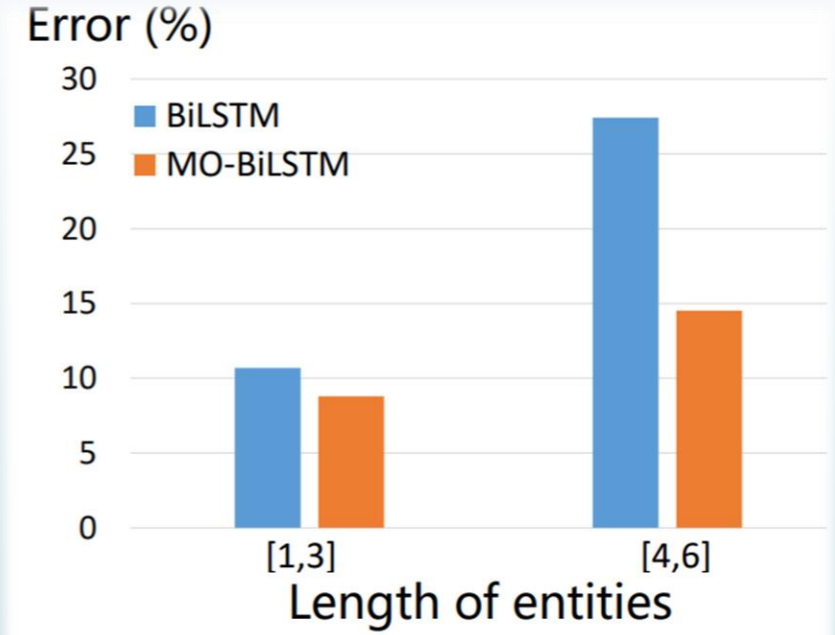


(b) Number of predicted entities belonging to boundary error.

the "boundary" error of MO-BiLSTM has a reduction rate of nearly 40% compared with BiLSTM

Error Analysis

the MO-BiLSTM model has a significant reduction in the recognition error of long entities from 27.42% to 14.52%.



(c) Percentage of error entities regarding the length of entities.

Conclusions

- Single order LSTM may lead to poor performance
- However, integrating the information from single order models achieves very competitive results
- The proposed MO-BiLSTM mainly helps in the prediction of segment boundaries and the recognition of long segments.

Thank You