Chapter 8 信息抽取（研究进展、现状&趋势）

什么是信息抽取？

信息抽取（Information Extraction）是指从非结构化/半结构化文本（如网页、新闻、论文文献、微博等），使用多种技术（如规则方法、统计方法、知识挖掘方法），提取指定类型的信息（如实体、属性、关系、事件、商品记录等），并通过信息归并、冗余消除和冲突消解等手段将非结构化文本转换为结构化信息，并将这些信息在不同的范围进行集成（知识去重、知识链接、知识系统构建等）的一项综合技术。

被抽取出来的信息通常以结构化的形式描述，可以为计算机直接处理

每一段文本内所包含的寓意可以描述为其中的一组实体以及这些实体相互之间的关联和交互，抽取文本中的实体和它们之间的语义关系也就成为了理解文本意义的基础

意义

- 实现对海量非结构化数据的分析、组织、管理、计算、查询和推理，并进一步为更高层面的应用和任务（如自然语言理解、知识库构建，智能问答 系统、舆情分析系统）提供支撑。
- 组织、管理和分析海量文本信息的核心技术和重要手段，是大数据时代的使 能技术,具有重要的经济和应用意义
- 构建可支撑类人推理和自然语言理解的大规模常识知识库的有效技术之一

应用

- 如舆情分析、舆情监控、网络搜索、智能问答系统、知识库构建、文本分析等

研究内容

命名实体识别(Named Entity Recognition, NER)

目的是识别文本中指定类别的实体，主要包括人名、地名、机构名、专有名词等的任务

包含部分

- 实体边界识别 —— 判断一个字符串是否是一个实体
- 实体分类 —— 将识别出的实体划分到预先给定的不同类别中去

主要难点 —— 表达不规律、且缺乏训练语料的开放域命名实体类别（如电影、歌曲名）

关系抽取 (Relation Extraction)

检测和识别文本中实体之间的语义关系，将表示同一关系的提及（mention）链接起来的任务

输出：通常是一个三元组（实体 1，关系类别，实体 2），表示实体 1 和实体 2 之间存在特定类别的语义关系

- 例子：句子"北京 是中国的首都、政治中心和文化中心"中表述的关系可以表示为（中国，首都，北京），（中 国，政治中心，北京）和（中国，文化中心，北京）。
- 语义关系类别可以预先给定（如 ACE 评测中的七大类关系），也可以按需自动发现（开放域信息抽取）

核心模块

- 关系检测 —— 判断两个实体之间是否存在语义关系
- 关系分类 —— 将存在语义关系的实体对划分到预先指定的类别中
- 关系发现（某些场景下） —— 主要目的是发现实体和实体之间存在的语义关系类别

事件抽取

从非结构化文本中抽取事件信息，并将其以结构化形式呈现出来的任务 —— 例子：从"毛泽东 1893 年出生于湖南湘潭"这句话中抽取事件(类 人物：毛泽东，时间：1893 年，出生地：湖南湘潭)

子任务

事件类型识别

- 判断一句话是否表达了特定类型的事件
- 事件类型决定了事件表示的模板，不同类型的事件具有不同的模板 —— 例如出生事件的模板是(人物， 时间，出生地)，而恐怖袭击事件的模板是(地点，时间，袭击者，受害者，受伤人数...)。

事件元素填充

- 事件元素指组成事件的关键元素
- 根据所属的事件模板，抽取相应的元素，并为其标上正确元素标签的任务

信息集成 (Information Integration)

原因

- 实体、关系和事件分别表示了单篇文本中不同粒度的信息
- 在很多应用中，需要将来自不同数据源、不同文本的信息综合起来进行决策

技术

共指消解技术

- 检测同一实体/关系/事件的不同提及，并将其链接在一起的任务
- 例如，识别"乔布斯是苹果的创始人之一，他经历了苹果公司几十年的起落与兴衰"这句话中的"乔布斯"和 "他"指的是同一实体

实体链接技术

- 目的是确定实体名所指向的真实世界实体
- 例如识别"苹果"和"乔布斯"分别指向真实世界中的苹果公司和其 CEO 史蒂夫·乔布斯

关键科学问题

- 自然语言表达的多样性、歧义性和结构性
- 目标知识的复杂性、开放性和巨大规模
- 多源异构信息的融合与验证

信息抽取方法

根据模型的不同 （重点）

基于规则的抽取方法

- 一个基于规则的抽取系统通常包括一个规则集合和规则执行引擎（负责规则的应用、冲突消解、优先级排 序和结果归并）
- 规则系统在抽取可控且表达规范的信息时非常有效
- 表现形式：正则表达式、词汇-语法规则、面向 HTML 页面抽取的 Dom Tree 规则等等
- 抽取规则可以通过人工编写得到或者使用学习方法自动学习得到
- 抽取规则的管理、冲突消解和优先级排序也是基于规则的信息抽取研究内容 —— 原因：为抽取一类特定信息，通常需要一系列相关的抽取规则，在实际情况中，通常会存在规则相互冲突或规则不一致的情况
- 研究重点、难点
  - 重点：构建更高效的规则执行引擎、更方便的规则开发平台、更具表达能力的规则表示语言
  - 难点：如何学习更精准的抽 取规则、如何消除抽取规则的歧义、如何自动评估规则的效果 （如 Bootstrapping 系统通常会遇到的语义漂移问题）

基于统计模型的抽取方法

通常将信息抽取任务形式化为从文本输入到特定目标结构的预测，使用统计模型来建模输入与输出之间的关联，并使用机器学习方法来学习模型的参数。

统计方法

- 最大熵分类模型、基于树核的 SVM 分类模型、隐马尔可夫模型、条件随机场模型（CRF）等等
- CRF是实体识别的代表性统计模型，它将实体识别问题转化为序列标注问题；
- 基于树核的关系抽取系统则将关系抽取任务形式化为结构化表示的分类问题。

与深度学习结合

- 相比传统的统计信息抽取模型，深度学习模型无需人工定义的特征模板，能够自动的学习出信息抽取的有效特征；同时神经网络的深度结 构使得深度学习模型具有更好的表达能力
- 在标注语料充分的情况下，深度学习模型往往能够取得比传统方法更好的性能

基于文本挖掘的抽取方法

- Web 中往往还存在大量的半结构的高质量数据源，这些结构往往蕴含有丰富的语义信息
- 半结构 Web 数据源上的语义知识获取（knowledge harvesting），如大规模知识共享社区（如百度百科、互动百科、维基百科）上的语义知识抽取，往往采用文本挖掘的方法
- 核心：构建从特定结构（如列表、Infobox）到目标语义知识（实体、关系、事件）的映射规则
- 由于映射规则本身可能带有不确定性和歧义性，同时目标结构可能会有一定的噪音，文本挖掘方法往往基于特定算法来对语义知识进行评分和过滤
- 只从容易获取且具有明确结构的语料中抽取知识，抽取出来的知识质量往往较高。但是仅仅依靠结构化数据挖掘无法覆盖人类的大部分语义知识，现有结构化数据源只能覆盖有限类别的语义知识，相比人类的知识仍远远不够
- 展望：如何结合文本挖掘方法（面向半结构化数据，抽取出的知识质量高但覆盖度低）和文本抽取方法（面向非结构化数据，抽取出的知识相比文本挖掘方法质量低但覆盖度高）的优点，融合来自不同数据源的知识，并将其与现有大规模知识库集成，是文本挖掘方法的研究方向之一。

根据对监督知识的依赖，信息抽取 方法可以划分为无监督方法、弱监督方法、知识监督方法和有监督方法

根据抽取对象的不同，可以划分为实体识别方法、关系抽取方法、事件抽取方法

局限性

- 在构建成本上，现有高 质量抽取系统往往依赖于标注语料，构建成本较高
- 在构建方式上，现有信息抽取系统依赖于许多预处理模块（如分词、词性标注、句法分析等），缺乏端到端的自动构建方式（随着深度神经网络的使用，已经有所改善），同时也容易受预处理模块性能的影响
- 在自适应性上，现有抽取系统的自适应性不强，往往在更换语料、更换领域、更换知识类别时会有一个大幅度的性能下降
- 在系统的性能上，现有信息抽取技术在抽取复杂结构（如事件、Taxonomy） 时性能仍然离实用有一定距离

发展方向

面向开放域的可扩展信息抽取技术

现状：现有监督抽取模型无法处理海量异质数据源上开放性和复杂知识的抽取

1.数据规模上的可扩展性；2.数据源类型上的可扩展性；3. 领域的可扩展性；4. 低构建成本（不能完全依赖有监督学习，要基于无监督技术、弱监督技术、知识监督技术等低成本构建技术）

自学习、自适应和自演化的信息抽取系统

- 研究面向开放域的数据源，研究自学习的信息抽取技术，在极少人工干预下构建高性能的终生学习信息抽取系统（Never End Learning System）
- 面向演化数据源，研究增量式的信息抽取技术，实现信息抽取系统的性能自检测和自动领域适应
- 研究信息抽取多任务管理技术，面向不同数据源、不同任务，自动的重用之前的信息抽取模块，并利用自学习技术构建高性能的抽取系统
- 研究融合人、信息、和计算机的信息抽取技术平台，充分利用人、计算机各自的优势，大幅提高抽取结果的可用性

面向多源异构数据的信息融合技术

原因：1）目前大部分信息抽取系统抽取结果都是碎片化、分散和不一致的，很难构建一个完整的、可解释的复杂知识系统模型；2）Web 文本规模巨大，质量参差不齐，导致信息抽取的结果存在冗余、冲突和错误，并存在一定程度的不确定性

目的：去除信息抽取结果的冗余、冲突和错误，并减少信息抽取结果的不确定性；通过将抽取出来的知识碎片组装成一个完整的全局系统，信息融合技术可以帮助我们构建一个完整的、解释性的知识系统，进而支撑更高层的智能应用，如医学药物分析、经济系统分析等等

研究包括跨文档、跨语言和跨媒体三个层次上的融合技术，包括 信息冒信度衡量、冗余信息去除、解决信息之间的冲突、减少抽取信息的不确定性，并构建 自动的缺失信息检测和补全技术

研究信息融合的全局机制，探索基于信息融合的复杂知识模型构建，如基于本体关系的知识图谱，基于因果关系的复杂因果网络，等等