# 信息抽取概述

信息抽取是一种从自然语言文本中自动提取有用信息的技术。其核心任务是识别和分类文本中的实体、关系和事件，并将这些信息转化为结构化数据。信息抽取的主要应用包括搜索引擎、知识图谱构建、情报分析和问答系统等。

## 命名实体识别

### 定义

命名实体识别（NER）是信息抽取的关键任务之一，旨在从文本中识别出具有特定意义的词或短语。根据其定义，命名实体可分为狭义和广义两种：

- 狭义命名实体指现实世界中具体或抽象的实体，如人（如张三）、机构（如中国中文信息学会、阿里巴巴网络技术有限公司）、地点等，通常用唯一标识符表示。
- 广义命名实体还包括时间（如12:00）、日期（如2017年10月17日）、数量表达式（如100）、金钱（如一亿美金）、产品名、疾病名、手术名、股票名等。

### 任务

NER任务主要包括两个步骤：

1. **实体边界识别**：确定实体的起始和结束位置。
2. **实体类别标注**：将识别出的实体分类到特定类别，如人名、机构名、地名、时间、日期、货币和百分比等。

### 特点

NER的难点在于：

- 时间、日期、货币和百分比的构成有较为明显的规律，识别相对容易。
- 人名、地名、机构名等识别难度较大，因其数量巨大、表达形式多样且没有严格的规律可循。
- 某些类型的实体名称用字灵活，表达形式多样。

### 方法

常见的NER方法包括：

- **基于词典的方法**：通过与预定义的词典进行匹配实现命名实体识别，典型方法包括正向最大匹配和反向最大匹配。
- **基于统计的方法**：包括生成式方法和判别式方法。生成式方法如隐马尔可夫模型（HMM），判别式方法如最大熵模型（Maxent）、支持向量机（SVM）、条件随机场（CRF）、卷积神经网络（CNN）、循环神经网络（RNN）以及结合LSTM和CRF的方法。
- **条件随机场（CRF）**：一种用于序列标注问题的概率模型，通过定义在边和节点上的特征函数进行实体识别。
- **基于LSTM+CRF**：结合双向LSTM和CRF模型，通过LSTM学习特征并由CRF进行序列标注，实现更高精度的命名实体识别。
- **基于阅读理解**：将NER任务建模为机器阅读理解任务，通过自然语言形式的问题识别不同类型的实体。
- **基于模板生成**：通过预定义模板生成自然语言判断句，利用预训练模型如BART进行命名实体识别。

## 关系知识抽取

### 任务

关系抽取旨在自动识别由一对概念和联系这对概念的关系构成的相关三元组。典型例子包括：

- "比尔盖茨是微软的CEO" 可以被抽取为三元组：`CEO(比尔盖茨, 微软)`
- "CMU坐落于匹兹堡" 可以被抽取为三元组：`Located-in(CMU, 匹兹堡)`

### 关系类别

关系抽取面临的一个重要挑战是关系类别的多样性。例如：

- **ACE评测语料**包括61种关系类别。
- **TAC-KBP**和**SemEval**评测语料也定义了多种关系类别。
- 真实环境中的关系类别数量更为庞大，如Freebase、DBpedia、NELL和Knowledge Vault等知识库中包含成千上万的关系类别。

### 难点

关系抽取的主要难点包括：

- **自然语言的多样性**：同一关系可以有多种不同的表述方式。
- **自然语言的歧义性**：相同的表述在不同语境下可以表示不同的关系。

### 方法

常见的关系抽取方法包括：

- **传统方法**：依赖于特征工程，如词汇特征、句法特征和内核特征。这些方法通常需要复杂的NLP工具来提取特征。
- **深度学习方法**：利用CNN、注意力机制和多示例学习等技术，从句子中自动学习特征进行关系分类和抽取。
- **大模型方法**：如GPT-RE，通过构建提示、任务感知样例检索和真实标签诱导推理等方法，利用大模型进行关系抽取。

# 知识图谱生命周期—知识获取

### 输入

知识图谱的知识获取过程需要以下输入：

- **知识本体**：定义领域内的概念及其关系。
- **海量数据**：包括结构化、半结构化和非结构化的数据源，如文本、垂直站点、百科和多模态数据。

### 输出

知识获取的输出是实例化的知识，包括：

- **实体集合**：如人名、地名、机构名等。
- **事件集合**：如某人做了什么事情。
- **实体关系/属性**：如"比尔盖茨是微软的CEO"。
- **事件关系**：如"地震导致海啸"。
- **过程知识**：如某个事件发生的前因后果。
- **常识知识**：如"水在0度以下会结冰"。

### 主要技术

知识获取主要依赖于信息抽取技术，通过自动化手段从各种数据源中提取和整理知识。

# 非结构化文本信息抽取示例

非结构化文本的信息抽取通常涉及从新闻报道、社交媒体等文本中提取有用的信息。例如，从一篇报道中提取关于地震的具体信息，包括时间、地点、震级、震源深度等。

# 信息抽取研究方向

**信息抽取（Grishman, 1997）**
信息抽取的研究方向包括从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出。

**命名实体识别**
命名实体识别的研究内容涉及如何定义和识别命名实体、如何处理命名性指称、名词性指称和代词性指称，以及如何在知识图谱中应用命名实体识别技术。

**关系知识抽取**
关系知识抽取的研究内容包括自动识别实体间的关系，生成关系三元组。具体任务包括关系分类、实体关系联合抽取、多关系抽取和远程监督关系抽取等。

# 基于大模型的生成式信息抽取

**背景**
随着大语言模型的出现，传统的预训练加微调的学习范式逐渐被上下文情景学习范式取代。大模型能够有效地建模各种任务之间的关联关系，通过文本生成实现统一范式的信息抽取。

**方法**

- **GPT-NER**：利用大模型进行命名实体识别。具体方法包括使用KNN方法从训练集中检索语义相似的样例，构造提示进行上下文学习，对每一个实体类型进行标注。
- **生成式信息抽取**：通过大模型进行开放式信息抽取任务，利用大模型强大的上下文学习能力，根据用户指令或少数样例完成抽取任务，解决数据稀缺问题。

# 多模态命名实体识别

**背景**
在社交媒体和短视频平台上，不仅存在文本信息，还存在大量的图片、视频和音频信息。相比于传统的文本形式，社交媒体和短视频平台上的文本具有长度短、表示不规范、各种语言和表情包混杂等特点。

**挑战**

- **利用何种模态的信息**：如语音、图像、视频。
- **如何融合多种模态的信息**：需要考虑多种模态之间的联系。
- **多模态数据中的噪声**：引入多模态信息可能会同时引入噪声。

**方法**

- **文本+图像**：通过视觉模态信息辅助文本命名实体识别，如利用视觉注意力模型识别社交媒体中的命名实体。
- **文本+语音**：利用语音信息（如停顿、音调、韵律）辅助命名实体识别，特别是在提供准确的词边界信息上有重要作用。

## 远程监督关系抽取

**起源**

远程监督关系抽取利用知识库对文本自动进行回标，生成远程监督数据集。这个方法无需人工标注，获取代价低，易于扩展到大规模的场景。

**方法**

- **多示例学习**：将远程监督关系抽取视为多示例学习问题，使用分段卷积神经网络进行预测。
- **基于注意力机制**：利用关系向量对每个句子进行查询，通过注意力机制为包中的每个句子赋予权重，联合利用所有句子信息进行预测。
- **强化学习**：利用强化学习方法优化远程监督关系抽取模型。

## 开放式关系抽取

**方法**

开放式关系抽取通过识别表达语义关系的短语，抽取实体间的关系。典型的方法包括：

- **TextRunner**：从大规模网络文本中抽取关系三元组，并计算其可信度。
- **ReVerb**：利用句法和统计数据过滤抽取出来的三元组，确保关系短语是以动词为核心的短语。

## 总结

大模型下的信息抽取表现出与传统方法相当的性能，特别是在零样本和小样本学习范式下，具有显著优势。然而，领域特定的信息抽取仍有很大研究价值，需要进一步提升大模型的鲁棒性和适应性。