

- ▶ 知识图谱的构建流程
- ▶ 实体识别
- ▶ 实体消歧
- ▶ 关系抽取
- ▶ **事件抽取**
  - ▶ 基本定义
  - ▶ 经典模型
- ▶ 开放域知识抽取
- ▶ 多模态知识抽取

# 事件定义

- ▶ 认知科学
  - ▶ 世界上发生的所有事情都被定义为事件，人们通过认识事件以及事件之间的联系来观察和了解世界
- ▶ 哲学
  - ▶ 事件被定义为现实世界中事实的具体表现
- ▶ 语言学
  - ▶ WordNet中事件被定义为在某个具体的时间或地点发生的事情
  - ▶ 事件是一个包含三部分信息的术语：谓词及其发生时间、发生的环境或条件
- ▶ 信息抽取评测会议**ACE**中对事件的定义：

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**事情或状态的改变**。

## ▶ 事件抽取

- ▶ 从描述事件的文本中**识别事件**，抽取用户感兴趣的**事件元素**并以结构化的形式呈现出来。
- ▶ 通常包含事件触发词的识别及分类、事件元素的识别及分类等子任务。

据路透社消息，英国**当地时间9月15日早8时15分**，位于伦敦西南地铁线District Line 的**Parsons Green地铁站**发生爆炸，目前已确定有多人受伤，具体**伤亡人数尚不明确**。目前，英国警方已将此次爆炸与起火定性为恐怖袭击。

### ▶ 恐怖袭击事件



触发词: 发生爆炸

时间: 当地时间9月15日早8时15分

地点: Parsons Green地铁站

攻击者: -

伤亡人数: -

# 事件类型及事件元素角色

▶ ACE定义的事件类型和子事件类型

事件类型 (event type)	子事件类型 (event subtype)
生命 ( life )	出生、结婚、离婚、伤害、死亡
移动 ( movement )	运输
联系 ( contact )	会面、打电话/写信
冲突 ( conflict )	袭击、游行
商务 ( business )	机构合并、破产声明、机构成立、机构终止
交易 ( transaction )	金钱转移、所有权转移
人事 ( personnel )	竞选、职位开始、职位结束、提名
司法 ( justice )	逮捕、执行、赦免、假释、罚款、宣告有罪、控告、听证、开释、判决、起诉、引渡、上诉

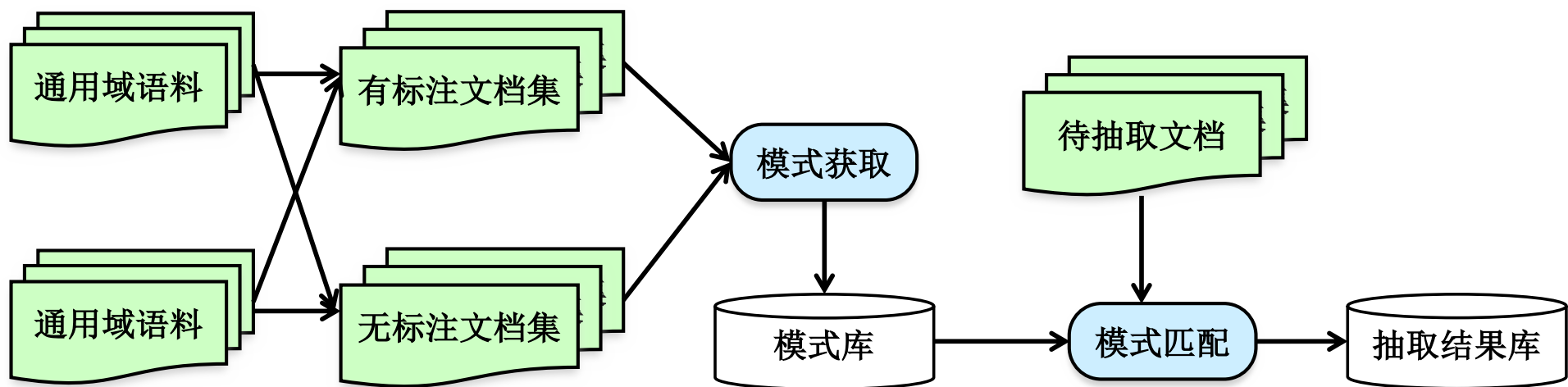
▶ ACE中部分事件元素角色

事件元素角色集合
人物、地点、时间、买家、卖家、价格、起点、重点、接受者、赠与者、袭击者、目标、受害人、原告、被告、陪审员...

# 基于模式匹配的事件抽取

## ► 基于模式匹配的事件抽取

- 在一些模式的指导下，对某种类型的事件进行识别和抽取
- 可分为两个步骤：模式获取、模式匹配



# 基于模式匹配的事件抽取

## ▶ AutoSlog

- ▶ **有监督**模式匹配方法的事件抽取系统（模式由人工标注）
- ▶ 基于“事件元素**首次提及之处**即可确定该元素与事件间关系”和“事件元素**周围的语句**中包含了事件元素在事件中的元素描述”两个假设
- ▶ 通过句法分析和模式匹配共同实现事件抽取

例句：王某在地铁站制造了爆炸。



规则：动词 <制造爆炸> 出现在主动结构，则对应的主语被标记为**嫌犯**

[1] Riloff, Ellen. "Automatically constructing a dictionary for information extraction tasks." AAAI. Vol. 1. No. 1. 1993.

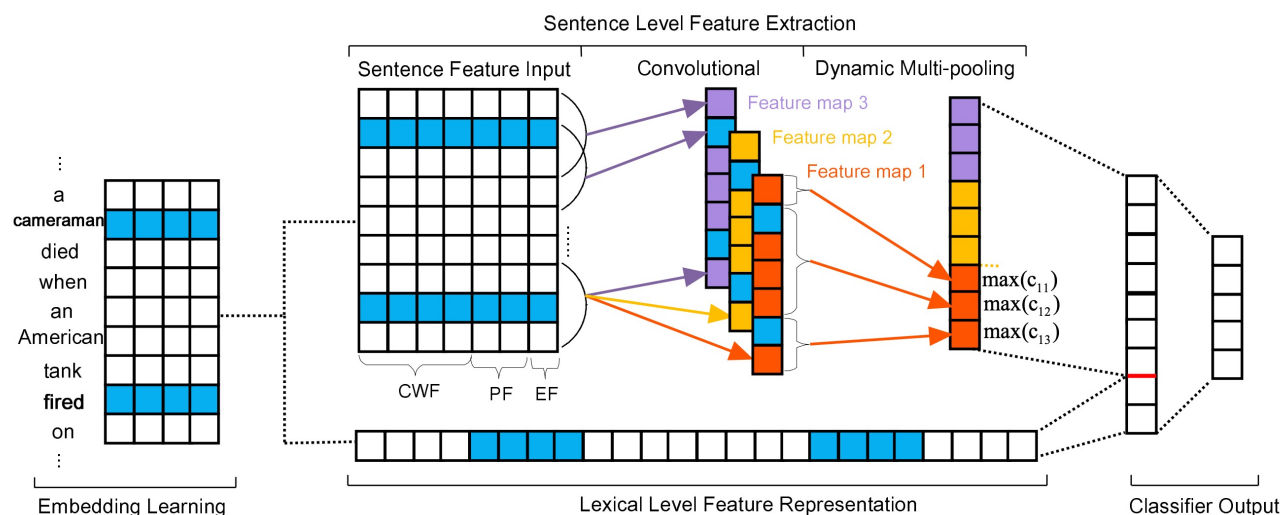
# 基于模式匹配的事件抽取

- ▶ 基于模式匹配的事件抽取
  - ▶ **模板**的准确性是影响整个方法性能的关键因素
  - ▶ 在**特定领域**中性能较好，便于理解和后续应用
  - ▶ 获取模板的过程费时费力，具有很强的专业性
  - ▶ 对于语言、领域、和文档形式等有较强的依赖，覆盖度和**可移植性较差**
- ▶ 基于机器学习的事件抽取
  - ▶ 基于特征工程的方法，通过搭配手动设计的特征和不同的分类器来完成
  - ▶ 基于神经网络的方法，自动从纯文本中提取特征，性能较好

# 基于神经网络的事件抽取

## ► DMCNN

- 模型主体结构为动态多池化卷积神经网络，和关系抽取模型PCNN类似
- 两阶段事件抽取：触发词分类、事件元素分类
- 不依赖句法解析等外部工具，且不需要人工特征设计

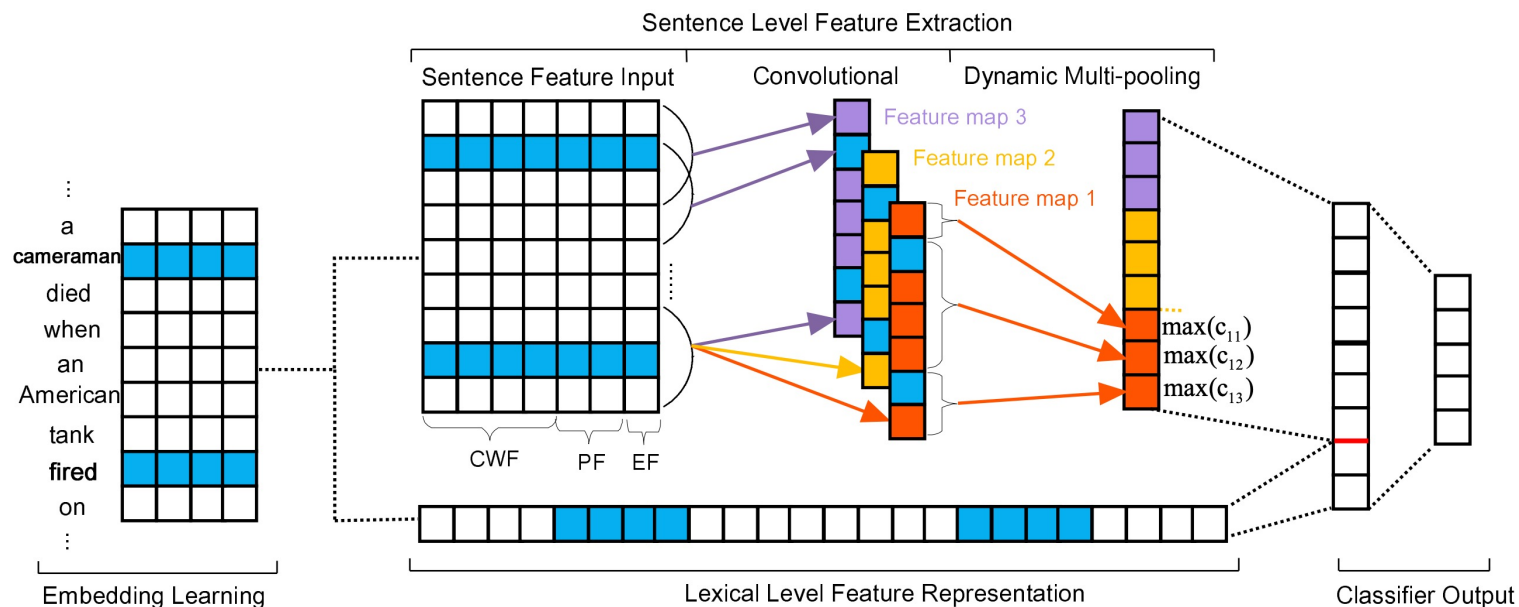


[1] Chen, Yubo, et al. "Event extraction via dynamic multi-pooling convolutional neural networks." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015.



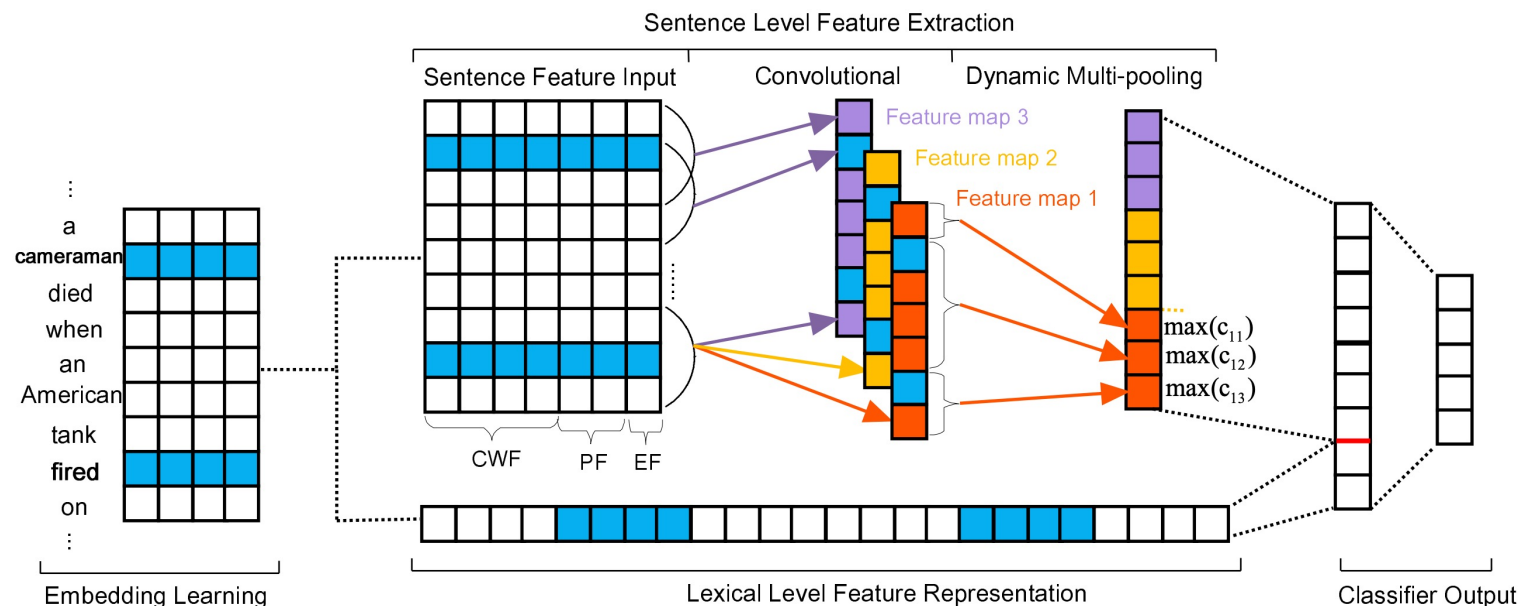
# 基于神经网络的事件抽取

- 两阶段采用相同的模型结构，以事件元素分类为例，给定触发词fired和待分类的候选事件元素词cameraman：



[1] Chen, Yubo, et al. "Event extraction via dynamic multi-pooling convolutional neural networks." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015.

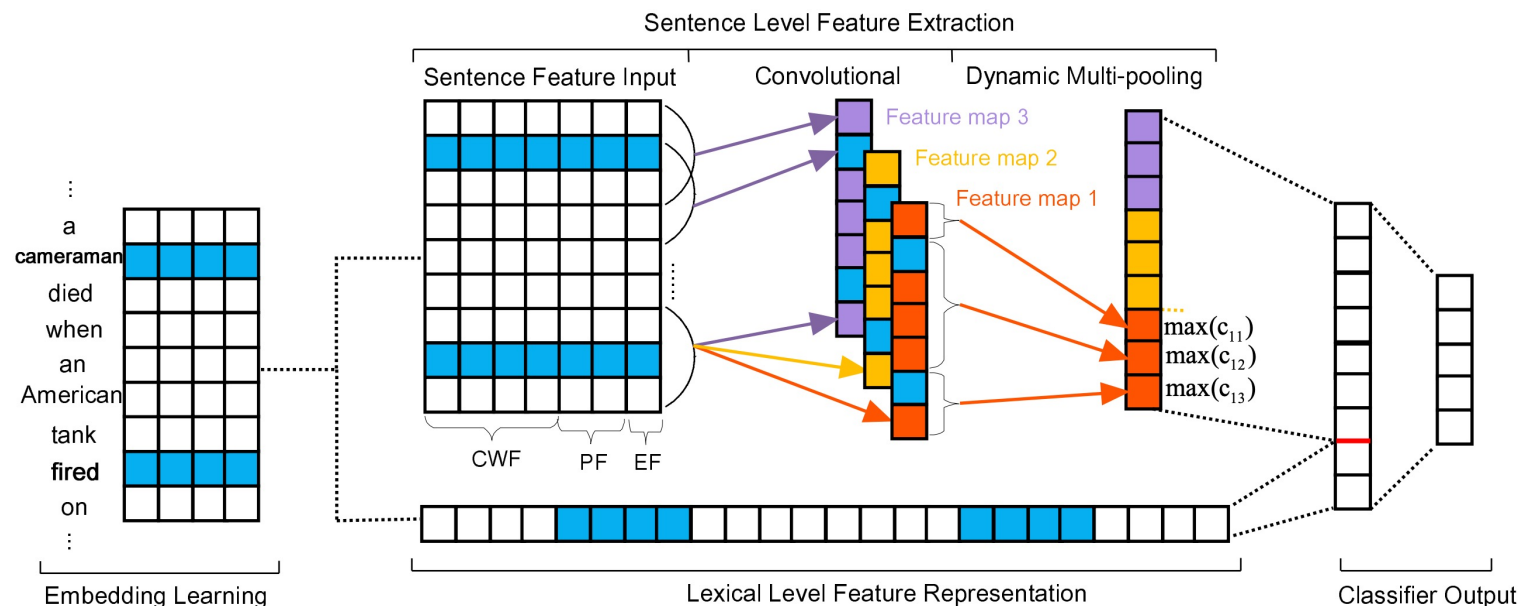
# 基于神经网络的事件抽取



- ▶ 词汇特征由skip-gram算法学习到的词向量而来
- ▶ 句子级特征输入
  - ▶ CWF(context-word feature): 每个词的词向量
  - ▶ PF(position feature): 每个词和触发词、待分类候选事件元素词的两个相对距离向量构成
  - ▶ EF(event-type feature): 触发词对应的事件类型向量

[1] Chen, Yubo, et al. "Event extraction via dynamic multi-pooling convolutional neural networks." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015.

# 基于神经网络的事件抽取



- ▶ Dynamic Multi-pooling layer和PCNN中结构类似，根据**触发词位置**和**待分类候选事件元素词**将句子划分成三段，然后对卷积后的输出进行分段池化
- ▶ 第一阶段对触发词进行分类时，此处动态pooling仅根据触发词位置将句子划分成两段

[1] Chen, Yubo, et al. "Event extraction via dynamic multi-pooling convolutional neural networks." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015.

## ▶ 事件抽取

- ▶ 主要分为通过事件**触发词**对事件进行识别及分类、通过候选词对事件元素进行识别及**角色分类**两个部分。
- ▶ 与实体识别等任务较为相关，**实体抽取**效果的好坏将直接影响事件抽取的结果。
- ▶ 事件抽取不仅需要底层语言学知识，还需要更为深层的**语义和篇章级知识**。
- ▶ 事件结构远比实体关系三元组复杂，事件的**Schema结构**对事件抽取有很强的约束作用。

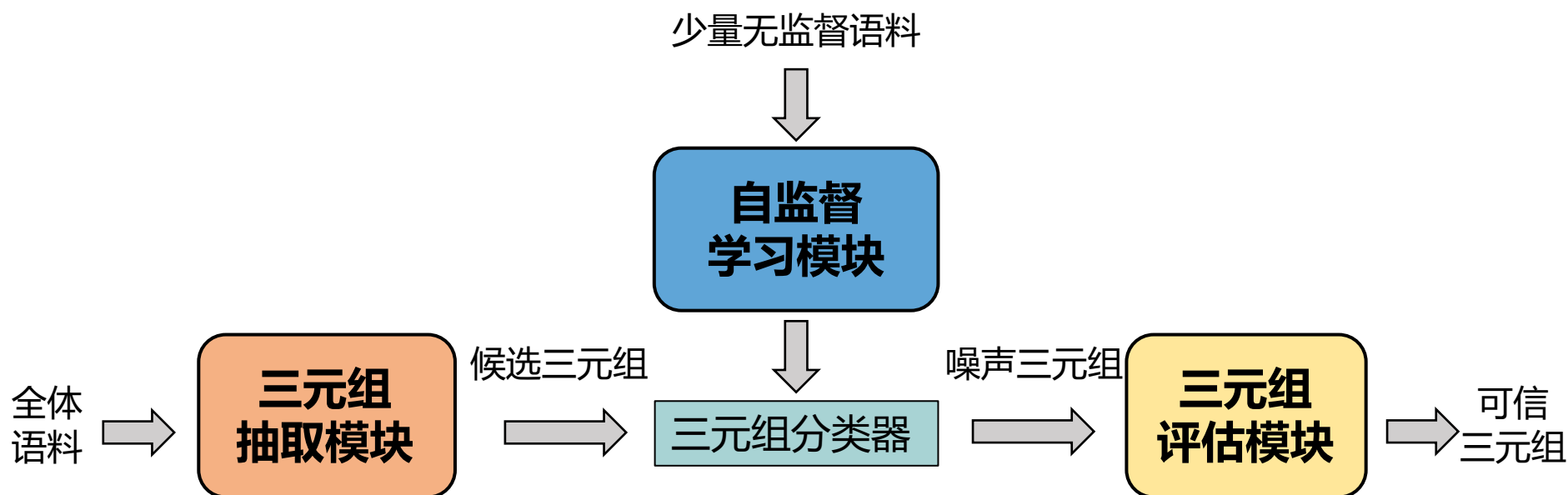
- ▶ 知识图谱的构建流程
- ▶ 实体识别
- ▶ 实体消歧
- ▶ 关系抽取
- ▶ 事件抽取
- ▶ **开放域知识抽取**
- ▶ 多模态知识抽取

- ▶ 大量开放语料库，例如
  - ▶ GigaWord (news texts)
  - ▶ PubMed (scientific articles)
  - ▶ World-Wide Web
  - ▶ Wikipedia-百度百科
- ▶ 知识不能事先确定或指定，大量新知识的发现和处理
  - ▶ Low Recall, Low Coverage
- ▶ 涉及开放域关系抽取、开放域事件抽取等多个任务，多采用无监督、半监督学习等方法。
- ▶ 对于很多开放域知识库的构建非常必要，如Cyc/Freebase等。



### ► OpenIE

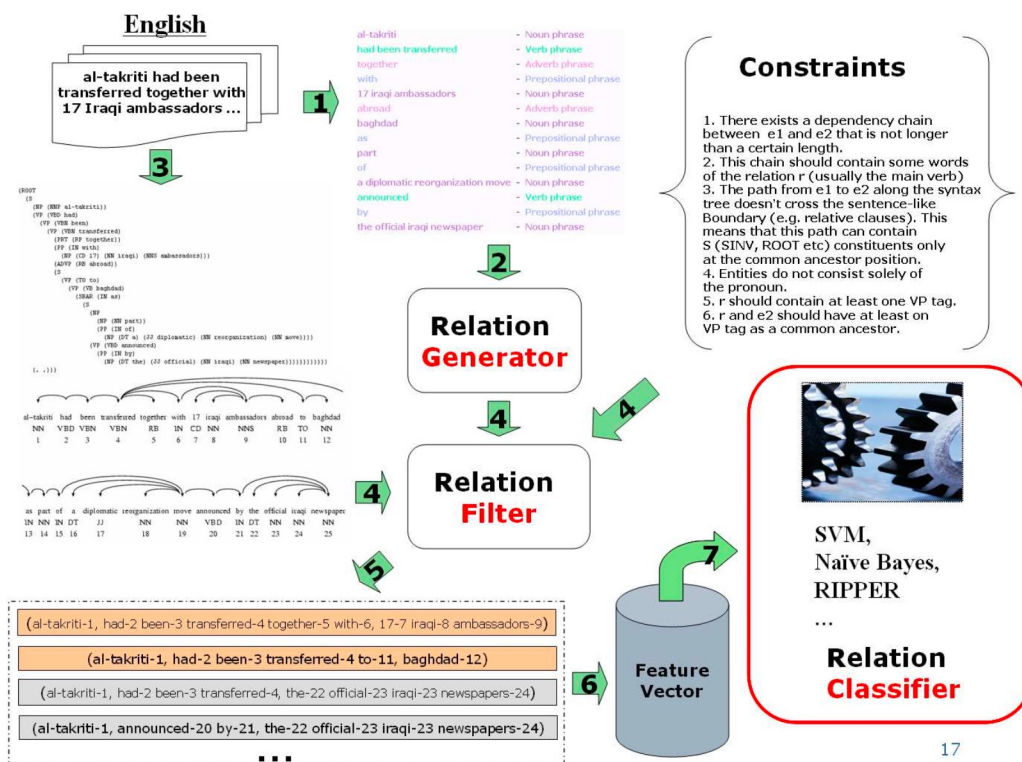
- 2007年由华盛顿大学构建的最早的开放领域实体关系抽取系统
- 主要包含自监督学习模块、三元组抽取模块、三元组评估模块三个模块



[1] Etzioni, Oren, et al. "Open information extraction from the web." Communications of the ACM 51.12 (2008): 68-74.

## ► 自监督学习模块

- 首先从语料中生成一些三元组，然后利用规则为三元组打上正确或错误标签，最后利用标注数据训练分类器（如朴素贝叶斯分类器等）





## ▶ Relation Generator

- ▶ 生成三元组( $e_1, r, e_2$ )
- ▶ 实体为名词短语构成
- ▶ 关系是句法结构树上连接两个实体的单词序列

## ▶ Relation Filter

- ▶ 将符合规则的三元组标记为正样例，不符合的标记为负样例
- ▶ 部分规则：

两个实体间的依存路径长度不能大于指定值。  
两个实体必须在同一个句子内。  
实体不能是代词。

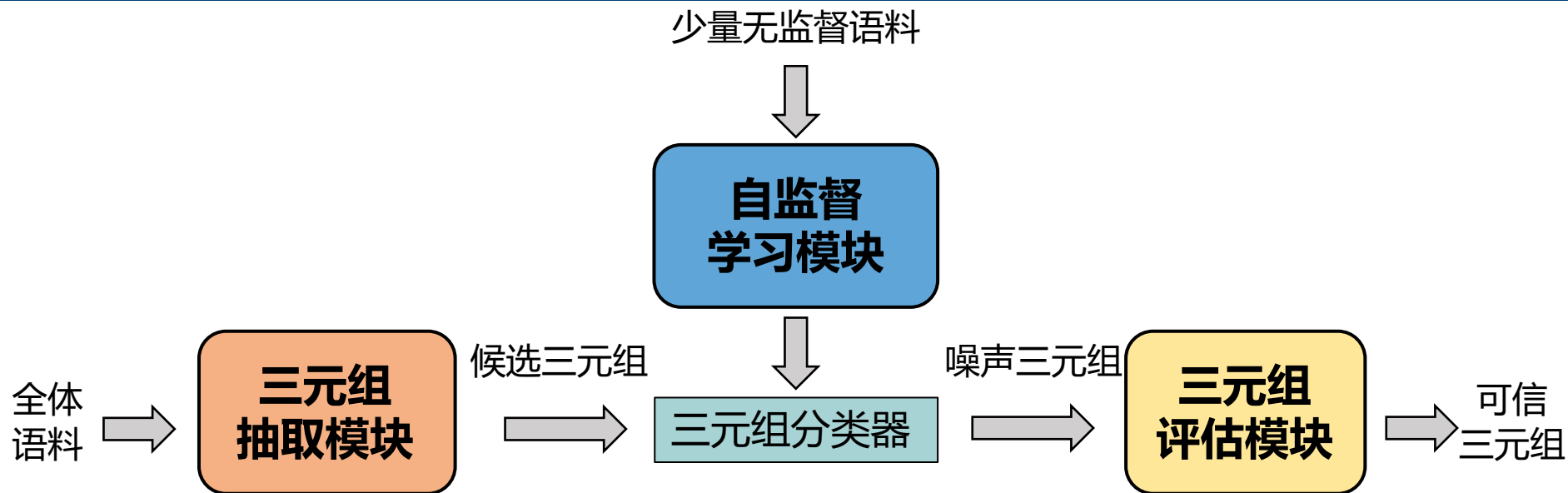
## ▶ Relation Generator

- ▶ 生成三元组( $e_1, r, e_2$ )
- ▶ 实体为名词短语构成
- ▶ 关系是句法结构树上连接两个实体的单词序列

## ▶ Relation Filter

- ▶ 将符合规则的三元组标记为正样例，不符合的标记为负样例
- ▶ 部分规则：

# 开放域关系抽取



## ▶ 三元组抽取模块

- ▶ 从全体语料中提取出所有可能的三元组，然后利用三元组分类器进行分类筛选

## ▶ 三元组评估模块

- ▶ 首先对相似的三元组进行合并，如两个实体相同且关系词的词根相同的三元组；然后根据网络数据的冗余性，计算合并三元组在网络文本中出现的次数，从而计算相应三元组的可信度

[1] Etzioni, Oren, et al. "Open information extraction from the web." Communications of the ACM 51.12 (2008): 68-74.

## ► 基于Bootstrapping的关系抽取

- 给定种子集合，如<姚明，叶莉>，进行下述步骤:

1. 从文档中抽取出包含种子实体的片段，如

- 姚明老婆叶莉简历曝光
- 姚明与妻子叶莉外出赴约
- 姚明携妻子叶莉赴约

提取出对应pattern:

- X 老婆 Y 简历曝光
- X 与妻子 Y 外出赴约
- X 携妻子 Y 赴约

2. 将抽取出的pattern去文档集中匹配

- 小猪与妻子伊万外出赴约

3. 根据pattern抽取出来的新文档加入种子库，迭代多轮直到不符合条件

## ► 特点

- 构建成本低，适合大规模构建
- 对初始种子集较为敏感，随着迭代次数增多容易出现语义漂移

[1] Etzioni, Oren, et al. "Open information extraction from the web." Communications of the ACM 51.12 (2008): 68-74.

- ▶ 知识图谱的构建流程
- ▶ 实体识别
- ▶ 实体消歧
- ▶ 关系抽取
- ▶ 事件抽取
- ▶ 开放域知识抽取
- ▶ **多模态知识抽取**

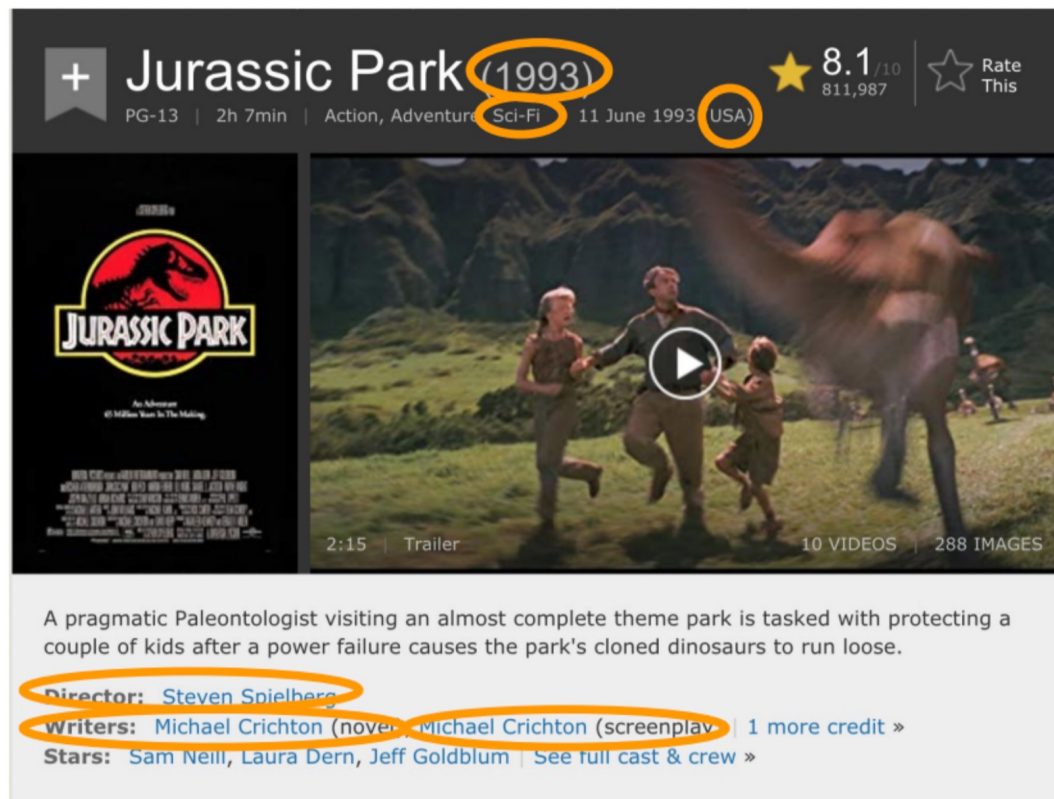
## ► 互联网上存在海量多模态数据

### *Jurassic Park* (film)

From Wikipedia, the free encyclopedia

*This article is about the 1993 film. For the franchise, see [Jurassic Park \(disambiguation\)](#).*

***Jurassic Park*** is a 1993 American science fiction adventure film directed by Steven Spielberg and produced by Kathleen Kennedy and Gerald R. Molen. It is the first installment in the *Jurassic Park* franchise, and is based on the 1990 novel of the same name by Michael Crichton and a screenplay written by Crichton and David Koepp. The film is set on the fictional island of Isla Nublar, located off Central America's Pacific Coast near Costa Rica. There, billionaire philanthropist John Hammond and a small team of genetic scientists have created a wildlife park of de-extinct dinosaurs. When



**Jurassic Park** (1993) ★ 8.1 /10 811,987 ☆ Rate This

PG-13 | 2h 7min | Action, Adventure, Sci-Fi | 11 June 1993 (USA)

2:15 | Trailer | 10 VIDEOS | 288 IMAGES

A pragmatic Paleontologist visiting an almost complete theme park is tasked with protecting a couple of kids after a power failure causes the park's cloned dinosaurs to run loose.

**Director:** Steven Spielberg

**Writers:** Michael Crichton (novel), Michael Crichton (screenplay) | 1 more credit »

**Stars:** Sam Neill, Laura Dern, Jeff Goldblum | See full cast & crew »

无结构化文本

VS.

半结构化页面

# 多模态知识抽取

## ▶ 互联网上存在海量多模态数据

<b>Directed by</b>	Steven Spielberg
<b>Produced by</b>	Kathleen Kennedy Gerald R. Molen
<b>Screenplay by</b>	Michael Crichton David Koepp
<b>Based on</b>	<i>Jurassic Park</i> by Michael Crichton
<b>Starring</b>	Sam Neill Laura Dern Jeff Goldblum Richard Attenborough Bob Peck Martin Ferrero

+

# Jurassic Park (1993)

PG-13 | 
 2h 7min | 
 Action, Adventure, Sci-Fi

11 June 1993 (USA)

★ 8.1<sup>/10</sup>  
811,987

☆ Rate This



An Adventure  
45 Million Years In The Making.



2:15 | Trailer

10 VIDEOS | 288 IMAGES

A pragmatic Paleontologist visiting an almost complete theme park is tasked with protecting a couple of kids after a power failure causes the park's cloned dinosaurs to run loose.

**Director:** Steven Spielberg

**Writers:** Michael Crichton (novel), Michael Crichton (screenplay) | [1 more credit »](#)

**Stars:** Sam Neill, Laura Dern, Jeff Goldblum | [See full cast & crew »](#)

# 半结构化页面 VS. 半结构化页面



► 互联网上存在海量多模态数据

surpass \$1 billion in ticket sales. The film won more than twenty awards, including three [Academy Awards](#) for its technical achievements in visual effects and sound design. *Jurassic Park* is considered a

Year ↕	Award ↕	Category ↕	Nominees ↕	Result ↕
1993	<a href="#">Bambi Awards</a> <sup>[154]</sup>	International Film	<i>Jurassic Park</i>	Won
	<a href="#">66th Academy Awards</a> <sup>[155]</sup>	<a href="#">Best Sound Editing</a>	<a href="#">Gary Rydstrom</a> and <a href="#">Richard Hymns</a>	Won
		<a href="#">Best Sound Mixing</a>	<a href="#">Gary Summers</a> , <a href="#">Gary Rydstrom</a> , <a href="#">Shawn Murphy</a> and <a href="#">Ron Judkins</a>	Won
		<a href="#">Best Visual Effects</a>	<a href="#">Dennis Muren</a> , <a href="#">Stan Winston</a> , <a href="#">Phil Tippett</a> and <a href="#">Michael Lantieri</a>	Won
		<a href="#">Best Director</a>	<a href="#">Steven Spielberg</a>	Won
	<a href="#">Saturn Awards</a> <sup>[147]</sup>	<a href="#">Best Science Fiction Film</a>	<i>Jurassic Park</i>	Won
		<a href="#">Best Special Effects</a>	<a href="#">Dennis Muren</a> , <a href="#">Stan Winston</a> , <a href="#">Phil Tippett</a> and <a href="#">Michael Lantieri</a>	Won
		<a href="#">Best Writing</a>	<a href="#">Michael Crichton</a> and <a href="#">David Koepp</a>	Won
		<a href="#">Best Actress</a>	<a href="#">Laura Dern</a>	Nominated
		<a href="#">Best Costumes</a>		Nominated
		<a href="#">Best Music</a>	<a href="#">John Williams</a>	Nominated
		<a href="#">Best Performance by a Young Actor</a>	<a href="#">Joseph Mazzello</a>	Nominated

无结构化文本      VS.      结构化表格

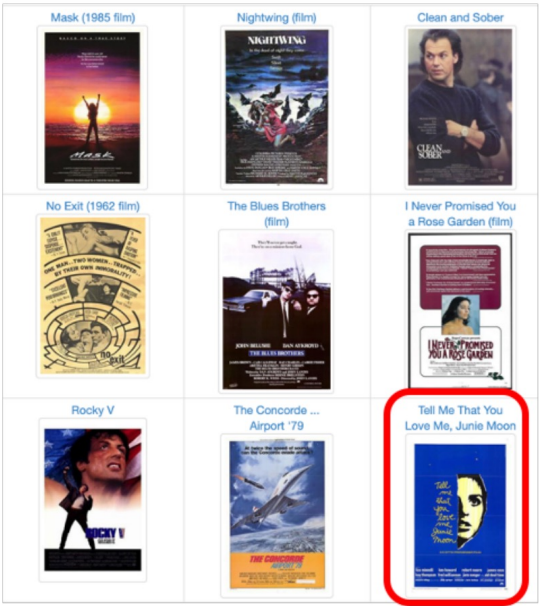
- ▶ 多模态知识抽取
  - ▶ 从多种不同模态结合的知识源中抽取出的知识(如实体三元组等)
  - ▶ 需要考虑不同模态的文本（无结构化文本、半结构化文本、结构化表格）
  - ▶ 需要考虑不同模态的信号输入（文本信息、视觉信息等）

## Multimodal Context

<a href="#">Steal This Movie!</a> The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".	<a href="#">La liceale</a> La liceale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini. ... Guida. It was followed by "La liceale nella classe dei ripetenti".
<a href="#">Sage Stallone</a> Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balboa Jr., the onscreen son of his father's title character. He did not, however, ... After that, he acted in lesser profile films.	<a href="#">Pierino contro tutti</a> Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an ... I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

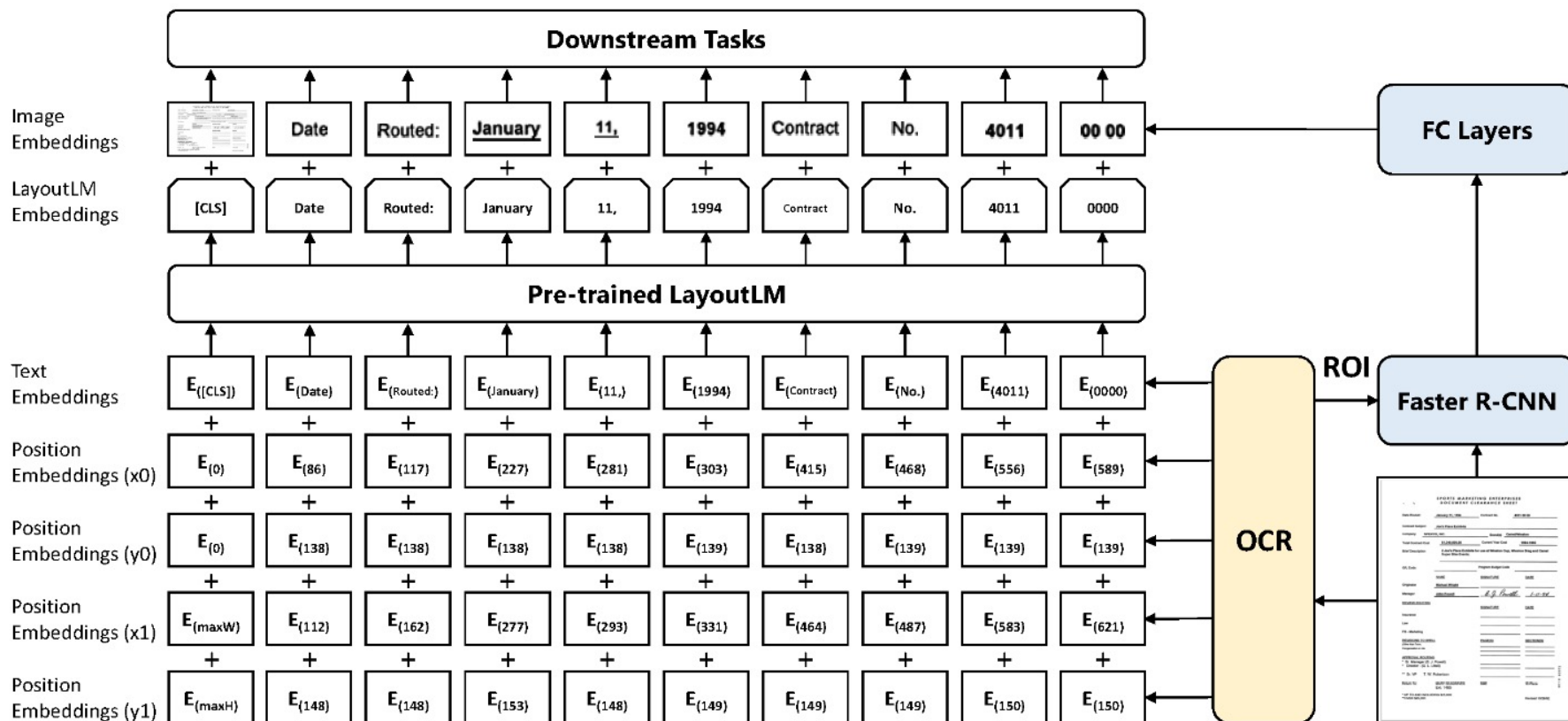
## Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	<a href="#">Tell Me That You Love Me, Junie Moon</a>	Jesse
1972	The Outside Man	Desk Clerk
...	...	...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck





## 通过预训练模型建模不同模态之间联系——LayoutLM



[1] LayoutLM: Pre-training of Text and Layout for Document Image Understanding Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, KDD 2020

- ▶ 知识图谱的构建流程
- ▶ 实体识别
- ▶ 实体消歧
- ▶ 关系抽取
- ▶ 事件抽取
- ▶ 开放域知识抽取
- ▶ 多模态知识抽取