Journal of Computer Applications
计算机应用, 2022, 42(10): 2975 - 2989
ISSN 1001-9081
CODEN JYIIDU
2022-10-10
http://www.joca.cn

文章编号:1001-9081(2022)10-2975-15
DOI:10.11772/j.issn.1001-9081.2021081542

事件抽取综述

马春明[1],李秀红[1*],李　哲[2],王惠茹[1],杨　丹[1]
(1. 新疆大学 信息科学与工程学院,乌鲁木齐 830046; 2. 香港理工大学 电子及资讯工程学系,香港 999077)
(*通信作者电子邮箱xjulxh@xju.edu.cn)

摘　要:将用户感兴趣的事件从非结构化信息中提取出来,然后以结构化的方式展示给用户,这就是事件抽取。事件抽取在信息收集、信息检索、文档合成、信息问答等方面有着广泛应用。从全局出发,事件抽取算法可以分为基于模式匹配的算法、触发词法、基于本体的算法以及前沿联合模型方法这四类。在研究过程中根据相关需求可使用不同评价方法和数据集,而不同的事件表示方法也与事件抽取研究有一定联系;以任务类型区分,元事件抽取和主题事件抽取是事件抽取的两大基本任务。其中,元事件抽取有基于模式匹配、基于机器学习和基于神经网络这三种方式,而主题事件抽取有基于事件框架和基于本体两种方式。事件抽取研究在中英等单语言上均已取得了优秀成果,而跨语言事件抽取依然面临着许多问题。最后,总结了事件抽取的相关工作并提出未来研究方向,以期为后续研究提供参考。

关键词:事件抽取;事件表示;元事件抽取;主题事件抽取;跨语言事件抽取
中图分类号:TP391.1　　**文献标志码:**A

Survey of event extraction

MA Chunming[1], LI Xiuhong[1*], LI Zhe[2], WANG Huiru[1], YANG Dan[1]
(1. *College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang* 830046, *China*;
2. *Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong* 999077, *China*)

Abstract: The event that the user is interested in is extracted from the unstructured information, and then displayed to the user in a structured way, that is event extraction. Event extraction has a wide range of applications in information collection, information retrieval, document synthesis, and information questioning and answering. From the overall perspective, event extraction algorithms can be divided into four categories: pattern matching algorithms, trigger lexical methods, ontology-based algorithms, and cutting-edge joint model methods. In the research process, different evaluation methods and datasets can be used according to the related needs, and different event representation methods are also related to event extraction research. Distinguished by task type, meta-event extraction and subject event extraction are the two basic tasks of event extraction. Among them, meta-event extraction has three methods based on pattern matching, machine learning and neural network respectively, while there are two ways to extract subjective events: based on the event framework and based on ontology respectively. Event extraction research has achieved excellent results in single languages such as Chinese and English, but cross-language event extraction still faces many problems. Finally, the related works of event extraction were summarized and the future research directions were prospected in order to provide guidelines for subsequent research.

Key words: event extraction; event representation; meta-event extraction; subject event extraction; cross-language event extraction

0　引言

事件抽取研究具有重大意义和实用价值,是不同学科发展和融合的需要。在实际生活中,事件抽取研究在信息收集、信息检索、文档合成、信息问答等方面有着广泛应用,促使自然语言处理技术的发展取得了重大突破。事件抽取可以定义为检测特定类型事件并识别有关信息,即事件类别识别和事件元素识别。

将事件句从文本中检测出来,然后根据其特征判断其所属类别,即事件类别识别。在事件句的检测过程中,一般使

收稿日期:2021-08-31;**修回日期:**2021-12-08;**录用日期:**2021-12-09。　**基金项目:**国家语委科研重点项目(ZDI135-96)。
作者简介:马春明(1997—),男,四川绵阳人,硕士研究生,主要研究方向:自然语言处理、事件抽取; 李秀红(1977—),女,山东威海人,副教授,博士,主要研究方向:自然语言处理、图像处理; 李哲(1992—),男,山东泰安人,博士研究生,主要研究方向:说话人识别、多模态语义分析; 王惠茹(1996—),女,新疆伊犁人,硕士研究生,主要研究方向:自然语言处理、图像处理; 杨丹(1996—),女,四川南充人,硕士研究生,主要研究方向:自然语言处理、图像处理。

用基于触发词的方法，在训练时实例化其中的每一个词，可以判定触发词是否存在于机器学习模型中。然而许多反例也被引进来，使正反例严重失衡。为解决上述问题，文献[1]中首先进行事件检测，然后对事件进行分类。这种方法是对部分事件进行特征选择，把特征选择中的正特征和负特征组合在一起，识别的效果较好。在基于触发词的方法中，不仅正反例严重失衡，还产生了数据稀疏性问题。为解决此问题，文献[2]中提出了一种全新的关于自动识别事件类别的算法。在事件句的分类问题中，主要使用了最大熵模型（Maximum Entropy Model，MEM）和支持向量机（Support Vector Machine，SVM）分类器进行分类。在进行候选事件句类别识别时，文献[3-4]中在基于二分类策略中均使用了以上两种分类器。在实际应用中，使用多元分类处理一个事件句属于多个事件类别的情况比使用二元分类更好，而用合适的事件特征来描述事件句以此提高分类的准确性是事件句分类的难点。文献[5]中利用选取词、上下文及其词典信息描述候选事件，在ACE（Automatic Content Extraction）2005上进行测试，该方法的F值为61.2%，效果良好。如果在原来的基础上引进依存分析，然后寻找触发词和别的词已有的句法关系，最后根据这个特征让事件句在支持向量机分类器上进行分类，该方法的F值为69.3%。为提高事件类别的相关识别率，未来研究将会重点放在分类器和事件特征的选取上。

识别出真正关于命名实体、时间表达式和属性值的事件元素，然后对它们进行正确的角色标注，即事件元素识别。事件句一般包含许多实体、时间表达式、属性值等事件信息。为了过滤真实的事件元素，必须首先识别并标注信息，对于信息理解会议（Message Understanding Conference，MUC）来说，这是很重要的研究内容。对于事件元素识别来说，如果事件信息识别及其标注在文本预处理时已经结束，事件元素识别在任务方面会产生和语义角色标签（Semantic Role Labeling，SRL）类似的效果。在一个句子中，动词（谓词）和有关联的不同短语的语句间有着语义关系，根据语义关系把语义角色信息给予这些句子的成分，即语义角色标注。例如施事、受事或者工具等。文献[6]中角色标注了任职事件和会见事件的元素，在条件随机场（Conditional Random Field，CRF）取得了良好的标注效果，这也说明事件元素和语义角色之间存在一定的联系。

文献[7]中在进行事件元素的识别时运用了上述联系。对于底层的模块，如分词以及句法分析等，很依赖这种联系；如果它们不够成熟，可能造成很多级联错误，对事件元素的识别有一定影响。为解决此问题，使用分类问题的思想来进行事件元素的识别，运用了MEM。在对候选元素进行描述时，从四种特征多方面进行：取词法、类别、上下文以及句法结构。为实现事件元素进行自动识别，运用了二元和多元两种分类策略[3]。

在最近的事件抽取研究中，文献[8]中提出了一种基于对比学习的预训练框架CLEVE，让预训练模型更好地从大型无监督数据中学习事件知识和对应的语义结构，从而在有监督和无监督的两种场景下都取得了良好结果。

本文从不同角度对事件抽取的研究现状进行了总结与展望。可大致分为5个部分：

1）从全局出发总结事件抽取算法以及评价方法，并介绍事件抽取所用的各种数据集以及与之相关的事件表示方法。

2）根据事件抽取的研究方向，详细介绍了元事件抽取和主题事件抽取的抽取方式以及使用不同抽取方式的研究现状。

3）介绍了中英文事件抽取的研究现状以及成果；跨语言事件抽取面临的问题及其解决方法，以及在未来研究中跨语言事件抽取的研究趋势。

4）根据不同研究角度，总结事件抽取相关技术，包括事件表示、元事件抽取、主题事件抽取、跨语言事件抽取的分类及特点。

5）事件抽取研究面临的问题以及未来研究趋势。

1 相关事件抽取算法及评价方法

事件抽取算法可分为四种：基于建立事件、事件句模板或者事件本体的模式匹配法；基于关键词的触发词法；基于领域本体的本体方法；把不同模型利用不同技术联合在一起的前沿联合模型方法。下文将对这四种事件抽取算法以及当前事件抽取主要的评价方法进行介绍。

1.1 基于模式匹配算法

以人工或自动构建的事件句子特征形式表示模板为指导的事件抽取，一般称为模式匹配。语义角色标注法与事件本体法是现有研究中最常用的构建模板的方法。

1.1.1 语义角色标注法

事件元素对应其语义角色，即语义角色标注法。对于实体、中心词词性以及关键词的层次，它们的语义约束在事件元素中完成定义。如果要使事件被匹配到，必要元素与相应的语义角色对应就会出现。首先预处理文本信息，然后在文本信息里进行语义角色标注，语义角色标注的语义信息通过词法分析对应得到；接着通过语义信息建立概念图，如果领域场景能被匹配到，就让规则库中的规则和映射规则一起匹配；最后，通过映射信息点实现抽取[9]。基于语义角色与概念图的抽取流程如图1所示。

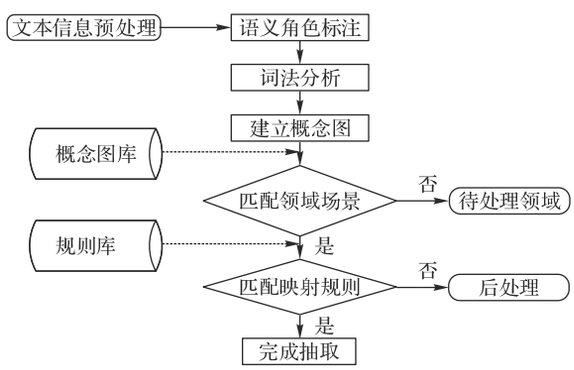

图1 基于语义角色与概念图的抽取流程

Fig. 1 Extraction flow based on semantic role and concept map

1.1.2 事件本体法

定义实体元素组、事件类别和事件的关系，从中得到特征项构建，再根据得到的特征项对事件和事件间的关系进行

挖掘,即事件本体法。文献[10]中提出了一种基于事件本体的文本特征重构方法,该方法表明了事件本体法的实用性。在构建事件的特征项时,包括两个流程:一是基于本体进行特征压缩,这主要指对同义项进行合并;二是基于本体进行扩充,这主要指在事件文本中,添加已经失去的事件元素特征。

在"中奖欺诈""网络色情""非法交易"三类语料库上比较了基于事件本体并且支持向量机的方法和只支持向量机的方法的准确性,实验结果如表1所示,与只支持向量机的方法(平均准确率为85.0%)相比,基于事件本体并且支持向量机的方法(平均准确率为78.7%)更加准确,这也说明了事件本体能让分类变得更准确。

表1　基于支持向量机与基于事件本体+支持向量机的文本分类结果

Tab. 1　Text classification results based on support vector machine and based on event ontology+support vector machine

方法	语料库类别	训练次数	正面测试语料库		负面测试语料库		准确率/%
			所有分类文档	正确分类文档	所有分类文档	正确分类文档	
基于支持向量机	中奖欺诈	57	58	46	58	42	74.1
	网络色情	36	36	32	36	27	81.9
	非法交易	49	48	41	48	36	80.2
基于事件本体和支持向量机	中奖欺诈	57	58	52	58	45	83.6
	网络色情	36	36	34	36	28	86.1
	非法交易	49	48	43	48	39	85.4

1.2　触发词法

触发词法也叫作事件关键词法。在统计处理事件句时,在句子的文本中有一类情况出现的事件句比较多,这种情况基本都是在句子文本中有某一种术语或者词汇,因此可以通过创建事件触发词词典,使得事件抽取出现更好的效果[11]。

创建事件触发词词典的方法有两类:一是在应用中,如果触发词的词量没发生多少变化,就基于领域经验由领域专家手工创建,不过这种方法很依赖领域专家的经验;二是根据词汇在事件句中已经存在的分析统计,把触发词从相应的事件句中提取出来,这类方法比第一类方法在触发词的查重率方面有所提高。在触发词字典中,也有两类方法进行系统应用:一是通过程序自动地读取建立的触发词库,这种方法比较灵活并且容易维护;二是在程序代码中直接写入触发词,这种方法不够灵活,必须通过对程序进行修改才能进行触发词的增减操作[12]。

1.3　基于领域本体的本体方法

领域本体事件基于专业领域的概念、领域概念的属性、方法及其概念之间的关系,但是这些概念可能并不仅仅是事件,甚至有些基本不包含事件。如果把某一领域的事件作为研究的对象,那么该领域概念可以用事件来表示,并且概念间关系对应于事件间关系;但在事件实体里面,元素之间的关系一般不存在[13]。在事件抽取算法中,都会有一个预处理阶段,这个阶段一般包括有分词、词性标注、去噪、特征提取等。通过本体例库里存在的命名实体以及命名实体之间的关系等语义信息,合并有联系的词,删掉无用信息构成领域实体;为了使特征项变少,可以合并同义概念,增加预处理性能。邻域本体通常和触发词、模式匹配、语义分析或者机器学习算法一起使用,即基于本体的事件抽取算法。

1.4　前沿联合模型方法

前沿联合模型方法是利用技术把不同的模型联合在一起。下面介绍三种联合模型。

1.4.1　模式识别和支持向量机联合

文献[14]中在进行模式识别时,使用了基于SVM的算法。在实验中设计了单分类器和多分类器两种算法,这是根据多元关系的特征进行研究,抽取事件的关系识别及其关系元。对于识别多元关系的全部角色,研究只使用了一种分类器,即单分类器算法;对于不一样语义约束的角色进行识别,研究在多种分类器上进行,即多分类器算法,实验结果表明,后者的算法效果比前者好。

1.4.2　机器学习和词嵌入联合

文献[15]中提出了一种抽取中文事件的方法。这属于商务事件抽取中的一种全新方法。在深度学习中,研究对模式、词嵌入技术以及机器学习模型进行集成。为扩展事件触发词的字典,运用了词嵌入以及事件触发词字典。在机器学习的算法中,引入了触发器特征,这种特征在字典中是存在的,研究使得事件类型识别变得更精细。

1.4.3　深度学习和词嵌入联合

文献[16]中提出了一种表示方法,该方法属于多重分布式表示,可应用在生物医学事件抽取中。在训练模型时,该方法中深度学习模型的输入使用了基于依赖的词嵌入和任务特征的分布式方法;在标记示例候选时使用了Softmax分类器。实验结果表明了该方法的先进性。

1.5　事件抽取评价方法

主流的事件抽取评价方法有两种:

1)微平均值法。

设P表示正确标注的数量与系统中进行标注的总数之比,即准确率;R表示正确标注的数量与按语料标准进行标注的总数之比,即召回率;F为它们的综合度量值。计算公式如式(1)所示:

$$F = \frac{2PR}{P + R} \tag{1}$$

2)错误识别代价法。

设L表示丢失率;M表示误报率;$Cmiss$表示一次丢失代价;Cfa表示一次误报代价;$Ltar$表示当系统作出肯定判断时的先验概率,一般为常值。错误识别代价C的计算公式如式(2)所示:

$$C = Cmiss*L*Ltar + Cfa*M(1 - Ltar) \tag{2}$$

在分析不同的算法效果时要运用不同的评价方法。通常单一的事件抽取都使用微平均值法来进行测评,而对于需要作出错误判断的事件比如话题追踪类任务等常用错误识别代价法来进行测评。

2 相关数据集

目前为止，事件抽取技术大多使用ACE2005数据集，但是它数据规模较小，具有严重的数据稀疏问题，因此后续研究又使用了其他数据集或者借助其他资源来解决数据集问题。

2.1 ACE2005数据集

ACE2005数据集是一种以阿拉伯文、英文以及中文作为培训数据并由关系、实体以及事件注释构成的不同类型的数据集。

ACE语料解决了实体、值、关系、时间表达式以及事件这5个子任务识别的问题，文档中存在的语言数据通过系统处理，这是子任务的要求。此外文档还要输出提到或者讨论子任务的信息。

下面是关于此版本中数据量、注释状态以及数据源缩略语信息：

1P：data subject to first pass（complete）annotation；

1P：须先通过（完整）注释的资料；

DUAL：data also subject to dual first pass（complete）annotation；

DUAL：数据也服从对偶第一遍（完整）注释；

ADJ：data also subject to discrepancy resolution/adjudication；

ADJ：资料也有经争议解决/裁定；

NORM：data also subject to TIMEX2 normalization；

NORM数据也要服从TIMEX2标准；

NW：Newswire；

NW：新闻专线；

BN：Broadcast News；

BN：广播新闻；

BC：Broadcast Conversations；

BC：广播对话；

WL：weblog；

WL：微博；

UN：Usenet；

UN：网络新闻；

CTS：Conversational、Telephone、Speech；

CTS：对话、电话、讲话。

adj、fp1、fp2、timex2norm文件夹分别表示不同的标注过程。ACE语料在所有任务上都是通过两个独立工作的标注器来进行标注。第一轮的标注成为1P，与之独立的双重第一轮标注成为DUAL。对于1P和DUAL来说，一个标注器完成文件的所有任务。文件是通过自动标注工作流程系统（Annotation Work-flow System，AWS）来进行分配的，而且文件分配是双盲的。Note：1P和DUAL在文件夹里都是以fp1和fp2来存放的，也就是说1P和fp1对应，DUAL和fp2对应。每个文件的1P和DUAL版本之间的差异由资深标注员或者小组负责人来进行裁决，从而得到一个高质量的gold standard文件。gold standard裁决文件被称为ADJ（即ADJ文件夹）。在裁决之后，TIMEX2值被标准化处理以后得到NORM。这个语料中的所有数据集都已经被NORM标注。表2为英文数据源的注释状态，表3为中文和阿拉伯文数据源的注释状态。

表2 英文数据源的注释状态

Tab. 2 Annotation status of English data sources

源	words				files			
	1P	DUAL	ADJ	NORM	1P	DUAL	ADJ	NORM
NW	60 658	57 807	33 459	48 399	128	124	81	106
BN	59 239	58 144	52 444	55 967	239	234	217	226
BC	46 612	46 110	33 874	40 415	68	67	52	60
WL	45 210	43 648	35 529	37 897	127	122	114	119
UN	45 161	44 473	26 371	37 366	58	57	37	49
CTS	47 003	47 003	34 868	39 845	46	46	34	39
合计	303 833	297 185	216 545	259 889	666	650	535	599

表3 中文和阿拉伯文数据源的注释状态

Tab. 3 Annotation status of Chinese and Arabic data sources

源	中文						阿拉伯文					
	chars			files			words			files		
	1P	DUAL	ADJ	1P	DUAL	ADJ	1P	DUAL	ADJ	1P	DUAL	ADJ
NW	127 319	124 175	121 797	248	242	238	61 287	56 158	53 026	239	226	221
BN	134 963	133 696	120 513	332	328	298	29 259	27 165	26 907	134	128	127
WL	71 839	68 063	65 681	107	101	97	21 687	20 181	20 181	60	55	55
合计	334 121	325 834	307 991	687	671	633	112 233	103 504	100 114	433	409	403

2.2 第四次信息理解会议数据集

第四次信息理解会议（Fourth Message Understanding Conference，MUC-4）事件抽取数据集包含1 700篇发生在拉丁美洲恐怖袭击的新闻报道。MUC-4数据集被切分为了1个dev集和4个测试集，其中dev集包含1 300篇文档，每个测试集中包含100篇文档。在使用MUC-4数据集时，使用了dev集中的1 300篇文章进行训练，test1+test2中的200篇文章作为dev集，test3+test4中的200篇文章作为测试集。

MUC-4包含4种类型的事件模板ARSON、ATTACK、BOMBING、KIDN。事件共用4种槽位Prepetrator、Instrument、Target和Victim。Prepetrator是Prepetrator Invdividual和Prepetrator Organization的组合。MUC-4数据集的标注样例如图2所示。

```
0. MESSAGE: ID                          TST4-MUC4-0006
1. MESSAGE: TEMPLATE                    1
…
4. INCIDENT: TYPE                       BOMBING
5. INCIDENT: STAGE OF EXECUTION         ACCOMPLISHED
6. INCIDENT: INSTRUMENT ID              "MINE"
7. INCIDENT: INSTRUMENT TYPE            MINE: "MINE"
8. PERP: INCIDENT CATEGORY              TERRORIST ACT
9. PERP:INDIVIDUAL ID                   "INSURGENTS"
10. PERP: ORGANIZATION ID               "FMLN"
…
```

图2 MUC-4数据集的标注样例

Fig. 2 Annotation samples of MUC-4 dataset

2.3 Freebase数据集

Freebase包含超过1.25×10^{8}个tuple关系元组、超过4 000种类别、超过7 000种属性,支持超大规模的collaborative data creation and maintenance,也就是支持信息之间的丰富关联并且赋予这种关联的使用。

Freebase里的数据包含非常多话题和类型的知识,如关于人类、媒体、地理位置等信息。同时Freebase不仅提供一个数据集或数据库,还提供较为便捷的访问方式。它支持面向对象的查询语言(Metaweb Query Language,MQL)与结构化的查询对象;还支持HTTPweb(Hyper Text Transfer Protocol web)端的访问和JSON(JavaScript Object Notation)数据格式的API(Application Program Interface)。

2.4 其他数据集

1)FrameNet(Frame Network)数据集是一个人读和机读的英语词汇数据库。它的基本思想很简单:事件、实体或者关系以及对参与者的描述叫作语义框架,而借助语义框架可以很好地对大部分单词含义进行理解。在ACE2005数据集中许多类别事件存在着数据稀疏问题;为了解决该问题,引入FrameNet数据集,在定义的事件类型里,让它与ACE2005数据集匹配,从而建立新的事件识别数据集。

2)TAC KBP(Text Analysis Conference, Knowledge Base Population)数据集2009—2018。TAC KBP是通过美国国防高级研究计划局进行资助的一种对实体链接的评测,TAC KBP数据集一般可用于事件抽取中,用手工进行标注,新闻与论坛是数据来源。

3)中文事件语料库(Chinese Emergency Corpus, CEC)属于生语料数据集,生语料来自互联网上5种突发事件的新闻报道,经过了一系列操作处理,最终把标注结果保存到语料库。该语料库总计332篇,全面标注了事件及其事件的要素。

3 事件表示

把信息通过结构化的形式表示出来,即事件。而把结构化形式的信息表示为计算机能够理解的形式称为事件表示,它促进了人工智能的发展,与事件抽取任务有着密切联系。人们早期基本使用离散的事件表示,后来开始研究以深度学习为基础,用神经网络来进行向量表示的稠密事件表示。

3.1 离散的事件表示

早期研究者们基本都使用由事件元素构成元组的离散事件表示。如文献[17]中使用三元组($\boldsymbol{O}_i$,$\boldsymbol{P}$,t)对事件进行表示,O表示给定对象集合,对象的谓词$\boldsymbol{O}_i \subseteq O$;$\boldsymbol{P}$表示对象与对象的关系或者属性;$t$表示事件的发生时间。文献[18]中则在事件表示中加入了角色元素,使用了六元组($\boldsymbol{P}$,$\boldsymbol{O}_1$,$\boldsymbol{O}_2$,$\boldsymbol{O}_3$,$\boldsymbol{O}_4$,t)进行标记,其中,$\boldsymbol{P}$为事件发生时的动作或者状态,也即对象与对象的关系或者属性;$\boldsymbol{O}_1$为不同数量事件的实施者;$\boldsymbol{O}_2$为不同数量事件作用的对象;$\boldsymbol{O}_3$为使不同数量事件发生的工具;$\boldsymbol{O}_4$表示一个或者多个地点;t为时间戳,也即事件的发生时间。文献[19]中使用了四元组($\boldsymbol{O}_1$,$\boldsymbol{P}$,$\boldsymbol{O}_2$,t),$\boldsymbol{P}$表示事件动作,也即对象与对象的关系或者属性;$\boldsymbol{O}_1$为实施事件者;$\boldsymbol{O}_2$为受事者,也即不同数量事件作用的对象;t为时间戳。一个事件仅有一个实施事件者和受事者。文献[20]中提出了一种事件表示方法。在脚本事件预测任务里,以时间为顺序将该方法与有关事件合成事件链。而在该方法中,构成以每个事件表示为动作并且动作和角色之间存在依存关系的二元组。由于角色在相同事件链中都是相同的,所以不用在事件表示中加入角色。

在离散的事件表示研究中,研究者们做了大量工作来对事件进行泛化,提出了基于语义的知识库,这很好地解决了离散事件表示所面临的稀疏性问题。例如文献[19]在事件元素中,基于WordNet(Word Network)把单词还原成词干,为得到泛化事件,把事件动作词泛化为一种类别名称,该类别名称存在于VerbNet(Verb Network)里。

3.2 稠密的事件表示

研究者们在深度学习技术不断发展的基础上对文本学习分布式的语义表示进行了探索。把字、词等文本单元嵌入向量空间,对于任意文本单元语义信息,由语义单元所在的向量空间位置确立,即分布式语义。在此基础上产生了稠密的事件表示,它的基础是预训练词向量,对此按照事件的结构进行语义组合。对于低维、稠密的向量,可计算事件的向量表示。稠密的事件表示分为两类:基于词向量参数化加法的事件表示和基于张量神经网络的事件表示。

3.2.1 基于词向量参数化加法的事件表示

对事件元素的词向量进行相加或拼接操作,再根据输入的参数化函数将它映射到事件空间向量,即基于词向量参数化加法的事件表示。文献[21]中提出对事件元素词向量进行操作,求取它的平均值。该方法属于基线方法。文献[22]中提出了一种向量表示方法,该方法拼接了事件元素词向量。文献[23]中提出一种词向量组合方法,组合前拼接了事件元素词向量,在多层全连接神经网络里面进行输入再组合操作。而文献[24]中忽视了组合事件元素的词向量,在文献[25-26]中直接用事件向量进行事件表示。不仅在事件表示中用事件元素向量的和或者平均值来表示,而且在不同的事件元素角色中出现相同词时使用不同词向量来表示。用$|V|$表示词表的大小,$|R|$表示角色的数量,H表示词向量的维数,三维张量$\boldsymbol{T} \in \mathbf{R}^{|V| \times |R| \times H}$由不同角色词向量组成。通过三个矩阵$\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$来表示三维张量$\boldsymbol{T}$,并且用$F$个一阶张量的乘积来表示张量的分解,减少了模型参数数量。如式(3)所示:

$$\boldsymbol{T}_{i,j,k} = \sum_{f=1}^{F} \boldsymbol{A}_{i,f} \boldsymbol{B}_{j,f} \boldsymbol{C}_{f,k} \tag{3}$$

设$\boldsymbol{r}$表示角色独热向量,$\boldsymbol{r}$和三维张量$\boldsymbol{T}$的切片相对应。$\boldsymbol{r}$和$\boldsymbol{T}$的切片$\boldsymbol{w}_r$如式(4)所示:

$$\boldsymbol{w}_r = \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{r}\boldsymbol{B})\boldsymbol{C} \tag{4}$$

最后,对于事件元素对应角色的词向量矩阵,可以在其中寻找其词向量,并且和所有事件元素词向量组合成事件向量。

3.2.2 基于张量神经网络的事件表示

对于基于词向量参数化加法的事件表示,虽然取得了良好效果,使词向量信息被完全利用,但对于建模事件元素来说,很难以实现交互,而且在建模时,事件表面形式的微小差异使之很困难。为了解决其中的问题,基于张量神经网络的事件表示被提出,该方法的事件元素通过双线性张量运算组合得到。

$\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbf{R}^d$ 表示两个事件元素向量，三维张量神经网络 $\boldsymbol{T} \in \mathbf{R}^{k \times d \times d}$，可得张量计算公式如式(5)所示：

$$\boldsymbol{v}_{\text{comp}} = \boldsymbol{v}_1^{\mathrm{T}} \boldsymbol{T}^{[1:k]} \boldsymbol{v}_2 \tag{5}$$

$\boldsymbol{v}_{\text{comp}}$ 的结果是 k 维向量，由向量 $\boldsymbol{v}_1$、$\boldsymbol{v}_2$ 以及矩阵 $\boldsymbol{T}_i$ 相乘得到 k 维向量里一个维度 i 上的元素。为了取得事件论元的交互，在双线性张量运算中，模型作了相乘运算；因此，虽然事件论元只有很小的表面区别，但是对于事件表示来说，语义上会有很大差别。

文献[27]中使用了三元组 $(\boldsymbol{O}_1, \boldsymbol{P}, \boldsymbol{O}_2)$，$\boldsymbol{P}$ 表示事件动作或者状态，$\boldsymbol{O}_1$ 为实施事件者，$\boldsymbol{O}_2$ 为受事者。研究考虑了它的事件结构，使用了神经张量网络模型，模型结构如图3所示。若使用 $\boldsymbol{O}_1$、$\boldsymbol{P}$、$\boldsymbol{O}_2$ 分别表示三种事件元素的词向量，即实施事件者 $\boldsymbol{O}_1$ 的词向量为 $\boldsymbol{O}_1$、事件动作或者状态 $\boldsymbol{P}$ 的词向量为 $\boldsymbol{P}$、受事者 $\boldsymbol{O}_2$ 的词向量为 $\boldsymbol{O}_2$，使用 $\boldsymbol{E}$ 表示组合两个向量的最终事件向量，$\boldsymbol{W}_i$ 和 $\boldsymbol{b}_i$ 均为张量参数。由张量运算、线性运算以及激活函数 f 组合起来，计算公式如式(6)~(8)所示：

$$\boldsymbol{R}_1 = f\left(\boldsymbol{O}_1^{\mathrm{T}} \boldsymbol{T}_1^{[1:k]} \boldsymbol{P} + \boldsymbol{W}_1 \begin{pmatrix} \boldsymbol{O}_1 \\ \boldsymbol{P} \end{pmatrix} + \boldsymbol{b}_1\right) \tag{6}$$

$$\boldsymbol{R}_2 = f\left(\boldsymbol{P}^{\mathrm{T}} \boldsymbol{T}_2^{[1:k]} \boldsymbol{O}_2 + \boldsymbol{W}_2 \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{O}_2 \end{pmatrix} + \boldsymbol{b}_2\right) \tag{7}$$

$$\boldsymbol{E} = f\left(\boldsymbol{R}_1^{\mathrm{T}} \boldsymbol{T}_3^{[1:k]} \boldsymbol{R}_2 + \boldsymbol{W}_3 \begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{R}_2 \end{pmatrix} + \boldsymbol{b}_3\right) \tag{8}$$

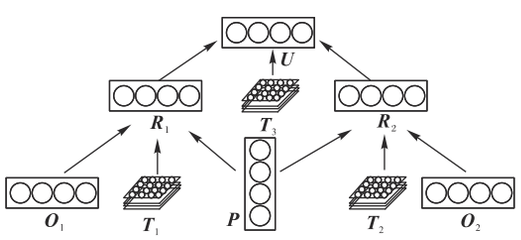

图3 神经张量网络模型结构

Fig. 3 Structure of neural tensor network model

文献[21]中同样使用了三元组 $(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o})$，考虑了事件结构，其中：$\boldsymbol{s}$ 表示主语，$\boldsymbol{p}$ 表示谓语，$\boldsymbol{o}$ 表示宾语，使用了谓词张量模型以及角色—因式张量模型，模型结构见图4。对谓语 $\boldsymbol{p}$ 用三维张量 $\boldsymbol{T}$ 进行建模。分别用 $\boldsymbol{s}$ 表示主语 $\boldsymbol{s}$ 的向量、$\boldsymbol{p}$ 表示谓语 $\boldsymbol{p}$ 的向量、$\boldsymbol{o}$ 表示宾语 $\boldsymbol{o}$ 的向量，事件向量 $\boldsymbol{e}$ 由主语向量 $\boldsymbol{s}$ 和宾语向量 $\boldsymbol{o}$ 通过张量 $\boldsymbol{T}$ 语义组合形成，它的每个元素 $\boldsymbol{e}_i$ 的计算公式如下：

$$\boldsymbol{e}_i = \sum_{j,k} \boldsymbol{T}_{ijk} \boldsymbol{s}_j \boldsymbol{o}_k \tag{9}$$

谓词张量(Predicate Tensor)模型通过张量 $\boldsymbol{T}$ 由谓语词向量 $\boldsymbol{p}$ 动态计算得出，然后由张量 $\boldsymbol{T}$ 语义组合主语和宾语。模型参数用 $\boldsymbol{W}$ 和 $\boldsymbol{U}$ 来表示，d 表示词向量维数，$\boldsymbol{W}$ 和 $\boldsymbol{U}$ 都是 $d \times d \times d$ 的三维张量，如式(10)~(11)所示：

$$\boldsymbol{T}_{ijk} = \boldsymbol{W}_{ijk} \sum_a \boldsymbol{T}_a \boldsymbol{U}_{ajk} \tag{10}$$

$$\boldsymbol{e}_i = \sum_{a,i,j,k} \boldsymbol{T}_a \boldsymbol{s}_j \boldsymbol{o}_k \boldsymbol{W}_{ijk} \boldsymbol{U}_{ajk} \tag{11}$$

角色-因式张量(Role-Factored Tensor)模型单独地对事件的主语及谓语、谓语及宾语进行语义组合，组合后的两个向量通过线性变换后相加得到事件向量，如式(12)~(14)所示：

$$\boldsymbol{vs}_i = \sum_{j,k} \boldsymbol{T}_{i,j,k} \boldsymbol{s}_j \boldsymbol{p}_k \tag{12}$$

$$\boldsymbol{vo}_i = \sum_{j,k} \boldsymbol{T}_{i,j,k} \boldsymbol{o}_j \boldsymbol{p}_k \tag{13}$$

$$\boldsymbol{e} = \boldsymbol{W}_{\text{S}} \boldsymbol{vs} + \boldsymbol{W}_{\text{o}} \boldsymbol{vo} \tag{14}$$

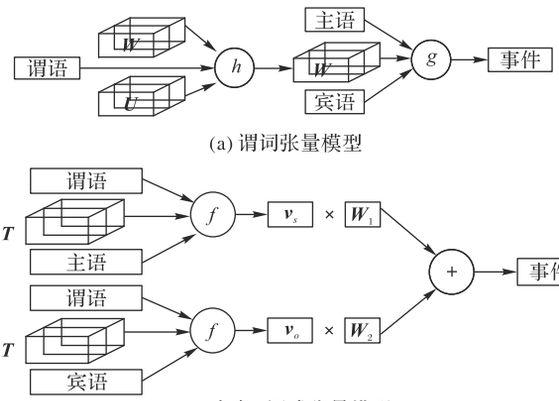

图4 谓词张量模型与角色-因式张量模型结构

Fig. 4 Structure of predicate tensor model and role-factored tensor model

文献[28]中使用了较小维度的张量值来分解低矢量的张量，使模型参数变少了。低秩张量分解运算的示意图见图5。用 $\boldsymbol{T}_1 \in \mathbf{R}^{k \times d \times r}$、$\boldsymbol{T}_2 \in \mathbf{R}^{k \times d \times r}$、$\boldsymbol{t} \in \mathbf{R}^{k \times d}$ 这三个参数来代替三阶张量参数 $\boldsymbol{T}$，而 $\boldsymbol{T}$ 的近似值为 $\boldsymbol{T}_{\text{appr}}$，$\boldsymbol{T}_{\text{appr}}^{[i]}$ 表示每一个切片，如式(15)所示：

$$\boldsymbol{T}_{\text{appr}}^{[i]} = \boldsymbol{T}_1^{[i]} \times \boldsymbol{T}_2^{[i]} + \text{diag}(\boldsymbol{t}^{[i]}) \tag{15}$$

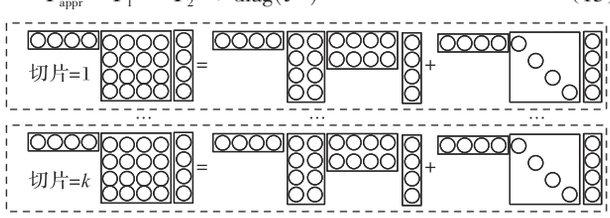

图5 低秩张量分解示意图

Fig. 5 Schematic diagram of low-rank tensor decomposition

在使用低矢量张量的分解时，不仅减少了模型参数，还能取得和以前模型差不多甚至更好的性能效果。

4 元事件抽取技术

元事件抽取方式有三类：基于模式匹配、基于机器学习和基于神经网络的元事件抽取。本章将对这三种类型进行详细介绍。

4.1 基于模式匹配的元事件抽取

模式的作用是在目标信息的上下文指定构成约束环，并且对语言和领域知识进行融合。在模式的指导下对元事件进行识别和抽取，即基于模式匹配的元事件抽取。为了使模式约束的信息得到满足，必须使用多种模式匹配算法进行抽取，构建模式是核心。基于模式匹配的元事件抽取分为两步：模式获取、元事件抽取，它的抽取框架见图6。

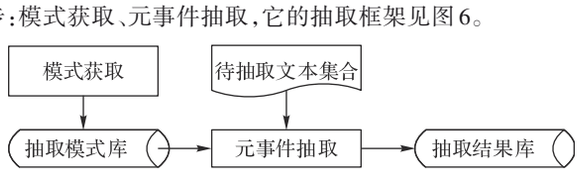

图6 基于模式匹配的元事件抽取框架

Fig. 6 Meta-event extraction framework based on pattern matching

在基于模式匹配的元事件抽取中,早期使用手工方法获取模式,这种方法费时间和人力,而且用户要有相当高的技能水平。文献[29]中对句型模板进行填充时建立了抽取规则,该规则是通过手工来确定的。在文本进行处理后,对事件信息进行抽取并填充句型模板。文献[30]中研究了自动获取模式,提出了一种学习方法,该方法基于领域无关概念知识库。在学习模式中,信息抽取(Information Extraction,IE)任务被用户定义,在没有分类和标准语料中,IE模式能自动被系统学习出来,降低了对用户的劳动力和技能的要求。文献[31]中将军事演习组块的识别和领域词典结合起来了,这是一种基于种子模式的自举方法。实验结果说明了该方法的有效性。

通常,使用模式匹配的方法来进行元事件抽取可以在特定领域内产生更好的结果;但是系统的可移植性不好,从一个领域移到另一个领域时,必须重新创建模式。建模既费时又费力,并且需要该领域的专家指导。尽管引入机器学习方法可以在一定程度上加快模式的获取,但是模式之间的冲突也是一个难题。此外,大多数可用的研究语义级别仍处于句法级别,并且语义级别仍需要改进。

4.2　基于机器学习的元事件抽取

4.2.1　基于机器学习的元事件抽取方法

基于机器学习的元事件抽取有两类方法:管道式抽取方法、联合学习方法。

管道式元事件抽取方法将抽取分为触发词以及论元识别等任务,它被转化为多阶段进行分类的问题。抽取的基础是触发词的识别,后面的抽取依赖触发词识别取得的成果。文献[3]中在抽取元事件时使用了管道式方法,分成触发词检测、论元检测、事件对齐以及事件关系检测四部分,并对它们进行特征选择,模型构建时选择了K近邻以及MEM算法,针对同一任务对两类算法进行性能对比。

由于在管道式方法中,先进行触发词检测再进行论元检测,论元信息在前者不能被考虑到,这对前者的精度有所影响。针对该问题,研究者们提出了联合学习方法。这种方法对各个任务都建立了一个联合学习的模型,使得在提取触发词与论元信息时,它们之间有相互促进的良好效果。文献[32]中使用了联合预测模型,使用带不精确搜索的结构化感知器来联合提取同一句子中同时发生的触发点和论据。根据当前模型w寻找最佳配置$z \in y$,$\boldsymbol{f}(x,y')$表示特征向量,如式(16)所示:

$$z = \operatorname{argmax} w \cdot \boldsymbol{f}(x,y');\ y' \in y(x) \tag{16}$$

感知器在线学习模型w,设$D = \{x^{(j)}, y^{(j)}\}_{j=1}^{n}$为训练实例集($j$索引当前训练实例)。在每次迭代中,$x$在当前模型下找到最优配置$z$,如果$z$不正确,则更新权值,如式(17)所示:

$$w = w + f(x,y) + f(x,z) \tag{17}$$

由于技术的挑战,还没有将联合产出结构作为一项单一任务进行预测的工作。而文献[33]中将实体识别和事件抽取作为一个联合任务进行,并用基于转移的神经方法进行建模。为了解决问题,研究使用了基于神经转换的框架建立了第一个模型,在状态转换过程中逐步预测复杂的关节结构,动作预测模型见图7。在该预测模型中,存储历史行为用栈A表示;存储的部分实体用栈e表示;维护未被处理的单词用缓冲区β表示;维护处理过的元素用栈σ表示;维护暂时从σ中出栈的元素;未来还会回栈的用队列δ表示;λ是一个变量,每次只提及一个元素。在标准基准上的结果显示了联合模型的优势,它给出了文献中最好的结果。

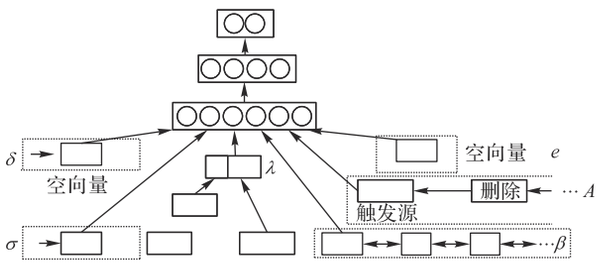

图7　动作预测模型

Fig. 7　Action prediction model

文献[34]中设计了一种基于跨度的事件提取器,采用联合学习抽取的方法对所有带注释的事件现象进行抽取。在新冠肺炎的预测任务中,自动提取的症状信息改善了测试结果的预测。该方法还将在事件抽取相关领域继续使用。

总而言之,尽管基于机器学习的元事件抽取方法对语料的内容格式不是很依赖,然而却存在着数据稀疏性问题,必须使用大规模语料。现如今的语料不能满足要求,使用人工标注又比较浪费人力资源;另外,机器学习的结果与特征选取有关。因此机器学习方法研究的重点是解决数据稀疏性问题和选择合适特征。

4.2.2　核心任务及面临问题

事件类别识别、分类和事件元素识别是元事件识别的两种核心任务。当识别元事件利用机器学习的方法时,元事件的分类及其文本分类存在差异,它的主要特点是分类简短,大部分是完整的句子。由于它是事件表述语句,因此语句中包含的信息量很大。

在事件元素的识别任务中,文献[35]中第一次引入MEM,实现了事件抽取。该模型在估计概率时使用了除所施加的约束以外尽可能少的假设原则。这些约束通常来自训练数据,表达特征和结果之间的某种关系。满足上述性质的概率分布是具有MEM的概率分布,它是唯一的,与最大似然分布一致,并具有指数形式,如式(18)所示:

$$p(o|h) = \frac{1}{Z(h)} \prod_{j=1}^{K} \alpha_j^{f_j(h,o)} \tag{18}$$

其中:o表示结果;h表示历史(或上下文);$Z(h)$是归一化函数。每个特征函数是二元函数。例如,在预测单词是否属于单词类时,o是true或false,h指的是周围的上下文。如式(19)所示:

$$f_j(h,o) = \begin{cases} 1, & o\text{为ture且前一个词为the} \\ 0, & \text{其他} \end{cases} \tag{19}$$

文献[36]中在研究语义角色标注时,用了CRF模型来做实验。这还有利于在TimeML(Time Markup Language)进行事件抽取,使得系统的性能大大提高了。为了使系统识别的能力提升,有时候让机器学习和模型匹配混合使用或者使用多个机器学习算法。如文献[3]中为了完成事件类别识别和

元素识别，把 MegaM 和 TiMBL（Tilburg Memory-Based Learner）这两类机器学习算法联系在一起，并在 ACE 语料库上进行了实验，证明了该方法比单一算法好。

以上对于事件的探测，都利用了触发词，但它只占全部词的小部分，致使在训练时许多反例被引进来，正反例严重失衡。并且在判断每个词的时候，增加了额外的计算量。为了解决此问题，文献[37]中在对事件类别进行识别时，采用了将触发词扩展与二元分类结合的方法。在相同特征下，分别测试文献[2]与文献[37]中的方法，实验对比结果如表4所示，表明了文献[2]中的方法更有优势。此外，在训练模型时，文献[2]中的词典收录了触发词并且扩展了同义词，解决了正反例严重失衡的问题，还使数据稀疏得到了缓解，在 ACE 数据集上的实验结果显示得了良好的效果。

表4 相同特征下不同方法的实验结果对比 单位：%

Tab. 4 Comparison of experimental results of different methods under same features unit:%

方法	类别	*R*	*P*	*F*
文献[2]方法	训练	43.06	58.29	49.53
	测试	38.91	52.36	44.64
文献[37]方法	训练	57.14	64.22	60.48
	测试	54.86	69.29	61.24

文献[2]和文献[4]在进行事件探测时不使用传统的基于触发词方法，而使用了基于事件实例方法。该方法识别实例为句子而不是词语，解决了正反例严重失衡的问题，数据稀疏也得到了缓解。在文献[2]的实验中，为把非事件句筛选掉，使用了二元分类器，再对取得的候选事件句进行分类，使用了多元分类器。在实验中，分别对8类事件类别以及33类事件子类别进行测试和训练，实验结果如表5。文献[4]中则将问题转化为聚类问题，以此得到事件句。

此外，文献[38]中提出了一个新的学习范式，将事件抽取转换成为一个机器阅读理解问题。该方法是将事件模式转换成一组自然问题，是一种基于网络的问答过程，以事件抽取的形式检索答案。实验结果显示了该方法在解决数据稀疏性和正反例失衡问题的优越性。

表5 文献[2]方法在不同事件类别上的实验结果 单位：%

Tab. 5 Experimental results of literature[2] method on different types of events unit:%

事件类别	类别	*P*	*R*	*F*
8类事件类别	测试	81.65	73.62	77.43
	训练	84.34	75.79	79.84
33类事件子类别	测试	74.24	65.34	69.51
	训练	76.35	64.26	69.79

4.3 基于神经网络的抽取方法

在元事件抽取方法中，结合神经网络进行抽取是一种主要方法，该方法属于有监督多元分类，该方法有特征选择以及分类模型两大流程。本文分别从使用特征的范围不同、模型学习方式不同、是否融合外部资源三方面对该方法进行描述。

4.3.1 根据使用特征的范围分类

句子级和篇章级是元事件抽取根据使用特征范围的分类。特征仅由句子内部得到的是句子级事件抽取，它的特征适用于全部事件抽取；特征里面有跨句子、跨文档信息的是篇章级事件抽取，它的特征适用于面向实际任务挖掘。

在句子级基于神经网络的事件抽取中，与传统离散特征的区别是它的特征是连续型向量，并在此基础上学习了更抽象的特征，该特征依托在各种各样神经网络模型上。如文献[5]中在事件抽取和事件识别任务中都使用了同样的方法，即神经网络方法。在传统卷积神经网络（Convolutional Neural Network，CNN）模型中，为使性能方面有所突破，加入了动态多池（dynamic multi-pooling）机制，构成了动态多池 CNN（Dynamic Multi-pooling CNN，DMCNN），DMCNN 的结构[5]如图8所示。

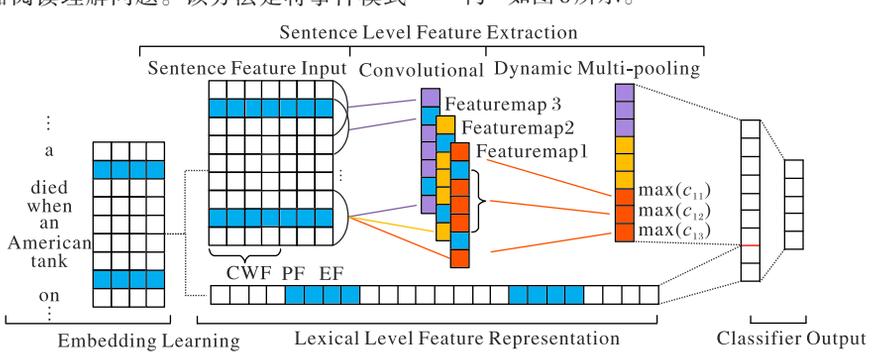

图8 动态多池卷积神经网络结构

Fig. 8 Dynamic multi-pool convolutional neural network structure

对于当前词，输入这个词和它前后的 c 个词的 embedding，通过 DMCNN 可以得到特征向量，再通过特征向量进行有监督训练完成抽取和识别。此外，对于事件抽取和识别，在初始表示每个单词时都选择了预训练词向量；在建模研究中，都对单词的语义和语法信息进行了组合。实验结果表明使用神经网络特征对句子级事件进行抽取可以取得良好效果。

在篇章级基于神经网络的事件抽取中，需要跨句子或跨文档信息，以此作为特征来完成任务。如文献[7]中首先研究端到端神经序列模型（带有预先训练的语言模型表示）如何在文档级角色填充提取中执行，以及捕获的上下文长度如何影响模型的性能。为了动态地聚集在不同粒度级别（例如句子级和段落级），提出了一种新的多粒度阅读器。

在多粒度阅读器模型结构嵌入层中，每个 token 通过单

词嵌入和上下文符号表征拼接表示;词嵌入使用 GloVe(Global Vectors for word representation)词向量模型,获得固定长度的预训练词向量。预训练语言模型表征已经被证明了拥有可以超出句子边界建模上下文的能力,并且在一系列自然语言处理任务上表现良好。在 MUC-4 事件抽取数据集上评估了该模型,结果表明最佳系统比以前的工作表现更好。多粒度阅读器模型结构如图 9 所示。该模型与 DMCNN 类似,均是由嵌入层到句子级别,再进行后续抽取和识别;而与 DMCNN 分类器提取结果不同的是该模型使用了融合机制再到 CRF 的过程。

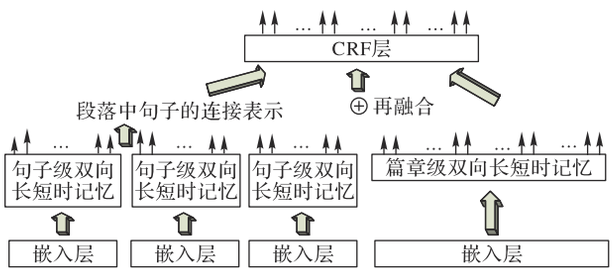

图 9　多粒度阅读器模型结构

Fig. 9　Multi-granular reader model structure

文献[39]中提出了一种文档级别的神经事件参数抽取模型,通过将任务公式转化为事件模板后的条件生成,还通过创建一个端到端的零触发事件提取框架表明了模型的移植性。

在以前的事件抽取研究中,大多数方法都直接基于触发词的有关特性进行研究,如一些分类的任务被用来辅助论元角色;但在对触发词进行识别的任务里,没有研究论元信息对它的作用。文献[40]中通过结合注意力模型,在事件识别里面成功地输入了论元信息,该注意力模型属于有监督论元。实验结果表明当识别事件触发词时,可以使用论元信息进行辅助。在该论元注意力模型中,在进行触发词的识别时,将论元信息直接与之结合起到辅助作用,这与在联合模型中间接地对触发词和论元信息进行结合然后共同辅助是不一样的。如果把事件检测当成多分类任务,而在句子中,将每一个符号全当成候选触发词,对候选触发词进行分类就是它的目标。

论元注意力模型由上下文表示学习和事件检测器两部分组成。其中,上下文表示学习的主要作用是通过注意机制获取上下文词汇的表示和实体类型信息的表示;事件检测器的作用是基于已经学习到的表示来对每一个候选词进行分类,也就是对事件进行分类。模型结构如图 10 所示,该模型与 DMCNN 均采用了分部分层次进行事件抽取的操作,最后均由分类器对结果进行输出。

4.3.2　根据模型学习的方式分类

根据模型学习方式分类的元事件抽取有流水线和联合模型。

流水线模型把元事件抽取分为触发词识别和论元识别等任务,依次完成全部任务。其中,在所有元事件抽取流程中,基础是触发词识别,它取得的成果将会对之后的工作产生很大影响。由于文献[3]中没有考虑到论元信息,其触发词的精确度有影响,因此研究者们提出了联合学习方法。

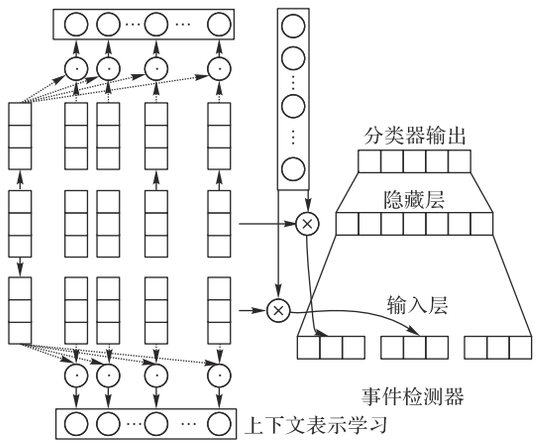

图 10　论元注意力模型结构

Fig. 10　Structure of argument attention model

文献[32]中在进行事件识别及其对论元角色进行分类时,采用了联合学习的方法。结构化感知机(structured perceptron)在研究中起到辅助作用,用来在联合学习中完成 2 个任务,即实体识别和实体对的关系分类。研究中还使用离散特征进行特征表示。该研究发现了联合学习方法比流水线方法效率更高,这在论元角色进行分类时更为突出。在基于神经网络的元事件抽取中,使用联合神经网络模型还简化了特征工程。文献[41]中在进行事件识别及其对论元角色进行分类时,也采用了联合学习的方法,设计了一种基于循环神经网络(Recurrent Neural Network, RNN)的模型。为进行特征表示,设计了局部和全局特征,其中,文本序列和局部窗口特征属于局部特征。在基于 RNN 的模型中传入句子表示,序列特征由此获得;局部窗口特征通过窗口里面的词向量获得。此外,还设计了记忆网络(Memory Network)模型进行建模,由此获取了全局特征,并且 2 个任务的性能也有所提升,取得了良好效果。

以前大多采用联合学习方法进行事件识别及其对论元角色进行分类,而文献[42]中首次对联合学习实体进行识别。在文档中抽取事件以及实体,在此环节通过联合推断让信息流贯穿 3 个子模块,并且在全局优化中为触发变量 t、论元角色变量 r 及实体变量 α 赋值,如式(20)所示:

$$\max_{t,r,\alpha}\sum_{i\in\mathcal{T}}E(t_i,r_i,\alpha)+\sum_{i,i'\in\mathcal{T}}R(t_i,t_{i'})+\sum_{j\in\mathcal{N}}D(\alpha_j) \tag{20}$$

式(20)由三部分组成:第一项是在事件内部结构模块的预估参数上单个事件置信度之和;第二项是事件对模块的预估参数上事件之间关系的置信度之和;第三项是实体识别的置信度之和。实验结果在置信度上取得了良好效果,该研究也在联合学习实体识别任务上取得了重大突破。

此外,文献[43]中提出了一种事件提取的可解释方法,通过为两个目标联合训练来缓解泛化和可解释之间的紧张关系。使用一个编码器-解码器架构,它联合训练一个用于事件提取的分类器以及一个规则解码器,生成解释事件分类器决策的语法-语义规则。在解释事件分类器中,有以下学习以及训练过程,如式(21)~(26)所示:

$$\boldsymbol{q}=\boldsymbol{W}_{\mathrm{q}}\boldsymbol{H}_{\mathrm{z}} \tag{21}$$

$$\boldsymbol{K}=\boldsymbol{W}_{\mathrm{K}}\boldsymbol{H}^{\mathrm{E}} \tag{22}$$

$$\boldsymbol{V} = \boldsymbol{W}_{\mathrm{v}}\boldsymbol{H}^{\mathrm{E}} \tag{23}$$

$$\boldsymbol{s} = \boldsymbol{q}\boldsymbol{K} \tag{24}$$

$$\boldsymbol{a} = \mathrm{Softmax}(\boldsymbol{s}) \tag{25}$$

$$\boldsymbol{C} = \boldsymbol{V} \odot \boldsymbol{a} \tag{26}$$

其中：$\boldsymbol{W}_{\mathrm{q}}$、$\boldsymbol{W}_{\mathrm{K}}$、$\boldsymbol{W}_{\mathrm{v}}$为学习矩阵，维数为200×200；$\boldsymbol{H}^{\mathrm{E}}$包含双向长短时记忆（Bi-directional Long Short-Term Memory，Bi-LSTM）的隐藏状态；$\boldsymbol{H}_z$是$\boldsymbol{H}^{\mathrm{E}}$中实体$z$的隐藏状态。将每个上下文向量$\boldsymbol{C}$与实体向量$\boldsymbol{H}$连接起来，并使用一个Softmax函数将连接的向量提供给两个前馈层，使用其输出预测该位置是否有触发器，使用二进制日志损失函数计算分类器的损失。这种方法可以用于半监督学习，并且当在由基于规则的系统生成的自动标记的数据上进行训练时，其性能得到了提高。

文献[44]中提出利用事件中参数的角色信息，设计一个分层策略网络（Hierarchical Policy Network，HPNet）来执行联合事件抽取（Event Extraction，EE）。整个事件处理过程是通过一个两级层次结构来完成的，该结构由两个用于事件检测和参数检测的策略网络组成，实现了子任务之间的深层信息交互，处理多事件问题更加自然。在ACE2005和TAC2015进行大量实验，分别使用MEM[35]、DMCNN[5]、HPNet[44]的实验结果如表6所示。从表6可以看出HPNet具有最先进的性能，并且对于具有多个事件的句子，优势更明显。

表6 ACE2005和TAC2015数据集上各个模型的结果对比 单位：%

Tab. 6 Results comparison of different models on ACE2005 and TAC2015 datasets unit：%

模型	ACE2005												TAC2005					
	事件触发识别			事件触发分类			事件参数识别			参数角色分类			事件触发识别			事件触发分类		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
MEM[35]	73.1	65.4	69.0	70.1	63.3	66.5	75.0	20.3	31.9	71.0	19.3	30.3	69.7	46.8	56.0	65.4	44.5	53.0
DMCNN[5]	79.6	67.2	72.9	74.3	62.9	68.1	69.1	51.8	59.2	62.8	45.0	52.4	77.4	48.7	59.8	71.3	45.8	55.8
HPNet[44]	81.3	77.2	79.2	80.1	75.7	77.8	70.2	53.8	60.9	64.6	50.7	56.8	78.2	55.6	65.0	70.9	54.8	61.8

4.3.3 根据是否融合外部资源分类

在元事件抽取任务中，大多使用ACE2005数据集，它含有很稀缺的有标记事件数据，但是标注质量不太好，而且规模很小、事件类型也很稀疏，这对完成事件抽取整体任务有很大影响，所以大量研究都试着使用外部资源来完成抽取。根据是否融合外部资源，可分成基于同源数据和融合外部资源两类。

文献[45]职工为解决事件类型稀疏的问题，使用了FrameNet数据集来辅助抽取。将ACE2005的事件类型上加入FrameNet里面的框架进行匹配，研究设计了全新的基于FrameNet的数据集，该数据集在事件识别等任务上取得了良好效果。

对从FrameNet检测到的事件进行间接评估，它基于的直觉是具有更高精度的事件预计会给基本模型带来更多的改进。使用自动检测到的事件扩充ACE语料，然后分别使用文献[5]方法、文献[40]方法、文献[41]方法、只使用人工神经网络（Artificial Neural Network，ANN）[45]、在ANN中加入FrameNet方法[45]共5种方法进行实验，结果如表7所示。可知文献[45]中的两个方法在FrameNet事件检测中的有效性。

表7 使用自动检测到的事件扩展训练数据的效果 单位：%

Tab. 7 Effect of expanding training data with events automatically detected unit：%

方法	*P*	*R*	*F*
文献[41]方法	71.8	66.4	69.0
文献[5]方法	75.6	63.6	69.1
文献[40]方法	75.3	64.4	69.4
ANN[45]	79.5	60.7	68.8
ANN+FrameNet[45]	77.6	65.2	70.7

此外，文献[46]中融合外部资源，研究设计了一个基于维基百科的事件数据集，该数据集使用了Freebase来辅助设计。在Freebase中，首先使用了统计方法找到在它任一事件类型里面的关键论元集合，然后通过在维基百科里面的每个句子，判断它里面是否存在Freebase里的任一事件实例的全部关键论元，以此来判断里面有没有存在事件。在存在事件的维基百科句子里使用了统计方法，以此找到每个Freebase事件类型里面的重要触发词。为对触发词进行筛选和对名词性的触发词进行扩展，还借用了FrameNet来辅助进行，最后得到了数据集。该数据集是从维基百科中得到的有标注的数据集，它被用来和ACE2005数据集一起训练模型。

为了获取事件抽取所需数据的方法，可用Freebase和FrameNet进行自动标注。任一事件类型的关键论元与触发词都可以通过以上方法探测得到，最后利用得到的关键论元与触发词来从文本中标注事件。该方法的体系结构如图11。

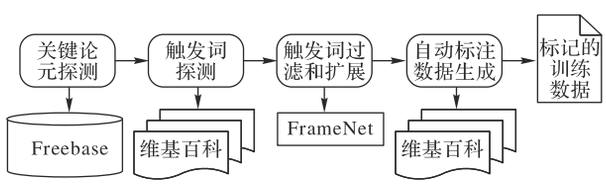

图11 自动标记数据的体系结构

Fig. 11 Architecture of automatic data labeling

文献[47]中使用了外部资源来建立批量事件数据集，该数据集是从维基百科和Freebase中建立的。为确定事件是否发生，该研究以是否含有关键论元来确定，这和以前在ACE2005中用触发词的方式来确定有所不同。此外，为获得每个事件类型的关键论元集，该研究也使用了统计方法，从Freebase里面进行抽取。设计中还含有事件抽取正例，这是从事件实例的关键论元的维基百科句子得来的。此外，为得到质量更高的有标注数据集，该研究还对远距离监督的进程实现了约束。

在以上研究中，事件抽取是直接抽取关键的论元，再加上论元大多是词组，因此将事件抽取转化成一个序列标注问题。目标是标出句子里的标签结构BIO（Beginning，Inside，

Outside),从而找到了这一堆实体,再去元数据 CVT(Compound Value Types)表里匹配即可。序列标注的模型使用 Bi-LSTM+CRF+整数线性规划(Integer Linear Programming,ILP)。Bi-LSTM 可以对于每个单独的词,很好地预测标签 BIO;CRF 的目标函数是整个序列的联合概率,可以让相邻的标签之间满足该有的依存规则;ILP 的过程是最大化目标函数,如式(27)所示:

$$\sum_{i,l,l'} v_{i,l,l'} * (P_{i,j} + A_{l,l'}) \tag{27}$$

其中:P 和 A 分别是 CRF 中的发射分数(emission score)和过渡分数(transition score);$P_{i,j}$ 表示标签 i 到标签 j 的概率;$A_{i,j}$ 表示标签 i 到标签 j 的过渡分数,是给定的参数。

此外,文献[48]中使用了外部资源来进行事件抽取,通过设计一种面向任务的对话系统,形成了一个由强化学习驱动的框架,实现了利用事件参数关系来进行事件抽取,并在 ACE2005 上评估了该方法的优越性。文献[49]中则是将事件抽取范例公式化为一个问答任务,基于问答系统以端到端的方式提取事件参数,实验结果表明了该方法的优越性。

5　主题事件抽取技术

元事件抽取只能在句子层面进行抽取,为了满足对一个及其以上的文档进行抽取,主题事件抽取应运而生,它是由一个以上的动作或者状态构成。为了对相同主题事件中的文档进行描述,需要确定进行描述的文档集合;并且在主题事件的集合里面有许多片段,需要将它们进行合并,这些是主题事件抽取的核心。将主题事件抽取分为基于事件框架和基于本体的主题事件抽取两类。

5.1　基于事件框架的主题事件抽取

对事件框架进行定义,将它结构化、层次化,然后对主题事件抽取进行指导,通过框架来阐述主题事件的各方面以及归纳事件信息,即基于事件框架的主题事件抽取。可以把框架当作一类知识表示的方法,可对有关概念的轮廓框架进行刻画。在人们处在一个新的状态时,会在人脑中进行搜索,从众多情景状态里面找到其中一个,让它来认识新事物。这些众多的情景状态就叫知识框架。对于事件侧面,在语义上能够对它进行分离,因此框架结构属于一种分类体系,把它用来对各种各样的事件侧面进行分隔。对于事件,需要用词语形容它的不同侧面,这样的词语称为“侧面词”。而分类体系可通过“侧面词”进行创建,这就是事件框架。对于框架方法,核心是要出现完全的事件框架体系;对于研究者们,研究的方向是提高构建框架的完整性和自动化程度,这也是研究的重点。

5.2　基于本体的主题事件抽取

在知识工程与人工智能中,本体是很重要的课题,主要用来得到有关的领域知识。关于领域知识,它们之间有共同理解,还能找到其中一起认可的词汇,对于这一系列词汇彼此之间的关系,能从各种各样的层次形式化模式里得到。根据本体的特点,很适合进行主题事件抽取。对于基于本体的主题事件抽取,主要是按照本体描述的信息来进行抽取,该信息包括概念、关系等,抽取的内容是文本里面的有关实体信息和侧面事件。抽取按照 3 步进行:建立领域本体,是后续抽取工作的基础;基于领域本体根据文本内容进行自动语义标注;基于语义标注进行抽取。

文献[50]中设计了一类基于本体的事件抽取。在建立本体的过程中,提出了领域层、类别层、事件层、扩展概念层 4 层模型。本体中所在领域的名称是领域层,许多个专家定义的类别层构成了它;任一类别都包括一系列事件集合;任一类别包含的事件类由事件层定义;事件和对象的概念以及对任一类事件相关的角色和概念及对应的子事件,这在扩展概念层进行定义。当对新闻事件进行抽取和在自动文摘中,可使用这个构建模型的本体,实验结果表明在中文气象这类新闻事件抽取时能更好地运用这个系统。

文献[51]中构建了一个进化的事件知识本体,以此探索从文本中自动获取事件知识的框架,指出未来研究将用此框架扩展数据,并将进化的事件本体扩展到大规模的事件实例中。

6　跨语言事件抽取

6.1　中文事件抽取

中文事件抽取存在着一系列问题:一方面是方法问题;另外一方面是语言特性问题,其中词句意合特性是首要问题。中文词语之间未曾出现显式间隔,并且分词之间显然存在着错误与误差。

在中文事件抽取中,文献[52]中指出触发词的不一致,并把该问题分为跨语言不一致以及内词语不一致两个类别。为解决上述问题,提出了两种方法:1)在基于词语的触发词标注中,可以对测试集里面分词不一致的触发词进行修正;使用训练集创建一个全局勘误表,该表可以对测试集进行修改。2)在基于字符的触发词标注中,可以对触发词检测进行操作,将它转变为序列标注问题。基于词语和字符的方法之间的性能比较如表 8 所示,实验结果表明基于字符的方法比基于词语的方法性能更好。

表 8　基于词语和字符的方法之间的性能比较　　**单位:%**

Tab. 8　Performance comparison between methods based on words and characters　　unit:%

方法	触发识别			触发标签			参数识别			参数标签		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
基于词语	68.1	52.7	59.4	65.7	50.9	57.4	56.1	38.2	45.4	53.1	36.2	43.1
基于字符	82.4	50.6	62.7	78.8	48.3	59.9	64.4	36.4	46.5	60.6	34.3	43.8

文献[53]中除了利用基于序列的字符标注法,还运用了 Bi-LSTM 以及 CRF,利用它们来抽取句子特征。在对上下文语义特征进行抽取时,还使用了 CNN,更好地完成了中文事件抽取。另外,中文事件抽取还存在着严重的数据稀疏问题,触发词相当多,而大量未知的触发词将会出现在测试集中。文献[54]中对未知的以及分词错误的触发词进行识别时,使用了中文语言组合语义以及语言一致性,使得系统性能有很大提升。

6.2 英文事件抽取

基于统计以及机器学习的方法是研究英文事件抽取的主要方法。文献[35]中使用了MEM来进行事件抽取研究，在命名实体等不复杂特征上具有很好成效。

文献[3]中将事件类型与触发词的识别进行等同，基于触发词进行事件抽取。在对事件类别和子类别进行识别时，除了使用触发词识别的二元分类以外，还使用了多元分类器，在ACE2005上显示了其效果很好。文献[55]中创建了关于跨文档的事件抽取系统，对当前句信息进行操作，在其基础上，把有关的文本背景知识植入进去。文献[56]中使用了文档级别信息，用它提升了系统性能。文献[32]中提出了一个联合学习模型，该模型基于结构化感知机，在实验中对事件触发词与论元进行学习然后抽取，该实验效果良好。

6.3 跨语言事件抽取

基于易得的大规模语料，事件抽取在中英等单语上已经取得足够优秀的成果，而跨语言事件抽取仍然面临着许多问题。

迄今为止，利用跨语言训练来提高性能的工作非常有限。为解决此问题，文献[57]中对众多双语平行语料进行操作，对跨语言谓词集进行抽取，接着使用抽到的谓词集对中英文事件抽取进行操作，以提高其召回率。文献[58]中对特征进行叠加，以此融合双语信息，还在中英文事件中都完成了触发词分类。文献[59]中则是提出了一种全新的跨语言事件抽取方法。这种方法训练了大量的语言，并通过语言特征的依赖性和不依赖性来促使性能提高。该方法不采用高质量的机器翻译或者手动对齐文档，因为给定目标语言是无法满足该需求的。

此外，跨语言还需解决缺乏标注数据给事件检测带来的挑战性问题，通过在不一样的语言之间传递知识，促使性能提升。以前的方法严重依赖并行资源，限制了适用性。为解决此问题，文献[60]中提出了跨语言检测的新方法，实现了并行资源的最小依赖。为了构建不同语言之间的词汇映射，设计了一种上下文依赖的翻译方法；为了解决语序差异问题，提出了一种用于多语言联合训练的共享句法顺序事件检测器。在两个标准数据集上进行了大量实验，实验结果表明该方法在执行不同方向的跨语言迁移和解决注解不足的情况下具有良好的效果。

从资源不足以及注释不足的语料库中进行复杂语义结构的识别（例如事件和实体关系）是很困难的，这已经成为了一个具有挑战性的跨语言事件抽取任务。为解决此问题，文献[61]中通过使用CNN，将所有实体信息片段、事件触发词、事件背景放入一个复杂、结构化多语言公共空间，然后从源语言注释中训练一个事件抽取器，并将它应用于目标语言。文献[62]中引入了一个图形注意力转换编码器（Graph Attention Transformer Encoder，GATE）。由于对句法分析的依赖，GATE产生了健壮性，有助于跨语言的传输。实验结果表明了该方法在跨语言事件抽取上的良好迁移效果。

基于以前的研究，很多小语种缺少标注语料。由于面临着语义表征等问题，面向小语种的跨语言事件任务成为目前研究的难点。

7 事件抽取技术总结

在事件抽取中，元事件抽取是动作状态级的，动作产生或状态发生变化，一般由动词驱动；而主题事件抽取是事件级别的，指的是核心或者与之有关的事件或者活动。表9详细总结了事件抽取与之相关的各项技术分类以及特点。

表9 事件抽取技术总结

Tab. 9 Summary of event extraction technologies

事件抽取任务	分类	特点
事件表示	离散的事件表示	事件表示为由事件元素构成的元组；面临稀疏性的问题
	稠密的事件表示	以预训练的词向量为基础；根据事件结构对事件元素的词向量进行语义组合；为事件计算低维、稠密的向量表示
元事件抽取	基于模式匹配的元事件抽取	在特定领域内产生更好的结果；系统的可移植性不好；需要建模，既费时又费力
	基于机器学习的元事件抽取	分类简短，大部分是完整的句子；由于是事件表述语句，因此语句中包含的信息量很大
	基于神经网络的元事件抽取	一种有监督多元分类任务；抽取方法包括2个步骤：特征选择和分类模型
主题事件抽取	基于事件框架的主题事件抽取	通过定义结构化、层次化的事件框架来指导主题事件的抽取；利用框架来概括事件信息
	基于本体的主题事件抽取	根据本体所描述的概念、关系、层次结构、实例等来抽取待抽取文本中所包含的侧面事件及相关实体信息
跨语言事件抽取	中文事件抽取	中文语言词语间没有显式间隔，而分词会带明显的错误和误差；存在触发词分词不一致、数据稀疏问题
	英文事件抽取	核心和主流方法是基于统计和机器学习的方法
	跨语言事件抽取	在多种语言上进行训练，并利用依赖于语言的特征和不依赖于语言的特征来提高性能

8 事件抽取面临的问题及未来研究趋势

8.1 面临的问题

事件抽取经过长期的发展已经取得了大量的研究成果，尤其在最近几年，随着社会化网络、电子商务应用的快速发展，事件抽取的研究进步更明显。但是从整体来看，还是存在以下问题需要解决：

1）目前研究事件抽取主要用的是ACE标注语料，但是定义事件类型有限。当前方法仅仅对特定类型事件有用，缺乏可移植性和可扩展性。

2）现阶段的事件框架体系不是通用的。仅通过人工来标注语料数据，费时费力且成本高昂，并且通过这种方式产生的事件语料数据规模小、类型少。

3）事件抽取依赖于子任务结果，为实现多任务联合，怎样设计神经网络模型是一大难点。

4）大量小语种缺少标注语料，面向小语种的跨语言事件

抽取面临着语义表征等问题。

8.2 未来研究趋势

在事件抽取技术的研究与发展过程中,尽管面临诸多挑战,也必将受到研究者越来越多的关注,并在未来的研究中呈现出以下趋势:

1)如今对事件抽取进行研究时,都是分开提取短语和依存句法分析信息的特征,怎样对这两种句法分析获取的信息进行全面分析,获得更有效的句法特征需要进一步研究。

2)在事件抽取中,对当前方法的局限性进一步分析;对任一子任务的影响程度进行量化。不仅需要提升句法分析这些基本任务性能,还需要使用新的方法与技术来提升事件抽取中任一子任务的精度。

3)如今对中文事件进行抽取时,大多都是基于现有语料的,实体信息都是已经标注好的语料,在没有标注好的生语料中抽取效果很不好。怎样对没有标注文本的中文事件进行抽取也值得进一步研究。

4)如何解决标注语料的缺失、面临语义表征等问题的面向小语种跨语言事件抽取是进一步研究的重点和难点。

9 结语

从当前研究来看,尽管研究者们对事件抽取技术已经进行了大量研究,在理论以及应用上都取得了许多成果,但依然没有达到实际应用的水平,事件抽取仍然存在大量需要研究的方向,同时还有许多问题需要解决,如如何更好地从无结构纯文本中自动抽取结构化事件知识等。研究者可能最需要关注的是可移植性以及系统性能问题;从作者自身角度上说,如今的事件抽取技术可能大多集中在某一领域进行研究,希望未来研究能渗透到不同领域,让事件抽取技术在多个领域实现创新和发展;诸如小样本和零样本这样的事件抽取研究甚少,希望未来研究能解决某些技术性难题,在这些方面有所贡献;主题事件抽取的研究尚未成熟,还面临着许多困难,能否借鉴神经网络以及外部资源来进行主题事件抽取是作者自身的一个猜想。

此外,事件抽取是自然语言处理的一个分支,它的研究价值已得到广泛重视和认可,不仅需要认识并研究它,还需要对比它和自然语言其他领域的区别和联系,以求创新来引导事件抽取研究的不断发展和进步。

参考文献 (References)

[1] 谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨:哈尔滨工业大学2008:14-89.(TAN H Y. Research on Chinese event extraction [D]. Harbin: Harbin Institute of Technology, 2008:14-89.)

[2] 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, 2010, 4(1):34-44.(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.)

[3] AHN D. The stages of event extraction [C]// Proceedings of the 2006 Workshop on Annotating and Reasoning about Time and Events. Stroudsburg, PA: Association for Computational Linguistics, 2006:1-8.

[4] NAUGHTON M, KUSHMERICK N, CARTHY J. Event extraction from heterogeneous news sources [C]// Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis. Menlo Park, CA: AAAI Press, 2006:1-6.

[5] CHEN Y B, XU L H, 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 (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2015:167-176.

[6] FENG X C, QIN B, LIU T. A language-independent neural network for event detection [J]. Science China Information Sciences, 2018, 61(9): No. 092106.

[7] DU X Y, CARDIE C. Document-level event role filler extraction using multi-granularity contextualized encoding [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020:8010-8020.

[8] WANG Z Q, WANG X Z, HAN X, et al. CLEVE: contrastive pre-training for event extraction [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2021:6283-6297.

[9] 杨选选,张蕾. 基于语义角色和概念图的信息抽取模型[J]. 计算机应用, 2010, 30(2):411-414.(YANG X X, ZHANG L. Information extraction based on semantic role and concept graph [J]. Journal of Computer Applications, 2010, 30(2):411-414.)

[10] LI C H, HU Y, ZHONG Z M. An event ontology construction approach to web crime mining [C]// Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway: IEEE, 2010:2441-2445.

[11] 丁效,宋凡,秦兵,等. 音乐领域典型事件抽取方法研究[J]. 中文信息学报, 2011, 25(2):15-20.(DING X, SONG F, QIN B, et al. Research on typical event extraction method in the field of music [J]. Journal of Chinese Information Processing, 2011, 25(2):15-20.)

[12] 吉久明,陈锦辉,李楠,等. 中文事件抽取研究文献之算法效果分析[J]. 现代情报, 2015, 35(12): 3-10.(JI J M, CHEN J H, LI N, et al. Effect analysis of Chinese event extraction method based on literatures [J]. Journal of Modern Information, 2015, 35(12): 3-10.)

[13] 吴奇. 基于领域本体的Web实体事件抽取问题研究[D]. 济南:山东大学, 2014:1-6.(WU Q. Research on domain ontology-based Web entity event extraction [D]. Jinan: Shandong University, 2014:1-6.)

[14] 赵小明,朱洪波,陈黎,等. 基于多分类器的金融领域多元关系信息抽取算法[J]. 计算机工程与设计, 2011, 32(7):2348-2351.(ZHAO X M, ZHU H B, CHEN L, et al. Multi-relation extraction in finance based on multi-classifier [J]. Computer Engineering and Design, 2011, 32(7):2348-2351.)

[15] HAN S Q, HAO X L, HUANG H L. An event-extraction approach for business analysis from online Chinese news [J]. Electronic Commerce Research and Applications, 2018, 28:244-260.

[16] WANG A R, WANG J, LIN H F, et al. A multiple distributed representation method based on neural network for biomedical event extraction [J]. BMC Medical Informatics and Decision Making, 2017, 17(S3): No. 171.

[17] KIM J. Supervenience and Mind: Selected Philosophical Essays [M]. New York: Cambridge University Press, 1993:336-357.

[18] RADINSKY K, DAVIDOVICH S, MARKOVITCH S. Learning

causality for news events prediction[C]// Proceedings of the 21st International Conference on World Wide Web. New York: ACM, 2012: 909-918.

[19] DING X, ZHANG Y, LIU T, et al. Using structured events to predict stock price movement: an empirical investigation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014:1415-1425.

[20] CHAMBERS N, JURAFSKY D. Unsupervised learning of narrative event chains[C]// Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2008:789-797.

[21] WEBER N, BALASUBRAMANIAN N, CHAMBERS N. Event representations with tensor-based compositions[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018:4946-4953.

[22] LI Z Y, DING X, LIU T. Constructing narrative event evolutionary graph for script event prediction[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. California: ijcai.org, 2018:4201-4207.

[23] GRANROTH-WILDING M, CLARK S. What happens next? event prediction using a compositional neural network model[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016:2727-2733.

[24] LEE I T, GOLDWASSER D. FEEL: featured event embedding learning[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018:4840-4847.

[25] TILK O, DEMBERG V, SAYEED A, et al. Event participant modelling with neural networks[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2016:171-182.

[26] HONG X D, SAYEED A, DEMBERG V. Learning distributed event representations with a multi-task approach[C]// Proceedings of the 7th Joint Conference on Lexical and Computational Semantics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 11-21.

[27] DING X, ZHANG Y, LIU T, et al. Deep learning for event-driven stock prediction[C]// Proceedings of the 24th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2015:2327-2333.

[28] DING X, LIAO K, LIU T, et al. Event representation learning enhanced with external commonsense knowledge[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4894-4903.

[29] 吴平博,陈群秀,马亮. 基于事件框架的事件相关文档的智能检索研究[J]. 中文信息学报, 2003, 17(6):25-30, 59. (WU P B, CHEN Q X, MA L. Study on intelligent retrieval of event relevant documents based on event frame[J]. Journal of Chinese Information Processing, 2003, 17(6): 25-30, 59.)

[30] 姜吉发. 一种跨语句汉语事件信息抽取方法[J]. 计算机工程, 2005, 31(2):27-29, 66. (JIANG J F. A method to do Chinese event IE from a multiple sentences' event narration[J]. Computer Engineering, 2005, 31(2):27-29, 66.)

[31] 贾美英,杨炳儒,郑德权,等. 基于模式匹配的军事演习情报信息抽取[J]. 现代图书情报技术, 2009(9):70-75. (JIA M Y, YANG B R, ZHENG D Q, et al. Sham battle information extraction based on pattern matching[J]. New Technology of Library and Information Service, 2009(9): 70-75.)

[32] 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 (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2013: 73-82.

[33] ZHANG J C, QIN Y X, ZHANG Y, et al. Extracting entities and events as a single task using a transition-based neural model[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. California: ijcai.org, 2019:5422-5428.

[34] LYBARGER K, OSTENDORF M, THOMPSON M, et al. Extracting COVID-19 diagnoses and symptoms from clinical text: a new annotated corpus and neural event extraction framework[J]. Journal of Biomedical Informatics, 2021, 117: No. 103761.

[35] CHIEU H L, NG H T. A maximum entropy approach to information extraction from semi-structured and free text[C]// Proceedings of the 18th National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2002: 786-791.

[36] LLORENS H, SAQUETE E, NAVARRO-COLORADO B. TimeML events recognition and classification: learning CRF models with semantic roles[C]// Proceedings of the 23rd International Conference on Computational Linguistics. [S. l.]: COLING 2010 Organizing Committee, 2010:725-733.

[37] 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. (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.)

[38] LIU J, CHEN Y B, LIU K, et al. Event extraction as machine reading comprehension[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2020: 1641-1651.

[39] LI S, JI H, HAN J W. Document-level event argument extraction by conditional generation[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2021: 894-908.

[40] LIU S L, CHEN Y B, 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 (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2017:1789-1798.

[41] 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: Association for Computational Linguistics, 2016:300-309.

[42] YANG B S, MITCHELL T M. Joint extraction of events and entities within a document context[C]// Proceedings of the 2016

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2016:289-299.

[43] TANG Z, HAHN-POWELL G, SURDEANU M. Exploring interpretability in event extraction: multitask learning of a neural event classifier and an explanation decoder[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Stroudsburg, PA: Association for Computational Linguistics, 2020:169-175.

[44] HUANG P X, ZHAO X, TAKANOBU R, et al. Joint event extraction with hierarchical policy network[C]// Proceedings of the 28th International Conference on Computational Linguistics. [S. l.]: International Committee on Computational Linguistics, 2020: 2653-2664.

[45] LIU S L, CHEN Y B, HE S Z, et al. Leveraging FrameNet to improve automatic event detection[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2016:2134-2143.

[46] CHEN Y B, LIU S L, 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 (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2017:409-419.

[47] ZENG Y, FENG Y S, MA R, et al. Scale up event extraction learning via automatic training data generation[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018:6045-6052.

[48] LI Q, PENG H, LI J X, et al. Reinforcement learning-based dialogue guided event extraction to exploit argument relations[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30:520-533.

[49] DU X Y, CARDIE C. Event extraction by answering (almost) natural questions[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2020:671-683.

[50] LEE C S, CHEN Y J, JIAN Z W. Ontology-based fuzzy event extraction agent for Chinese e-news summarization[J]. Expert Systems with Applications, 2003, 25(3):431-447.

[51] MAO Q R, LI X, PENG H, et al. Event prediction based on evolutionary event ontology knowledge[J]. Future Generation Computer Systems, 2021, 115: 76-89.

[52] 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. Stroudsburg, PA: Association for Computational Linguistics, 2009:209-212.

[53] ZENG Y, YANG H H, FENG Y S, et al. A convolution BiLSTM neural network model for Chinese event extraction[C]// Proceedings of the 2016 International Conference on Computer Processing of Oriental Languages. Cham: Springer, 2016: 275-287.

[54] LI P F, ZHOU G D, ZHU Q M, 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: Association for Computational Linguistics, 2012: 1006-1016.

[55] JI H, GRISHMAN R. Refining event extraction through unsupervised cross-document inference[C]// Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2008: 254-262.

[56] LIAO S 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. Stroudsburg, PA: Association for Computational Linguistics, 2010:789-797.

[57] JI H. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning[C]// Proceedings of the 2009 Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics. Stroudsburg, PA: Association for Computational Linguistics, 2009:27-35.

[58] ZHU Z, LI S S, ZHOU G D, et al. Bilingual event extraction: a case study on trigger type determination[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2014:842-847.

[59] HSI A, YANG Y M, CARBONELL J, et al. Leveraging multilingual training for limited resource event extraction[C]// Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. [S. l.]: COLING 2016 Organizing Committee, 2016:1201-1210.

[60] LIU J, CHEN Y B, LIU K, et al. Neural cross-lingual event detection with minimal parallel resources[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2019:738-748.

[61] SUBBURATHINAM A, LU D, JI H, et al. Cross-lingual structure transfer for relation and event extraction[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2019:313-325.

[62] AHMAD W U, PENG N Y, CHANG K W. GATE: graph attention transformer encoder for cross-lingual relation and event extraction[C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2021:12462-12470.

This work is partially supported by Scientific Research Key Project of National Language Commission (ZDI135-96).

MA Chunming, born in 1997, M. S. candidate. His research interests include natural language processing, event extraction.

LI Xiuhong, born in 1977, Ph. D. , associate professor. Her research interests include natural language processing, image processing.

LI Zhe, born in 1992, Ph. D. candidate. His research interests include speaker recognition, multi-modal semantic analysis.

WANG Huiru, born in 1996, M. S. candidate. Her research interests include natural language processing, image processing.

YANG Dan, born in 1996, M. S. candidate. Her research interests include natural language processing, image processing.