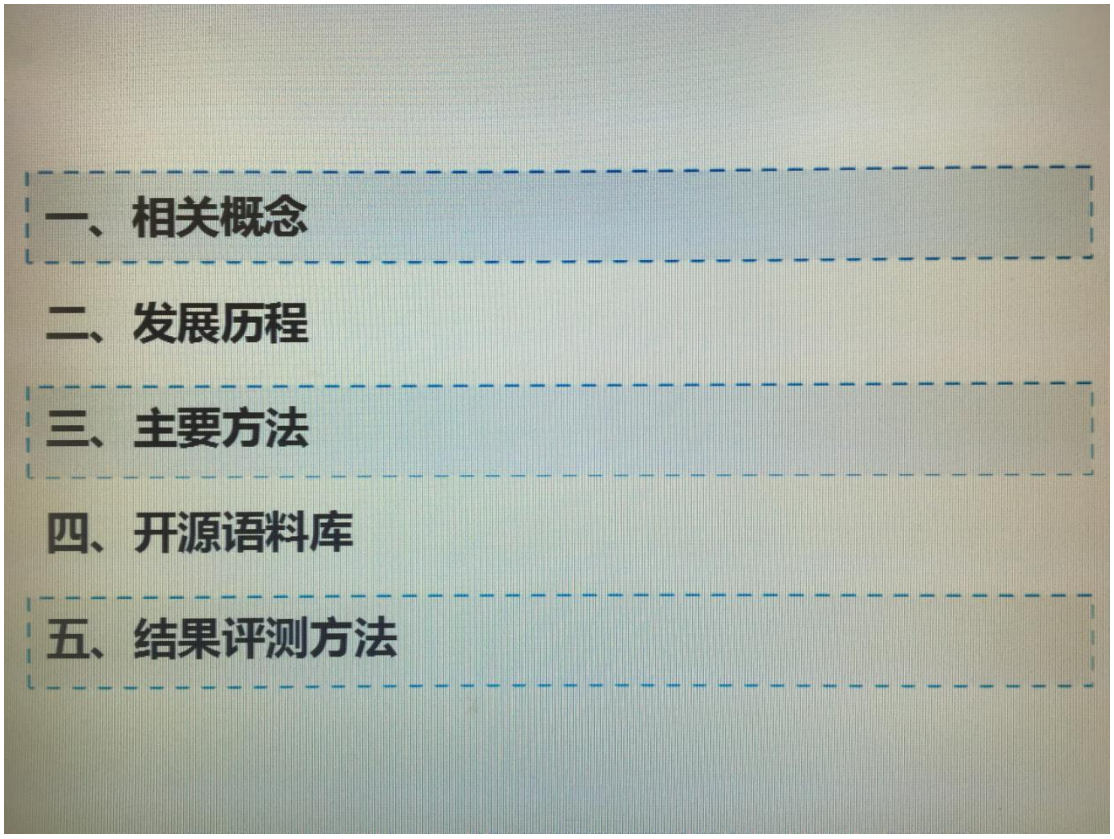
事件抽取技术方案简介

本文主要介绍事件抽取技术的相关概念和主要方法等方面内容，主要用于交流学习，可能有些地方理解不够深入，欢迎指正。文末附上了所看的参考文献，具体每篇文章所用的方法也进行了总结，由于篇幅问题，并未在此文中呈现，有需要可进一步联系我。



一、相关概念

- **事件抽取：**

  ACE2005将该项任务定义为：识别特定类型的事件，并进行相关信息的确定和抽取，主要的相关信息包括：事件的类型和子类型、事件元素角色等。

- **两大核心子任务：**

  （1）事件的检测和类型识别：触发词（event trigger）的抽取

  （2）事件元素的抽取：参数（event argument）的抽取

- **事件类型：**

  ACE中定义了8个大的事件类型和33个事件子类型（33+1个None类），每种事件类别对应着唯一的事件模板，每个事件有四个属性（模态、倾向性、普遍性、时态）。

- **事件元素：**

  实体：7种类型（人、组织、位置、地缘政治实体、设施，车辆、武器），每个还有子类型。

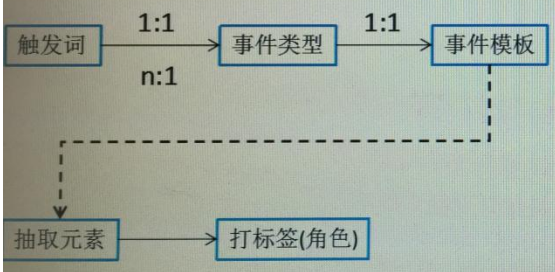
  时间、values集合(联系方式，数值，职称，犯罪类型，和句子类型)。

一、相关概念

基本处理过程：

1、事件类别识别

2、事件元素识别



例子：

“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（Life/Be-Born）事件模板中的三个元素标签，即：Person、Time以及Place。

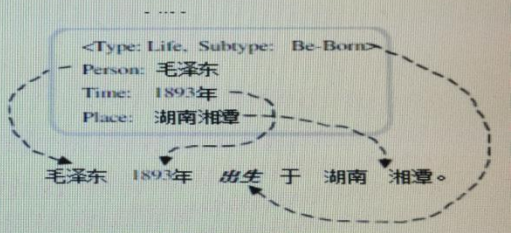


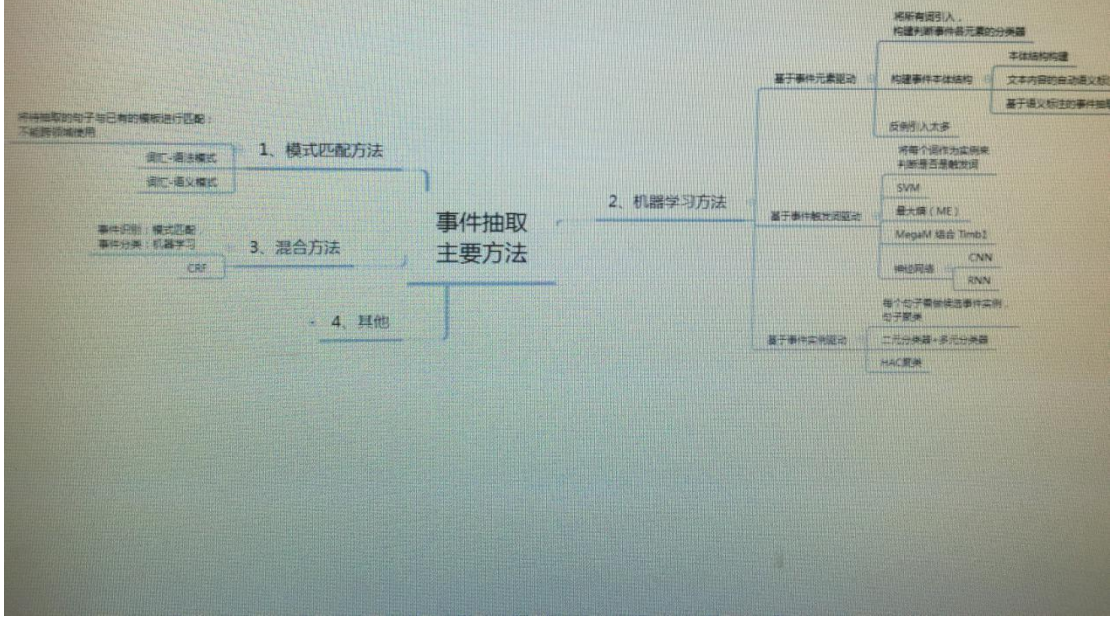
图1 “出生”事件的基本组成要素

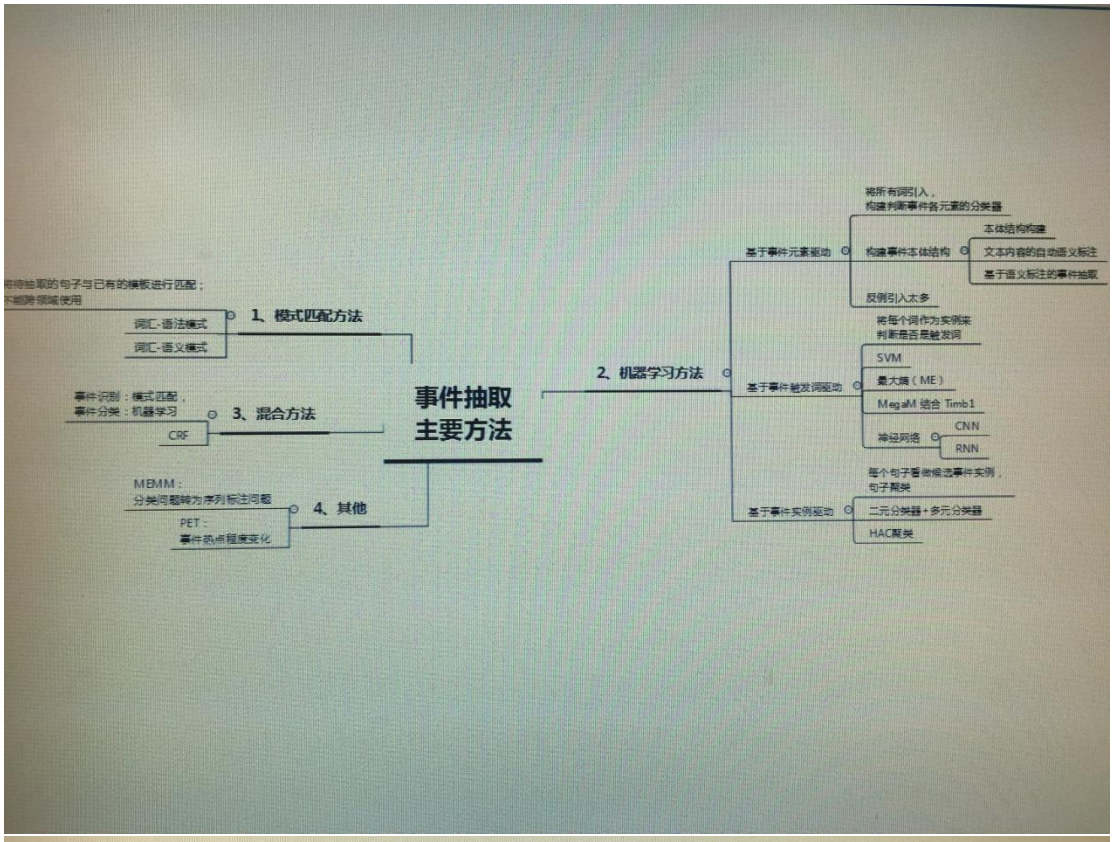
二、发展历程

20世纪80年代末开始蓬勃发展，两个比较重要的评测会议：MUC和ACE。

- MUC：1987-1998年，针对某一特定领域和场景，提供预先定义好的模板进行填充。
- ACE：2000年开始，不针对某个领域，不提供预先定义好的模板。自动抽取预料中出现的实体、关系和事件等内容。

三、主要方法





三、主要方法

多数系统是由机器学习的方法辅助基于模式的方法，以解决缺乏专家知识或应用的问题，从而提高提取性能。

	基于模式匹配	基于机器学习
概念	基于一定的模式（上下文约束环境），将待抽取的句子与已有的模板进行匹配。	将事件类别及事件元素的识别转换为分类问题。基于短语或句子层级的信息。
优点	针对特定领域能取得较高性能。	与领域无关，移植性好。
缺点	移植性较差。	需要大规模的标准语料（已标注），否则会有严重的数据稀疏。
核心步骤	抽取模式构建(机器学习自动获取模式)	重点：分类器和特征的选择 （1）事件类别的识别； （2）分类事件元素识别
其他	常见的事件抽取系统ExDisco、GenPAM、 FSA	半监督/无监督、特征选取、利用篇章级或跨篇章

三、主要方法—模式匹配

模式匹配：依赖于预先手工定义的模板（词汇-语义模式、词汇-语法模式）。

（1）词汇-语法模式：

将句子按照词汇-语法模式分割[26]；将词汇-语法模式应用在发现政治和金融领域之间大量关系和事件上[1]；采用词汇-语法模式来定义文本中参数结构[38]；将词汇-语法匹配与语义角色标签结合使用[13]；

（2）词汇-语义模式：

在模式识别中，概念的特定含义和关系缺乏时，或者是缺乏模式描述时（例如无法进行词汇-语法模式描述），可以采用词汇-语义模式。

考虑到生物领域的概念语义，构建概念识别器，用来抽取医疗事件[6]；应用词汇-语义模式构建一个事件识别框架[35]；自动化预警系统：使用词汇-语义模式进行概念匹配，使用词典（单词列表）增强依赖关系链，从而只要在相同的句子中发生传达句子中组成概念的语法相关链表达式就会匹配概念[3]。

四、开源语料库

- 触发词扩展：
  使用哈工大信息检索研究室的《同义词词林（扩展版）》自动扩充种子触发词。

Step 1:
For every seed trigger t in the "seed trigger-event type" table, find all its senses in *TongYiCi CiLin* (extension version)
Step 2:
If n or more than n words in a synset are contained in the "seed trigger-event type" table, and these words have the same event type, then all the words in the synset will be extended as triggers, and given the same event type of t. Here, n is called the expansion threshold.
Step 3:
Filter the extended triggers for multiple event types for the final "seed trigger-event type" binary pair table.

Fig. 4 Automatic trigger expansion algorithm

四、开源语料库

- MUC 会议和ACE 会议所提供的语料基本上是针对**通用领域**。
- 卡耐基梅隆大学标注了485 个电子板报构成的**学术报告**通知数据集。
- 北京语言大学标注了4 类**突发事件**（地震、火灾、中毒、恐怖袭击）文本。
- 哈工大社会计算与信息检索研究中心对**音乐领域**典型事件（举办演唱会、发行专辑）进行了标注。
- 哈工大社会计算与信息检索研究中心与浙江核新同花顺网络信息股份有限公司联合标注了**金融领域**信息抽取语料4000 句。

五、结果评测方法

对于事件提取的结果评测一般采用MUC会议的评测标准，包括三个指标：正确率P、召回率R和F值。其中，F值是P值和R值得调和平均，一般选取其作为衡量结果综合性能的指标。

1、正确率(Precision)
提取的信息中正确的信息占比，查准率。

2、召回率(Recall)
所有信息中正确信息的占比。查全率，召回目标类别的比例。

1. 正确率 = 提取出的正确信息条数 / 提取出的信息条数
2. 召回率 = 提取出的正确信息条数 / 样本中的信息条数

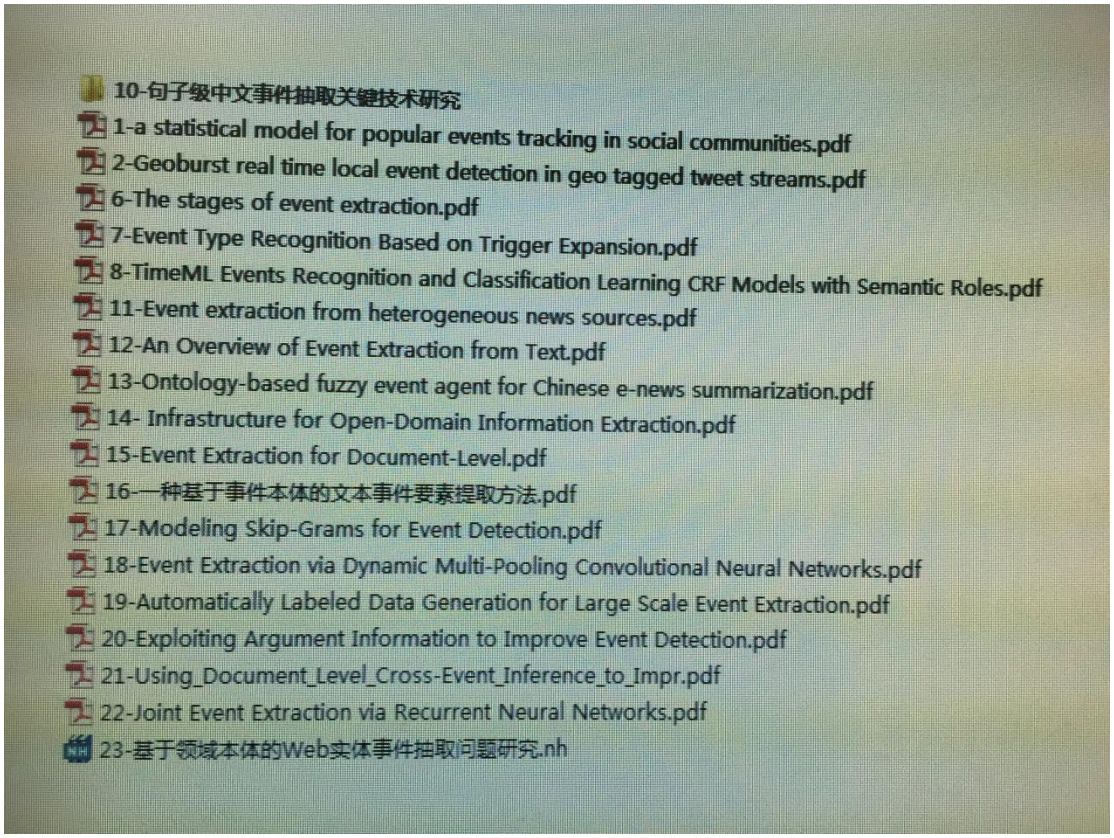
3、F值(F-score/F-Measure)
调和平均，综合反应整体的指标。

F-Measure是Precision和Recall加权调和平均：

$$F=\frac{(\alpha^2+1)P*R}{\alpha^2(P+R)}$$

当参数α=1时，就是最常见的F1，也即

$$F1=\frac{2*P*R}{P+R}$$


10-句子级中文事件抽取关键技术研究
1-a statistical model for popular events tracking in social communities.pdf
2-Geoburst real time local event detection in geo tagged tweet streams.pdf
6-The stages of event extraction.pdf
7-Event Type Recognition Based on Trigger Expansion.pdf
8-TimeML Events Recognition and Classification Learning CRF Models with Semantic Roles.pdf
11-Event extraction from heterogeneous news sources.pdf
12-An Overview of Event Extraction from Text.pdf
13-Ontology-based fuzzy event agent for Chinese e-news summarization.pdf
14- Infrastructure for Open-Domain Information Extraction.pdf
15-Event Extraction for Document-Level.pdf
16-一种基于事件本体的文本事件要素提取方法.pdf
17-Modeling Skip-Grams for Event Detection.pdf
18-Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks.pdf
19-Automatically Labeled Data Generation for Large Scale Event Extraction.pdf
20-Exploiting Argument Information to Improve Event Detection.pdf
21-Using_Document_Level_Cross-Event_Inference_to_Impr.pdf
22-Joint Event Extraction via Recurrent Neural Networks.pdf
23-基于领域本体的Web实体事件抽取问题研究.nh