# 综述和概念

综述资料：

<https://blog.csdn.net/muumian123/article/details/81746583>

## 事件抽取任务分类

参考文献：[1]高强,游宏梁.元事件与主题事件抽取技术研究综述[C].//中国科学技术情报学会.第25届全国计算机信息管理学术研讨会论文集.2011:114-121.

事件抽取任务可分为元事件（**Meta Event**）抽取及主题事件（**Topic Event**）抽取。当前的事件抽取研究主要面向元事件，而主题事件抽取的研究成果较少。

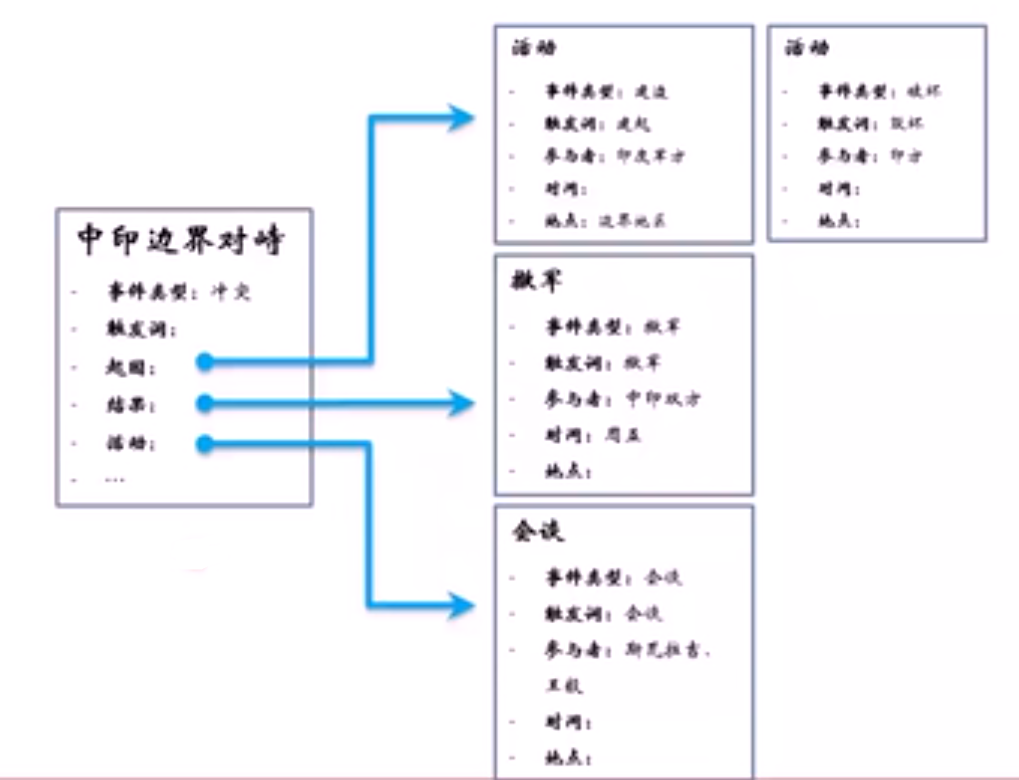
### Meta Event

表示一个动作的发生或状态的变化。往往由动词驱动,也可以由能表示动作的名词等其它词性的词来触发,它包括参与该动作行为的主要成份（如人物、地点、时间等），元事件是主题事件的基本组成单位。

### Topic Event

就是说主题事件包括多个元事件。包括一类核心事件或活动,以及所有与之直接相关的事件和活动，可以由多个元事件片段组成。

感觉就是事件嵌套

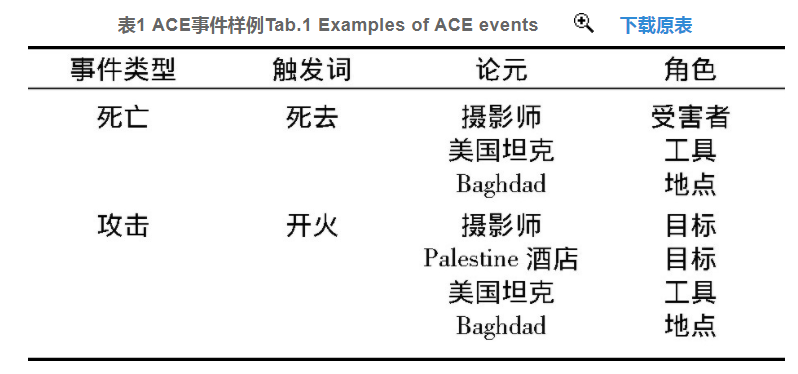


## 一些词的概念

这些词是基于自动内容抽取 (Automatic Content Extraction, ACE) 评测中的英文事件抽取任务涉及的相关术语及任务, 对其展开如下的研究阐释。

* 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) 。
* 触发词 (**Trigger**) 。表示事件发生的核心词, 多为动词或名词。
* 论元 (**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。
* 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。

举个例子便于理解：





## 子任务介绍

参考文献：

[1]秦彦霞,张民,郑德权.神经网络事件抽取技术综述[J].智能计算机与应用,2018,8(03):1-5+10.

[2]Chen Y , Yang H , Liu K , et al. Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention Mechanisms[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[3] <https://www.zhihu.com/question/314000659/answer/613879751>

事件抽取任务可分解为4个小子任务，两个大子任务

* **Event detection (ED)**(事件识别任务)
  + **Trigger Identification** (触发词识别)
  + **Trigger Type Classification** (事件类型分类)
* **Argument Extraction (AE)(**论元角色分类任务)
  + **Argument Identification** (论元识别)
  + **Argument Role Classification** (角色分类任务)

**Event detection**判断句子中的每个单词归属的事件类型, 是一个基于单词的**多分类任务**。

事件类型是数据集标注时规定的

**Argument Extraction**任务则是一个基于词对的多分类任务, 判断句子中任意一对触发词和实体之间的角色关系。

## 抽取方法分类

NLP的任务很多都是分为以下两种方式构建模型和流程

### 流水线Pipeline

每一个子任务都单独用一个模型，造成误差累计

### 联合抽取Joint

用一个模型多任务学习，模型会比较复杂

# 数据集

ACE2005用的最多，但是得花钱，感觉最终选择哪个Corpus还是应该问问老师或者助教的建议，或者再深入调研一下

后来找到个CEC是开源的，感觉可以作为候选

## LDC简介

LDC里有很多NLP数据集，比如下面提到的ACE2005

官网：

<https://www.ldc.upenn.edu/>

浏览全部语料：

<https://catalog.ldc.upenn.edu/byyear>

简介：

<https://www.cnblogs.com/plwang1990/p/4206085.html>

中文的CLDC:

<http://www.chineseldc.org/>

## ACE2005

用这个的特别多，ACE2005语料库是LDC发布的由实体，关系和事件注释组成的各种类型的数据，有中文语料，有点老，要花钱，好像是2400美元一年

语料获取链接：

<https://catalog.ldc.upenn.edu/LDC2006T06>

介绍：

<https://www.jianshu.com/p/71ed0d780210>

参考资料：

<https://blog.csdn.net/qq_39304851/article/details/104694179>

## CEC

这是我找到的第一个开源的语料…我觉得这个可以

<https://github.com/shijiebei2009/CEC-Corpus>

## 其他

科大讯飞2020组织过一个事件抽取比赛，貌似给了语料

<https://zhuanlan.zhihu.com/p/150190165?from_voters_page=true>

TAC

<https://tac.nist.gov/2017/KBP/Event/index.html>

收费的：

<https://www.datatang.com/dataset/info/text/83>

TimeML、MUC这两个事件抽取的会议也可能会提供语料

# 开源代码

就先找到了两个可能会之后借鉴的project

深度学习的方法（基于pytorch写的，提供的测试数据是英文文本）

<https://github.com/Hanlard/Transformer-based-pretrained-model-for-event-extraction>

基于模板匹配的方法

<https://github.com/jialin666/EventExtraction>

这里顺便给出一个事件抽取技术方案简介（<https://zhuanlan.zhihu.com/p/32488490>）：

