元事件抽取研究综述

高李政 周 刚 罗军勇 兰明敬

(数学工程与先进计算国家重点实验室 郑州 450001)

摘 要 事件抽取是信息抽取领域的一个重要研究方向,在情报收集、知识提取、文档摘要、知识问答等领域有着广泛应用。对当前事件抽取领域研究得较多的元事件抽取进行了综述。首先,简要介绍了元事件和元事件抽取的基本概念,以及元事件抽取的主要实现方法。然后,重点阐述了元事件抽取的主要任务,详细介绍了元事件检测过程,并对其其他相关任务进行了概述。最后,总结了元事件抽取面临的问题,在此基础上展望了元事件抽取的发展趋势。

关键词 元事件抽取,信息提取,元事件检测,触发词

中图分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.002

Survey on Meta-event Extraction

GAO Li-zheng ZHOU Gang LUO Jun-yong LAN Ming-jing

(State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China)

Abstract Event extraction is an important branch of information extraction. It is widely used in intelligence collection, knowledge extraction, document summarization, question answering, etc. This paper presented a survey on meta-event extraction which attracts most attention in event extraction. It firstly interpreted the basic concepts of meta-event and meta-event extraction and briefly introduced the main implementation methods of meta-event extraction. Then it focused on the tasks of meta-event extraction, highlighted the task of meta-event identification, and briefly introduced other tasks. Finally, it summarized the problems of current meta-event extraction and predicted the trend of meta-event extraction.

Keywords Meta-event extraction, Information extraction, Meta-event identification, Trigger

1 引言

随着信息技术的飞速发展,网络已经成为人们获取信息的最主要的媒介。网络承载的数据呈现出爆炸式增长的状态,表现为数据种类繁多多样、结构复杂、规模庞大。“数据爆炸”使得人们迷失在数据的海洋中,无法区分有用知识和无效噪音,人工分析、处理数据变得异常困难,自动化的信息抽取技术应运而生。信息抽取是利用计算机从文本、音频、视频、图像等数据中自动抽取结构化信息的过程。事件抽取是信息抽取领域的一个重要研究方向,同时也是该领域最具挑战性的一个课题。事件抽取一直吸引着许多研究机构和研究者,MUC会议和ACE会议将事件抽取作为典型任务。

事件抽取在网络舆情监控、突发事件告警、情报收集领域有着重要应用。网络舆情变化通常是由某些热点社会事件引发的,事件抽取技术可以在第一时间发现这些热点事件,从而为预测网络舆情变化提供帮助。互联网传播消息的速度很快,如果能够及时地从互联网中挖掘突发事件,将为政府部门做好应对赢得时间。在情报收集领域,事件抽取技术可以帮

助情报分析人员从大量的低价值情报数据中自动获取事件信息,大大减小情报人员的工作量,在数据量急剧膨胀的今天,自动化的事件抽取技术显得尤为重要。

简单来讲,事件是指对所发生的事情的描述,描述的要素通常包括事情发生的时间和地点、事情的内容和状态,以及与事情相关的参与者(或称为实体)等,通常用自然语言表达这些要素。如果事件只描述简单的动作或状态的改变,则称这类事件为元事件,如“恐怖分子袭击了美国世贸大楼”。如果事件描述的是事情的发展过程,则称这类事件为主题事件。

事件抽取是从某种语言的文本中获得事件要素的过程,是对表达事件的信息进行抽取的过程。事件抽取可分为元事件抽取和主题事件抽取两大类,目前研究得较多的是元事件抽取,本文重点关注元事件抽取技术。

2 元事件抽取的主要实现方法

根据实现方式,元事件抽取可以分为两类:基于模式匹配的方式和基于机器学习的方式。下文将对这两种方法进行简要介绍。

到稿日期:2018-12-24 返修日期:2019-02-25 本文受国家自然科学基金项目(61309007)资助。

高李政(1990-),男,博士生,CCF学生会员,主要研究方向为信息抽取、事件抽取,E-mail:gao1440429064@foxmail.com;周刚(1974-),男,教授,主要研究方向为数据挖掘、知识图谱,E-mail:gzhoughzhou@126.com(通信作者);罗军勇(1964-),男,教授,主要研究方向为数据挖掘、社交网络分析;兰明敬(1982-),男,副教授,主要研究方向为知识图谱、舆情分析。

2.1 基于模式匹配的元事件抽取

基于模式匹配的元事件抽取是指在元事件模板的指导下对元事件进行检测和信息提取,元事件模板主要用于指明构成目标信息的上下文约束环境。基于模式匹配的元事件抽取包含两个基本步骤:模式获取和元事件抽取。

这种方法的关键在于元事件模板的构建。最初,模板构建主要依靠手工进行,这种方法费时费力,需要较强的专业知识。研究人员尝试用机器学习的方法自动构建元事件模板,取得了良好的效果。总的来说,基于模式匹配的元事件抽取在特定领域往往能够取得较好的检测效果,但其跨领域和可移植性较差,因此近年来研究人员更多地利用机器学习的方法进行元事件抽取。

2.2 基于机器学习的元事件抽取

基于机器学习的元事件抽取建立在统计模型的基础上,将元事件检测和论元识别转换成分类问题。这种方法主要是选择合适的特征值与分类器来完成相关的分类。基于机器学习的元事件抽取可以分为管道式方法(pipeline)和联合学习方法。

2.2.1 管道式方法

管道式方法把元事件抽取分为触发词识别和论元识别等任务,所有的任务依次被执行。其中,触发词识别是整个元事件抽取的基础,后续工作依赖于之前任务的成果。文献[1]提出将元事件抽取分为触发词检测、论元检测、事件对齐、事件关系检测4个部分,并提出利用管道式的方法进行事件抽取。分别选择每个任务的特征,使用K近邻和最大熵算法构建模型,比较两种算法在相同任务中的性能。

特定的事件类型一般对应特定的论元,因此触发词信息对于论元识别很重要。另一方面,特定的论元通常也对应特定的事件,由于管道式方法中触发词检测总是在论元检测之前,在进行触发词检测时无法考虑论元信息,从而限制了触发词检测的精度,因此研究者们提出了联合学习方法。

2.2.2 联合学习方法

联合学习方法对触发词识别、论元识别等任务构建一个联合学习模型,使得触发词和论元信息可以相互促进彼此的提取效果。

2013年,文献[2]提出了一个联合预测模型,该模型通过结构预测的方法,将触发词识别和论元识别相结合,从而更有效地利用局部特征来提高抽取的效果;此外,将不同元事件的触发词、论元之间的依赖关系作为全局特征来辅助元事件抽取。实验结果证明,该方法的抽取效果优于普通的管道式方法。

2016年,文献[3]基于RNN构建了一个联合学习模型。其方法主要包括两部分:首先利用RNN将事件句表示成向量形式,句子中每个Token对应的向量采用3种向量的级联:第一种是对应词的词向量,第二种是对应词的类型向量,第三种是句子中的依赖特征;然后利用RNN对句子进行编码,单向的RNN会将某个输入及其之前的信息转化为向量表示,但这丧失了输入之后的信息,因此,该方法同时使用反向RNN,从句尾开始,将某输入及其之后的信息转化为向量表示,然后将正反向量的级联作为某时刻输入的上下文表示。

由于普通的RNN会忽略长期信息,文献[3]采用LSTM存储长期信息。

3 元事件抽取的主要任务

元事件抽取的基本任务分为元事件检测和论元提取,此外还包括元事件关系检测、元事件对齐等外延任务。当前大部分研究聚焦于元事件检测,对后续任务的研究较少。下文重点对元事件检测的已有方法进行阐述,然后对论元提取、元事件关系检测、元事件对齐、元事件模板提取等任务进行了简要介绍。

3.1 元事件检测

元事件检测是元事件抽取的基础,包括元事件识别和类型检测。当前的元事件检测方法主要分为基于触发词的方法和基于事件实例的方法两种。

3.1.1 基于触发词的元事件检测

ACE认为触发词是元事件语句中最能描述该元事件发生的词语。最初的元事件抽取方法通常将元事件检测等价于触发词检测,将句子中的每个词当作候选触发词,利用词性特征、上下文语义特征、语法特征、实体特征等来判断该候选触发词对应的事件类型,文献[1]提出的方法是这种方法的典型代表。这种方法的问题在于:

(1)将每一个候选触发词作为分类实例,引入了大量反例,造成了正反例严重不均衡。训练语料中每种事件对应的触发词较少,导致召回率较低。

(2)触发词的歧义性。由于词语的多义性,同一个触发词可能对应不同类型的事件。例如,对于“他走了”这句话,很难判断其表示的是他去世了还是离开了。人们在写作时会有意地进行含蓄表达,这进一步增大了触发词识别的难度。

(3)中文触发词检测会受到分词效果的影响。中文语句的词语之间没有分隔符,需要进行分词,但由于中文结构复杂且灵活,分词结果经常出现错误,导致触发词检测出现问题。

(4)现有方法大多基于ACE标注数据,规模小、分布不均匀,导致模型训练效果较差。

针对这些问题,研究人员分别提出了不同的解决方法。

(1)触发词及事件类型扩展

文献[4]采用中文《同义词词林》对中文元事件进行触发词扩展,认为同义词通常触发相同类型的元事件,因此利用中文《同义词词林》扩展触发词集合,建立触发词表,提升了触发词检测的召回率;使用触发词的上下文信息来进一步确定元事件所属类型。

文献[5-7]利用中文组合语义学,从动词构词结构和形态结构两方面来识别未知触发词。其主要思想为:具有相同核心词素及相同形态结构的词语往往具有相似的语义。例如,“出任”和“担任”这两个词虽然词形不同,但是核心词素(“任”)相同,且拥有相同的形态结构,因此它们的语义也相近,都触发任职事件。

传统的触发词检测需要在有监督的条件下进行,因此需要大量的标记数据。ACE提供的guidelines中包含许多元事件的定义、描述以及事件样例,因而也包含大量触发词。文献[8]考虑抽取guidelines中的触发词项作为种子,为每一类元事件构建对应的触发词种子列表,然后将待检测句子中的每

个文本 Token 当作一个触发词,通过评估其与种子集合的相似度来识别元事件类型,该方法只需要很少的标记数据。

针对事件类型扩展,文献[9]提出了一种基于迁移学习的方法,认为相同类型的事件实例和事件模板在语义上也是相似的。因此,利用 CNN 将事件实例和对应的事件模板映射到同一语义空间中,并最小化两者间的距离,从而学习到一个通用的映射函数。某些知识库(如 FrameNet)中存在大量的事件模板,但是缺乏对应的事件实例。利用映射函数可以将这些事件模板映射到语义空间中,从而在缺乏标记数据的情况下实现事件类型的扩展。

(2) 触发词消歧

单独使用触发词特征很难实现消歧,研究人员考虑增加 3 种不同层次的特征,分别为句子级别、文档级别以及跨文档级别。

1) 句子级别。细粒度的实体信息可以为元事件类型识别提供依据^[10]。例如,对于“Trump beat Hillary”,如果我们知道 Trump 和 Hillary 都是美国总统候选人,那么几乎可以确认这一句子指的是“特朗普在总统选举中打败了希拉里”,而不是“特朗普殴打了希拉里”。

文献[11]认为,特定类型的实体往往出现在特定的元事件中,因此利用实体的类型信息可以减少触发词的歧义。首先将实体作为查询项,从大规模的文本资源中查询相关文档;然后利用这些文档的全局信息产生实体类型描述,并通过实体类型描述来对实体进行聚类;最后根据已知元事件以及实体类型信息对测试实例进行检测。

2) 文档级别。文档的上下文通常可以为元事件的类型识别提供依据。人们在阅读过程中遇到意义含糊的句子时,都会在这句话的上下文中寻找信息。文献[12]利用同一文档中不同元事件之间的关联性对歧义性进行判断。例如,当攻击事件发生时,发生死亡事件或者受伤事件的概率比发生其他事件的概率高。该方法的关键在于如何利用上下文信息。文献[12]首先抽取置信度较高的元事件,然后利用元事件之间的关系来对存在歧义的元事件进行判断。文档的主题通常与其包含的元事件类型有较强的相关性,例如娱乐类的新闻中通常不会出现攻击事件。因此,主题信息也可以作为消歧的依据。文献[13]同时考虑局部的实体信息、语义信息、候选触发词信息,以及全局的事件关系、事件-主题关系信息。本文利用局部特征对事件类型进行初步判断,在此基础上提取事件的全局特征,利用概率软逻辑构建模型,对事件类型进行最终判定。

3) 跨文档级别。文献[14]认为,在主题相关的文档集合中,触发词和论元的类型分布具有相似性。因此,首先对主题相关文档进行聚类,然后利用基准系统对这些文档进行元事件抽取,并利用抽取结果对具有歧义性的元事件进行判断。

此外,文献[15]认为触发词消歧与词义消歧^[16](Word Sense Disambiguation, WSD)具有很强的相似性,两者都是对语句中的词进行分类;其区别仅在于触发词消歧的类型标记是事件类别,而 WSD 的类型标记是词义。文献[15]提出了基于转移学习的方法,即利用 WSD 学习到的知识来辅助事件检测。

(3) 基于字的触发词检测

由于中文分词效果不理想,候选触发词与真实触发词的边界经常不一致。文献[17]提出了基于字的触发词检测方法,该方法将每个字当作一个检测实例,利用最大熵马尔科夫模型(MEMM)判断每个字的位置:位于触发词内(标记为 I)、位于触发词起始位置(标记为 B)或位于触发词外(标记为 O);联合进行分词识别和触发词识别,改善了中文触发词识别的效果。文献[18]在文献[17]的基础上,采用条件随机场模型(CRF),解决了 MEMM 的标记偏置问题。此外,文献[7]利用上下文一致性来避免中文分词结果对触发词检测的影响。

(4) 训练语料的扩充

文献[19]考虑利用 FrameNet^[20] 框架扩充训练语料,并利用事件框架信息辅助元事件抽取。文章认为 FrameNet 定义的框架与 ACE 定义的元事件在结构上很相似。元事件由触发词和论元组成,而框架由词汇单元和框架元素组成,框架的词汇单元在功能上与元事件的触发词接近,框架中的元素相当于元事件中的论元。此外,FrameNet 中的许多框架描述的是某种类型的元事件。因此,文献[19]首先研究如何构建元事件与框架之间的映射,利用映射关系从 FrameNet 中提取大量的元事件数据,有效扩展了训练语料;在此基础上进行模型训练,提高了元事件抽取的性能。文献[21]在文献[19]的基础上结合深度学习构建模型,进一步提高了抽取性能。

Freebase^[22] 是一个著名的知识库,利用 CVT(Compound Value Types)结构存储了大量结构化的元事件实例。文献[23]考虑如何从 Freebase 中提取元事件来扩充元事件语料,由于 Freebase 的元事件结构只包含事件论元,缺乏触发词,因此考虑通过弱监督的方法从维基百科文本语料中恢复触发词。文章认为,如果句子 A 包含了元事件 E 的所有论元,则 A 极有可能是 E 的事件描述句。事实上,事件论元往往散布在多个语句中,单个语句包含所有论元的情况极少出现。因此,文章改用关键论元对事件句进行定位,然后通过简单的规则对触发词进行初步筛选,最后利用 FrameNet 对触发词进行过滤和扩展。

在事件检测的过程中,经常会遇到某些事件数据比较丰富(记为 D_A),而另一些事件数据比较稀少(记为 D_T)的情况,训练数据稀少会影响模型的训练效果。针对这个问题,文献[24]提出利用迁移学习的方法借助 D_A 数据训练 D_T 数据的检测模型。该模型分为两步:首先利用 D_A 数据来训练初始模型,以获得初始的参数和模型;然后利用 D_T 对第一步得到的模型继续进行训练,从而得到最终模型。

此外,研究人员还提出了其他方法来改善触发词检测的性能。文献[25]认为,特定的元事件对应特定的论元;反之,特定的论元也对应特定的元事件。文章在元事件类型未知的前提下,利用模式匹配的方法提取事件句中的核心论元和辅助论元。这种论元提取方法的准确率高、召回率低,其虽然不能识别事件句中的所有论元,但提取出的论元信息对于元事件检测的性能提升已有明显效果。在完成元事件检测之后,对论元进行再次提取,并将此次的结果作为论元检测的最终结果。

文献[26]利用英汉双语协同训练的方法来提高事件检测的精度。该方法的核心思想是将事件的英文数据和中文数据分别看作事件数据的两个视图。文献[27]认为,同一事件的不同语言版本通常可以实现信息互补。首先,不同语言中表达相同含义的句子往往包含相同的语义成分;其次,某个词在一种语言中有歧义,但在另一种语言中可能没有歧义。因此,文章同时考虑多语言信息,利用注意力模型控制目标语言流向源语言的信息,从而实现多语信息的整合。

针对管道模型在元事件检测时无法利用论元信息的问题,研究人员提出了联合模型。然而,由于元事件中的论元数量比触发词数量多得多,因此联合模型更倾向于论元检测。文献[10]在相关实验的基础上发现,联合模型对于论元检测的提升效果明显高于元事件检测的效果,在某些情况下联合模型中的元事件检测性能甚至出现下降。因此,文章考虑在元事件检测中直接利用论元信息。由于句子中论元信息比非论元信息的重要性更高,因此文章采用注意力机制来提高触发词检测时论元信息的权重,结果取得了良好的检测效果。

随着词向量^[28-30]在自然语言处理领域的成功应用,研究人员开始考虑将词向量技术应用于元事件检测。文献[31]把词向量、位置向量以及实体类型向量拼接为一个整体,按照词的顺序将向量有序排列。文章利用 CNN 构建元事件检测模型,且取得了良好的效果。文献[32]在文献[31]的基础上做了进一步改进,认为文本中经常出现一句话包含多个元事件的情况,而同一句话中的元事件往往存在很大关联,利用元事件之间的关联关系可以提升元事件检测的效果。传统 CNN 采用的 Max-Pooling 策略只能抽取出最重要的元事件,无法得到其他元事件的信息。文献[32]提出了一种 DMCNN(Dynamic Multi-pooling Convolutional Neural Network)算法,其采用了动态多池策略,可以同时抽取出多个元事件的重要信息,改善了元事件检测的效果。

基于 CNN 模型的事件检测方法通常只考虑候选触发词临近连续窗口内的其他单词,这样可能会漏掉句子中的一些重要结构。针对该问题,文献[33]提出随机抽取句子中的单词组合作为 k-grams 进行卷积,提高了事件检测的性能。文献[34]认为随机抽取 k-grams 可能会引入一些噪音,因此提出利用句子中的句法依赖关系来确定候选 k-grams。首先对事件句进行句法依存分析,然后在句法依存图上进行卷积操作。传统的 CNN 无法处理图结构,文献[34]采用了 GCN^[35](Graph Convolutional Networks)的方法。卷积网络适合处理局部信息,不适合处理长距离的依赖信息,因此文献[34]在进行卷积之前,首先利用 Bi-LSTM 进行表示学习,然后将得到的结果利用 GCN 进行处理。

文献[36]提出了一种 dbRNN(Dependency Bridge Recurrent Neural Network)模型,将语句中的依存关系信息和序列信息进行整合,检测结果优于只依赖序列信息的 RNN 模型的处理结果。模型的核心改进在于更新隐藏状态时考虑了结点的依存关系。每个结点可能与多个结点存在依存关系,且依存关系的方向可能是前向或者后向。模型为每种依存关系赋予权重,当前结点的依存信息是所有依存结点的隐藏状态的加权平均。模型增加了一个门结构来控制依存关系的影

响。文献[37]提出了一种 DAG-GRU 模型,该模型的思路与 dbRNN 很相似,都是利用 RNN 将语句中的依存关系信息和序列信息进行整合,其最大的不同在于,DAG-GRU 不需要额外增加门结构,而是利用注意力机制来控制依存关系信息所占的比重。

由于语言之间存在差异,每种事件检测方法通常只针对特定的语言。文献[38]提出了一种混合神经网络模型,该模型不依赖于特定的语言资源,可以从语料中自动学习特征,因此可以用来构建多语言的事件检测器。文章将语句信息分为局部的块信息和词语上下文信息两种,并利用 CNN 来获取语句的块信息,利用 Bi-LSTM 来获取上下文信息。将两种信息进行级联,最后利用 Softmax 层来对事件类型进行判断。

上述基于神经网络的方法通常只关注事件句内的信息,忽视了事件句所在的整个文档的信息。根据上文所述,考虑文档的主题以及上下文信息能够进一步提升事件检测的性能。因此,文献[39]提出了一种 DLRNN(Document Level Recurrent Neural Networks)模型,该模型首先利用 PV-DM 模型^[40]对文档进行向量表示,然后将文档向量与事件句中的词向量拼接在一起,利用 Bi-LSTM 处理得到的向量序列,最后利用 Softmax 层对事件类型进行判断。

3.1.2 基于事件实例的元事件检测

基于触发词的元事件检测把每个候选触发词当作分类实例,引入了大量反例,造成正反例严重不平衡;此外,它还存在触发词歧义以及分词效果不佳等问题。因此,研究人员考虑将元事件检测的层级扩展到句子级别。

文献[41]首先对候选事件句进行特征提取及特征表示,然后利用二分类的方法对新闻语料中的事件句和非事件句进行区分,把每个事件句当作检测实例进行聚类。由于聚类过程非监督,该方法相当于只对事件实例进行了划分,没有对聚类结果进行标记,也没有进一步抽取论元等事件信息。文献[42]在其基础上进一步利用多分类的方法对事件句进行类型识别。

事件实例向量通常采用向量空间模型的方法进行表示,向量的每一维对应一个词语。由于事件句通常很短,向量的大部分取值都为 0,导致向量相似度的计算效果较差。文献[43]提出利用《知网》计算中文词汇的语义相似度,并进行一定的加权来获得事件的相似度,然后利用 k-medoids 算法对事件进行聚类。

事件实例是由参与者、时间、地点、触发词等元素构成的。文献[44]尝试利用多种方法将事件元素向量化,然后将事件元素向量拼接成事件实例向量。为每种事件类型人工挑选少量的种子事件,通过计算待检测事件与种子事件之间的相似度来确定待检测事件的类型。

3.2 其他重要任务

本节主要对论元提取、元事件对齐(Meta-Event Coreference)、元事件关系检测、元事件模板提取进行简要介绍。

(1)论元提取。论元包括实体、时间、地点、数值等。论元提取是信息抽取的一个重要研究方向,已经有了很多比较成熟的方法,如实体的识别和对齐、时间表达式的抽取、语义角色标记等,论元识别主要依靠这些成熟的方法,其面临的主要

问题是论元缺失。一个事件的论元往往位于一篇文章的多个语句中,而现有方法大多只针对单个句子,无法得到事件的所有论元。为了解决跨语句论元的提取问题,文献[45]提出了一个文档级论元检测联合模型。首先利用最大熵方法构建句子级的论元和角色分类器,然后利用整数线性规划构建联合模型。为了更好地表示元事件之间的关系以及元事件与论元的关系,文章引入了各种层次、各种类型的约束条件,最终取得了良好的效果。文献[46]提出了一种解决新闻报道中时间论元缺失问题的方法。一篇新闻报道通常描述一个主题事件,主题事件是由不同的元事件组成的。这些元事件之间相互关联,不仅有因果关系,还包括时序关系。为了简洁,新闻报道中经常会省略某些元事件的发生时间,而根据这些元事件之间的关系就可以从已知时间的元事件去推测时间缺省元事件的发生时间。文章提出了两种方法来构建模型,即基于规则的方法和基于统计学习的方法。

(2)元事件对齐。元事件对齐的目的是确定哪些事件句指的是同一元事件。元事件是一种包含触发词、时间、地点、参与者等要素的复杂结构。两个事件句指向同一元事件的前提是共有的各事件要素都相同或一致。首先,触发词应当属于同一事件类型。然后,事件发生的时间地点应当一致。时间和地点的描述粒度很可能不统一,例如事件句 A 和 B 指向同一元事件,但是 A 的时间是“2018 年 6 月 6 日”,B 的时间是“2018 年 6 月”,元事件对齐首先需要判断这两个时间指的是同一时间。同理,地点的描述粒度也可能不统一,例如 A 的地点为北京,而 B 的地点为海淀区,虽然描述粒度不同,但它们实际指的是同一个地点。最后,事件参与者应当相同,主要涉及到实体对齐的相关技术。文献[47]利用层次聚类的方法对文档中的事件实例进行对齐,采用最大熵模型学习事件实例的相似度评价函数。针对已有方法忽略了事件属性信息的问题,对事件属性特征进行细致的分析和概括,并将其加入到相似度评价函数中。

(3)元事件关系检测。元事件之间主要有因果关系和时序关系。某个元事件的发生可能由一个或多个元事件引起,称其为元事件之间具有因果关系。元事件发生的先后和伴随关系称为元事件的时序关系。元事件关系的抽取可以分为同句、同文异句、异文 3 类。同句是指两个元事件存在于同一个句子中,这种情况下两个元事件有很大概率存在相互关系,对其进行关系检测也较为容易。同文异句是指两个元事件存在于同一个文档中,但属于不同的句子。由于属于同一文档,两个事件有一定的概率存在因果关系或者时序关系。位于不同文档中的两个元事件存在关系的可能性很小,抽取的难度最大。

(4)元事件模板提取。元事件模板是对元事件的基本描述,它规定了某种类型的元事件包含何种触发词、论元(及论元角色)以及事件属性(例如伤亡事件中的伤亡人数)等。元事件模板是元事件检测领域的重要资源,通常由人工定义。然而,人工定义元事件模板需要耗费大量的人力和物力,且定义的事件类型较少,不适用于大规模开放域事件的提取。文献[48]提出自动化的元事件模板提取方法,首先对开放域中的元事件进行预提取。开放域中由于缺乏统一的元事件类型

标注体系,无法通过基准方法提取元事件。文章首先利用 AMR(Abstract Meaning Representation)^[49]对候选元事件语句进行解析;其次利用 WSD(Word Sense Disambiguation)对候选触发词进行消歧,并利用 OntoNotes 以及 FrameNet 对触发词进行初步筛选,挑选与触发词存在 AMR 关系的词组作为论元;然后利用词向量表示的方法将候选触发词与候选论元表示成向量形式,利用张量表示元事件结构;最后根据论元和触发词对元事件进行聚类,将每个类中最靠近类中心的事件实例的类型作为类标记,利用 PropBank, VerbNet 和 FrameNet 对论元的角色进行标记。每个类对应一个元事件模板,每个类包含的论元对应模板中的槽。

4 面临的问题及未来的研究方向

经过十几年的发展,元事件抽取领域已经取得了丰硕的研究成果,但元事件抽取依然存在一些问题。

(1)当前研究主要基于 ACE 定义的规范,使用 ACE 提供的标注语料。由于 ACE 定义的元事件类型有限,现有方法通常只针对特定类型的事件,方法的可移植性和可扩展性不强。

(2)现有方法大多基于触发词对元事件进行检测。基于触发词的元事件检测把每个候选触发词当作分类实例,引入了大量反例,造成正反例严重不均衡;此外,它还存在触发词歧义以及分词效果不佳等问题。

(3)现有研究大部分聚焦于元事件检测,对论元提取、元事件关系检测、元事件对齐等任务的研究还不充分。

(4)元事件抽取的粒度较小,抽取的事件信息不完整。在实际需求中,我们更多地需要获取事件的起因、经过、结果,以及完整的事件要素等信息,而这些信息在单个元事件中是无法获得的。

针对上述问题,本文预测了元事件抽取领域在未来的研究方向。

(1)针对现有方法可移植性和可扩展性不理想的问题,未来的研究可能从两个方面着手。一方面是对事件类型进行扩展,对事件语料进行扩充;另一方面是研究开放域元事件抽取方法。

(2)针对基于触发词的元事件检测存在的问题,未来的研究可能转向基于事件实例的元事件检测,而句向量技术的发展和将会推动该检测技术的发展。

(3)在元事件检测技术逐渐成熟的基础上,未来的研究将更多地关注论元提取、元事件关系检测、元事件对齐等任务。

(4)主题事件描述的是事件的发展过程,它是多个元事件的有序集合,未来的事件抽取将更多地关注主题事件的抽取。

结束语 本文首先介绍了元事件及元事件抽取的基本概念;其次简要介绍了元事件抽取的主要实现方法;然后重点阐述了元事件抽取的主要任务,主要对元事件检测进行了详细介绍,对其他相关任务进行了概述;最后总结了元事件抽取面临的问题,并在此基础上展望了元事件抽取的发展趋势。

从目前的研究现状来看,元事件抽取领域还有许多方面值得研究,同时也有不少问题亟待解决。基于事件实例的元事件抽取以及主题事件抽取将是未来事件抽取领域的发展方向。

参考文献

- [1] AHN D. The stages of event extraction[C]//Proceedings of the Workshop on Annotating and Reasoning about Time and Events. Association for Computational Linguistics, Stroudsburg, PA; ACL, 2006;1-8.
- [2] LI Q, JI H, HUANG L. Joint event extraction via structured prediction with global features[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA; ACL, 2013;73-82.
- [3] NGUYEN T H, CHO K, GRISHMAN R. Joint event extraction via recurrent neural networks[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies. Stroudsburg, PA; ACL, 2016;300-309.
- [4] ZHAO Y Y, QIN B, CHE W X, et al. Research on Chinese Event Extraction[J]. Journal of Chinese Information Processing, 2008, 22(1):3-8. (in Chinese)
赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1):3-8.
- [5] LI P, ZHOU G. Employing morphological structures and sememes for Chinese event extraction[C]//Proceedings of COLING 2012. 2012;1619-1634.
- [6] CHEN C, NG V. Joint modeling for chinese event extraction with rich linguistic features[C]//Proceedings of COLING 2012. 2012;529-544.
- [7] LI P, ZHOU G, ZHU Q, et al. Employing compositional semantics and discourse consistency in Chinese event extraction[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA; ACL, 2012;1006-1016.
- [8] BRONSTEIN O, DAGAN I, LI Q, et al. Seed-Based Event Trigger Labeling: How far can event descriptions get us? [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg, PA; ACL, 2015;372-376.
- [9] HUANG L, JI H, CHO K, et al. Zero-shot transfer learning for event extraction[J]. arXiv:1707.01066, 2017.
- [10] LIU S, CHEN Y, LIU K, et al. Exploiting argument information to improve event detection via supervised attention mechanisms [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA; ACL, 2017;1789-1798.
- [11] HONG Y, ZHANG J, MA B, et al. Using cross-entity inference to improve event extraction[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics; Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011;1127-1136.
- [12] LIAO S, GRISHMAN R. Using document level cross-event inference to improve event extraction[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010; 789-797.
- [13] LIU S, LIU K, HE S, et al. A Probabilistic Soft Logic Based Approach to Exploiting Latent and Global Information in Event Classification[C]//Thirtieth AAAI Conference on Artificial Intelligence. Menlo Park, CA; AAAI, 2016;2993-2999.
- [14] JI H, GRISHMAN R. Refining event extraction through unsupervised cross-document inference [C]// Proceedings of ACL-08; HLT. Stroudsburg, PA; ACL, 2008;254-262.
- [15] LU W, NGUYEN T H. Similar but not the Same: Word Sense Disambiguation Improves Event Detection via Neural Representation Matching[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA; ACL, 2018;4822-4828.
- [16] RAGANATO A, BOVI C D, NAVIGLI R. Neural sequence learning models for word sense disambiguation[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA; ACL, 2017;1156-1167.
- [17] CHEN Z, JI H. Language specific issue and feature exploration in Chinese event extraction[C]// Proceedings of Human Language Technologies; The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume; Short Papers. Association for Computational Linguistics, 2009;209-212.
- [18] HOU L B, LI P F, ZHU Q M. Study of Event Recognition Based on CRFs and Cross-event [J]. Computer Engineering, 2012, 38(24):191-195. (in Chinese)
侯立斌, 李培峰, 朱巧明. 基于 CRFs 和跨事件的事件识别研究 [J]. 计算机工程, 2012, 38(24):191-195.
- [19] 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. Stroudsburg, PA; ACL, 2016;2134-2143.
- [20] BAKER C F, FILLMORE C J, LOWE J B. The berkeley framenet project[C]//Proceedings of the 17th international conference on Computational linguistics. Association for Computational Linguistics, 1998;86-90.
- [21] SPILIOPOULOU E, HOVY E H, MITAMURA T, et al. Event Detection Using Frame-Semantic Parser[C]// Meeting of the Association for Computational Linguistics. Stroudsburg, PA; ACL, 2017;15-20.
- [22] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York; ACM, 2008;1247-1250.
- [23] CHEN Y, LIU S, ZHANG X, et al. Automatically labeled data generation for large scale event extraction[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA; ACL, 2017;409-419.
- [24] NGUYEN T H, FU L, CHO K, et al. A two-stage approach for extending event detection to new types via neural networks[C]// Proceedings of the 1st Workshop on Representation Learning for NLP. Stroudsburg, PA; ACL, 2016;158-165.
- [25] LI P F, ZHOU G D, ZHU Q M. Semantics-Based joint model of

- Chinese event trigger extraction[J]. *Journal of Software*, 2016, 27(2):280-294. (in Chinese)
- 李培峰,周国栋,朱巧明. 基于语义的中文事件触发词抽取联合模型[J]. *软件学报*, 2016, 27(2):280-294.
- [26] CHEN Z,JI H. Can one language bootstrap the other:a case study on event extraction[C]// *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. Association for Computational Linguistics, 2009:66-74.
- [27] LIU J,CHEN Y,LIU K, et al. Event detection via gated multilingual attention mechanism[C]// *Thirty-Second AAAI Conference on Artificial Intelligence*. Menlo Park, CA: AAAI, 2018.
- [28] MIKOLOV T,SUTSKEVER I,CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// *Advances in neural information processing systems*. Cambridge, MA: MIT Press, 2013:3111-3119.
- [29] MIKOLOV T,CHEN K,CORRADO G, et al. Linguistic regularities in continuous space word representations[C]// *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies*. Stroudsburg, PA: ACL, 2013:746-751.
- [30] PENNINGTON J,SOCHER R,MANNING C. Glove: Global vectors for word representation[C]// *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA: ACL, 2014:1532-1543.
- [31] NGUYEN T H,GRISHMAN R. Event detection and domain adaptation with convolutional neural networks[C]// *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Stroudsburg, PA: ACL, 2015: 365-371.
- [32] CHEN Y,XU L,LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]// *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Stroudsburg, PA: ACL, 2015: 167-176.
- [33] NGUYEN T H,GRISHMAN R. Modeling skip-grams for event detection with convolutional neural networks[C]// *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: ACL, 2016:886-891.
- [34] NGUYEN T H,GRISHMAN R. Graph convolutional networks with argument-aware pooling for event detection[C]// *Thirty-Second AAAI Conference on Artificial Intelligence*. Menlo Park, CA: AAAI, 2018.
- [35] KEARNES S,MCCLOSKEY K,BERNDL M, et al. Molecular graph convolutions; moving beyond fingerprints[J]. *Journal of Computer-aided Molecular Design*, 2016, 30(8):595-608.
- [36] SHA L,QIAN F,CHANG B, et al. Jointly Extracting Event Triggers and Arguments by Dependency-Bridge RNN and Tensor-Based Argument Interaction [C] // *Thirty-Second AAAI Conference on Artificial Intelligence*. Menlo Park, CA: AAAI, 2018.
- [37] ORR J W,TADEPALLI P,FERN X. Event Detection with Neural Networks: A Rigorous Empirical Evaluation[J]. arXiv:1808.08504, 2018.
- [38] FENG X,QIN B,LIU T. A language-independent neural network for event detection [J]. *Science China Information Sciences*, 2018, 61(9):092106.
- [39] DUAN S,HE R,ZHAO W. Exploiting document level information to improve event detection via recurrent neural networks [C]// *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. Berlin: Springer, 2017: 352-361.
- [40] LE Q,MIKOLOV T. Distributed representations of sentences and documents [C] // *International Conference on Machine Learning*. New York: ACM, 2014:1188-1196.
- [41] NAUGHTON M,KUSHMERICK N,CARTHY J. Event extraction from heterogeneous news sources[C]// *Proceedings of the AAAI Workshop Event Extraction and Synthesis*. Menlo Park, CA: AAAI, 2006:1-6.
- [42] XU H L,CHEN J X,ZHOU C L, et al. Research on Event Type Identification for Chinese Event Extraction[J]. *Mind and Computation*, 2010, 4(1):34-44. (in Chinese)
- 许红磊,陈锦秀,周昌乐,等. 自动识别事件类别的中文事件抽取技术研究[J]. *心智与计算*, 201, 40(1):34-44.
- [43] XU X Y,LI B C,ZHANG X F, et al. News Text Event Extraction Driven by Event Sample [J]. *Computer Science*, 2011, 38(8):232-235. (in Chinese)
- 许旭阳,李弼程,张先飞,等. 基于事件实例驱动的新闻文本事件抽取[J]. *计算机科学*, 2011, 38(8):232-235.
- [44] PENG H,SONG Y,ROTH D. Event detection and co-reference with minimal supervision[C]// *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: ACL, 2016:392-402.
- [45] LI P,ZHU Q,ZHOU G. Joint Modeling of Argument Identification and Role Determination in Chinese Event Extraction with Discourse-Level Information[C]// *IJCAI*. San Francisco: Morgan Kaufmann, 2013:2120-2126.
- [46] GUPTA P,JI H. Predicting unknown time arguments based on cross-event propagation[C]// *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Stroudsburg, PA: ACL, 2009: 369-372.
- [47] CHEN Z,JI H,HARALICK R. A pairwise event coreference model, feature impact and evaluation for event coreference resolution[C]// *Proceedings of the Workshop on Events in Emerging Text Types*. Association for Computational Linguistics, Stroudsburg, PA: ACL, 2009.
- [48] HUANG L,CASSIDY T,FENG X, et al. Liberal event extraction and event schema induction[C]// *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: ACL, 2016:258-268.
- [49] BANARESCU L,BONIAL C,CAI S, et al. Abstract meaning representation for sembanking[C]// *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Stroudsburg, PA: ACL, 2013:178-186.