# NLP学习算法2

- NER tricks
  - trick 1：领域词典匹配
    - 场景：对于某些 常见短语，可以 采用 词典匹配 的方式。
    - 方法：构建一个 常见短语 的 词典，比如 药物、疾病 等，然后采用 flashtext 进行 关键词匹配；
    - 优点：
      - 能够准确的挖掘出 常见短语；
      - 效率更快
    - 缺点：
      - 对于有些 嵌套实体，如果 长实体未包含在词典中，那么将匹配到 短实体；
      - 词典收集工作量大
  - trick 2：规则抽取
    - 场景：对于一些 规定句式，可以 采用 规则匹配 的方式。
    - 方法：构建一些 规则模板库，比如 “ 去|到|抵达|经过 ”、“ 能够|可以 治疗 ” 等；
    - 优点：
      - 对于某些固定句式，这种方法匹配度高；
      - 效率快；
    - 缺点：
      - 会出现干扰词，eg: “ 去|到|抵达|经过|访 ” 抽取 "特朗普和第一夫人访华" -> (特朗普和第一夫人,)、(华，)；
      - 需要手工制定规则；
  - trick 3：词向量选取：词向量 or 字向量？

- 词向量
  - 方式：首先对句子进行分词，然后训练 所有词语 的向量表示，最后利用 这些 词向量 训练模型；
  - 优点：
    - 能够 帮助 模型 学习 句子中 词汇关系；
  - 缺点：
    - OOV 问题；
    - 维护成本高；
    - 如果分词效果不好，那么词向量的质量 将受影响；
- 字向量
  - 方式：首先对句子按字切分，然后训练 所有字的 向量表示，最后利用 这些 字向量 训练模型；
  - 优点：
    - 解决了 词向量的 OOV 问题；
    - 减少人工维护成本；
    - 不用分词；
    - 在训练数据质量较差的时候（比如口语化较多，错别字较多，简称缩写较多等），采用字向量的效果好于词向量；
  - 缺点：
    - 学不出 词语间 的 关系；
  - 解决方法：
    - 利用 具有 双向 的 特征提取器 能够 缓解 该功能，eg: bilstm、bert 等；
- trick 4：特征提取器 如何选择？
  - 短句子：
    - 模型：LSTM、BiLSTM、CNN、IDCNN

        - 优点：
            - 在句子较短的情况下， 模型能够 捕获 句子中词语间的依赖关系
    - 长句子：
        - 模型：Bert
        - 优点：
            - 在句子较长的情况下， 由于 LSTM、BiLSTM、CNN、IDCNN 会出现 长距离依赖问题，所以性能下降；
- trick 5：专有名称 怎么 处理？【注：这一点来自于 命名实体识别的几点心得 】
    - 场景：#1机组1A锅炉磨煤机故障，#2机组2C炉磨煤机故障。 实体是磨煤机。
    - 方法：在训练ner模型时，可以将一类专业名词改写成一个符号表示
    - 具体操作：
        - #1机组、#2机组、#3机组...是一类机组名词，可用符号<Unit>表示；
        - 1A锅炉，1A炉，1B炉，1C锅炉...是一类锅炉专业名词，可用<Speciality>符号表示；
    - 转化后：
        - <Unit><Speciality>磨煤机故障，标注：[OOBIIOO]
- trick 6：标注数据 不足怎么处理？【这个问题可以说是现在很多小厂最头疼的问题】
    - 问题介绍：随着模型的愈发精细复杂，需要训练的参数日益庞大，但其训练所需的人工标注数据却因为标注成本的问题难以得到相应地增长。
    - **方法一：远程监督标注数据**

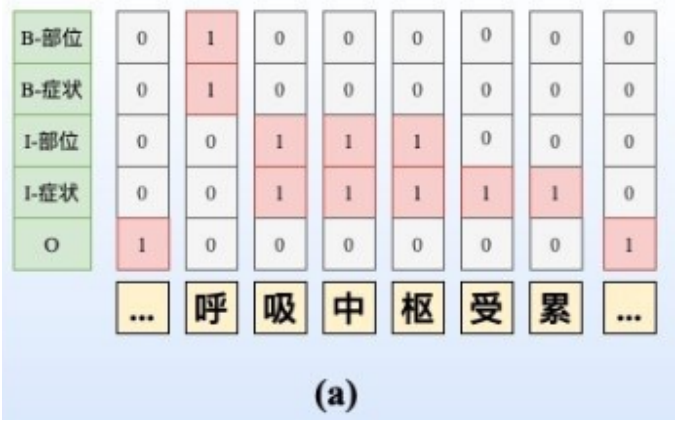
- 思路：使用远程监督的方法来得到大量的远程监督标注数据
- 问题：有限覆盖问题（Limited Coverage）。由于用于远程监督的知识库规模有限，大量的实体存在于文本中而未出现在知识库中，导致在远程监督时，将这些未出现在知识库中的实体标注为非实体，从而产生大量的假负例；

- **方法二：优化模型**
  - 思路：限制参数量，从而使得模型能够在较小的标注数据集上也能够完成训练；
- **方法三：采用主动学习方法**
  - 思路：
    - step 1：先标注一小部分数据，利用这部分标注数据训练模型；
    - step 2：利用模型去标注 未标记数据；
    - step 3: 利用 查询函数 筛选 信息量最大的数据；（方法：信息熵计算不确定样本法、多模型投票选取争议性最高样本法）；
    - step 4：由人工进行标注，并加入到 标注数据中，回到 step 1，直到 样本足够大，或者 模型预测值趋于平衡；
  - 优点：
    - 减少标注成本。由于 选取的数据 所包含的信息量较高，所以减少数据 标注成本（本人之前做过一个小实验，发现采用主动学习方法筛选出的总样本的30%左右，训练出的命名实体识别模型性能与全量训练效果相近）；
    - 数据质量高；
  - 缺点：

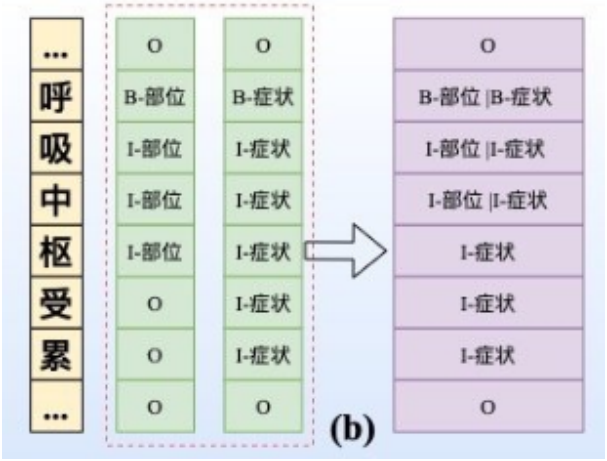
- 如果 查询函数 选取 不对，可能吃力不讨好，也就是 选取的样本存在偏差，比如选到了 离群点。

- **方法四：迁移学习 【这种方法也蛮常见，问题就是风险太高】**
  - *迁移学习定义：将某个领域或任务上学习到的知识或模式应用到不同但相关的领域或问题中。*
- **方法五：预训练+自训练 【Self-training Improves Pre-training for Natural Language Understanding】**
  - 背景知识：
    - 方法：
      - 预训练（Pre-training）从广义上来讲，是指先在较大规模的数据上对模型训练一波，然后再在具体的下游任务数据中微调。大多数情况下，预训练的含义都比较狭窄：在大规模无标注语料上，用自监督的方式训练模型。这里的自监督方法一般指的是语言模型；
        - 通常语言模型说的是，一元、二元、三元模型，但是bert的模型语言模型似乎不太一样。深入浅出讲解语言模型 - 知乎 (zhihu.com)
      - 自训练是说有一个Teacher模型Ft和一个Student模型Fs，首先在标注数据上训练Ft，然后用它对大规模无标注数据进行标注，把得到的结果当做伪标注数据去训练Fs。
    - 相同点：用到了大规模无标注的数据
    - 区别：
      - 预训练始终对针对一个模型进行操作，而自训练却用到了两个模型；
      - 预训练是直接从无标注数据中学习，而自训练是间接地从数据中学习；

- 思路：
  - 将一个预训练模型（本文使用RoBERTa_Large）在标注数据上训练，作为教师模型Ft；
  - 使用Ft从海量通用语料中提取相关领域的数据，一般是 抽取出 置信度较高的样本；
  - 用Ft对提取的数据作标注；
  - 用伪标注语料训练学生模型Fs。
- **方法六：实体词典+BERT相结合**
  - 利用实体词典+BERT相结合，进行半监督自训练 **【注：参考 资料11】 **
- trick 7：嵌套命名实体识别怎么处理 【注：参考 资料3】
  - 7.1 什么是实体嵌套?
    - 实体嵌套是指在一句文本中出现的实体，存在某个较短实体完全包含在另外一个较长实体内部的情况，如“南京市长”中地名“南京”就嵌套在职务名“南京市长”中。
  - 7.2 与 传统命名实体识别任务的区别
    - 传统的命名实体识别任务关注的都是平坦实体（Flat entities），即文本中的实体之间不交叉、不嵌套。
  - 7.3 解决方法：
    - **7.3.1 方法一：序列标注**
    - 多标签分类
      - 思路：命名实体识别本来属于基于字的多分类问题，嵌套实体需要将其转化为 多标签问题（即每个字有多种标签，如下图所示）
      - 问题：
        - 学习难度较大
        - 容易导致label之间依赖关系的缺失




- 合并标签层
  - 思路：采用CRF，但设置多个标签层，对于每一个token给出其所有的label，然后将所有标签层合并
  - 问题：
    - 指数级增加了标签；
    - 对于多层嵌套，稀疏问题较为棘手；




- **7.3.2 方法二：指针标注**
- 层叠式指针标注
  - 思路：设置 C 个指针网络


- MRC-QA+指针标注

- 思路：构建query问题指代所要抽取的实体类型，同时也引入了先验语义知识，如下图（d）所示。在文献中就对不同实体类型构建query，并采取指针标注，此外也构建了 **1 矩阵来判断span是否构成一个实体mention。

(d) MRC-QA+指针标注

- **7.3.3 方法三：多头标注**
- 构建 span 矩阵
  - 思路：构建一个 ** 的Span矩阵
  - 说明：如图，Span{呼}{枢}=1，代表「呼吸中枢」是一个部位实体；Span{呼}{累}=2，代表「呼吸中枢受累」是一个症状实体；
  - 问题：
    - 如何构造Span矩阵问题
    - 如何解决0-1标签稀疏问题


- 嵌套实体的2篇SOTA之作：
  - ACL20的《Named Entity Recognition as Dependency Parsing》采取Biaffine机制构造Span矩阵 **【注：具体可以参考 资料12】**https://arxiv.org/pdf/1909.01441.pdf；
  - EMNLP20的HIT **【注：具体可以参考 资料13】**https://aclanthology.org/2020.emnlp-main.486.pdf则通过Biaffine机制专门捕获边界信息，并采取传统的序列标注任务强化嵌套结构的内部信息交互，同时采取focal loss来解决0-1标签不平衡问题。

- **7.3.4 方法四：片段排列**
- 十分直接，如下图（f）所示。对于含T个token的文本，理论上共有 = (+1)/2 种片段排列。如果文本过长，会产生大量的负样本，在实际中需要限制span长度并合理削减负样本。


- trick 8：为什么说「词汇增强」方法对于中文 NER 任务有效?
  - 动机：虽然**基于字符的NER系统通常好于基于词汇（经过分词）的方法**，但**基于字符的NER没有利用词汇信息**，而**词汇边界对于实体边界通常起着至关重要的作用。**
  - 目标：**如何在基于字符的NER系统中引入词汇信息**
  - 思路：
    - 方法一：设计一个动态框架，能够兼容词汇输入 **【注：具体可以参考 资料6-10】**
    - 方法二：采用多种分词工具和多种句法短语工具进行融合来提取候选实体，并结合词典进行NER
- trick 9：NER实体span过长怎么办?
  - 动机：如果NER任务中某一类实体span比较长（比如医疗NER中的手术名称是很长的），直接采取CRF解码可能会导致很多连续的实体span断裂;
  - 解决方法：
    - 加入规则进行修正

        - 引入指针网络+CRF构建多任务学习。指针网络会更容易捕捉较长的span，不过指针网络的收敛是较慢的，可以对CRF和指针网络设置不同学习率，或者设置不同的loss权重。
- trick 10: NER 标注数据噪声问题?
    - 动机：NER 标注数据存在噪声问题，导致模型训练效果差
    - 方法：
        - 方法一：对训练集进行交叉验证，然后人工去清洗这些“脏数据”
        - 方法二：将noisy label learning应用于NER任务，惩罚那些噪音大的样本loss权重 **【注：具体可以参考 资料12】**
- trick 11： 给定两个命名实体识别任务，一个任务数据量足够，另外一个数据量很少，可以怎么做?
    - 动机：NER 标注数据 有些类别 标注数据量 较少；
    - 方法：
        - 重采样
        - loss惩罚
        - Dice loss
        - 若 该类实体属于 长尾实体（填充率低），可以挖掘相关规则模板、构建词典库
- trick 12： NER 标注数据不均衡问题?
    - 迁移学习
        - 假设：数据量足够任务为 T1，数据量很少任务为 T2。
        - 思路
            - 首先在任务T1中训练模型，然后模型利用之前学习任务所得的知识，应用于任务T2。也就是说模

型在任务T1学习知识（特征、权重），然后推广这一知识（特征、权重）至任务T2（明显数据更少）。

- 半监督策略，即引入虚拟对抗；
  - 思路：随机生成一个扰动，然后进行导数，L2规范化等处理，生成当前所需要的扰动，将其加入到model原始input embedding，再次求损失；模型最终优化loss + loss（带扰动）;使得模型鲁棒性更强，准确率更高

- 事件抽取
  - 原理篇
    - 1.1 什么是事件?
      - 事件在不同领域中有着不同的含义， 对于事件目前还没有统一的定义。在 IE ( Information Extraction) 中， 事件是指在某个特定的时间片段和地域范围内发生的， 由一个或多个角色参与， 由一个或多个动作组成的一件事情， 一般是句子级的。在 TDT ( Topic Detection Tracking) 中， 事件是指关于某一主题的一组相关描述， 这个主题可以是由分类或聚类形成的。
    - 1.2 什么是事件抽取?
      - 事件抽取技术是从非结构化的信息中抽取出用户感兴趣的事件， 并以结构化的形式呈现给用户。
      - **1、Closed-domain**
      - 事件抽取使用预定义的事件模式从文本中发现和提取所需的特定类型的事件。事件模式包含多个事件类型及其相应的事件结构。D.Ahn首先提出将ACE事件抽取任务分成四个子任务:触发词检测、事件/触发词类型识别、事件论元检测和参数角色识别。我们使用ACE术语来介绍如下事件结构:
        - **「事件提及」**：描述事件的短语或句子，包括触发词和几个参数。

- 「**事件触发词**」：最清楚地表达事件发生的主要词，一般指动词或名词。
- 「**事件论元**」：一个实体，时间表达式，作为参与者的值和在事件中具有特定角色的属性。
- 「**论元角色**」：论元与它所参与的事件之间的关系

- **2、Open domain**
- 在没有预定义的事件模式的情况下，**开放域**事件抽取的目的是从文本中检测事件，在大多数情况下，还可以通过提取的事件关键词聚类相似的事件。事件关键词指的是那些主要描述事件的词/短语，有时关键词还进一步分为触发器和参数。
  - 「**故事分割**」：从新闻中检测故事的边界。
  - 「**第一个故事检测**」：检测新闻流中讨论新话题的故事。
  - 「**话题检测**」：根据讨论的主题将故事分组。
  - 「**话题追踪**」：检测讨论先前已知话题的故事。
  - 「**故事链检测**」：决定两个故事是否讨论同一个主题。
- 前两个任务主要关注事件检测;其余三个任务用于事件集群。虽然这五项任务之间的关系很明显，但每一项任务都需要一个不同的评价过程，并鼓励采用不同的方法来解决特定问题。
- 事件抽取涉及自然语言处理、机器学习、模式匹配等多个学科，对于相关学科理论的完善和发展起到积极的推进作用。同时，在情报研究工作中事件抽取技术能帮助情报人员从海量信息中快速提取相关事件，提高了情报研究工作的时效性，并为开展定量情报分析提供技术支撑。事件抽取在情报研究领域具有广阔的应用前景。

- 1.3 ACE测评中事件抽取涉及的几个基本术语及任务是什么?
- 1、**实体(Entity)**。属于某个语义类别的对象或对象集合。其中包括:人(PER)、地理政治区域名( GPE)、组织机构

（ORG)、地名(LOC)、武器(WEA)、建筑设施(FAC)和交通工具( VEH)。

- 2、**事件触发词(Event Trigger)**。表示事件发生的核心词,多为动词或名词。
- 3、**事件论元(Event Argument)**。事件的参与者,主要由实体、值、时间组成。值是一种非实体的事件参与者,例如工作岗位( Job-Title)。和实体一样,ACE05也标记出了句子中出现的值和时间。下文中,即将实体、值、时间统称为实体。
- 4、**论元角色(Argument Role)**。事件论元在事件中充当的角色。共有35类角色,例如,攻击者( Attacker)、受害者( Victim)等。
- 其中，我常用的ACE 2005定义了8种事件类型和33种子类型。其中,大多数事件抽取均采用33种事件类型。事件识别是**基于词**的**34类**( 33类事件类型+None)多元分类任务,角色分类是**基于词**对的**36类**(35类角色类型+None)多元分类任务。这里,参考文献

- 1.4 事件抽取怎么发展的?
  - 从事件抽取的发展历史来看,事件抽取的研究几乎与信息抽取的研究同时开始。 20世纪七、八十年代 ,耶鲁大学就针对新闻报道如地震 、工人罢工等领域或场景 ,开展有关故事理解的研究, 并根据故事脚本理论建立信息抽取系统,就是针对事件抽取的研究 ,开创了事件抽取研究的先河 。但是真正推进事件抽取研究进一步发展的动力主要是相关的评测会议的推动。
  - 消息理解会议(MessageUnderstandingConference, MUC)对事件抽取这一研究方向的确立和发展起到了巨大的推动作用。 MUC定义的抽取任务的各种规范以及确立的评价体系已经成为事件抽取研究事实上的标准,同时也为事件抽取技术的研究奠定了坚实的基础。 MUC是由美国国防高级研究计划委员会(Defense Advanced Research Projects Agency, DARPA)资助 ,从 1987年开始到 1998年 , 会议共举行了 7届 , 具体的历次会议信息如表 1所示 。当前， 由 MUC 定义的

概念、模型和技术规范对整个信息抽取领域起着主导作用， 其主要的评测项目是从新闻报道中提取特定的信息，填入某种数据库中， 事件抽取 ( Scenario Template， ST) 始终是这一会议的评测项目之一。MUC 会议的很多研究都是探索性的， 对信息抽取领域起到了巨大的推动作用，并为事件抽取的研究打下了坚实的基础。每一届 MUC 都针对一个特定领域和场景，并且提供预先定义好的模板 ( Template) 进行填充， 填充之后的模板形成了对文本核心事件的整体描述。

表 1 MUC会议发展历程

| 历届会议 | 时间 | 相关评测任务 |
| --- | --- | --- |
| MUC-1 | 1987. 05 | 没有明确的任务定义，也没有制定评测标准。 |
| MUC-2 | 1989. 05 | 规定了模式以及槽的填充规则，抽取任务被明确为一个模式填充的过程。 |
| MUC-3 | 1991. 05 | 抽取任务是从新闻报道中抽取拉丁美洲恐怖事件的信息，定义的抽取模式由 18个槽组成。 |
| MUC-4 | 1992. 06 | 仍然是从新闻报告中抽取恐怖事件信息。但抽取模式变得更复杂了，总共由 24个槽组成。 |
| MUC-5 | 1993. 08 | 抽取的任务更加复杂，而且包含了两个目标场景，并引入了嵌套的模式结构。 |
| MUC-6 | 1995. 09 | 除了模式填充任务外，又增加了命名实体识别、共指关系确定和模式元素填充 3个任务。 |
| MUC-7 | 1998. 04 | 在 MUC-6的基础上又增加了一个新的任务——模式关系任务。 |

- 在强烈的应用需求下 ,来自美国国家标准技术研究所 (NIST) 组织的 ACE评测会议应运而生, 这项评测真正推动了事件抽取研究的发展 。从 1999年开始酝酿, 2000年正式开始启动 。研究的主要内容是自动抽取新闻语料中出现的实体 、关系 、事件等内容, 即对新闻语料中实体 、关系、事件的识别与描述。
- 与 MUC相比, ACE评测不针对某个具体的领域或场景, 采用基于漏报 (标准答案中有而系统输出中没有)和误报 (标准答案中没有而系统输出中有 )为基础的一套评价体系 , 还对系统跨文档处理 (CrossDocumentProcessing)能力进行评测。这一新的评测会议把事件抽取技术研究引向新的高度 。具体的历次会议信息如表 2所示 。ACE 会议作为 MUC 会议的延伸， 是事件抽取领域最具影响力的评测会议， 该会议从 2000 年到 2007 年共举办了 7 届。目前大多数研究都是围绕 ACE 的评测任务开展，它把事件抽取的研究推向一个新的高度。会议研究的主要内容是自动抽取新闻语料中出现的实体、关系、事件等内容。ACE 定义的事件属于元事件的范畴， 包括事件类别和事件元素的识别。与 MUC 相比， ACE 评测不针对某个具体的领域或场景， 也不提供预先定义好的模板， 而是强调对文本基本意义或基本概念的刻画， 因此所定义的任务显得更为细致和深入。用户指定要检测的事件的类别， 系统给出检测文本中这些

事件的出现， 但最后的输出并未形成对核心事件的整体描述， 并且 ACE 同 MUC 一样都是篇章级 ( Document Level) 的事件抽取， 不涉及跨文档抽取。

表 2　ACE会议发展历程

| 历届会议 | 时间 | 相关评测任务 |
| --- | --- | --- |
| ACE-Pilot | 2000. 05 | 任务是实体检测与识别，评测语料是英语。 |
| ACE-1 | 2000. 11 | 同 ACE-Pilot |
| ACE-2 | 2002. 11 | 任务是实体检测与识别和关系检测与识别，评测语料是英语。 |
| ACE2003 | 2003. 09 | 任务同 ACE-2 评测的语料是英语、中文和阿拉伯语。 |
| ACE2004 | 2004. 09 | 任务是实体检测与识别、关系检测与识别和事件检测与识别（即事件抽取）评测的语料是英语、中文和阿拉伯语。 |
| ACE2005 | 2005. 08 | 任务是实体检测与识别、价值检测与识别、时间表达识别与标准化、关系检测与识别以及事件检测与识别，评测的语料是英语、中文和阿拉伯语。 |
| ACE2007 | 2006. 10 | 任务是实体检测与识别、价值检测与识别、时间表达识别与标准化、关系检测与识别、事件检测与识别以及实体翻译，评测的语料是英语、中文和阿拉伯语。 |
| ACE2008 | 2008. 05 | 任务被分为两组，一是文档内的任务，另一个是跨文档的任务，评测的语料是英语和阿拉伯语。 |

- 自 2009年 , ACE成为文本分析会议(TextAnalysisConference, TAC)中的一个任务 。 TAC主要由 3个评测任务组成 ,主要目的是促进自然语言处理技术发展和相关的应用 。
- 总而言之,从 20世纪七 、八十年代开始,事件抽取一直发展到今天 ,已经走过了几十年的研究历程 ,其所取得的进步与上述评测会议的推动密不可分 ,但从评测会议公布的结果来看, 抽取的精度离实用还相差甚远, 在领域扩展性和移植性方面的表现还不能令人满意, 注定今后事件抽取技术的研究还有很长的路要走。

- 1.5 事件抽取存在什么问题?
  - 1）对实体、关系识别、语法分析等相关技术的底层技术研究不够成熟，导致级联错误。事件抽取是在实体和关系识别的基础上发展起来的。它在某种程度上取决于实体、关系识别和文本预处理的效果，但是这些基础技术仍然不够成熟。并且，目前缺乏对子任务输出结果的评估及矫正技术。
  - 2）事件抽取系统的现场可伸缩性和便携性并不理想。例如，有关中文事件抽取的相关研究主要集中在生物医学、微博、新闻、紧急情况等方面。其他领域和开放领域的研究很少。关于领域和跨语言事件抽取技术的研究很少。
  - 3）缺乏大规模成熟的语料和标注语料，需要进一步完善。手动标注语料库既费时又费力，而且缺少语料库限制了事件抽取技术研究的发展。因此，大型语料库的自动构建技术方法需要进一步研究。

- 4）如何设计神经网络模型以实现多任务联合是一大难点。

- 基本任务篇
  - 2.1 触发词检测
    - 2.1.1 什么是触发词检测?
      - 表示事件发生的核心词，多为动词或名词
    - 2.1.2 触发词检测有哪些方法?
      - 现有的检测事件句的方法主要是基于触发词的方法。Grisman、赵妍妍等都是采用这种方法来发现文本中的事件句。在这类方法中,将每个词作为一个实例来训练并判断是否为触发词的机器学习模型,但引入了大量的反例,导致正反例严重不平衡。为了解决了上述问题,哈尔滨工业大学的谭红叶提出了一种基于局部特征选择和正负特征相结合的事件检测与分类方法,取得了不错的识别效果。厦门大学的许红磊等人也提出了一种新的事件类别自动识别算法,很好地克服了传统基于触发词方法所带来的正反例失衡和数据稀疏问题。
  - 2.2 类型识别
  - 2.2.1 什么是类型识别?
  - ACE2005 定义了8种事件类型和33种子类型。其中，大多数事件抽取均采用33 种事件类型。事件识别是基于词的34 类(33类事件类型+None) 多元分类任务，角色分类是基于词对的36 类(35类角色类型+None) 多元分类任务。
  - 事件类别识别是指从文本中检测出事件句,并依据一定的特征判断其所归属的类别。不难看出,事件类别识别是典型的分类问题,其重点在于事件句的检测和事件句的分类。
  - 2.2.2 类型识别有哪些方法?
  - 在已有的研究中 ,事件句分类主要采用最大熵模型(MaximumEntropyModel, MEM)和支持向量机

(SupportVectorMachine, SVM)。赵妍妍和许红磊等人分别使用上述两种分类器基于二元分类策略实现了候选事件句的类别识别,但二元分类的最大缺陷就是无法处理一个事件句属于多个事件类别的情况 ,多元分类应该是更合理的选择 。事件句分类的难点主要是如何选择合适的描述事件句的特征提高分类精度 。赵妍妍等人选取词法 、上下文和词典信息等语言学特征对候选事件进行描述 ,在 ACE2005中文语料上取得F-值为 61.2%的效果 。付剑锋等人在此基础上引入依存分析发掘触发词与其它词之间的句法关系 ,并以此为特征在 SVM分类器上对事件句进行分类 F-值提高到 69.3%

- 可见,事件类别的识别率还有很大的提升空间,选择更加合适的分类器以及事件特征进一步提高识别效果仍有待于下一步研究与探讨。
- 2.3 角色识别
- 2.3.1 什么是角色识别?
- 事件角色识别是事件抽取中又一核心任务 。该任务主要从众多命名实体(Entity)、时间表达式 (Time Expression)和属性值 (Value)中识别出真正的事件元素 ,并给予其准确的角色标注。事件句中通常包含大量的 Entity、TimeExpression和 Value等事件信息, 要想从中筛选出真正的事件元素, 首先要把所有信息识别并标注出来,而这也正是 MUC会议的主要研究内容 。在事件元素识别中 ,假定在文本预处理阶段已完成事件信息的识别与标注 。
- 事件的参与者，主要由实体、值、时间组成。值是一种非实体的事件参与者，例如工作岗位。
- 2.3.2 角色识别有哪些方法?
- 事件角色识别与语义角色标注 (SemanticRoleLabeling, SRL)任务有一定的相似性。所谓语义角色标注是根据一个句子中的动词(谓词)与相关的各类短语等句子成分之间的语义关系而赋予这些句子成分的语义角色信息 ,如

施事、受事 、工具或附加语等。于江德等探索了基于条件随机场 (Conditional Random Fields, CRFs)的方法对任职事件和会见事件的事件元素进行角色标注 ,取得了不错标注效果,也从侧面揭示了事件元素与语义角色之间存在着一定的对应关系 。吴刚等利用这种对应关系实现了事件角色的识别,然而该方法依赖的底层模块较多,如 : 分词、句法分析、SRL等 ,如果底层处理模块不够成熟,将会导致级联错误过多 ,影响事件元素识别效果 。赵妍妍等是将事件元素识别看作分类问题 ,使用最大熵模型 ,选取词法、类别、上下文和句法结构等 4类特征多角度地描述候选元素 ,采用二元分类和多元分类两种策略实现了事件元素的自动识别 。

- 2.4 论元检测
- 2.4.1 什么是论元检测?
- 事件论元在事件中充当的角色。共有35类角色，例如，攻击者 、受害者等。
- 2.4.2 论元检测有哪些方法?

以上内容整理于 幕布文档