事件知识抽取

事件

定义

事件是特定时间点或时间段、特定地域范围内，由一个或多个角色参与的动作或状态的变化。事件在自然语言处理中的定义可以和日常生活中的理解相似，例如婚礼、地震等。一个事件的发生通常包含以下几个要素：

- **时间**：事件发生的时间点或时间段。
- **地点**：事件发生的具体位置。
- **角色**：参与事件的主体，例如人、机构、设备等。
- **动作**：角色执行的行为或事件发生的情况。

事件类型与元素

- **事件类型**：代表事件的类别。例如“地震事件”表示地震发生的事件类型，“结婚事件”表示婚礼事件类型。
- 事件元素：构成事件的具体细节，包括时间、地点、角色等。例如在“1992年10月3日，奥巴马与米歇尔在三一联合基督教堂结婚”中：
  - **时间**：1992年10月3日
  - **地点**：三一联合基督教堂
  - **角色**：奥巴马（配偶），米歇尔（配偶）

示例

例如，文本“1992年10月3日，奥巴马与米歇尔在三一联合基督教堂结婚”可以被解析为：

- **实体抽取**：识别文本中的实体，如“1992年10月3日”，“奥巴马”，“米歇尔”，“三一联合基督教堂”。
- **关系抽取**：识别实体之间的关系，如“奥巴马”和“米歇尔”之间的夫妻关系。
- **事件抽取**：确定事件类型和相关元素，如结婚事件，时间为1992年10月3日，地点为三一联合基督教堂，参与者为奥巴马和米歇尔。

术语

- **事件描述（Event mention）**：文本中描述事件的部分。例如“奥巴马与米歇尔在三一联合基督教堂结婚”。
- **事件触发词（Event Trigger）**：指示事件发生的关键词，如“结婚”。
- **事件元素（Event argument）**：事件中的重要组成部分，如时间、地点、角色。
- **元素角色（Argument role）**：事件元素在事件中的角色，如“时间”、“地点”、“参与者”。

基于机器学习的方法

神经网络方法

神经网络在事件抽取中具有显著优势，主要在于其能够自动学习特征，并处理复杂的上下文关系。常用的神经网络方法包括：

- 卷积神经网络（CNN）：通过卷积层捕捉局部特征，用于事件检测和领域适应。CNN在提取空间特征方面表现出色。
  - 示例：Yubo Chen等人（2015）的工作利用动态多池卷积神经网络进行事件抽取，显著提升了事件检测的性能。

- 循环神经网络（RNN）：擅长处理序列数据，特别是长距离依赖关系。RNN常用于语言无关的事件检测和事件预测。
    - 示例：Xiaocheng Feng等人（2016）提出了语言独立的神经网络，用于事件检测，能够处理多种语言的文本数据。
- 多层感知机（MLN）：通过多层网络结构结合框架和事件信息，提高事件检测的性能。
    - 示例：Shulin Liu等人（2017）利用FrameNet框架中的信息改进了事件检测的性能。

生成式方法
将事件抽取任务转化为生成问题，使用预训练模型如T5、BART进行微调。生成式方法可以自动生成事件描述，但也面临一些挑战：

- **人工构造提示固定不变**：预定义的提示无法适应不同的事件类型，限制了模型性能。
- **未考虑事件类型之间的关联**：在处理多个事件时，未能有效利用事件类型之间的关系。

基于动态前缀微调的事件抽取
动态前缀微调通过为每个事件类型初始化一组向量（静态前缀），并通过多头注意力机制动态调整前缀向量，建模事件类型之间的关联性，提升模型性能。

基于代码生成的事件元素抽取
将事件元素抽取任务转化为代码生成任务，利用大语言模型（如GPT-3）的代码生成能力，实现事件元素的准确提取。这种方法可以更好地处理复杂的事件结构和依赖关系。

事件关系抽取

任务
事件关系抽取旨在自动识别和分类事件之间的关系，包括同指关系、因果关系、时序关系和子事件关系。

- **同指关系**：识别文本中提到的同一事件。
- **因果关系**：识别一个事件引发另一个事件的因果关系。
- **时序关系**：识别事件发生的时间顺序。
- **子事件关系**：识别复杂事件中的子事件。

共指关系
共指关系识别涉及找到不同文本中提到的同一事件。例如，多个新闻报道可能描述同一事故，但使用不同的表达方式。

因果关系
因果关系识别通过利用外部知识和图神经网络增强模型性能。例如，使用知识库中的描述性和关联性知识，构建因果推理模型。

时序关系
时序关系识别使用BERT和图神经网络进行建模，通过分析事件描述的时间信息，确定事件发生的顺序。

子事件关系
子事件关系识别在复杂事件中提取子事件。例如，战争中的不同战役可以看作是整体战争事件的子事件。

脚本知识抽取

脚本

定义

脚本是特定上下文场景中的事件序列，用于表示过程性知识。脚本可以帮助理解和预测事件的顺序和依赖关系。例如：

- **餐馆脚本**：包括点餐、上菜、结账等一系列事件。
- **手机故障脚本**：包括发现问题、联系客服、维修等步骤。

脚本事件学习

脚本事件学习涉及如何从事件序列中学习和预测事件，包括：

- **脚本事件排序**：将无序的事件序列重新排列为有序的事件序列。
- **脚本事件预测**：给定事件链，从候选列表中预测最合适的后续事件。
- **脚本事件生成**：根据目标生成符合目标的一系列事件。

脚本事件生成方法

增强、检索、生成三阶段方法

- **增强**：利用外部知识库（如ATOMIC）预训练增强生成模型（如T5）。
- **检索**：通过检索器获取相关的事件序列。
- **生成**：根据目标和检索结果生成事件序列。

多粒度知识联合抽取

背景

现有信息抽取方法面临挑战：

- 难以为每个任务设计特定架构。
- 训练独立模型限制了知识共享。
- 构建不同知识源和数据集成本高。

需要设计通用信息抽取框架，对不同任务统一建模。

通用结构生成：UIE

动机

将实体、关系、事件和情感等任务建模为统一的文本到结构生成框架，适用于全监督、低资源和小样本等场景。

方法

- **结构化抽取语言（SEL）**：统一表示不同的抽取结构，如实体位置、实体关系。
- **结构模式指导器（SSI）**：基于模式的提示机制，控制模型进行不同任务。

预训练数据集构建

通过构建文本-结构数据集、结构数据集和文本数据集，训练统一的信息抽取模型，提升模型的泛化能力。

基于标注指南和大模型的联合抽取

动机
提升大模型理解复杂标注指令的能力，实现任务模式理解和人类定义的对齐。

方法
将输入输出表示为代码形式，把标注指南表示为注释形式，通过微调模型遵循指令进行生成，提升大模型在信息抽取任务上的表现。

总结

信息抽取任务尚未被大语言模型（如ChatGPT）彻底解决，仍需提升模型的指令理解能力、任务模式理解和事实准确性。未来的发展方向包括从句子级到跨篇章的信息抽取，以及从单模态到多模态的信息抽取。具体策略包括：

- **提升模型理解复杂标注指令的能力**：通过更精细的提示和注释，提高大模型的准确性。
- **实现任务模式理解和人类定义的对齐**：确保大模型能够正确理解和执行任务，避免生成不准确的信息。
- **从句子级到跨篇章的信息抽取**：在更大范围内进行信息抽取，捕捉更全面的事件信息。
- **从单模态到多模态的信息抽取**：结合文本、图像、视频等多种模态的信息，提供更完整的事件描述。