SLK-NER: Exploiting Second-order Lexicon Knowledge for Chinese NER

Dou Hu [1] and Lingwei Wei [2,3]

[1]National Computer System Engineering Research Institute of China, Beijing, China

[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

2020/7/17

Outline

- Introduction
- The Proposed Method
- Experiments and Analysis
- Conclusions

Outline

- **Introduction**
- The Proposed Method
- Experiments and Analysis
- Conclusions

Background

- NER aims to locate and classify named entities into predefined entity categories in the corpus.
- Named Entity Recognition (NER) is a fundamental task for various downstream applications, such as information retrieval, question answering, etc.
- Word boundaries in Chinese are ambiguities and word segmentation errors have a negative impact on identifying Name Entity.
- Character-based taggers can outperform word-based counterparts
- Integrate **external lexicon knowledge** into **character-based models**
- However
  - The lexical words may introduce erroneous information
  - Word boundary conflicts (lead to wrongly matched entities)

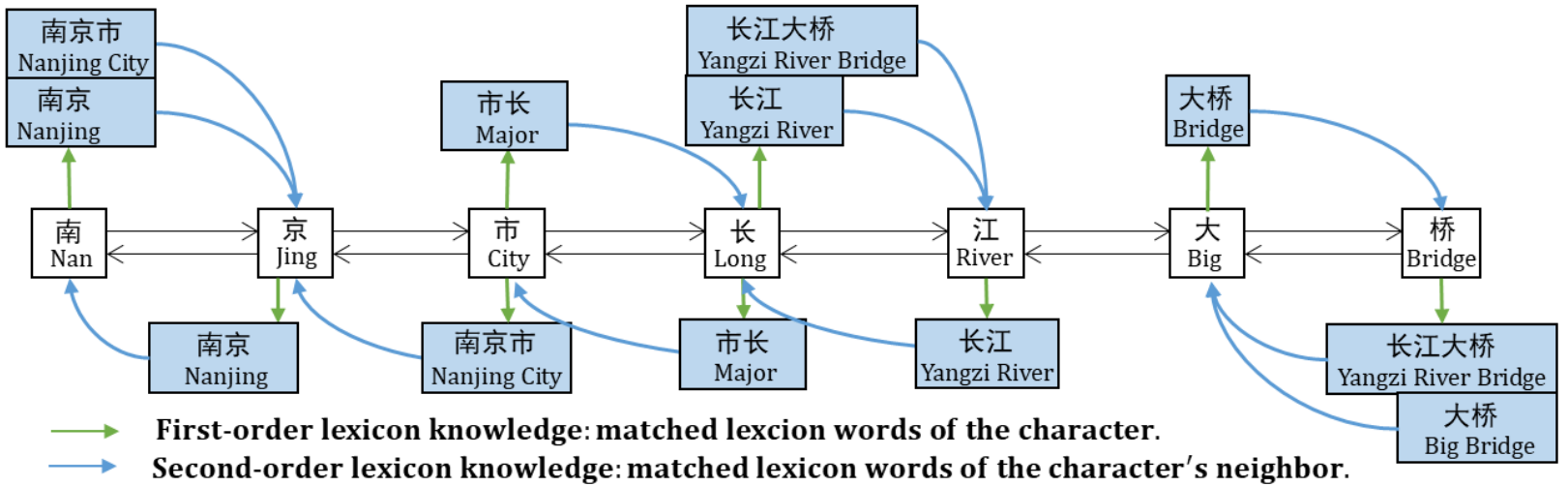
Recent Progress

- **Sequence-based methods**
  - Zhang et al. introduced a lattice LSTM to model all potential words matching a sentence to exploit explicit word information.
  - Liu et al. integrated word boundary features into input character vector via four strategies.
  - Gui et al. extend rethinking mechanism to relieve word boundary conflicts.
- **Weakness**
  - They performed with **simple first-order lexicon knowledge**, which provides **insufficient word information**

Recent Progress

- **Graph-based methods**
  - Gui et al. proposed a GNN-based method to explore multiple graph-based interactions among characters, potential words, and semantics.
  - Sui et al. proposed a collaborative graph network to assign both self-matched and the nearest contextual lexical words.
  - Ding et al. proposed a multi-digraph structure to learn the contextual information of the characters and the lexicon.
- **Weakness**
  - They explored the lexicon knowledge with graph **where higher-order information introducing negative words may disturb the identification**.

Motivation



- First-order lexicon knowledge only contains the lexical features of the characters itself, which cannot offer adequate word information.
- The conflict caused by this deficiency mainly comes from the middle of the named entity.
- Second-order lexicon knowledge: the neighbor's lexicon knowledge of the character.

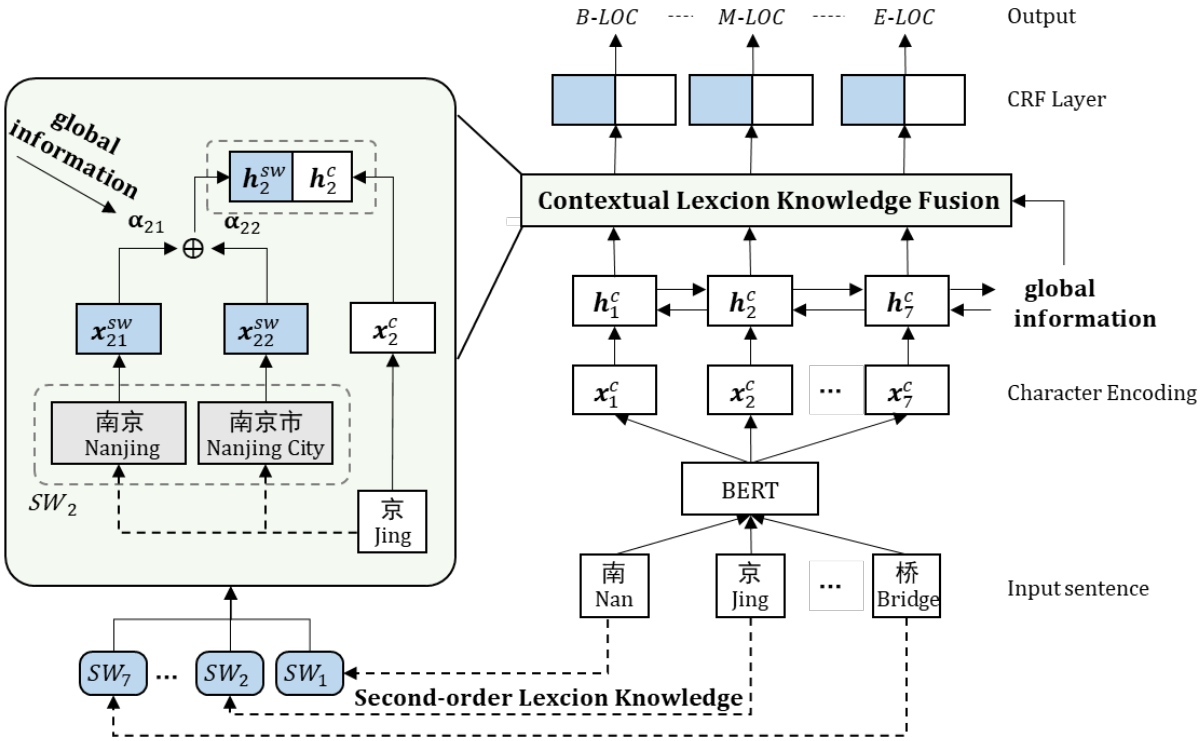
Contribution

- A new **insight** about second-order lexicon knowledge (SLK) of the character. SLK provides sufficient lexicon knowledge into characters in sentences and is capable of relieving the challenge of word boundary conflicts.
- We propose **a Chinese NER model** named SLK-NER with a novel strategy to integrate lexicon knowledge into the character-based model.
- Experimental results demonstrate the efficiency of SLK and our model significantly outperforms pervious methods, achieving state-of-the-art over three public Chinese NER datasets.

Outline

- Preamble
- **The Proposed Method**
- Experiments and Analysis
- Conclusions

An Overview of Our Approach



- Character-based sentences are encoded via character encoding layer. To integrate more lexicon knowledge, we construct the SLK for each character.
- A fusion layer with the global attention information is used for fusing different SLK
- A standard CRF model is employed for decoding labels.

SLK: Second-order Lexicon Knowledge

- An input sentence $s = \{c_1, c_2 \ldots, c_n\}$
- The lexicon $\mathcal{D}$
- First Lexicon Knowledge of i-th character

$$\mathcal{FW}_i = \overrightarrow{\mathcal{FW}}_i \cup \overleftarrow{\mathcal{FW}}_i$$

  - $\overrightarrow{\mathcal{FW}}_i$ and $\overleftarrow{\mathcal{FW}}_i$ denote a set of words obtained by matching all possible forward and backward subsequences in the lexicon, respectively

- Second-order Lexicon Knowledge of i-th character

$$\mathcal{SW}_i = \overrightarrow{\mathcal{FW}}_{i-1} \cup \overleftarrow{\mathcal{FW}}_{i+1}, i \in [1, n].$$

- For example

南京市长江大桥
Nanjing City Yangtze river bridge

南京市(Nanjing City)
南京(Nanjing)

SLK-NER

- **Character Encoding Layer**
  - Encode each character c_i in the sentence to a vector

$$\mathbf{x}_i^c = BERT(c_i).$$

  - To capture more contextual information, we apply Bi-GRU:

$$\mathbf{h}_i^c = GRU(\mathbf{x}_i^c), i \in [1, n].$$

  - Global Features

$$\mathbf{g} = \mathbf{h}_n^c.$$

- **Lexicon Knowledge Encoding Layer**
  - Pre-trained word embedding

$$\mathbf{x}_{ij}^{sw} = \mathbf{e}^w(sw_{ij})$$

SLK-NER

- **Contextual Lexicon Knowledge Fusion**
  - Global contextual information is introduced to extract SLK
  - For the j-th word in the matching set $\mathcal{SW}_i$ of i-th character,

$$\mathbf{u}_{ij} = \mathbf{W}_u \mathbf{x}_{ij}^{sw} + \mathbf{b}_u$$

  - Measure the importance of lexical word

$$\alpha_{ij} = \frac{exp(\mathbf{u}_{ij}^T \mathbf{g})}{\sum_j exp(\mathbf{u}_{ij}^T \mathbf{g})},$$

$$\mathbf{h}_i^{sw} = \sum_j \alpha_{ij} \mathbf{x}_{ij}^{sw}.$$

  - The final representation

$$\mathbf{r}_i = [\mathbf{h}_i^{sw}; \mathbf{h}_i^c].$$

SLK-NER

- **Decoding and Training**
  - A CRF is used to make sequence tagging

$$\mathbf{O} = \mathbf{W}_o\mathbf{R} + \mathbf{b}_o$$

$$p(y|s) = \frac{exp(\sum_i(\mathbf{O}_{i,y_i} + \mathbf{T}_{y_{i-1},y_i}))}{\sum_{\hat{y}} exp(\sum_i \mathbf{O}_{i,\hat{y}_i} + \mathbf{T}_{\hat{y}_{i-1},\hat{y}_i}))}$$

  - Objective function

$$L = -\sum_j log(p(y_j|s_j)).$$

Outline

- Preamble
- The Proposed Method
- **Experiments and Analysis**
- Conclusions

Experimental Settings

Data sample

```
吴 B-NAME
重 M-NAME
阳 E-NAME
， O
中 B-CONT
国 M-CONT
国 M-CONT
籍 E-CONT
， O
大 B-EDU
学 M-EDU
本 M-EDU
科 E-EDU
...
```

- **Datasets**
  - **OntoNote4** is a multilingual corpus in the news domain.
  - **Weibo** dataset consists of annotated NER messages drawn from Sina Weibo.
  - **Resume** dataset is composed of resumes collected from Sina Finance.

TABLE I
THE STATISTICS OF THE DATASETS.

Dataset	Training	Validation	Testing
OntoNotes4	15724	4301	4346
Weibo	1350	270	270
Resume	3821	463	477

- **Evaluation Metrics**
  - Precision/Recall/F1

Baselines

- **General sequence labeling models**
  - BiLSTM-CRF (Huang et al., 2015)
  - BERT(Devlin et al., 2019)
  - CAN(Zhu et al., 2019)
- **Sequence-based models**
  - WC-LSTM(Liu et al., 2019)
  - LR-CNN(Gui et al., 2019)
  - Lattice-LSTM(Zhang et al., 2018)
- **Graph-based models**
  - CGN(Sui et al., 2019)
  - LGN(Gui et al., 2019)
  - MG-GNN(Ding et al., 2019)

Experimental Results

TABLE II
EXPERIMENTAL RESULTS(%) ON THREE DATASETS.

Models	OntoNotes4			Weibo			Resume		
	P	R	F1	P	R	F1	P	R	F1
BiLSTM-CRF[13]	72.0	75.1	73.5	60.8	52.9	56.6	93.7	93.3	93.5
BERT[12]	**78.0**	80.4	79.2	61.2	63.9	62.5	94.2	95.8	95.0
CAN[14]	75.1	72.3	73.6	55.4	63.0	59.3.	95.1	94.8	94.9
LGN[9]	76.1	73.7	74.9	-	-	60.2	95.3	95.5	95.4
MG-GNN[11]	74.3	76.2	75.2	**63.1**	56.3	59.5	-	-	-
CGN[10]	75.1	74.5	74.8	-	-	63.1	-	-	-
LatticeLSTM[5]	76.4	71.6	73.9	53.0	62.3	58.8	94.8	94.1	94.5
WC-LSTM[8]	76.1	72.9	74.4	52.6	**67.4**	59.8	95.3	95.2	95.2
LR-CNN[7]	76.4	72.6	74.5	-	-	59.9	**95.4**	94.8	95.1
SLK-NER	77.9	**82.2**	**80.2**	61.8	66.3	**64.0**	95.2	**96.4**	**95.8**

Strategies Analysis

- **Lexicon Knowledge Types**

TABLE III
EXPERIMENTAL RESULTS (%) OF DIFFERENT ENCODING STRATEGIES ON THREE DATASETS.

Encoding Strategy	OntoNotes4			Weibo			Resume		
	P	R	F1	P	R	F1	P	R	F1
using SLK	**77.9**	82.2	**80.2**	**61.8**	66.3	**64.0**	**95.2**	**96.4**	**95.8**
using FLK	76.6	**82.9**	79.8	**61.8**	64.6	63.2	95.1	96.2	95.6
using SLK and FLK	76.4	82.7	79.6	60.6	63.6	62.1	94.9	96.2	95.5
no lexicon	77.7	81.3	79.6	56.7	**66.5**	61.2	94.2	96.1	95.1

- The character-based model performs poorly without lexicon knowledge
- Adding SLK outperforms significantly on F1 in all datasets
- When using both FLK and SLK, the F1 declines over three datasets

Strategies Analysis

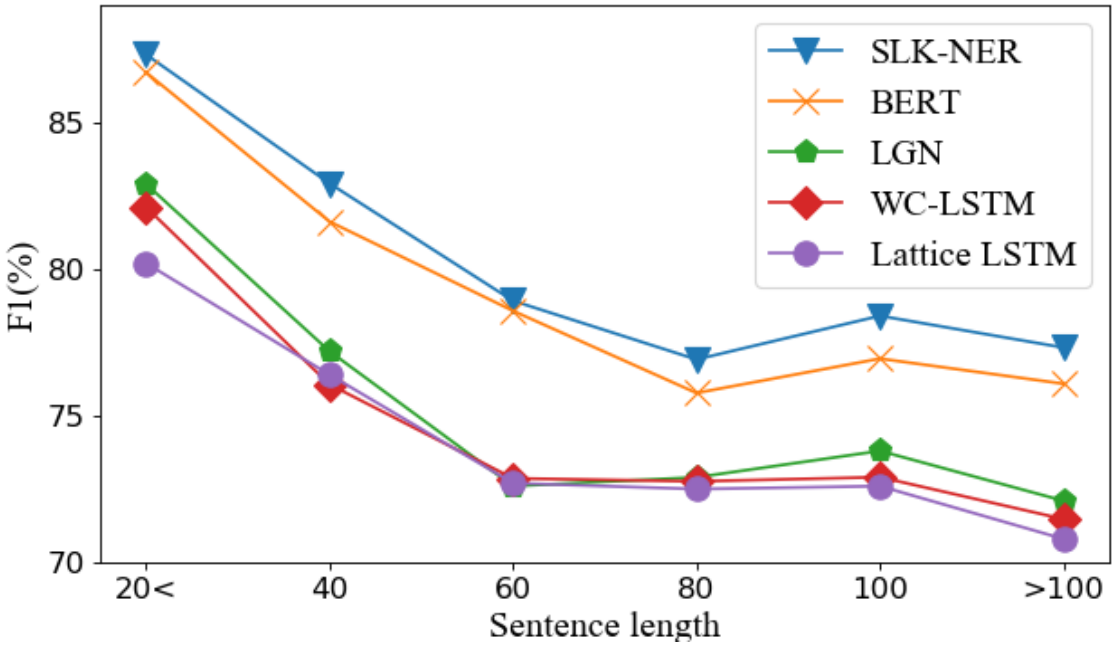
- **Lexicon Knowledge Encoding**

TABLE IV
EXPERIMENTAL RESULTS (%) OF DIFFERENT FUSION STRATEGIES ON THREE DATASETS.

Fusion Strategy	OntoNotes4			Weibo			Resume		
	P	R	F1	P	R	F1	P	R	F1
Global-Attention	77.9	**82.2**	**80.2**	**61.8**	**66.3**	**64.0**	**95.2**	**96.4**	**95.8**
Self-Attention	77.2	81.2	79.1	55.9	60.1	57.9	94.2	96.3	95.2
Shortest Word First	77.1	81.5	79.2	55.8	57.7	56.7	93.9	96.1	95.0
Longest Word First	77.1	81.6	79.3	57.6	56.9	57.3	94.7	96.1	95.4
Average	**78.6**	80.8	79.7	56.4	58.4	57.3	94.3	96.3	95.3

- Global attention in our model achieves best performance on F1 score
- Our model can combine more informative features to determine the word boundary and effectively alleviate the negative influence of word boundary conflicts

Sentence Length Analysis



- F1 against sentence length on OntoNotes4 dataset.
- BERT and SLK-NER outperform significantly than other baselines
- SLK-NER obtains a higher F1 over different sentence lengths

Outline

- Preamble
- The Proposed Method
- Experiments and Analysis
- **Conclusions**

Conclusion

- We present a new insight about second-order lexicon knowledge (SLK) of the character. SLK can provide sufficient lexicon knowledge into characters in sentences and is capable of relieving the challenge of word boundary conflicts.
- We propose a Chinese NER model named SLK-NER with a novel strategy to integrate lexicon knowledge into the character-based model. SLK-NER can enable to capture more beneficial word features with the help of global context information via attention mechanism.
- Experimental results demonstrate the efficiency of SLK and our model significantly outperforms pervious methods, achieving state-of-the-art over three public Chinese NER datasets.

Thank you !

Dou Hu and Lingwei Wei

hudou18@mails.ucas.edu.cn
weilingwei@iie.ac.cn

Datasets and code: https://github.com/zerohd4869/SLK-NER