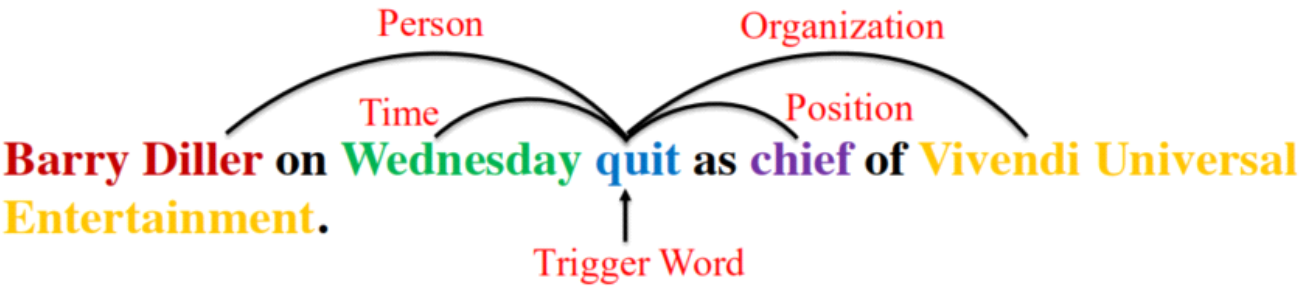
SEU 知识抽取-事件抽取

2019年11月25日 9:03

- **事件抽取**
  - **事件的定义**

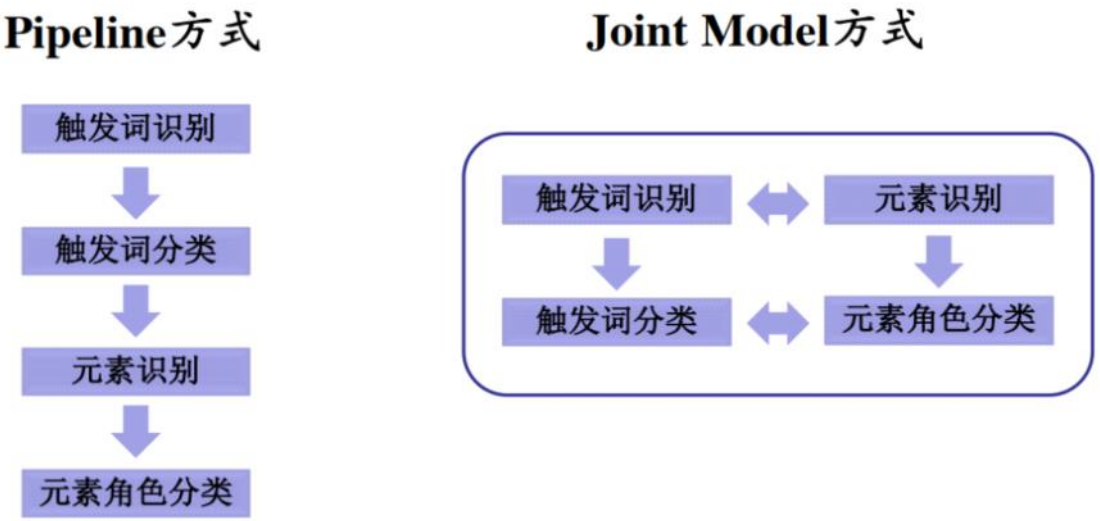
    事件是发生在某个特定的时间点or时间段、某个特定的地域范围内，由一个或多个角色参与的一个或者多个动作组成的状态的改变。
  - **事件抽取**

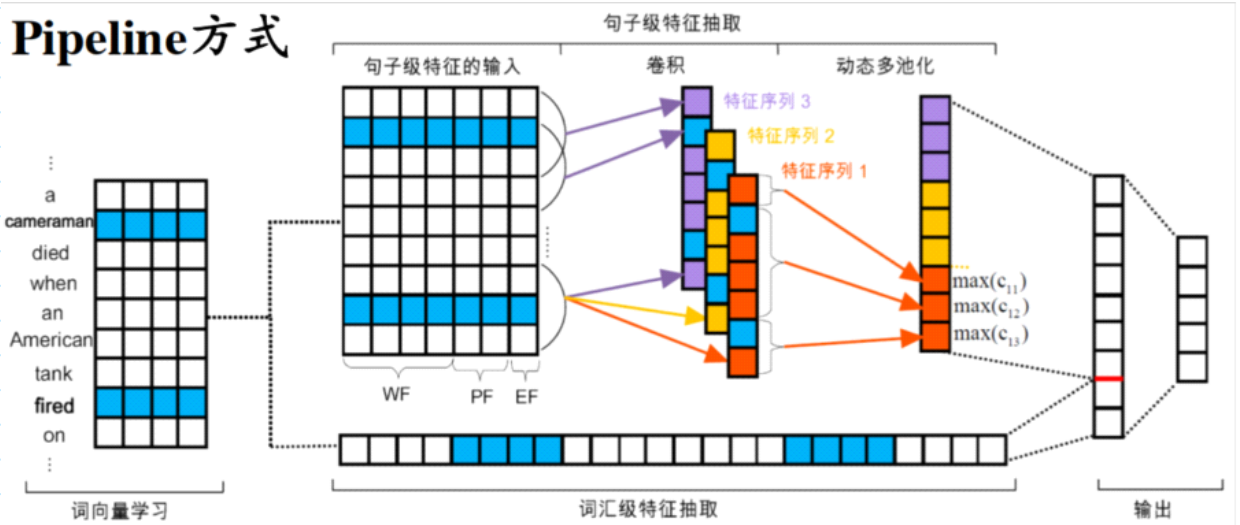
    从自然语言文本中抽取出来用户感兴趣的事件并以结构化的形式展示出来。如什么人、组织，在什么时间，在什么地方，做了什么事情。
  - **相关术语**
事件描述Event Mention：	描述事件的句子
事件触发词Event Trigger：	标记事件类型的词汇
事件元素Event Arugment：	事件的参与者
事件角色Event Role：	元素在事件句中扮演的角色
  - **事件抽取相关子任务**
    - 事件发现
事件触发词检测	Event Trigger Detection
事件触发词分类	Event Trigger Typing
    - 事件元素抽取
事件元素识别	Event Argument Identification
事件元素角色识别	Event Argument Role Identification



Trigger	quite("Personnel/End-Position" event)	
Argument	Role=Person	Barry Diller
	Role=Organization	Vivendi Universal Entertainment
	Role=Position	Chief
	Role=Time-within	Wednesday(2003-03-04)

- **事件抽取的方法**
  - 基于模板和规则的方法
    - 定义语义框架和短语模式来表示特定领域事件的抽取模式
  - 基于机器学习的方法
    - 传统的分类任务
    - 依赖依存分析、句法分析、词性标注等传统的NLP工具
  - 基于深度学习的方法
    - Pipeline
    - Joint Model
  - 深度学习的优势
    - 减少了对外部NLP工具的依赖，建模形成了端到端的系统
    - 模型使用词向量作为输入，词向量本身蕴含了丰富的语义信息
    - 避免了人工去设计大量的特征





- **Lexical-level（词汇级别特征）**
  - S1:Obama beats McCain ->Elect Event
  - S2：Tyson beats his opponent ->Attack Event
- **Sentence-level（句子级别特征）**
  - 上下文特征（WF）：Argument Candida 的上下文
  - 位置特征（PF）：当前词和需要预测类型的candida的相对位置
  - 事件类型特征（EF）：将触发词的事件类型编码为特征辅助元素识别与分类

- **Dynamic Multi-Pooling**

  传统的Max Pooling直接对每个feature map做一个max操作来提取句子中最有用的信息，dynamic pooling是将每个feature map根据**候选元素**和**触发词**来进行分割操作，即把每个feature map根据元素和触发词切分成三块，然后分别计算max value

- **Encoding Phase**
  - Word Embedding：使用word2vec训练的词向量
  - Entity Type Embedding：将实体类型作为特征并编码为向量
  - Dependency Tree Relation Embeddings：One-hot Embedding，当在依赖关系树中存在与Wi相连的边时，则第i维的值为1

Prediction Phase:

- **Trigger Prediction**
  - h_i: the hidden vector to encapsulate the global context of the sentence.
  - L_i^{trg}: the local context vector for w_i.
  - G_{i-1}^{trg}: the memory vector from the previous step.
  - Output: A Feed-Forward Neural Network with a softmax layer.
- **Argument Prediction**
  - h_i and h_{i_j}: the hidden vectors to capture the global context of the input sentence for w_i and e_j.
  - L_{ij}^{arg}: the local context vector for w_i and e_j.
  - B_{ij}: the hidden vector for the binary feature vector V_{ij}.
  - $G_{i-1}^{arg}[j]$ and $G_{i-1}^{arg/trg}[j]$: the memory vectors for e_j.
  - Output: A Feed-Forward Neural Network with a softmax layer.

- **Joint Model相比于Pipeline方式的优势：**
  - 避免了误差累计传播导致模型性能下降的问题
  - 使用一个模型同时抽取出所有的事件信息
  - 使用从整体结构中学到的全局特征来提升对局部信息的预测能力

- **FrameNet**
  - 语言学家定义的语义框架
  - 一个词法数据库：事件、实体、关系和参与者
  - 包含1000+框架、1w多个词法单元、15w条标注句子

- **为什么要在事件抽取中使用强化学习?**

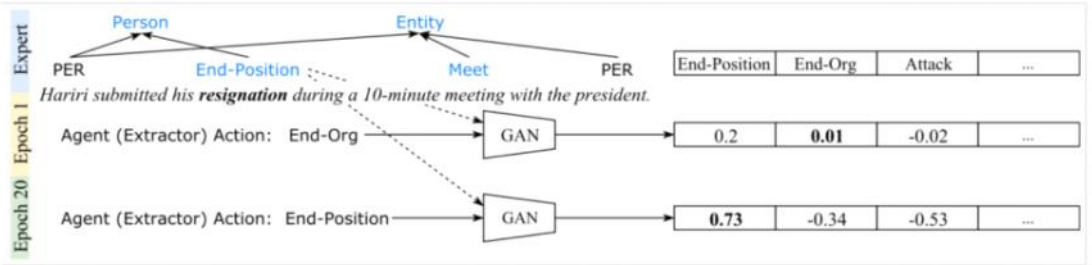
  传统的事件抽取模型中，都是在有监督的背景下训练模型学习正确的label，而没有对有错误的label进行深入分析。这样的模型最终学习到的是如何正确的标记句子中的实体，但是碰到有歧义的实体时，模型将无法正确区分。因此，提升模型的性能不仅在于教会模型去学习正确的label，对错误label的学习也非常重要。通过让机器去模拟人的行为，对正确的行为给予奖励，错误的行为获得惩罚，从而将强化学习的思想引入到事件抽取任务中。

- **Event Extraction with Generative Adversarial Imitation Learning（GAIL)**

  核心思想：

  使用强化学习的思想通过Q-learning的方式对句子做序列标注，标记出句子里面的entity 和trigger 然后使用policy gradient判定事件中entity 对应的argument role。

  训练过程中强化学习使用到的reward 有GAN动态生成。



- **Epoch1: 错误的将“resignation”标注为“End-Org”.**
- **Epoch20:通过GAN不断扩大在正确行为和错误行为上发放奖励之间的差距，使得抽取器在接下来的训练中正确标记实体.**