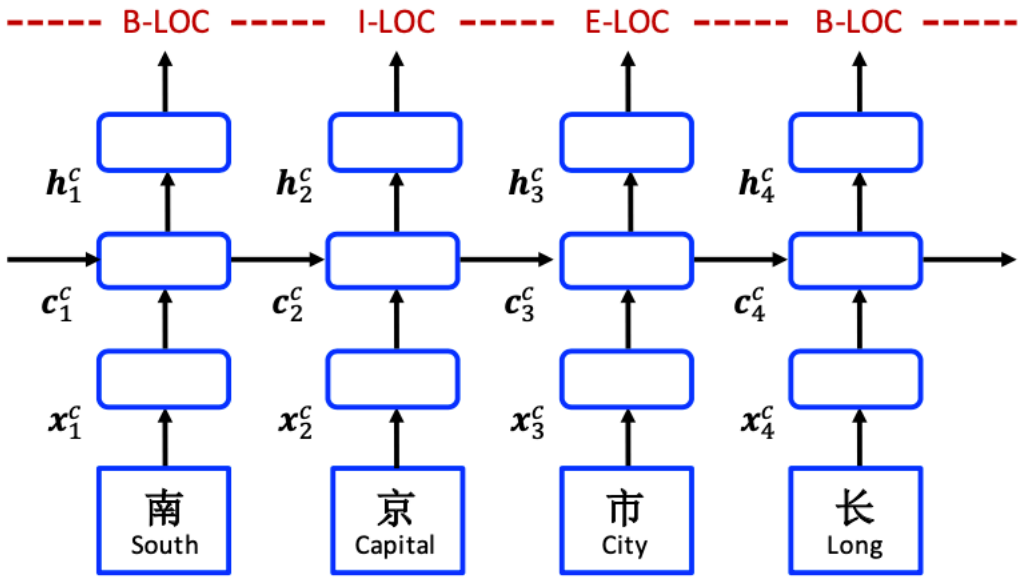
中文NER任务

许润昕

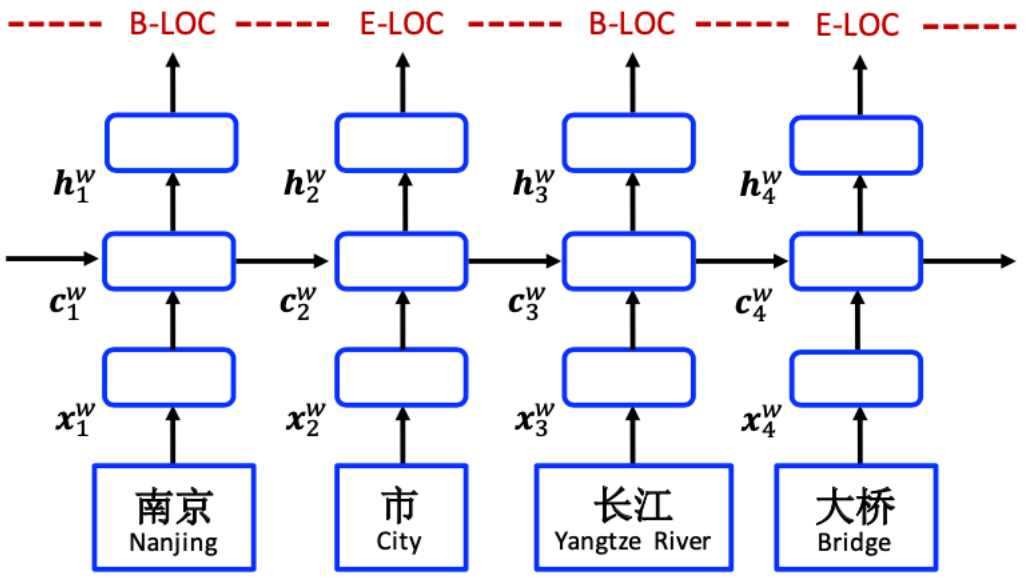
中文NER任务与英文NER任务的区别与特点：中文句子不包括Word Boundary。

英文：Nanjing Yangtze River Bridge

中文：南京市长江大桥



(a) Character-based model.



(b) Word-based model.

丢失word boundary信息

Error propogation

Chinese NER Using Lattice LSTM （ACL2018）

- 引入词典的中文NER开山之作

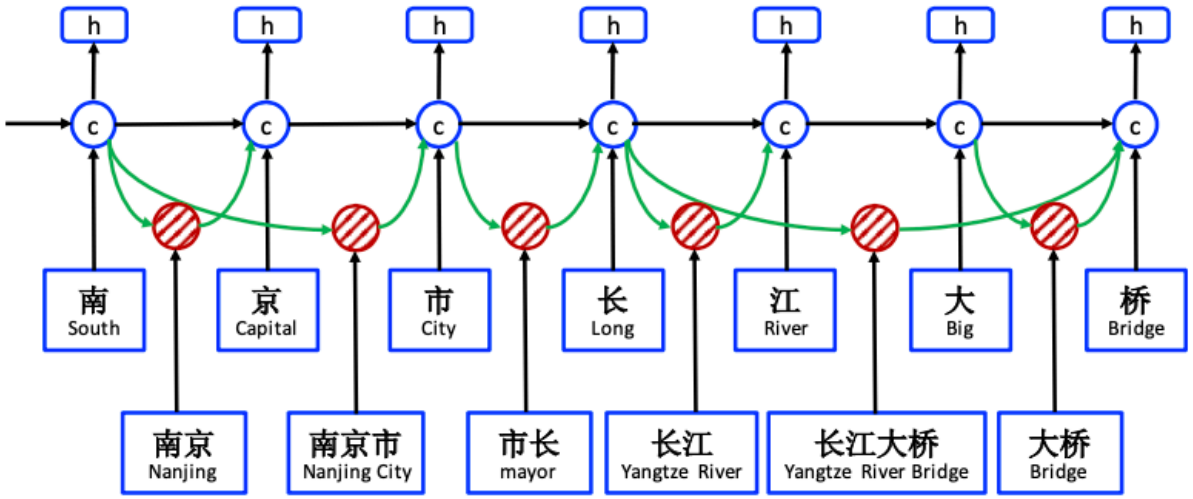


Figure 2: Lattice LSTM structure.

Models	P	R	F1
Chen et al. (2006a)	91.22	81.71	86.20
Zhang et al. (2006)*	92.20	90.18	91.18
Zhou et al. (2013)	91.86	88.75	90.28
Lu et al. (2016)	–	–	87.94
Dong et al. (2016)	91.28	90.62	90.95
Word baseline	90.57	83.06	86.65
+char+bichar LSTM	91.05	89.53	90.28
Char baseline	90.74	86.96	88.81
+bichar+softword	92.97	90.80	91.87
Lattice	**93.57**	**92.79**	**93.18**

Table 6: Main results on MSRA.

Chinese NER Using Lattice LSTM （ACL2018）

- 存在的问题：
  - 每个cell依赖的node数量不一样，比较难并行处理
  - 只能应用于RNN这类模型，无法推广
  - 信息感知有损失，只有最后一个character能够感知lexicon
  - 信息单向流动，前面的没法感知后面的context
- 后续研究基本上在follow它，并对他进行改进

Leverage Lexical Knowledge for Chinese Named Entity Recognition via Collaborative Graph Network（EMNLP2019）

- 解决lattice LSTM的两个问题：信息单向流动，感知context能力弱；只有最后一个character能够感知lexicon存在。
- 解决方法：用三个图+GAT，无向图

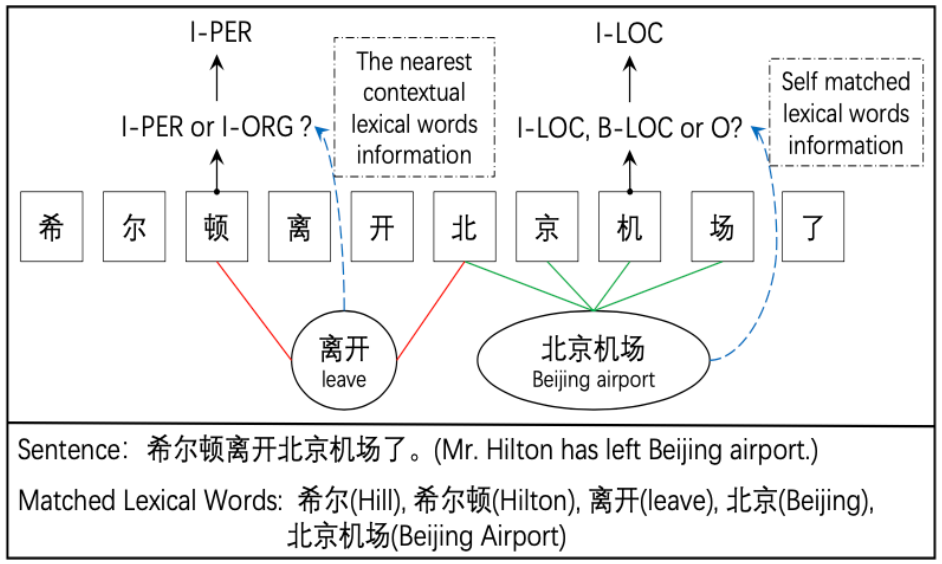
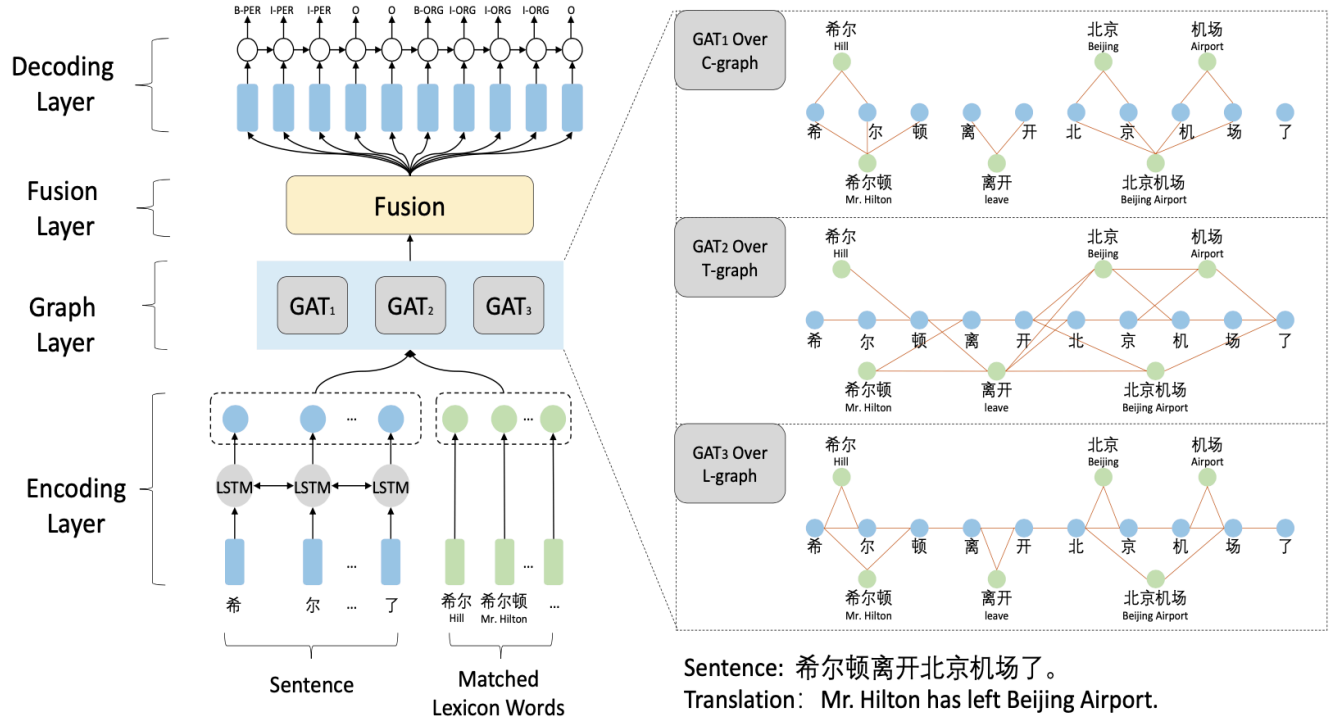


Figure 1: An example sentence integrating the nearest contextual lexical words (red line) and self-matched lexical words (green line)



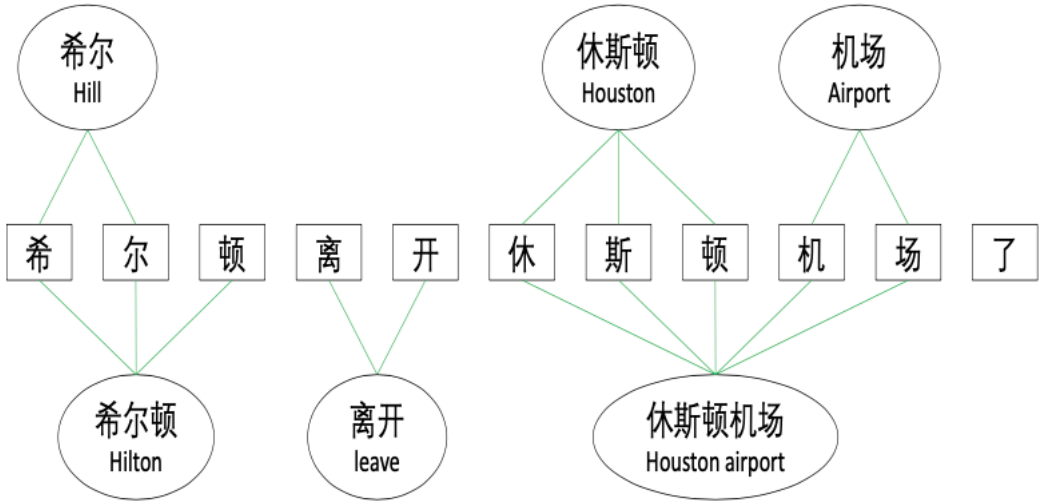


Figure 2: Word-Character Containing graph

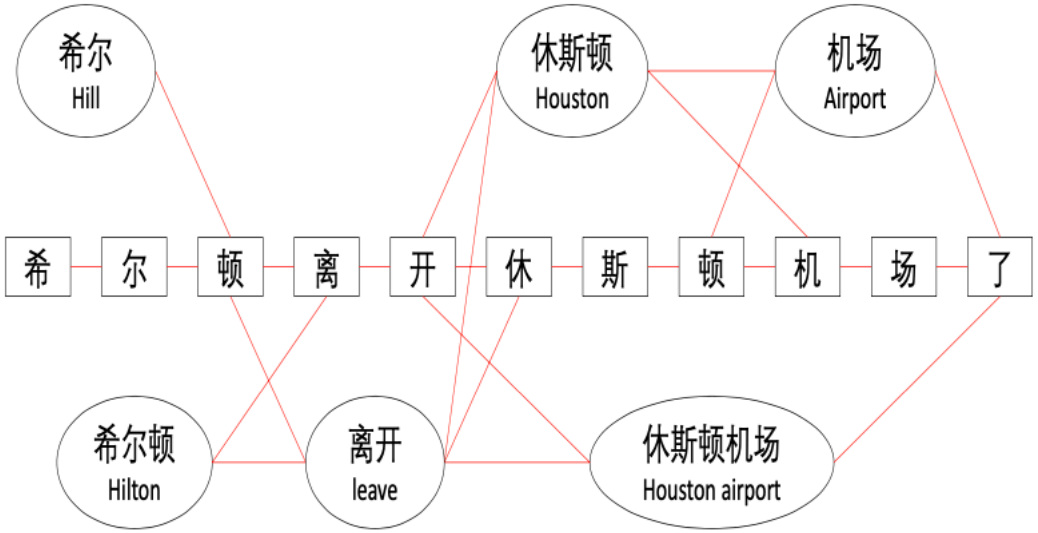


Figure 3: Word-Character Transition graph

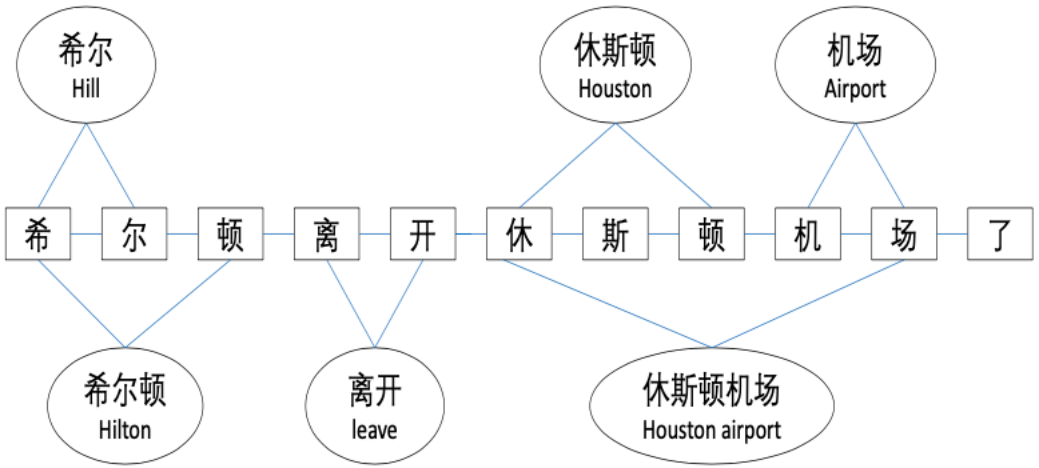
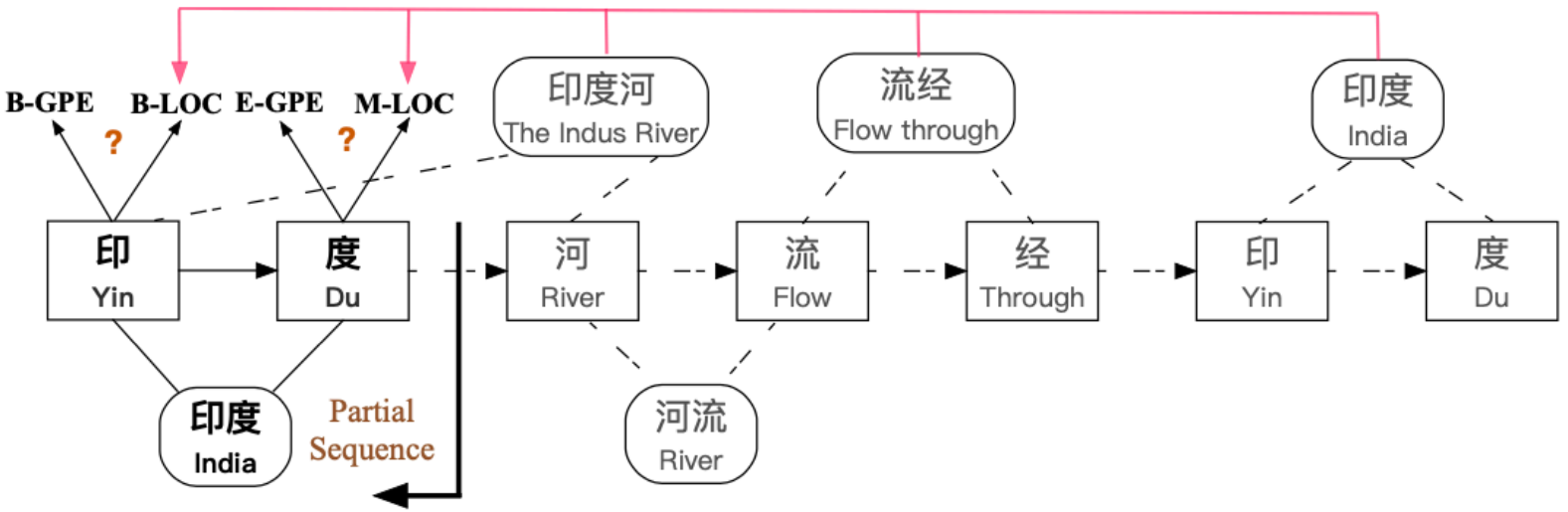


Figure 4: Word-Character Lattice graph

A Lexicon-Based Graph Neural Network for Chinese NER （EMNLP2019）

- 解决lattice LSTM的两个问题：信息单向流动，感知context能力弱；只有最后一个character能够感知lexicon存在。
- 解决方法：图，global node



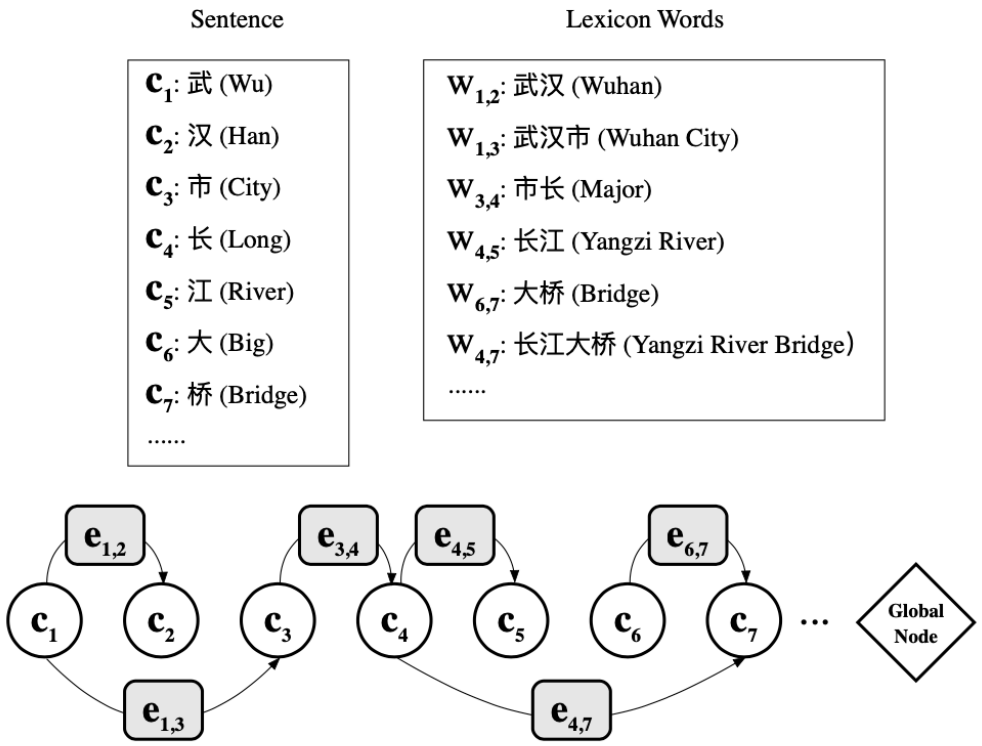
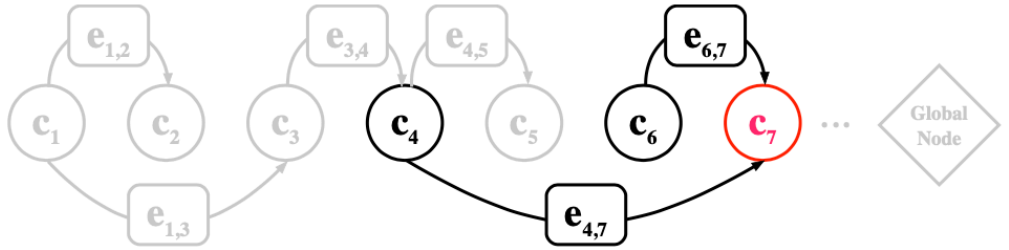
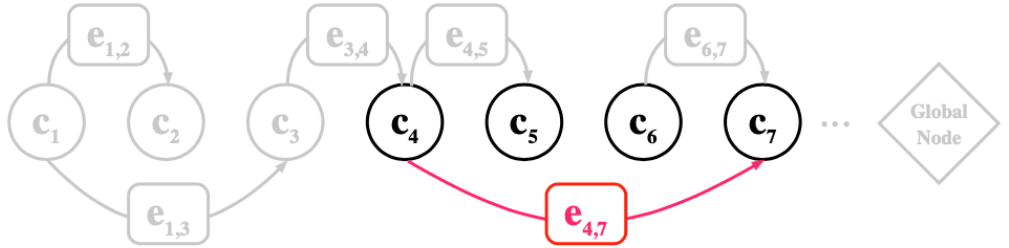


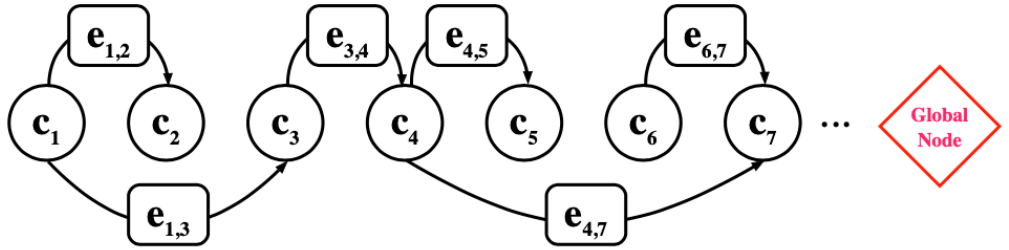
Figure 2: Illustration of graph construction.



(a) $e \rightarrow c$



(b) $c \rightarrow e$

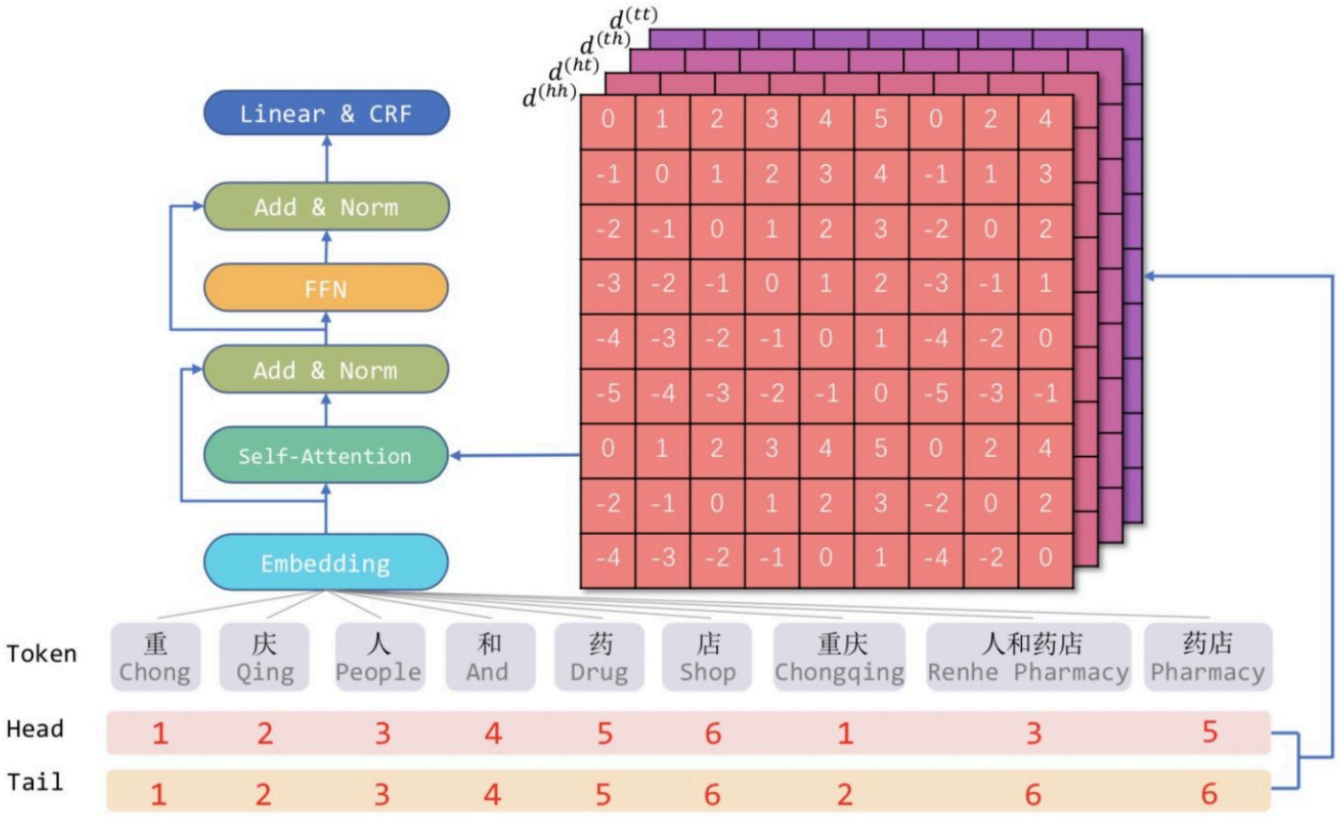


(c) $c, e \rightarrow g$

Figure 3: Aggregation in LGN. Red indicates the element that is being updated, and black indicates other elements involved in the aggregation. The aggregation results are then used in update modules.

FLAT: Chinese NER Using Flat-Lattice Transformer（ACL2020）

- 把图变成更规整的线性结构；RNN的长程依赖问题
- 模型关键是position embedding



FLAT结构

$$d_{ij}^{(hh)} = head[i] - head[j],$$
$$d_{ij}^{(ht)} = head[i] - tail[j],$$
$$d_{ij}^{(th)} = tail[i] - head[j],$$
$$d_{ij}^{(tt)} = tail[i] - tail[j],$$

$$\mathbf{A}_{i,j}^{*} = \mathbf{W}_q^{\top}\mathbf{E}_{x_i}^{\top}\mathbf{E}_{x_j}\mathbf{W}_{k,E} + \mathbf{W}_q^{\top}\mathbf{E}_{x_i}^{\top}\mathbf{R}_{ij}\mathbf{W}_{k,R} + \mathbf{u}^{\top}\mathbf{E}_{x_j}\mathbf{W}_{k,E} + \mathbf{v}^{\top}\mathbf{R}_{ij}\mathbf{W}_{k,R},$$

CNN-Based Chinese NER with Lexicon Rethinking（IJCAI2019）

- Lattice LSTM速度慢，中间词无法感知lexicon
- 通过堆叠的CNN捕捉所有可能n-gram

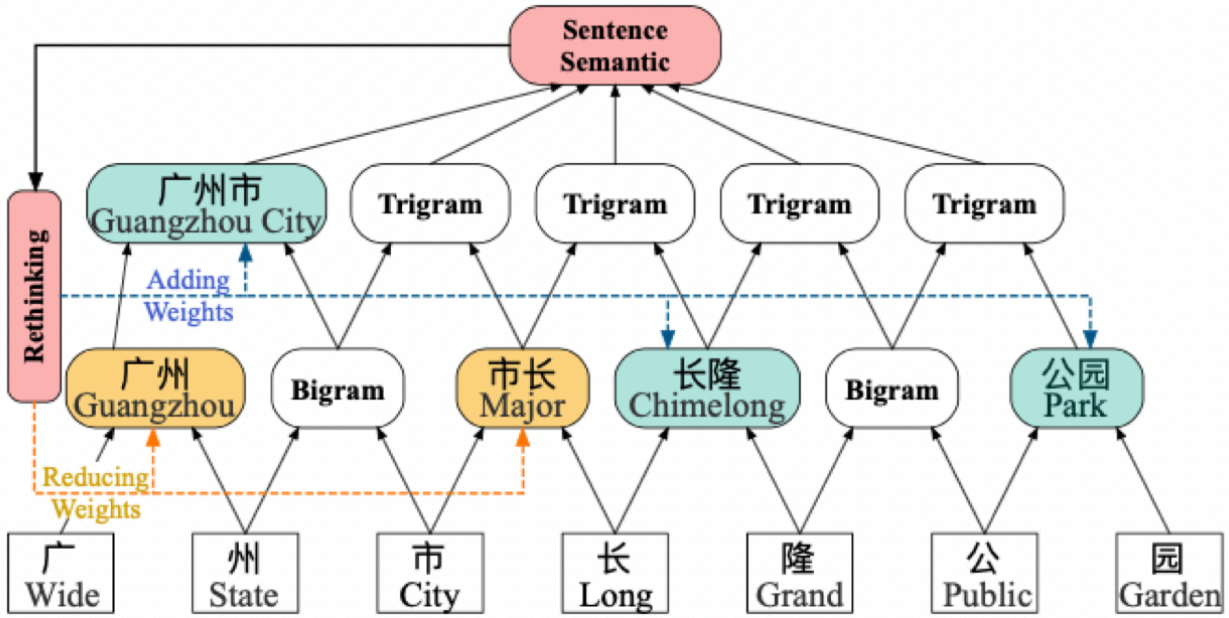


Figure 2: Schematic of CNN model incorporating lexicons with a rethinking mechanism.

$$s_m^l = \sum_{d=1}^{D} X_m^{l'}[d], \quad \alpha_m^l = \frac{\exp(s_m^l)}{\sum_{l=1}^{L} \exp(s_m^l)}$$

$$X_m^{att} = \sum_{l=1}^{L} \alpha_m^l X_m^{l'}.$$

An Encoding Strategy Based Word-Character LSTM for Chinese NER（NAACL2019）

- 不改变模型结构，而是在embedding层面来解决Lattice LSTM的问题

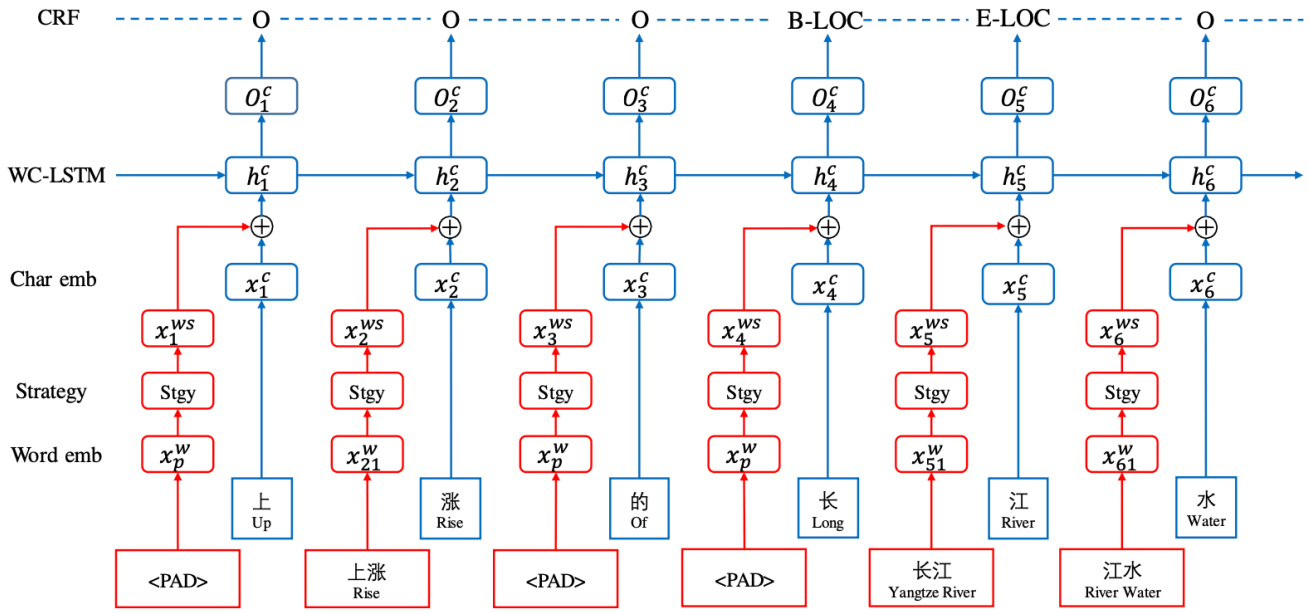


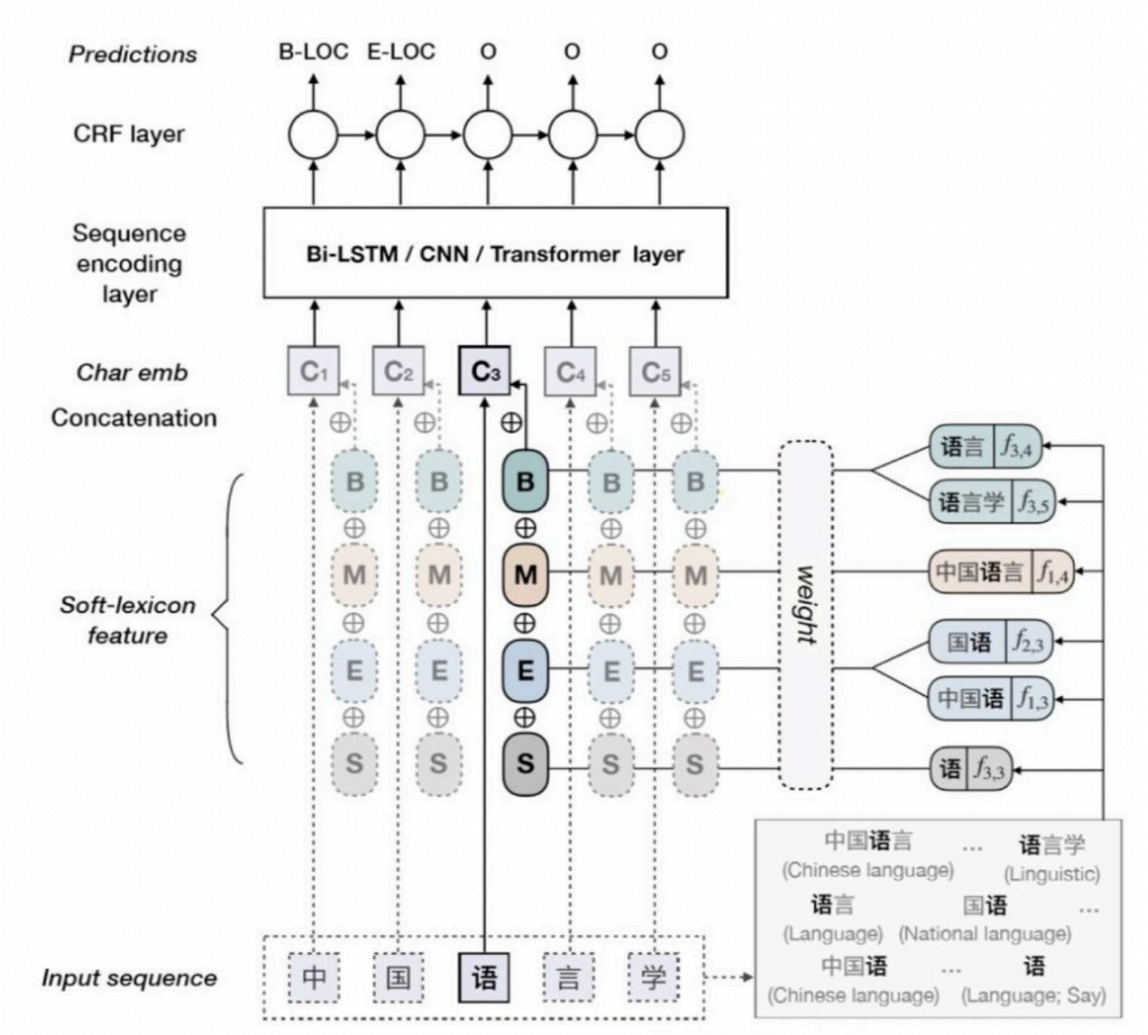
Figure 2: The architecture of our unidirectional model. The blue part can be seen as a standard character-based model but with a word-character LSTM(WC-LSTM), and the red part indicates the process of encoding word information into a fixed-size representation. Word information is integrated into the end character of the word. Where "<PAD>" denotes padding value; "Stgy" denotes a certain encoding strategy and ⊕ denotes concatenation operation.

如果有多个lexicon在同一个位置：

1. shortest word
2. longest word
3. average
4. self-attention

Simplify the Usage of Lexicon in Chinese NER（ACL2020）

- 在前面工作基础上，每个位置考虑BMES的所有lexicon



A Neural Multi-digraph Model for Chinese NER with Gazetteers（ACL2019）

- 采用的词典是实体词典，与前面那些工作不同

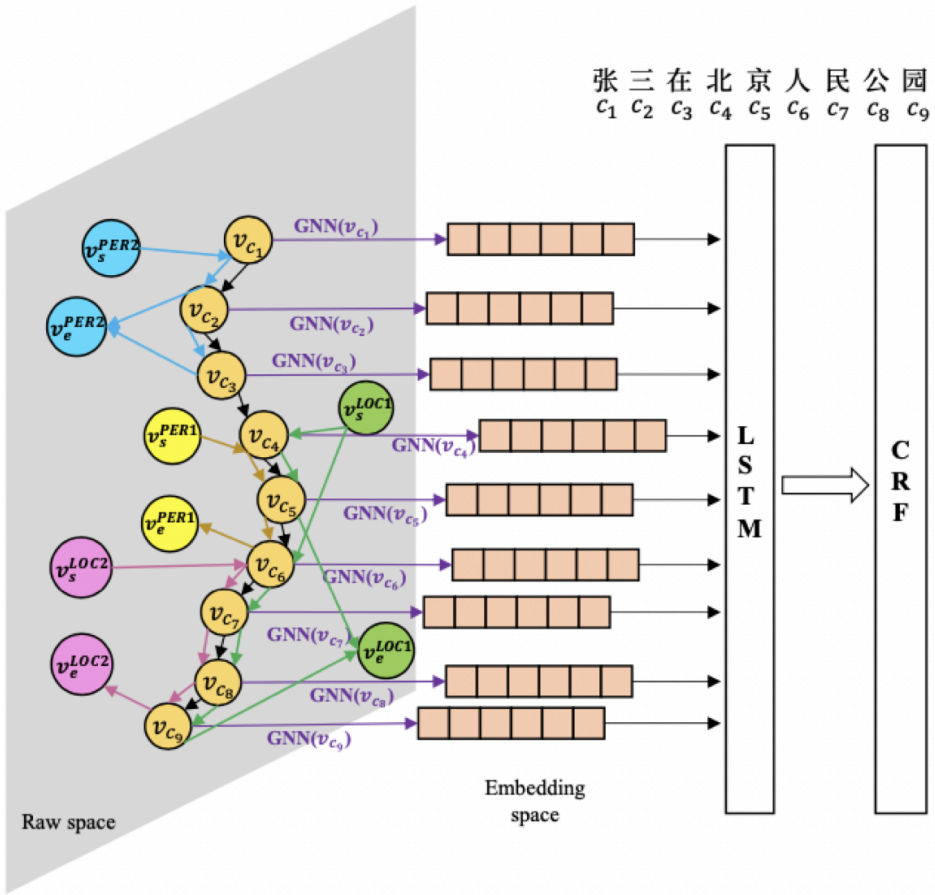


Figure 2: System architecture

词汇增强范式	方法	特点	存在问题
Dynamic Architecture 设计相应结构以融入词汇信息	Lattice LSTM	开篇之作，设计兼容的LSTM将词汇信息引入中文NER任务	推断效率低，无法捕捉长距离依赖，存在一定程度的信息损失问题。
	LR-CNN	采取CNN进行堆叠编码，采取rethink机制解决词汇冲突问题	
	CGN	构建基于协作的图网络（GAN），充分利用词汇信息	将NER任务转化为node分类任务，但需要RNN作为底层编码器来捕捉顺序性，结构复杂。
	LGN	构建局部和全局聚合的图网络，充分利用词汇信息	
	FLAT	通过设计位置向量引入词汇信息，利用transformer捕捉长距离依赖、提高推断效率。	
Adaptive Embedding 模型无关，具备可迁移性	WC-LSTM	通过四种encoding策略对Lattice LSTM输入静态编码	存在信息损失，仍然采取LSTM进行编码
	Multi-digraph	引入实体词典，通过多图结构更好地显示建模字符和词典的交互	
	Simple-Lexicon	通过Soft-lexicon方法引入词汇信息， 简单直接	

	lexicon	Ontonotes	MSRA	Resume	Weibo
biLSTM	----	71.81	91.87	94.41	56.75
Lattice LSTM	词表1	73.88	93.18	94.46	58.79
WC-LSTM	词表1	74.43	93.36	94.96	49.86
LR-CNN	词表1	74.45	93.71	95.11	59.92
CGN	词表2	74.79	93.47	94.12	63.09
LGN	词表1	74.85	93.63	95.41	60.15
Simple-Lexicon	词表1	75.54	93.50	**95.59**	61.24
FLAT	词表1	**76.45**	94.12	95.45	60.32
FLAT	词表2	75.70	**94.35**	94.93	**63.42**
BERT	----	80.14	94.95	95.53	68.20
BERT+FLAT	词表1	**81.82**	**96.09**	**95.86**	**68.55**

https://zhuanlan.zhihu.com/p/142615620

中文NER任务与英文NER任务的区别与特点：中文句子不包括Word Boundary。

解决方法：character-level + 借助外部词典注入知识

1. 通过修改模型结构实现（Lattice LSTM，GNN）
2. 通过修改embedding实现