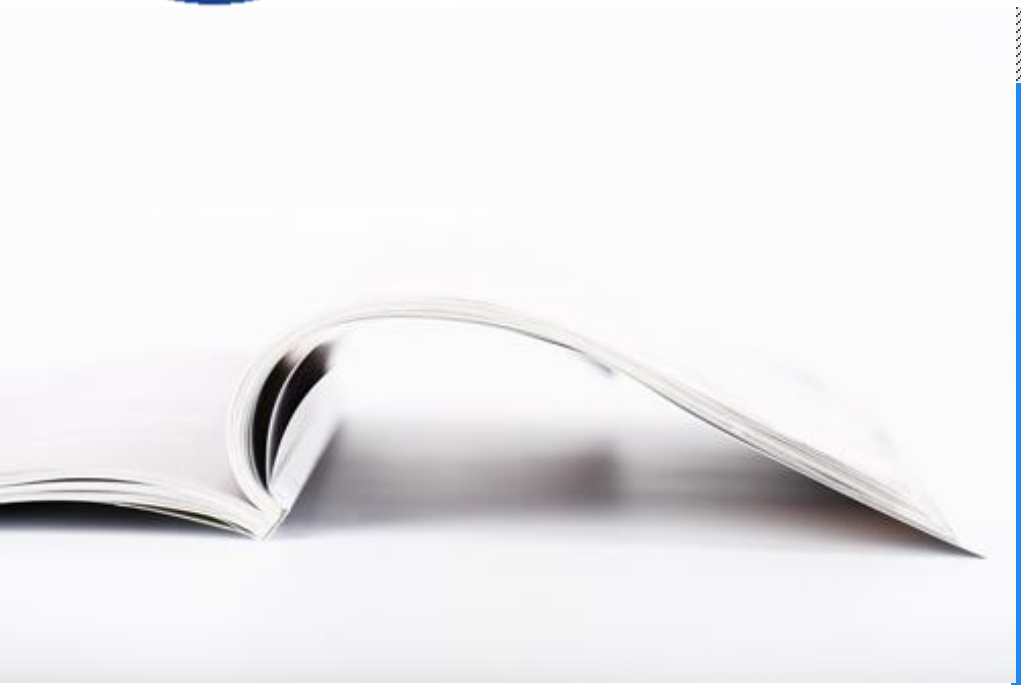


Web信息处理与应用

第十节

关系抽取

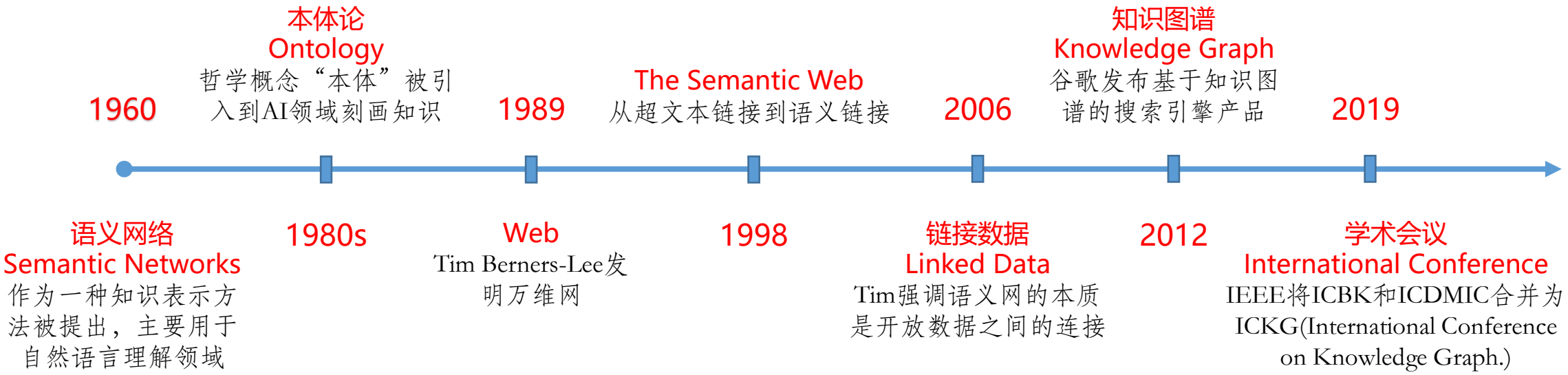
徐童

2021.11.15

上节回顾

- **从知识到知识关联**

- 知识图谱的发展历程：从语义网络与本体论衍生而来
  - 语义网络的部分内容在第三节有所提及，例如：同义词/相关词，WordNet



上节回顾

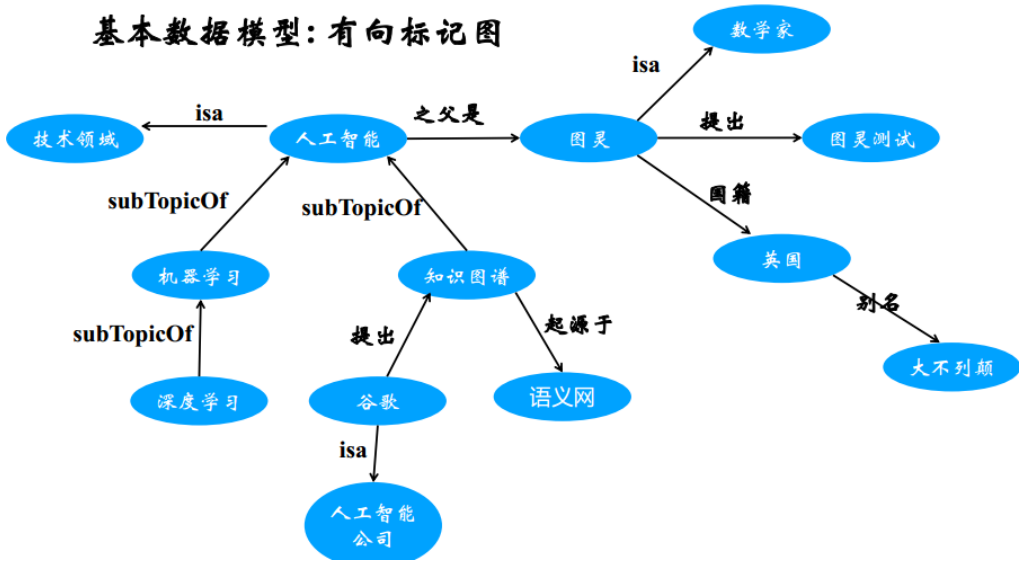
- **信息抽取的基本任务**

- MUC-7所定义的5类基本的信息抽取任务：抽取实体，确定关系
  - **命名实体NE**、模板元素TE、共指关系CR、**模板关系TR**、背景模板ST
- **命名实体NE**（实体抽取）
  - 命名实体是文本中基本的信息元素，是正确理解文本的基础
- **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
    - 例如，职务（Post_of）、雇佣关系（Employee_of）、生产关系（Product_of）等

上节回顾

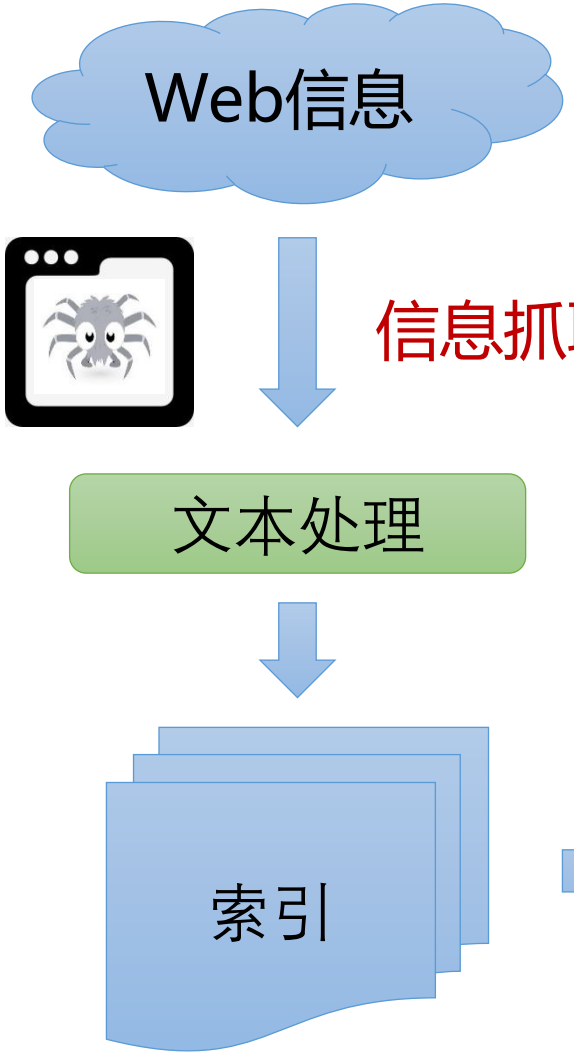
- **知识图谱的基本形式**

- 由前述各种网络衍生而来，知识图谱呈现出类似的基本形式
- 由节点和节点之间的边组成，节点表示概念（或实体），边表示关系（或属性）。
- 在数学上，知识图谱表现为一个有向图。
- 点和边组成知识图谱的基本单位：**三元组**
  （头实体-关系-尾实体）



本节主题

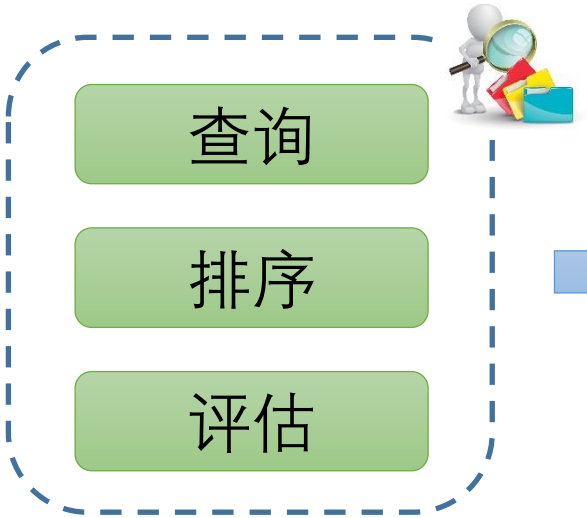
- 本课程所要解决的问题

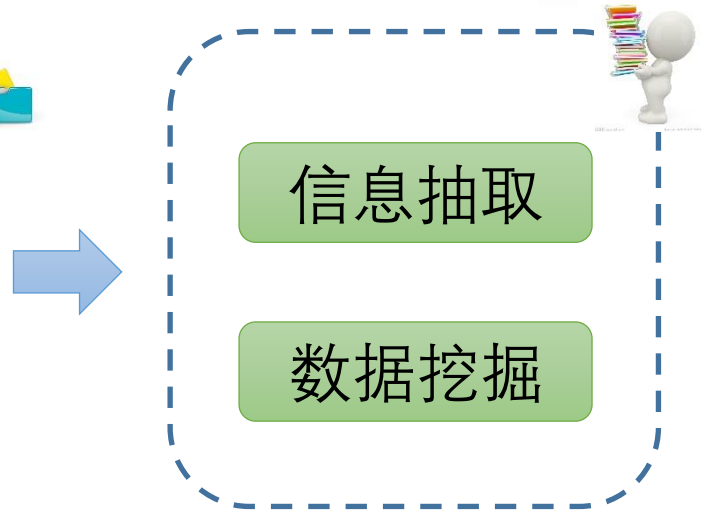


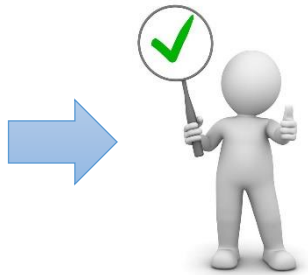
第九个问题：

如何从文档中提取关系，

实现知识实体的关联？







本节目录

- **关系抽取**
  - 关系抽取方法
  - 开放关系抽取
  - 远程监督方法
- 关系补全
- 事件抽取

- **关系抽取的基本概念**
- 1997年，MUC-7上首次引入了关系抽取任务（Template Relation）
- 从文本中识别出两个实体（或多个实体）之间存在的事实上的关系。

...在文本中检测实体之间语义关系的一种技术...

—— 引自维基百科

- 例如：“朋友们好啊，我是混元形意太极门掌门人马保国”
  - 从中我们可以抽取出以下模板关系
    - Post_of (掌门人，马保国)
    - Employee_of(混元形意太极门，马保国)

- **关系抽取的意义**

- （1）关系抽取是搜索引擎发现和关联知识的重要渠道
  - 鸡腿和减肥如何关联？基于“热量”与体重之间的关系

- **关系抽取的意义**

- （2）关系抽取是知识库构建与知识关联的基础性手段
  - 在理解关系的基础上建立结构化知识库，为关联和推理奠定基础

中国共产党职务		
前任： **赵紫阳**	**中国共产党中央委员会总书记** 1989年 - 2002年	继任： **胡锦涛**
前任： **邓小平**	**中国共产党中央军事委员会主席** 1989年 - 2004年	
前任： **芮杏文**	**中国共产党上海市委员会书记** 1987年 - 1989年	继任： **朱镕基**

- **关系抽取的意义**

- （3）关系抽取是支持问答系统、推荐系统等应用的有力工具
  - 不仅推荐得准，而且推荐得丰富多彩，推荐得有理有据

- **关系如何表达**

- 二元组<subject, objects>
  - 适合**特定领域关系**抽取，类似二部图，例如企业收购关系：<Microsoft, Nokia>
- 三元组<subject, predicate, object>
  - 适合**不定类型关系**抽取，最为常见、基础的表达形式，例如企业之间的商业关系抽取
  - <Microsoft, acquisition, Nokia> / <Microsoft, cooperation, Intel>
- 多元组，引入更多外部信息，例如<subject, predicate, object, time>
  - 目前在时态关系抽取上研究较多，例如<Biden, as-president, USA, [2021, 2024?]>
- 在非事先约定的情况下，一般采用三元组来表达

- **常见的关系数据源**

- 结构化程度较高的数据
  - 如Wikipedia，黄页等

- **常见的关系数据源**

- 结构化程度较高的数据
  - 或借助WordNet、大词林等词库

- **常见的关系数据源**

- 可用作关系抽取实验的公开数据集
  - ACE 2005：包含7大类，599篇文档。可用于实体和关系检测任务。
    - 其中包含的关系有：局部整体关系(PART-WHOLE)，地理位置关系(PHYS)，类属关系(GEN-AFF)，转喻关系(METONYMY)，制造使用关系(ART)，组织结构从属关系(ORG-AFF)，人物关系(PER-SOC)
  - SemEval-2010 Task 8：包含9类关系，10,717个样本。
    - 由于这9类关系是有向的，因此可视作9x2+1=19类关系（多一个其他）
  - NYT：从1987至2007年间纽约时报中抽取的1.18M个句子。包含53种关系和数十万个实体对。通过 Stanford NER 工具进行标注。

本节目录

- 关系抽取
  - 关系抽取方法
  - 开放关系抽取
  - 远程监督方法
- 关系补全
- 事件抽取

- **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
  - 回忆一下上节课提过的“模板元素”和“槽”的概念
    - 例如：马保国掌门指出，这两个年轻人不讲武德。
      - TE：两个年轻人是不讲武德的（属性）
      - 由类似模板：AAA是BBB的，可知AAA是实体，BBB是属性

- **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，同样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**

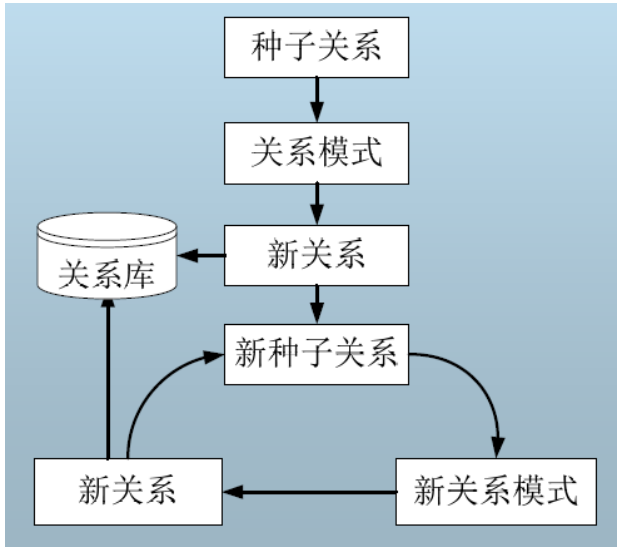
- 描述前述的<X, IS_A, Y>关系，可采用以下规则
  - Rule 1: “Y such as X”： Universities such as MIT and CMU ……
  - Rule 2: “X or other Y”： Apples or other fruits ……
  - Rule 3: “Y including X”： Machine learning methods including SVM and CRF……
  - Rule 4: “Y, especially X”： Most students, especially Ph.D. candidates ……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

- **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**

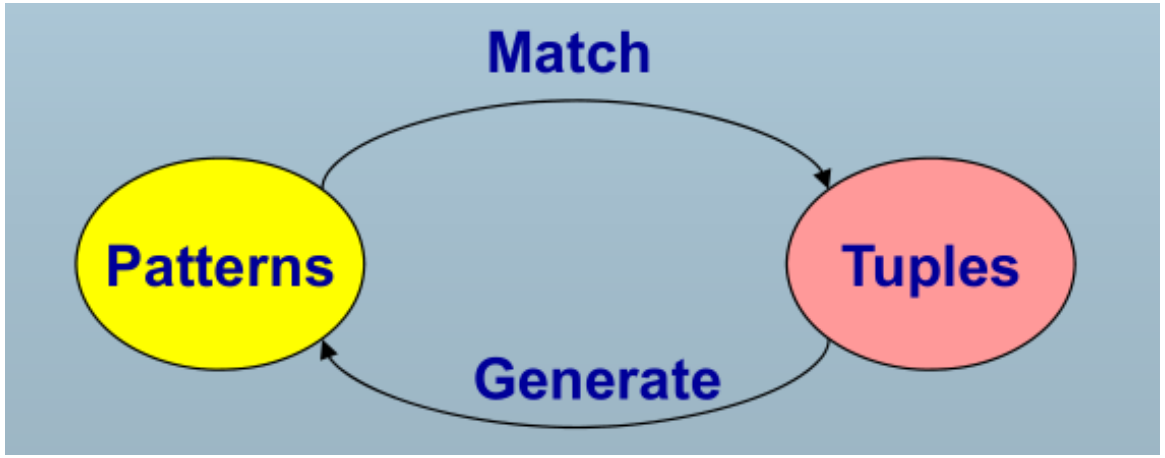
- 在表示这些三元组时，存在着某种重复的“模式”。

- **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**

- 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**

  - 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

此时捕捉到了新信息

- **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
  - 将同一关系的不同实例在网页上所呈现的不同Occurrence中，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

注：我们会在第十一节课有关关联规则的介绍中，再次见到这两个概念

- **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。
  - 移值性差，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

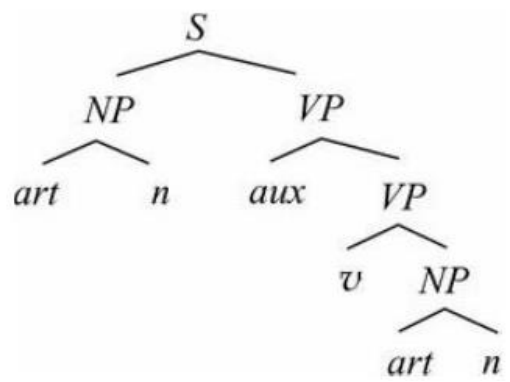
- **基于机器学习的关系抽取**

- 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。

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*

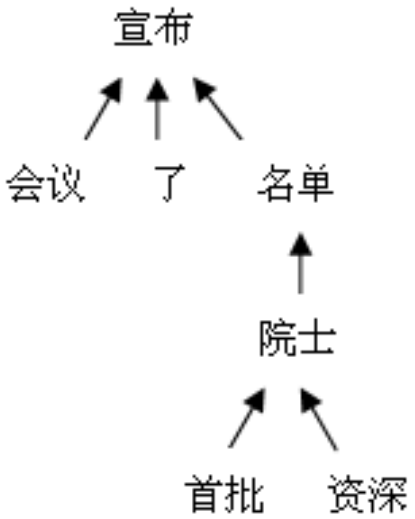
- **基于机器学习的关系抽取**

- 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。

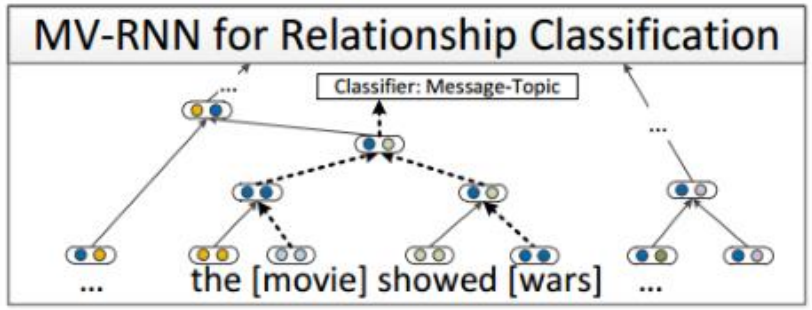


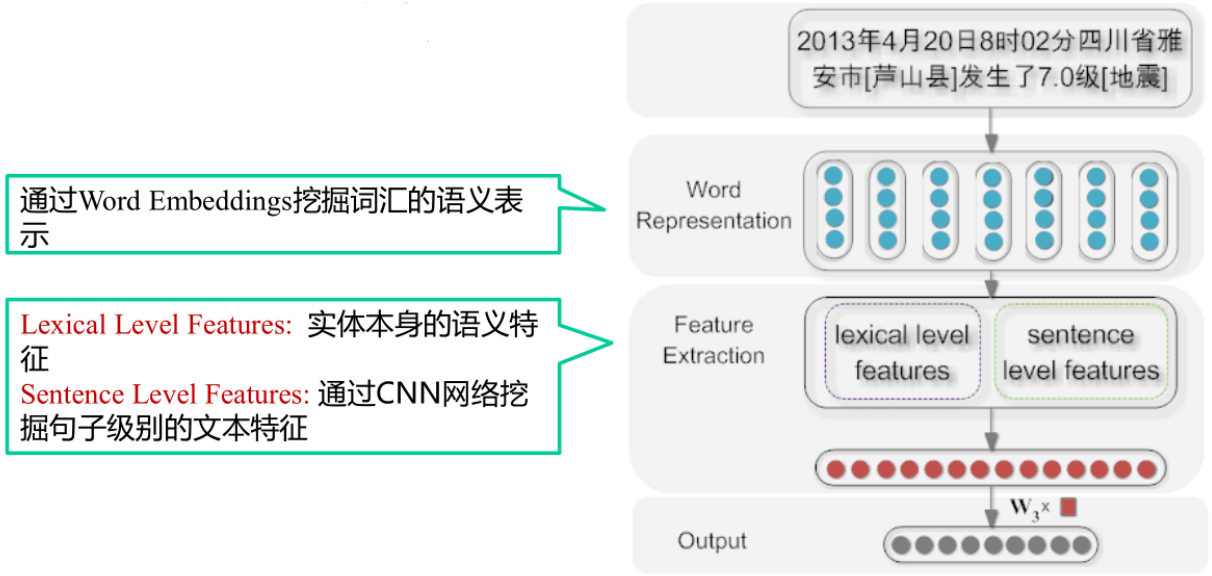
- **基于机器学习的关系抽取**

- 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，例如，将对关系实例表示为某种结构的树（例如语法树），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。
  - 常用方法包括浅层树核（Zelenko , 2003）、依存树核（Culotta, 2004)、最短依存树核（Bunescu, 2005）、卷积树核（Zhou, 2007）等。
    - 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。

本节目录

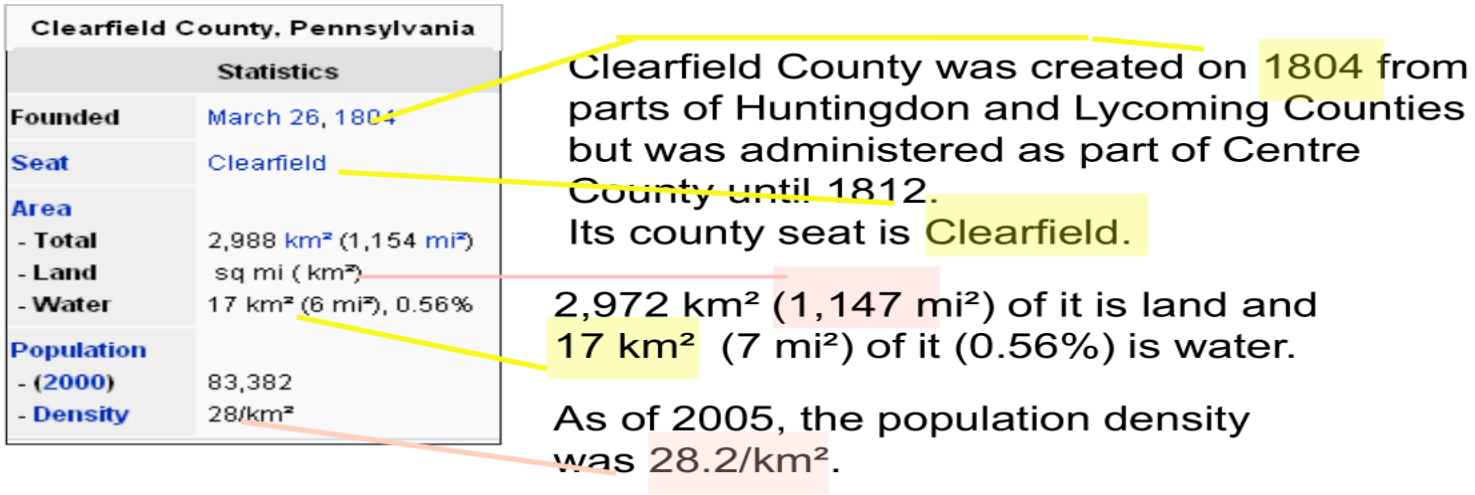
- **关系抽取**
  - 关系抽取方法
  - **开放关系抽取**
  - 远程监督方法
- 关系补全
- 事件抽取

- **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回标，产生训练语料



  - 参考资料：赵军老师《开放域事件抽取》报告

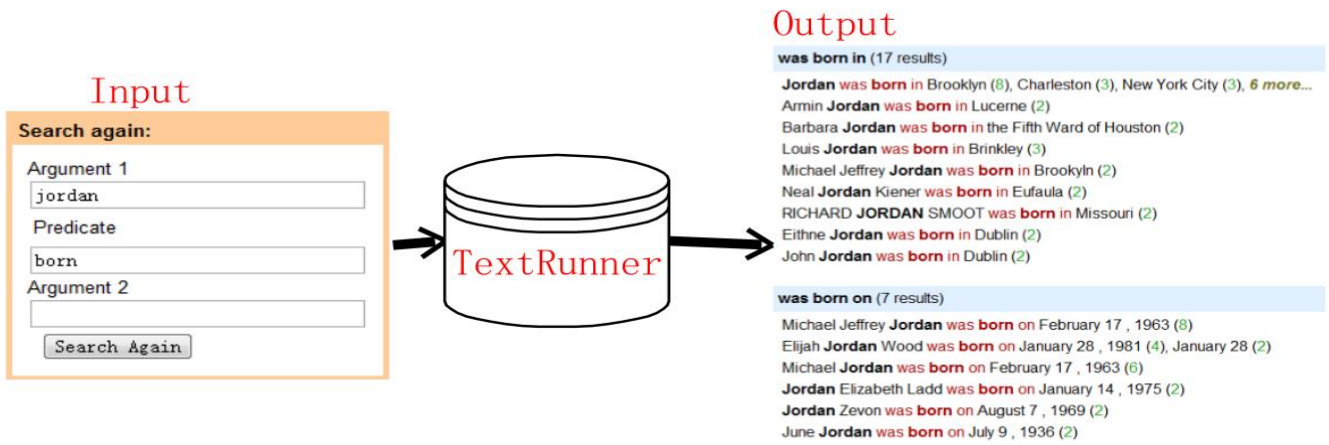
• 基于句法的开放关系抽取

- 同时，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
  - 例如，以下三个句子均表达了类似的实体间关系
    - (华为，总部位于, 深圳)
    - (华为，总部设置于，深圳)
    - (华为，将其总部建于，深圳)
  - 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。
  - 在抽取三元组关系的同时，对于关系可信度进行评估。



- 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**

- 基于以上思路，可通过对于关系短语的句法结构约束来抽取关系

$$V \mid VP \mid VW^*P$$
V = verb particle? adv?
W = (noun | adj | adv | pron | det)
P = (prep | particle | inf. marker)

- 优点：无需预先定义关系类别（这个动词就是关系）
- 缺点：语义没有归一化，同一关系存在多种不同表达
  - 例如，前述“总部位于”、“总部设置于”、“将总部建于”等

- 参考资料：赵军老师《开放域事件抽取》报告

本节目录

- **关系抽取**
  - 关系抽取方法
  - 开放关系抽取
  - **远程监督方法**
- 关系补全
- 事件抽取

- **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**

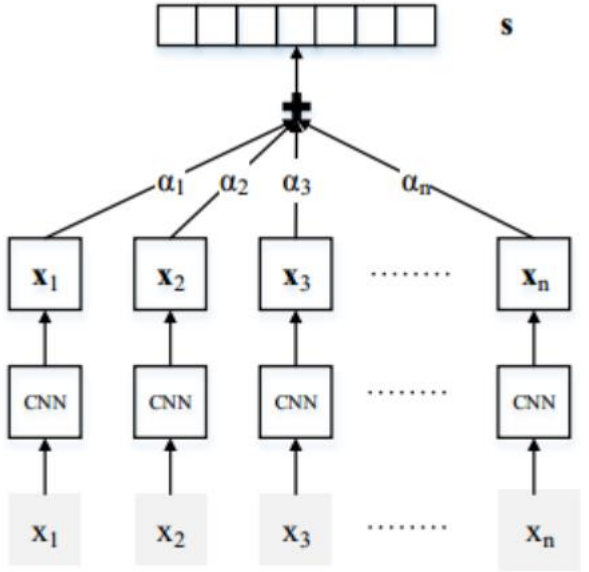
- 远程监督的优化方案（1）：动态转移矩阵
  - 尽管噪音数据不可避免，但是对噪音数据模式进行统一描述是可能的。
    - 例如，一个人的工作地点和出生地点很有可能是同一个地点，这种情形下远程监督就很有可能把born-in和work-in这两个关系标签打错。
  - 解决方案：引入一个动态转移矩阵，描述各个类之间相互标错的概率。
  - 在利用算法得到的关系分布的基础上乘以这一转移矩阵，即可得到更为准确的关系分类结果。

- **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。

- **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CNN等技术，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。



本节目录

- 关系抽取
  - 关系抽取方法
  - 开放关系抽取
  - 远程监督方法
- **关系补全**
- 事件抽取

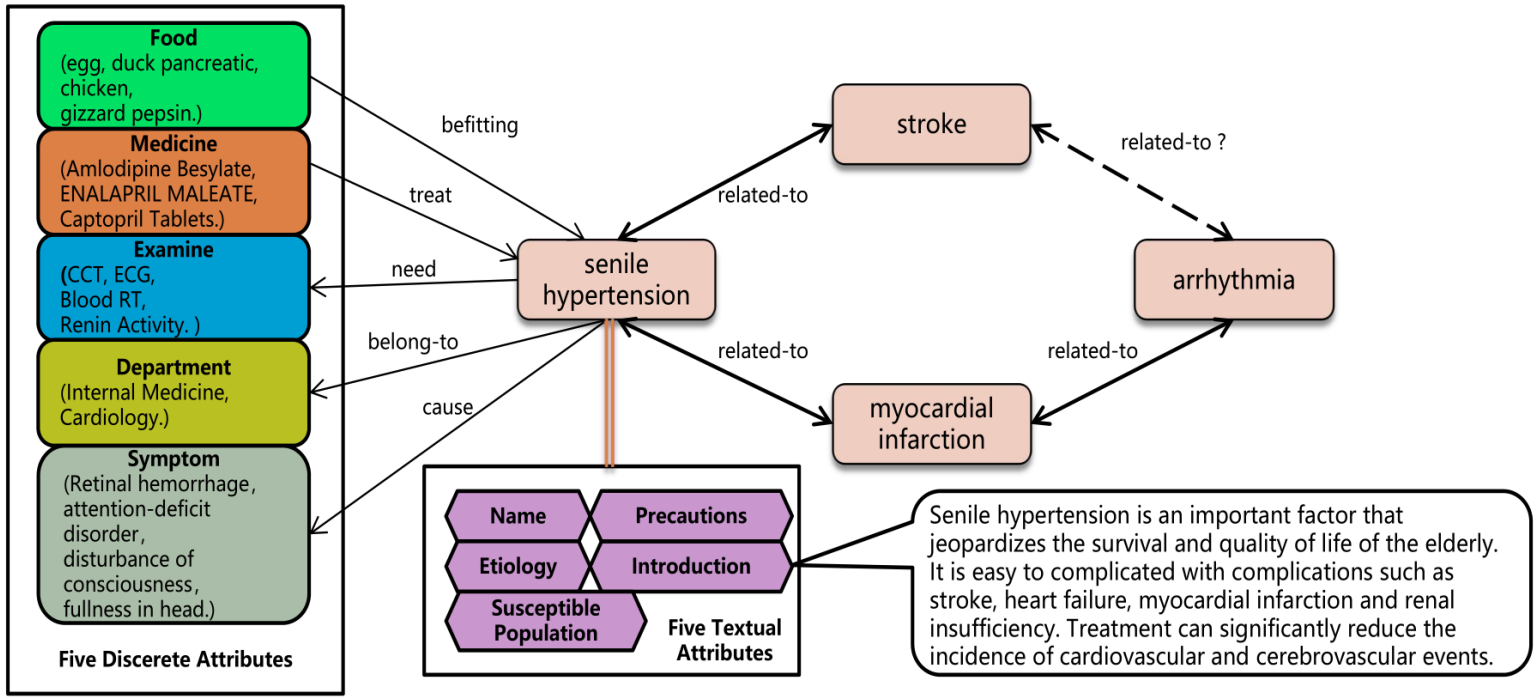
- **知识图谱的补全**

- 虽然在关系抽取方面已有诸多进展，想要挖掘完整的关系依然难以实现
  - 人工方法构建图谱成本高，限制了图谱的规模，而自动生成效果有限
  - 现有图谱规模仍较为稀疏，大量隐含关系尚未被充分挖掘
  - 实体和关系处于不断演化和拓展中，新的关系不断生成
- 基于现有图谱对关系进行补全，从而获得更完整的图谱，是当下的研究热点
  - 通过挖掘尚未被发现的潜在关系，或提供新的知识和新的路径
    - 例如，新的药物治疗方式，尚未被发现的并发症等

关系补全

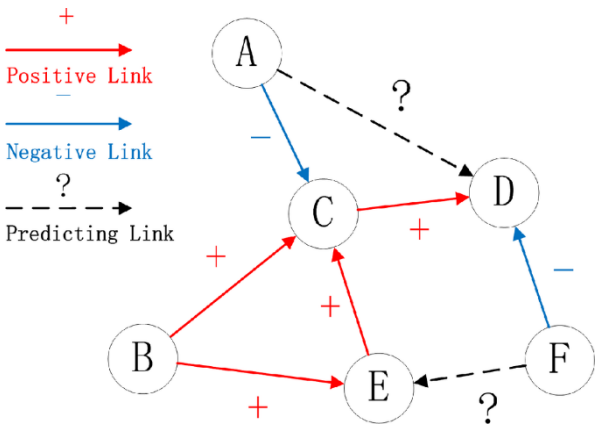
• 知识图谱的补全

- 基于现有图谱对关系进行补全，从而获得更完整的图谱，是当下的研究热点
  - 例如，新的药物治疗方式，尚未被发现的并发症等



关系补全与链接预测

- 某种意义上，知识图谱的关系补全问题可转化为经典的链接预测问题
  - 已知图结构中的节点和部分边，推测其他可能存在的边
  - 目前，针对这类问题已有大量的研究工作
    - 例如，仅考虑网络结构，基于已有边推测未知边
      - 基本的原则：朋友的朋友很可能也是朋友
    - 拓展：在网络结构基础上考虑符号性
      - 朋友的朋友是朋友，敌人的敌人也是朋友
    - 再拓展：不仅考虑结构，也考虑属性
      - 朋友具有属性上的相似性



- **关系补全与链接预测**

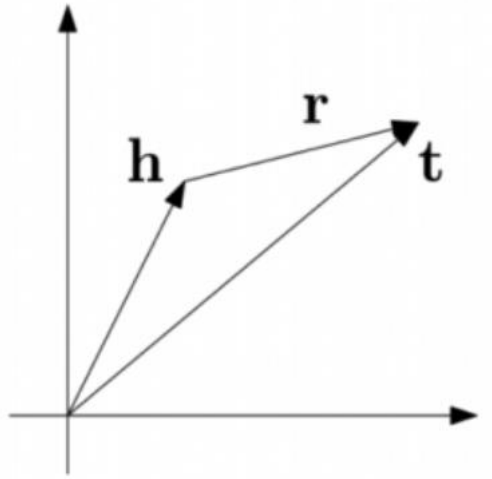
- 然而，链接预测方法未必胜任关系补全问题，两者仍存在明显差异
  - 知识图谱中的实体具有不同的类型和属性
  - 知识图谱中的关系具有不同的类型和属性
    - 不同关系的链接之间未必具有相互揭示效果
    - 部分实体之间不存在关系，或部分关系不合理
      - 例如，两种药都可以治疗某种疾病，但是两种药本身存在互斥

- **基于知识图谱嵌入的补全方法**

- 近年来流行的Trans系列模型，通过将知识图谱嵌入到一个连续的向量空间，同时保留一定的图中信息，逐渐受到广泛关注
- 代表性工作1：TransE模型
  - A Bordes, et al., Translating embeddings for modeling multi-relational data, NIPS 2013
  - 核心思想是一种翻译模型，将三元组（h, r, t）中的关系 r 视作从 h 到 t 的翻译，通过不断调整表征，使 h+r 接近 t

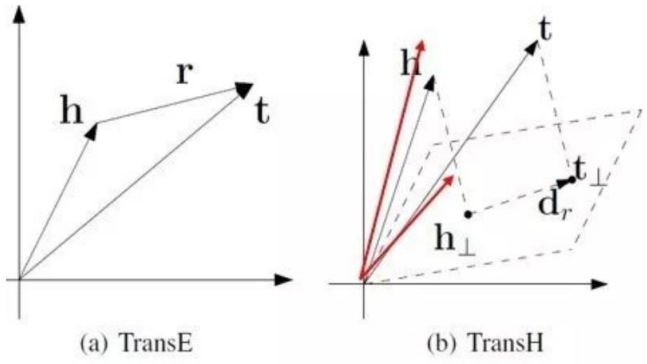
$$L = \sum_{(h,r,t)\in\Delta} \sum_{(h',r',t')\in\Delta'} \max(f_r(h,t) + \gamma - f_{r'}(h',t'), 0).$$

  - 其思路借鉴了word2vec，利用了词向量的“平移不变性”
    - 例如，C(中国) - C(北京) = C(美国) – C(华盛顿)

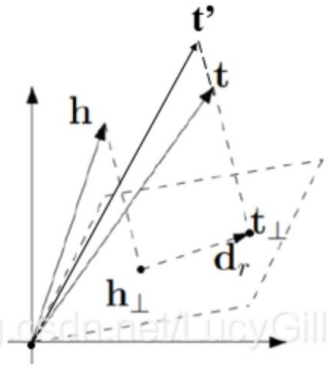


- **基于知识图谱嵌入的补全方法**

- 代表性工作2：TransH模型
  - Z Wang, et al., Knowledge Graph Embedding by Translating on Hyperplanes, AAAI 2014
  - TransE模型无法解决多对一或一对多的关系，会导致“多”的一方向量表征过于相似
  - TransH对这一问题进行了修正，不再严格要求 h + r 接近t，而只需要保证 h 和 t 在关系平面上的投影在一条直线上即可。
  - 通过这种方法，一个实体在不同关系下可以有不同的表示。



两个不同的尾实体，在关系平面上投影在同一直线→



- **关系补全与链接预测**

- 其他Trans系列模型概述
  - TransR模型（AAAI 2015）：用于表述不同关系对应实体不同属性的情况
  - TransD模型（ACL 2015）：基于动态矩阵，描述同一种关系的不同语义
  - TransA模型（ArXiv 2015）：基于马氏距离，区别对待不同维度特征
  - TransG模型（ACL 2016）：基于贝叶斯非参混合模型描述关系多语义问题
  - Transparse模型（AAAI 2016）：拓展TransR，解决不同关系稀疏不同问题

本节目录

- 关系抽取
  - 关系抽取方法
  - 开放关系抽取
  - 远程监督方法
- 关系补全
- **事件抽取**

上节回顾

- **信息抽取的基本任务**

- **场景模板ST**（事件抽取）
- 又称事件，是指实体发生的事件
  - 例如：会议 (Time<…>, Spot<…>, Convener<…>, Topic<…>)
  - 常见的新闻事件描述模板 5W1H
    - Who 、 When 、 Where 、 What 、 Why 、 How

上节回顾

- **信息抽取的基本任务**

- ST事件抽取结果示例

```
<EventTemplateInstatnces>
    <ConferenceInfo>
        <Time>4日晚 ( 1998-01 )</Time>
        <Spot> 意大利</Spot>
        <Converner> 普罗迪</Converner>
        <Title>由意外长、内政和国防部长参加的紧急会议
        </Title>
    </ConferenceInfo>
</EventTemplateInstatnces>
```

会议时间 Time	4 日晚 (1998-01)	
会议地点 Spot	意大利	
召集人 Convener	姓名/团体名称 Name	普罗迪
	机构、职位 Org/Post	意大利总理
会议名/标题 Conf-Title	由意外长、内政和国防部长参加的紧急会议	

- **事件抽取的概念**

- 事件是信息的一种表现形式，其定义为特定的人、物，在特定时间和特定地点相互作用所产生的客观事实。
  - 例如，可对应先前所说的5W1H基本要素
  - 一般信息呈现为句子级别（相比之下，关系往往表现为短语级别）
  - 推理任务（事件时序推理、事理推理）的前提
  - ACE中对于事件的定义如下：
    - *An event is a specific occurrence involving participants. An event is something that happens. An event can frequently be described as a change of state.*

- **事件抽取的基本要素**

- 通常而言，事件往往包含以下基本要素：
  - 事件触发词：表示事件发生的核心词，多为动词或名词
    - 相应的，事件触发词的检测与分类是事件抽取的基本任务

Example:

- *Henry[argument] was injured, and then passed away soon*

Detection: injured
Typing: Injure
Argument: Henry

Detection: passed away
Typing: Die
Argument : Henry

- **事件抽取的基本要素**

- 通常而言，事件往往包含以下基本要素：
  - 事件类型：与触发词相对应，往往可以通过触发词分类加以识别
    - 例如，前例中的触发词Pass away对应着“死亡”的事件类型
  - 事件元素：事件的参与者，主要由实体、时间等组成。
    - 例如，前例中的Henry是事件的主体
  - 事件元素角色：事件元素在事件中充当的角色。
    - 例如，前例中的Henry在事件中是一个“受害者”的角色

- **事件抽取的模板**

- 通过触发词识别和分类，判定事件及其类型后，可以借助模板实现抽取。
  - 在上一节中，我们提到过**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
  - 其中，模板元素通过槽（Slots）描述了命名实体的基本信息
    - 槽的内容可包括名称、类别、种类等

- **事件抽取的模板**

- 在选定相应的模板之后，通过事件元素与事件元素角色的识别，将相应的元素填入模板合适的槽（Slot）内，即完成了事件抽取。

模版元素	实体类型	描述
Person-Arg	PER	结婚的人
Time-Arg	TIME-within	结婚时间
Place-Arg	GPE LOC FAC	结婚地点

- **事件抽取的模板**

- 在选定相应的模板之后，通过事件元素与事件元素角色的识别，将相应的元素填入模板合适的槽（Slot）内，即完成了事件抽取。
  - 案例：刚才有个朋友问我，马老师发生甚么事了
    - 元素/描述：人物（朋友），时间（刚才），事件（提问发生了甚么事）

- **限定域事件抽取**

- 某种意义上，限定域事件抽取与预定义关系抽取存在相似之处，即预先定义好目标事件的类型及每种类型的具体结构（包含哪些具体的事件元素）。
- 因此，限定域事件抽取可以采用基于模式匹配的方法实现
  - 可以采用完全规则的方法实现，即完全通过人工标注方式获得模式
  - 也可以采用弱监督的模式匹配，即不需要对语料进行完全标注，只需要人工对语料进行一定的预分类或者制定少量种子模式
    - 类似于前述的DIPRE方法，迭代式获得更完善的语料和模板

- **限定域事件抽取**

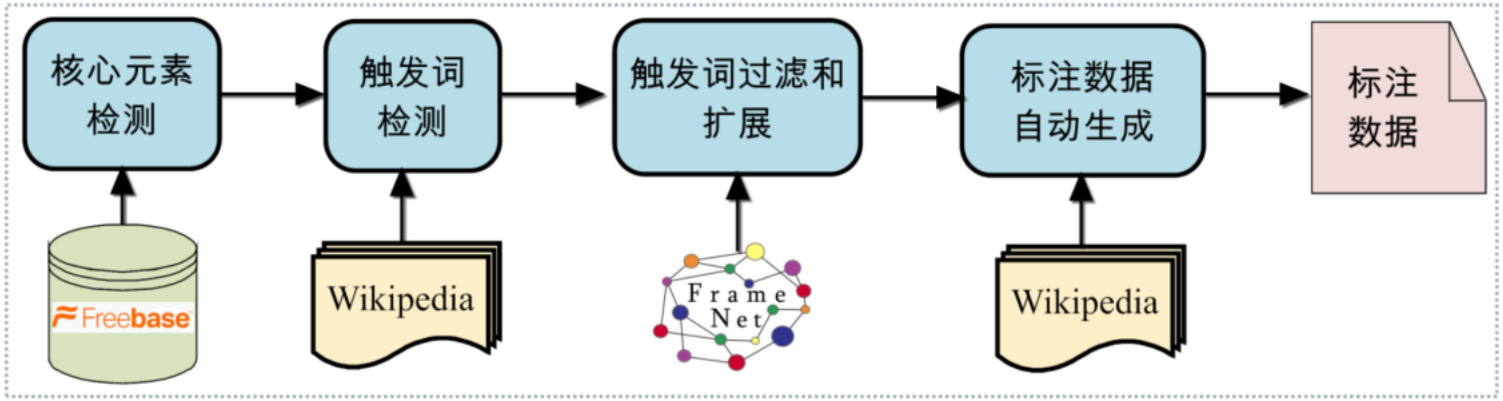
- 同样的，限定域事件抽取也可采用基于机器学习的方法实现。
- 例如，采用有监督学习方式，将事件抽取转化为一个多分类问题（传统艺能）
  - 基于特征工程的方法：将事件实例转换成分类器可以接受的特征向量
  - 基于神经网络的方法：自动从文本中获取特征进而完成事件抽取
- 同样，也可以采用远程监督等弱监督的方式，实现限定域的事件抽取

- **开放域事件抽取**

- 限定域事件与预定义关系面临相似的问题：种类有限，维护困难。
- 如何在开放域环境下，自动识别未知结构与类型的事件？
  - 一种思路是采用无监督方法（从而摆脱对于标注语料的依赖），通过聚类找到潜在的事件簇
  - 该思路基于分布假设理论：
    - 候选事件触发词或者候选事件元素具有相似的语境，那么这些候选事件触发词倾向于触发相同类型的事件，相应的候选事件元素倾向于扮演相同的事件元素。
  - 然而，无监督事件抽取没有规范语义标签，难以映射到现有知识库。

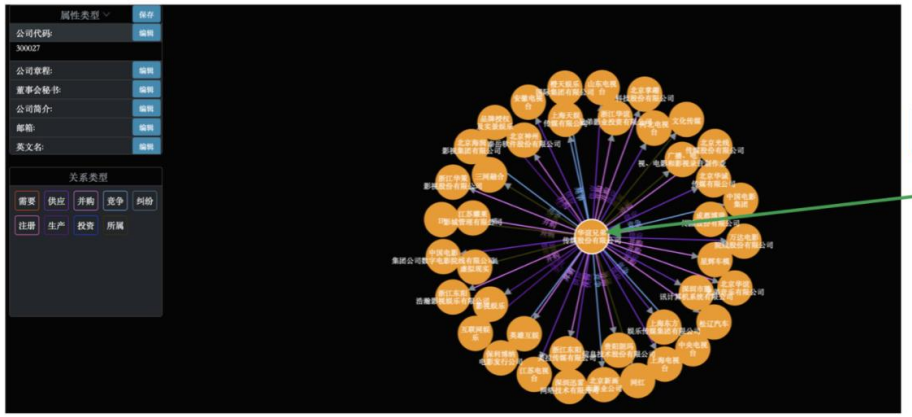
- **开放域事件抽取**

- 另一种解决方法：与开放关系抽取中的“知识监督”方案类似
  - 主要挑战在于现有知识库中缺乏事件触发词信息，如何获取？
  - Y Chen, et al., Automatically Labeled Data Generation for Large Scale Event Extraction, ACL 2017
  - 定义事件核心元素，通过初步回标找到触发词并进行过滤和扩展

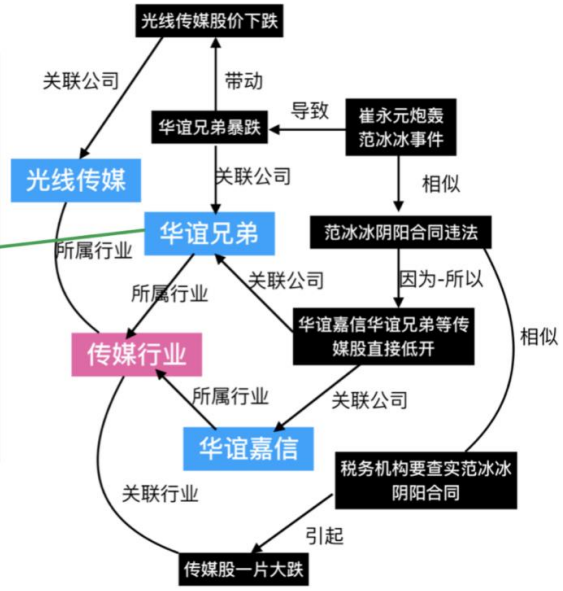


- **从知识图谱到事理图谱**

- 以事件代替实体作为节点，可以将知识图谱拓展为事理图谱
  - 知识图谱与事理图谱的结合具有丰富的应用价值



公司金融知识图谱



因果事件图谱

- **知识图谱表示学习**
  - OpenKE
  - 包含TransE，TransH，TransR在内的开源知识表示工具包。
- **关系抽取**
  - OpenNER
  - 开源的关系抽取工具包，包含句子级别关系抽取，篇章级关系抽取，少样本关系抽取。

本章小结

关系抽取

- 关系抽取
  - 概述与基本方法
  - 开放关系抽取
  - 远程监督方法
- 关系补全：从链接预测到Trans系列模型
- 事件抽取
  - 基本概念与基本要素，限定域/开放域下的事件抽取

tongxu@ustc.edu.cn