



中文命名实体识别：任务与方法

复旦大学

黄萱菁

命名实体识别

识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等

Republican Trump replaces Obama as President of the United States.

↓
NER

Republican Trump replaces Obama as President of the United States.

组织名

人名

人名

地名

命名实体识别-作用

从无结构的文本中获取**实体**知识



2017苹果秋季新品发布会于美国当地时间**9月12日**上午10点在Apple Park新总部的**史蒂夫·乔布斯剧院**举行苹果新品发布会。此次发布会由**苹果公司CEO蒂姆·库克**主持，发布的产品有**iPhone X、iPhone8 & iPhone 8 Plus、Apple Watch Series 3、Apple TV**。

发布会时间 Time	9月12日	
发布会地点 Spot	史蒂夫·乔布斯剧院	
发布会主持人 Convener	个人姓名/ 团体名称 Name	蒂姆·库克
	机构、职位 Org/Post	CEO
发布会名称 Conf-Title	2017苹果秋季新品发布会	
发布产品 Product	iPhone X、iPhone8、iPhone 8 Plus、Apple Watch Series 3、Apple TV	

命名实体识别-作用

南京/市/长江大桥/建成/了

分词

verb art adj noun

Make a better world

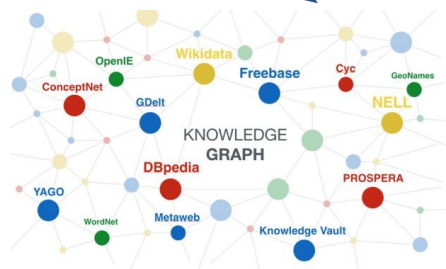
词性标注

When Sebastian Thrun **PERSON** started at Google **ORG** in 2007 **DATE**, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American **NORP** car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun **PERSON**, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode **ORG** earlier this week **DATE**.

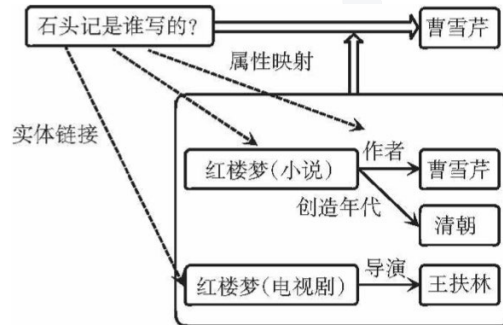
命名实体识别



提升信息检索的质量



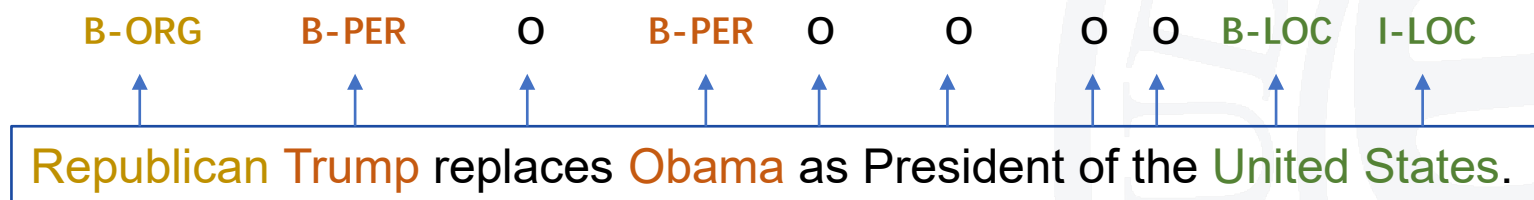
构建知识图谱



辅助智能问答

命名实体识别-整体思路

- 过程：
- (1) 实体边界识别
 - (2) 确定实体类别（人名、地名、机构名或其他）



B: 实体的开始

I: 实体的中间部分或者结尾

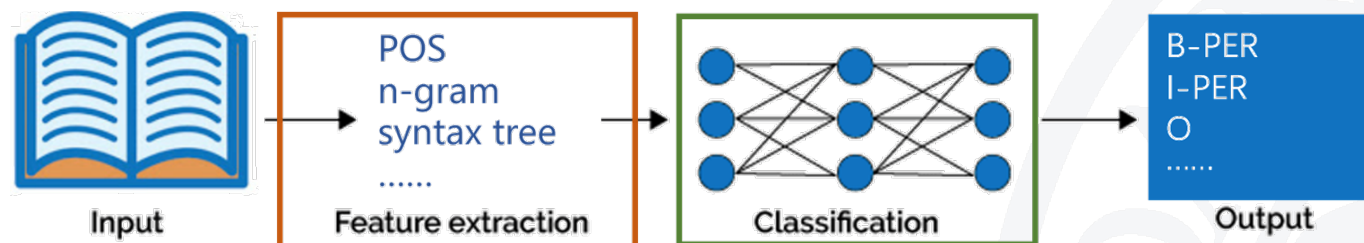
O: 实体之外

PER: 人名

ORG: 组织名

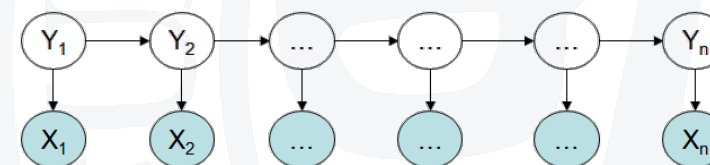
LOC: 地名

基于机器学习的命名实体识别

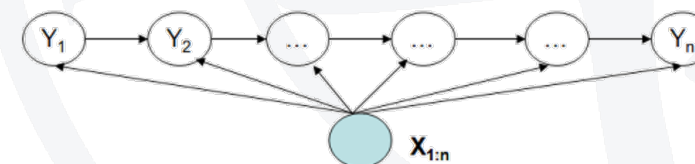


Features	Examples
Case	- Starts with a capital letter - Word is all uppercased - The word is mixed case (e.g., ProSys, eBay)
Punctuation	- Ends with period, has internal period (e.g., St., I.B.M.) - Internal apostrophe, hyphen or ampersand (e.g., O'Connor)
Digit	- Digit pattern (see section 3.1.1) - Cardinal and Ordinal - Roman number - Word with digits (e.g., W3C, 3M)
Character	- Possessive mark, first person pronoun - Greek letters
Morphology	- Prefix, suffix, singular version, stem - Common ending (see section 3.1.2)
Part-of-speech	- proper name, verb, noun, foreign word
Function	- Alpha, non-alpha, n-gram (see section 3.1.3) - lowercase, uppercase version - pattern, summarized pattern (see section 3.1.4) - token length, phrase length

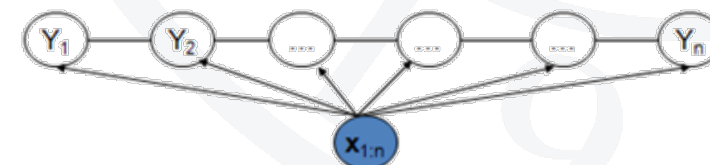
HMM



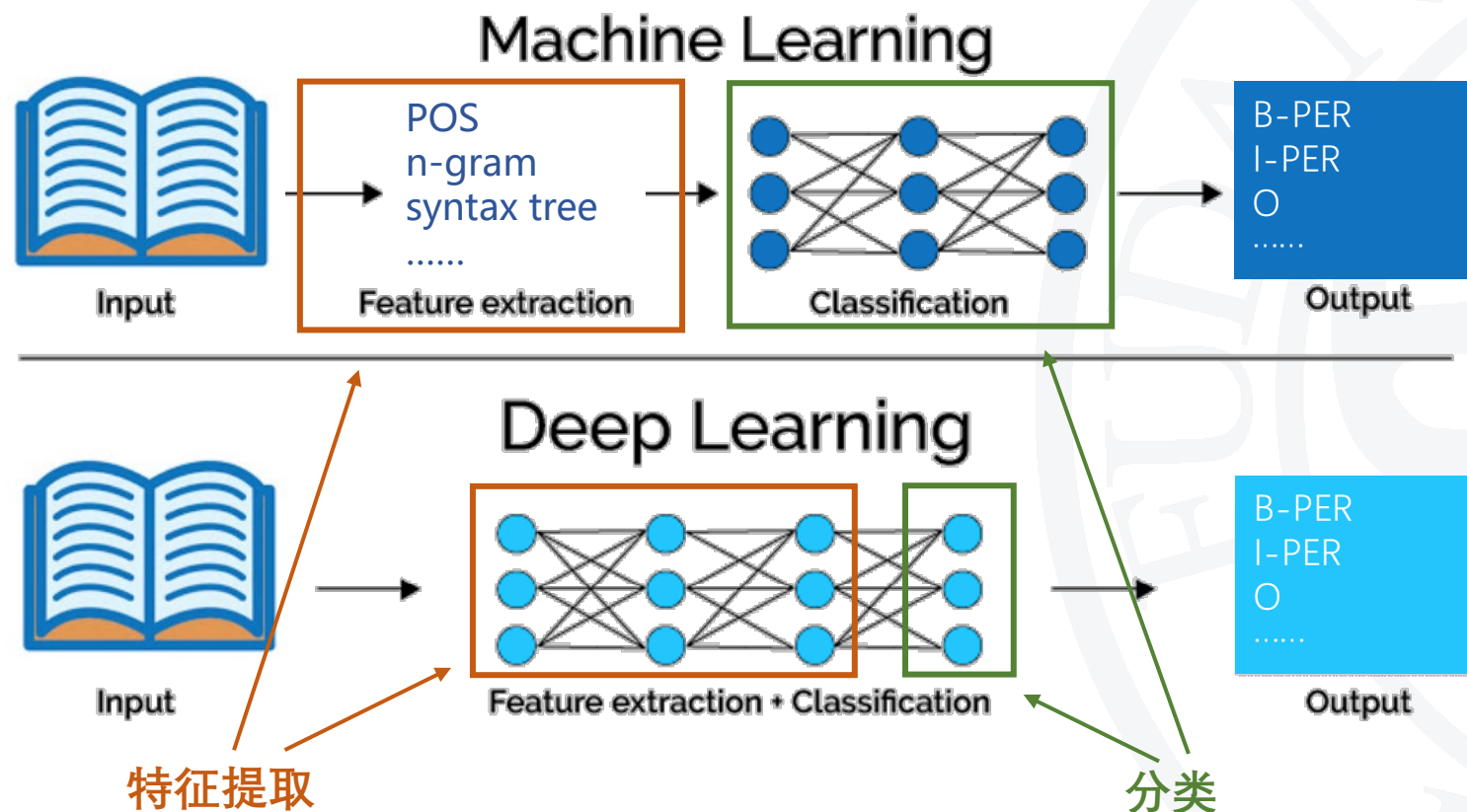
MEMM



CRF

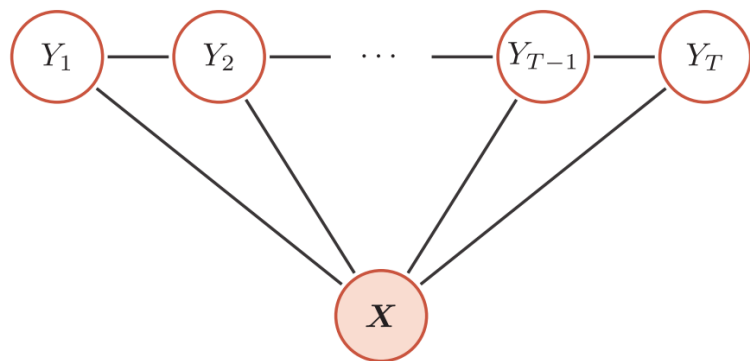


基于深度学习的命名实体识别

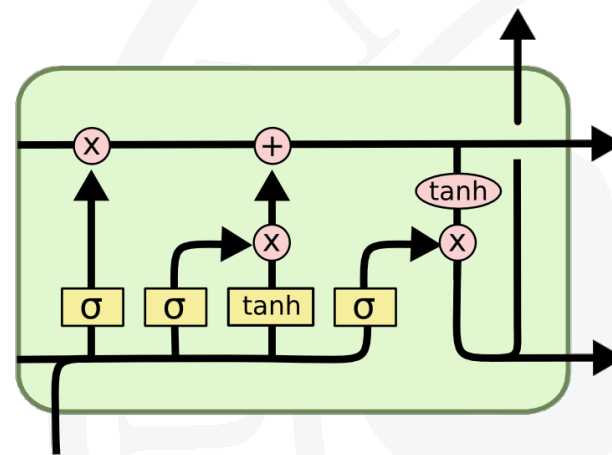


深度学习能够自动学习到对模型有益的特征
传统的机器学习方法需要需要繁杂的特征工程, 而深度学习则不需要

基于深度学习的命名实体识别



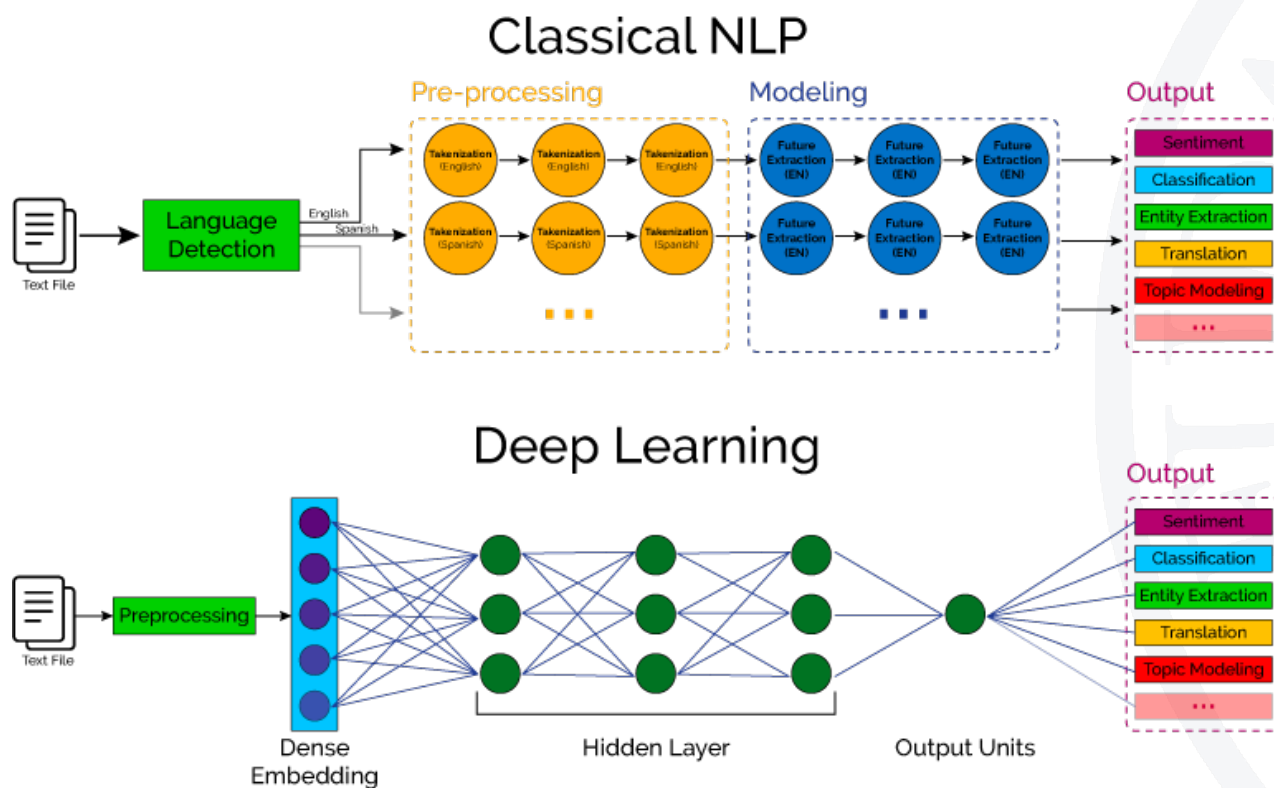
线性链条件随机场



长短时记忆神经网络

NER 受益于 DL 的高度非线性
相比于传统的线性模型(线性 HMM 和 线性链 CRF), 深度学习模型能够学到更复杂的特征

基于深度学习的命名实体识别

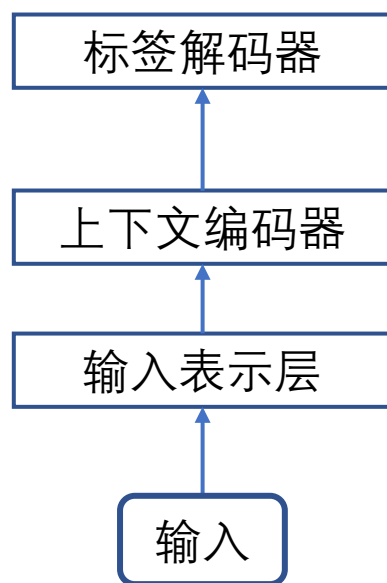


Pipeline

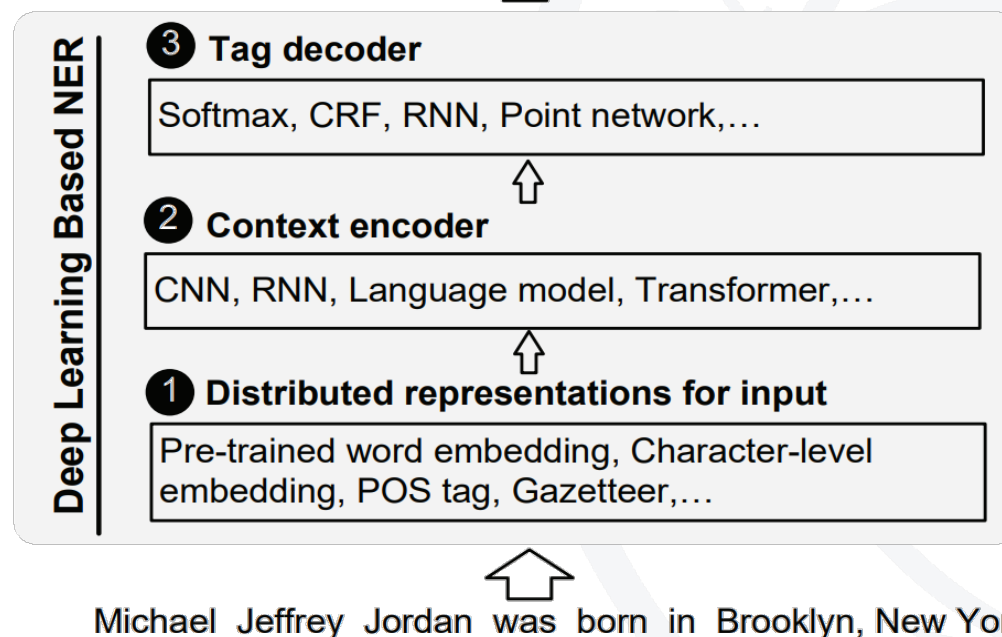
End-to-End

深度学习可以端到端地搭建模型, 这允许我们搭建更复杂的 NER 系统

基于深度学习的NER-通用框架



B-PER I-PER E-PER O O O S-LOC O B-LOC E-LOC O
Michael Jeffrey Jordan was born in Brooklyn , New York .

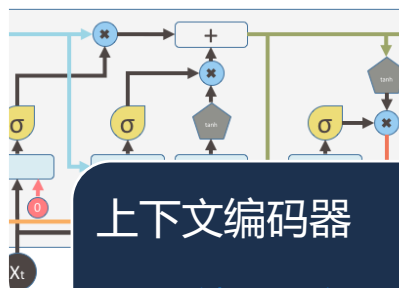


基于深度学习的NER模型



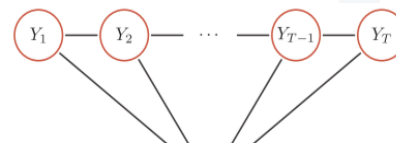
输入表示层

- 词级别
- 字符级别
- 预训练语言模型



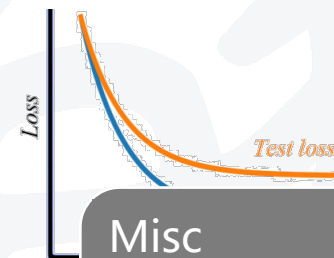
上下文编码器

- 循环神经网络
- 卷积神经网络
- Transformer
- 中文NER



标签解码器

- MLP+softmax
- 条件随机场 CRF
-

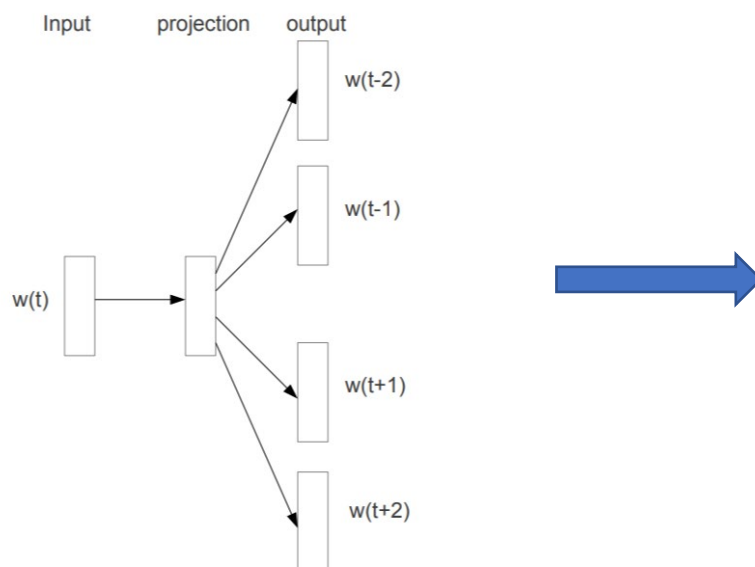


Misc

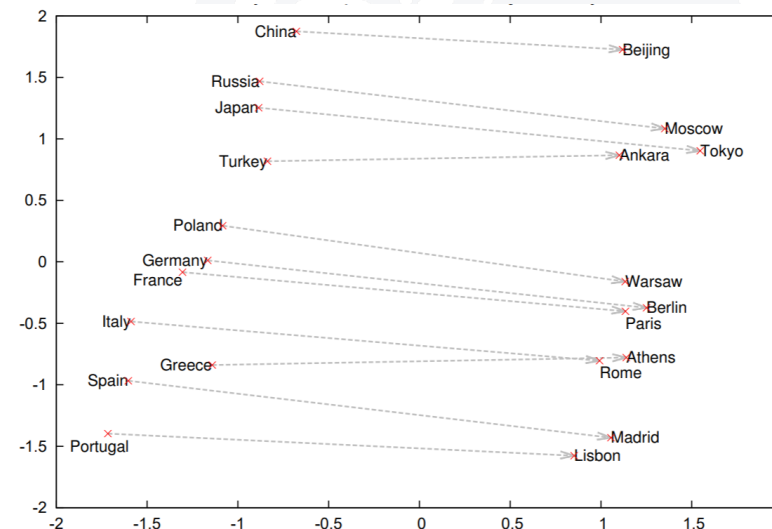
- 多模态NER
- NER的可解释性
-

输入表示层 (离散->连续)

- Word Embedding
 - Glove / Word2Vec / Random Init



Word2Vec (skip-gram)

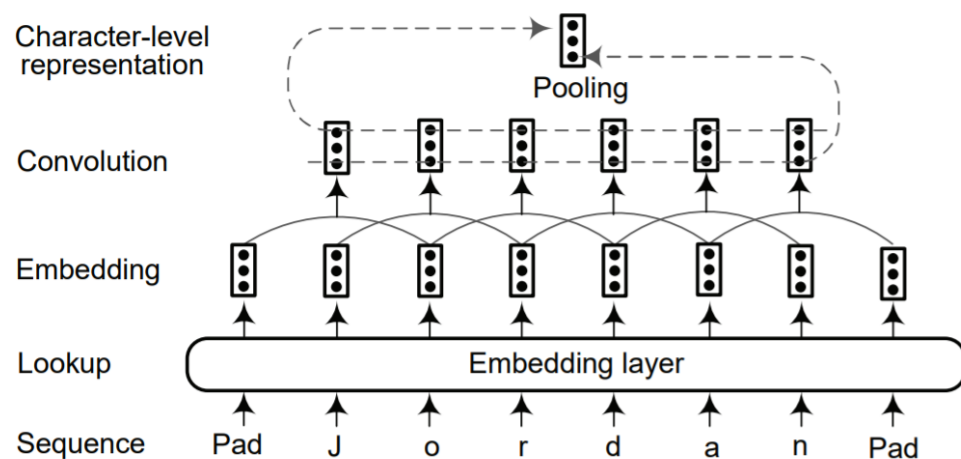


Good Representation :
Country and Capital Vectors

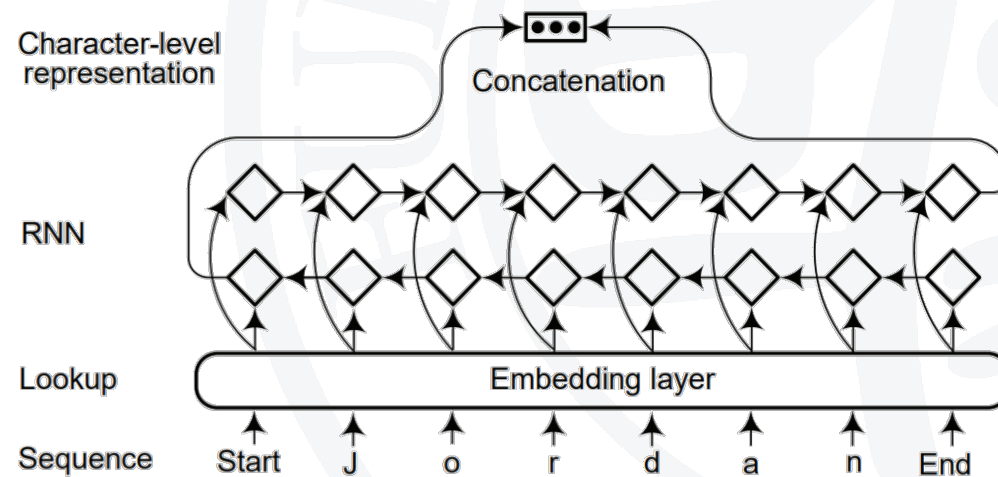
输入表示层 (离散->连续)

- Character embedding (less UNK, morphology)

RNN/CNN/Transformer based char-level representation



基于CNN的字符级别表示

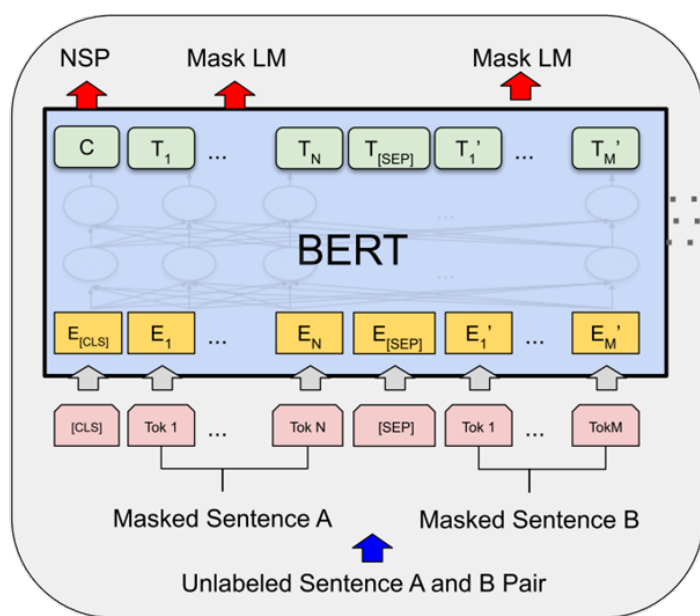


基于RNN的字符级别表示

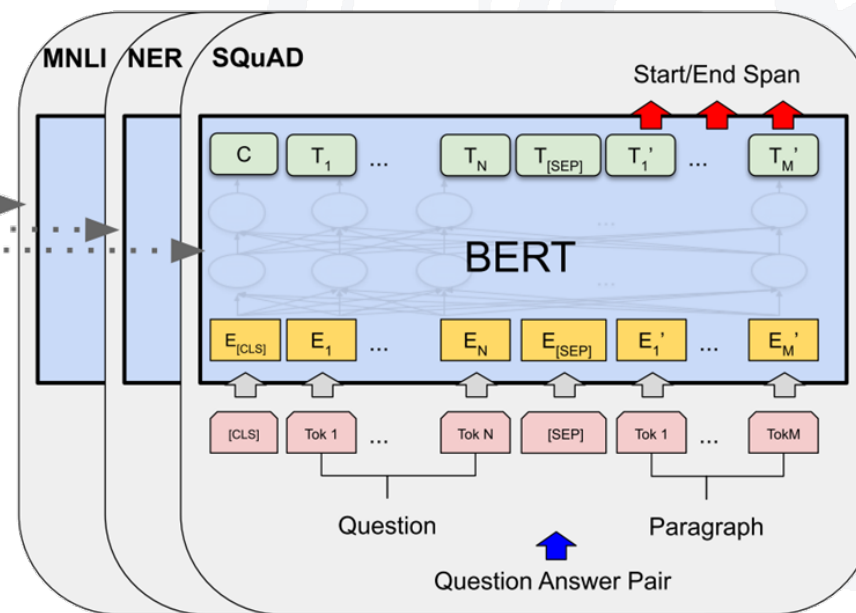
输入表示层 (离散->连续)

➤ Pretrained Contextual Embedding (contextual)

- ELMo / BERT (我在苹果公司工作/我吃了苹果)



Pre-training



Fine-Tuning

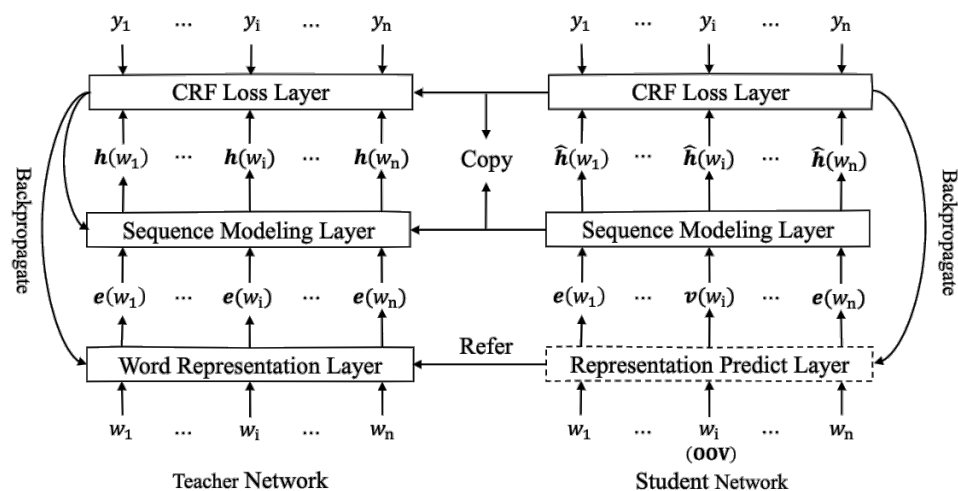
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019

Deep contextualized word representations, NAACL 2018

自动学习OOV的表示

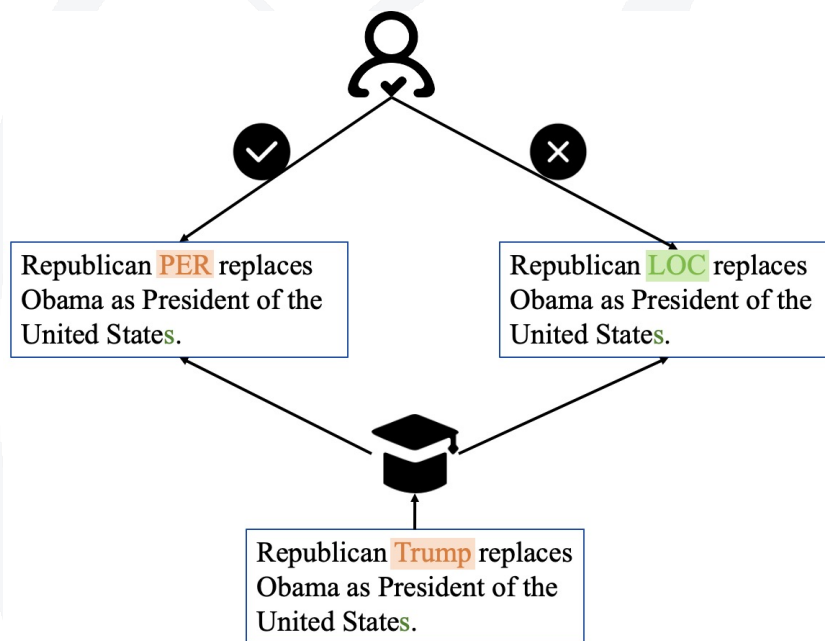
研究动机

- 在测试集中经常会出现一些在训练集中没有出现过的OOV词
- 在序列标注任务中，模型在OOV词上的表现相较于非OOV词一般差很多



模型

- 训练一个普通的序列标注模型，称为**Teacher Network**
- 训练一个OOV词表示模型(**Student Network**)：利用一个词的上下文和该词的字符序列预测该词表示，使得待预测的词表示在Teacher Network中有好的表现

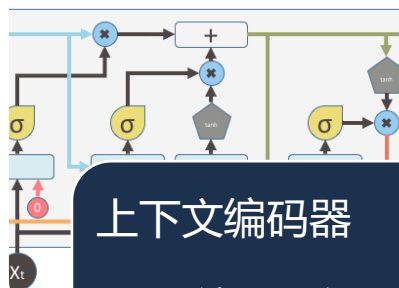


基于深度学习的NER模型



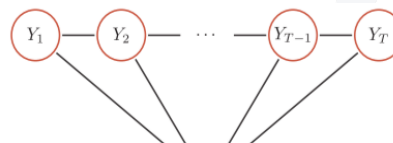
输入表示层

- 词级别
- 字符级别
- 预训练语言模型



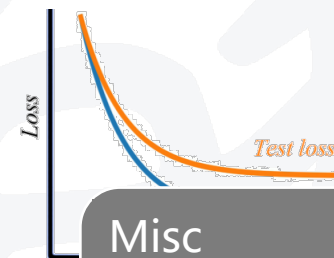
上下文编码器

- 循环神经网络
- 卷积神经网络
- Transformer
- 中文NER



标签解码器

- MLP+softmax
- 条件随机场 CRF
-

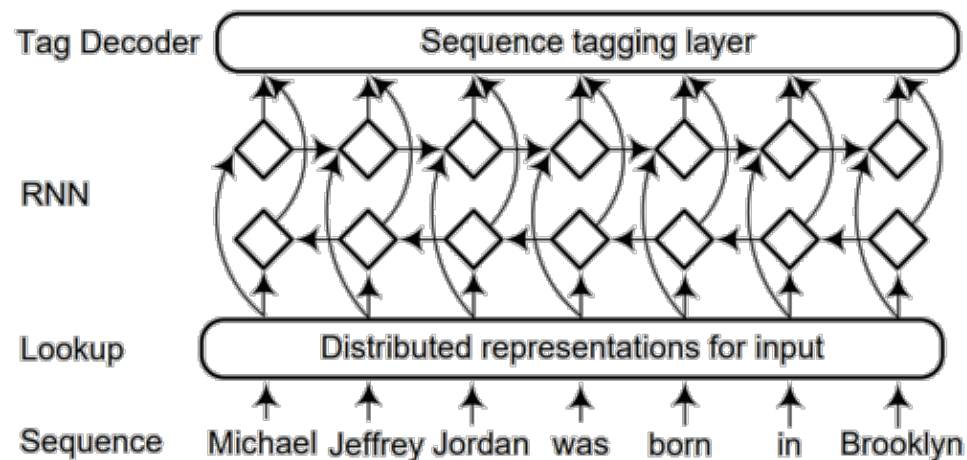


Misc

- 多模态NER
- NER的可解释性
-

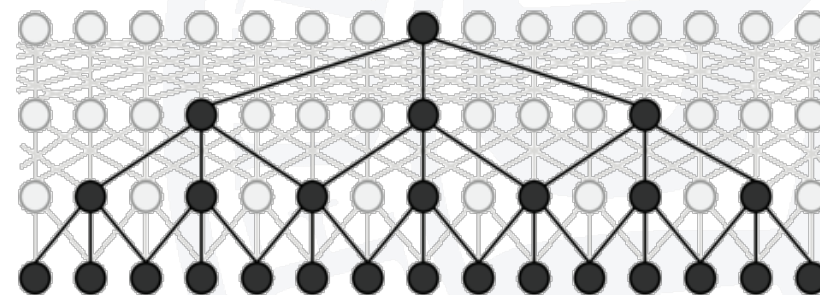
上下文编码器

- Recurrent Neural Network (RNN)
序列性建模文本，性能稳定，速度慢



Bi-LSTM for NER

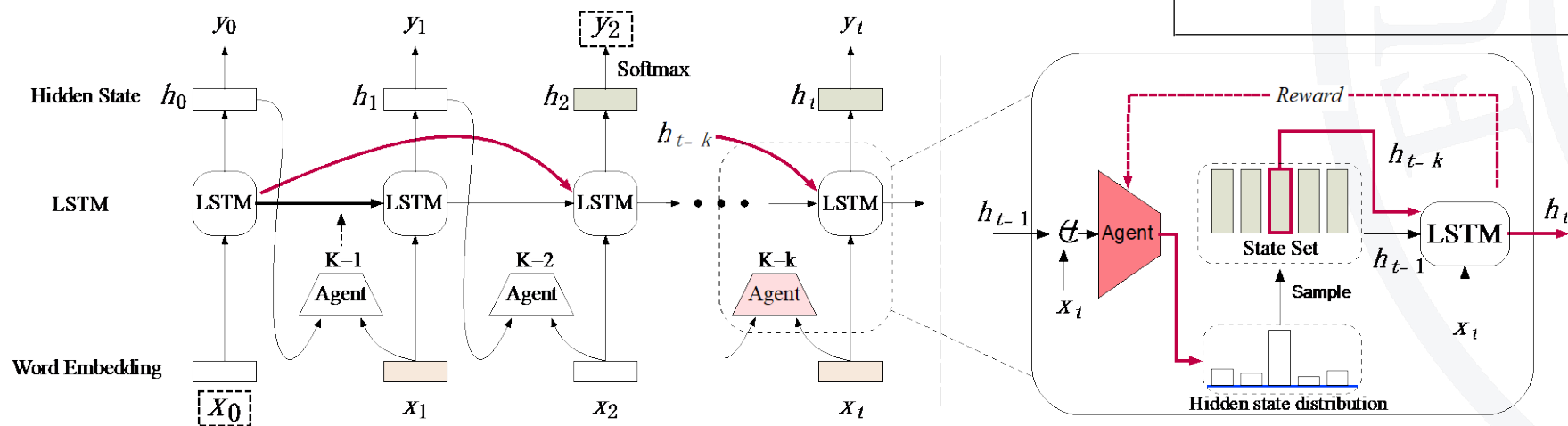
- Convolutional Neural Network (CNN)
并行运行速度快，难建立长距离依赖



Dilated CNN

建模复杂的词语依赖关系-基于动态跳跃的 LSTM

- 句子中词语的依赖关系非常复杂
- RNN 序列建模句子很难拟合复杂的依赖关系



姚明出生于1980年9月12日，上海。

明天特朗普将会晤德克萨斯州的乔斯威尔。

11月6日我参观了创建于1902年的南京大学。

LSTM 计算转移方程时，不只是基于前一个隐状态利用强化学习，在前序的隐状态中挑选最合适的

基于Transformer的上下文编码器

- Transformer (Encoder)

速度快，能直接建立长距离依赖，小数据上稍差

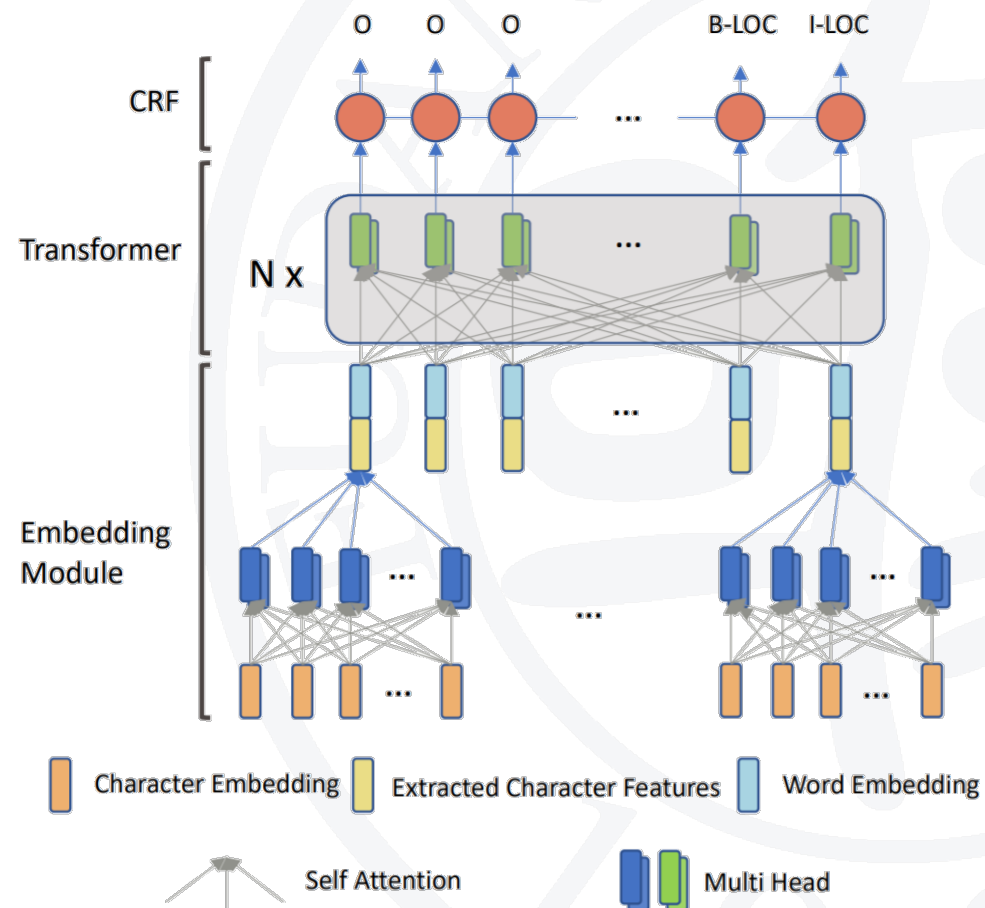
Louis Vuitton founded Louis Vuitton Inc. in 1854 .
PER PER ORG ORG ORG TIME

$$Q, K, V = HW_q, H, HW_v,$$

$$R_{t-j} = [\dots \sin(\frac{t-j}{10000^{2i/d_k}}) \cos(\frac{t-j}{10000^{2i/d_k}}) \dots]$$

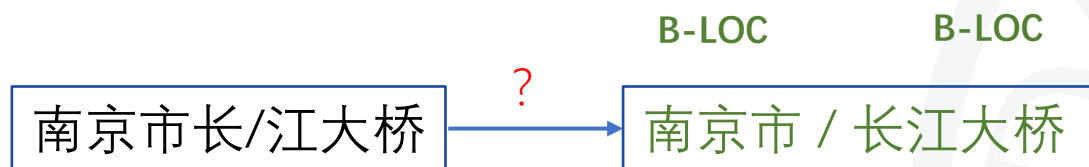
$$A_{t,j}^{rel} = Q_t^T K_j + Q_t^T R_{t-j} + \mathbf{u}^T K_j + \mathbf{v}^T R_{t-j},$$

$$\text{Attn}(Q, K, V) = \text{softmax}(A^{rel})V,$$

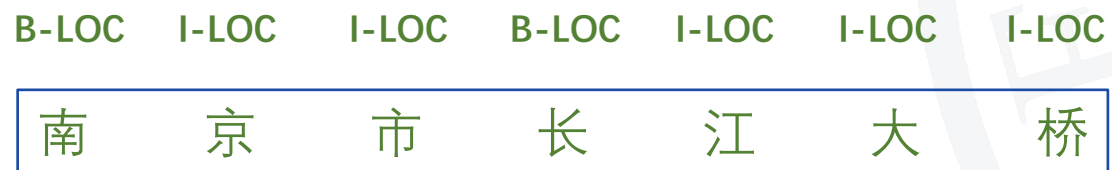


中文NER

分词->NER的pipeline



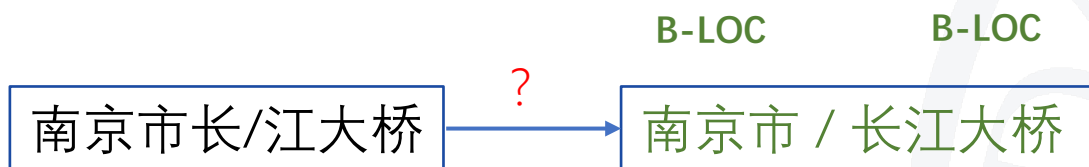
字级别的NER



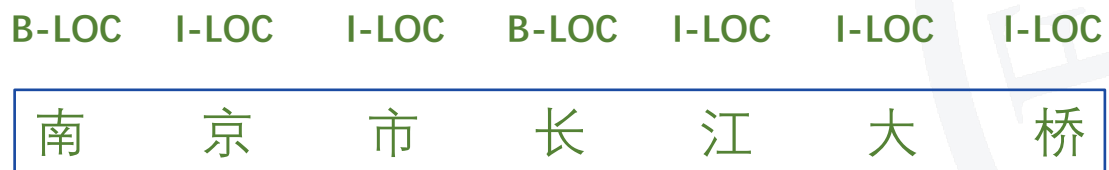
缺少 { 分词的边界信息
词义信息

中文NER

分词->NER的pipeline



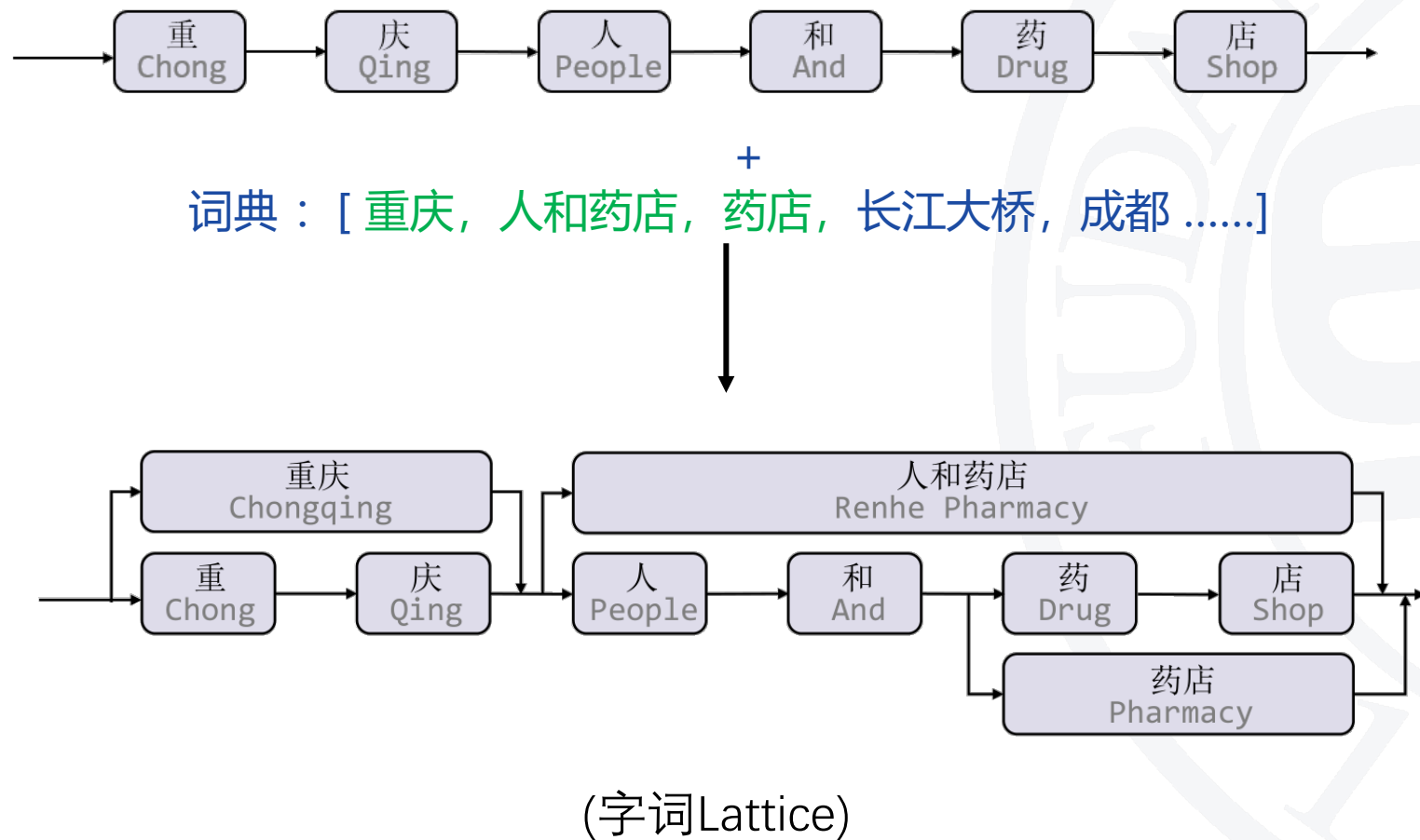
字级别的NER



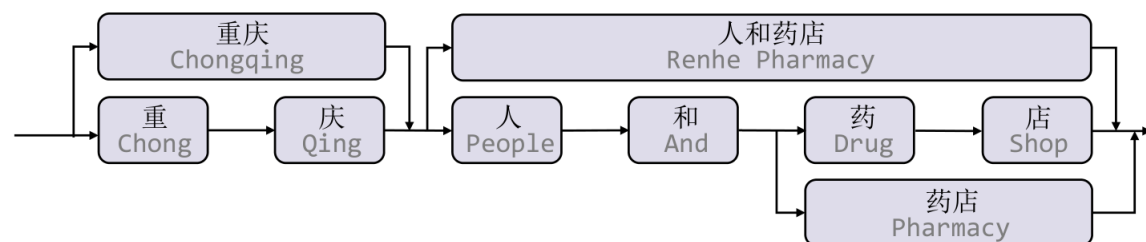
缺少 { 分词的边界信息
词义信息

如何在**不分词**的情况下向**字级别**的序列中加入**词信息**?

词典匹配



词典匹配



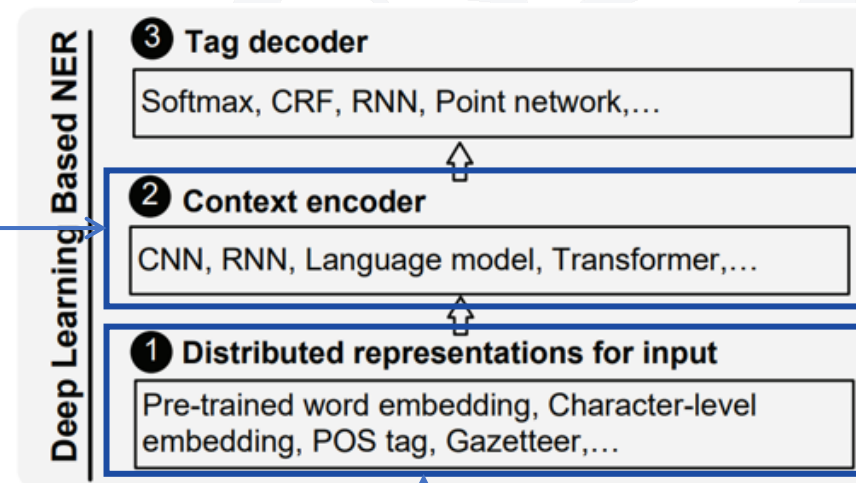
北	0.1	0.6	0.9	-0.2
京	-0.3	0.4	0.5	-0.1
市	0.6	0.7	0.4	0.4

(不用词典匹配信息)

北	0.1	0.6	0.9	-0.2
京	-0.3	0.4	0.5	-0.1
市	0.6	0.7	0.4	0.4

北京	0.3	0.8	0.7	0.5
----	-----	-----	-----	-----

(增加额外的向量,
改动上下文编码器)

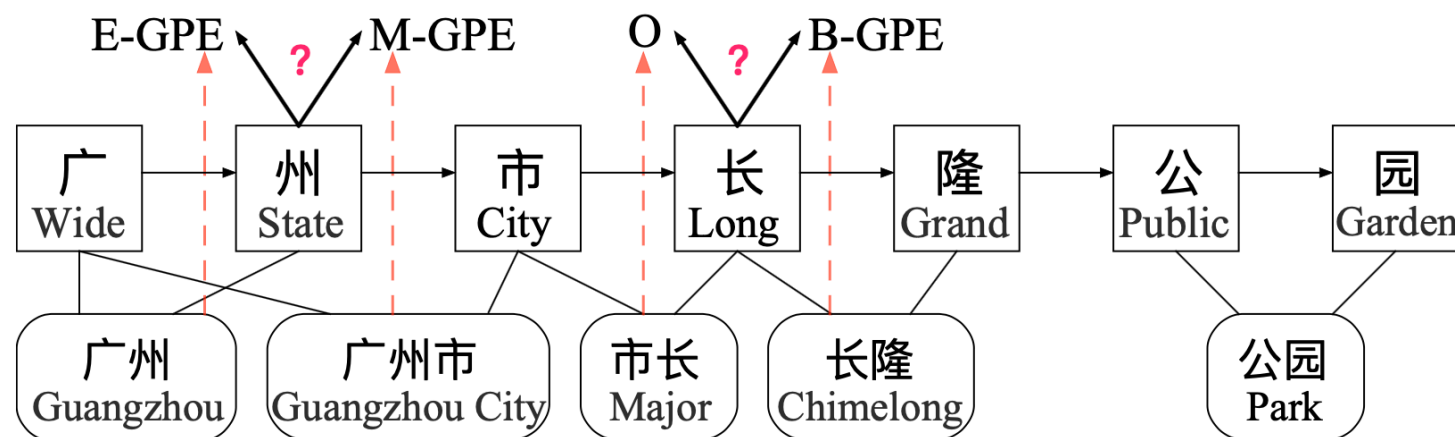


北	0.1	0.6	0.9	-0.2	-0.2
京	-0.3	0.4	0.5	-0.1	0.3
市	0.6	0.7	0.4	0.4	0.8

(在输入表示层做文章)

在CNN中加入反思机制

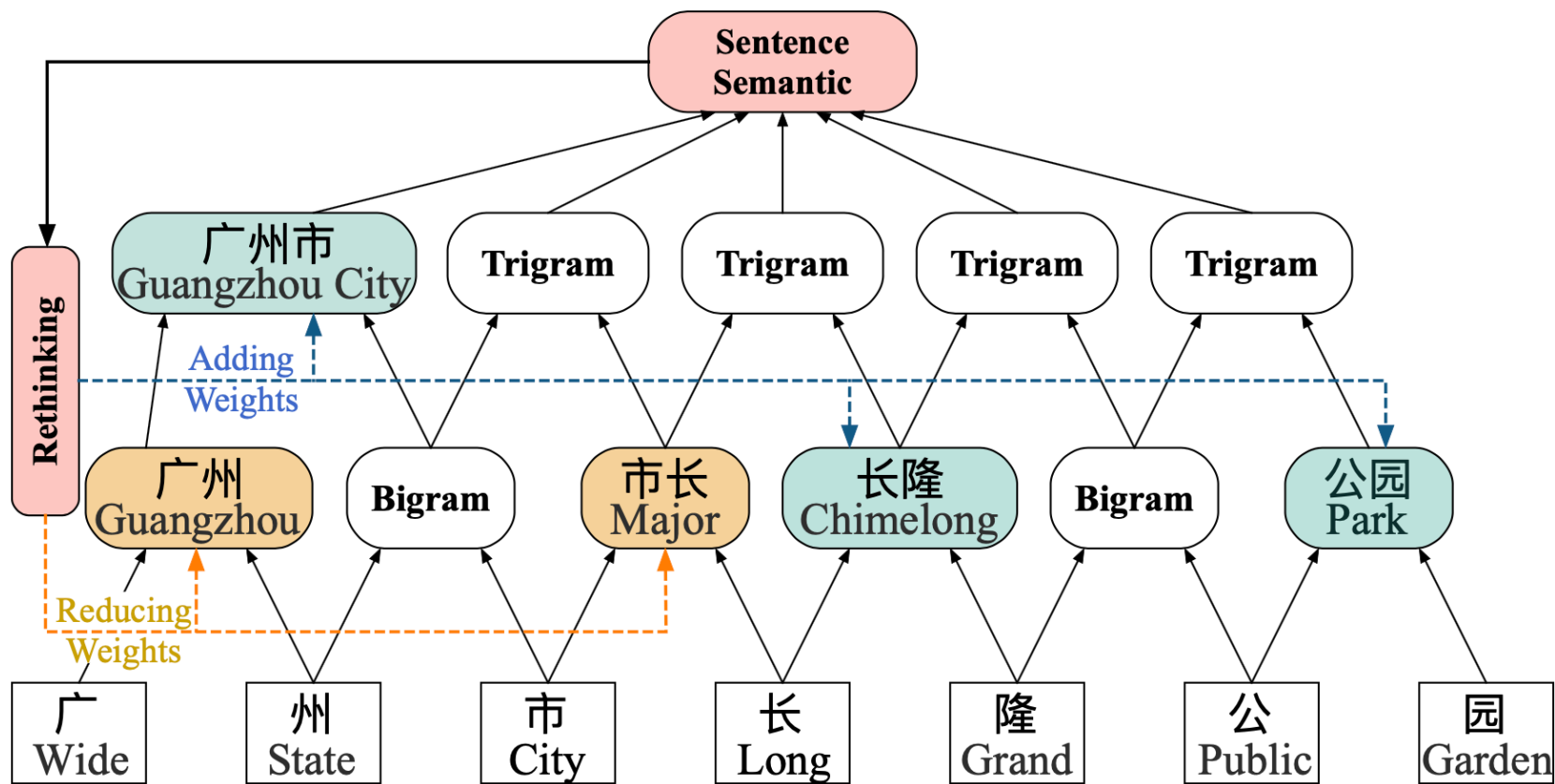
- **研究动机**
 - ✓ 对于结合词典的中文NER，基于RNN的模型由于在时序方向无法并行，效率较低→设计**基于CNN**的中文NER+词典模型，解决并行**效率**问题
 - ✓ 中文NER加入词典时存在**词边界冲突**→引入**rethink机制**，通过高层信息的反馈加以解决



1. 基于RNN的方法需要依次编码整个输入序列，效率较低
2. 引入词典信息时，由于词边界不确定造成的冲突

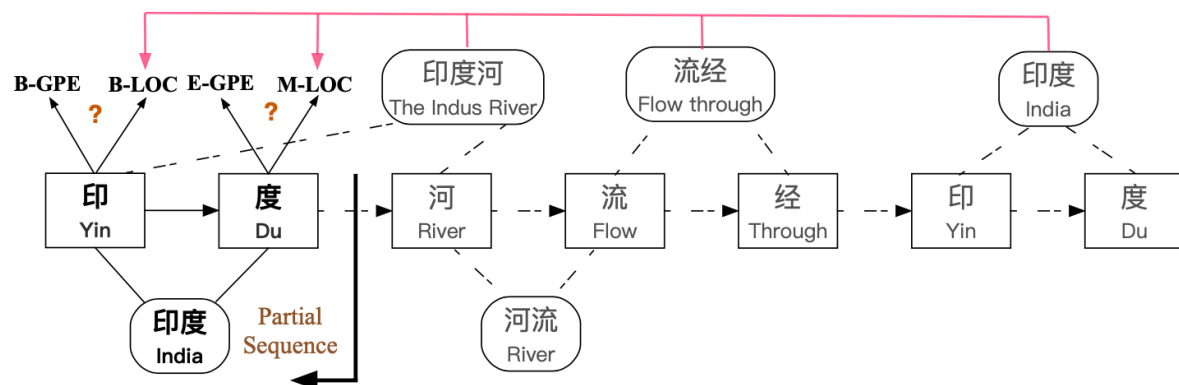
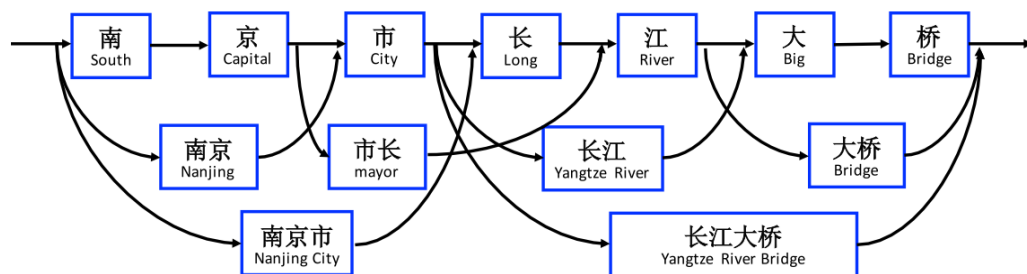
基于 CNN 反思机制的中文 NER

- ✓ 在CNN每一层编码对应长度词语
- ✓ 利用Rethink机制将高层信息反馈至低层，指导低层权重分配



用图结构替代序列结构

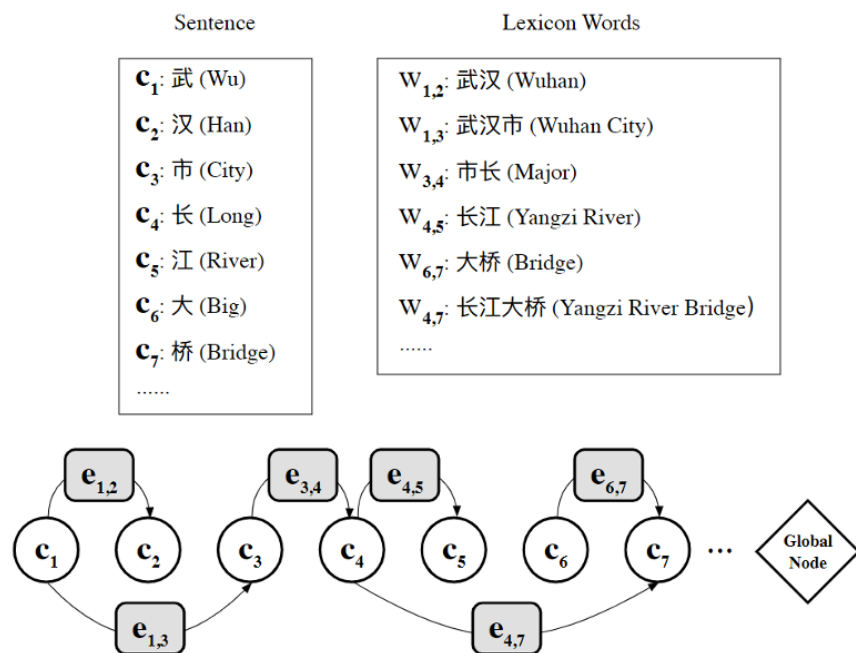
- 研究动机
 - 自然语言不是严格的序列结构
 - 如果缺少上下文和全局信息，基于RNN的模型很难处理词语的组合歧义



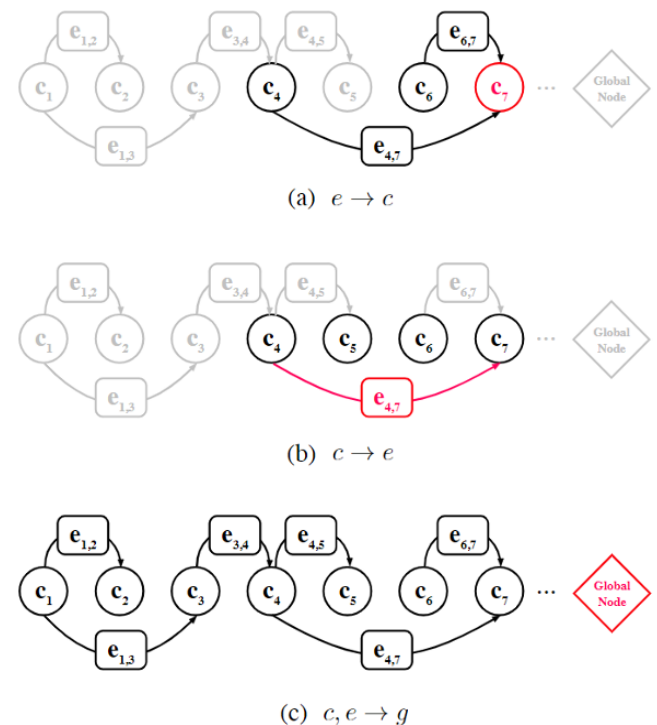
基于图的中文 NER

• 模型

1. 用带有全局结点的GNN来建模句子
2. 利用词典中的词语构造边



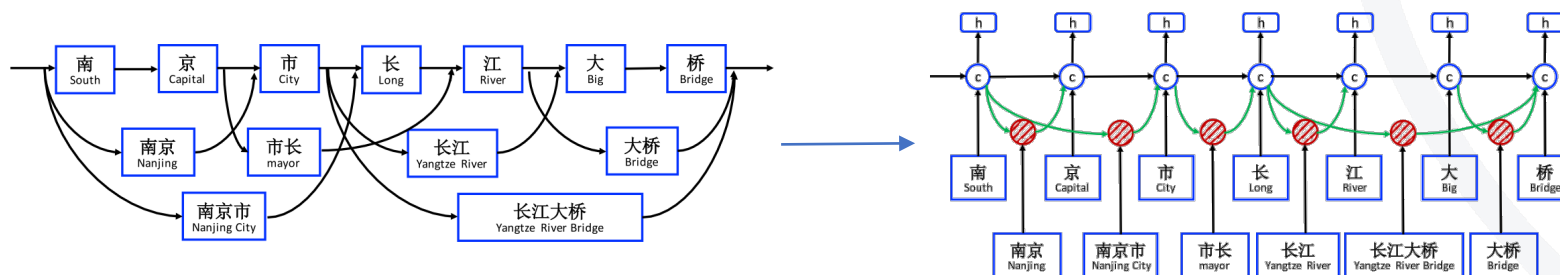
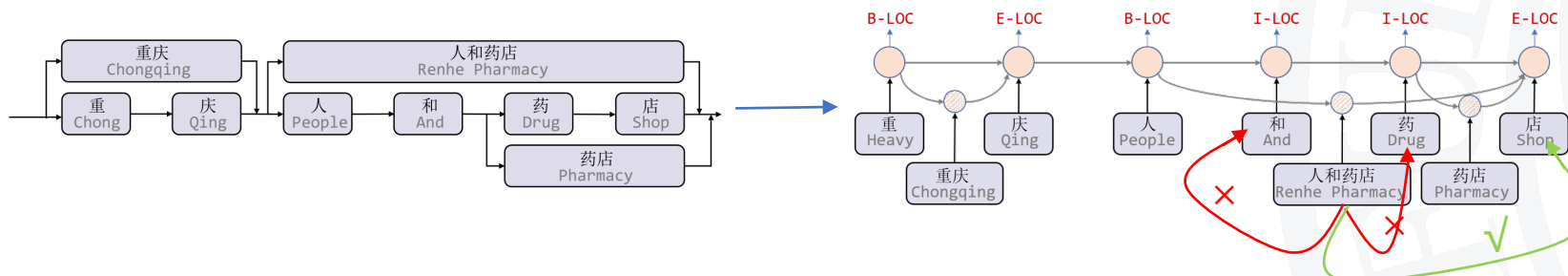
3. 使用消息传递机制 (Aggregate + Update) 更新结点状态



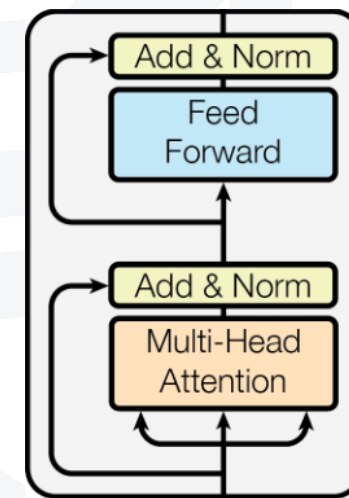
基于Transformer的中文NER

- 研究动机

- 很多中文NER模型需要根据改变自身结构去表征动态的输入结构，这导致运行速度下降
- 之前模型的信息传递效率不高

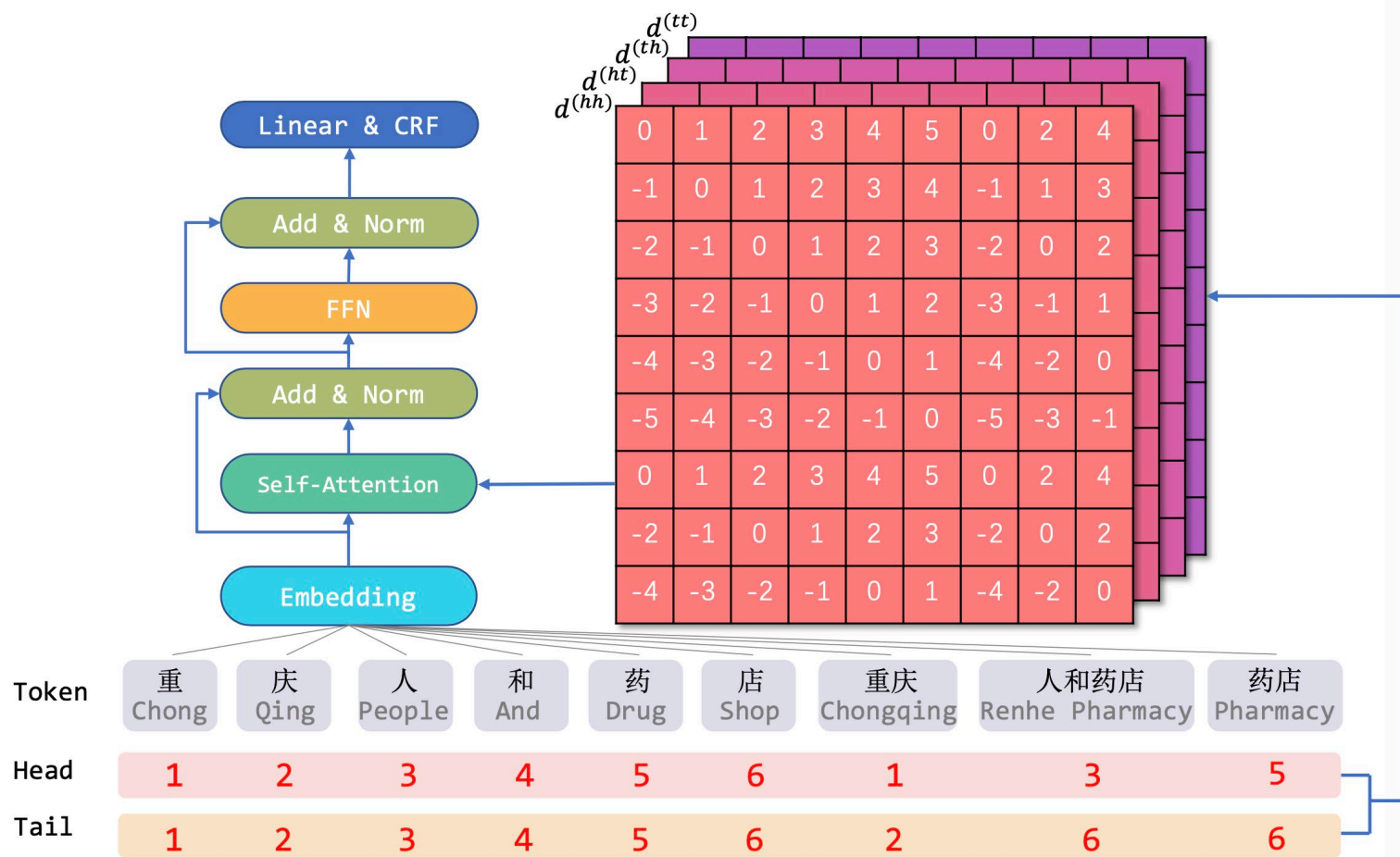


Lattice LSTM



Transformer,
YES!

基于Transformer的中文NER

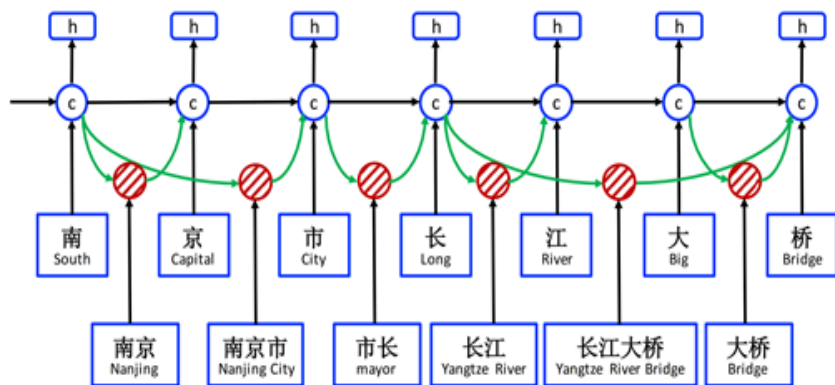


- 将字与匹配到的词同时作为 Transformer 的输入
- 设计了一种针对字词 lattice 的位置表示方法来编码输入的结构
- 两者相结合能达到不需要动态改变模型结构就能表征输入结构的效果，同时信息传递效率非常高

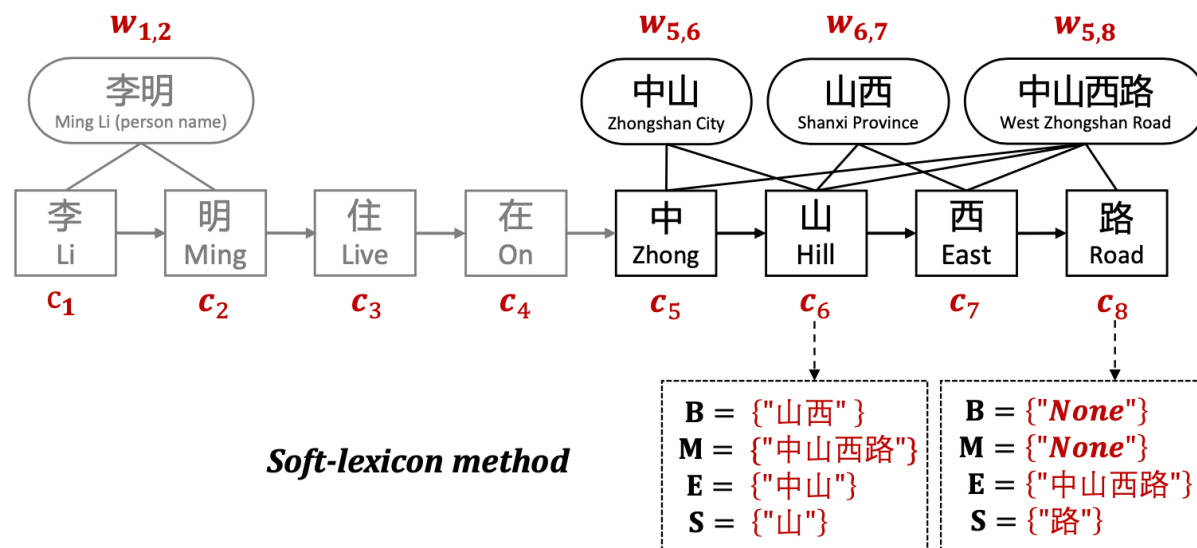
将中文词典信息结合到输入表示层

研究动机

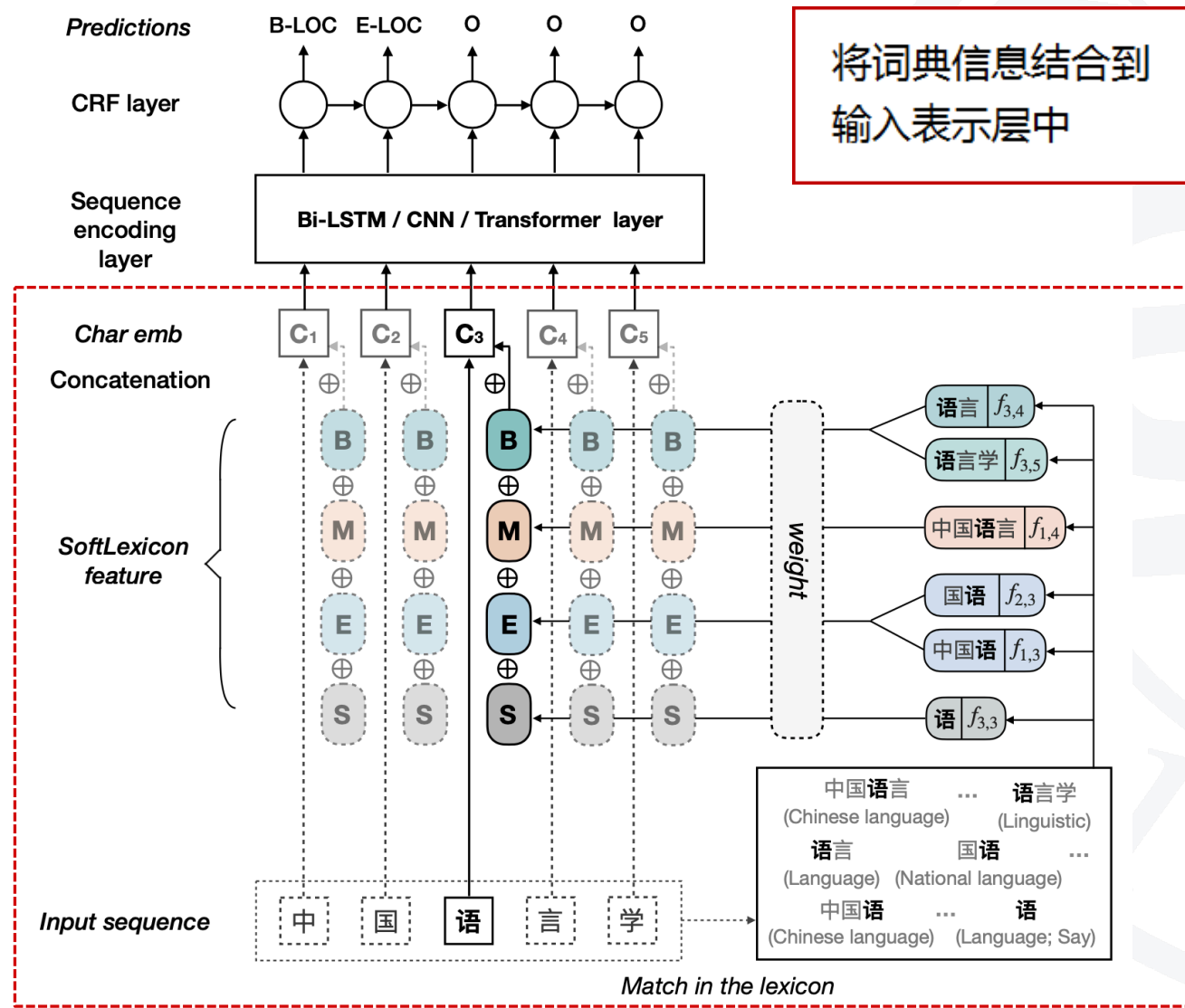
- 以往中文NER结合词典的工作大多关注**上下文编码器**，通过改造模型结构的方法引入中文词典信息
这种方法的问题在于：
 - 上下文编码层结构复杂，模型**解码效率低**
 - 针对特定网络结构（如LSTM/CNN）改进，**无法迁移**
- 解决方案：将中文词典信息引入**输入表示层**，能够大幅提升解码速度，且能够应用于任意神经网络结构



Lattice-LSTM
复杂的模型结构



将中文词典信息结合到输入表示层

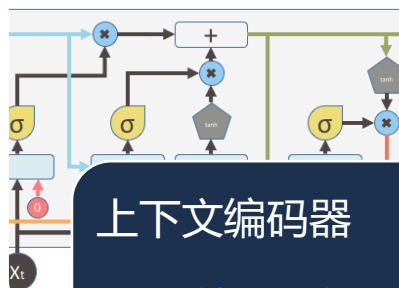


基于深度学习的NER模型



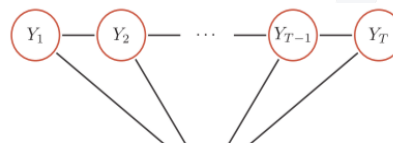
输入表示层

- 词级别
- 字符级别
- 预训练语言模型



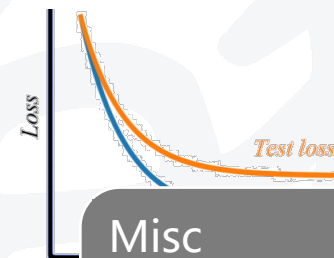
上下文编码器

- 循环神经网络
- 卷积神经网络
- Transformer
- 中文NER



标签解码器

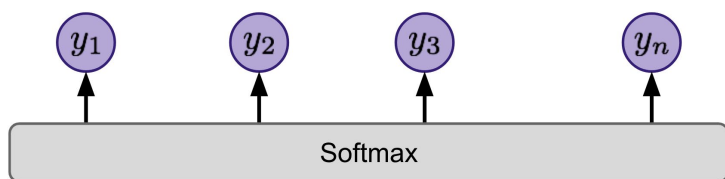
- MLP+softmax
- 条件随机场 CRF
-



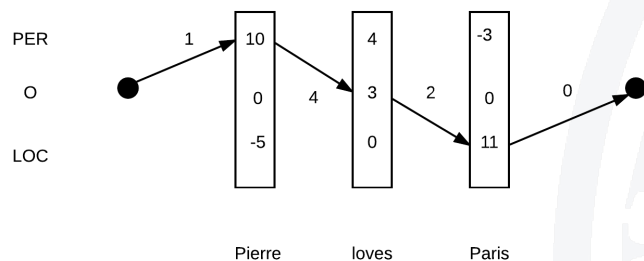
Misc

- 多模态NER
- NER的可解释性
-

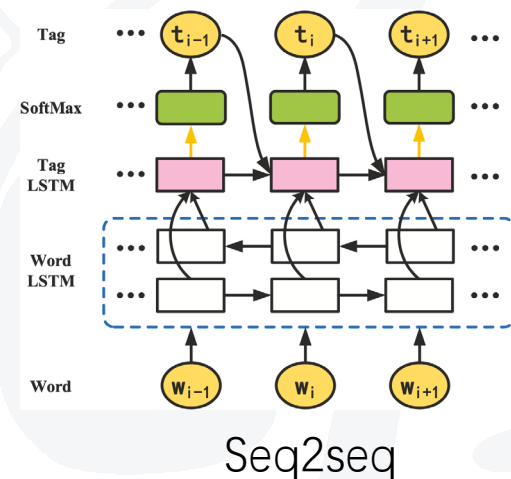
标签解码器



Softmax

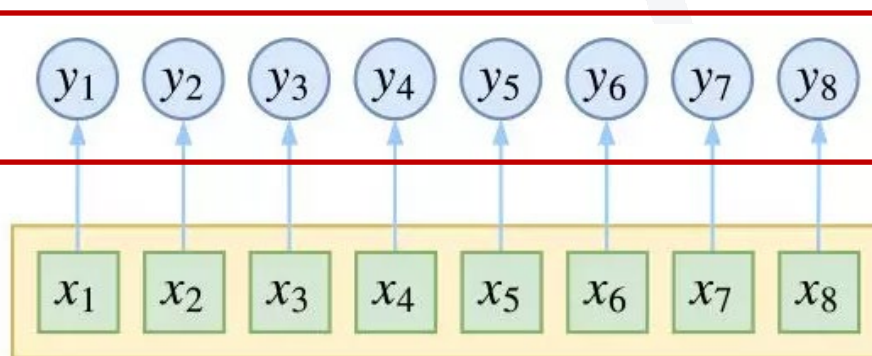


CRF



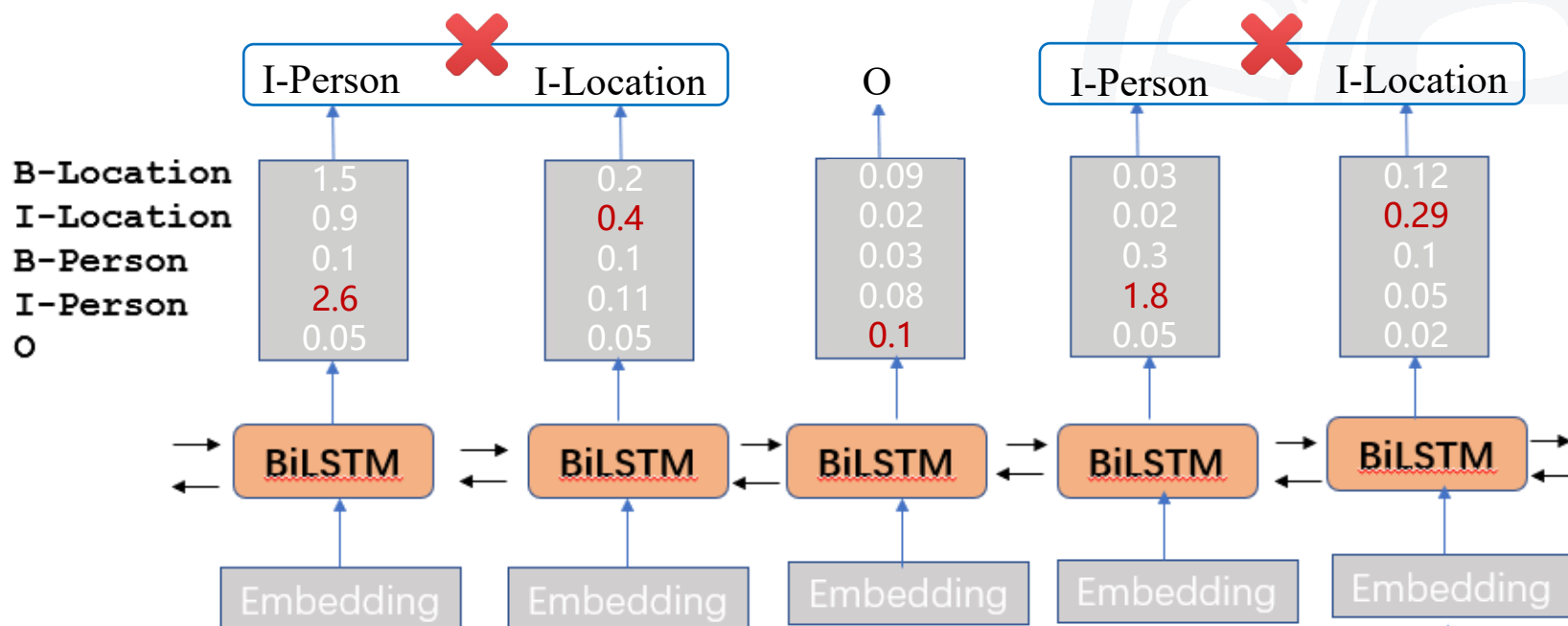
标签解码层

文本编码层



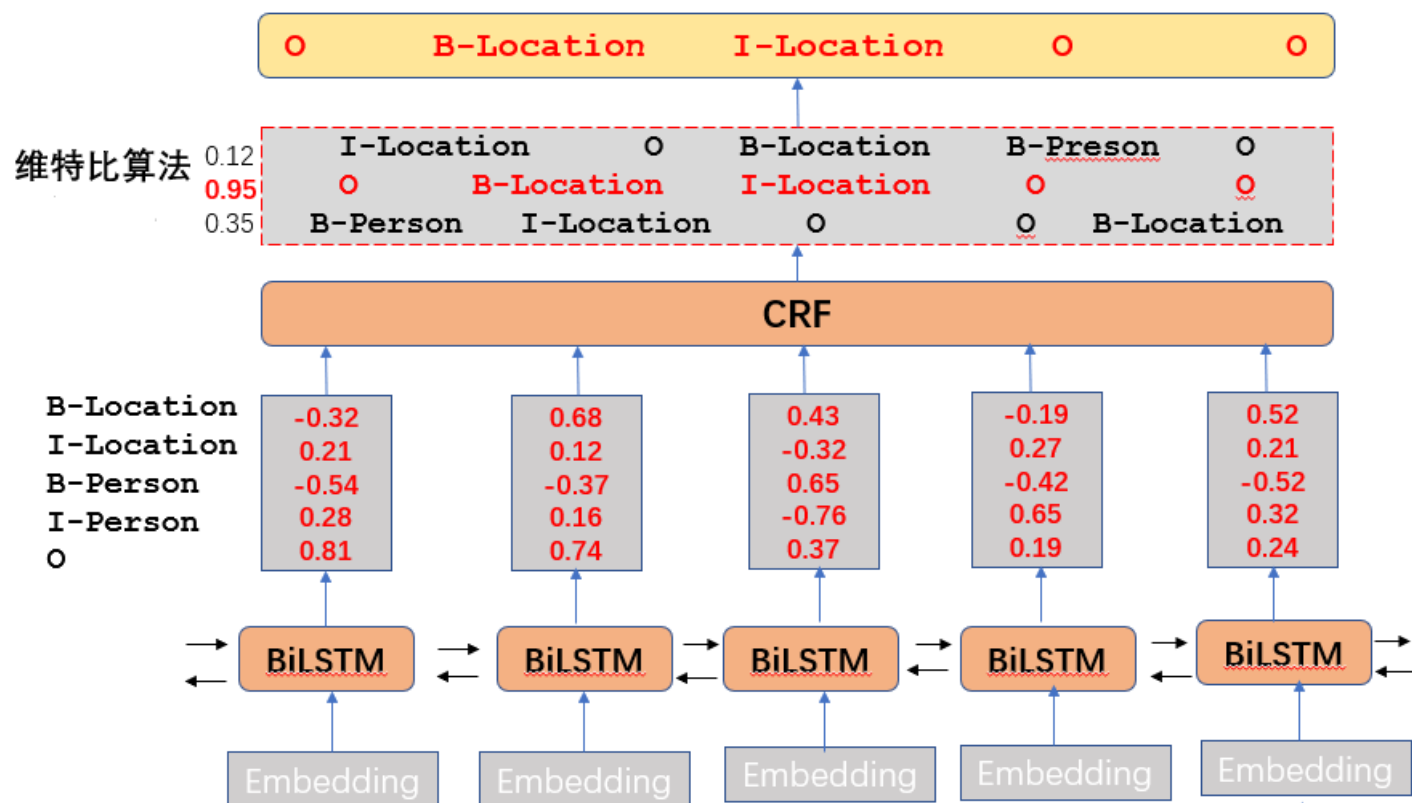
Softmax-标签解码

- 优点: **并行**解码
- 缺点: 无法建立标签间的依赖关系



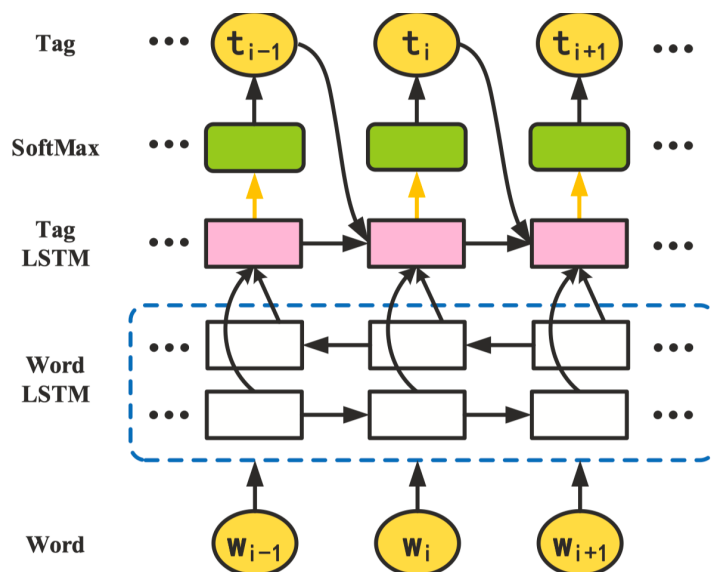
条件随机场-标签解码

- 优点：建模**相邻标签**的依赖关系
- 缺点：**维特比**解码速度较慢



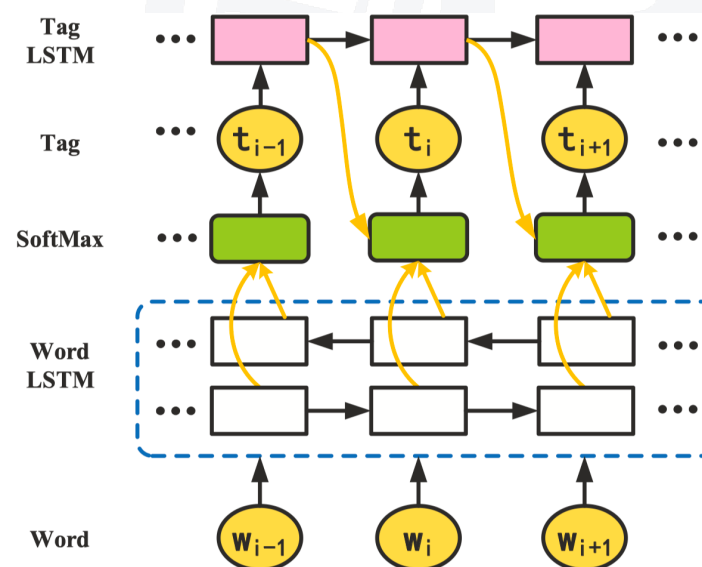
Seq2seq-标签解码

- 优点：建模**长距离**的标签依赖
- 缺点：**错误传播，序列解码**较慢



The₁ credit₂ card₃ you₄ wo₅ n't₆ want₇ to₈ do₉ **without**₁₀ .₁₁
DT NN **NN** PRP MD RB VB TO VB **IN** .

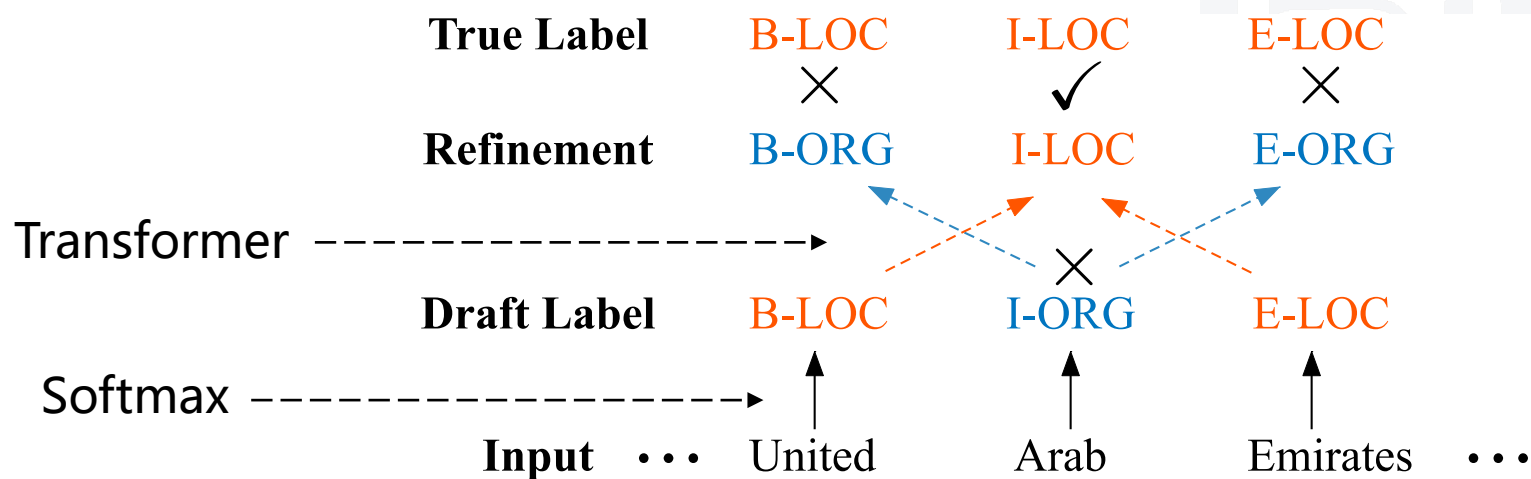
We₁ 'll₂ have₃ to₄ do₅ **without**₆ .₇
PRP MD VB TO VB **RB** .



基于模型不确定性的标签解码

- 研究动机

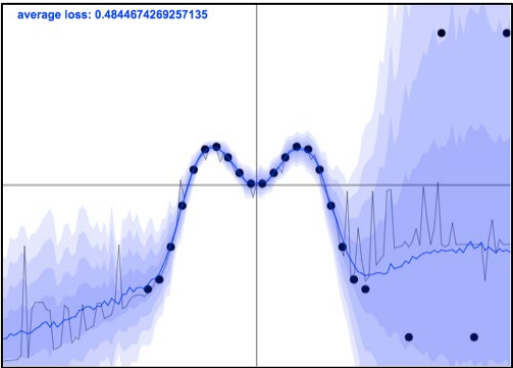
- Softmax 草稿标签, **并行解码**
- Transformer 标签修正, **并行解码**, **长距离依赖**, **错误传播**



基于模型不确定性的标签解码

- 研究动机

- Softmax 草稿标签, **并行解码**
- Transformer 标签修正, **并行解码, 长距离依赖**
- 贝叶斯模型不确定性**防止错误传播**

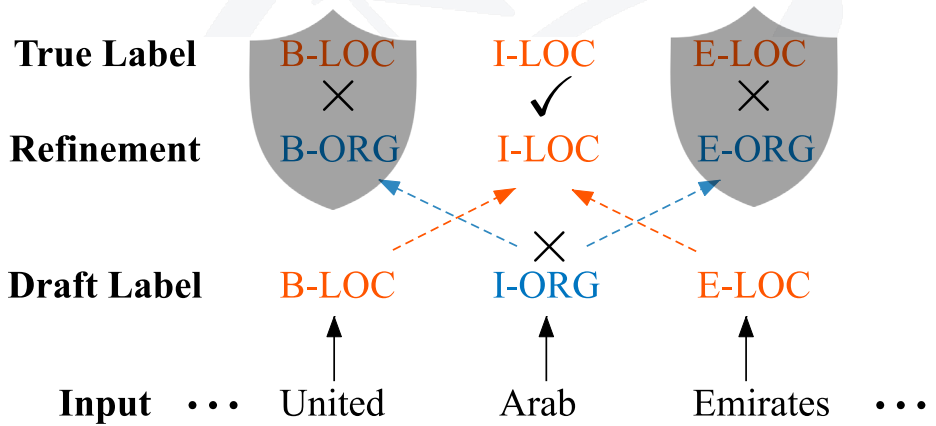


Bayesian NNs for
Uncertainty Estimation



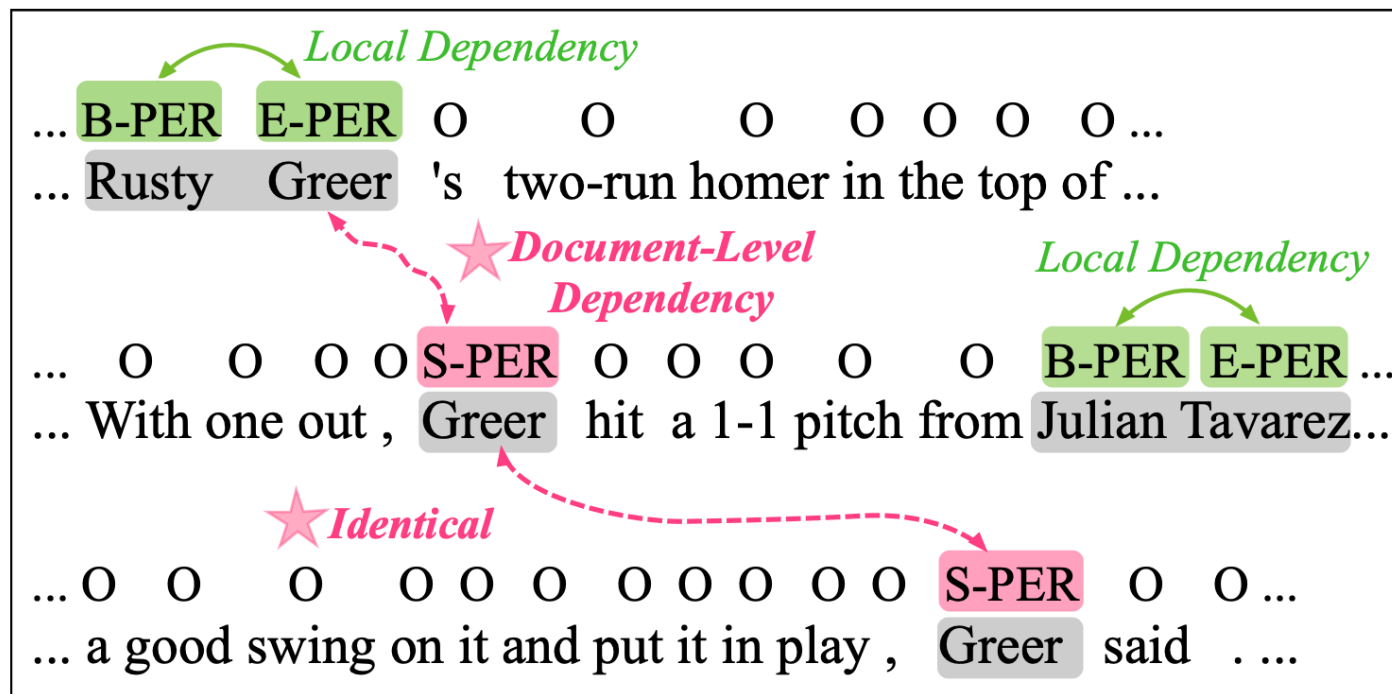
Draft	Uncertainty	Refinement	#Tokens
✓	0.018	✓ → ✗	39
✗	0.524	✗ → ✓	54

Table 1: Results of LAN with uncertainty estimation evaluated on CoNLL2003 test dataset. ✓ refers to the correct prediction, and ✗ refers to the wrong prediction.



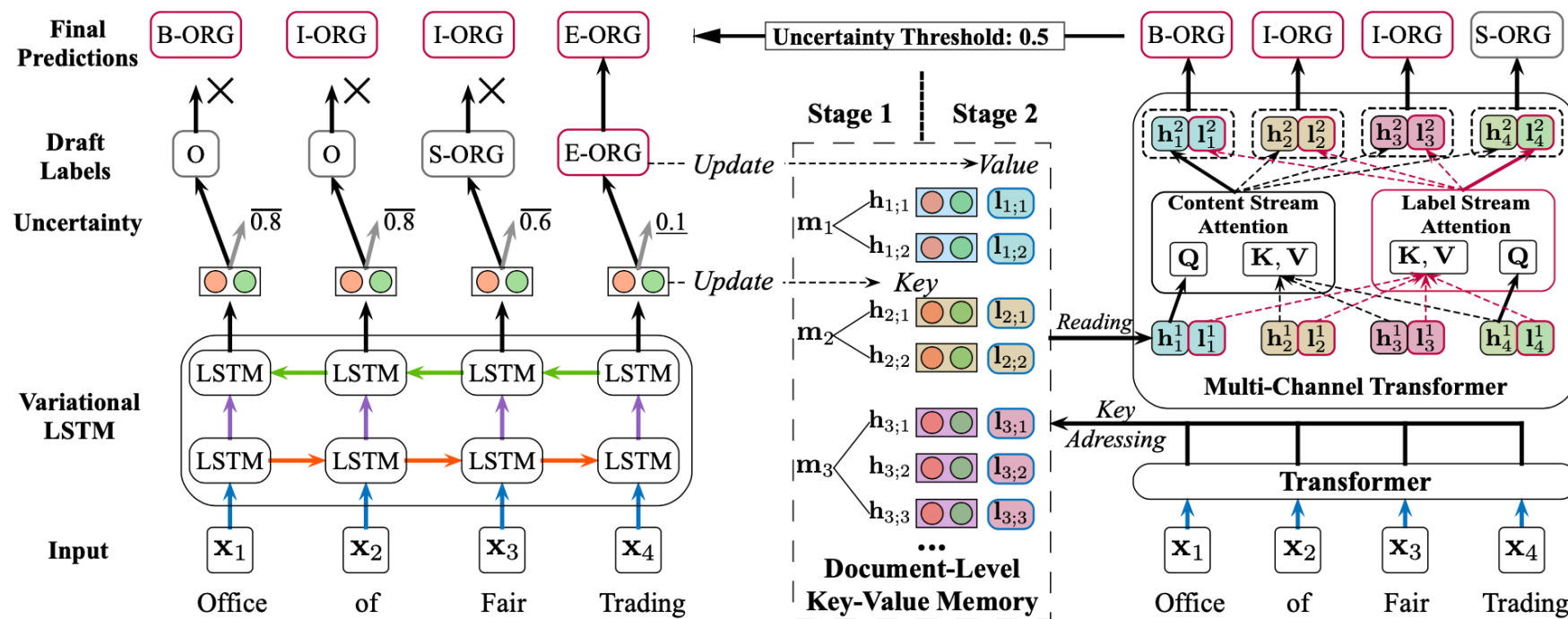
文档级的NER

- 研究动机
- ✓ 以前工作没有显式建模文档级别的标签依赖关系
- ✓ CRF只能建模局部依赖，无法建模句子以及文档级别的长距离依赖



文档级的NER: two-stage模型

1. 用Bayesian Neural Networks衡量第一阶段预测的不确定性
2. 用memory network保存文档级别tokens的上下文和草稿标签信息
3. 用Two-Stream Transformer结合第一阶段的预测和文档级别信息进行并行refine
4. 设置阈值进行两阶段结果的整合

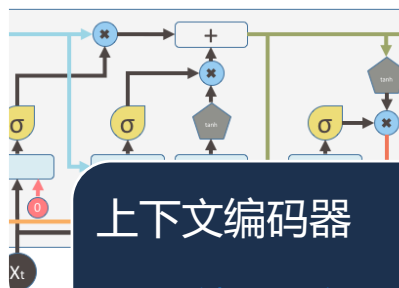


基于深度学习的NER模型



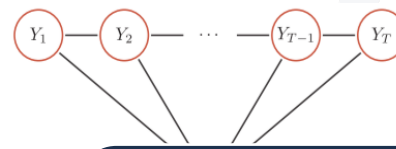
输入表示层

- 词级别
- 字符级别
- 预训练语言模型



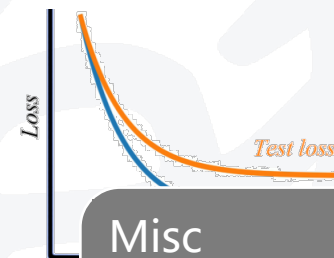
上下文编码器

- 循环神经网络
- 卷积神经网络
- Transformer
- 中文NER



标签解码器

- MLP+softmax
- 条件随机场 CRF
-



Misc

- 多模态NER
- NER的可解释性
-

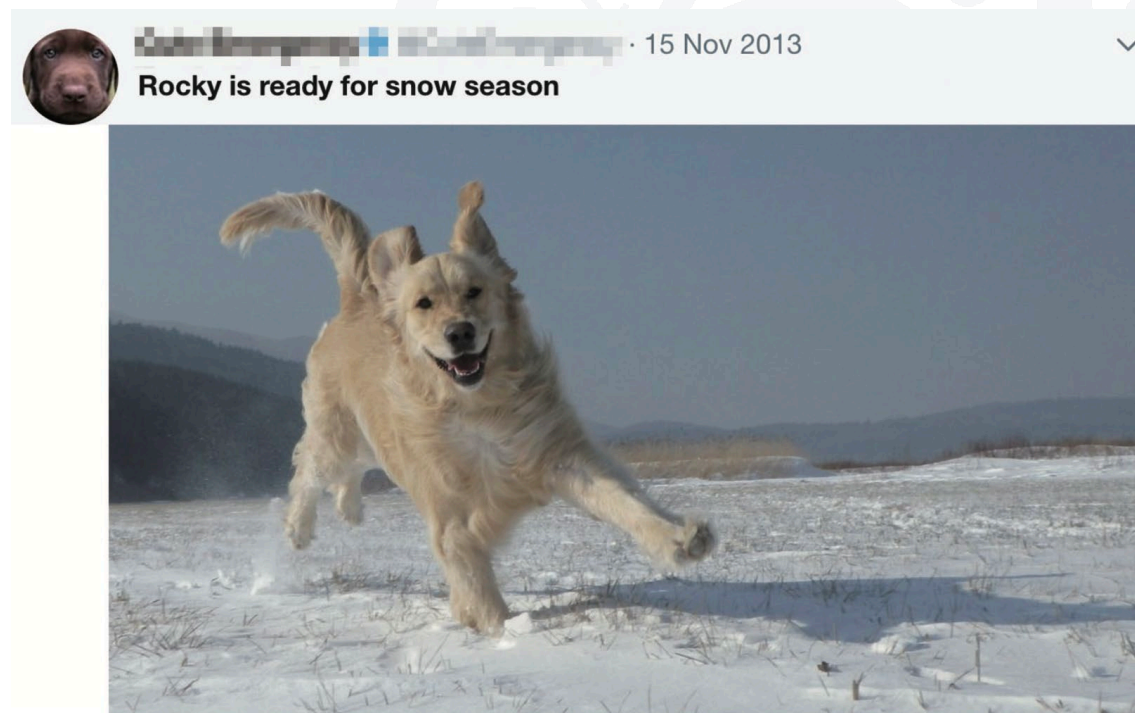
多模态命名实体识别

•研究动机

- ✓ 图片信息在NER任务中能有效起到实体消歧的作用
- ✓ 图片信息却很少在NER任务中被考虑，目前并不存在标注的多模态NER数据集

•主要贡献

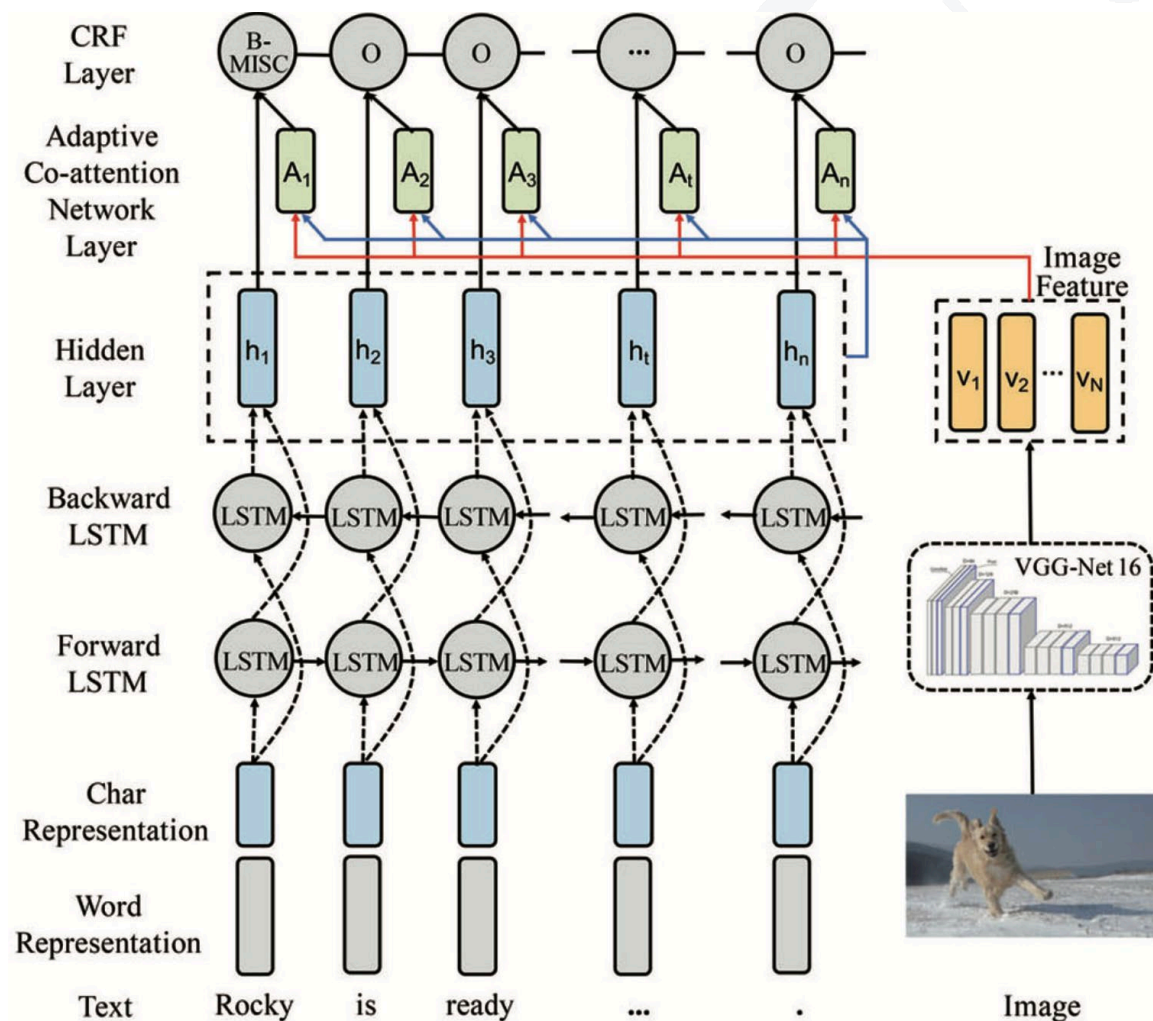
- 最先考虑多模态命名实体识别
- 标注了多模态命名实体识别数据集
- 提出有效的融合图片和文本信息的框架



多模态推文的例子。"Rocky"是狗的名字(MISC)，然而只考虑文本信息，通常会将"Rocky"识别为人名(PER)。

多模态命名实体识别

- 模型:
 - 基于 CNN+Bi-LSTM+CRF 框架
 - 用 VGG-Net 提取图片特征
 - 加入 adaptive co-attention network, 来融合文本特征和图片特征



反思神经模型的泛化能力：以命名实体识别为例

What should we do in the post-BERT era?



Rethink the **Generalization** of Neural NER Models



Diagnose the **bottleneck** of existing NER system

反思神经模型的泛化能力：以命名实体识别为例

问题一 我们的模型确实具有泛化能力，还是只是假装理解并进行了一些浅层次的模板匹配？

指标1 **实体覆盖率** 测试集中的实体在训练集中，并被标记为相同类型的比例（ECR）

标注错误的检测及修复

Model	CoNLL	ReCoNLL
(Devlin et al. 2018)	92.80	-
(Peters et al. 2018)	92.22	-
(Akbik, Blythe, and Vollgraf 2018)	93.09	-
(Akbik, Bergmann, and Vollgraf)	93.18	-
Our Implementation		
GloVe	89.10	89.85
ELMo	91.79	92.65
BERT	91.34	92.16
FLAIR	93.03	93.78

在CoNLL 2003 及其修复集上的测试性能 (F1)

细粒度评估测试集

Datasets	Embed-layer		Entity Coverage Rate					
	Char	Word	Overall	1	(0.5, 1)	(0, 0.5]	$C \neq 0$	$C = 0$
CoNLL	CNN	-	76.42	79.94	86.99	78.84	69.74	77.61
	FLAIR	-	89.98	95.30	95.58	82.39	72.16	90.39
	ELMo	-	91.79	97.61	95.98	85.15	71.43	92.22
	BERT	-	91.34	97.72	95.17	86.66	77.83	92.37
	-	Rand	78.43	95.05	94.75	73.54	37.97	66.40
	-	GloVe	89.10	98.44	96.31	81.34	57.80	87.23
	CNN	Rand	82.88	94.13	94.48	74.25	47.78	78.91
	CNN	GloVe	90.33	98.32	95.94	80.33	59.67	89.74
	ELMo	GloVe	92.46	98.08	96.46	86.14	69.79	93.08
	FLAIR	GloVe	93.03	98.56	96.38	87.07	73.58	93.42
WNUT	CNN	-	20.88	45.99	67.01	40.25	19.14	19.74
	FLAIR	-	41.49	81.15	88.14	54.36	39.56	43.44
	ELMo	-	43.70	88.72	90.83	55.56	44.19	43.32
	BERT	-	44.08	77.75	81.61	49.74	34.65	41.92
	-	Rand	14.97	60.62	83.84	50.00	3.90	4.77
	-	GloVe	37.28	89.29	92.62	45.65	35.34	35.15
	CNN	Rand	22.29	48.88	71.43	39.08	16.75	18.83
	CNN	GloVe	40.72	86.12	92.24	49.74	26.67	40.06
	ELMo	GloVe	45.33	90.38	89.92	56.57	37.8	46.58
	FLAIR	GloVe	45.96	90.52	89.92	61.69	42.07	48.38

在CoNLL和WNUT细分数据集上的测试性能
(采用不同的预训练策略)

反思神经模型的泛化能力：以命名实体识别为例

问题二 数据集的哪个因素可以区分泛化良好的神经网络和泛化一般的神经网络？

指标2: 实体覆盖率期望 (EECR)

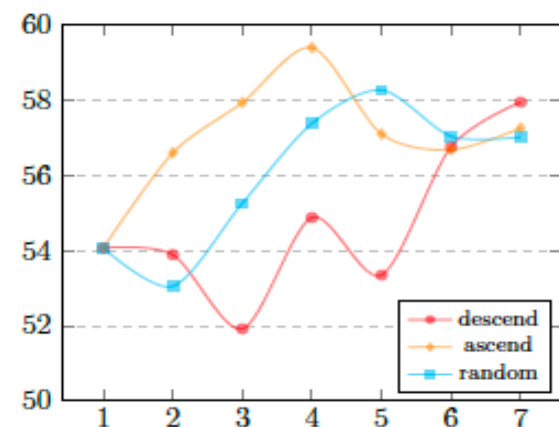
指标3: 实体上下文覆盖率 (CCR)
属于同一类型但分别来自训练和测试集的实体的上下文相似性

杂交数据集的泛化分析

Matrix	Train	WN.	Co.	BN	WB	MZ	NW	BC	P-row
M_{F1}	WN.	46.6	16.7	12.0	13.8	11.4	6.50	11.7	0.98
	Co.	22.2	70.2	19.4	17.5	12.7	19.8	17.3	0.93
	BN	24.4	35.2	65.7	35.4	28.6	36.2	42.3	0.93
	WB	19.6	25.6	28.0	55.7	17.3	28.3	25.8	0.83
	MZ	15.3	25.0	29.7	21.8	67.8	32.2	26.6	0.89
	NW	22.2	24.2	32.6	29.3	24.6	68.4	26.9	0.87
	BC	28.6	29.1	41.1	43.6	25.5	33.2	70.3	0.90
P-col		0.93	0.97	0.97	0.83	0.96	0.94	0.96	0.88
M_{ρ}	WN.	1.00	0.26	0.29	0.21	0.07	0.24	0.31	0.96
	Co.	0.325	1.00	0.35	0.37	0.29	0.36	0.37	0.94
	BN	0.35	0.33	1.00	0.49	0.32	0.45	0.65	0.95
	WB	0.27	0.22	0.43	1.00	0.29	0.34	0.52	0.91
	MZ	0.15	0.15	0.24	0.15	1.00	0.31	0.23	0.99
	NW	0.20	0.27	0.38	0.34	0.34	1.00	0.41	0.92
	BC	0.28	0.22	0.52	0.49	0.35	0.36	1.00	0.87
P-col		0.95	0.95	0.91	0.89	0.97	0.92	0.92	0.78
M_{ϕ}	WN.	1.00	0.06	0.08	0.11	0.07	0.03	0.12	0.99
	Co.	0.176	1.00	0.20	0.35	0.20	0.13	0.27	0.91
	BN	0.19	0.22	1.00	0.56	0.34	0.18	0.65	0.88
	WB	0.29	0.26	0.41	0.88	0.42	0.23	0.70	0.70
	MZ	0.32	0.28	0.43	1.00	1.00	0.33	0.68	0.57
	NW	0.36	0.22	0.42	0.90	0.53	1.00	0.69	0.70
	BC	0.26	0.22	0.40	0.52	0.32	0.19	1.00	0.90
P-col		0.85	0.97	0.96	0.45	0.93	0.93	0.83	0.89

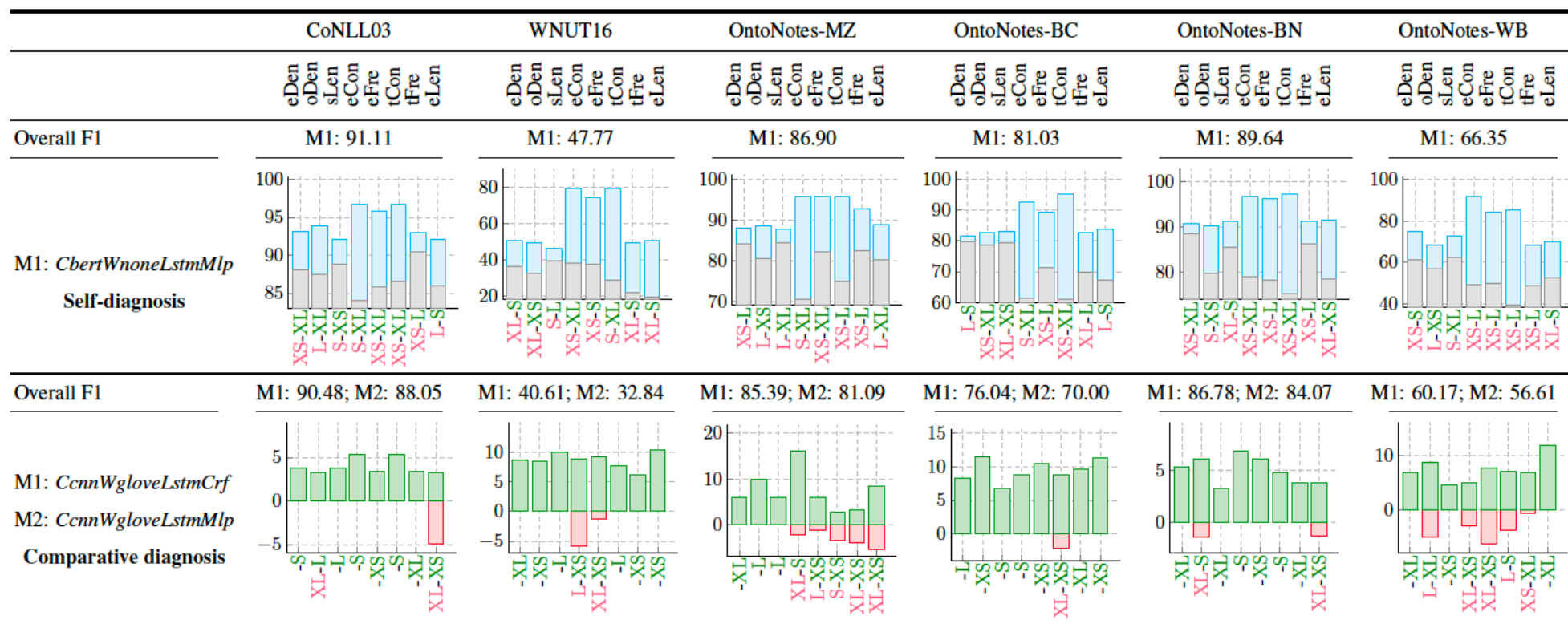
跨领域设置下的F1 score, EECR, 及CCR

数据排列问题



以三种不同的顺序引入源域之后F1-score的变化
EECR分数降序（红色），升序（橙色）和随机顺序（蓝色）

命名实体识别的可解释评估



对不同的NER系统进行自诊断和互诊断

| 总结&展望

- 世界知识对于实体识别有重要意义，将世界知识引入词表示学习，并加强未登录词的表示至关重要
- 现有模型在获取短文本和非规范文本的语义表示上仍有不足，利用多模态信息等外部知识和迁移学习是可行方案
- 嵌套实体、非连续实体在医学、法律等领域广泛存在，设计统一的标签解码方案是待解决的问题
- 可解释性评估是我们理解模型的工作原理，工作机制的基础，可以帮助融合多个模型的不同优势，对不同的应用需求设计针对性的模型方案

Thank You for Attention