第五章 事件知识学习

1. 任务定义、目标和研究意义

事件（Event）的概念起源于认知科学，广泛应用于哲学、语言学、计算机等领域[Quine, 1985;Trabasso, 1985;Zwaan, 1999;Chemero, 2000;Zacks, 2001;Glasbey, 2004;Fernando, 2007]。遗憾的是，目前学术界对此尚且没有公认的定义，针对不同领域的不同应用，不同学者对事件有不同的描述。在计算机科学的范畴内最常用的事件定义有如下两种：

- 第一种源自信息抽取领域，最具国际影响力的自动内容抽取评测会议（Automatic Content Extraction, ACE）对其定义为：事件是发生在某个特定时间点或时间段、某个特定地域范围内，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[Doddington et.al., 2004]。
- 第二种源自信息检索领域，事件被认为是细化的用于检索的主题。美国国防高级计划研究委员会主办的话题检测与追踪（Topic Detection and Tracking, TDT）评测指出：事件是由某些原因、条件引起，发生在特定时间、地点，涉及某些对象，并可能伴随某些必然结果的事情[Allan et.al., 1998a]。

虽然两种定义的应用场景和侧重点略有差异，但均认为事件是促使事物状态和关系改变的条件[Dong et.al., 2010]。目前已存在的知识资源（如维基百科等）所描述实体及实体间的关联关系大多是静态的，事件能描述粒度更大的、动态的、结构化的知识，是现有知识资源的重要补充。此外，很多认知科学家们认为人们是以事件为单位来体验和认识世界的，事件符合人类正常认知规律，如维特根斯坦在《逻辑哲学论》中论述到“世界是所有事实，而非事物的总和”[Ludwig, 2001]。因此，事件知识学习，即将非结构化文本中自然语言所表达的事件以结构化的形式呈现，对于知识表示、理解、计算和应用均意义重大。接下来，本文将沿着上述两种定义对事件知识学习的任务、挑战、研究现状和趋势进行梳理和展望。

1.1 任务定义

为了方便叙述，本文称针对第一种定义的相关研究为事件识别和抽取，针对第二种定义的相关研究为事件检测与追踪。

事件识别和抽取研究如何从描述事件信息的文本中识别并抽取出事件信息并以结构化的形式呈现出来，包括其发生的时间、地点、参与角色以及与之相关的动作或者状态的改变，核心的概念有：

- **事件描述（Event Mention）**：客观发生具体事件的自然语言描述，通常

是一个句子或者句群。同一事件可以有很多不同的事件描述，可能分布在同一文档的不同位置或不同的文档中。

- **事件触发词（Event Trigger）**：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，在 ACE 评测中事件触发词一般是动词或名词。
- **事件元素（Event Argument）**：事件的参与者，是组成事件的核心部分，与事件触发词构成了事件的整个框架。事件元素主要由实体、时间和属性值等表达完整语义的细粒度单位组成。
- **元素角色（Argument Role）**：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。
- **事件类型（Event Type）**：事件元素和触发词决定了事件的类别。很多评测和任务均制定了事件类别和相应模板，方便元素识别及角色判定。

事件检测与追踪旨在将文本新闻流按照其报道的事件进行组织，为传统媒体多种来源的新闻监控提供核心技术，以便让用户了解新闻及其发展。具体而言，事件发现与跟踪包括三个主要任务：分割，发现和跟踪，将新闻文本分解为事件，发现新的（不可预见的）事件，并跟踪以前报道事件的发展。事件发现任务又可细分为历史事件发现和在线事件发现两种形式，前者目标是从按时间排序的新闻文档中发现以前没有识别的事件，后者则是从实时新闻流中实时发现新的事件。

1.2 公开评测和数据集

1.2.1 事件识别和抽取

事件识别和抽取最早可以追溯到 20 世纪 70 年代耶鲁大学 Roger 等开展的故事理解相关研究，他们应用故事脚本理论从新闻报道中抽取工人罢工、地震等事件[Roger, 1978]。随着信息抽取技术的不断发展，事件识别和抽取也受到越来越多的关注，主要推动力是一系列相关国际评测会议的开展以及语料资源的丰富。

1.2.1.1 公开评测

消息理解会议（Message Understanding Conference, MUC）是公认最早的信息抽取评测会议[Chinchor and Marsh, 1998]，由美国国防高级研究计划委员会（Defense Advanced Research Projects Agency, DARPA）于 1987 年首次举办。MUC 要求从非结构化文本中抽取信息填入预定义模板中的槽，包括实体、实体属性、实体间关系、事件和充当事件角色的实体。从 1987 年到 1997 年的 7 届 MUC 评测，除了任务种类和槽值数量，还增加了模版嵌套、多语言抽取等。

前文提到的 ACE 评测会议由美国国家标准技术研究所（National Institute of Standards and Technology, NIST）从 1999 年 7 月开始准备，2000 年首次召开，到 2008 年共召开了八次，后被并入文本分析会议（Text Analysis Conference, TAC）。从 2004 年起，事件抽取成为 ACE 评测的主要任务，其中的事件是预定义类型的

句子级事件，每个事件都标注了事件触发词、事件类型、事件子类型、事件元素和元素角色信息，2007 年评测任务中增加了时间表达式的识别和归一化。ACE 吸引了很多学者参与并设计测评系统，现在很多流行的事件抽取工具都是针对 ACE 测评研发的，ACE 评测对事件识别和抽取技术的研究具有非常深远的影响。

知识库生成测评（Knowledge Base Population, KBP）隶属于 TAC 会议，主要研究从自然语言文本中抽取信息并链接到现有知识库的相关技术。事件抽取为 KBP 的一项重要任务，2014 年首次加入评测，目前已成功举办四届，2018 届正在筹备。KBP 中事件类型和实体角色遵从 ACE 2005 的定义，每类子事件类型都定义了各自的事件元素，并在此基础上极大丰富了事件抽取任务内容，主要包括：事件识别、事件消歧、事件元素抽取和链接、事件元素验证和链接。2016 年起，KBP 评测增加了以文档为单位的识别，语料也从英文扩展至汉语和西班牙语。

此外还有一些限定领域事件识别和抽取的公开评测，例如，东京大学组织的 BioNLP 是面向生物医学事件抽取的最权威的评测，2009 年到 2013 年共举办三届，目标是从生物医学文献中抽取出事件触发词、事件类型和事件元素等生物事件信息，其他类似评测不再赘述。

1.2.1.2 数据集

由于事件结构的复杂性和自然语言表达的灵活多样性，目前还没有形成统一的事件框架体系。另外，语料依赖人工标注，标注过程耗时、费力、成本高，因此事件类型较少、整体规模也不容易扩大。相关研究多是在各国际评测和公开语料的推动下展开，下面主要介绍 ACE 数据集、TimeBank 语料和中文事件语料库。

ACE 语料来源包括新闻专线、广播新闻、广播会话、网络日志、论坛数据及电话会话，美国宾夕法尼亚大学的语言数据组织（Linguistics Data Consortium, LDC）对源文本进行标注，生成的事件抽取数据集包含英文 599 篇，中文 633 篇。ACE 共定义了 8 大类 33 小类的事件。除基本的触发词、事件类型、事件子类型、事件元素和元素角色信息，ACE 还为每个事件标注了如下四种属性：

- 极性（Polarity），即肯定的事件和否定的事件。
- 时态（Tense），包括过去发生的事件，现在正在发生的事件，将来即将发生的事件以及无法确定时态的事件。
- 指属（Genericity），包括特指（Specific）事件和泛指（Generic） 事件。
- 形态（Modality），包括语气非常肯定的事件（Asserted Event）和信念事件（Believed Event）、假设事件（Hypothetical Event）和其它事件。

在此基础上，KBP 英文语料同样包含 599 篇标注文档，由人工过滤来确保每种事件类型都有多个实例，并且针对长句子进行了截断。KBP 2016 提供了 200 篇标注的英文文档、20 万词的中文文档以及 12 万词的西班牙文文档用于评测，但并未提供训练语料。

TimeBank 语料[Pustejovsky et.al., 2003]是由面向问答系统的时间和事件识别会议 (Time and Event Recognition for Question Answering Systems, TERQAS)提供，主要应用于识别和抽取事件的时间元素及事件之间的时序关系。语料包括来源于 DUC、ACE 和 PropBank 中的 300 篇新闻报道。TimeBank 不关心时间之外的事件元素，主要标注了事件、时间及其类型、时间信号词以及事件间的时序关系等。其中，事件通过事件类、事件时态以及事件状态三个属性来描述，事件类型分为 Occurrence、Perception、Reporting、Aspectual、State、Intensional State、Intensional Action 和 Modal 等 8 种。

中文事件语料库（Chinese Event Corpus, CEC）由上海大学语义智能实验室构建，旨在填补中文突发事件语料库的空白，包含 CEC-1 和 CEC-2 两个语料库[Meng, 2015]。CEC-1 是针对 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，但标注的粒度相对较大，尤其是对事件元素，且未对事件关系分类。为弥补上述不足，CEC-2 选取 333 篇关于地震、火灾、交通事故、恐怖袭击及食物中毒五类突发事件的互联网新闻报道作为待标注语料。标注过程中不仅覆盖了语料库中的所有事件，而且在中文句法分析和语义分析后进行标注，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。除标注语料，CEC-2 还保留了未标注的原始语料，其中记录了语料来源、标题、主体等信息。

1.2.2 事件检测与追踪

借鉴 MUC 的成功经验，DARPA 主办了 TDT 评测，旨在以事件的形式组织新闻报道，对其进行研究和评测[Allan, 2012]。话题（Topic）是 TDT 中的最基本的概念，起初与事件具有相同的含义，即指由某种原因引起的，发生在特定时间点或者时间段，在某个地域范围内，并可能导致某些必然结果的一个事件；后来演变为包括一个核心事件以及与之直接相关的事件的集合。

1998 年举行的首届 TDT 主要是针对中英文两种语料进行新闻报道切分、话题识别和话题追踪三项评测，第二届增加了新事件识别和报道关系识别评测。五个子任务均与事件检测追踪研究密切相关，新事件识别就是从给定的大量文档中识别出首次报道或者以前没有识别出来的事件。对新闻报道的切分就是将大量的文档聚成不同类别，涉及新事件识别、历史事件识别和报道关系识别。随后的历届 TDT 评测（2000 年—2004 年）都包含上述五个子任务，且将评测语言扩展到中文、英文和阿拉伯文。

LDC 为 TDT 系列评测提供了 TDT-pilot 和 TDT-2 到 TDT-5 五种语料。需要指出的是，TDT 语料标注方法与 ACE 等评测的标注完全不同：TDT-2 和 TDT-3 采用 YES、BRIEF 和 NO 三类标签分别表示当前报道内容与事件绝对相关、部分相关和不相关，TDT-4 和 TDT-5 则简化为 YES 和 NO 两种[Hong et.al., 2007]。

2. 研究内容和关键科学问题

事件知识学习是一项综合研究，需要比较深入的自然语言处理方法和技术作为支撑。相对于其他抽取和识别任务（如实体识别、关系抽取），事件识别和抽取更加复杂且富有挑战性，其难点主要表现在以下几个方面：

认知层面：事件具有复杂的内部结构。事件抽取不仅要识别出事件触发词和事件类别，还要识别出事件所涉及的所有元素并判断其在事件中扮演的角色。相较于实体和关系，事件涉及更多的实体和值，而且事件中各个元素间具有复杂关系和结构。因此需要对事件描述文本更深层次的理解。

语言层面：事件的表述是灵活的、具有歧义的。同一事件会有不同的描述和报道，例如“离开”既可以触发移动事件，也可以触发离职事件。同一事件的元素也可能会出现在多个句子、段落或者篇章中，一个句子或者一篇文章会描述多个不同但是相关或者不相关的事件。因此自然语言的灵活多变和歧义性对面向非结构化文本的事件抽取提出了很大的挑战。

方法层面：事件抽取会遇到错误累积的问题。事件抽取一般依赖于词法、句法分析等基本的自然语言处理工具，但实际中许多自然语言处理工具性能并不高，低性能的工具引入的错误会降低事件抽取系统的性能。

语料层面：标注语料规模小、数据稀疏。事件结构的复杂性和表述方式的歧义性导致人工标注事件的成本高、一致性差、耗时费力。因此，现有事件抽取相关数据集普遍规模较小，数据稀疏问题严重，对抽取的性能造成了很大的影响。

对于事件检测和追踪，虽然着眼点比事件识别和抽取要稍显宏观，但二者在认知、语言、方法和语料层面的挑战是高度统一的。

3. 技术方法和研究现状

考虑到事件识别和抽取、事件检测和追踪两个任务的处理对象、着眼点和技术路线的差异，本节分别对其主流的方法和现状进行梳理。

3.1 事件识别和抽取

根据抽取方法，事件抽取可以分为基于模识匹配的事件抽取和基于机器学习的事件抽取。接下来首先依此分类梳理事件抽取的国内外相关研究工作，然后对目前关注度较高的中文事件抽取相关研究进行介绍。

3.1.1 基于模式匹配的方法

基于模式匹配的方法是指对某种类型事件的识别和抽取是在一些模式的指导下进行的，模识匹配的过程就是事件识别和抽取的过程。采用模式匹配的方法进行事件抽取的过程一般可以分为两个步骤：模式获取和模式匹配。模式准确性是影响整个方法性能的重要因素，按照模式构建过程中所需训练数据的来源可细

分为基于人工标注语料的方法和弱监督的方法。

3.1.1.1 基于人工标注语料的方法

顾名思义，此类方法的模式获取完全基于人工标注的语料，学习效果高度依赖于人工标注质量。Ellen 等基于“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角色描述”两个假设开发的事件模式抽取系统 AutoSlog 就属于这个范畴[Riloff, 1993]。Kim 和 Moldovan 开发的 PALKA 是另一个典型代表[Kim and Moldovan, 1995]，他们假设“特定领域中高频出现的语言表示方式是可数的”，提出用语义框架和短语模式结构来表示特定领域中的模式，用语义树来表示语义框架、用短语链模型来表示短语模式。通过融入 WordNet 的语义信息，PALKA 在特定领域可取得接近纯人工抽取的效果。

3.1.1.2 弱监督的方法

这类方法不需要对语料进行完全标注，只需人工对语料进行一定的预分类或制定种子模式，由机器根据预分类语料或者种子模式自动进行模式学习。例如 Ellen 等研发的 AutoSlog 升级版 AutoSlog-TS 系统[Riloff and Shoen, 1995]就只需在人工预分类的语料上进行训练，可以解决标注标准不一致的问题，同时也降低了模式训练的准备工作量。欧洲委员会联合研究中心研发的 NEXUS 系统则使用无监督聚类的方式对语料进行预处理[Piskorski et.al., 2001; Tanev et.al., 2008]。

Yangarber 等研发的 ExDisco 通过匹配优质模式获取与待抽取事件相关的语料，利用人工制定的种子模式和经过一定预处理语料迭代来寻找新的匹配模式，省去了对语料进行人工标注或者预分类，只需提供少量的模式种子，大大减少了工作量[Yangarber et.al., 2000]。Chai 等则在其提出的模式抽取系统 TIMES 中引入了领域无关的概念层次知识库 WordNet，提升模式学习的泛化能力，并通过人工或者规则进行词义消歧，使最终的模式更加准确[Chai and Biermann, 1998]。GenPAM 系统在由特例生成泛化模式的学习过程中，有效利用模式间的相似性实现词义消歧，最大限度地减少了人工的工作量和对系统的干预[Jiang, 2005]。

总体而言，基于模式匹配的方法在特定领域中性能较好，知识表示简洁，便于理解和后续应用，但对于语言、领域和文档形式等均有不同程度的依赖，覆盖度和可移植性较差。

3.1.2 基于机器学习的方法

基于机器学习的方法建立在统计模型基础上，一般将事件抽取建模成多分类问题，因此研究的重点在于特征和分类器的选择。根据利用信息的不同可以分为基于特征、基于结构和基于神经网络三类主要方法。

3.1.2.1 基于特征的方法

基于特征的方法研究重点在于如何提取和集成具有区分性的特征，从而产生描述事件实例的各种局部和全局特征，作为特征向量输入分类器。该类方法多用于阶段性的管道抽取，即顺序执行事件触发词识别和元素抽取，从特征类型（或来源）上又可细分为利用句子级信息的方法和利用篇章级信息的方法。

句子级信息：Chieu 等首次将最大熵模型应用于事件抽取，使用了 unigram、bigram、命名实体等简单词法特征[Chieu and Ng, 2002]。Ahn 提出在事件抽取过程中同时使用 Timbl 和 MegaM 两种模型，并抽取候选词相关的词法特征、上下文特征、实体特征、句法特征和语言学特征，在事件触发词识别和元素抽取两个阶段都取得了不错的效果[Ahn, 2006]。

篇章级信息：Ji 等提出了跨文档事件抽取框架[Ji and Grishman, 2008]，其主要思想是对于一个句子级的抽取结果不仅要考虑当前的置信度，还要考虑与待抽取文本相关的文本对它的影响。具体实现时通过人工设置的 9 条推理规则定量地度量相关文本对当前抽取结果的影响，从而帮助修正原有的句子级事件抽取结果。该方法的优秀表现使得后来很多学者借鉴其利用篇章信息和背景知识的思想，相继出现了跨文本事件抽取的改进[Liao and Grishman, 2010]和跨实体事件抽取系统[Hong et.al., 2011]等。为了能更好地应用全局信息，Liu 等提出了利用全局信息（如事件的相关性）和更精确的局部信息（如实体类型）相结合的基于概率软逻辑推断的方法用于事件分类[Liu et.al., 2016a]。该方法首先利用局部信息做出初步分类，进而收集全局信息，学习事件和话题间，事件与事件间的共现信息，最后结合局部信息给出的初步分类和全局信息进行全局推理。

3.1.2.2 基于结构预测的方法

与基于特征适用的阶段性的管道抽取不同，基于结构的方法将事件结构看作依存树，抽取任务则相应地转化为依存树结构预测问题，触发词识别和元素抽取可以同时完成。例如，Li 等考虑到传统管道式事件抽取方法中多个步骤会导致错误传递以及忽略了事件触发词与事件元素之间的相互影响，首次提出基于结构感知机的联合模型同时完成事件触发词识别和事件元素识别两个子任务，并通过 beam search 缩小搜索解空间[Li et.al., 2013a]。为了利用更多的句子级信息，Li 等提出利用结构预测模型将实体、关系和事件进行联合抽取[Li et.al., 2014]。

3.1.2.3 基于神经网络的方法

上述两种方法在特征提取的过程中都依赖依存分析、词性标注、句法分析等传统的自然语言处理工具，容易造成误差累积，而且有很多语言没有自然语言处理工具。2015 年起，如何利用神经网络直接从文本中获取特征进而完成事件抽取成为研究热点。Chen 等[Chen et.al., 2015]和 Nguyen 等[Nguyen and Grishman, 2015]相继提出利用卷积神经网络模型（Conventional Neural Networks, CNN）抽

取特征来完成两阶段的识别任务。Feng 等提出利用循环神经网络（Recurrent Neural Networks, RNN）进行事件检测，取得了很好的性能，但没有探索循环神经网络在事件元素抽取阶段的效果[Feng et.al., 2016]。为了更好地考虑事件内部结构和各个元素间的关系，Nguyen 等将联合抽取模型与 RNN 相结合，利用带记忆的双向 RNN 抽取句子中的特征，并联合预测事件触发词和事件元素，进一步提升了抽取效果[Nguyen et.al., 2016]。

3.1.2.4 弱监督的方法

上述方法无一例外地需要大量的标注样本，而人工标注数据耗时费力、一致性差，尤其是面向海量异构的网络数据时，问题就更加明显。而无监督方法得到的事件信息没有规范的语义标签（事件类别，角色名称等），很难直接映射到现有的知识库中。因此，弱监督方法也是事件抽取中的一个重要分支。Chen 等提出利用部分高质量的标注语料训练分类器，然后利用初步训练好的分类器判断未标注的数据，选取高置信度的分类样本作为训练样本，通过迭代自动扩充训练样本[Chen and Ji, 2009]。Liao 等在相关文档中使用自训练的（Self-Training）的半监督学习方法扩展标注语料，并利用全局推理的方法考虑样例的多样性进而完成事件抽取；进一步提出同时针对词汇和句子两个粒度训练最大熵分类器，并用协同训练（Co-training）的方法扩展标注数据，进而对分类器进行更充分的训练[Liao and Grishman, 2011a; 2011b]。Liu 等利用 ACE 语料训练的分类器去判定 FrameNet 中句子的事件类别，再利用全局推断将 FrameNet 的语义框架和 ACE 中的事件类别进行映射，进而利用 FrameNet 中人工标注的事件样例扩展训练数据以提升事件检测性能。目前基于弱监督的事件抽取方法还处于起步阶段，亟需能自动生成大规模的、高质量的标注数据的方法提升事件抽取的性能[Liu et.al., 2016b]。

3.1.3 中文事件抽取

目前国内外事件抽取相关的研究大部分都是面向英文文本的英文事件抽取，面向中文文本的中文事件抽取工作才刚刚起步，主要面临技术和数据两方面的挑战。技术层面，中文的词句是意合的，词语间没有显式分隔符，而且中文实词在时态和形态上也没有明显变化，因此面向中文的事件抽取研究在基础自然语言处理层面具有天然的劣势。数据层面，由于起步较晚，缺乏统一的、公认的语料资源和相关评测，极大制约了中文事件抽取的研究。尽管如此，近些年中文事件抽取在公开评测、领域扩展和跨预料迁移方面也取得一定进展。

3.1.3.1 公开评测

基于 ACE 人工标注的中文数据集，国内学者探索了中文事件抽取的若干关键研究点。除了在模型方面的创新[Chen and Ng, 2012;Li et.al., 2012a;2013b]，在中文语言特性的利用方面，Li 等通过中文词语的形态结构、同义词等信息捕获更

多的未知触发词，进而解决中文事件抽取面临的分词错误和训练数据稀疏等问题；进一步细分中文事件触发词内部的组合语义（复合、附加和转化），进而提高系统的性能[Li et.al., 2012b]。Ding 等利用聚类的方法自动生成新事件类型的语料，在抽取过程中特别地考虑了待抽取文本的 HowNet 相似度[Ding et.al., 2013]。

3.1.3.2 领域扩展

除了公开评测，国内很多机构均面向实际应用展开特定领域的事件抽取研究，覆盖突发灾难、金融、军事、体育、音乐等多个领域。例如，Zhou 等针对金融领域事件中的收购、分红和贷款三个典型事件，提出自动构建抽取规则集的方法进行中文金融领域事件抽取 [Zhou, 2003]；Liang 等利用事件框架的归纳和继承特性实现对灾难事件的抽取[Liang and Wu, 2006]；其他不再一一详述。

3.1.3.3 跨语料迁移

由于目前中文事件抽取缺少公认语料，很多学者尝试利用现有大量的高质量英文标注语料辅助中文事件抽取。Chen 等首次提出该想法并利用跨语言协同训练的 Bootstrap 方法进行事件抽取[Chen and Ji, 2009]。Ji 提出基于中英文单语事件抽取系统和基于并行语料两种构建跨语言同义谓词集合的方法辅助进行中文事件抽取[Ji, 2009]，Zhu 等利用机器翻译同时扩大中文和英文训练语料，联合利用两种语料进行事件抽取[Zhu et.al., 2014]。Hsi 等联合利用符号特征和分布式特征的方法，利用英文事件语料提升中文事件抽取的性能[Hsi et.al., 2016]。

3.2 事件检测和追踪

事件检测和追踪研究的主流方法包括基于相似度聚类和基于概率统计两类。

3.2.1 相似度聚类法

基于相似度的方法首先需要定义相似度度量，而后基于此进行聚类或者分类。Yang 等提出在 TDT 中用向量空间模型（Vector Space Model, VSM）对文档进行表示，并提出了组平均聚类（Group Average Clustering, GAC）和单一通过法（Single Pass Algorithm, SPA）两种聚类算法[Yang et.al., 1998]。GAC 只适用于历史事件发现，它利用分治策略进行聚类。SPA 可以顺序处理文档并增量式产生聚类结果，能同时应用于历史事件发现和在线事件发现。在此基础上，Yang 等还提出利用衰减函数和时间窗口对事件聚类进行约束[Yang et.al., 1999]。Allan 等也在 TDT 的评测中尝试了 Single-Pass 的方法并取得较好的性能[Allan et.al., 1998b]。

后续研究中，有些工作尝试寻找新的距离度量方式，如 Hellinger 距离[Brants et.al., 2003]，但主要的还是尝试更改文档特征的提取方法来提升效果。如 Yang 等提出根据类别信息重新计算命名实体和非命名实体的权重[Yang et.al., 2002]。张等提出了一种基于 TF-IDF 的改进算法，利用 WordNet 中词汇关系对原文本向量进行补充 [Zhang, 2006]。Kumaran 等提出使用制定规则的方法为文档打分，复杂

度要小于分别计算文档向量，所以速度较快[Kumaran and Allan, 2004]。对于语法规则的规范文档的处理方面，上述算法已经可以取得比较不错的效果，但现实情况是很多文本是被大量普通用户创造，文本长短不一，且内容、格式和语法等方面均不规范，导致对这些不规范文档进行 TDT 十分有挑战性。Guille 等提出一种针对 Twitter 的事件检测与跟踪的算法，利用社交网站的评论和转发等特性尽可能的获取足够多的信息[Guille et.al., 2014]，类似的还有 Shamma 等[Shamma et.al., 2011]和 Benhardus 等[Benhardus and Kalita, 2013]的工作，均尝试解决社交平台上的短小不规范文档的 TDT 研究。

总体而言，基于相似度的模型用途广泛，计算速度通常比较快，但缺乏对于统计规律的利用。

3.2.2 概率统计法

概率统计方法通常使用生成模型，由于需要大量数据的支持，所以这种方法更加适用于历史事件检测。对比基于相似度聚类的模型，这类模型虽然复杂，但当数据量充足时，通常可以取得更好的准确率。基于概率的方法是目前 TDT 中的研究热点，主要分成两个方向，一是针对新闻等比较正式的规范文档，另一个则用于不规则或没有规律的非规范文档，下面分开阐述。

对新闻等规范文档，文中一般包含有完整的时间、地点、人物等信息，找出这些要素可以帮助建立新闻之间的关联。Li 等[Li et.al., 2012b]提出的生成模型将事件和文章均表示成<人物，地点，关键字，时间>，事件和文章的区别是文章的时间是时间点，而事件是时间段。人物，地点，关键字提取出来后，使用朴素贝叶斯的思想求出生成数据的分布。最近的相关研究是 Ge 等人提出的 BINets[Ge et.al., 2016a]。BINets 是一个边上带有权值的图，图中的每个结点代表一段时间反常出现的词，两个结点之间权重由统计规律得出，即两个单词一起反常出现的概率越高，它们对应结点之间的权重越高。利用 BINets，Ge 又提出了基于 BINets 进行聚类的方法[Ge et.al., 2016b]，利用 BINets 找到某特定事件的中心位置，再通过 BINets 的权值进行聚类。

不规范文档方面，算法经常是基于 LDA 等主题模型的变体建立文档间的联系，Blei 等对一些变体的特点进行了总结[Blei and Lafferty, 2006]。Griffith 提出了通过在一段时间窗口内的后验概率可以估计在这段时间内对事件的支持程度[Griffiths and Steyvers, 2004]。Hall 等提出通过计算事件在以年为单位的离散时间上的分布的后验概率来计算事件分布的强度变化趋势[Hall et.al., 2008]。Mei 等在其提出的基于主题模型的算法中充分利用了时间信息，提高了事件追踪的准确率[Mei and Zhai, 2005]。Hu 等引入突破点的概念（即事件突发或发生重大转折的时间点），通过新闻切分、分析、演化关系发现等步骤检测突破点，得出新闻事件的时序摘要[Hu et.al., 2011]。徐等基于 LDA 进行改进，通过对不同时间段分别建

模，可以分析该事件热度的变化并持续追踪该事件[Xu et.al., 2016]。

3.3 事件知识库构建

前文曾经提到，已有知识图谱，如 DBpedia，Yago 和 Wikidata 等均侧重于实体的客观属性及实体间的静态关联，缺乏结构化的事件数据。事件知识学习的最终目的就是从非结构化的文本数据中抽取结构化的事件表示，构建事件知识库弥补现有知识图谱的动态事件信息缺失问题。目前事件知识库构建的研究处于起步阶段，基础就是上述两方面研究，基于句子级的事件抽取和文档级的事件发现。

3.3.1 基于句子级的事件抽取

Wang 等构造了一种基于本体的新闻事件模型 NOEM，利用事件的类型、时间、空间、结构、因果、媒体六个方面特征描述新闻事件的 5WlH[3]语义要素[Wang and Zhao, 2012]。将抽取的关键事件语义要素自动扩充到本体后，可构成事件知识库，支持事件语义层次的应用。与现有事件模型的比较以及实际应用结果显示，NOEM 能够有效描述单个新闻文档中的关键事件、语义要素以及它们之间的关联，具有很强的形式化知识表达、应用集成和扩展能力。NewsReader 中，Rospocher 等提出一个以事件为中心的知识图谱表示，利用基于深度学习的 NLP 技术包括实体链接、语义角色标注等抽取不同语言的新闻当中的事件，并且将实体链接到已有的知识库 DBpedia 中，自动构建事件知识图谱[Rospocher et.al., 2016]。Tao 等提出了事件立方体（Eventcube）的概念，他们从新闻中抽取关键词（包括时间，人物，地点和事件等）构建词网络，基于该网络提出一种话题生成模型构建层次话题，每个话题对应词网络结构中的子图，并且支持多维度的搜索[Tao et.al., 2013]。Rouces 等构建了 FrameBase，提出了 N 维语义框架表示，是一种新的事件的表示方法，解决了传统 RDF 表示中三元组只能包含两个实体的局限。他们将句子看作实例，句子中的时间，事件类型，人物，地点等都是属于和实例关联的元素[Rouces et.al., 2015]。

3.3.2 基于文档级的事件发现

Event Registry[Rupnik et.al., 2016]从多语言的新闻文档中抽取事件，将相似的新闻文档聚类，从每个类中获取一个宏观的事件，抽取事件中的人物、时间、地点等要素。类似的研究工作还包括：Kuzey 等将每个新闻文档看成一个节点，并通过新闻之间的相似度建立节点之间的边，形成图的结构，基于图对新闻文档进行聚类，每个类作为一个事件，同时得到事件之间的时序和层次关系[Kuzey et.al., 2014]；Hoxha 等利用基于文档的词袋表示，对新闻按照所描述的事件进行聚类，识别新闻中的实体并和已有的知识库进行实体链接，形成事件知识库[Hoxha et.al., 2016]。NewsMiner 通过 LDA 模型将新闻按照事件组织，并分析新

[3] Who（何人）、When（何时）、Where（何地）、What（何事）、Why（何因）、How（何种方式）

闻和评论之间的联系，在对事件，话题，以及实体之间的关系深入分析的基础上提供新闻多刻面搜索。随着相似事件的不断重复发生，事件知识可以通过增量学习得到积累完善[Hou et.al., 2015a]。在相似事件增量学习中，Hu 等将已有相似事件作为先验指导新事件知识学习并保证新事件的相对独立性，提出基于先验的狄利克雷过程混合模型，模型鼓励但不强制先验知识相关话题的出现，且允许新话题的出现[Hu et.al., 2015a;2015b]。除了新闻报道，维基百科页面的目录表蕴含了层次话题，Hu 等针对不同事件维基页面的层次话题的多样性，提出基于概率的贝叶斯网络结构学习方法，将事件维基页面的话题结构化信息和文本描述都以概率的方式建模为网络图的边上的权重[Hu et.al., 2015c]。

4. 技术展望与发展趋势

4.1 事件识别和抽取的发展趋势

通过 3.1 节的综述可以发现，事件抽取在 2002 年前基本会被形式化为模式发现和匹配，2002 年至 2013 年间，基于机器学习方法成为了主流，极大地提高了准确度并且降低了邻域迁移成本。2013 年以来，随着神经网络在图像领域取得的巨大成功，越来越多的研究者开始转向基于神经网络的事件抽取，为事件抽取任务的提升，特别是预定义的从非结构化文本中进行事件抽取任务的提升带来了新的契机。

分步抽取到联合抽取：事件抽取的目标往往是很多样的，通常均会将任务拆分为几个步骤完成，最普遍的分解方式是 ACE 在 2005 年测评中定义的事件触发词识别、事件触发词分类、事件元素识别和事件元素分类四个阶段。近年来，更多工尝试将四个传统过程整合成更少的步骤，如前文提到的 Chen 和 Nguyen 的工作[Nguyen et.al., 2016; Chen and Ng, 2012]。从更高层面上讲，其他信息抽取任务（如实体抽取、关系抽取）也可以和事件抽取进行联合学习，在之后的研究过程中，联合抽取以避免分步噪音积累的思路一定会更加普遍。

局部信息到全局信息：事件抽取研究初期更多的考虑是当前词自身的特征，但研究者逐渐开始利用不同词之间的联系，从而获取更多的全局信息来完成事件抽取任务，例如 Li 等提出利用整数线性规划的方法联合抽取方法和为解决中文事件抽取中的成员缺失问题而提出联合利用句子、上下文和相关文档的中的相关事件和共指事件信息的事件抽取方法[Li et.al., 2013b]。此外还有前文提到的 Ji 等 2008 年首次提出的跨文档事件抽取中借助篇章信息和背景知识的思想[Ji and Grishman, 2008]。可以看出事件抽取考虑的信息越来越多样化和全局化。

人工标注到半自动生成语料：目前的语料多是英文语料，中文和其他语言的语料非常稀少。且由于事件本身的复杂程度，人工标注大量的语料十分困难。因此，越来越多的学者开始思考如何利用现有的语料迭代生成更多语料。目前主流

的解决思路是利用英文语料辅助另一种语言语料的生成，做跨语言迁移学习。另一种可能的解决思路是借鉴外部知识来自动扩展语料，例如 Chen 等研究如何基于世界知识和语言学知识大规模自动生成事件语料[Chen and Ji, 2009]。不管是哪种途径，事件抽取肯定会向如何减少人工参与即可取得良好效果的方向发展。

4.2 事件检测和追踪的发展趋势

事件检测和追踪方面，基于 LDA 等主题模型的研究逐渐成为主流，相关研究的主要发展趋势包括两个方向：一是非参数化，放宽对话题数目的限制；二是多数据流共同建模，有效利用不同数据间的互补信息。

非参数化：Ahmed 等人在动态话题模型基础上，利用层次化狄利克雷过程（Hierarchical Dirichlet Process, HDP）放宽对话题数目的限制，提出无限动态话题模型（iDTM），该模型理论上允许同一个时间片内生成无限个新闻话题，方便对事件发展过程中话题产生、话题重要程度变化以及话题消亡等情况的建模[Ahmed and Xing, 2010]。Cui 等和 Gao 等在 iDTM 基础上增加了两种话题演化行为：分裂和融合，即一个话题可以分裂成多个子话题，多个话题也可能合并为一个大话题，通过对模型的调整使其能够在保证原有功能基础上对上述两种行为进行描述，最后根据话题的生命周期以及周期内强度变化生成时序结构化摘要，并提供可视化展示[Cui et.al., 2011; Gao et.al., 2011] 。

多流交互：Hong 等扩展 LDA 算法，同时对多个社交媒体上的媒体流，例如 Twitter 和 Yahoo，进行事件检测，其思路是利用 LDA 分别在多个数据流上检测主题，再利用主题联系各个数据流[Hong et.al., 2011]。Wang 等将不同新闻媒体流定义成协同文本流（Coordinated Text Streams），建模的过程中考虑流间的相互增益，通过估计给定时间点话题的出现概率来检测突发话题，最后以不同流内共同的突发话题为关键点将新闻流在时间线上对齐[Wang et.al., 2007]。Wang 等将异步文本流的对齐和话题抽取放到同一框架中，模型通过引入一个自增强过程，有效地利用异步文本的语义关联性对齐和时间信息的关联性，将时序对齐和话题建模整合为统一目标函数[Wang et.al., 2009]。除了新闻和用户生成内容的文本信息，Lin 等进一步考虑了用户间关系（即社区信息）。他们将新闻的传播以及社区的形成（即用户关系的建立）形式化为一个联合推理问题，分别使用混合话题模型和高斯马尔可夫随机场（Gaussian Markov Random Field）刻画用户生成内容的产生和用户间影响力的变化对问题进行建模求解[Lin et.al., 2011]。针对新闻和社交媒体间跨数据流的相互依赖，Hou 等提出特定事件内新闻和用户生成内容相互影响分析问题，并分别利用基于话题距离[Hou et.al., 2015b]和格兰杰因果测试的影响发现方法[Hou et.al., 2016]结合新闻传播学、语言学和社会认知学对新闻和用户生成内容间的相互影响进行量化分析。

参考文献

[Ahmed and Xing, 2010] Ahmed A, Xing E P. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA: AUAI Press, 2010. 20–29.

[Ahn, 2006] Ahn D. The stages of event extraction. [J]. In Proceedings of the Workshop on Annotating and Reasoning About Time and Events, 2006: 1–8.

[Allan, 2012] Allan J. Topic detection and tracking: event-based information organization[J]. Springer Science and Business Media, 2012, 12.

[Allan et.al., 1998a] Allan J, Carbonell J G, Doddington G, et al. Topic detection and tracking pilot study final report[M]. 1998.

[Allan et.al., 1998b] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking. [J]. In Proceedings of the 21st Annual International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval, 1998: 37–45.

[Benhardus and Kalita, 2013] J. Benhardus and J. Kalita, “Streaming trend detection in twitter,” IJWBC, vol. 9, no. 1, pp. 122–139, 2013.

[Blei and Lafferty, 2006] D.M. Blei, J.D. Lafferty. Dynamic Topic Model [C]. Proceedings of the International Conference on Machine Learning, 2006:113-120

[Brants et.al., 2003] Brants T, Chen F, Farahat A. A system for new event detection[C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 330-337

[Chai and Biermann, 1998] Chai J Y, Biermann. AW. Learning and generalization in the creation of information extraction systems. [J]. Citeseer, 1998.

[Chemero, 2000] Chemero A. What events are. [J]. Ecological Psychology, 2000, 12(1): 37–42.

[Chen and Ji, 2009] Chen Z, Ji H. Can one language bootstrap the other: a case study on event extraction. [J]. In Proceedings of the North American Chapter of the Association for Computational Linguistics 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, 2009: 66–74.

[Chen and Ji, 2009] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction[J]. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009: 209–212.

[Chen and Ng, 2012] Chen C, Ng V. Joint modeling for Chinese event extraction with rich linguistic features[J]. In Proceedings of International Conference on Computational Linguistics, 2012: 529–544.

[Chen et.al., 2015] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In ACL-IJCNLP.

[Chieu and Ng, 2002] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text. [J]. In Eighteenth National Conference on Artificial Intelligence, 2002:786–791.

[Chinchor and Marsh, 1998] Chinchor N, Marsh E. Muc-7 information extraction task definition[J]. In Proceeding of the seventh message understanding conference, Appendices, 1998: 359–367.

[Cui et.al., 2011] Cui W, Liu S, Tan L, et al. Textflow: Towards better understanding of evolving topics in text. Visualization and Computer Graphics, IEEE Transactions on, 2011, 17(12):2412–2421.

[Ding et.al., 2013] Ding X, Qin B, Liu T. Building Chinese event type paradigm based on trigger clustering. [J]. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013: 311–319.

[Doddington et.al., 2004] Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ace) program-tasks, data, and evaluation. [J]. In Proceedings of the International Conference on Language Resources and Evaluation, 2004, 2: 1.

[Dong et.al., 2010] Dong Z, Dong Q, Hao C. Hownet and its computation of meaning[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010: 53-56.

[Feng et.al., 2016] Feng X, Huang L, Tang D, et al. A language-independent neural network for event detection[J]. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 66.

[Fernando, 2007] Fernando T. Observing events and situations in time. [J]. Linguistics and Philosophy, 2007, 30(5): 527–550.

[Gao et.al., 2011] Gao Z J, Song Y, Liu S, et al. Tracking and connecting topics via incremental hierarchical dirichlet processes. Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, BC, Canada, 2011. IEEE. 1056–1061.

[Ge et.al., 2016a] Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, Ming Zhou: Event Detection with Burst Information Networks. In COLING 2016.

[Ge et.al., 2016b] Tao Ge, Lei Cui, Baobao Chang, Sujian Li, Ming Zhou, Zhifang Sui: News Stream Summarization using Burst Information Networks. In EMNLP 2016.

[Glasbey, 2004] Glasbey S. Event structure, punctuality, and when. [J]. Natural language semantics, 2004, 12(2): 191–211.

[Griffiths and Steyvers, 2004] T.L. Griffiths, M Steyvers. Finding scientific topics [C] Proceedings of the National Academic Science USA, 2004, 101(1):5228-5235

[Guille et.al., 2014] Guille A, Favre C. Mention-anomaly-based event detection and tracking in twitter[C]//Advances in Social Networks Analysis and Mining, 2014 IEEE/ACM International Conference on. IEEE, 2014: 375-382.

[Hall et.al., 2008] D. Hall, D Jurafsky, C.D. Manning. Studying the History of Ideas Using Topic Models [C]. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008:363-371

[Hong et.al., 2007] Hong Y, Zhang Y, Liu T, et al. Topic detection and tracking review[J]. Journal of Chinese information processing, 2007, 6(21): 77–79.

[Hong et.al., 2011] Hong L, Dom B, Gurumurthy S, et al. A time-dependent topic model for multiple text streams. Proceedings of the 17th ACM International Conference on Knowledge Discovery in Data Mining, California, USA, 2011. 832–840.

[Hong et.al., 2011] Hong Y, Zhang J, Ma B, et al. Using cross-entity inference to improve event extraction[J]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, 1: 1127–1136.

[Hou et.al., 2015a] Hou L, Li J, Wang Z, et al. Newsminer: multifaceted news analysis for event search[J]. Knowledge-Based Systems, 2015, 76: 17-29.

[Hou et.al., 2015b] Hou L, Li J, Li X L, et al. Measuring the influence from user-generated content to news via cross-dependence topic modeling[C]//International Conference on Database Systems for Advanced Applications. Springer International Publishing, 2015: 125-141.

[Hou et.al., 2016] Hou L, Li J, Li X L, et al. Detecting Public Influence on News Using Topic-Aware Dynamic Granger Test[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, 2016: 331-346.

[Hoxha et.al., 2016] Hoxha K, Baxhaku A, Ninka I. Bootstrapping an Online News Knowledge Base. International Conference on Web Engineering. Springer International Publishing, 2016: 501-506.

[Hsi et.al., 2016] Hsi A, Yang Y, Carbonell J, et al. Leveraging multilingual training for limited resource event extraction. [J]. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 1201–1210.

[Hu et.al., 2011] Hu P, Huang M, Xu P, et al. Generating breakpoint-based timeline overview for news topic retrospection. Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, BC, Canada: IEEE, 2011. 260–269.

[Hu et.al., 2015a] Linmei Hu, Juanzi Li, Xiaoli Li, Chao Shao, Xuzhong Wang. TSDPMM: Incorporating Prior Topic Knowledge into Dirichlet Process Mixture Models for Text Clustering. EMNLP 2015: 787-792.

[Hu et.al., 2015b] Linmei Hu, Chao Shao, Juanzi Li, Heng Ji. Incremental learning from news events. Knowledge-Based System. 89: 618-626 (2015).

[Hu et.al., 2015c] Linmei Hu, Xuzhong Wang, Mengdi Zhang, Juan-Zi Li, Xiaoli Li, Chao Shao, Jie Tang, Yongbin Liu. Learning Topic Hierarchies forWikipedia Categories. ACL (2) 2015: 346-351.

[Ji, 2009] Ji H. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning[J]. In Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, 2009: 27–35.

[Ji and Grishman, 2008] Ji H, Grishman R. Refining event extraction through unsupervised Cross-document inference[C]. Proceedings of ACL-08 . HLT Columbus, USA. HLT, 2008: 254-262.

[Jiang, 2005] Jifa J. An event ie pattern acquisition method. [J]. Computer Engineering, 2005, 31(15): 96–98.

[Kim and Moldovan, 1995] Kim J, Moldovan D I. Acquisition of linguistic patterns for knowledge-based information extraction. [J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(5): 713–724.

[Kumaran and Allan, 2004] Kumaran G, and Allan J. Text Classification and Named Entities for New Event Detection. In Proceedings of the 27th Annual International ACM SIGIR Conference, New York, NY, USA. ACM Press, 297–304. 2004.

[Kuzey et.al., 2014] E. Kuzey, J. Vreeken, G. Weikum, a fresh look on knowledge bases: Distilling named events from news, In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, 2014, pp. 1689–1698.

[Liang and Wu, 2006] Liang H, Wu P. Information extraction system based on event frame. [J]. Journal of Chinese Information Processing, 2006, 20(2): 40–46.

[Liao and Grishman, 2010] Liao S, Grishman R. Using document level cross-event inference to improve event extraction. [J]. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 789–797.

[Liao and Grishman, 2011a] Liao S, Grishman R. Can document selection help semi-supervised learning? a case study on event extraction[J]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, 2: 260–265.

[Liao and Grishman, 2011b] Liao S, Grishman R. Using prediction from sentential scope to build a pseudo co-testing learner for event extraction. [J]. In Proceedings of 5th International Joint Conference on Natural Language Processing, 2011: 714–722.

[Li et.al., 2012a] Li P, Zhu Q, Diao H, et al. Joint modeling of trigger identification and event type determination in Chinese event extraction. [J]. In Proceedings of the International Conference on Computational Linguistics, 2012: 1635–1652.

[Li et.al., 2012b] Li P, Zhou G, Zhu Q, et al. Employing compositional semantics and discourse consistency in Chinese event extraction[J]. In Proceedings of the 2012

Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 1006–1016.

[Li et.al., 2013a] Li P, Zhu Q, Zhou G. Joint modeling of argument identification and role determination in Chinese event extraction with discourse-level information. [J]. In Proceedings of International Joint Conference on Artificial Intelligence, 2013: 612–622.

[Li et.al., 2013b] Li P, Zhu Q, Zhou G. Argument inference from relevant event mentions in Chinese argument extraction[J]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013: 1477–1487.

[Li et.al., 2014] Li Q, Ji H, Hong Y, et al. Constructing information networks using one single model[J]. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1846–1851.

[Lin et.al., 2011] Lin C X, Mei Q, Han J, et al. The joint inference of topic diffusion and evolution in social communities. Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, BC, Canada, 2011. IEEE. 378–387.

[Liu et.al., 2016a] Shulin Liu, Kang Liu and Jun Zhao, A Probabilistic Soft Logic Based Approach to Exploit Latent and Global Information in Event Classification, AAAI 2016

[Liu et.al., 2016b] Liu S, Chen Y, He S, et al. Leveraging framenet to improve automatic event detection[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, 1: 2134-2143.

[Ludwig, 2001] Ludwig Wittgenstein: Tractatus logico-philosophicus[M]. Oldenbourg Verlag, 2001.

[Mei and Zhai, 2005] Mei Q, Zhai C. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. Proceedings of the 11th ACM International Conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, 2005. ACM. 198–207.

[Meng, 2015] Meng J. Research on event extraction technology in the field of unexpected events. [J]. Master's thesis, Shanghai University, 2015.

[Nguyen and Grishman, 2015] Nguyen H T, Grishman. R. Event detection and domain adaptation with convolutional neural networks[J]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 365–371.

[Nguyen et.al., 2016] Nguyen T H, Cho K, Grishman R. Joint event extraction via recurrent neural networks. [J]. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 300–309.

[Piskorski et.al., 2001] Piskorski J, Tanev H, Atkinson M, et al. Online news event extraction for global crisis surveillance [J]. In Transactions on computational collective intelligence V, 2001: 182–212.

[Pustejovsky et.al., 2003] Pustejovsky J, Hanks P, Sauri R, et al. The timebank corpus. [J]. In Proceedings of the Corpus linguistics, 2003: 40–49.

[Quine, 1985] Quine W V O. Events and reification. [J]. Actions and events: Perspectives on the philosophy of Donald Davidson, 1985: 162–171.

[Riloff, 1993] Riloff E. Automatically constructing a dictionary for information extraction tasks[C]//AAAI. 1993: 811-816.

[Riloff and Shoen, 1995] Riloff E, Shoen J. Automatically acquiring conceptual patterns without an annotated corpus. [J]. In Proceedings of the Third Workshop on Very Large Corpora, 1995, 3.

[Roger, 1978] Wilensky R. Understanding goal-based stories. [J]. Technical report, DTIC Document, 1978.

[Rospocher et.al., 2016] Rospocher M, van Erp M, Vossen P, et al. Building event-centric knowledge graphs from news[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2016, 37: 132-151.

[Rouces et.al., 2015] Rouces J, de Melo G, Hose K. Framebase: Representing n-ary relations using semantic frames[C]. European Semantic Web Conference. Springer International Publishing, 2015: 505-521.

[Rupnik et.al., 2016] Rupnik J, Muhic A, Leban G, et al. News across languages-cross-lingual document similarity and event tracking[J]. Journal of Artificial Intelligence Research, 2016, 55: 283-316.

[Shamma et.al., 2011] D.A. Shamma, L. Kennedy, and E.F. Churchill, “Peaks and persistence: modeling the shape of microblog conversations,” in CSCW, 2011, pp. 355–358.

[Tanev et.al., 2008] Tanev H, Piskorski J, Atkinson M. Real-time news event extraction for global crisis monitoring. [J]. In Proceedings of the International Conference on Application of Natural Language to Information Systems, 2008: 207–218.

[Tao et.al., 2013] Tao F, Lei K H, Han J, et al. Eventcube: multi-dimensional search and mining of structured and text data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 1494-1497.

[Trabasso, 1985] Trabasso T, Broek P V D. Causal thinking and the representation of narrative events. [J]. Journal of memory and language, 1985, 24(5): 612–630.

[Wang and Zhao, 2012] 王伟, 赵东岩. 中文新闻事件本体建模与自动扩充[J]. 计算机工程与科学, 2012, 34(4): 171-176.

[Wang et.al., 2007] Wang X, Zhai C, Hu X, et al. Mining correlated bursty topic patterns from coordinated text streams. Proceedings of the 13th ACM International Conference on Knowledge Discovery in Data Mining, San Jose, California, USA, 2007. ACM. 784–793.

[Wang et.al., 2009] Wang X, Zhang K, Jin X, et al. Mining common topics from multiple asynchronous text streams. Proceedings of the 2nd ACM International

Conference on Web Search and Data Mining, Barcelona, Spain, 2009. ACM. 192–201.

[Xu et.al., 2016] 徐佳俊，杨 飏，姚天昉，付中阳：基于 LDA 模型的论坛热点话题识别和追踪. In Journal of Chinese information proceeding Vol. 30. No 1. 2016

[Yangarber et.al., 2000] Yangarber R, Grishman R, Tapanainen P, et al. Automatic acquisition of domain knowledge for information extraction[C]//Proceedings of the 18th conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 2000: 940-946.

[Yang et.al., 1998] Yang Y, Pierce T, Carbonell J. A study of retrospective and on-line event detection. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998. 28–36.

[Yang et.al., 1999] Yang Y, Carbonell J G, Brown R D, et al. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems and Their Applications, 1999, 14(4):32–43.

[Yang et.al., 2002] Yang Y, Zhang J, Carbonell J, et al. Topic-conditioned novelty detection[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 688-693.

[Zacks, 2001] Zacks J M, Tversky B. Event structure in perception and conception. [J]. Psychological bulletin, 2001, 127(1): 3.

[Zhang, 2006] 张阔. 新闻挖掘关键技术研究. 博士学位论文, 2006

[Zhou, 2003] Zhou J. Research on financial event extraction technology-based on automatic rule acquisition [J]. Master's thesis, Tsinghua University, 2003.

[Zhu et.al., 2014] Zhu Z, Li S, Zhou G, et al. Bilingual event extraction: a case study on trigger type determination. [J]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 842–847.

[Zwaan, 1999] Zwaan R. Five dimensions of narrative comprehension: The event-indexing model. [J]. Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso, 1999, 5(1): 93–110.

第六章 知识存储与查询

1. 任务定义、目标和研究意义

简单地说，“知识图谱”就是以图（Graph）的方式来展现“实体”、实体“属性”，以及实体之间的“关系”。目前知识图谱普遍采用了语义网框架中 RDF[W3C, 2014]（Resource Description Framework，资源描述框架）模型来表示数据。语义网是万维网之父蒂姆·伯纳斯-李（Tim Berners-Lee）在 1998 年提出的概念[Wikipedia, 2018]，其核心是构建以数据为中心的网络，即 Web of Data；这是相对于我们目前的万维网是 Web of Pages 而提出的。语义网的核心是让计算机能够理解文档中的数据，以及数据和数据之间的语义关联关系，从而使得机器可以更加智能化地处理这些信息。因此我们可以把语义网想象成是一个全球性的数据库系统，也就是我们通常所提到的 Web of Data。本报告将从数据管理的角度去介绍在知识存储和查询方面的研究和应用问题。

RDF 是用于描述现实中资源的 W3C 标准。它被设计为提供一种描述信息的通用方法，这样就可以被计算机应用程序读取并理解。现实中任何实体都可以表示成 RDF 模型中的资源，比如图书的标题、作者、修改日期、内容以及版权信息。资源以唯一的 URI（统一资源标识——Uniform Resource Identifiers，通常使用的 URL 是它的一个子集）来表示，不同的资源拥有不同的 URI。这些资源可以用来作为知识图谱中对客观世界的概念、实体和事件的抽象。

图 1 给出了一个 RDF 数据实体示例，用来表示现实中一个著名欧洲哲学家亚里士多德（Aristotle）。在 RDF 数据模型中，亚里士多德就能通过亚里士多德头像上方所示的 URI 来进行唯一标识。客观世界的概念、实体和事件很多都是有属性。图 1 中亚里士多德头像下方给出的属性和属性值描述了亚里士多德这个资源所对应的人的名字是“亚里士多德”。此外，客观世界中不同概念、实体和事件相互之间可能会有各种关系，所以 RDF 模型中不同资源之间也是会存在关系。比如，图 1 给出了亚里士多德和另一个表示希腊城市卡尔基斯（Chalcis）所对应的资源通过一个 placeOfDeath 关系连接了起来，描述了亚里士多德死于卡尔基斯这个事实。

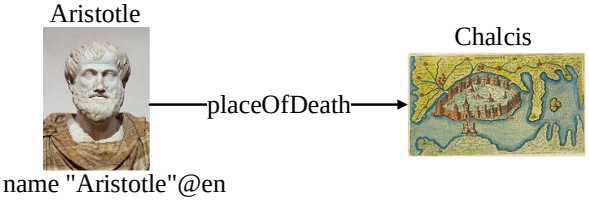

图 1. 示例 RDF 资源