## 2021-2022学年秋季学期

# 自然语言处理 Natural Language Processing



授课教师: 胡玥

助 教: 李运鹏

## 中国科学院大学网络空间安全学院专业核心课

## 自然语言处理 Natural Language Processing

## 第9章 信息抽取

授课教师: 胡玥

授课时间: 2021.12

# 内容提要

- 9.1 信息抽取概述
- 9.2 实体识别与抽取
- 9.3 实体消歧
- 9.4 关系抽取
- 9.5 事件抽取

## 9.5 事件抽取

## ■ 事件抽取

## 本节内容:

- 1. 事件抽取概述
- 2. 句子级事件抽取
- 3. 文档级事件抽取
- 4. 事件抽取资源简介

#### ■ 事件的定义

事件是发生在某个特定的时间点或时间段、某个特定的地域范围内,由一个或者多个角色参与的一个或者多个动作组成的事情或者状态的改变

事件由事件触发词(Trigger)和描述事件结构的元素(Argument)构成

触发词: 引发某类事件的一个或几个关键词,大部分为动词

事件结构的元素:不同类型事件结构元素不同,一般由事件模板确定

例1: 输入: 毛泽东 1893 年 出生 于 湖南 湘潭

输出:

事件:

Trigger: 出生

Type: Life, Subtype: Be-Born

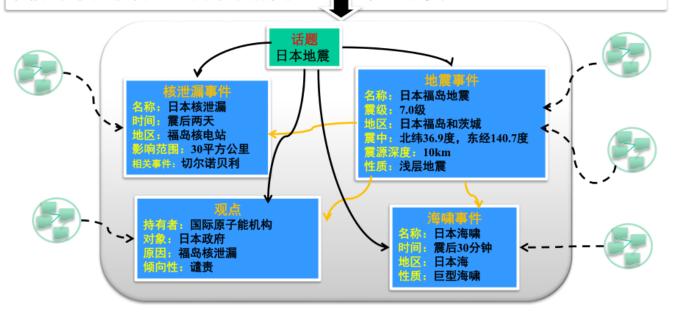
Person: 毛泽东 Time: 1893年

Place: 湖南湘潭

"出生"是该事件的触发词,所触发的事件类别(Type)为Life,子类别(Subtype)为Be-Born;事件的三个组成元素"毛泽东","1893年"、"湖南湘潭",分别对应着该类(CLife/Be-Born)事件模板中的三个元素标签即:Person,Time以及Place

#### 例2:

2011年4月11日17点16分,日本东北部的福岛和茨城地区发生里氏7.0级强烈地震(震中北纬36.9度、东经140.7度,即福岛西南30公里左右的地方,震源深度10公里,属于浅层地震)当局已经发布海啸预警震后约30分钟后在日本海地区发生巨型海啸,同时造成福岛核电站出现核泄漏震后第十天,国际原子能机构对于日本政府反应迟钝进行了谴责



摘自:赵军,知识图谱综述

#### ■ 事件相关概念:

事件描述(Event Mention):是指对一个客观发生的具体事件进行的自然语言形式的描述,通常是一个句子或者句群

事件触发词(Event Trigger): 是指一个事件描述中最能代表事件发生的词,是决定事件类别的重要特征,往往是动词或者名词

事件元素/论元角色(Event Argument): 是指事件中的参与者,是组成事件的核心部分,它与事件触发词构成了事件的整个框架

元素角色(Argument Role):是指事件元素与事件之间的语义关系,也就是事件元素在相应的事件中扮演什么角色

事件类别(Event Type): 事件元素和触发词决定了事件的类别

#### ■ 事件抽取任务(Event Extraction, EE):

事件抽取是NLP领域中一种经典的信息抽取任务,在商业、军事等领域的情报工作中应用非常广泛。其任务主要是从描述事件信息的文本中抽取出用户感兴趣的事件信息并以结构化的形式呈现出来(从无结构化数据中抽取结构化事件信息)

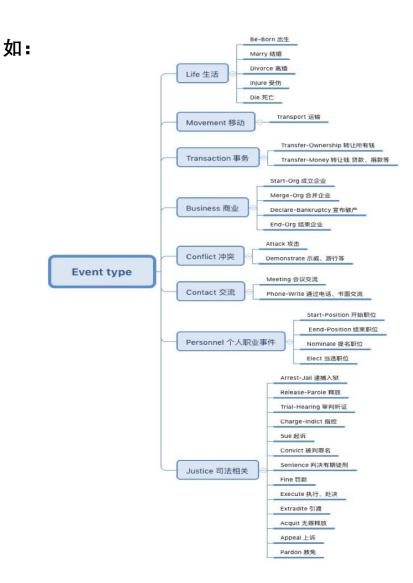
传统事件抽取具体内容:识别一段文本中事件触发词(trigger)、事件触发词所属类型(event type)、<u>事件要素(argument)</u>,其中事件要素包括每个要素涉及的角色(role)。

#### 传统事件抽取任务:事件类别,子类别,事件元素(角色)确定

如,自动内容抽取会议(Automatic Content Extraction, ACE) ACE2005中 标注数据集: 事件类型: 8大类33小类

事件类型(Event Type)	子事件类型(Event Subtype)
生命 (Life)	出生 (Be-Born)、结婚 (Marry)、 离婚 (Divorce)、伤害 (Injure)、死亡 (Die)
移动 (Movement)	运输(Transport)
联系 (Contact)	会面 (Meet)、打电话/写信 (Phone-Write)
冲突 (Conflict)	袭击(Attack)、游行(Demonstrate)
	机构合并 (Merge-Org)、
商务 (Business)	破产声明(Declare-Bankruptcy)、
	机构成立 (Start-Org)、机构终止 (End-Org)
交易 (Transaction)	金钱转移(Transfer-Money)、
X (Hallsaction)	所有权转移(Transfer-Ownership)
人事 (Personnel)	竞选 (Elect)、职位开始 (Start-Position)、
/ (Fersonner)	职位结束 (End-Position)、提名 (Nominate)
	逮捕 (Arrest-Jail)、执行 (Execute)、
	赦免 (Pardon)、假释 (Release-Parole)、
司法 (Justice)	罚款 (Fine)、宣告有罪 (Convict)、
HIA (Justice)	控告 (Charge-Indict)、听证 (Trial-Hearing)、
	开释 (Acquite)、判决 (Sentence)、起诉 (Sue)、
	引渡 (Extradite)、上诉 (Appeal)

#### 不同小类事件涉及不同的论元角色(对应着唯一的事件模板)



Type/Subtype	Template
Life/Be-Born	Person, Time-Within, Place
Business/Merge-Org	Org, Time, Place
Contact/ Meet	Entity, Time, Duration, Place
Transport:	Origin, Artifact, Transporter Vehicle, Destination, Agent,

事件抽取流程:首先识别触发词及类别,然后按识别的的触发词类别(类型/子类)确定模板,按模板要素从给定文本中抽取事件各角色论元

例1: 输入: 毛泽东1893年出生于湖南湘潭

触发词检测: 触发词: 出生

触发词类别: life / Be-Born

#### life / Be-Born 类模板:

Type/Subtype	Template
Life/Be-Born	Person, Time-Within, Place
Business/Merge-Org	Org, Time, Place
Contact/Meet	Entity, Time, Duration, Place
Transport:	Origin, Artifact, Transporter Vehicle, Destination, Agent,

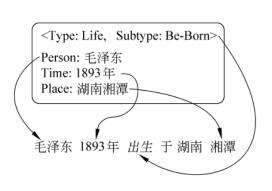
输出: 事件:

Trigger: 出生

Type: Life, Subtype: Be-Born

Person: 毛泽东 Time: 1893年

Place: 湖南湘潭



例2: 输入: Saddam's family left that city three days ago (Event mention)

触发词检测: 触发词: left

触发词类别: Movement / Transport

#### Movement / Transport 类模板:

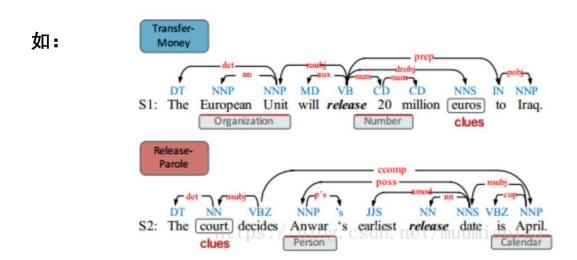
Type/Subtype	Template
Life/Be-Born	Person, Time-Within, Place
Business/Merge-Org	Org, Time, Place
Contact/Meet	Entity, Time, Duration, Place
Transport:	Origin, Artifact, Transporter Vehicle, Destination, Agent,

#### 输出:

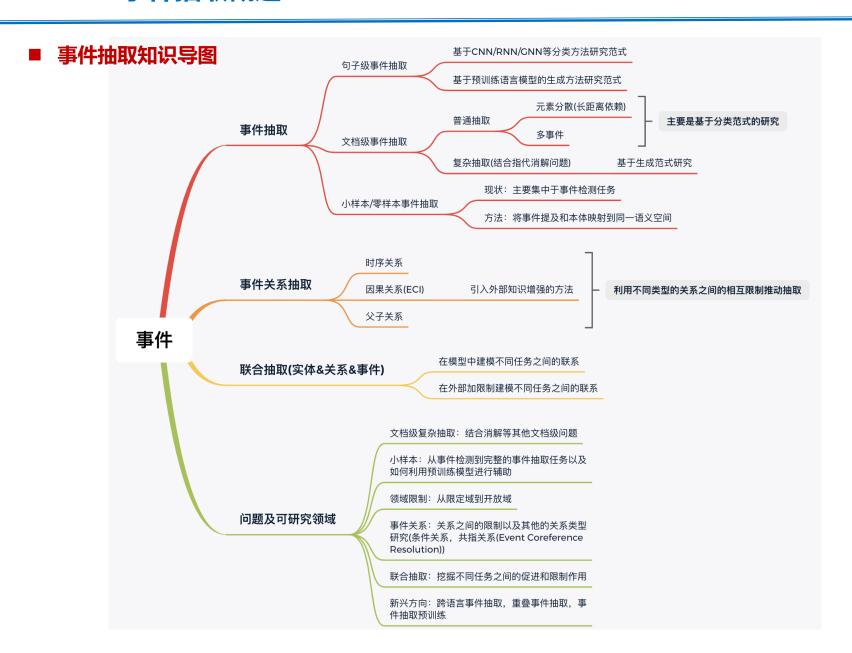
Trigger	left	left		
Event type	Movement Transport	Movement Transport		
	Role=Artifact	Saddam's family		
Arguments	Role=Origin	that city		
	Role=Time-Within	three days ago		

#### 事件触发词识别及分类挑战:

事件触发词识别及分类是事件提取的一个关键且非常具有挑战性的子任务,相同的事件可能以各种触发词形式出现,而同一个词可能在不同的上下文中表示不同的事件类型。



- · S1中:"release"是动词的概念和"Transfer-Money"事件的触发
- 词S2中: "release"是名词的概念和"Release-Parole"事件的触发词



#### ■ 事件抽取实现方法:

1. 流水线方式: 先识别触发词及类别, 然后抽取论语角色

优点: 能够简化整个事件抽取任务

缺点: 容易错误级联 (第一步识别的错误导致第二步识别准确度更底)

2. 联合抽取方式: 同时识别触发词和论元角色

优点: 能够产生Triggers与Arguments之间的双向信息流交互

缺点: 同时抽取模型更复杂,泛化性能差

#### 分章节只讲授流水线方式,分为:

- 句子级事件抽取
- 文档级事件抽取

## 9.5 事件抽取

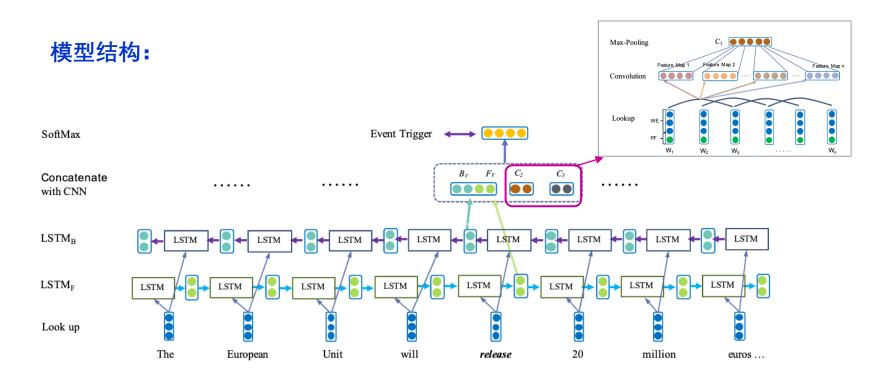
## ■ 事件抽取

## 本节内容:

- 1. 事件抽取概述
- 2. 句子级事件抽取
- 3. 文档级事件抽取
- 4. 事件抽取资源简介

#### **★ HNN (RNN&CNN)**

基本思想:利用Bi-LSTM捕获上下文特征,CNN捕获局部特征共同用于事件类型检测。



输入: 句子

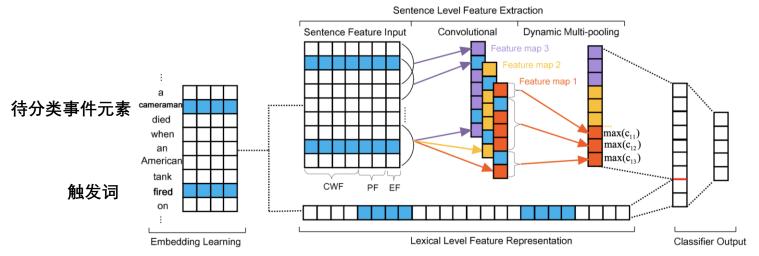
输出: 触发词及其类别

#### **★ Event Extraction via DMCNN**

Pipeline 方法: 事件类型检测模块 → 事件元素分类模块

■ 假设:已完成 触发词识别和分类,进行事件元素识别

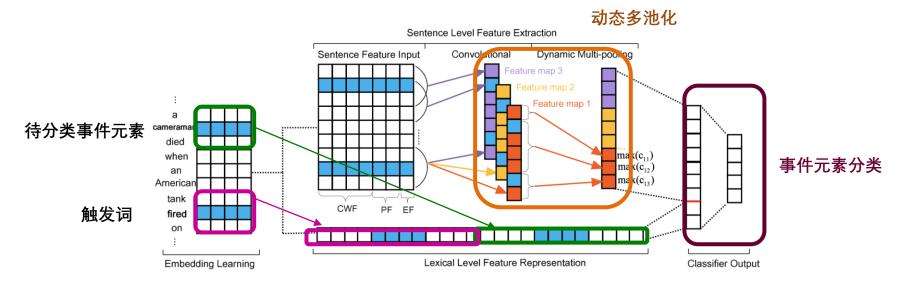
模型结构:



输入: 句子(确定了触发词和指定了要识别元素的词)

输出: 指定元素词类别

#### 模型结构:



- 注: 触发词识别与前面模型架构相同, 略有以下差异:
  - 只有待分类触发词的向量和上下文表示
  - 只根据待分类触发词划分为两部分池化
  - 输出为触发词分类

## 实验:

Methods	<i>'</i>	<b>Frigge</b> :	r	Trigge	er Ident	ification	A	rgume	nt	Argu	ıment	
Wethods	Ident	ificatio	on(%)	+ Cla	assifica	tion(%)	Ident	ificatio	on(%)	Rol	e(%)	
	P	R	F	P	R	F	P	R	F	P	R	F
Li's baseline	76.2	60.5	67.4	74.5	59.1	65.9	74.1	37.4	49.7	65.4	33.1	43.9
Liao's cross-event		N/A		68.7	68.9	68.8	50.9	49.7	50.3	45.1	44.1	44.6
Hong's cross-entity		N/A		72.9	64.3	68.3	53.4	52.9	53.1	51.6	45.5	48.3
Li's structure	76.9	65.0	70.4	73.7	62.3	67.5	69.8	47.9	56.8	64.7	44.4	52.7
DMCNN model	80.4	67.7	73.5	75.6	63.6	69.1	68.8	51.9	59.1	62.2	46.9	53.5

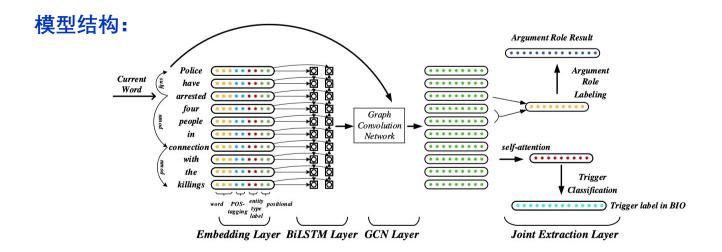
Stage	1/1 (%)	1/N (%)
Trigger	72.7	27.3
Argument	76.8	23.2

Stage	Method	1/1	1/N	all
Stage	Wictiou	$F_1$	$F_1$	$F_1$
	Embedding+T	68.1	25.5	59.8
Trigger	CNN	72.5	43.1	66.3
	DMCNN	74.3	50.9	69.1
	Embedding+T	37.4	15.5	32.6
Argument	CNN	51.6	36.6	48.9
	DMCNN	54.6	48.7	53.5

#### **★** JMEE(GCN):

目标:从一个句子中抽取出多个事件触发器和参数 (arguments)

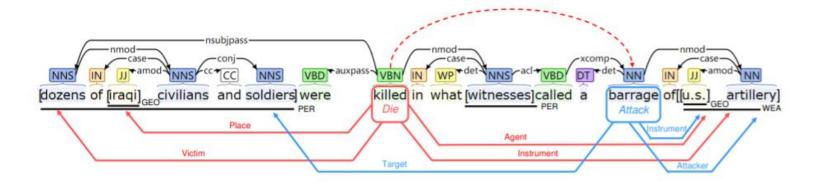
基本思路: 把多个事件的触发器放在一起来考虑,提高事件抽取的正确性。



输入: 句子

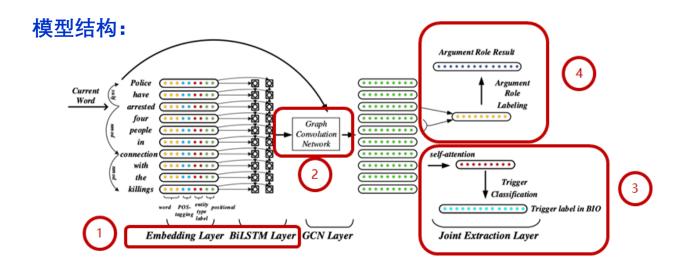
输出: 各触发词及其相关参数

#### 例:



killed是事件Die的触发器,4条红线表示它的4个参数; barrage是事件Attack的触发器,3条蓝线表示它的3个参数。按句子的序列顺序,从killed跳到barrage需要6步。根据句法依存树的话,只需要3步。3个arcs组成了一个shortcut path,使用较少的跳数,将句法依存的信息流从killed节点转移给barrage节点。

#### **★ JMEE(GCN):**



#### JMEE模型由4个模块组成:

- 1. word representation: 使用向量来表示句子
- syntactic graph convolution network module: 根据句法结构引入shortcut arcs,
  进行卷积操作;
- 3. self-attention trigger classifification module: 捕获一个句子中多个事件间的关联;
- 4. argument classifification: 预测每个实体在对应类型的事件中扮演的角色

#### **★ MQAEE:**

基本思想:用任务转换的方式把事件抽取看作是一个多轮问答任务,根据事件抽取的任务定义,可以认为先从输入的文本中询问:文本包含什么事件类型?接着询问:该事件发生的地点是?该事件发生的时间是?将回答的结果整合下,即可完成事件抽取任务。

事件抽取任务要素: 触发词识别+触发词分类 (Yes/No) +事件元素抽取

#### MQAEE中将事件抽取转成三类问题:

第一类问题是: Which word is the trigger word? (for Trigger identification)

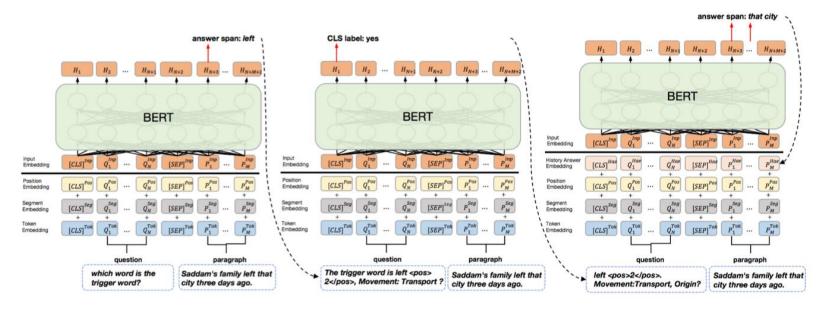
第二类问题是: The trigger word is left <pos> 2 </pos>, movement:

transport? (for Trigger classification)

第三类问题是: left <pos> 2 </pos>. Movement : transport, time-within?

其中<pos>number</pos>表示触发词在文本处的位置。

#### 模型结构:



#### 实验结果:

Model	Tri-Id	Tri-Cls	Arg-Id	Arg-Cls
Model	F1 (%)	F1 (%)	F1 (%)	F1 (%)
JointBeam	N/A	64.2	38.0	35.0
RBPB	N/A	67.8	55.4	43.8
JointEventEntity	N/A	68.7	50.6	48.4
dbRNN	N/A	69.6	57.2	50.1
DYGIE++	N/A	68.9	54.1	51.4
DYGIE++(ens)	76.5	73.6	55.4	52.5
MQAEE	74.5	71.7	55.2	53.4
MQAEE(ens)	77.4	73.8	56.7	55.0

## 9.5 事件抽取

## ■ 事件抽取

## 本节内容:

- 1. 事件抽取概述
- 2. 句子级事件抽取
- 3. 文档级事件抽取
- 4. 事件抽取资源简介

## 3.文档级事件抽取

★ DCFEE: (从句子级事件抽取扩充到文档级)

文档级事件抽取:当前的事件抽取方法大多都是抽取句子中的事件,但是大多数情况下,一个事件是由多个句子描述的。在金融领域,91%的事件论元是分布在多个句子中的

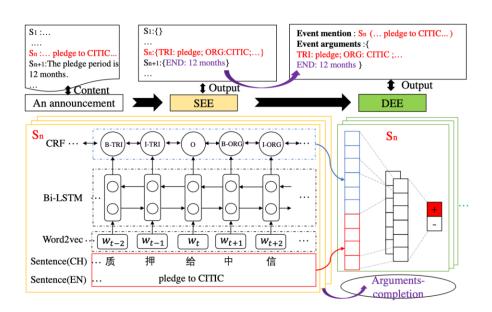


Example of an Equity Freeze event triggered by "frozen" and containing five arguments.

## 3.文档级事件抽取

事件抽取系统:包括基于句子级标注数据的句子级事件抽取(SEE)和基于文档级标注数据的文档级事件抽取(DEE)

#### 模型结构:



- 句子级事件抽取:从单个句子中抽取出事件论元和事件触发词(用序列标注实现)
- 文档级事件抽取:包括两部分:从文档中发现事件提及的关键事件检测模型和基于邻居句子的抽取结果进行元素补充完成文档级抽取模型

## 3.文档级事件抽取

#### 应用实例



不同颜色的单词代表不同的事件参数类型,下划线句子代表文档中提到的事件。如图所示,可以从非结构化文本(关于股权冻结的公告)中获得一个完整的股权冻结事件。

## 9.5 事件抽取

## ■ 事件抽取

## 本节内容:

- 1. 事件抽取概述
- 2. 句子级事件抽取
- 3. 文档级事件抽取
- 4. 事件抽取资源简介

#### ■ 典型的事件抽取会议及语料

- · MUC评测会议
- TDT评测会议
- ACE (KBP)

ACE: 2000~2008; KBP: 2014~2020

#### ■ 领域相关的事件抽取会议及语料

- · BioNLP评测会议-面向生物医学
- · TimeBank语料库-面向问答系统
- · CEC (Chinese Event Corpus) 语料库-国内外突发事件
- · 文档级事件抽取数据集
  - ▶ Doc2EDAG-来自金融网站
  - ▶ RAMS- Reddit上的新闻文档
  - WikiEvents-英文维基百科的事件相关的新闻文档

#### 特点:

给定事件类型 专家标注语料, 语料质量高 抽取的目标类别已经定义好 有公认的评价方式

#### MUC评测会议:

- MUC-2: 在从海军军事情报中抽取事件并填入预定义的模板中。 模板中包括了事件类型、时间、地点、参与者等 10 个槽
- MUC-3: 针对拉丁美洲的恐怖袭击的报道进行评测,模板的槽也增加到了 18 个
- MUC-4: 沿用了 MUC-3 的语料,模板的槽增加到了 24 个
- MUC-5:加入了对日文的评测。主要针对微电子技术领域中芯片技术的进展情况和金融领域中的公司合资情况两类文本进行抽取(金融领域定义了11个模板47个槽)
- MUC-6: 加入汉语、日语、西班牙语、新增: 命名实体识别、共指关系确定和模板元素填充三项任务
- MUC-7: 采用 200 篇新闻报道作为评测语料。抽取任务分为领域相关和无关两个阶段

#### TDT评测会议:

话题识别与跟踪(Topic Detection and Tracking)

- 从TDT-2开始,增加了新事件识别和报道关系识别两个子任务,包含中文、英文和阿拉伯文三种评测语言。
- 发布了 TDT-pilot、TDT-2、TDT-3、TDT-4 和 TDT-5 共 5 种语料
- 由于 TDT 中话题的粒度比事件的粒度大,因此 TDT 的语料标注方法与ACE 等评测的标注方法完全不同。TDT-2 和 TDT-3 采用了"YES"、"BRIEF"和"NO"三类标签分别表示当前报道的内容 与话题绝对相关、部分相关和不相关。而 TDT-4 和 TDT-5 则只采用了"YES" 和"NO"两种标签

#### ACE2005:

自动内容抽取会议(Automatic Content Extraction, ACE)目前使用最广泛的事件抽取标注数据集; 事件类型: 8大类33小类.

1635篇 = 英文: 599篇(16375句); 中文: 603篇; 阿拉伯文: 403篇

事件类型(Event Type)	子事件类型(Event Subtype)
生命 (Life)	出生 (Be-Born)、结婚 (Marry)、 离婚 (Divorce)、伤害 (Injure)、死亡 (Die)
移动 (Movement)	运输(Transport)
联系 (Contact)	会面 (Meet)、打电话/写信 (Phone-Write)
冲突 (Conflict)	袭击(Attack)、游行(Demonstrate)
商务 (Business)	机构合并(Merge-Org)、 破产声明(Declare-Bankruptcy)、
	机构成立 (Start-Org)、机构终止 (End-Org)
交易(Transaction)	金钱转移(Transfer-Money)、
	所有权转移(Transfer-Ownership)
人事 (Personnel)	竞选 (Elect)、职位开始 (Start-Position)、 职位结束 (End-Position)、提名 (Nominate)
司法 (Justice)	逮捕(Arrest-Jail)、执行(Execute)、 赦免(Pardon)、假释(Release-Parole)、 罚款(Fine)、宣告有罪(Convict)、 控告(Charge-Indict)、听证(Trial-Hearing)、 开释(Acquite)、判决(Sentence)、起诉(Sue)、 引渡(Extradite)、上诉(Appeal)

一个事件可能涉及0个或多个论元。ACE2005语料中所有类别共涉及到35个角色。不同类型的事件涉及不同的论元角色,

类、子类及论元角色	类、子类及论元角色
Justice :	Life:
Acquit/Pardon: Defendant, Adjudicator,	Be-Born/Divorce/Marry: Person
Crime	Die/Injure: Victim, Instrument, Agent
Appeal/Charge-Indict/Trial-Hearing:	
Adjudicator, Defendant,	Movement:
Prosecutor, Crime	Transport: Origin, Artifact, Agent,
Arrest-Jail/ Execute: Person, Agent,	Vehicle, Destination,
Crime	Transport er
Convict: Adjudicator, Defendant, Crime	Business:
Extradite: Origin, Person, Agent,	Declare-Bankruptcy/ End-Org/ Merge-
Destination, Crime	Org: Org
Fine: Entity, Money, Adjudicator, Crime	Start-Org: Org, Agent
Release-Parole: Entity, Person, Crime	<i>y</i> 0, 0
Sentence: Sentence, Adjudicator,	Personnel:
Defendant, Crime	Start-Position/ End-Position: Entity,
Sue: Plaintiff, Adjudicator, Defendant, Crime	Person, Position
Crime	Nominate/ Elect: Person, Agent,
Conflict:	Po siti on
Attack: Attacker, Instrument, Target	Transaction:
Demonstrate: Entity	Transfer-Money: Recipient, Giver,
,	Money, Beneficiary
Contact:	Transfer-Ownership: Seller, Artifact,
Meet: Entity, Duration	Beneficiary, Price, Buyer
Phone-Write: Entity	<u> </u>
注解: 表中没有重复列出时间和地点这两个	个一般性论元角色,其中黑体字表示事件大
类, 黑斜体表示事件子类。	

样本标注了事件触发词、事件类型、事件子类型、事件元素和事件元素扮演的角色

#### KBP评测任务:

知识库生成测评任务(Knowledge Base population, KBP) 属于TAC(Text Analysis Conference): https://tac.nist.gov/

- KBP 2014 提供了 599 个标注文档,主要由新闻报道以及数据等构成,由人工过滤来确保每一种事件类型都有多个实例,并且针对长句子进行了截断
- KBP 2015 扩充为 5 个子任务(事件识别、事件识别和消歧、事件消歧、事件 元素抽取与链接、事件元素的验证和链接)
- KBP 2016新添了以文档为单位的识别,提供了200 个标注的英文文档、
  200000 个词的中文文档以及 120000 个词的西班牙文文档,但未提供任务相关的训练数据

#### KBP2020

- ➤ Task 1: 从多媒体文档流中抽取知识元素提及,并聚类生成文档级知识图谱
- ➤ Task 2: 融合文档的知识图谱形成独立于原文档的知识库
- ➤ Task 3: 根据知识图谱生成假设,每个假设代表文档流的语义一致性解释

#### BioNLP评测会议: (从2009-2013年举办了三次)

- 面向生物医学事件抽取的最权威的会议
- 将生物医学事件分为9个类别,每个事件类别具有相应的触发词,不同类别的事件具有1个或多个主题。以及0或多个不同类型的事件元素

#### TimeBank语料库

- 面向问答系统的时间和事件的识别会议
- 主要任务是识别和抽取事件的时间元素以及事件之间的时序关系。
- 语料分别来源于 DUC、ACE 和 PropBank 中收集的新闻报道
- 事件通过事件类(Event Class)、事件时态(Event Tense)以及事件 状态(Event Aspect)三个属性来描述
- 将事件划分为 Occurrence、Perception、 Reporting、Aspectual、
  State、Intensional State、Intensional Action 和 Modal 等 8 种类型
- 不关心时间之外的事件元素, 其语料中包含 300 篇新闻

#### CEC (Chinese Event Corpus) 语料库

#### **CEC-1**:

- 针对 200 篇国内外突发事件的中文新闻报道的标注,包含了 1228 个 句子、3133 个事件和 4878 个事件元素
- 其中标注的颗粒相对较大,尤其是对事件元素的标注,且没有涉及事件关系分类的研究

#### CEC-2:

- 从互联网上收集了333 篇关于地震、火灾、交通事故、恐怖袭击以及 食物中毒等五类突发事件的新闻报道作为待标注语料
- 不仅标注了所有事件,还进行了句法分析和语义分析标注,符合中文的特点,还能对标注后的语料进行一致性检查,保证语料标注的质量
- 不仅包括已标注语料还包括未标注的原始语料。未标注的语料中有关 于语料特性的描述,如语料来源、标题、主体等

#### CEC语料推动了面向中文突发事件的相关研究的发展

#### 文档级事件抽取数据集

#### Doc2EDAG

来源:中文ChFinAnn金融网站上新闻文档

包含5种事件类型,每种事件类型含有6-9个事件角色

链接: https://github.com/dolphin-zs/Doc2EDAG

#### RAMS

来源: Reddit上的新闻文档

链接: http://nlp.jhu.edu/rams/

#### WikiEvents

来源:英文维基百科的事件相关的新闻文档

链接: https://github.com/raspberryice/gen-arg

#### 相关比赛

CCKS(全国知识图谱与语义计算大会):每年都有涉及事件抽取的相关比赛,包括但不限于金融篇章级抽取、事件抽取抽取、医疗领域事件抽取等

LIC (语言与智能技术竞赛):每年都有涉及信息抽取的相关比赛,包括篇章级抽取、多模态抽取等

#### 事件抽取相关论文整理:

https://mp.weixin.qq.com/s/6IUhL7sptWcLGbh62\_mAlA

# 潮海各位!



