SEU 实体链接

2019年12月2日 10:13

- 实体链接是指将文档中出现的文本片段，即实体指称（entity mention）链向其在特定知识库（Knowledge Base）中相应条目的过程
  - 例如：

    86年的电视剧西游记是对小说西游记最经典的改编。
  - 链接结果：

    86年的电视剧西游记（1986年杨洁执导央视版电视剧）是对小说西游记（中国古典长篇小说）最经典的改编。

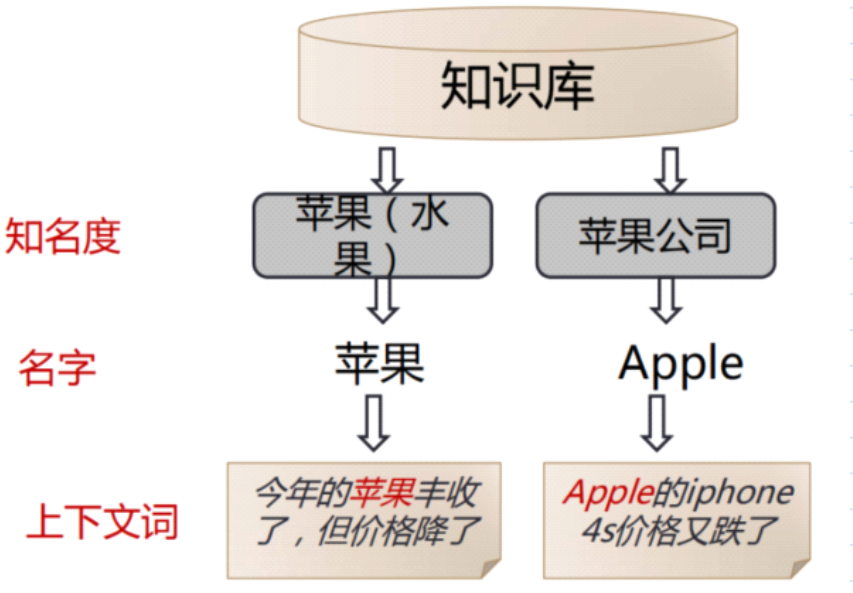
    【对其中的专有名词进行正确的标注】

- **实体链接的应用场景**
  - 文本分类和聚类
  - 信息检索
  - 知识库构建
  - 智能问答
  - ……
- **实体链接的步骤**
  - 命名实体识别
  - 词义消歧
- **实体链接的方法**
  - 基于概率生成模型的方法
  - 基于主题模型的方法
  - 基于图方法

- 基于深度学习的方法
- 无监督方法
- ……

- **基于概率生成模型的方法**
  - 人们在进行链接工作时，使用了大量关于实体的知识：
    - 实体的知名度
    - 实体的名字分布
    - 实体的上下文分布
  - 提出了实体-提及模型来融合上述异构知识
  - 一个实体的名字通常是固定的，并且以一定的概率出现
  - 指称上下文与实体越匹配，则越可能连接到对应的实体
    - 苹果 上下文包含 性能 续航等，则有可能指科技公司
    - 苹果 上下文包含 口感 色泽等，则有可能指水果
  - 利用M&W相似度可以计算出候选实体与上下文中其他实体的相关性

$$\rho^{MW}(a,b) = 1 - \frac{\log(\max(|in(a)|,|in(b)|)) - \log(|in(a) \cap in(b)|)}{\log(|W|) - \log(\min(|in(a)|,|in(b)|))}$$



基于上述模型，实体e是提及m目标实体的概率

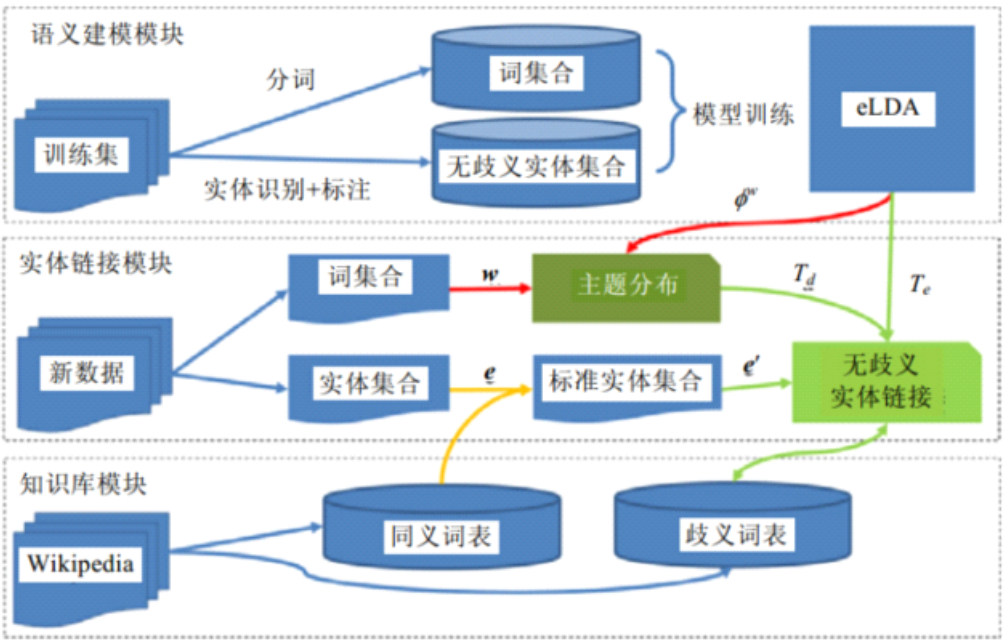
$$P(m,e) = P(s,c,e) = P(e)P(s \mid e)P(c \mid e)$$

基于上述模型，实体e是提及m目标实体的概率

$$P(m,e) = P(s,c,e) = P(e)P(s \mid e)P(c \mid e)$$

知名度　名字概率　上下文概率

- **基于主题模型的方法**
  - 强假设：
  - 同一篇文本中的实体应当与文本的主题相关：
    - 和科技、手机等有关文章也更有可能苹果公司，而不是水果
- **基于主题的知识推理**



- **基于图的方法**
  - 实体相似度计算
    - 根据实体属性值的数据类型使用不同相似度计算方法来度量他们之间的相似性再使用聚合函数初始化实体间的相似度矩阵
  - 图模型构建
    - 根据实体类型，基于中计算的得到的相似度确定候选链接单元，将所有候选单元作为关联图中的顶点，再基于各实体间的语义关系，确定候选链接单元间的关联（生成关联图中的边）

- 重启随机游走

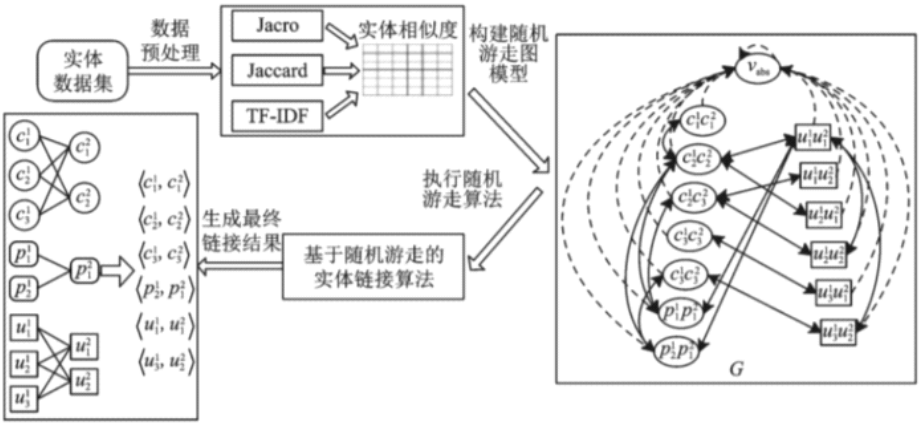


图 1　基于随机游走的实体链接模型

- 重启随机游走

  基本思想就是给定一个图，游走者从某个顶点或一些列顶点开始遍历该图。在任意一个顶点，游走者对于下一步行动有两种选择：

  - 以概率1-c随机选择一条关联到当前顶点的边以游走到某一个邻居顶点
  - 以概率c随机跳转到图中任意一个点。每次游走后，均将得到一个概率分布，将该概率分布作为下一次游走的输入，反复迭代。当满足一定前提条件时，该概率分布将会收敛到一个稳定值

- 基于深度学习的方法

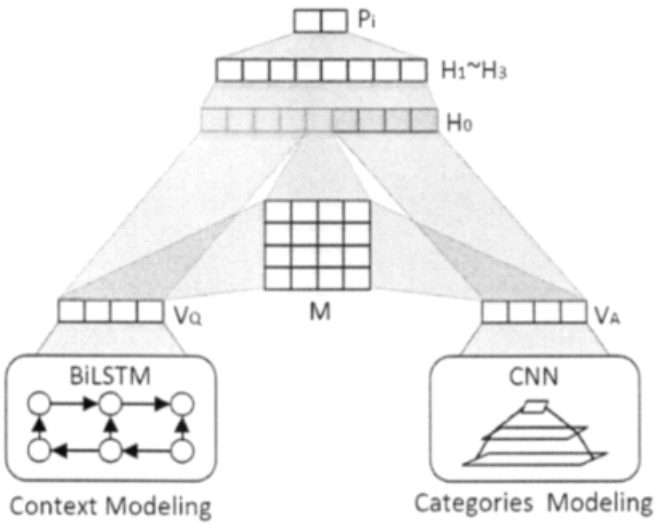


图 4.6 文本实体链接神经网络整体架构

- 借助BiLSTM、CNN等计算相似度后再基于图：

表 4.5 中文数据集上的实体链接对比试验结果

数据集	用例总数	文档总数	候选集平均大小

- **借助BiLSTM、CNN等计算相似度后再基于图：**

表 4.5 中文数据集上的实体链接对比试验结果

数据集		用例总数		文档总数		候选集平均大小
训练集		743,978		232,493		6.032
测试集		55,716		15,058		5.966
LIEL		DSRM		Ours		
Micro	Macro	Micro	Macro	Micro	Macro	
0.7063	0.7189	0.7434	0.7296	**0.8107**	**0.8211**	
DSRM + Ours		Prior		Prior + RWR		
Micro	Macro	Micro	Macro	Micro	Macro	
0.7776	0.7821	0.6983	0.6844	0.7191	0.7123	

  - 基于无监督的方法

    AAAI 18 阿里Colink

    - **协同训练算法**
      - 在该框架中定义两个不同的模型：一个基于属性的模型fatt和一个基于关系的模型frel
      - 这两个模型会进行二元分类预测，将一组给定实体对分类为正例（链接的）或负例（非链接的）
      - 该协同训练算法以迭代的方式不断增强这两个模型