Web信息处理与应用

第十三节

知识抽取与表达（下）

徐童

2023.11.27

- **信息抽取的内容**

- 核心的8字方针：“**抽取实体，确定关系**”
  - 实体：即命名实体，指文本中的基本构成块，如人、机构等
  - 属性：实体的特征，如人的年龄、机构的类型等
  - **关系**：实体之间存在的联系，也称事实，如公司和地址之间的位置关系、公司与人之间的雇佣关系
  - 事件：实体的行为或实体参与的活动

- **信息抽取的基本任务**

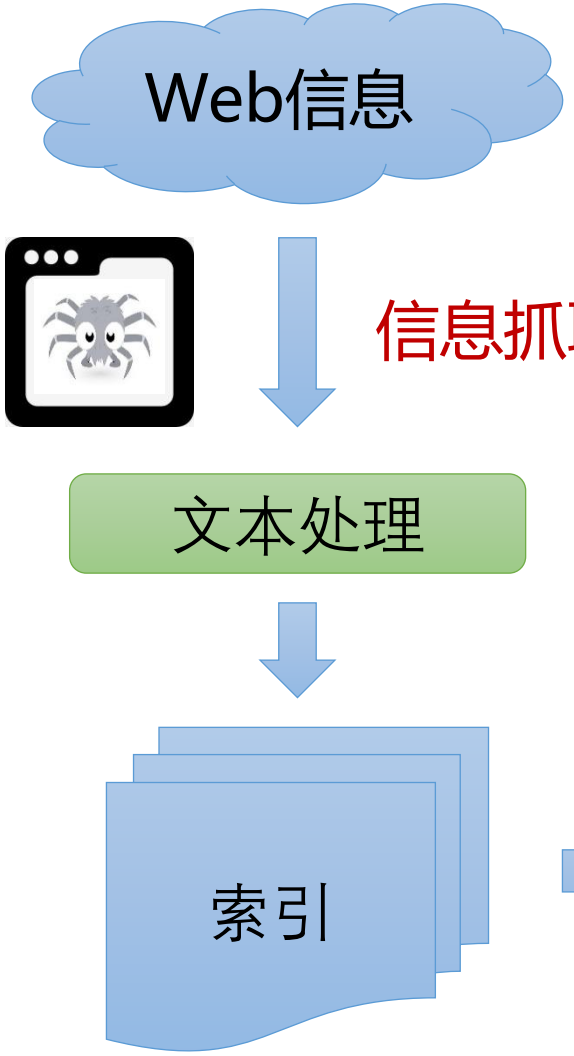
- **模板关系TR**（关系抽取）
- 实体之间的各种关系，又称为事实
- 通过关系抽取，将实体关联起来，并为推理奠定基础
  - 例如，职务（Post_of）、雇佣关系（Employee_of）、生产关系（Product_of）等
    - 如：
      - Post_of (老板，郝哥)
      - Employee_of(水果摊，郝哥)

- **信息抽取的基本任务**

- **模板元素TE**（属性抽取）
- 模板元素又称为实体的属性，目的在于更加清楚、完整地描述命名实体
- 通过槽（Slots）描述了命名实体的基本信息
  - 槽：名称、类别、种类等
    - 例如：郝哥表示，这都是大棚的瓜，你嫌贵我还嫌贵呢。
      - TE：瓜是大棚里产的（属性）

本节主题

- 本课程所要解决的问题



第十二个问题：

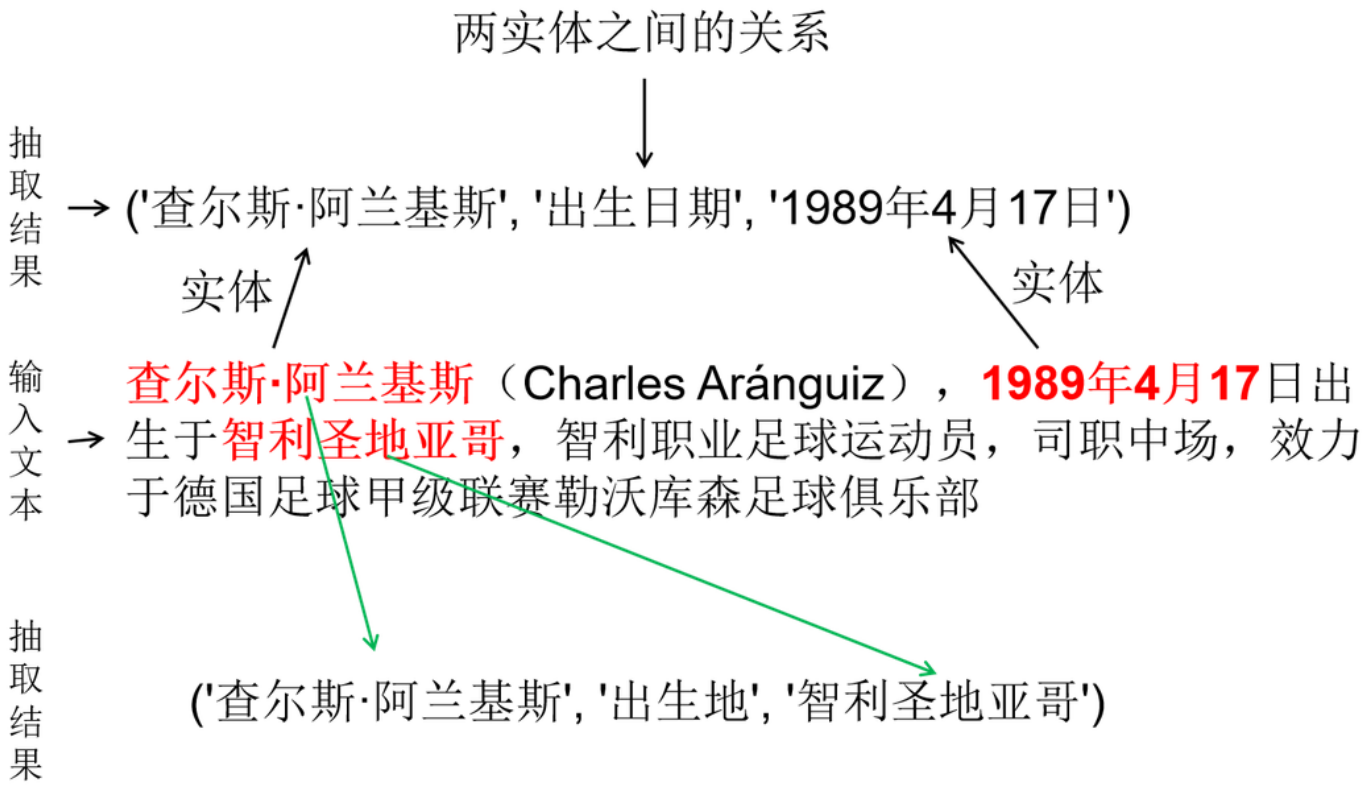
如何从文档中提取关系?

本节目录

- 关系抽取
  - 关系抽取方法
  - 远程监督方法
  - 开放关系抽取

- **关系抽取背景和定义**

- 关系抽取的概念是1988年在MUC大会上提出的，是信息抽取的基本任务之一，旨在识别出文本实体中的目标关系，是构建知识图谱的重要技术环节。



- **基本的关系抽取方法**

- 基本的关系抽取方法可大致分为以下三类
  - 基于规则的关系抽取
    - 纯手工定制规则，通过匹配从文本中寻找关系
  - 基于模式的关系抽取
    - 从种子关系中获得模式，再由模式寻找更多种子，迭代优化
  - 基于机器学习的关系抽取
    - 将关系抽取问题转化为分类问题，通过训练模型加以求解

- **基于规则的关系抽取**

- 根据欲抽取关系的特点，首先基于已有知识，手工设定一些词法、句法和语义模式规则，然后再从自由文本中寻找相匹配的关系实例类
  - 基于规则的关系抽取，需要从文本中寻找体现特定含义的规则
  - 例如：<X, IS_A, Y>关系抽取（同义词/上下位词关系）
    - Dog is a member of canid.（狗是犬科家族的一员）

```
dog, domestic dog, Canis familiaris
   => canine, canid
      => carnivore
        => placental, placental mammal, eutherian, eutherian mammal
          => mammal
            => vertebrate, craniate
              => chordate
                => animal, animate being, beast, brute, creature, fauna
                  => ...
```

- **基于规则的关系抽取**

- 描述前述的<X, IS_A, Y>关系，可采用以下规则
  - Rule 1: “Y such as X”： Universities such as MIT and CMU ……
  - Rule 2: “X or other Y”： Apples or other fruits ……
  - Rule 3: “Y including X”： Machine learning methods including SVM and CRF……
  - Rule 4: “Y, especially X”： Most students, especially Ph.D. candidates ……
  - ……

- **基于规则的关系抽取**

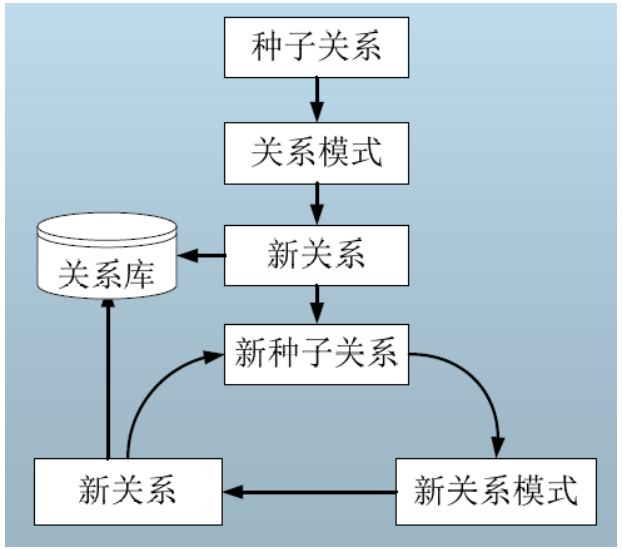
- 基于规则方法的优点：
  - 无需训练，实现简单
  - 人工规则有高准确度
  - 可以针对特定的垂直领域
  - 在小规模数据集上容易实现

- **基于规则的关系抽取**

- 基于规则方法的缺点：
  - 通常针对特定领域的特定关系抽取任务，可以根据想抽取的关系的特点设计针对性的规则，但部分任务可能很难制定规则。
  - 基于手工规则的方法需要领域专家构筑大规模的知识库，这不但需要有专业技能的专家，也需要付出大量劳动，因此这种方法的代价很大。
  - 知识库构建完成后，对于特定领域的抽取具有较好的准确率，但移植到其他领域十分困难，效果往往较差。

- **基于模式的关系抽取**

- 首先由种子关系生成关系模式，然后基于关系模式抽取新的关系，得到新关系后，从中选择可信度高的关系作为新种子，再寻找新的模式和新的关系。
  - 如此不断迭代，直到没有新的关系或新的模式产生。



套娃?

- **基于模式的关系抽取**

- 代表性方法1：DIPRE
  - 双重迭代模式关系提取（*Dual Iterative Pattern Relation Extraction*），由谷歌联合创始人Sergey Brin于1998年提出。（同年，PageRank诞生）
  - 其大致思路在于先给定一些已知关系类型的种子实体对，找到出现了这些实体对的Occurrences，再学习Occurrences的模式（Pattern）。
  - 进而，根据学到的模式，寻找更多符合该模式的数据，并加入到种子集合中，不断迭代这个过程以实现关系抽取。

- **基于模式的关系抽取**

- DIPRE的基本元素

  小问题：这里为什么是二元组的形式？

  - 元组：表示关系实例，如<Foundation, Isaac Asimov> — <Title, Author>
  - 模式：包含常量和变量，例如 ?x , by ?y的形式（可表示“title” by “author”）
- DIPRE的基本假设
  - 元组往往广泛存在于各个网页源中
  - 元组的各个部分往往在位置上是接近的
  - 在表示这些元组时，存在着某种重复的“模式”

- **基于模式的关系抽取**

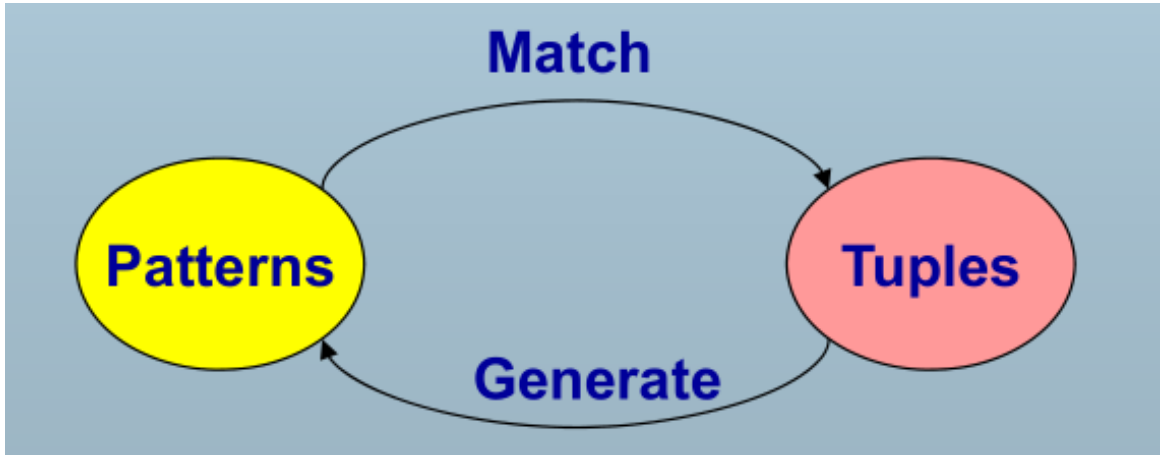
- 在表示这些三元组时，存在着某种重复的“模式”。

- **基于模式的关系抽取**

- 一种启发式方法（类似基于规则的基本方法）
  - 通过检查部分网站来获取潜在的“模式”，并利用正则表达式来描述
    - 例如：letter = [A-Za-z.]，title = letter{5,40}，author = letter{10,30}，(title) by (author)
  - 这种方法的缺陷是明显的（与基于规则的方法类似）：
    - 网站/系统存在特殊性，某个网站上的模式未必适用于其他网站
    - 与规则类似的问题：人类不可能穷尽所有的潜在模式

- **基于模式的关系抽取**

- 更好的方法：DIPRE正式登场
  - 考虑模式和元组实例之间的双重影响关系
    - 既要找到符合模式的元组，也要找到用于生成元组的模式



• 基于模式的关系抽取

- 从更多的元组实例中提炼种子元组，再去构建新的模式

北京邮电大学周安福教授学术报告会

发布时间：2020-09-10 浏览次数： 245

计算机科学技术专家讲座（九）——周安福

发布日期：2022-05-30 发布人： 点击量： 222

报告题目：基于强化学习的低延迟视频传输研究

报告时间：2022年6月2日，9:30

报告方式：腾讯会议

会议码：553-66[illegible]948

报告人：周安福 北京邮电大学教授

报告人简介：

北京邮电大学计算机学院教授、博士生导师，入选国家级青年人才计划。他的研究方向为物联网与移动计算，包括毫米波感知、毫米波传输、AI-driven低延迟视频传输等。先后承担了国家自然科学基金、国家重点研发等科研项目，研究成果发表在ACM SIGCOMM、ACM MobiCom、USENIX NSDI、IEEE/ACM Trans. on Networking、IEEE Trans. on Mobile Computing等一流国际会议和期刊上，部分成果已在产业界大规模应用。先后获得中国计算机学会CCF-Intel青年学者奖、ACM中国新星奖、阿里巴巴优秀学术合作奖、中国电子学会科技进步一等奖等。

- **基于模式的关系抽取**

- DIPRE的算法流程
  - 首先，输入一组种子元组实例R，如若干<title, author>的实体对
  - 其次，基于种子实例集合R，找到这些元组在网页中出现的内容O（Occurrence），注意寻找的时候保留上下文信息（Surrounding Context）
  - 进而，基于找到的元组实例O，生成模式P
  - 最后基于生成的模式，找到更多的元组实例R
    - 此时可选择停止，或返回第二步继续基于新实例生成新模式
      - 注意，此时生成的新模式可能与之前的模式有所差异!

- **基于模式的关系抽取**

- Occurrence的概念与实例
  - Occurrence的直译为“出现”，可以理解成元组在网页中的呈现形式
  - 一般而言，只有元组的元素在网页中非常接近，才视作Occurrence
    - 避免因间隔太远而可能导致的语义不相关问题

Occurrence→

 Foundation by Isaac Asimov (1951)

- url = http://www.scifi.org/bydecade/1950.html
- order = [title,author] (or [author,title])
  - denote as 0 or 1
- prefix = “ ” (limit to e.g., 10 characters)
- middle = “ by ”
- suffix = “(1951) ”
- occurrence =

('Foundation', 'Isaac Asimov', url, order, prefix, middle, suffix)

- **基于模式的关系抽取**

- 模式的概念与实例
  - 对于同一关系的不同实例在网页上所呈现的不同Occurrence，将相同内容保留下来，不同内容采用**通配符**取代，即可得到近似的模式

 Foundation by Isaac Asimov (1951)
<p> Nightfall by Isaac Asimov (1941)

- order = [title, author] (say 0)
- shared prefix =
- shared middle = by
- shared suffix = (19
- pattern = (order, shared prefix, shared middle, shared suffix)

- **基于模式的关系抽取**

- URL的前缀（Prefix）所起到的潜在作用
  - 如前所述，模式往往仅限特定网站/体系内，跨网站则模式可能不适用
  - 如何判定属于同一个网站/体系内？
- 回顾：网页排序部分提到的Hilltop算法的基本概念之一：非从属组织网页
  - 满足以下两种情况，将被视作从属组织网页
    - 主机IP地址的前三个字段相同，如182.61.200.X （百度）
    - URL中的主域名段相同，如 XXX.ustc.edu.cn

- **基于模式的关系抽取**

- 将URL的前缀（Prefix）引入模式中，用于描述模式的限定范围

http://www.scifi.org/bydecade/1950.html occurence:
 Foundation by Isaac Asimov (1951)

http://www.scifi.org/bydecade/1940.html occurence:
<p> Nightfall by Isaac Asimov (1941)

shared *urlprefix* = http://www.scifi.org/bydecade/19
pattern = (urlprefix,order,prefix,middle,suffix)

← 仅限此类网站可使用该模式

- **基于模式的关系抽取**

- 基于DIPRE算法，生成模式的基本步骤
  - 首先，将Occurrence归纳为Order（元素的顺序）和Middle（中间部分）
  - 其次，定义模式如下：
    - 模式的Order和Middle，即为 Occurrence 集合的Order和Middle
    - 模式的URLPrefix、Prefix、Suffix，分别为Occurrence集合中最长的公共（Shared）URL前缀与前、后缀。
      - 其他部分采用通配符填充

- **基于模式的关系抽取**

- 代表性方法2：Snowball
  - Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections, Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000: 85-94.
  - 基本思想在于对DIPRE算法的提升
    - 仅信任支持度和置信度较高的模式，从而保证模式质量

- **基于模式的关系抽取**

- 代表性方法2：Snowball
  - 支持度与置信度的计算方式
    - 支持度（Support），即满足每个模式的元组的数量
      - 将少于一定数量元组支持的模式予以删除
    - 置信度（Confidence），按照如下公式计算：

$$Conf(P) = \frac{P.positive}{(P.positive + P.negative)}$$

      - 即考虑符合该模式的元组，确实符合相应关系的概率

- **基于模式的关系抽取**

- 代表性方法2：Snowball
  - 置信度的计算实例（来自于原论文）
    - 例如，基于模式P = <{}, ORGANIZATION, <“,”, 1>, LOCATION, {}>
    - 可以得到以下三个实例
      - “Exxon, Irving, said”
      - “Intel, Santa Clara, cut prices”
      - “invest in Microsoft, New York-based analyst Jane Smith, said”
    - 其中，前两个符合原关系，而最后一个与事实不符，因此为Negative，置信度为2/3.

- **基于模式的关系抽取**

- 基于模式方法的优缺点：
  - 不同算法的差异主要在于模式生成方法和匹配方法。
  - 适合某种特定的具体关系的抽取，如校长关系、首都关系。
  - 基于字面的匹配，没有引入更深层次的信息，如词性、句法、语义信息等。
  - 难以确保模式的可靠性，需要人工复核。
  - 移植性差，必须为每一个具体的关系生成自己的识别模式。

- **基于机器学习的关系抽取**

- 采用机器学习方法关系抽取模型，先通过标注语料库训练得到一个判别模型，再利用该模型对自由文本中出现的关系实例进行识别。
- 往往将关系抽取问题变换为一个分类问题（二分类或者多分类），然后采用机器学习中常用的分类器来解决。
  - 通常采用基于特征或基于核函数的方法加以解决

- **基于机器学习的关系抽取**

- 基于特征的方法
  - 基于特征向量，然后使用SVM 、最大熵（ME ）等进行分类
  - 关键在于特征集的确定而不是机器学习方法
  - 难点在于如何找出适合关系抽取的、有效的词汇、句法或语义特征
    - 常用的特征包括单词本身、词性、分析树或依存树等。

2013年4月20日8时02分四川省雅安市[芦山县]$_{e1}$ 发生了7.0级[地震]$_{e2}$

震中 (e1,e2)

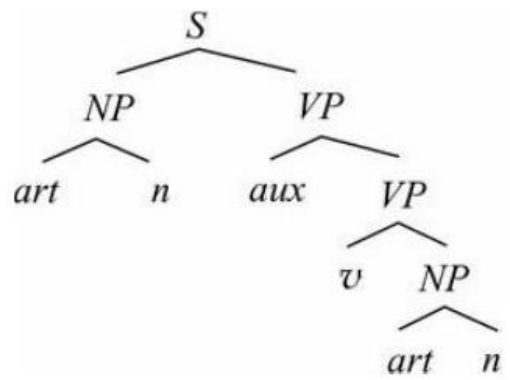
Words: *芦山县*$_{m11}$, *地震*$_{b1}$, *发生*$_{b2}$, *在*$_{21}$

Entity Type: $Noun_{m1}$, $Location_{m2}$

Parse Tree: *Location-VP-PP-Noun*

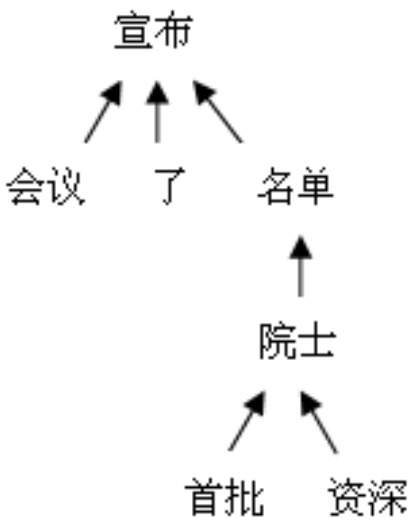
- **基于机器学习的关系抽取**

- 引申知识：句法分析树
  - 句法结构分析是指对输入的单词序列（一般为句子）判断其构成是否合乎给定的语法，分析出合乎语法的句子的句法结构。
  - 句法结构一般用树状数据结构表示，通常称为句法分析树。
  - 例如，句子The can can hold the water的分析树如右图：
    - 其中，S表示句子，NP和VP表示名词/动词短语
    - art表示冠词，n表示名词
    - aux表示助动词，v表示动词



- **基于机器学习的关系抽取**

- 引申知识：句法依存树
  - 句法依存树用于描述各个词语之间的依存关系。也即指出了词语之间在句法上的搭配关系，这种搭配关系是和语义相关联的。
  - 句法依存树的每个结点都是一个词语，需要分析识别句子中的“主谓宾”、“定状补”等语法成分。
  - 例如，“会议宣布了首批资深院士名单”的依存树如右图所示，可知词“宣布”支配“会议”、“了”和“名单”，故可以将这些支配词作为“宣布”的搭配词。

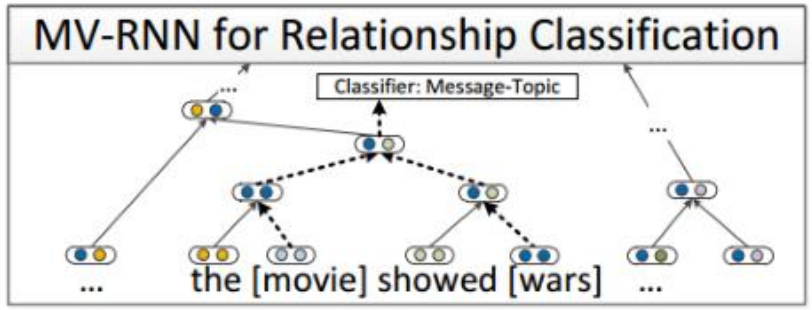


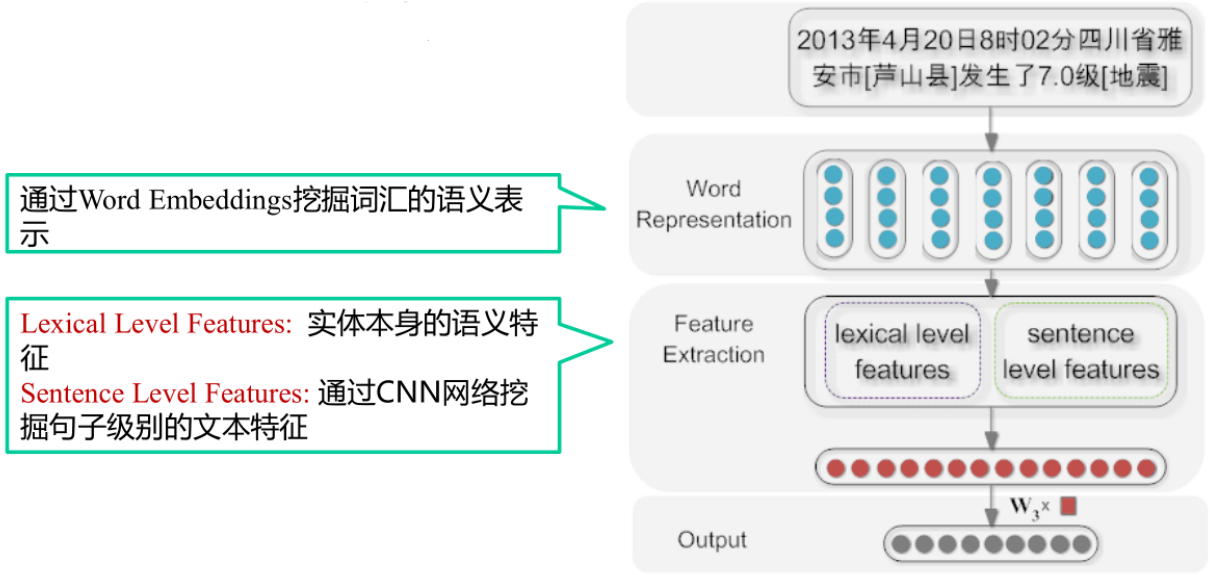
- **基于机器学习的关系抽取**

- 基于人工特征的方法存在若干缺陷
  - 对于缺少NLP处理工具和资源的语言，无法提取文本特征
  - NLP处理工具引入的“错误累积”
  - 人工设计的特征不一定适合当前任务

- **基于机器学习的关系抽取**

- 基于深度学习技术，可以在一定程度上摆脱对于人工特征的依赖
  - 基于CNN技术学习文本语义特征，同时保持句子级别的结构信息
    - D Zeng et al., Relation classification via convolutional deep neural network, COLING 2014



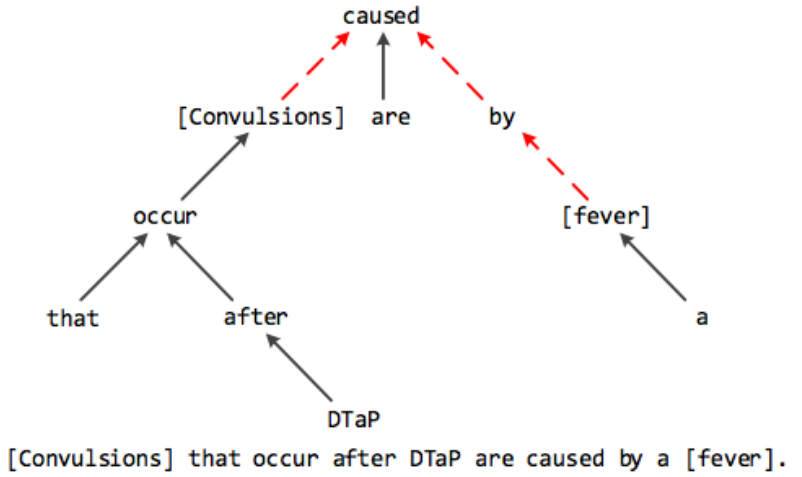


- **基于机器学习的关系抽取**

- 基于核函数的方法
  - 不需要构建特征向量，而是使用核函数来计算两个关系实例的相似性。
  - 核函数的概念：
    - 某些样本在低维空间时线性不可分，通过非线性映射将其映射到高维空间的时候则线性可分，但非线性映射的形式、参数等难以确定。
    - 核函数的目的，在于将高维空间下的内积运算转化为低维空间下的核函数计算，从而避免高维空间可能遇到的“维度灾难”问题。

- **基于机器学习的关系抽取**

- 基于核函数的方法
  - 在关系抽取时，往往是利用句子的结构特性进行抽取
    - 通常，将对关系实例表示为某种结构的树（例如句法树），并通过计算与结构相关的核函数的值来计算实例之间的相似度。
    - 例如， Bunescu等人提出两个实体之间的关系，可以由其依存树图上两个节点之间的最短路径作为核函数来加以判别。



- Convulsions和Fever，红色路径表示依存路径↑

本节目录

- **关系抽取**
  - 关系抽取方法
  - **远程监督方法**
  - 开放关系抽取

- **远程监督的由来和意义**

- 面向文本的关系抽取方法，最大的难点在于获取足够数量的、高质量的标注
  - 人工标注开支过高，且存在主观性问题；而算法标注可能有累积误差
- 如何借助某种启发式方法，方便快捷地扩充训练数据？
  - M Mintz, et al., Distant supervision for relation extraction without labeled data, ACL 2009
  - **远程监督**的思想：如果某个实体对之间具有某种关系，那么，所有包含这个实体对的句子都是用于描述这种关系。

- **远程监督的基本思路**

- 例如，我们已知“马云”和“阿里巴巴”之间是创始人/董事长关系
- 那么，我们默认以下包含这一实体对的句子，均描述这一关系
  - “马云再谈悔创阿里巴巴：再有一次机会,尽量不把公司做这么大”
  - “马云：不当阿里巴巴董事长,但绝不等于我不创业了 ”
  - “港交所披露阿里巴巴集团招股书：马云持股6.1%”
- 接下来，我们将这些语料打包，从中训练用于关系识别的模型，并进而用于判断更多的实体对之间的关系。
  - 某种意义上，这一迭代思路类似于前面介绍的DIPRE算法。

- **远程监督的局限性**

- 从上面的例子中我们可以看到，这一过程具有非常明显的局限性
  - 语义漂移（Semantic Draft）现象：不是所有包含该实体对的句子都表达该关系，错误模板会导致关系判断错误，并通过不断迭代放大错误
    - 例如，如果基于“港交所披露阿里巴巴集团招股书：马云持股6.1%”这个句子进行训练，那么所有控股关系都会被错判为“创始人”。
  - 如何解决?
    - 可通过人工校验，在每一轮迭代中观察挑出来的句子，把不包含这种关系的句子剔除掉，但开支实在过高。

- **远程监督的优化方案**

- 远程监督的优化方案（1）：动态转移矩阵
  - 尽管噪音数据不可避免，但是对噪音数据模式进行统一描述是可能的。
    - 例如，一个人的工作地点和出生地点很有可能是同一个地点，这种情形下远程监督就很有可能把born-in和work-in这两个关系标签打错。
  - 解决方案：引入一个动态转移矩阵，描述各个类之间相互标错的概率。
  - 在利用算法得到的关系分布的基础上乘以这一转移矩阵，即可得到相对更为准确的关系分类结果。
    - 然而，采用这种方式随机性较高，并不能完全保障其可靠性。

- **远程监督的优化方案**

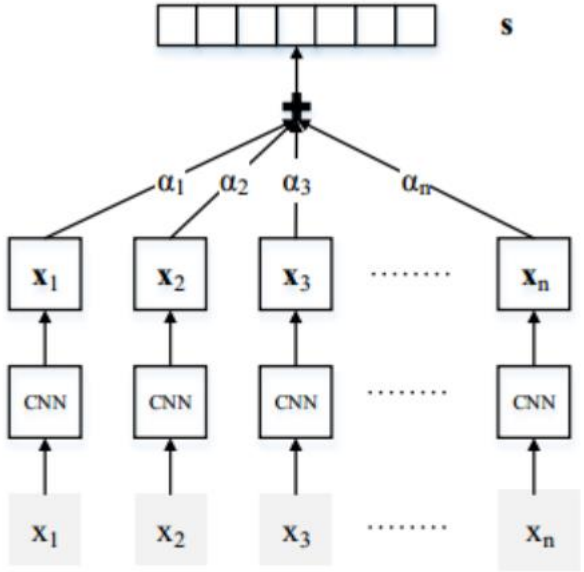
- 远程监督的优化方案（2）：规则学习
- 远程监督试图通过使用知识库作为监督来源，从文本中提取实体之间的关系。这种启发式方法因噪声的存在可能会导致一些句子被错误地标记。
- 针对这一问题提出一个新的生成模型，直接模拟远程监督的启发式标签过程。
  - 其中，设计相应的否定模式列表NegPat(r)，专门用于去除错误的标签，即某些关系的判断是否为错误。
  - 对于单一关系的判断，可以通过这种方式进行比较高效的复检。但规则的生成成本较高。

- **远程监督的优化方案**

- 远程监督的优化方案（3）：注意力机制
  - 即使是被打入同一个包里的句子，不同句子对于训练关系判别模型的贡献度也不相同，这一贡献度可以采用注意力模型加以衡量。
  - 采用深度学习技术，获取对于整个句子的表示。
  - 进而，通过注意力机制，将最能表达这种关系的句子们挑选出来。
    - 最为有效的办法，但依赖于一个高质量的样本集合。

- **远程监督的优化方案**

- 远程监督的优化方案（3）：注意力机制
  - Y Lin, et al., Neural Relation Extraction with Selective Attention over Instances，ACL 2016
  - 首次把注意力机制引入关系抽取的远程监督算法，刷新了当时的SOTA。
  - 把注意力机制和CNN句子编码器结合，有效地缓解了远程监督算法中的样本错误标注问题。
  - 论文作者中有知乎网红刘知远老师。



参考资料：https://zhuanlan.zhihu.com/p/521881255

本节目录

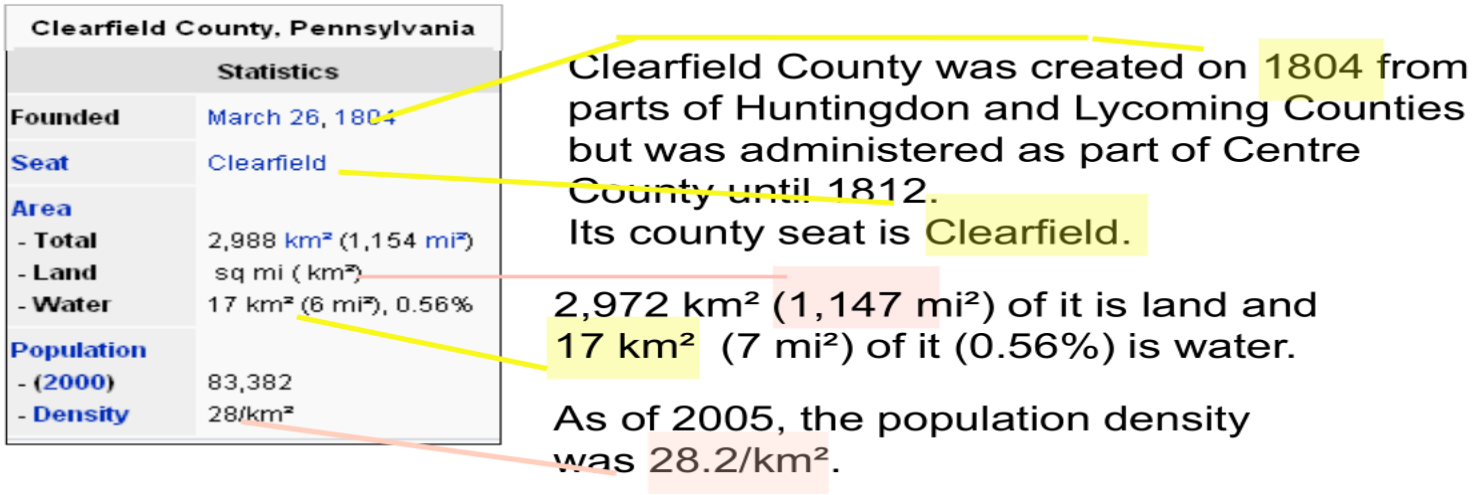
- **关系抽取**
  - 关系抽取方法
  - 远程监督方法
  - **开放关系抽取**

- **从模板关系到开放关系**

- 前面所介绍的关系抽取任务，往往针对预先定义好的关系。
  - 例如，MUC-7中有关模板关系的定义，如employee_of 等
- 然而，海量网络文本资源往往包含着更为复杂、丰富的实体关系类型，预先定义的模板关系已无法涵盖。
  - 同时，现有关系抽取研究受到关系类型与训练语料的双重限制。
- 突破封闭的关系类型限定与训练语料约束，从海量的网络文本中抽取更为丰富的实体关系三元组，已成为当下的热门需求。

- **基于知识监督的开放关系抽取**

- 一种思路是通过Wikipedia等结构化知识库，从文本中抽取关系信息
  - F Wu, et al., Autonomously Semantifying Wikipedia, CIKM 2007
  - 着重利用Wikipedia中的InfoBox，抽取已知的关系信息
  - 基于关系信息，对维基百科条目文本进行标注，产生训练语料



  - 参考资料：赵军老师《开放域事件抽取》报告

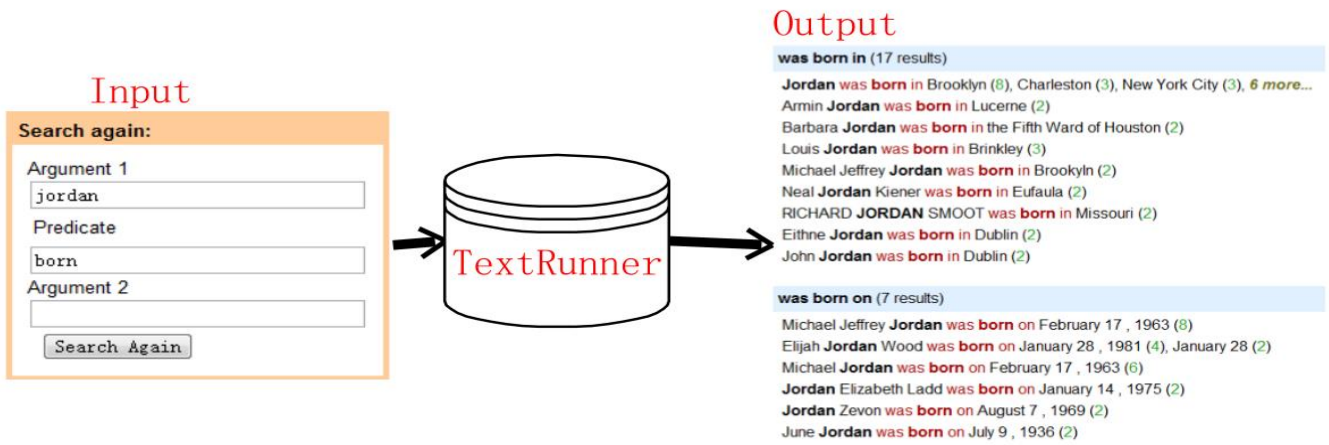
- **基于句法的开放关系抽取**

- 另一种思路，通过识别表达语义关系的短语，可以来抽取实体之间的关系
  - 可以采用类似DIPRE算法的思路，从抽取出的语料中提炼模式
  - 对于抽取出来的三元组，可通过句法和统计数据等来实现过滤
    - 关系短语应当是一个以动词为核心的短语
    - 关系短语应当匹配多个不同实体对（只有一个实体的短语不可用）

- 参考资料：赵军老师《开放域事件抽取》报告

- **基于句法的开放关系抽取**

- 基于句法的开放关系抽取的原型系统实例：TextRunner
  - M Banko, et al., Open Information Extraction from the Web, IJCAI 2007
  - 用户输入特定的实体或谓词，利用搜索引擎返回与之相关的句子。
    - 基于返回的句子整理相应语料



- 参考资料：赵军老师《开放域事件抽取》报告

本章小结

关系抽取

- 关系抽取
  - 概述与基本方法
  - 远程监督方法
  - 开放关系抽取

tongxu@ustc.edu.cn