专用命名实体抽取

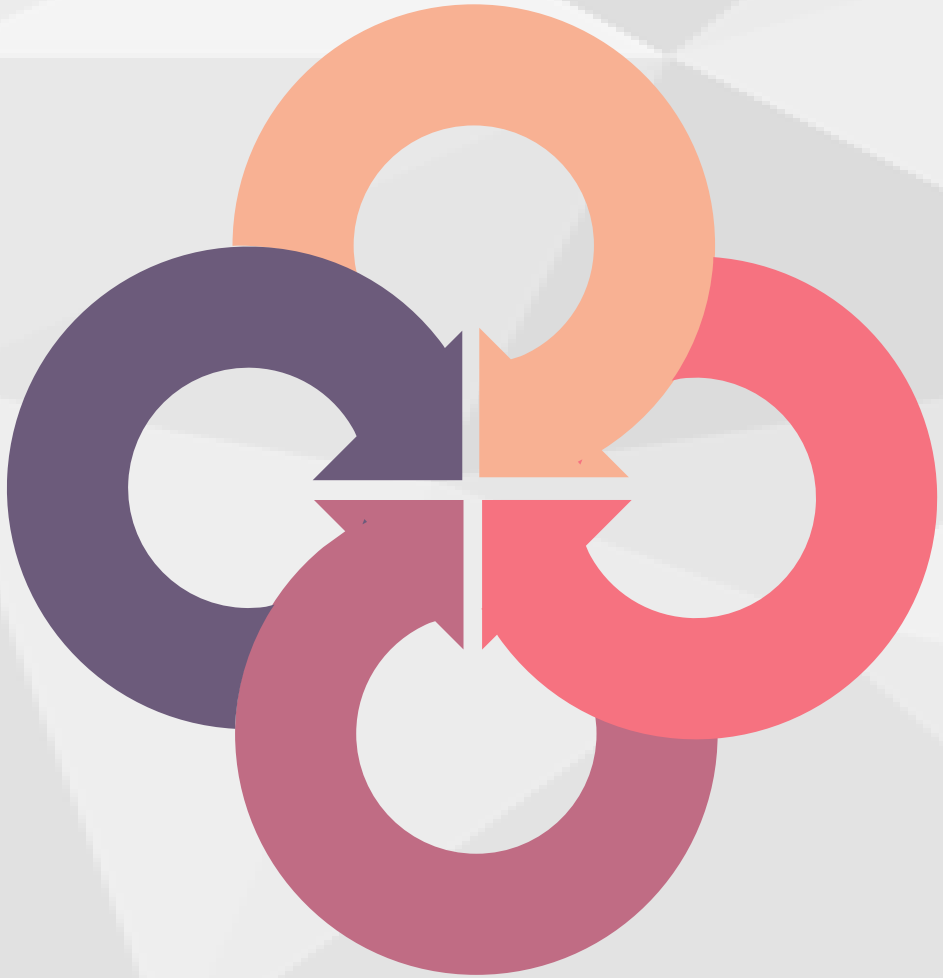
组员：毛禹川
沈潇雨
郑灵杰
侯子睿
陆皓霖

CONTENTS

1 引言

毛禹川

定义——专用命名实体



命名实体

以名称为标识的实体
人名、地名、组织机构名.....

专用

强调专用性
某个领域中特有的命名实体

举例

医疗领域
上呼吸道感染 感冒

打开思路

专用命名实体其实就在我们身边

定义——抽取

NER：Named Entity Recognition，即命名实体识别，是自然语言处理的一项基础任务，是众多自然语言处理任务的重要基础工具。

专用命名实体抽取所使用的技术与NER所使用的技术几乎完全一样，只不过是在命名实体的目标范围上有所区别。

专用命名实体抽取强调命名实体的专用性，比命名实体识别任务具有更高的难度。

定义——正确识别

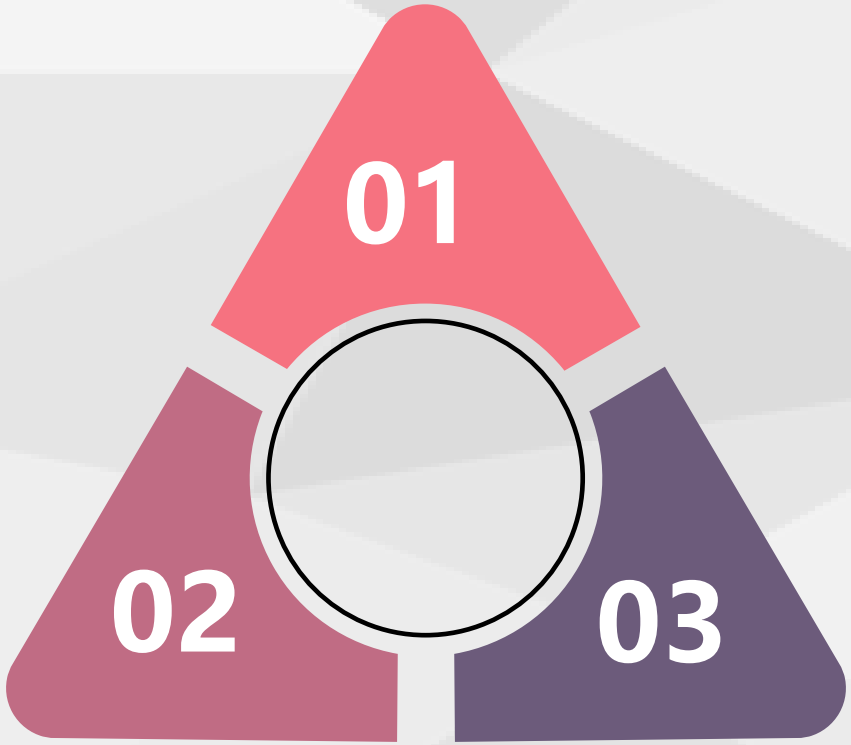
实体的边界是否正确

实体的类型是否标注正确

一个专用命名实体抽取场景

当O2体积分数位于区间(ϕ(O2)<0.4%)，SOF4产气量随着O2体积分数的增加而缓慢增加；结合ArcGIS统计分析工具，采用当前比较成熟的大数据分析技术，实现了对覆冰的趋势分析、与设计冰厚对比分析及预警；
月曰`补〕}笋、小瓷套问隙U(kV)工放电压31人V(11卜)工放电压2了.skV阀终尸100夕〔屯~~卜~,一一.~目`.~一工一一山~~山一么一I主11·毛18rZ一15划夕节二价t(声5)图3图5了O高阻值井联电刚J.J日Ut}(1、V雇小走人入并琢电阻U(kV)享户下一一福1,`1418t吸琳5)异阀片101:18t切5)图〔宝图8U(kV)护瓷弃/U(kV)_于冒毖一,毛高准值一许状1七吐户刁片30l0艘卜二`二二`三一`足一。
由以上分析可知，壁温越高，溶液所对应的饱和溶解度越低，浓度差越大，沉积率越大，污垢单位面积的沉积速度也就越快。
节点之间以对应的热阻相连，形成了定子铁心段基本单元网络(见图1，其中热网络节点之间连线表示热阻)。
图7A区控制脉冲仿真Fig.7SimulationofcontrolpulseinsectionA表3为采用逻辑法和计数法完成上述脉冲发生时所占用的CPLD资源情况。
t????t1时刻电阻降到2??，进入燃弧阶段。

挑战



歧义识别

如果文本有两种或两种以上的解释方式，就会产生歧义，但在特定语境中只有一种解释是正确的。

新词识别

面对训练语料库中没有的新命名实体，NER难以将其识别出来

领域适应

目前命令实体识别只是在有限的领域和有限的实体类型中取得了较好的成绩。但这些技术无法很好地迁移到其他特定领域中

挑战——歧义识别（英语）

MAC

MAC是一个多义词，请在下列义项上选择浏览（共15个义项） | 添加义项 +

- **苹果电脑**
- 最高容许浓度
- 加拿大彩妆品牌
- 使命召唤4职业玩家
- 麦康凯琼脂培养基
- 手动报警站
- 多址接入信道
- 平均气动弦
- 特摄剧《雷欧·奥特曼》中的地球防卫队
- 膜攻击复合物
- 消息认证码（带密钥的hash函数）
- 重大不利改变条款
- 最低肺泡有效浓度
- 介质访问控制层
- 多原链

收起∧

挑战——歧义识别（汉语）

原句

货拉拉拉不拉拉布拉多?

英语

Can Lalamove pick up Labrador?

汉语分词

货拉拉/拉不拉/拉布拉多

命名实体：

货拉拉：公司 拉布拉多：狗的品种

挑战——新词识别

01 元宇宙
——科技领域

02 非同质化代币
——金融领域

03 逆全球化
——政治领域

04 游戏人生
——人文领域

挑战——领域适应

一方面，由于特定领域资源匮乏造成训练语料缺失，导致模型训练很难直接开展。

另一方面，由于不同领域的专用命名实体往往不同，训练好的语言模型无法很好地在领域之间迁移。

数据集

CoNLL 2003

语种：英语、德语
实体类型：人名、地名、机构名、其他
数据来源：新闻

CoNLL 2002

语种：西班牙语
实体类型：人名、地名、机构名、其他
数据来源：新闻

OntoNotes 5.0 / CoNNLL 2012

语种：英语、阿拉伯语、中文
实体类型：人名、地名、机构名等18种
数据来源：电话、新闻、广播、博客等

ACE 2004

语种：英语、阿拉伯语、中文
用于完成实体、 关系、 事件抽取等任务

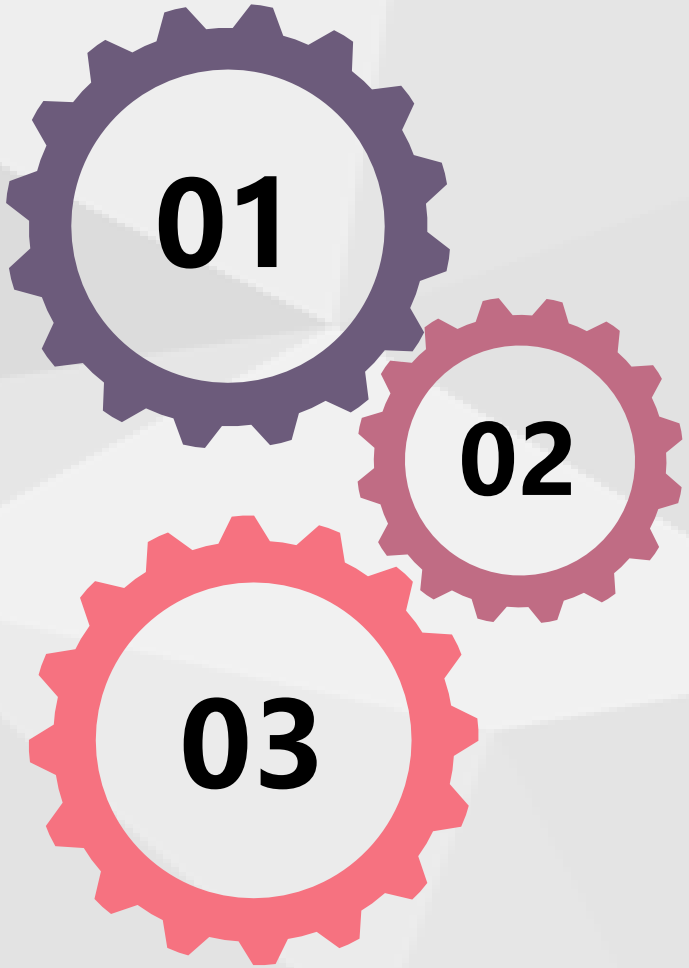
ACE 2005

语种：英语、阿拉伯语、中文
用于完成实体、 关系、 事件抽取等任务

标注方法

IOB 标注法

是早期 CoNLL数据集采用的标注法，标签形式为A-XXX。A代表该词在命名实体中的位置，I为 内部，O为外部，B为开始位置。XXX代表命名实体的类别。



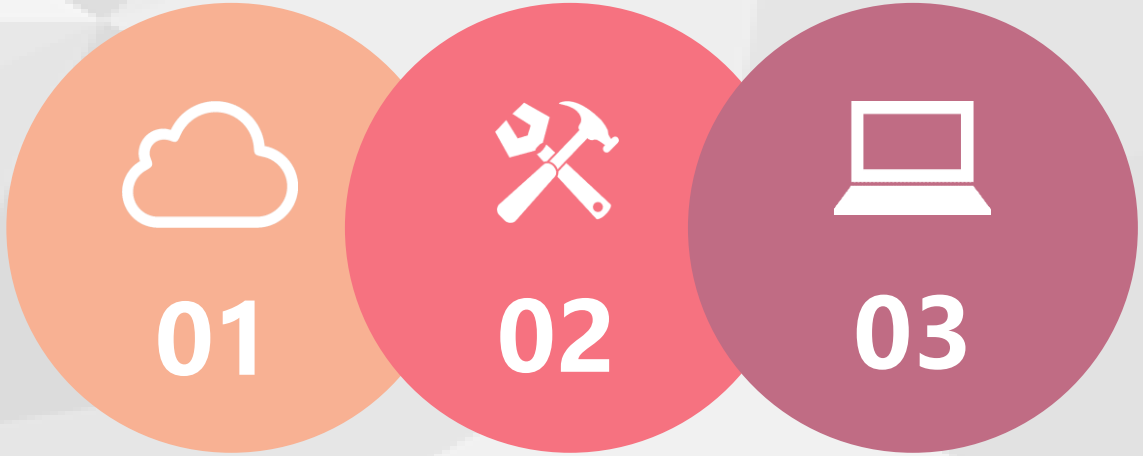
BIOES标注法

目前最通用的命名实体标注方法，在 IOB方法上增加了两个位置标签，E代表结束位置，S代表该词单独形成一个命名实体。

Makeup标注法

OntoNotes 使用的标注方法，用XML标签把命名实体框出来， 标注相应的类型。

评价指标



精确率(Precision)

$$Precision = \frac{TP + TN}{TP + FN + FP + TN}$$

召回率(Recall)

$$Recall = \frac{TP}{TP + FN}$$

F1 值(F1-Measure)

$$\frac{1}{F1} = \frac{1}{Recall} + \frac{1}{Precision}$$

2 发展历程

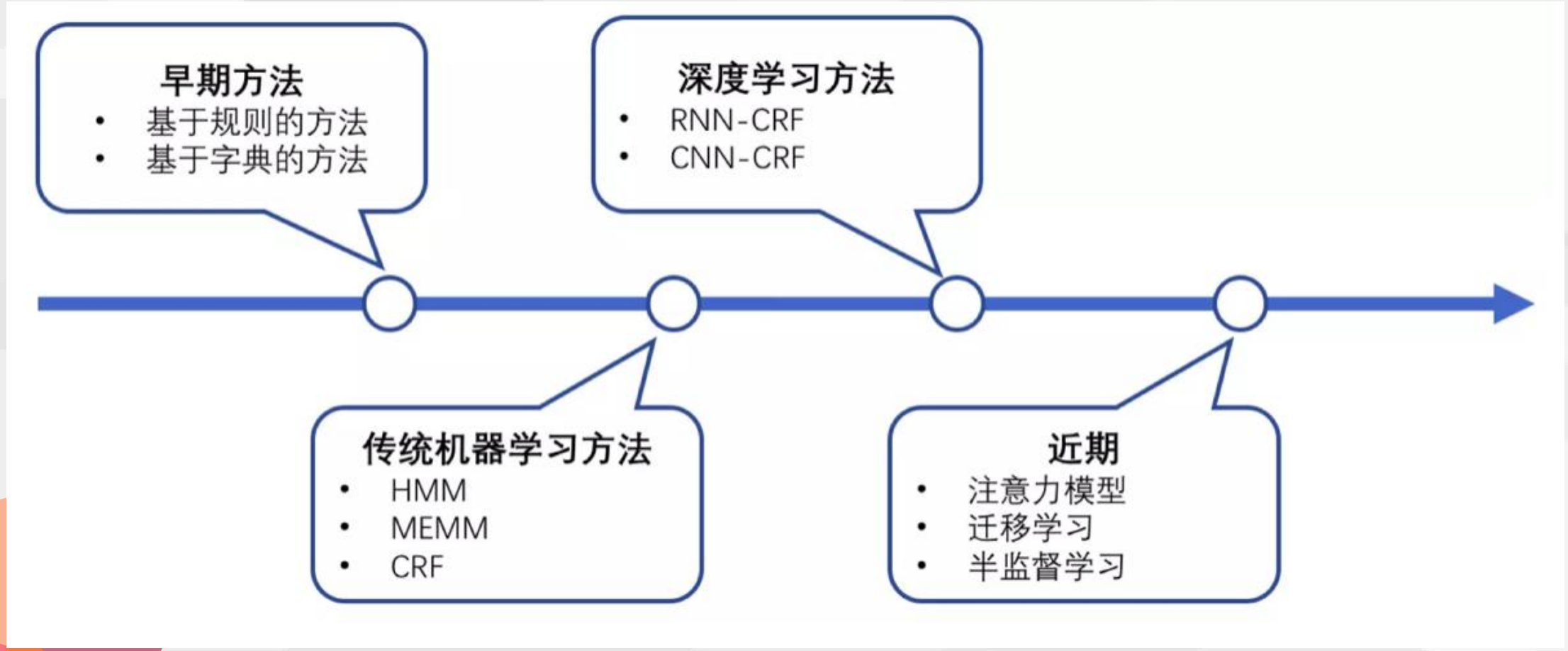
沈潇雨、郑灵杰

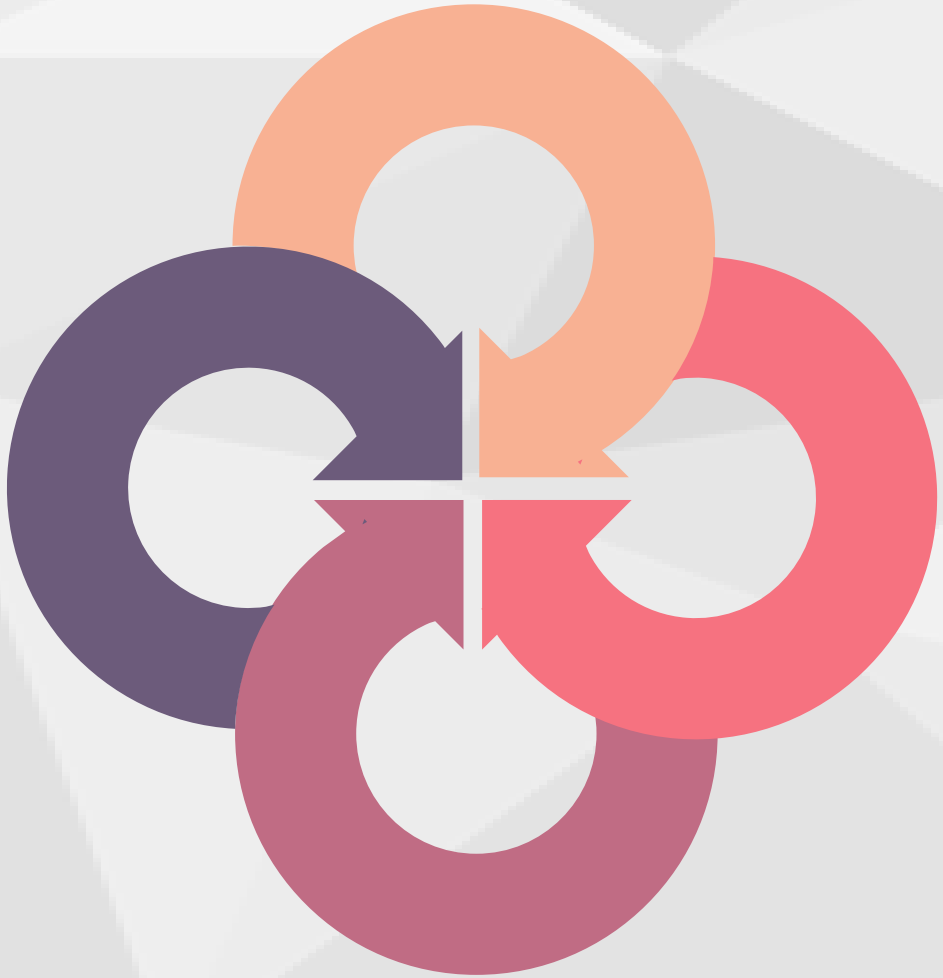
开端

1991年：第七届IEEE人工智能应用会议，Rau：关于“抽取和识别公司名称”的有关研究文章

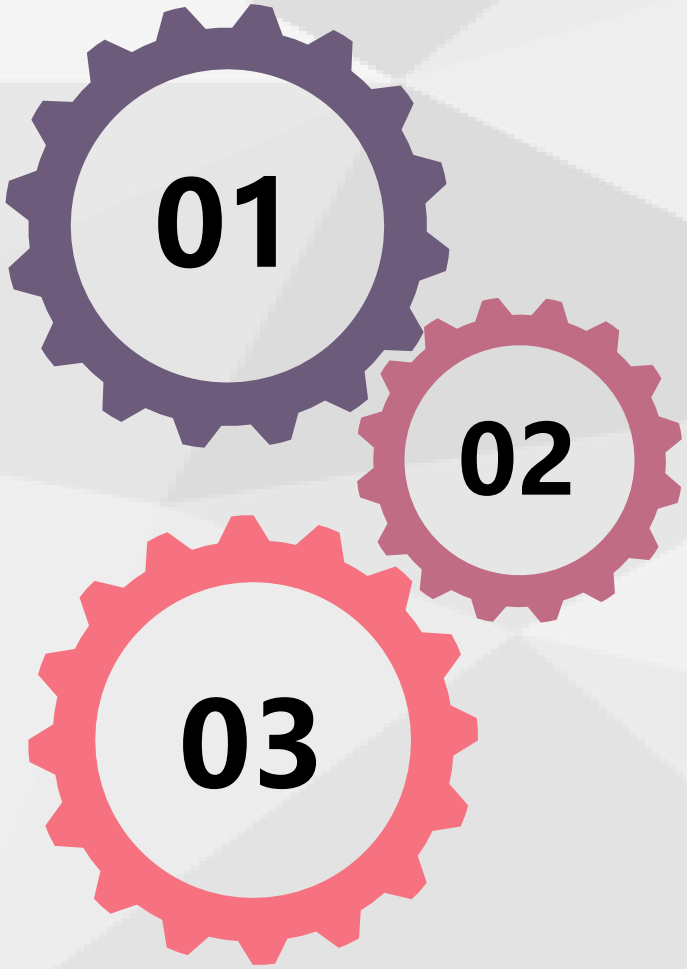
1996年，“命名实体（Named Entity，NE）”一词首次用于第六届信息理解会议。

随后出现了一系列信息抽取的国际评测会议这些评测会议对命名实体识别的发展有极大的推动作用。


早期方法
基于规则的方法
基于字典的方法
深度学习方法
RNN-CRF
CNN-CRF
传统机器学习方法
HMM
MEMM
CRF
近期
注意力模型
迁移学习
半监督学习



- 基于规则的方法
- 基于统计的机器学习方法
- 深度学习方法



01基于规则的方法

基于规则的实体识别

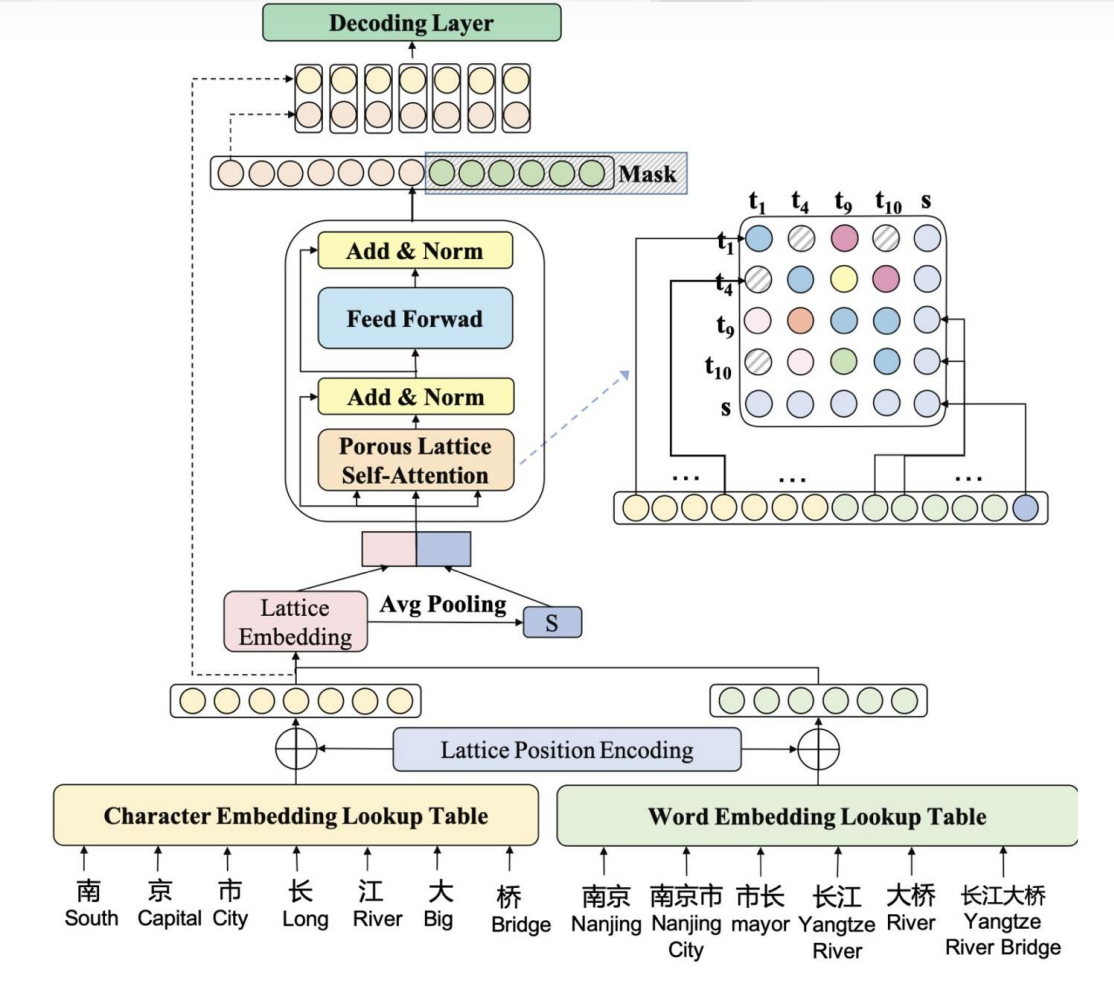
基于规则的实体识别方法是根据文本特点与定制规则特点匹配的方式完成实体识别。

常用的定制规则方式：

- 根据命名实体的构词规则
- 根据匹配标点符号的规则
- 根据命名实体
- 上下文词性用词特点的规则等

基于规则的实体识别

例：以词典为规则：



将词典中的每个词与被处理文档之间逐一匹配

- 字符串多模匹配：
  - trie树
  - 记录长度集合的最长匹配

- 切词匹配【准确率高，但时空效率低，召回率低】

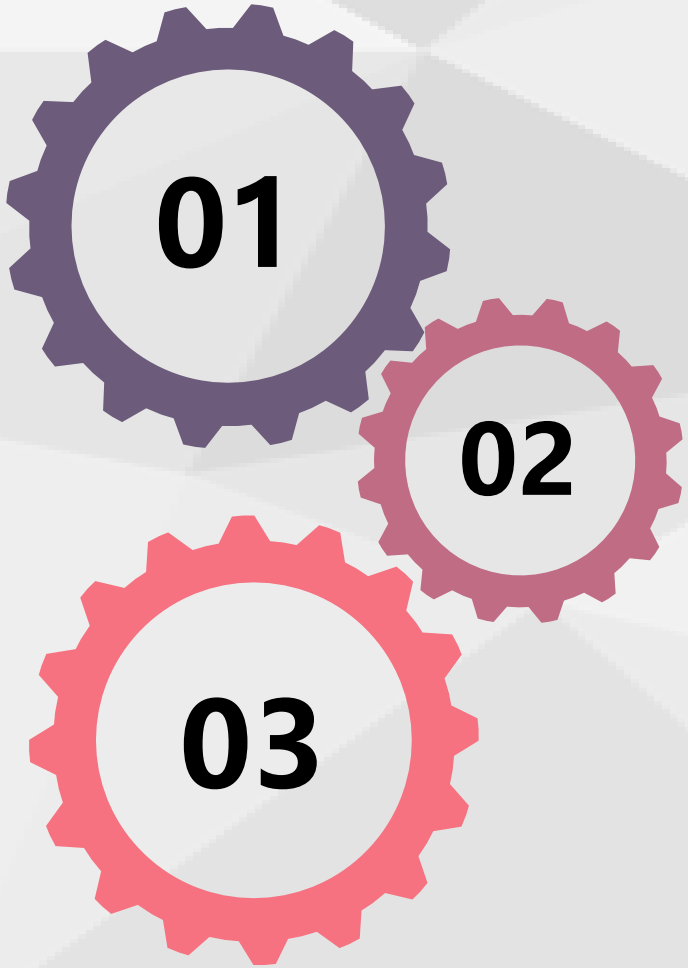
基于规则的实体识别

优点

- ✓ 使用简单
- ✓ 结果准确率较高
- ✓ 特别是对于数字和时间日期实体利用规则匹配的方式能获得较好的识别效果

缺点

- ☐ 规则库的建立需要花费大量时间和人力
- ☐ 不同的实体类型需要定制相应的规则，移植性差。



02基于统计的机器学习方法

基于统计的机器学习方法

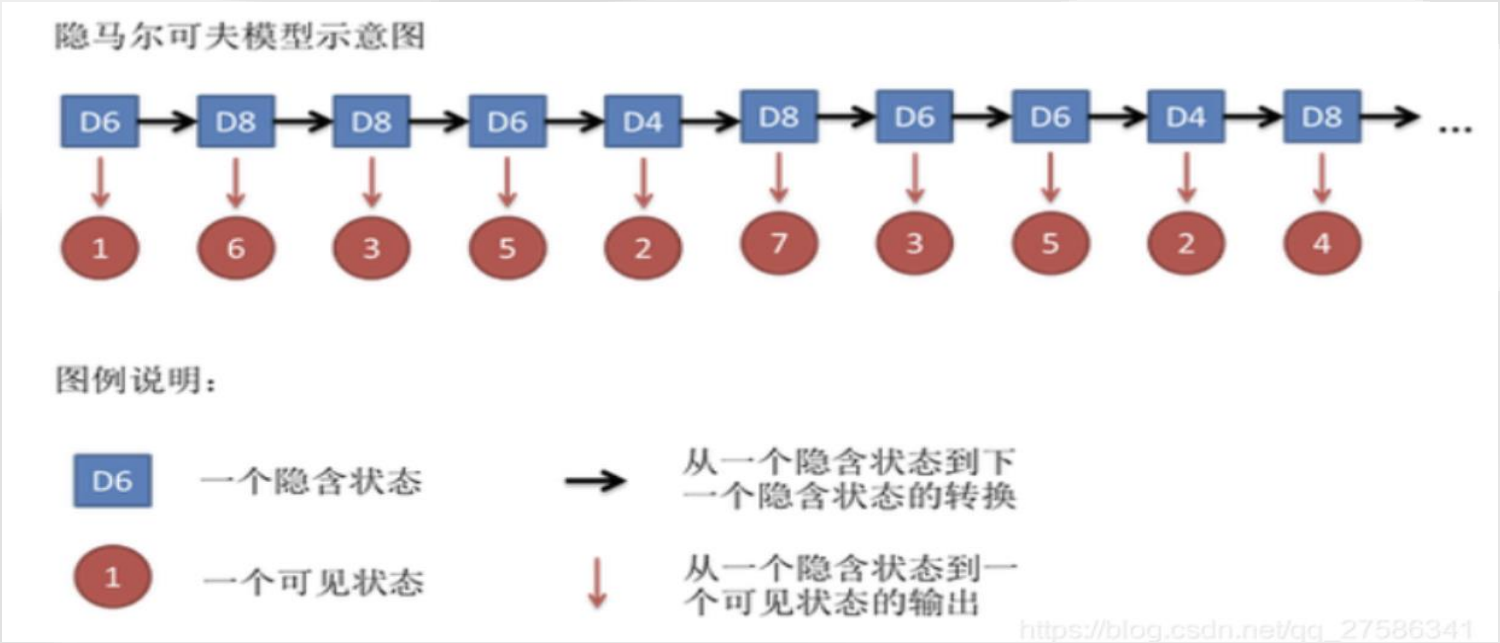
基于统计机器学习的方法是从给定的、已标注好的训练集出发，通过人工构建特征，并根据特定的模型对文本中每个词进行标签标注，实现命名实体识别

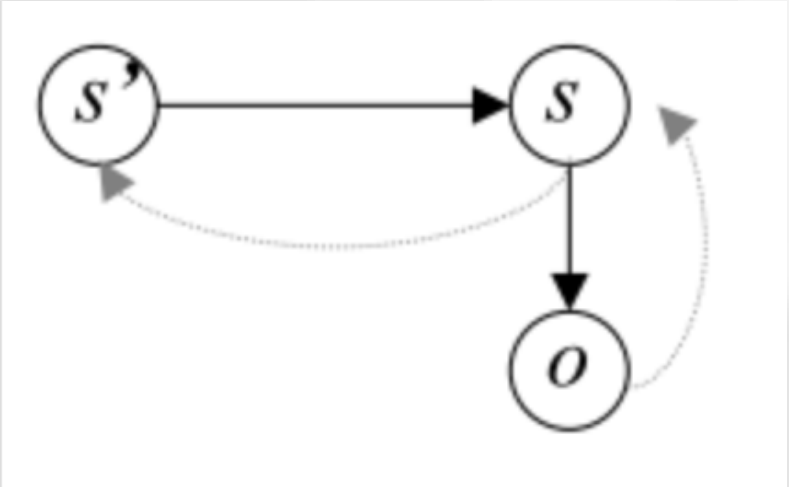
基于统计机器学习的模型：

- 隐马尔可夫模型HMM
- 最大熵马尔可夫模型MEMM
- 支持向量机SVM
- 条件随机场 CRF

隐马尔可夫模型HMM

- 隐马尔可夫模型是统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程；
- 隐马尔可夫模型中的两条序列：
  - 可以直接通过观测得到的观察序列，在NER中指每一个词语本身
  - 隐含的状态转移序列，指每个词语背后的标注



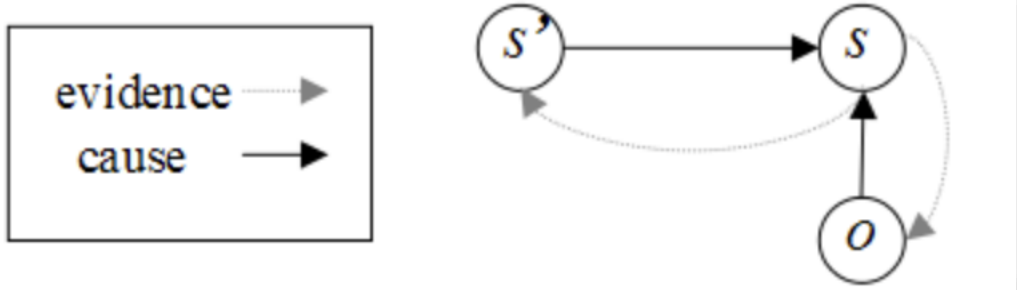


最大熵马尔可夫模型MEMM

最大熵模型：在给定训练数据的条件下对模型进行极大似然估计或正则化极大似然估计：

$$P_w(y|x) = \frac{exp\left(\sum_i w_i f_i(x,y)\right)}{Z_w(x)}$$

MEMM：直接学习条件概率 $P(s_t|s_{t-1}, o_t)$
其当前状态依赖于前一状态与当前观测



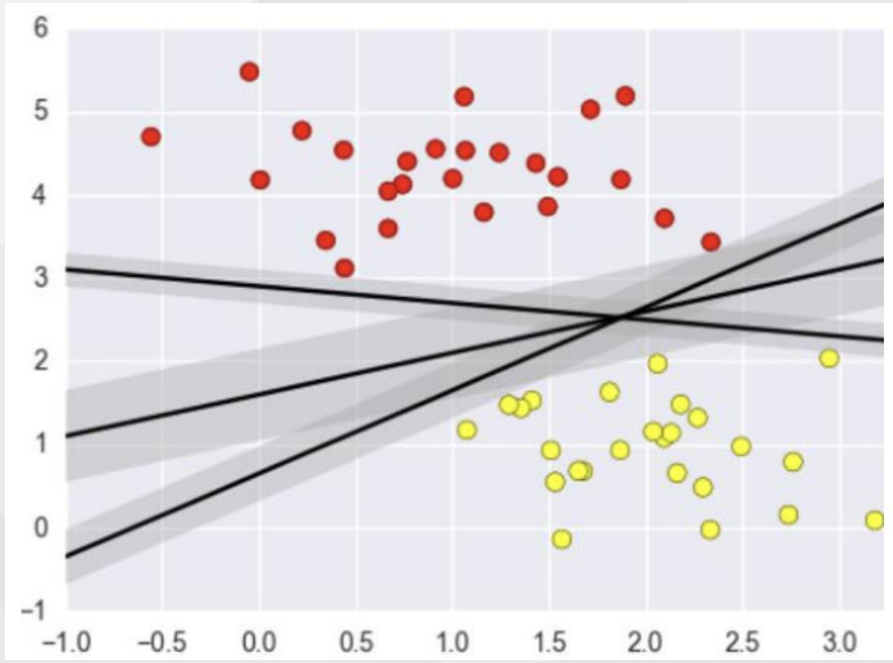
MEMM是一个概率模型，即在给定的观察状态和前一状态的条件下，出现当前状态的概率。

支持向量机SVM

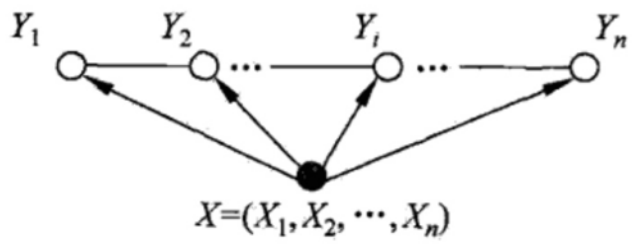
支持向量机是一种二分类模型：将实例的特征向量映射为空间中的一些点，并画出一条线，以“最好地”区分这两类点，以至如果以后有了新的点，这条线也能做出很好的分类。

它的基本模型是定义在特征空间上的间隔最大的线性分类器。

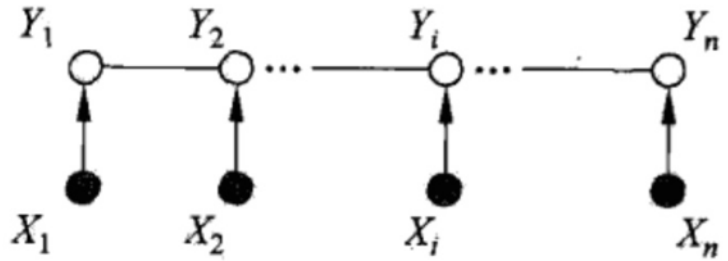
- 线性可分支持向量机
- 线性支持向量机
- 非线性支持向量机



条件随机场 CRF



线性链条件随机场



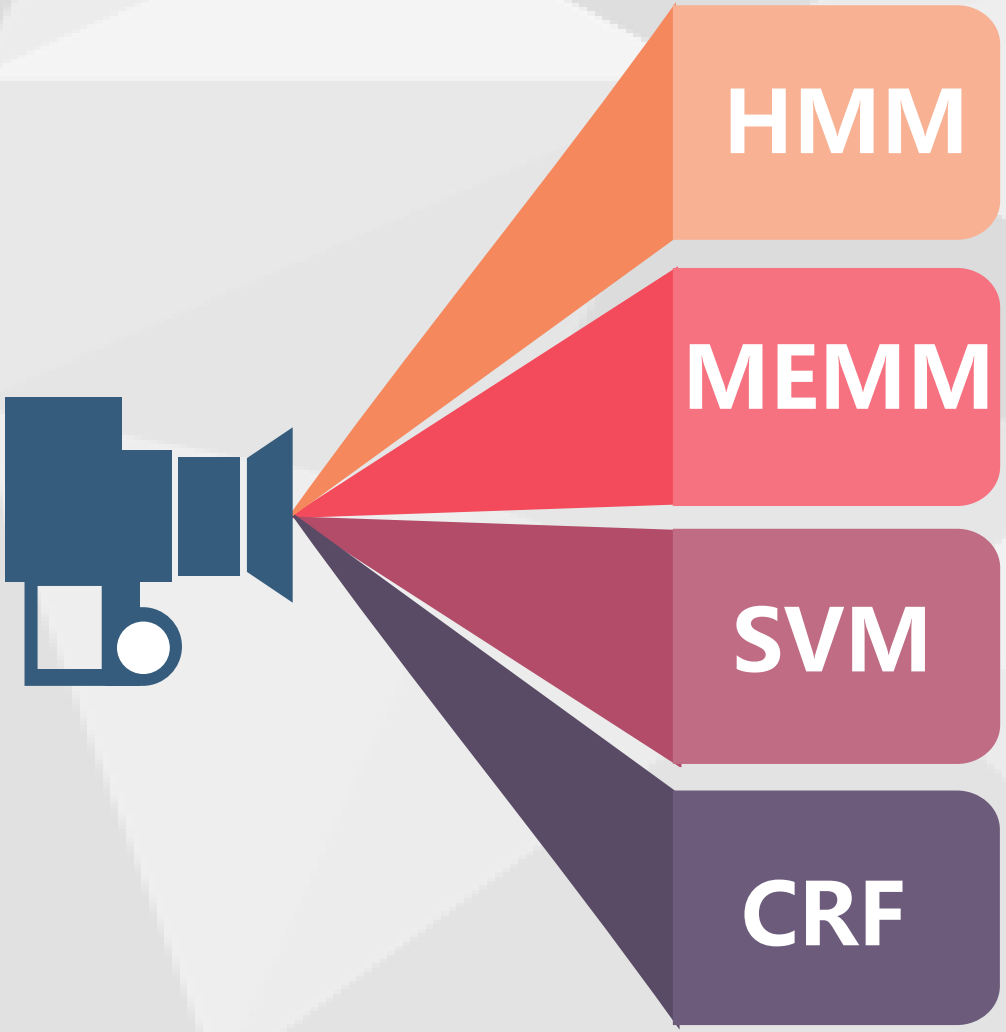
X和Y有相同的图结构的线性链条件随机场

条件随机场是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型，其特点是假设输出随机变量构成马尔可夫随机场。

- 每一个HMM模型都等价于某个CRF

- 概率计算：前向-后向算法
- 学习算法：拟牛顿法BFGS
- 预测算法：维特比算法

基于统计的机器学习技术比对

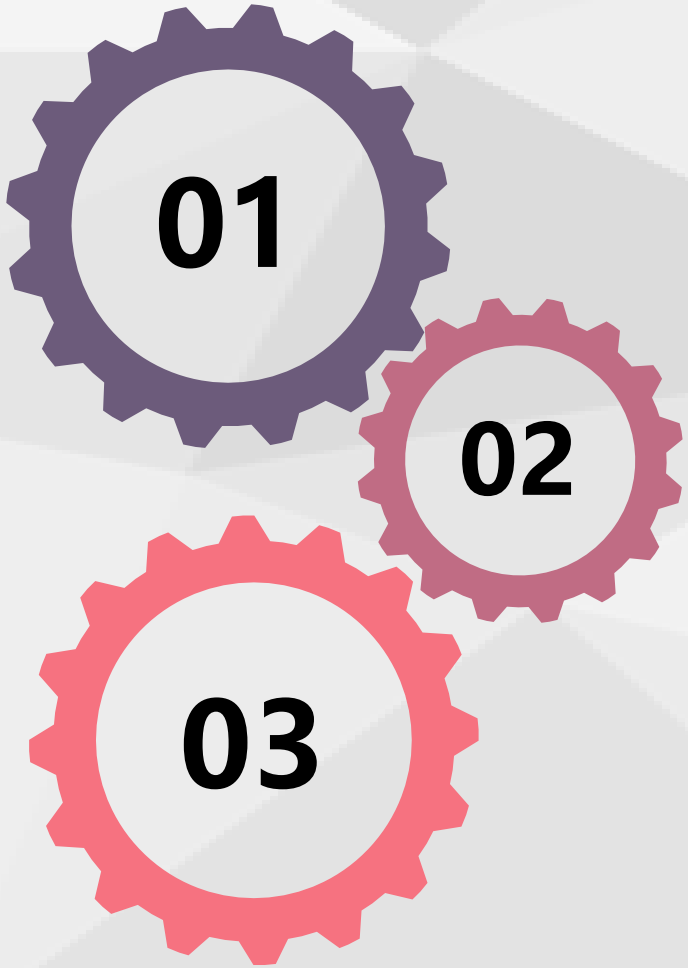


是最早提出的基于统计机器学习的实体识别技术，结构简单，适合小数据集，训练效率高；
但该算法并不能考虑上下文信息

解决了 HMM 输出独立性的问题，并具有较为灵活的特征选取方式；
但因为其局部归一化计算而出现了标签偏置问题

相对于其他3 种方法较为简单，降低了计算维数；
但对参数和核函数选择要求高，且不适用于训练大规模样本

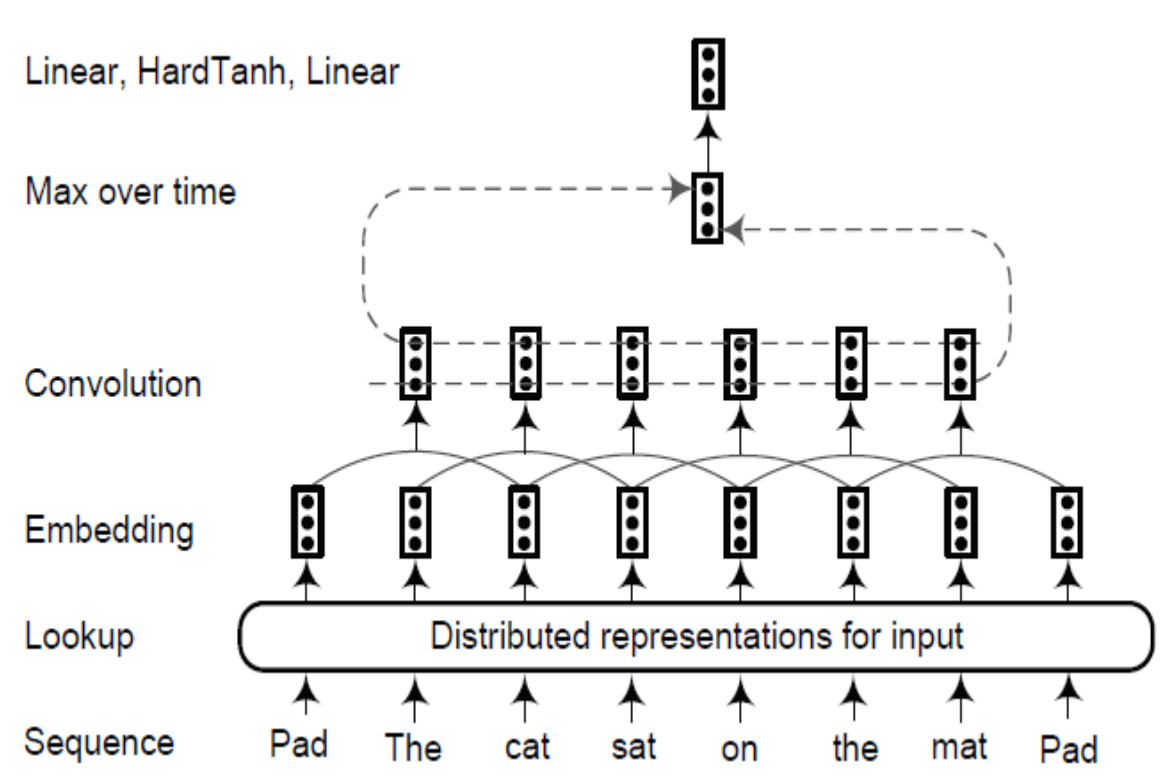
可利用上下文信息，全局归一化求得最优解，具有较好的识别性能；
但收敛速度慢、复杂度高，训练时间较长



03深度学习方法

深度学习方法

CNN



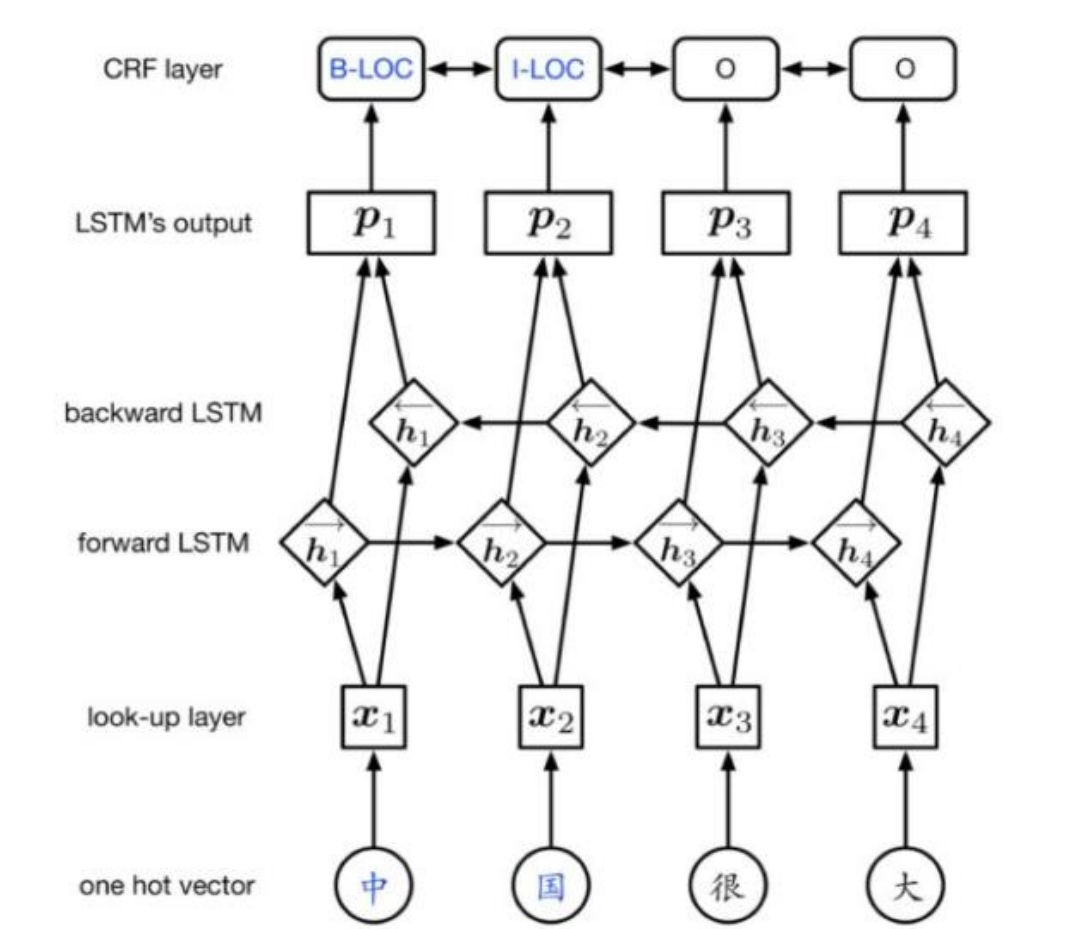
利用CNN去获取某个词的局部特征，然后通过最大或平均池化去获取全局特征，之后输入到标签解码器中获取标签分布。

CNN的优势是在GPU上的并行运算比较强大，训练速度比较快。

CNN缺点：CNN 的上下文信息取决于窗口的大小，虽然不断地增加 CNN 卷积层最终也可以达到使每个 token 获取到整个输入句子作为上下文信息的目的，但是其输出的分辨表现力太差。

深度学习方法

Bi-LSTM + CRF



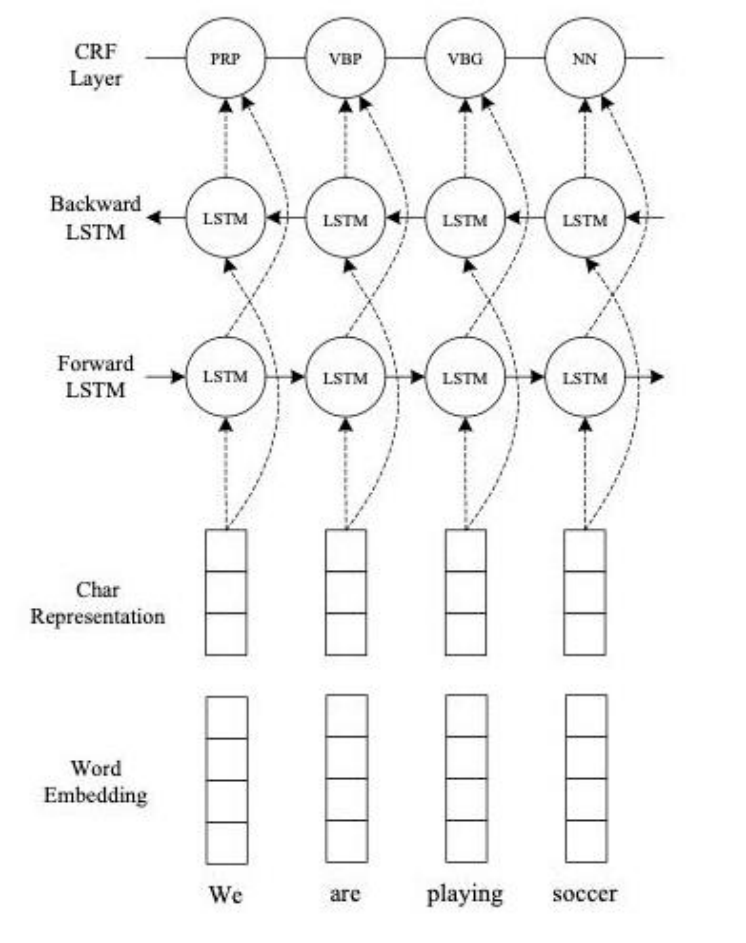
Bi-LSTM-CRF模型主要由Embedding层（主要有词向量，字向量以及一些额外特征），双向LSTM层，以及最后的CRF层构成。

Bi-LSTM-CRF 模型可以有效地利用过去和未来的输入特征。借助 CRF 层，它还可以使用句子级别的标记信息。

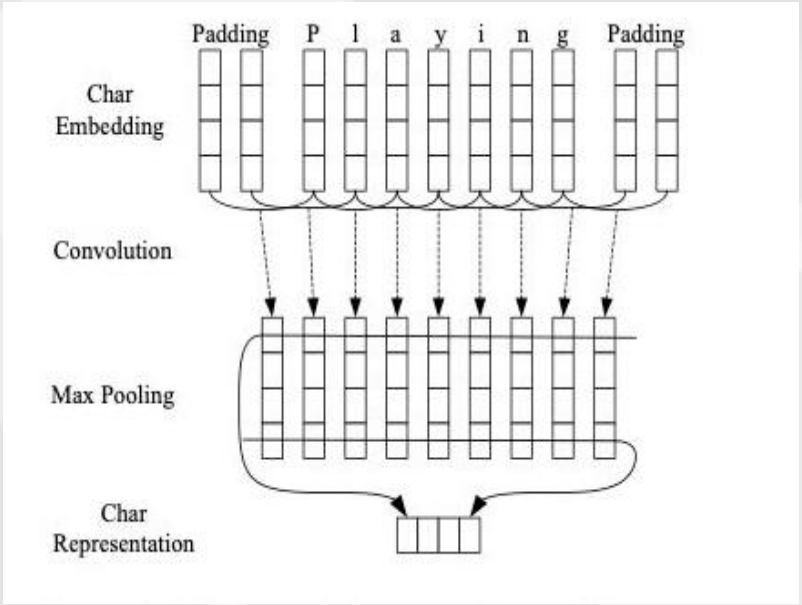
Bi-LSTM本质是一个序列模型，跟NER这种序列标注任务有很好的相性，但在对GPU并行计算的利用上就比不上CNN那么强大。

深度学习方法

CNN + Bi-LSTM + CRF

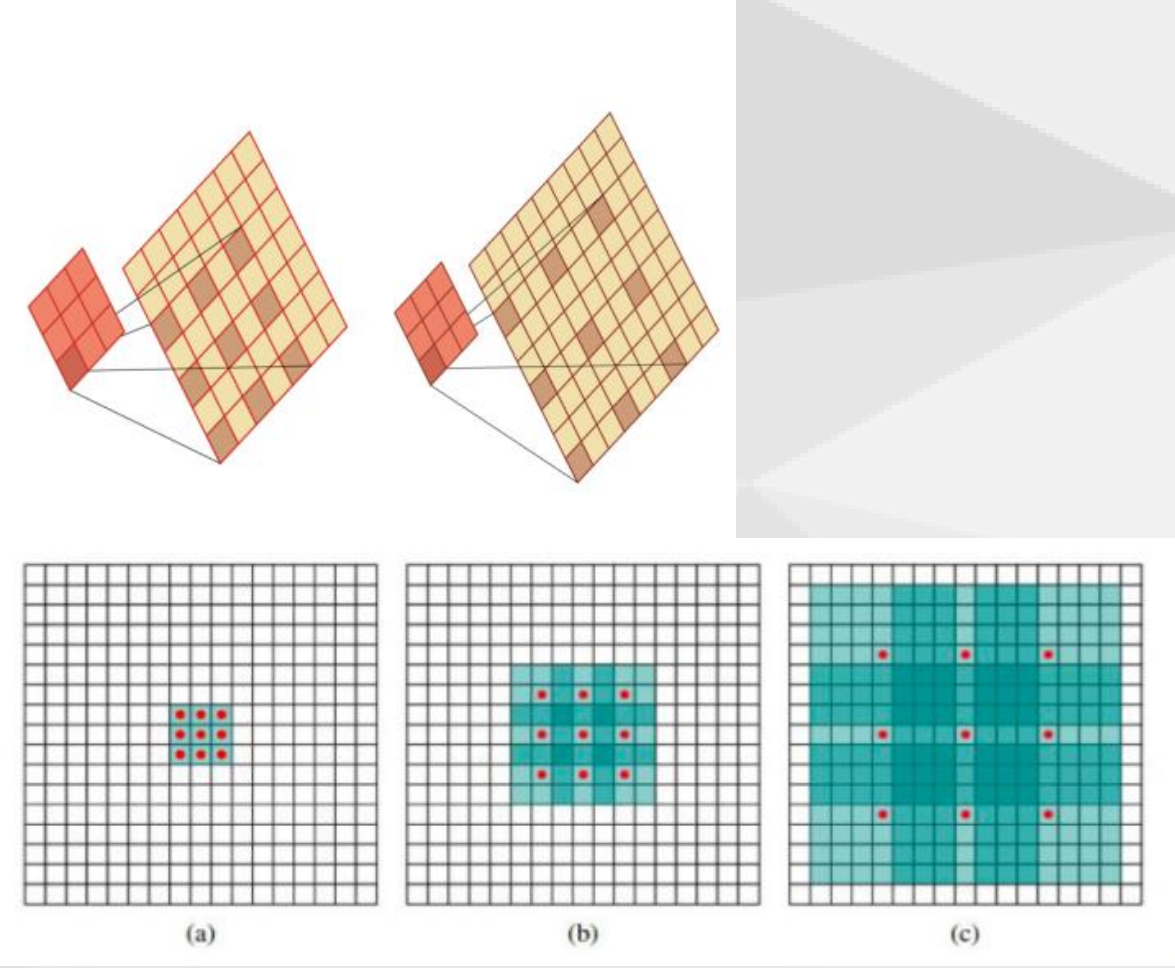


上一种方法的改进，通过CNN获取**字符级**的词表示（CNN是一个非常有效的方式去抽取词的形态信息（例如词的前缀和后缀）进行编码的方法），然后将CNN的字符级编码向量和词级别向量拼接，输入到Bi-LSTM + CRF网络中。



深度学习方法

IDCNN-CRF

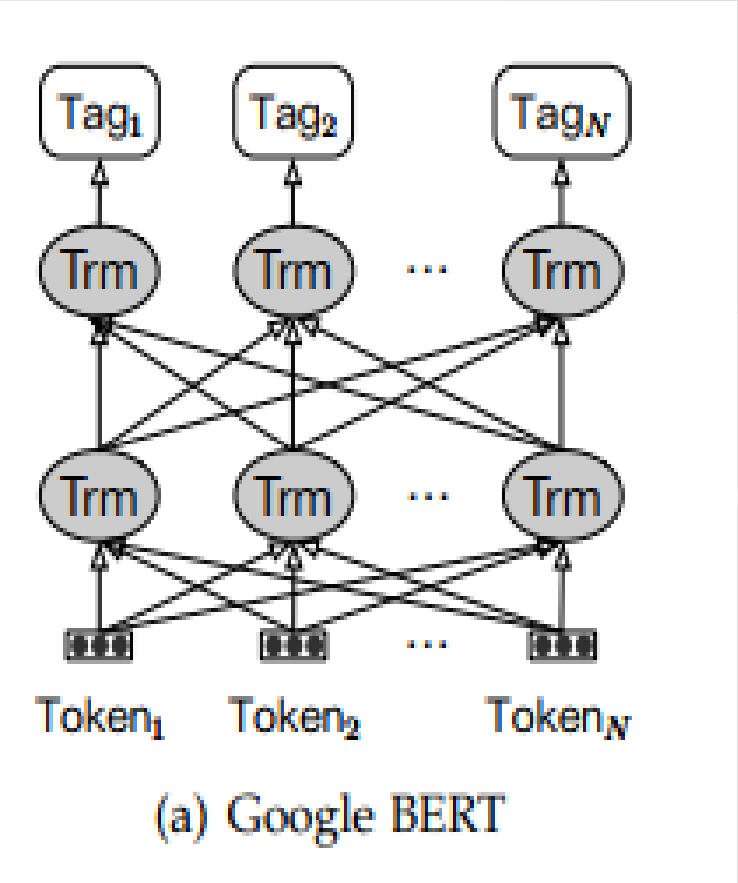


引入膨胀卷积（或叫空洞卷积），一方面可以引入CNN并行计算的优势，提高训练和预测时的速度；另一方面，可以减轻CNN在长序列输入上特征提取能力弱的劣势。

Dilated width（膨胀宽度）会随着层数的增加而指数增加。这样随着层数的增加，参数数量是线性增加的，而感受野却是指数增加的，这样就可以很快覆盖到全部的输入数据。

在保证和 Bi-LSTM-CRF 相 当的正确率，IDCNN-CRF带来了 14-20 倍的提速。句子级别的解码提速 8 倍。

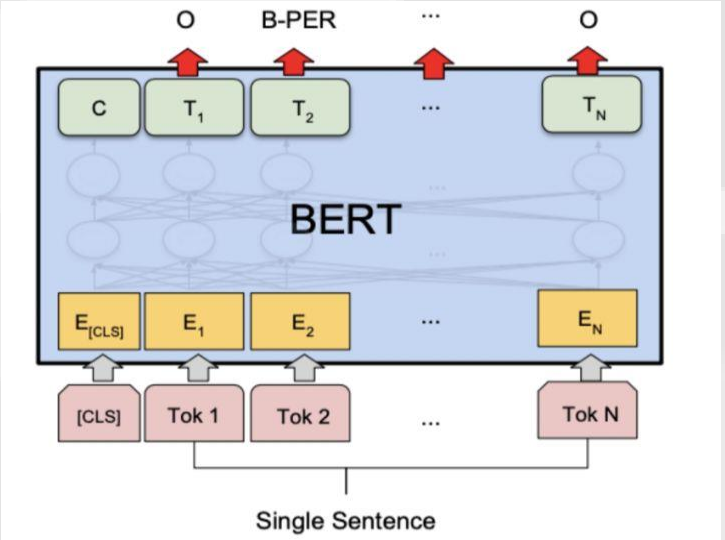
BERT——基于transformer



特点：深层、双向Transformer

基于fine-tuning的方式（对预训练的模型进行微调），每一层都联合考虑到了左边和右边的上下文信息。

Transformer的特征提取能力要远远的优于LSTM，且Transformer易于并行计算，能够很好地捕获长距离的信息。



3

前沿进展

侯子睿

前沿进展分享

ACL2021：（9）**信息抽取**（60篇）；

聚焦方向：

⭕少样本学习(Few-shot)

1.Leveraging Type Descriptions for Zero-shot Named Entity Recognition and Classification

2.Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data

⭕弱监督学习

1.BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition

⭕嵌套、不连续问题

1.Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition

2.A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition

⭕特定领域问题

1.A Neural Transition-based Joint Model for Disease Named Entity Recognition and Normalization

关于 中文领域 NER

不同于英文NER，中文NER通常以字符为单位进行序列标注建模。这主要是由于中文分词存在误差，导致基于字符通常要好于基于词汇（经过分词）的序列标注建模方法。

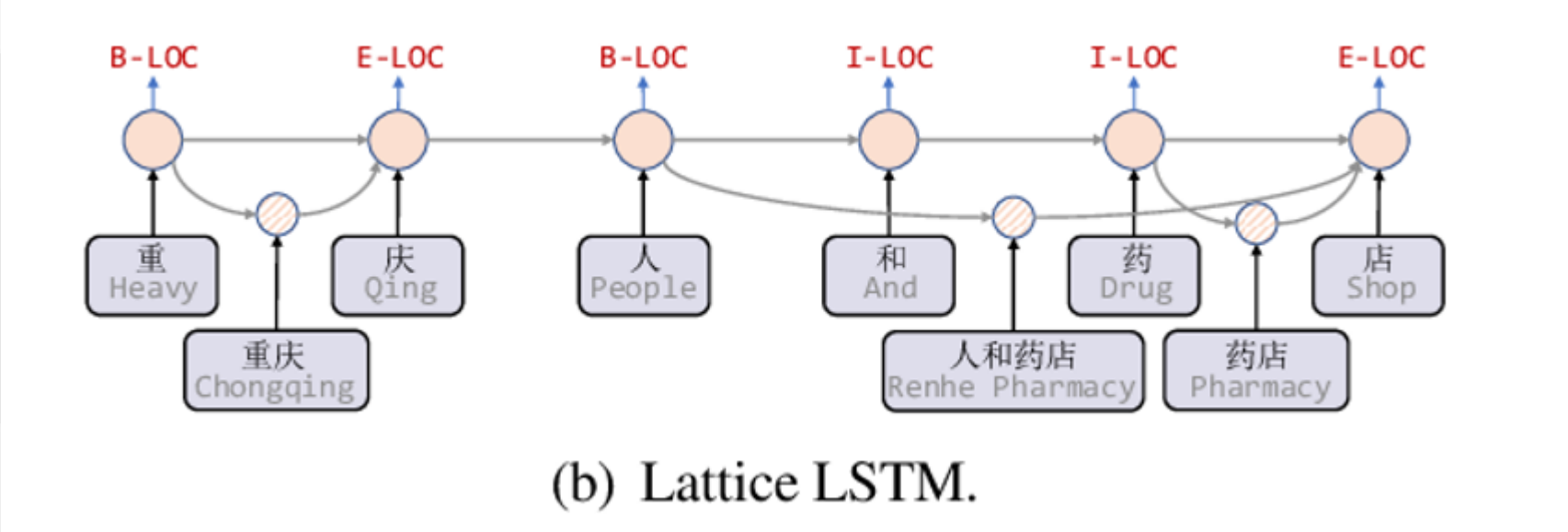
词汇增强：

1. Dynamic Architecture：设计一个动态抽取框架，能够兼容词汇输入；
2. Adaptive Embedding：基于词汇信息，构建自适应Embedding；与模型框架无关。

	模型名称	简介
Dynamic Architecture	Lattice LSTM	通过Lattice结构将词汇信息注入模型
	LR-CNN	多尺度CNN提取特征，回溯机制解决词汇冲突
	CGN	写作图网络，融合多个图结构
	LGN	全局与局部聚合更新的图网络
	FLAT	利用Transformer进行encode，相对位置编码融合词汇信息
Adaptive Embedding	WC-LSTM	四种策略进行静态的固定编码
	Multi-digraph	使用多图结构建模字符和词典的联系
	Simple-Lexicon	使用soft-lexicon，将词汇信息融入到Embedding
	LEBERT	使用Lexicon Adapter layer将词汇信息融合到BERT底层

▲ 词增强NER模型简介

Lattice结构



(b) Lattice LSTM.

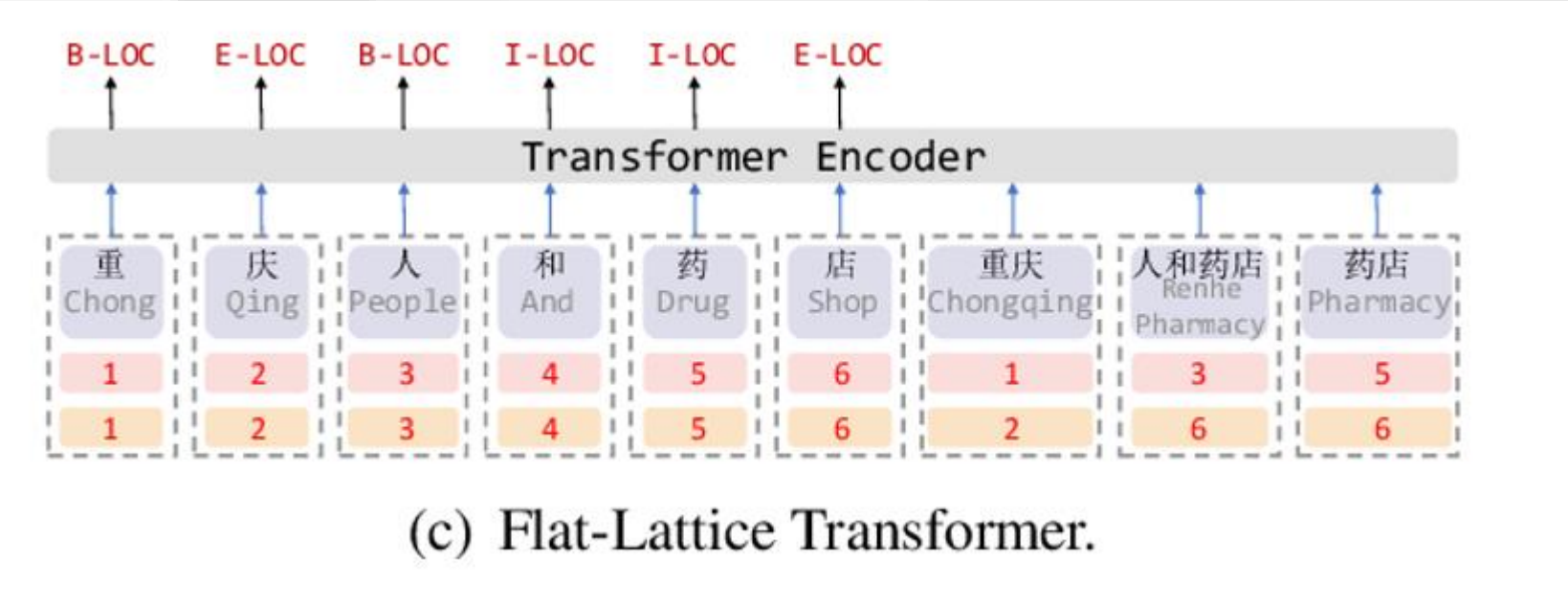
Lattice LSTM引入了word cell结构，对于当前的字符，融合以该字符结束的所有word信息。

信息损失：

- 每个字符只能获取以它为结尾的词汇信息。
- 由于RNN特性，采取BiLSTM时其前向和后向的词汇信息不能共享。
- Lattice LSTM并没有利用前一时刻的记忆向量 ，即不保留对词汇信息的持续记忆。

参考论文：FLAT: Chinese NER Using Flat-Lattice Transformer

Flat结构



(c) Flat-Lattice Transformer.

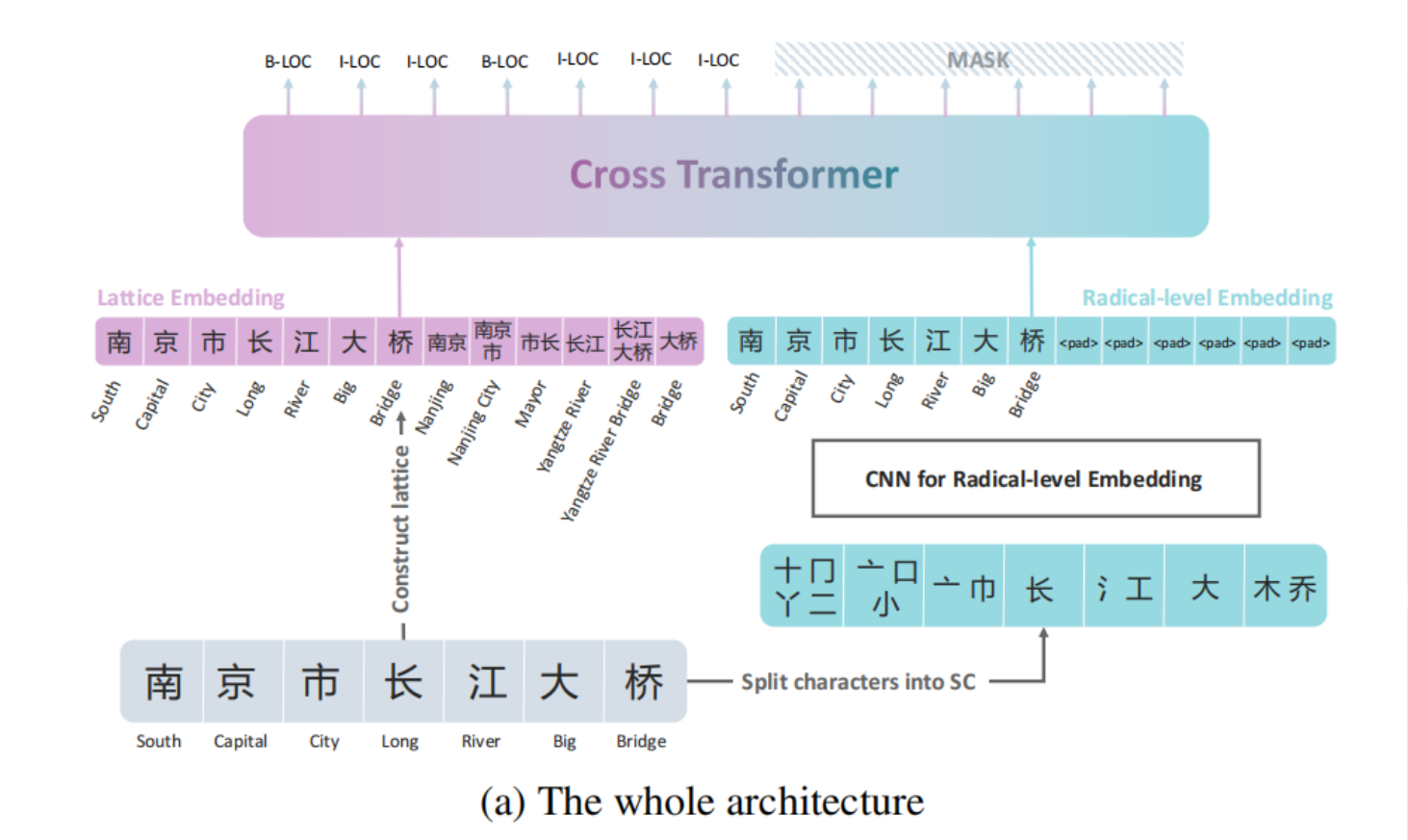
- Transformer采取全连接的自注意力机制可以很好捕捉长距离依赖，由于自注意力机制对位置是无偏的，因此Transformer引入位置向量来保持位置信息。

- 受到位置向量表征的启发，FLAT设计了一种巧妙position encoding来融合Lattice 结构，如上图所示，对于每一个字符和词汇都构建两个head position encoding 和tail position encoding，这种方式可以重构原有的Lattice结构。

参考论文：FLAT: Chinese NER Using Flat-Lattice Transformer

基于多元数据的双流Transformer编码模型

参考论文：MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition



(a) The whole architecture

- 汉字中包含的偏旁部首等结构可以代表某些含义。因此，作者提出在模型中融合进汉字的结构信息（例如部首等）

使用CNN提取Radical-level特征

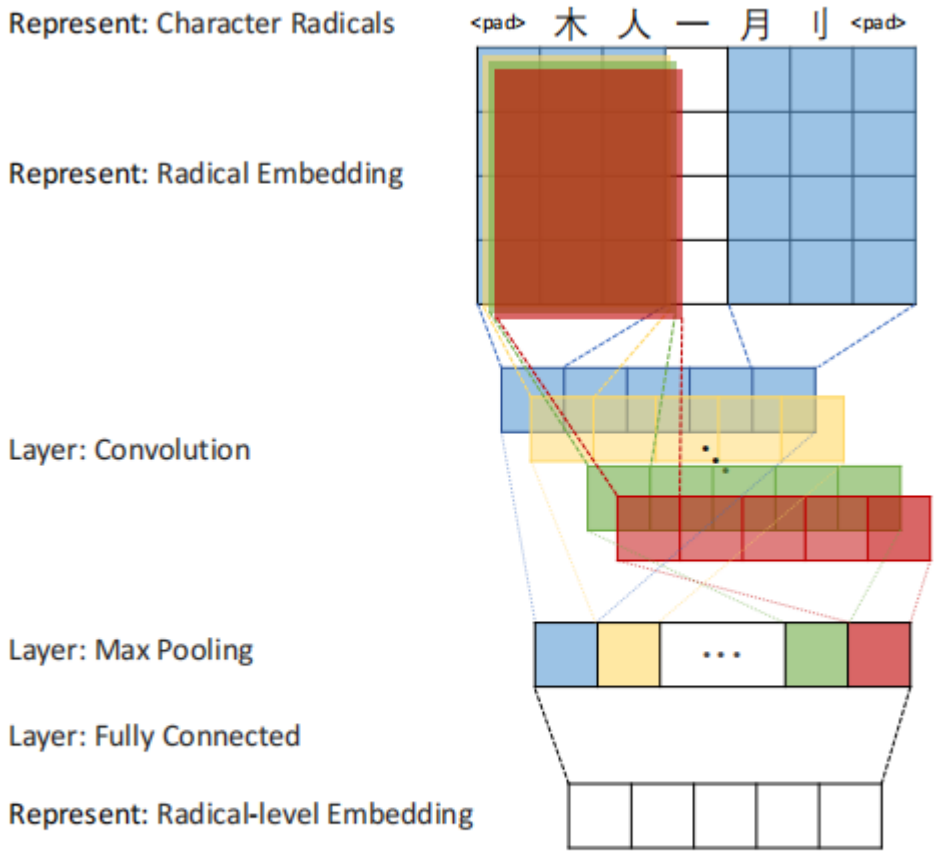


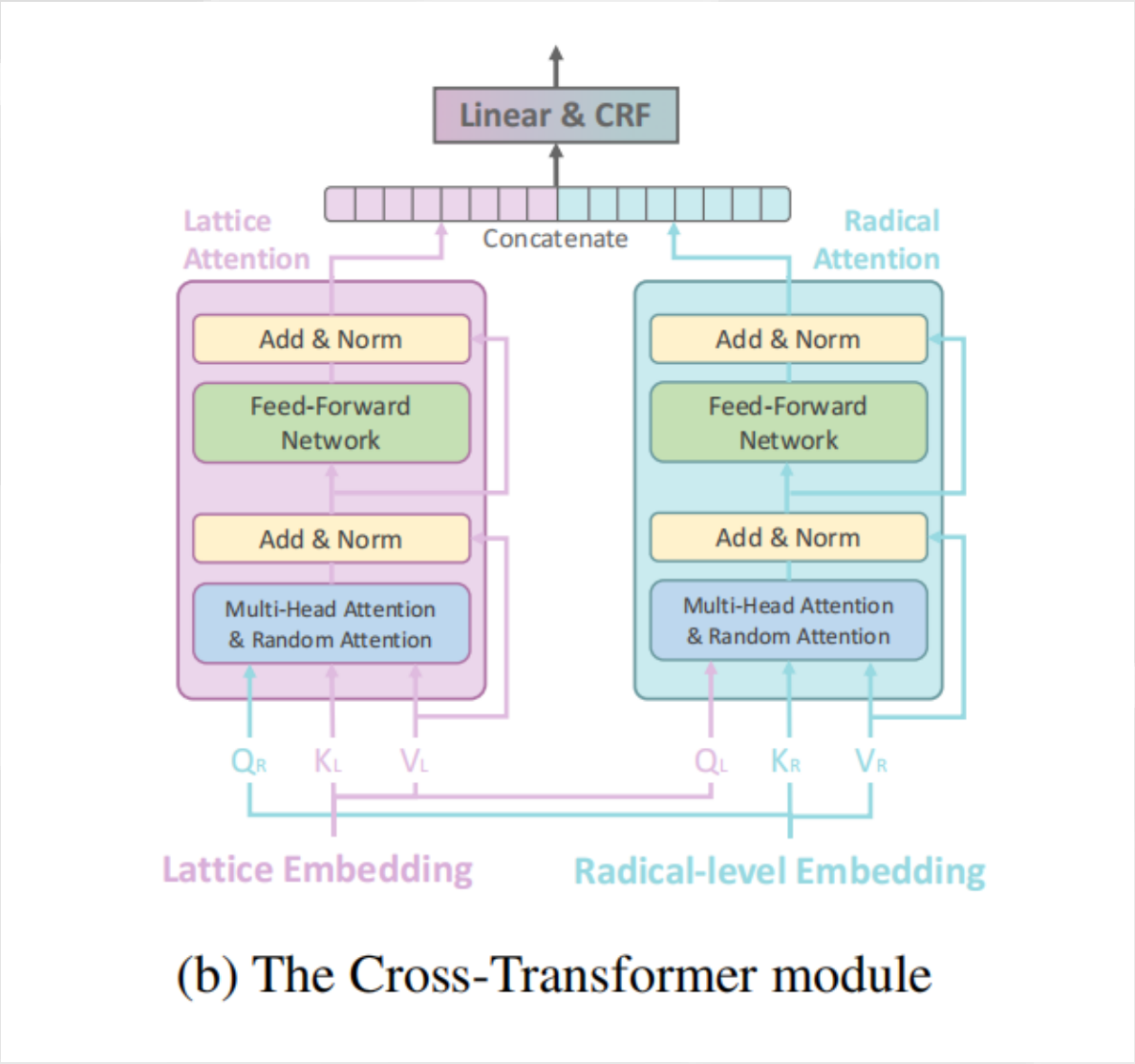
Figure 3: CNN for radical feature extraction.

Character	CR	HT	SC
题 (topic)	页	是页	日一走页
榆 (elm)	木	木俞	木人一月刂
渡 (ferry)	氵	氵度	氵广廿又
脸 (face)	月	月佥	月人一ツ一

- 对原始文本中的汉字进行拆字，然后把得到的字根特征输入到 CNN 特征提取器当中，然后使用最大值池化和全连接网络得到每个汉字的 Radical-level 特征。

参考论文：MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition

Cross-Transformer模块



(b) The Cross-Transformer module

参考论文：MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition

前沿进展分享

实验结果展示

Models	NE	NM	Overall
Peng and Dredze (2015)	51.96	61.05	56.05
Peng and Dredze (2016)*	55.28	62.97	58.99
He and Sun (2017a)	50.60	59.32	54.82
He and Sun (2017b)*	54.50	62.17	58.23
Cao et al. (2018)	54.34	57.35	58.70
Lattice-LSTM	53.04	62.25	58.79
CAN-NER	55.38	62.98	59.31
LR-CNN	57.14	**66.67**	59.92
LGN	55.34	64.98	60.21
PLT	53.55	64.90	59.76
SoftLexicon (LSTM)	59.08	62.22	61.42
Baseline	-	-	60.32
MECT	**61.91**	62.51	**63.30**
BERT	-	-	68.20
BERT + MECT	-	-	**70.43**

Table 4: Results obtained on Weibo (%).

Models	P	R	F1
Zhang and Yang (2018)$^{\mathcal{A}}$	93.72	93.44	93.58
Zhang and Yang (2018)$^{\mathcal{B}}$	94.07	94.42	94.24
Zhang and Yang (2018)$^{\mathcal{C}}$	93.66	93.31	93.48
Zhang and Yang (2018)$^{\mathcal{D}}$	94.53	94.29	94.41
Lattice-LSTM	94.81	94.11	94.46
CAN-NER	95.05	94.82	94.94
LR-CNN	95.37	94.84	95.11
LGN	95.28	95.46	95.37
PLT	95.34	95.46	95.40
SoftLexicon (LSTM)	95.30	**95.77**	95.53
+ bichar	95.71	**95.77**	95.74
Baseline	-	-	95.45
MECT	**96.40**	95.39	**95.89**
BERT	-	-	95.53
BERT + MECT	-	-	**95.98**

Table 5: Results obtained on Resume (%). For Zhang and Yang (2018), $\mathcal{A}$ represents word-based LSTM', $\mathcal{B}$ indicates 'word-based + char + bichar LSTM', $\mathcal{C}$ represents the 'char-based LSTM' model, and $\mathcal{D}$ is the 'char-based + bichar + softword LSTM' model.

前沿进展分享

实验结果展示

Models	P	R	F1
Yang et al. (2018)§	65.59	71.84	68.57
Yang et al. (2018)§*†	72.98	**80.15**	76.40
Che et al. (2013)§*	77.71	72.51	75.02
Wang et al. (2013)§*	76.43	72.32	74.32
Zhang and Yang (2018)♬§	**78.62**	73.13	75.77
Zhang and Yang (2018)♬¶	73.36	70.12	71.70
Lattice-LSTM	76.35	71.56	73.88
CAN-NER	75.05	72.29	73.64
LR-CNN	76.40	72.60	74.45
LGN	76.13	73.68	74.89
PLT	76.78	72.54	74.60
SoftLexicon (LSTM)	77.28	74.07	75.64
+ bichar	77.13	75.22	76.16
Baseline	-	-	76.45
MECT	77.57	**76.27**	**76.92**
BERT	-	-	80.14
BERT + MECT	-	-	**82.57**

Table 6: Results on Ontonotes 4.0 (%), where '§' denotes gold segmentation and '¶' denotes auto segmentation.

Models	P	R	F1
Chen et al. (2006)	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
Lattice-LSTM	93.57	92.79	93.18
CAN-NER	93.53	92.42	92.97
LR-CNN	94.50	92.93	93.71
LGN	94.19	92.73	93.46
PLT	94.25	92.30	93.26
SoftLexicon (LSTM)	94.63	92.70	93.66
+ bichar	**94.73**	93.40	94.06
Baseline	-	-	94.12
MECT	94.55	**94.09**	**94.32**
BERT	-	-	94.95
BERT + MECT	-	-	**96.24**

Table 7: Results obtained on MSRA (%).

- 结果显示，相比于 baseline 模型 FLAT，在模型中加入汉字结构特征以后，性能有了一定提升。并且在小规模数据集（例如 weibo）或者多类别数据集（Ontonotes 4.0）上，模型的提升更加显著。

4

Demo 展示

陆皓霖

crf模型——中文数据

选用中文病例数据，每个病例由四部分组成：一般项目、病史特征、诊疗过程、出院情况。每个部分由对应的数据和标注（起始位置、终止位置、类别）组成，如

入院后完善各项检查，给予右下肢持续皮牵引，应用健骨药物治疗，患者略发热，查血常规：白细胞数12.18*10^9/L，中性粒细胞百分比92.00%。给予应用抗生素预防感染。复查血常规：白细胞数6.55*10^9/L，中性粒细胞百分比74.70%，红细胞数2.92*10^12/L，血红蛋白94.0g/L。考虑贫血，指示加强营养。建议患者手术治疗,患者拒绝手术治疗。继续右下肢牵引，患者家属要求今日出院。

```
右下肢   12   14   身体部位
健骨药物      23   26   治疗
发热      33   34   症状和体征
血常规   37   39   检查和检验
白细胞   41   43   检查和检验
中性粒细胞百分比      58   65   检查和检验
抗生素   77   79   治疗
血常规   87   89   检查和检验
白细胞   91   93   检查和检验
中性粒细胞百分比      107  114  检查和检验
红细胞数      122  125  检查和检验
血红蛋   139  141  检查和检验
贫血      153  154  疾病和诊断
右下肢   183  185  身体部位
```

```
if label == '身体部位':
    for j in range(begin_position, end_position + 1):
        if j == begin_position:
            original_data_list[j][1] = 'B-body'
        else:
            original_data_list[j][1] = 'I-body'
elif label == '症状和体征':
    for j in range(begin_position, end_position + 1):
        if j == begin_position:
            original_data_list[j][1] = 'B-sas'
        else:
            original_data_list[j][1] = 'I-sas'
elif label == '治疗':
    for j in range(begin_position, end_position + 1):
        if j == begin_position:
            original_data_list[j][1] = 'B-tre'
        else:
```

训练与结果

训练crf模型（sklearn）

```
crf = sklearn_crfsuite.CRF(
    algorithm='lbfgs',
    c1=0.1,
    c2=0.1,
    max_iterations=100,
    all_possible_transitions=True
)
crf.fit(X_train, Y_train)
# joblib.dump(crf, "crf.joblib")
Y_pred = crf.predict(X_test)
labels = list(crf.classes_)
labels.remove('O')
# labels.remove('B-che')
# labels.remove('I-che')
metrics.flat_f1_score(Y_test, Y_pred,
                      average='weighted', labels=labels)
```

	precision	recall	f1-score	support
B-body	0.917	0.858	0.887	2477
I-body	0.904	0.850	0.876	4685
B-che	0.959	0.944	0.951	2143
I-che	0.950	0.914	0.932	4507
B-sas	0.940	0.957	0.948	1575
I-sas	0.943	0.954	0.948	1698
B-tre	0.835	0.678	0.748	298
I-tre	0.913	0.812	0.859	1558
micro avg	0.930	0.889	0.909	18941
macro avg	0.920	0.871	0.894	18941
weighted avg	0.929	0.889	0.908	18941

咳(B-sas) 嗽(I-sas) 6 + 月 ， 咳(B-sas) 痰(I-sas) 1 0 + 天
头(B-sas) 痛(I-sas) 7 月 余 ， 加 重 1 月
腹(B-sas) 胀(I-sas) 2 年 ， 检 查 发 现 食 管 癌 1 周
左 乳 癌 术 后 放 化 疗 后 2 年 ， 发 现 肝(B-body) 肺(B-body) 骨(I-body) 转 移 9 月 ， 解 救 化 疗 8 周 期 后 3 月 ， 发 现 进 展 半 月

右(B-body) 肺(I-body) 癌 术 后 1 年 半
发 现 左(B-body) 侧(I-body) 下(I-body) 牙(I-body) 龈(I-body) 溃(I-body) 烂(I-body) 8 月 ， 加 重 伴 疼(B-sas) 痛(I-sas) 1 月

ScispaCy

spaCy: Industrial-strength NLP

spaCy is a library for **advanced Natural Language Processing** in Python and Cython. It's built on the very latest research, and was designed from day one to be used in real products.

spaCy comes with pretrained pipelines and currently supports tokenization and training for **60+ languages**. It features state-of-the-art speed and **neural network models** for tagging, parsing, **named entity recognition**, **text classification** and more, multi-task learning with pretrained **transformers** like BERT, as well as a production-ready **training system** and easy model packaging, deployment and workflow management. spaCy is commercial open-source software, released under the MIT license.

127 Sci spaCy

This repository contains custom pipes and models related to using spaCy for scientific documents.

In particular, there is a custom tokenizer that adds tokenization rules on top of spaCy's rule-based tokenizer, a POS tagger and syntactic parser trained on biomedical data and an entity span detection model. Separately, there are also NER models for more specific tasks.

ScispaCy——预训练模型

病例

```
clinical_note = "Patient resting in bed. Patient given azithromycin without any difficulty. Patient has audible wheezing, \
states chest tightness. No evidence of hypertension. Patient denies nausea at this time. zofran declined. \
Patient is also having intermittent sweating associated with pneumonia. Patient refused pain but tylenol still given. \
Neither substance abuse nor alcohol use however cocaine once used in the last year. Alcoholism unlikely.\
Patient has headache and fever. Patient is not diabetic. No signs of diarrhea. \
Lab reports confirm lymphocytopenia. Cardaic rhythm is Sinus bradycardia. Patient also has a history of cardiac injury. \
No kidney injury reported. No abnormal rashes or ulcers. Patient might not have liver disease. \
Confirmed absence of hemoptysis. Although patient has severe pneumonia and fever, \
test reports are negative for COVID-19 infection. COVID-19 viral infection absent."
```

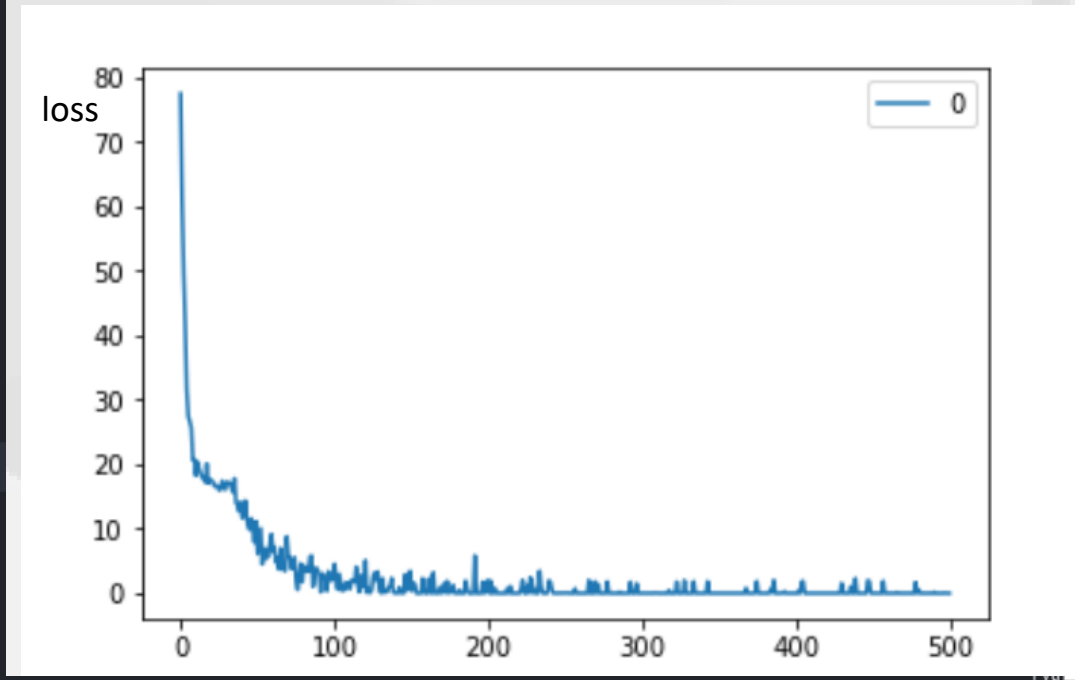
0.3s

Model	Description
en_core_sci_sm	A full spaCy pipeline for biomedical data.

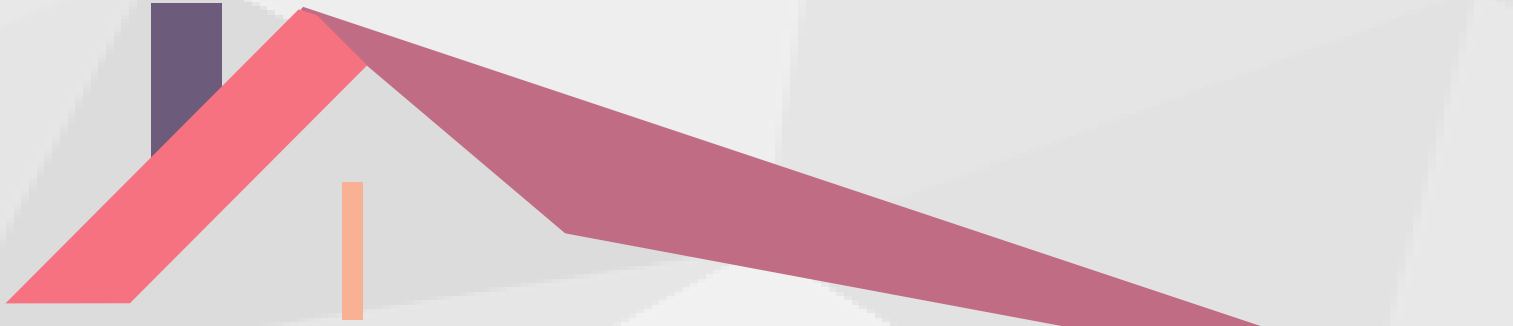
patient rest in bed . patient give azithromycin CHEMICAL without any difficulty . patient have audible wheezing DISEASE , state ch tightness DISEASE . no evidence of hypertension DISEASE . patient deny nausea DISEASE at this time . zofran CHEMICAL decline . patient be also have intermittent sweating associate with pneumonia DISEASE . patient refuse pain DISEASE but tylenol CHEMICAL still give . neither substance abuse DISEASE nor alcohol CHEMICAL use however cocaine CHEMICAL once use in the last year . alcoholism DISEASE unlikely . patient have headache DISEASE and fever DISEASE . patient be not diabetic DISEASE . no sign of diarrhea DISEASE . lab report confirm lymphocytopenia DISEASE . cardaic rhythm be sinus bradycardia DISEASE . patient also have a history of cardiac injury DISEASE . no kidney injury DISEASE report . no abnormal rash DISEASE or ulcer DISEASE . patient might not have liver disease DISEASE . confirmed absence of hemoptysis DISEASE . although patient have severe pneumonia DISEASE and fever DISEASE , test report be negative for covid-19 infection DISEASE . covid-19 viral infection DISEASE absent .

手动训练模型

```python
from spacy.training import Example
from spacy.util import minibatch, compounding
import random
pipe_exceptions = ["ner", "trf_wordpiecer", "trf_tok2vec"]
unaffected_pipes = [
    pipe for pipe in nlp.pipe_names if pipe not in pipe_exceptions]
with nlp.disable_pipes(*unaffected_pipes):
    sizes = compounding(1.0, 4.0, 1.001)
    for itn in range(100):
        random.shuffle(TRAIN_DATA)
        batches = minibatch(TRAIN_DATA, size=sizes)
        losses = {}
        for batch in batches:
            for text, annotations in batch:
                doc = nlp.make_doc(text)
```



Hepatitis DISEASE is a disease which causes inflammation of the liver and it can also cause jaundice DISEASE . Tuberculosis DISEASE is caused by a bacterium called Mycobacterium tuberculosis. AIDS DISEASE is the late stage of HIV infection that occurs when the body's immune system is badly damaged because of the virus.Typhoid is a bacterial infection that can lead to a high fever, diarrhea, and vomiting. Cancer DISEASE is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body.Chikungunya is a viral disease transmitted to humans by infected mosquitoes. Pneumonia DISEASE is an infection that inflames the air sacs in one or both lungs. Malaria DISEASE is a disease caused by a parasite.



THANK YOU