Information Extraction

* 定义

从非结构化/半结构化的文本中提取指定类型的信息（如：实体、属性、关系、事件、商品记录），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。

* EX

从恐怖事件信息中抽取出：时间、地点、袭击者、受害人、袭击目标、后果。

* 核心研究内容

Named Entity Recognition

Relation Extraction

Information Integration事件抽取和信息集成

* Named Entity Recognition

1. 研究内容
   1. 目的：识别文本中指定类别的实体。比如：人名、地名、机构名、专有名词。
   2. 命名实体识别系统：通常包括两个部分

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。

难点：表达不规律；缺乏训练语料

·Relation Extraction

1. 研究内容
   1. 目的：检测和识别文本中实体间的语义关系
   2. Ex:“北京是中国的首都，政治中心和文化中心。” 得到关系：（中国，首都，北京），（中国，政治中心，北京），（中国，文化中心，北京）。

关系抽取的输出常常是一个三元组

* 1. 语义关系类别：

可以预先给定,ex,ACE评测中的七大类关系：人物、组织、设施、处所、地理政治实体、车辆、武器。

也可以按需自动发现,即开放域信息抽取

* 1. 核心问题：

关系检测：判断两个实体间是否存在语义关系

关系分类：把存在语义关系的实体对 划分到预先制定的类别中

* 事件抽取

1. 研究内容
   1. 定义：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
   2. 例子：“毛泽东1893年出生于湖南湘潭”

抽取后：{类型：出生；人物：毛泽东；事件：1893年；出生地：湖南湘潭}

* 1. 组成：
     1. 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的模板；不同事件有不同的模板
     2. 事件元素识别：根据事件类型的模板，抽取相应元素并为其标上正确标签
* 信息集成

1. 研究内容
   1. 定义：将来自不同数据源、不同文本的信息综合起来进行决策
   2. 主要技术：
      1. 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

Ex:“乔布斯是苹果公司的创始人之一，他经历了苹果公司十几年的起落和兴衰”

要能识别出：“乔布斯”和”他“是同一个实体

* + 1. 实体链接技术

确定实体名所指向的真实世界实体人物、组织、设施、处所、地理政治实体、车辆、武器

Ex：上句中的“苹果”指代苹果公司，“乔布斯”指代苹果公司CEO史蒂夫·乔布斯。

* 1. 面临的问题
     1. 自然语言表达的特点
        1. 多样性：同一种意思有多种表达方式。

Ex:“总部位置”可以用“X的总部位于Y”.”X的总部坐落于Y”“作为X的总部所在地，Y……”

* + - 1. 歧义性：同一表达在不同上下文的意思也不同

“苹果“可以指公司、也可以指水果

* + - 1. 结构性：自然语言具有内在结构

“我从北京飞到了上海”和“我从上海飞到了北京”

* + 1. 人类知识的特点
       1. 复杂性、开放性、巨大规模

人类知识多种多样，知识之间相互关联，知识之间的不同的结构关系（Taxonomy结构，Part-of结构，因果关系网络）

* + - 1. 关键问题：构建可表示、可建模、能处理知识复杂性、开放性和巨大规模的技术
* 技术方法
  + 基于规则的抽取方法
    - 系统构成：

规则集合：规则有不同的表现形式：正则表达式、词汇-语法规则、面向HTML页面抽取的Dom Tree规则

规则可以由人工编写or使用学习方法得到

抽取规则开发平台：UIMA Ruta by Apache

规则执行引擎：负责规则的应用、冲突消解、优先级排序、结果归并

* + - 擅长的领域
      * 可控且表达规范的信息。
      * Ex:文本中的事件、电话号码、邮件地址
  + 基于统计模型的抽取方法
    - 用于抽取文本中目标信息的统计方法：
      * 最大熵分类模型
      * 基于树核的SVM分类模型：将关系抽取任务表示成结构化表示的分类问题
      * 隐马尔可夫模型
      * 条件随机场模型CRF:是实体识别的代表性统计模型，把实体识别问题转化为序列标注问题
    - 基于深度学习的模型
      * 卷积神经网络
      * 时序神经网络
      * 递归神经网络
      * 优点：无需人工定义的特征模板，能自动学习出特征。

在标注语料充分的情况下，性能比传统的要好

但需要大量的标注语料来学习