·理论探索·

中文事件抽取研究文献之算法效果分析

吉久明 陈锦辉 李 楠 孙济庆
(华东理工大学科技信息研究所,上海200237)

〔摘 要〕事件抽取是指识别文本中描述在某个时间(或时间段),某个地点或地区,由一个或多个角色参与的某动作的事件。首先对我国关于事件抽取研究的文献进行了总结,给出事件抽取的主要方法及模型。并针对文献中对这些事件抽取方法的效果进行统计分析,探讨各种事件抽取方法或模型的效果及适用性。经过对现有研究文献的统计,结论为:当前有关事件抽取的研究仍在继续,主要集中于金融资讯、会议信息、突发事件、个人简历等来自网页、微博微信等自媒体信息或军事法律等专业文件的事件抽取,所采用的算法包括SVM、CRF、ME、模式匹配、聚类算法等;CRF 算法应用与个人简历事件抽取效果最好,采用模式匹配算法的有效文献量相对较多,触发词方法的综合效果较优于模式匹配算法,但较多领域存在触发词算法的查全率较低的问题。

〔关键词〕中文事件;事件抽取;信息抽取;统计分析;情报分析
DOI:10.3969/j.issn.1008-0821.2015.12.001
〔中图分类号〕G253.1 〔文献标识码〕A 〔文章编号〕1008-0821(2015)12-0003-08

Effect Analysis of Chinese Event Extraction Method Based on Literatures

Ji Jiuming Chen Jinhui Li Nan Sun Jiqing
(Institute of Science and Technology Information, East China University of Science and Technology, Shanghai 200237, China)

〔Abstract〕Event Extraction is a process which extracts a series of events which consists of time, place and character. First, the paper summarized the literature about event extraction, and gave the main method and model of event extraction. Meantime, it analyzed the results of these methods and explore the effect and application of these methods. The result was that, the research of Event Extraction was going on, which is focused on the internet, the We-Media, and Professional documents. The internet is about Financial information, conference information, unexpected events, personal resume. The We-Media includes Microblogs and WeChat. The professional documents are about Military, law. the algorithm have SMV, CRE, ME, Pattern Match, Cluster. the number of literature using pattern matching algorithm is more, the comprehensive effect of trigger method is better than that of pattern matching algorithm, but the recall ratio of trigger method is low in areas.

〔Key words〕chinese event; event extraction; information extraction; statistical analysis; information analysis

当各种新闻充斥着互联网时,人们常常容易迷失方向,因此迫切希望能够直接浏览到从新闻中提取出来的简单直接的结构化的事件以及与之相关的后续事件,以备决策分析,而不是一堆辞藻堆砌的信息。事件抽取正是在这样的背景下产生的,它是信息抽取领域的重要研究方向之一,主要由计算机程序自动识别文本中描述在某个时间(或时间段),某个地点或地区,由一个或多个角色参与的某动作的事件。由于事件抽取涉及命名实体识别、命名实体之间的关系识别、事件之间的关系识别等技术,且由于中文具有博大精深的文化含义和语法灵活性,使得中文事件抽取的难度更大,至今仍是ACE(Automatic Content Extract)会议的主要研究目标之一[1]。本文首先介绍中文事件抽取的

收稿日期:2015-11-16
基金项目:国家社科基金资助项目“面向知识服务的学科领域术语语义分析及应用研究”(项目编号:13BTQ053)的研究成果之一。
作者简介:吉久明(1969-),女,部门主任,研究馆员,博士,硕士生导师,研究方向:知识集成。

基本思想及识别效果测评方法，进而介绍几种常用的事件抽取方法，最后对现有的研究中文事件抽取的重要文献中设计的算法效率进行统计分析，以期对中文事件抽取提供一些参考性的意见，促进特殊文本事件提取研究的发展。

1 事件抽取

1.1 基本思想

事件抽取就是要将某句子文本中所描述的非结构化的事件识别出来。其基本流程如图1 所示：

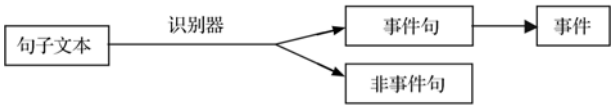

图1 事件识别的基本思想

情报分析领域还需对事件类型进行区分或将事件的元素进一步提取出来进行结构化表示供决策分析用。

信息抽取研究领域 (Information Extraction) 判断句子文本是否为事件句的依据一般为：该句文本中包含时间、地点、人物、动作、主题等基本的事件元素，如“周华健2008 年新年倒计时演唱会12 月31 日在上海举行”，由事件元素“周华健2008 年新年倒计时演唱会”、“12 月31 日”、“在上海”、“举行”构成事件句。ACE (Automatic Content) 会议则依据句子文本中是否含有事件触发词和描述事件结构的元素来判断事件句[2-3]，如“毛泽东1893 年出生于湖南湘潭”，由事件触发词“出生”及事件元素“毛泽东”、“1893 年”、“湖南湘潭”构成事件句。表面上看，两者的主要差别在于是否含有事件触发词，而事件触发词一般以动词或介词为主，因此两种判断依据基本一致。

由于语言表达的多样化及生动性需要，事件句中的事件元素往往存在不同的特征和模式，不同主题事件所包含的事件元素不同（如：识别句子中的场景描述[4]），其识别难度也不同，因此现有的研究一般针对具体的文本或事件主题设计识别任务，各种任务采用不同的方法。一般有两类基本方法：基于规则的方法或基于统计的方法[5]。

1.2 话题追踪与事件抽取

话题追踪（Topic Detection Tricking，TDT）涉及多个相关事件抽取，任务的目的是以大规模新闻流为研究和操作对象，通过监控新闻报道描述的话题，发现某类核心事件并跟踪其后续报道事件，由美国国防高级研究计划局（Defense Advanced Research Projects Agency）于1996 年提出[6]。其中，话题包括一个核心事件或活动（一个相互关联的事件集），以及所有与之直接相关的事件或活动。此处的事件是由某些原因、条件引起，发生在特定时间、地点，并可能伴随某些必然结果的一个特例。这类任务一般首先将新闻语料流切分成独立的报道，从报道中进一步识别某话题的多个事件子句，除报道切分及话题表示模型技术外，事件抽取也是关键技术之一[7]。

1.3 命名实体识别与事件抽取

命名实体是指文本中具有特定意义的实体，主要包括人名（Person）、地名（Location）、机构名（Organization）、日期（Data）、时间（Time）、百分数（Percentage）、货币（Money value）[8] 及身份、领域专业特有的术语，如：物质或蛋白质名称、化学分子式、生物化学反应、检测方法、化学仪器、药品名称、剂量等。事件的主要构成为事件元素，不同的事件识别任务中事件元素的类别不完全相同，除常用的人名、机构名、物质名、地点等命名实体外，还有其他的命名实体（演唱会名称）、事件动作、事件发生原因及其引起的后果等。因此，一些研究借助命名实体识别事件句[9]。另一方面，由于某些领域事件句具有相对明显的特征，也有研究将命名实体识别的任务建立在事件句模板的基础上，首先识别事件句，进而依据事件句模板识别其中的命名实体[10-11]。

1.4 效果测评

在事件抽取应用中通常采用两种不同的效果评价方法：基于召回率（记为R）准确率（记为P）的微平均（记为F）值法或基于丢失率（记为L）误报率（记为M）的错误识别代价（记为C）法。其中，

$$F = 2 * PR/(P + R)$$

$$C = Cmiss * L * Itar + Cfa * M * (1 - Itar)$$

Cmiss 为一次丢失的代价，Cfa 为一次误报的代价，Itar 为系统作出肯定判断的先验概率，通常根据具体应用设定为常值。上述公式表明，两种效果测评方法之间不存在简单的逆反关系，因此在分析不同评价方法下的两种不同算法的效果时应进行适当的换算。

微平均值法一般多用于单一事件抽取任务中，如：突发事件、门户网站、金融资讯的事件抽取。对于话题追踪任务而言，相对于正确率，人们对系统作出的错误判断往往更为敏感，这些错误包括：本应为是的判断为否（丢失），本应为否的判断为是（误报），因此常采用错误识别代价作为效果评价方法[6]。另外，事件抽取的各种算法在实际应用中，除考虑其识别结果的正确率外，还应该考虑算法的复杂程度及其可实现性。一些抽取效果好的算法往往是以牺牲时间为代价的。一些算法可能由于硬件要求太高，或训练时间太长而不具备可行性。

鉴于话题追踪任务的复杂性，本文主要关注已有的单个事件抽取算法的效果情况，故本文中所统计的研究文献均采用微平均值法评价算法的效果。

2 几种常用的事件抽取算法

如前所述，基本的事件抽取方法一般有两种：基于规则的方法及基于统计的方法。基于规则的方法首先建立事

件或事件句的模板或本体实现事件抽取，此类方法多应用于事件句或事件具有明显的特征，容易对其进行普遍形式化描述，如演剧院网站上显示的音乐会或电影院场次信息。基于统计的方法一般将事件抽取问题转化为句子文本的分类问题，应用此类算法抽取事件的句子文本或事件没有明显的特征，或者虽然具有一定的特征，但其形式多样、不断变化，不易于简单枚举，因此使用基于机器学习的统计类算法得出事件句的模式特征，实现事件抽取，主要有HMM、CRF、SVM、ME 等方法。

除上述基本方法外，还有基于命名实体抽取的方法。命名实体抽取方法的思路主要有两种：一是基于领域本体；二是基于事件关键词及语义知识或领域本体。

为区分起见，本文将基于建立事件或事件句模板、事件本体的事件抽取方法通称为模式匹配法，基于领域本体的事件抽取方法通称为本体方法，基于事件关键词的事件抽取方法通称为触发词法。

严格意义上，隐马尔科夫（HMM）、条件随机场（CRF）、支持向量机（SVM）、最大熵（ME）等机器学习算法也是模式匹配算法，这些算法已经被广泛应用于命名实体识别研究，只需将待构建模式的对象从较短的文本串调整为句子文本后，即可进一步推广至事件抽取应用中[12-15]，鉴于篇幅的原因本文不再展开此类算法的具体描述，而重点介绍基于模式匹配、基于本体、触发词的事件抽取算法。

2.1 基于模式匹配的事件抽取算法

本文将通过手工或自动构建的有关事件句特征形式化表示的模板指导事件抽取的方法通称为模式匹配，已有研究中比较典型的事件句模板构建方法有两类：语义角色标注、事件本体法。

2.1.1 语义角色标注法

语义角色标注法[4]将事件元素与相应的语义角色对应，并对事件元素定义实体、中心词词性和关键词层次的语义约束，匹配中，只要与必要元素对应的语义角色全部出现，即认为匹配到事件。如："[运动员(agent)][一周后(tmp)]将要[参加(V)]正式[比赛(patient)]"，这句话中，"参加"是谓词，"运动员"和"比赛"分别是其"施事者"和"受事者"，"一周后"表示其发生的时间，如果将这个句子的形式改变为"[一周后(tmp)][运动员(agent)]将要[参加(V)]正式[比赛(patient)]"，句中各部分的语义角色并没有发生变化。如果定义事件的必备元素为施事者、受事者、时间和谓词，该句子文本即为事件句。

语义角色标注法对于语义角色与事件元素映射关系相对固定且同时实现事件元素提取的事件抽取应用比较适合。该方法实施的关键是语义角色标注及构建语义角色与事件元素之间的映射，目前已经有多个较成熟的系统能实现自动语义角色标注，而语义角色与事件元素之间的映射关系主要基于标注语料中的事件句的语义角色统计得到，语义角色标注一般基于句法分析及领域知识。

2.1.2 事件本体法

事件本体法[16]首先定义事件的实体元素组、事件类别及事件之间的关系，进而获得事件的特征项构建，最后基于事件特征项挖掘事件及事件间的关系。事件实体元组一般包括活动元（Action），参与者元（Participant），时间元(Time)，位置地点元（Location），仪器设备元（Instrument）and 事物元（good），其中参与者元为施事或受事者。具有相同特征的事件属于同一类，事件种类因领域不同而不同。事件之间的关系主要有两类：类关系（class-related）或非类关系（non-class-related），非类关系又可分为组分(component)关系、原因结果（cause-effect）关系、跟随(follow)关系。如网络犯罪领域的事件抽取任务中，主要事件类别有犯罪、搜查、抓捕、审问等，一旦发生了诈骗事件（因），将引起公安机关的检查和搜查事件（果），抓住犯罪嫌疑人后，立即对其审讯（跟随）。事件特征项的构建包括基于本体的特征压缩与扩充两项工作，前者主要指同义词合并，后者主要指基于事件本体补充句子文本中缺失的事件元素特征（如："审讯"类犯罪事件"他接受了审讯"中，"警察"为事件的主要参与者之一）。

基于本体的特征项压缩或扩充过程使得事件句子的特征向量的语义完备性和准确性，但扩充特征时应结合上下文的其他事件类，又因其提供事件类别及关系的定义，因此该方法更适合话题追踪任务。

2.2 触发词法

触发词法[9]也称事件关键词法，通过对事件句的统计分析后发现，出现某类术语或词汇的句子文本中事件句的概率非常高，如：出现"发生"、"袭击"、"研制"、"生产"、"举行"、"举办"、"开幕"等动词词汇的句子"今年三月份在地铁3 号线发生乘客猝死事件"、"周杰伦将于2010 年6 月11 日在台北小巨蛋举办周杰伦超时代演唱会"等基本为事件句，因此通过建立事件触发词词典进行事件抽取能取得较好的效果。

建立事件触发词词典的方法一般有两种：一是由领域专家基于领域经验手工构建，此种方法过于依赖专家经验，适合事件句的触发词量变化不大的应用；二是通过已有事件句中词汇的分析统计，提取事件句触发词，相对手工构建，此种方法更容易提高触发词的查全率。系统应用触发词字典时也有两种方法，一是建立触发词库，由程序自动读取，此种方法方便灵活，便于维护；二是直接将触发词写在程序代码中，此种方法较呆板，一旦需要对触发词进行增减即需要修改程序。

2.3 基于(领域)本体的事件抽取算法

领域本体基于专业领域概念，定义概念的属性、方法

及概念间的关系[17]。这些概念并非仅为事件，甚至基本不涉及事件。当研究对象为某领域的事件时，事件即为该领域的概念，而概念间的关系即为事件间的关系，但基本不涉及事件实体元素之间的关系。当前，主要的事件抽取算法基本都需要经历分词、词性标注、去噪、特征提取等预处理环节。基于领域本体的事件抽取算法，主要利用本体实例库中的命名实体及其间关系等丰富的语义信息，去除无用的冗余信息、合并相关联的词使之成为领域实体、合并同义概念（如：“吸毒者”和“瘾君子”）减少特征项，提高预处理的效果，因此领域本体常与模式匹配、触发词、机器学习算法或语义分析算法结合使用。如：某收购事件中的“收购金额50 000 00 US dollars”一般被切分为“收购”、“金额”、“50”、“000”、“00”、“US”、“dollars”，经过本体实例查询，合并为“收购金额”、“50 000 00”、“US dollars”。

领域本体的语义丰富性决定了基于领域本体的事件抽取方法能够同时完成事件句识别、事件分类、事件元素提取3 项工作，但因构建工程复杂，当前成熟的领域本体库普遍缺乏，实际普及应用需要一定的时间。

3 事件抽取算法的效果及适用性分析

为了帮助研究者更快更方便地了解各种事件抽取算法的适用性，本文对各个算法的效果做一定的比较分析。使用中文检索式：题名或关键词= 事件抽取 OR 信息抽取 OR 话题识别，外文检索式：（Event Extraction）or（Information Extraction），分别在中国知网、Engineering Village 中检索2000 年以来的中外文文献，筛选中文核心期刊及EI 收录的研究事件抽取的论文（即判断句子文本为事件句），标注事件抽取算法、领域、语料、所选文本特征类型或特点，对各算法进行分析统计结果如下：

3.1 事件抽取主要算法

2000 年以来，对中文进行事件抽取研究主要以基于模式匹配的事件抽取算法和触发词法为主。进一步选择包含事件（准确率（P）、召回率（R））或效果（F）指标的文献19 篇中文核心期刊或EI 收录文献（以下简称“有效文献样本集合”）中使用较多的算法总体情况进行分析，结果如图2：

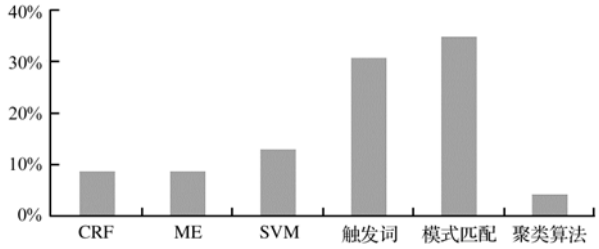

图2 文献中使用的事件抽取算法占比分布图

图2 表明，在本文所统计的有效文献集合中，触发词和模式匹配方法的使用频率最高。实际上，为了提高事件抽取的效果，实际应用中存在将两个或两个以上的算法进行组合应用，即首先使用某算法（如先使用模式匹配法过滤非事件句后再用SVM 法识别事件句，使用KNN 算法提取触发词再用触发词法抽取事件）进行处理，再使用其他算法进一步处理以提高事件抽取的准确度。

3.2 事件抽取语料库分析

本文收集的相关文献所使用的语料均为中文语料，其中，中文文献大多直接选取一定数量的相关领域语料，主要涉及特定领域新闻资讯（如金融资讯、突发事件等）、军事、个人简历等信息，这类语料一般具有很强的领域背景，表明已有研究侧重利用事件抽取算法解决实际应用问题。而“有效文献样本集合”中的外文文献的中文语料库近39 % 是“ACE 2005 Chinese corpus”，近23 % 文献的语料针对新闻，如人民网（www.people.com.cn）、百度新闻等知名新闻类网站，而非某一类具体行业的网站。

3.3 事件抽取算法效果分析

进一步对“有效文献样本集合”文献中所涉及的不同领域事件抽取算法中表现最优的算法以及同一类算法在不同领域抽取事件的效果进行分析统计，结果见表1、表2、表3、表4。

表1 算法整体统计效果分析

数 据	准确率P	召回率R	F 指数
最大值	0.961	0.9891	0.973
最小值	0.3818	0.4144	0.5203

表2 不同抽取算法表现最好领域对照表

基本算法	领 域	最高F	最高P	语料(测/训)	测试方法	文献号
模式匹配	犯罪信息		0.850	1/1	开 放	[16]
触发词法	法 律	0.8423	0.7503	2	开 放	[18]
CRF	个人简历	0.973	0.961	3，4/3	开 放	[19]
ME	突发事件	0.8426	0.8568	5/5	封 闭	[20]
SVM	金融资讯	0.6915	0.8783	6/7	开 放	[21]

表3 不同领域表现最好的算法对照表

统计领域	基本算法	F 值	P 值	语料(测/ 训)	测试方法	文献号
金融资讯	模式匹配+ SVM	0. 6915	0. 8783	6/7	开 放	[21]
突发事件	ME	0. 8426	0. 8568	5/5	封 闭	[20]
门户网站	模式匹配	0. 8079	0. 7859	8/9	开 放	[4]
医疗领域	触发词	0. 5312	0. 7396	10/10	封 闭	[22]
军事领域	模式匹配	0. 8144	0. 7776	11/12	开 放	[23]
个人简历	CRF	0. 973	0. 961	3，4/3	开 放	[19]
ACE 会议	触发词	0. 7840	0. 818	13/14	开 放	[24]
犯罪信息	模式匹配		0. 850	1/1	封 闭	[16]
法律信息	触发词	0. 8423	0. 7503	2	开 放	[18]
交 通	CRF	0. 5203	0. 3818	15/15	封 闭	[25]

表4 相同领域不同抽取算法F 值对照表

领 域	最高F 值	准确率P	主算法	特 征	语料(测/ 训)	测试方法	文献号
金融资讯	0. 6915	0. 8783	模式匹配+ SVM	关键词、事件触发词、上下文	6/7	开放	[21]
突发事件	0. 8426	0. 8568	ME	上下文特定位置的语言成分	5/5	封闭	[20]
门户网站信息	0. 8079	0. 7859	模式匹配	概念首义原、句子语言单位	8/9	开放	[4]
	0. 682	0. 714	模式匹配	利用树的平行结构特性抽取表格页面事件，使用总结模式抽取详情页面事件	16/16	封闭	[26]
	0. 7044	0. 6956	SVM	刻画一个事件发生的有代表性的特征，构成候选事件实例表示，构造二元分类器对事件实例与非事件实例进行自动识别	17/17	封闭	[27]
	0. 73	0. 7	触发词	触发词权重排序，选前3 个	18/18	封闭	[5]
医疗领域	0. 5312	0. 7396	触发词	上下文信息、蛋白质——触发词对特征、根路径特征	10/10	封闭	[22]
军事领域	0. 8144	0. 7776	模式匹配	领域动词、领域介词和领域实体	11/12	开放	[23]
个人简历	0. 973	0. 961	CRF	who，when，where，what，how	3，4/3	开放	[19]
ACE 会议	0. 7976	0. 6641	聚 类	实体类型、各实体间的关系	19/19	封闭	[28]
	0. 7840	0. 818	触发词	语义特征、上下文特征、位置特征	13/14	开放	[24]
	0. 682	0. 699	触发词	事件触发词内部的语义结构和核心词素	20/21	开放	[29]
	0. 7785	0. 7335	模式匹配	事件与非事件的特征	22/22	封闭	[30]
	0. 6230	0. 5727	触发词+ SVM	根据触发词的特征（Information，NE Context，Lexical Context） 来识别事件句	23/23	封闭	[31]
	0. 702	0. 714	触发词+ ME	事件触发词、事件类别	24/25	开放	[32]
		0. 6815	模式匹配	根据句子长度、位置等因素计算出句子SS_i值，然后取N－Best 的句子作为候选事件句	26/26	封闭	[33]
犯罪信息		0. 850	模式匹配	ICTCLAS3. 0，TF ＊IDF 词频选择文本向量特征词，特征词基于犯罪事件本体同义词合并且补充缺失特征（少于50 个字，含有至少3 个事件元素)	1/1	开放	[16]
法 律	0. 8423	0. 7503	触发词	事件触发词同义词词典	2	开放	[18]
交通信息	0. 5203	0. 3818	CRF	先构建若干特征模板，进行语料训练后得到合适的CRF 特征模型	15/15	封闭	[25]

语料列表：

1．the ‘People’s Daily’ in January 1998 。

2．自建语料，包含1 268 篇文档，其中共有3 008 个句子。

3．互联网上下载的785 份英文个人简历。

4．785 份简历文档和785 份非简历文档（选自路透社RCV－1 英文文档集合）。

5．自有语料库中挑选5 起国内外重大突发事件。

6．从百度新闻爬取的577 篇公司收购相关新闻，其中的76 篇。

7．从百度新闻爬取的577 篇公司收购相关新闻，其余的319 篇。

8．实验数据选自1998 年《人民日报》语料库的部分数据和实验室收集的数据共166 篇相关领域的文章和信息，共计1 253 条例句，随机选取400 条作为测试数据。

9．实验数据选自1998 年《人民日报》语料库的部分数据和实验室收集的数据共166 篇相关领域的文章和信息，共计1 253 条例句，其余853 条信息。

10．BioNLP'11 的GE 语料。

11．搜狐军事频道与武器装备有关的语料1～300 共300 篇。

12．搜狐军事频道与武器装备有关的语料301～600 共300 篇。

13．ACE2005 Chinese corpus，60 篇测试文档。

14．ACE2005 Chinese corpus，573 篇训练文档。

15．2000 条微博信息（2013 年8 月1 日到8 月7 日）和2013 年8 月13 日到17 日的Revolution Daily 上的交通信息。

16．挑选15 个事件信息发布网站，从中挑选10 个表格页面和50 个详情页面共900 个页。

17．从网上下载的500 篇新闻文本。

18．5 000 条www.chinatimes.com 新闻报道，时间从2002 年3 月到5 月。

19．ACE2005 Chinese corpus。

20．ACE2005 Chinese corpus 剩下的33 篇文档。

21．ACE2005 Chinese corpus 中的随机挑选567 篇文档。

22．ACEGrop2005。

23．ACE 2005 Chinese training corpus。

24．ACE 2005 Chinese corpus，剩下的66 篇文档作为测试语料。

25．ACE 2005 Chinese corpus，随机挑选567 篇文档作为训练语料。

26．ACE2005 training dataset 和取自南方周末2007 年11 月版的864 篇新闻报道。

事实上，语料的质量和结构不同都会影响到算法的效果，但本文认为不会产生根本性的影响，但测试模式对算法效果的评价有较大影响，本文的“有效文献样本集合”中，采用封闭测试的占52.63％，因此实际选择算法时还应考虑测试模式对算法效果的影响。

表1 显示各领域的事件抽取算法准确率最高可达0.961[19]（利用CRF 算法提取个人简历中的事件）；最低为0.3818[25]（利用CRF 算法提取交通领域交通事故事件）；召回率最高可达0.9891[28]（利用模式匹配法抽取ACE 会议语料中的事件）；最低为0.4144[22]（基于触发词算法抽取生物医学领域事件）。召回率略高于准确率；综合指数F 值最高为0.973[19]，最低为0.5203[25]。

表2 显示了在所参与调研的文献中，各抽取算法表现较好的领域。模式匹配算法在犯罪信息领域抽取事件的准确率P 值可达0.85；触发词算法在法律领域的综合效果较好，F 值可达0.8423，准确率P 值可达0.7503；CRF、ME、SVM 3 个算法中，CRF 在个人简历领域取得的效果较好，F 值可达0.973，其次是ME 算法在突发事件领域F 值和P 值分别可达0.8426、0.8568，而SVM 算法在金融领域抽取事件时，过滤掉一些明显非事件句后，进一步识别事件句[21] P 值虽然达0.8783，但召回率较低，F 值仅为0.6915。除突发事件领域中应用最大熵方法（ME）抽取事件为封闭测试外，其余均为开放测试，因此ME 算法的开放测试效果可能有所降低。

表3 显示了不同的领域中所采用的各种事件抽取算法中，效果较好的事件抽取算法。从中可以看出，在金融资讯事件抽取领域，将模式匹配算法与SVM 算法组合使用取得的综合效果（F＝0.6915）不及在门户网站、军事领域信息中抽取事件的效果好（F＝0.8079、0.8144），但准确率（P＝0.8783）却优于上述两个领域的事件抽取（P＝0.7859、0.7776），即抽取金融资讯事件的查全率较低，这可能与金融资讯中往往同时报道多个关联事件，部分事件元素缺失影响判别效果的缘故；使用模式匹配方法抽取犯罪信息中的事件虽然也取得了85％的准确率，但因为采用了封闭测试，因此实际开放测试效果有待考证；触发词方法抽取医疗领域事件的封闭测试综合效果（F＝0.5312）不及ACE 会议[24]及法律事件抽取的开放测试的效果好（F＝0.7840、0.8423），但准确率的差距不太大（分别为0.7396、0.818、0.7503），文献［24］采取了KNN 算法提高了触发词的质量，所取得的准确率最好，但查全率也不是很好，法律信息中所含有的动词等触发词相对较明确，因而查全率较高；CRF 方法在个人简历与交通信息事件抽取的综合效果及准确率差距都很大，F 值分别为0.973 和0.5203，这可能是因为个人简历事件相比交通事件的特征模板更容易构建的缘故；另外，虽然采用最大熵方法（ME）抽取突发事件信息的封闭测试效果也不错，但仅略高于触发词法抽取法律信息的开放测试效果，因而其实际效果不一定好。

表4 给出了在所参考的文献中采用的各种抽取算法以及它们所取得的效果值，其中在金融资讯、突发事件、医疗、军事、个人简历、犯罪信息、法律、交通信息领域所涉及的“有效文献”数分别仅为1 篇，所以重点分析以门户网站信息和ACE 会议语料为研究对象的有效文献。抽取门户网站事件信息的研究中，已有文献分别涉及模式匹配、SVM 和触发词方法，其中基于概念首义原、句子语言单位

的模式匹配方法取得的效果最好（F = 0.8079），但概念首义原的思想与触发词思想有异曲同工之处。ACE 会议语料事件抽取的研究中，虽然采用触发词法抽取事件取得的综合效果（F = 0.784） 仅略高于模式匹配所取得效果（F = 0.7785）、略低于聚类算法的效果（F = 0.7976），但后两者均为封闭测试。综合门户网站及ACE 会议语料事件抽取的研究成果，可以发现，触发词方法抽取事件的效果优于模式匹配方法。

3.4 算法时间复杂度分析

算法时间复杂度主要指完成某一任务所需要的时间（也称所耗费的时间），一般而言，算法越复杂，其时间复杂度越高，推广应用的难度越大。但随着计算机硬件性能的大幅度提高，算法的时间复杂度似乎变得不太重要。在本文所涉文献中，大多侧重提高算法的P、R、F 值，很少关注抽取任务的实际耗时情况。事实上，CRF 与SVM 方法的特征规模（向量维度） 均会大大影响运行效率[34-35]；语义角色标注方法需要句法分析技术[3]，其运行效率因句法分析方法的不同而不同。基本触发词方法由于运算简单，运行效率相对较高，但若结合KNN 等算法提取触发词，其运行效率将会受到一定的影响[24]。

4 研究展望

通过前述分析可以发现，当前中文文本事件抽取的实践主要集中于金融资讯、会议信息、突发事件、个人简历等来自网页、微薄或微信等自媒体信息或军事法律等专业领域文本的事件抽取，仍采用SVM、CRF、ME、模式匹配（非机器学习类）、聚类等在命名实体识别领域相对较传统的算法为主，除应用CRF 抽取个人简历类格式规范的文本中事件取得优异的效果外，F 值普遍低于0.9。

统计表明，事件一般含有触发词，前述分析也表明触发词方法的综合效果较优于单纯的模式匹配算法。虽然触发词方法的总体准确率不高，但在ACE 会议综合性语料中结合语义特征、上下文等特征所取得的准确率超过了0.8，在专业领域文本中应该可以进一步提高。另一方面，从基于ACE 会议语料的实践结果看，对于综合性强的领域文本，事件触发词的提取存在查全率不高的问题。笔者对一些专业领域文本中事件触发词分析后发现，专业领域文本中的事件触发词的数量基本有限。此外，触发词方法能够较大程度地过滤非事件句，大大提高后续处理的效率。因此，笔者认为，对于专业领域文本事件抽取问题，优先选择触发词方法。

参 考 文 献

[1] https：// www.ldc.upenn.edu/collaborations/past - projects/ace [EB].

[2] 高强，游宏梁. 事件抽取技术研究综述[J]. 情报理论与实践，2013，36（4）：114-117，128.

[3] 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.

[4] 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.

[5] Bao Jiana，Li Tingyu，Yao Tianfang. Event Information Extraction Approach based on ComplexChinese Texts [C] //IEEE Computer Society. 445 Hoes Lane - P. O. Box 1331，Piscataway，NJ 08855 - 1331，United States：IEEE Computer Society，2012：61-64.

[6] 张珏，刘云. 话题识别与跟踪技术的发展与研究[J]. 北京电子科技学院学报，2008，16（2）：77-79.

[7] 张晓艳. 新闻话题表示模型和关联追踪技术研究[M]. 北京：解放军出版社，2013.

[8] 滕青青，吉久明，郑荣廷，等. 基于文献的中文命名实体识别算法适用性分析研究[J]. 情报杂志，2010，（9）：157-161，169.

[9] 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.

[10] Zhang Xiuhong，Gong Zhe. Information extraction based on event driven fromtemplate web pages [C] //Springer Verlag. Tiergartenstrasse 17，Heidelberg，D - 69121，Germany：Springer Verlag，2013，211 LNEE：515-523.

[11] Jiang Bo，Zhu Mengxia，Wang Jiale. Ontology - based information extraction of crop diseases on Chinese web pages [J]. Academy Publisher，2013，8（1）：85-90.

[12] 潘超，杨良怀，龚卫华，等. 模式匹配研究进展[J]. 计算机系统应用，2010，（11）：265-277.

[13] Bao Jiana，Li Tingyu，Yao Tianfang. Event Information Extraction Approach based on Complex Chinese Texts [C] //IEEE Computer Society. 445 Hoes Lane - P. O. Box 1331，Piscataway，NJ 08855 - 1331，United States：IEEE Computer Society，2012：61-64.

[14] 黄发良，钟智. 用于分类的支持向量机[J]. 广西师范学院学报：自然科学版，2004，21（3）：75-78.

[15] 李丽双，党延忠，廖文平，等. CRF 与规则相结合的中文地名识别[J]. 大连理工大学学报，2012，（2）：285-289.

[16] Li Cunhua，Hu Yun，Zhong Zhaoman. An Event Ontology Construction Approach To Web Crime Mining [C] //IEEE Computer Society. 445 Hoes Lane - P. O. Box 1331，Piscataway，NJ 08855 - 1331，United States：IEEE Computer Society，2010，（5）：2441-2445.

[17] 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.

[18] Ding Xiaoshan，Li Fang，Zhang Dongmo. Causal Relation Recognition between Sentence - based Events [C] //IEEE Computer Society. 445 Hoes Lane - P. O. Box 1331，Piscataway，NJ 08855 - 1331，United States：IEEE Computer Society，2011：1688-1693.

[19] 李劲，张华，辜希武. 面向个人简历的事件抽取和检索框架[J]. 计算机科学，2012，39（7）：154-160，174.

[20] 韩永峰. 网络新闻突发事件信息抽取技术研究[D]. 郑州：中国人民解放军信息工程大学，2012.

[21] 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.
[22] 徐谦. 基于半监督方法的生物医学事件抽取的研究[D]. 大连: 大连理工大学, 2013.
[23] 张练. 领域信息抽取相关技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2010.
[24] Fu Jianfeng, Liu Zongtian, Zhong Zhaoman, Shan Jianfang. Chinese event extraction based on feature weighting [J]. Asian Network for Scientific Information, 2010, 9 (1): 184 – 187.
[25] Xiong Jiaxi, Hao Yonggang, Huang Zheng. Civil Transportation Event Extraction from Chinese Microblog [C] //IEEE Computer Society. 445 Hoes Lane – P. O. Box 1331, Piscataway, NJ 08855 – 1331, United States: IEEE Computer Society, 2013: 577 – 582.
[26] 何一鸣. 网页事件信息抽取研究[D]. 上海: 复旦大学, 2010.
[27] 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.
[28] Lin Ruqi, Chen Jinxiu, Xu Honglei, et al. A multi – information fusion approach to unsupervised Chinese event extraction [C] //IEEE Computer Society. 445 Hoes Lane – P. O. Box 1331, Piscataway, NJ 08855 – 1331, United States: IEEE Computer Society, 2010.
[29] Li Peifeng, Zhou Guodong. Employing morphological structures and sememes for chinese eventextraction [C] //COLING 2012 Organizing Committee. Powai, Mumbai, 400076, India: COLING 2012 Organizing Committee, 2012: 1619 – 1634.
[30] Yang Xiaofang, Chen Jinxiu, Lin Ruqi. Event detection and type recognition using self – training [C] //Springer Verlag. Tiergartenstrasse 17, Heidelberg, D – 69121, Germany: Springer Verlag, 2011, 227 CCIS (PART 4): 9 – 16.
[31] Wang Wei, Zhao Dongyan, Zou Lei, et al. Extracting 5 W1H event semantic elements from Chinese online news [C] //Springer Verlag. Tiergartenstrasse 17, Heidelberg, D – 69121, Germany: Springer Verlag, 2010, 6184 LNCS: 644 – 655.
[32] Li Peifeng, Zhu Qiaoming, Diao Hongjun, Zhou guodong. Joint modeling of trigger identification and event type determination in chinese event extraction [C] //COLING 2012 Organizing Committee. Powai, Mumbai, 400076, India: COLING 2012 Organizing Committee, 2012: 1635 – 1652.
[33] Wang Wei, Zhao Dongyan, Wang Dong. Chinese News Event 5 W1H Elements Extraction using Semantic Role Labeling [C] //IEEE Computer Society. 445 Hoes Lane – P. O. Box 1331, Piscataway, NJ 08855 – 1331, United States: IEEE Computer Society, 2010: 484 – 489.
[34] 冯元勇, 孙乐, 李文波, 等. 基于单字提示特征的中文命名实体识别快速算法[J]. 中文信息学报, 2008, 22 (1): 104 – 110.
[35] 李丽双, 黄德根, 陈春荣, 等. SVM 与规则相结合的中文地名自动识别[J]. 中文信息学报, 2006, 20 (5): 51 – 57.

(本文责任编辑: 马 卓)

第七届“信息资本、产权与伦理国际学术研讨会(ICPE – 7)”在东北师范大学召开

2015 年12 月1 日, 第七届“信息资本、产权及伦理国际学术研讨会”在东北师范大学净月校区开幕, 来自中国、美国、日本及中国台湾的专家学者和师生代表280 多人参加了会议, 共同研讨了在移动互联网和大数据时代, 信息安全、数据隐私等问题。本届研讨会由东北师范大学计算机科学与信息技术学院主办,《图书情报工作》杂志社和《现代情报》杂志社协办。

开幕式由东北师范大学计算机科学与信息技术学院副院长王战林主持, 东北师范大学副校长韩冬育、东北师范大学计算机科学与信息技术学院院长马志强致欢迎辞, 台湾世新大学资讯传播学系林志凤博士、北京大学信息管理系主任李广建教授, 吉林省图书馆馆长赵瑞军研究馆员、吉林省科技信息研究所所长郝屹研究员分别致贺辞。

在主会场主旨发言环节, 台湾世新大学资讯传播学系林志凤博士以信息素养教学小组的研究为主题进行了陈述; 美国南佛州大学信息学院葛瑟吉博士围绕信息素养有关问题进行了讲述; 日本鹤见大学图书馆、档案和信息系长塚隆教授针对云计算服务进行了精彩的演讲。在两个分会场, 来自国内外20 余位参会学者分别就信息社会公共图书馆的服务、信息流畅性、电子商务构建、公共文化政策、数据公共图书馆数字资源版权等热点问题发表了演讲, 参会师生在相关问题上进行了深入的探讨与交流。

为了提高图情系学生的科研水平和创新能力, 促进学生之间的学术交流, 会议期间还举办了“大学生学术成果获奖作品海报展”, 共有34 位同学展示了自己的作品, 其中11 位同学在会上做了发言。

为期两天的研讨会于12 月02 日圆满落下帷幕, 与会代表对本届研讨会的成功举办表示祝贺, 也对2016 年将在台湾世新大学举办的“ICPE – 8”表达了期待。