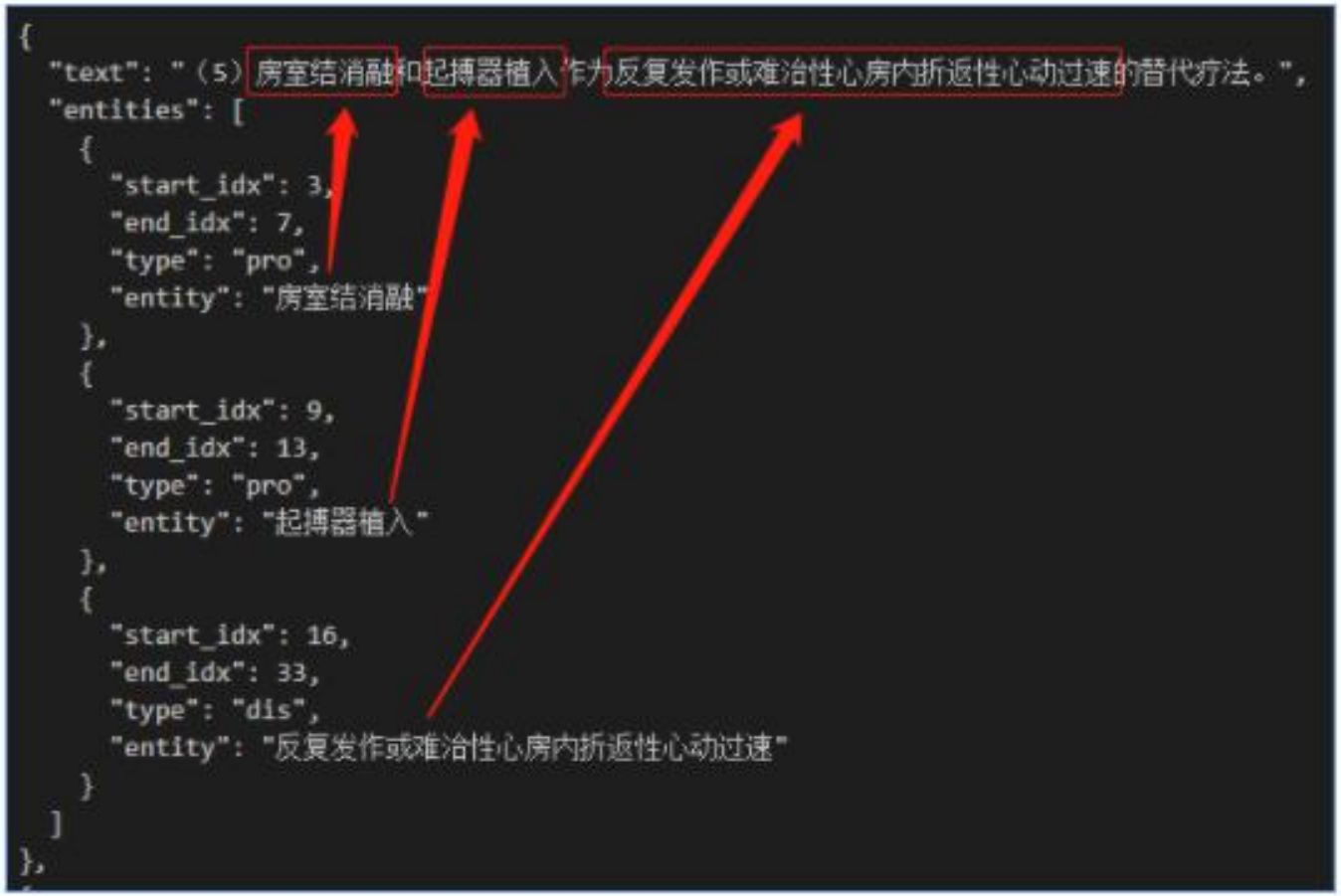
命名实体识别基础、标注方法、方法、工具、挑战 介绍

分享人：杨夕

一、什么是命名实体识别

- 介绍：从文本中自动识别出来命名的实体

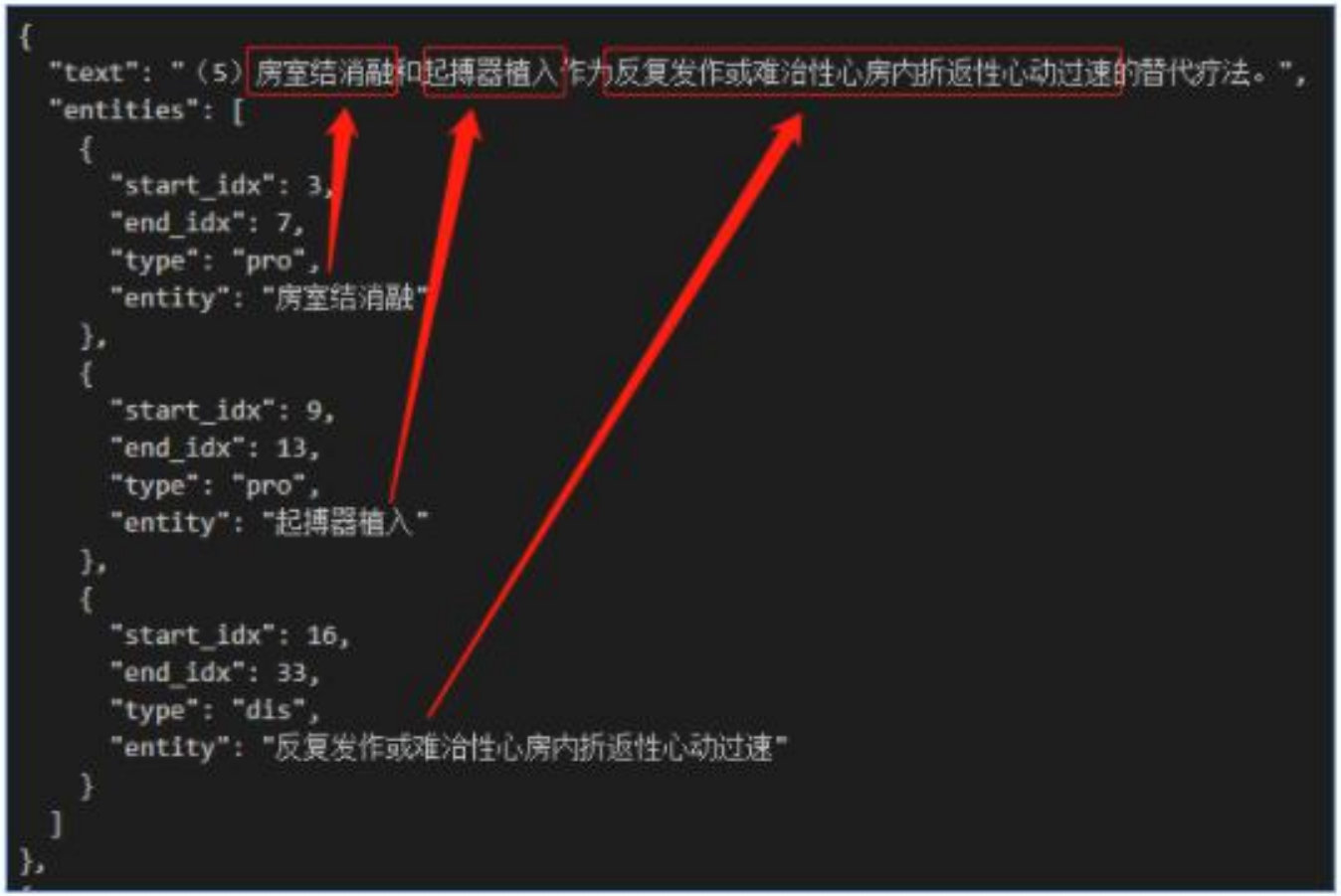


二、命名实体识别常用标注方式

标准	位置	说明	说明
IOB2	B	Begin	实体开始
	I	Inside	实体内
	O	Outside	实体外
IOBES	B	Begin	实体开始
	I	Inside	实体内
	E	End	实体结束
	O	Outside	实体外
	S	Single	单字实体

二、命名实体识别常用标注方式

- 位置标注方式



三、命名实体识别方法

- 基于规则方法
  - 代表技术：词典 + 规则
  - 核心思想：匹配规则

三、命名实体识别方法

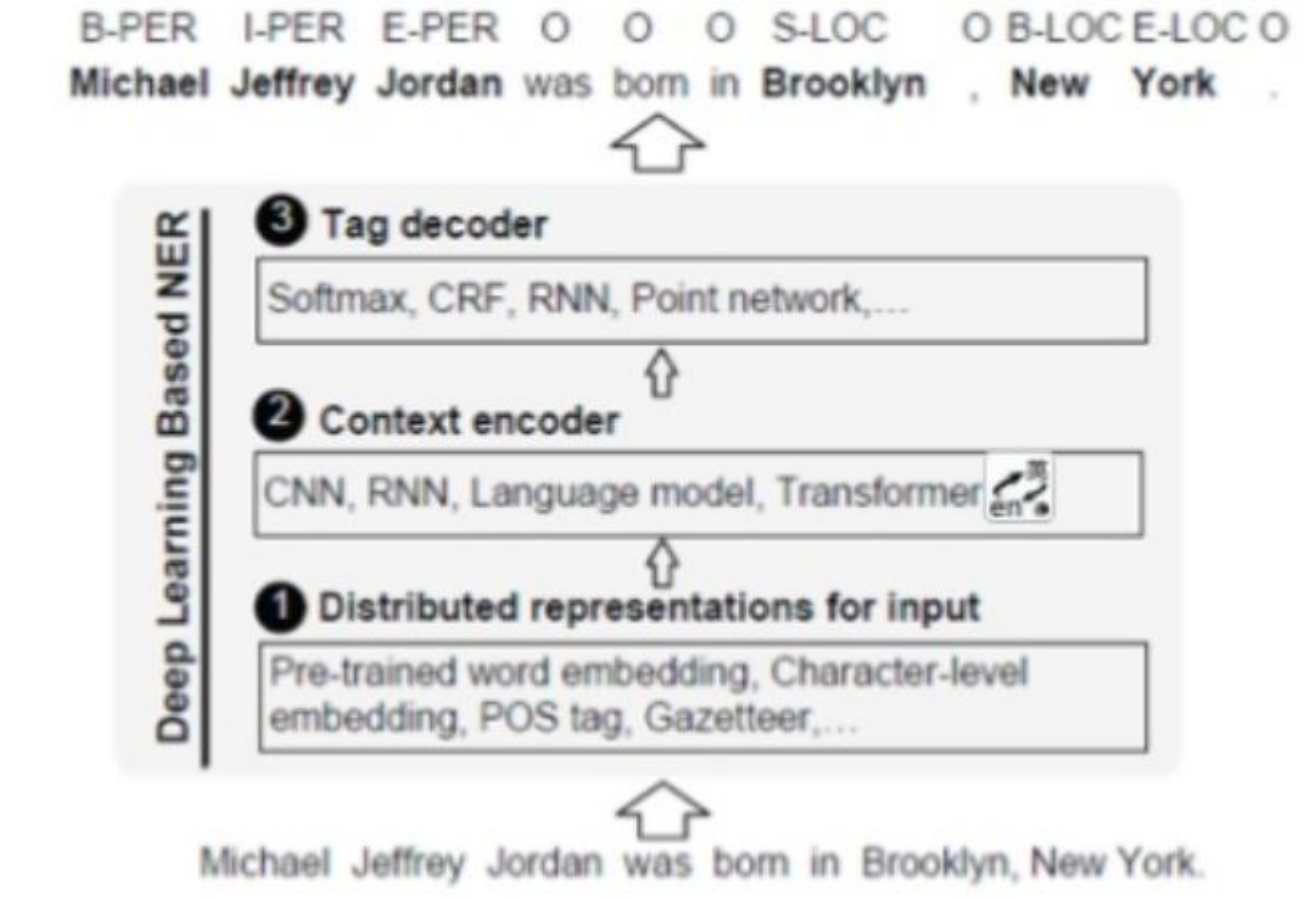
- 基于统计机器学习方法
  - 核心思想：选择概率最大的
  - 代表技术：HMM、MEMM、CRF

$$P(\text{姓名} \mid \text{刘启}) = P(X \mid \text{刘})\, P(M \mid \text{启})\, P(XM)$$

$$P(\text{姓名} \mid \text{刘启林}) = P(X \mid \text{刘})\, P(M \mid \text{启})\, P(M \mid \text{林})\, P(XMM)$$

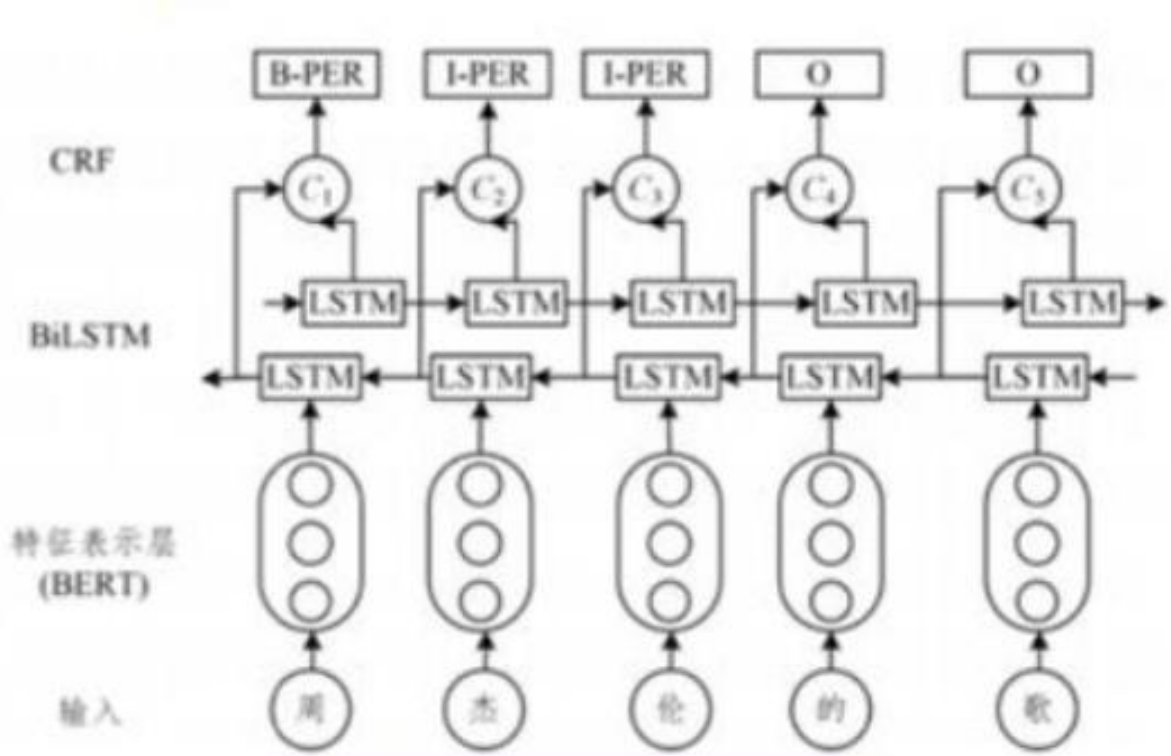
三、命名实体识别方法

- 基于深度学习方法
  - 核心思想：全部表示+选择概率最大的实体。
  - 代表技术：BiLSTM-CRF、IDCNN-CRF



三、命名实体识别方法

- 基于 Attention方法
  - 核心思想：重点表示+选择概率最大的实体
  - 代表技术：Transformer-CRF、BERT-CRF、BERT-LSTM-CRF



BERT-BiLSTM-CRF模型结构

四、命名实体识别工具

- HanLP

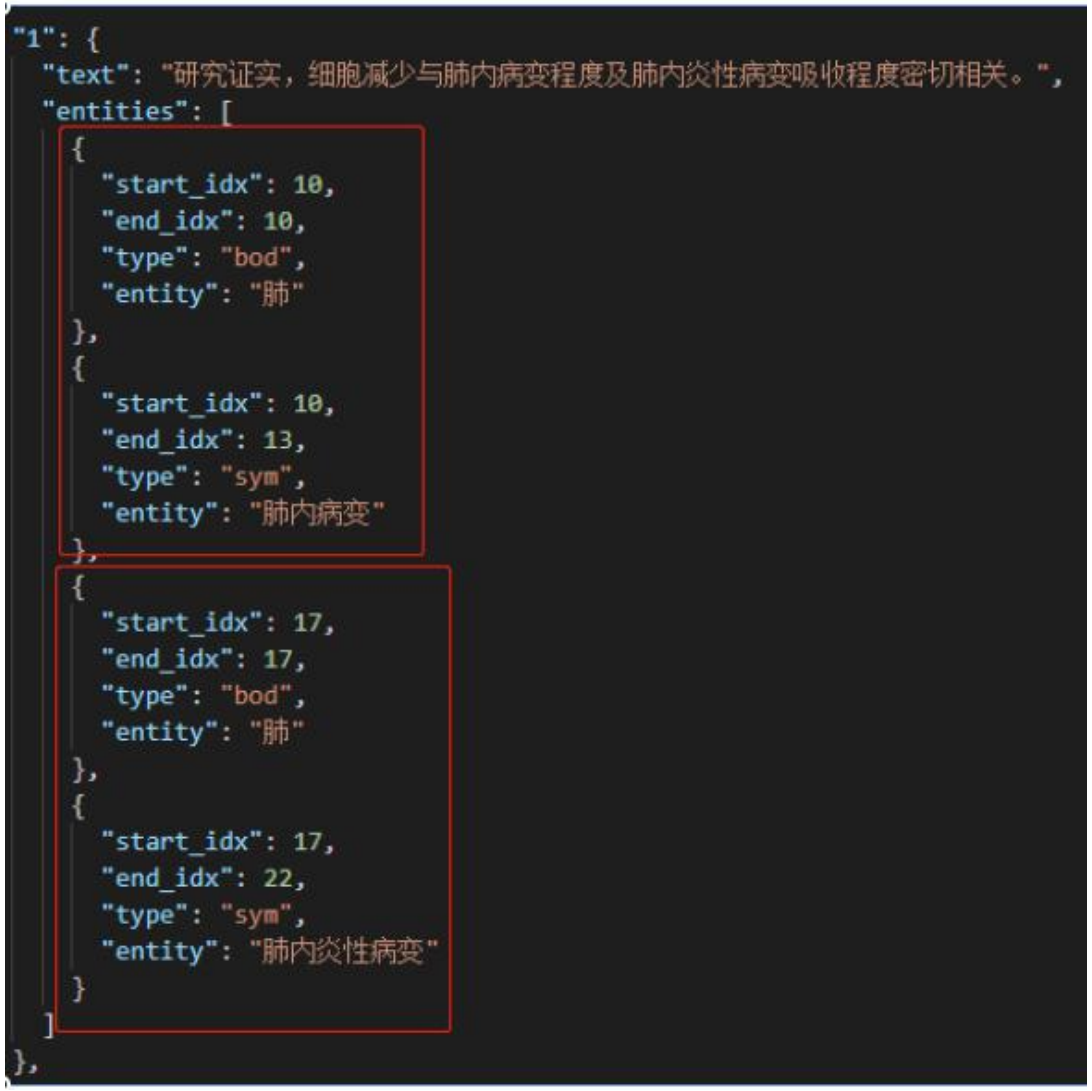
  （https://github.com/hankcs/pyhanlp）

- CRF++

  （https://github.com/taku910/crfpp）

	算法	原理	特点
HanLP 命名实体识别	中国人名识别	HMM-Viterbi	速度
	地名识别		
	音译人名识别	层叠隐马	
	日本人名识别		
	机构名识别		
	感知机命名实体识别	感知机	精度
	CFR命名实体识别	CRF	

五、命名实体识别挑战——实体嵌套问题

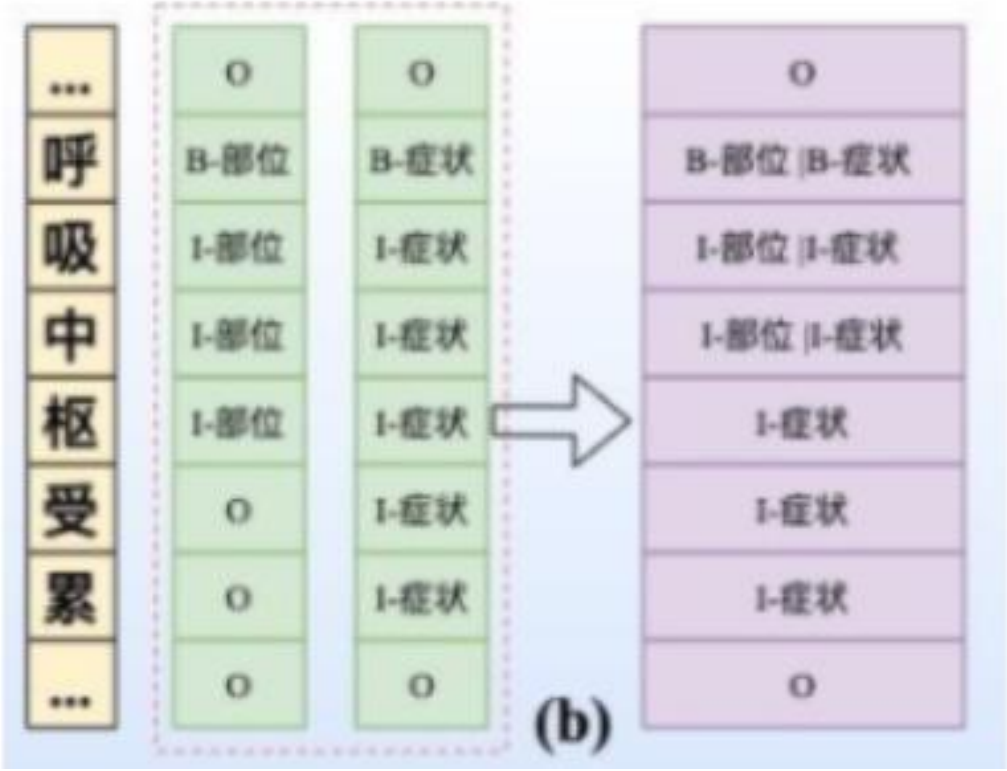


五、命名实体识别挑战——实体嵌套问题

- 解决方法一：多标签分类
  - 介绍：将分类任务的目标从单标签变成多标签
  - 缺点：
    - 学习难度较大，也会容易导致label之间依赖关系的缺失
    - 模型学习的目标设置过难，阈值定义比较主观，很难泛化；
    - 修改后的Schema预测的结果，复原回实体的时候又不再具有唯一性了

五、命名实体识别挑战——实体嵌套问题

- 解决方法二：合并标签层
  - 好处：最后的分类任务仍然是一个单分类，因为所有可能的分类目标我们都在Schema中覆盖
  - 缺点：
    - 指数级增加了标签，导致分布过于稀疏，很难学习；
    - 对于多层嵌套，需要定义非常多的复合标签；
    - 修改后的Schema预测的结果，复原回实体的时候又不再具有唯一性了；



五、命名实体识别挑战——实体嵌套问题

- 解决方法三：指针标注
  - 层叠式指针标注：设置c个指针网络

五、命名实体识别挑战——实体嵌套问题

- 解决方法三：指针标注
  - MRC-QA+指针标注：构建query问题指代所要抽取的实体类型，同时也引入了先验语义知识

五、命名实体识别挑战——实体嵌套问题

- 解决方法四：多头标注
  - 介绍：构建一个 C 的Span矩阵
  - 问题：如何构造Span矩阵、以及解决0-1标签稀疏问题
  - 方法：
    - Named Entity Recognition as Dependency Parsing
    - GlobalPointer：用统一的方式处理嵌套和非嵌套NER

五、命名实体识别挑战——实体嵌套问题

- 解决方法五：片段排列
  - 问题：对于含T个token的文本，理论上共有 =(+1)/2 种片段排列。如果文本过长，会产生大量的负样本，在实际中需要限制span长度并合理削减负样本。

六、命名实体识别挑战——实体非连续性问题

方法：合并抽取，在利用规则切分


```
{
  "text": "皮损可呈银屑病样、玫瑰糠疹样、脂溢性皮炎样、荨麻疹样、离心性环形红斑样等。",
  "entities": [
    {
      "start_idx": 0,
      "end_idx": 34,
      "type": "sym",
      "entity": "皮损可呈银屑病样、玫瑰糠疹样、脂溢性皮炎样、荨麻疹样、离心性环形红斑样"
    }
  ]
},
```

六、命名实体识别挑战——实体歧义问题

苹果：水果？企业？

Thank You!!!