·应用研究·

基于触发词优先级的事件抽取研究

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

〔摘 要〕本文将触发词分为时间类和非时间类，对触发词提取算法进行改进，以一定量导电塑料行业新闻为基础语料构建两类触发词词表，并采取时间类触发词优先的事件句识别策略。基于该触发词词表对导电塑料和太阳能行业新闻语料进行事件句识别算法有效性实验，开放测试的召回率和准确率分别超过 98% 和 95%。该结果表明：将触发词进行基于时间特性的分类，并优先使用时间类触发词提取事件句，能取得显著的效果。

〔关键词〕事件句；抽取；触发词优先
DOI:10.3969/j.issn.1008－0821.2016.12.009
〔中图分类号〕TP391 〔文献标识码〕A 〔文章编号〕1008－0821（2016）12－0046－04

Event Extraction Based on the Priority of Trigger Words

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

〔**Abstract**〕This paper focused on efficient event sentences extraction algorithm. A trigger phrase extraction algorithm to extract events sentence based on time or nontime trigger word was presented. Some trigger words were extracted from industry news corpus about conductive plastics based on the algorithm，the opening beta on industry news corpus about conductive plastics and solar higher than 98% recall ratio and 95% accuracy ratio，which indicated the effectiveness of algorithm.

〔**Key words**〕event sentence extraction；trigger word priority；time trigger word；nontime trigger word

当前有关事件抽取研究中主要以特定行业新闻事件抽取研究为主，包括金融、军事、法律、交通等行业，所采用的方法包括模式匹配、触发词和本体方法，触发词方法的使用频率最高，且综合效果（微平均）较优于单纯的模式匹配算法[1]。

触发词也称事件关键词，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含有事件句的概率非常高，如：包含“发生”、“袭击”、“研制”、“生产”、“举行”、“举办”、“开幕”等动词的句子，“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获得事件句集合再进行事件抽取能取得较好的效果。

一般地，收集事件触发词的方法主要有两种：一是建立特定的触发词模型，通过已有事件句中词汇的分析统计，提取事件句触发词；二是由领域专家基于领域经验手工构建。手工构建方法主要依赖专家经验，一方面需要较长时间、较多人力投入，另一方面动态增加的海量事件文本也是一大挑战，触发词模型方法正好弥补了手工方法的缺陷，理论上更容易提高新增事件文本中触发词的查全率，但触发词模型的有效性仍有待进一步提升，如文献[2]，在构建相似词汇链的基础上，提出了一种基于词汇的 TFIDF 值、在文中的位置及相似词汇链长度的事件关键词模型，提取一定数量的事件关键词，对较大样本的实验有约 74% 的事件关键词为可接受的[2]，自动提取事件触发词仍需要进一步研究[3-5]。

收稿日期：2016－10－29
基金项目：国家社会科学基金项目“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）、教育部人文社会科学研究青年基金项目“面向语义出版的富语义模型构建与应用研究”（项目编号：15YJC870014）研究成果之一。
作者简介：吉久明（1969－），女，研究馆员，博士，硕士生导师，研究方向：知识集成。

触发词方法强调了触发词对提高事件抽取召回率的重要意义，尤其对于触发词表相对固定的行业领域事件抽取而言，能提高事件抽取的效率；模式匹配方法更突出句法或语义角色对提高事件抽取准确率的重要意义，对于句法或语义角色相对固定的事件抽取任务，同样能提高事件抽取的效率。因此，将触发词方法和模式匹配方法结合使用，理论上应能取得较好的召回率、准确率，也能保证一定的效率水平，但实际效果还有待提高，如文献[4]或文献[7]基于触发词及其邻近特征的动态权重的 KNN 算法或支持向量机判别事件句，准确率分别为 81.8% 和 87.8%[6-7]，且由于涉及特征选择，实际执行效率不理想。笔者曾设计了一种基于触发词句型模版的行业新闻事件句提取算法，由于过分强调“词”的形式，所选择的多数触发词所提取的句子准确率很低，如“研制”事件抽取准确率仅为 61.19%，因此需要事先编制大量的触发词句型模版，尽管如此，仍仅有少量触发词句型模版抽取事件的准确率超过 80%[8]。但笔者发现：若将“研制”改为“研制了”、“制造”改为“制造了”，则仅基于该两种触发词的事件抽取准确率即可提高到 90% 以上。

因此笔者认为，对于触发词表相对固定的行业新闻，提高基于触发词的事件抽取准确率的方法主要在于提高触发词的“专指性”，即利用中文语言的特征寻找具有很强的事件提示功能的触发词或词组合。而若简单地统计事件中的高频词，则不易获得这类触发词。故本文将进一步研究获得高“专指性”新闻事件触发词的有效方法，进而提高基于触发词的事件句抽取的准确率。

1 基于触发词的行业事件抽取

1.1 语料特征分析——以导电塑料行业新闻为例

以“导电塑料”及其同义词或近义词为检索词，利用搜索引擎检索相关导电塑料行业新闻，共获得 658 条语料记录。根据新闻撰写的规定和相关理论，新闻导语句一般会报导新闻的五个要素——何时、何地、何人、何事、何因[9]，这五要素正是新闻事件句的必备元素，而后续的文字则是对新闻事件的补充说明，因此理论上可从导语部分抽取事件。但事实上 658 篇语料中，事件句分散在导语及第 2、3、4、5、6 句，如下列语料的第②句为行业新闻事件句。

①生意社 6 月 8 日讯：想象一下，把一个 USB 端口插入一张纸，将它变为一个平板电脑。②这可能需要一段时间，但是北卡罗莱纳州立大学的研究人员已经按照这些想法去研究如何将传导纳米涂层应用于简单的纺织品，如梭织棉布，或者甚至一张纸。

因此，为减少大量非事件句对事件抽取的干扰，本文暂针对新闻语料的前 6 句研究新闻事件句触发词的提取。

1.2 行业新闻事件句触发词词表构建

事件即某时发生在某地的某事，或某人某时在某地参与（见证、实施、做出、取得了）了某动作（决定、成果），对于新闻事件而言，其中的时间元素必不可少，但由于语境的关系，常有事件句的时间元素被省略的现象。如下列语料：

①人民网上海 2 月 10 日电：（记者姜泓冰）防伪纸币、穿戴设备……柔性电子技术研究已成国际热点。②近日，复旦大学一团队……取得突破性进展……。③复旦大学信息科学与工程学院仇志军副教授……，相关论文已发表于 1 月 27 日出版的国际权威性学术期刊《自然－通讯》（Nature Communications）。

该语料中第②句、第③句均为事件句，两句讲述的是同一件事，但第③句的时间元素被省略了。若以“取得”或“提出”为触发词提取事件，则两句均被命中，需要进一步依据其出现的次序进行甄别；而若以“近日”作为新闻事件触发词提取事件句，则可忽略第③句。

同时，笔者注意到以下现象：①通过设定触发词准确率阈值的方法能提高整体的事件提取准确率。例如，限定在训练语料中的准确率超过 95% 的候选词为触发词，则整体准确率将超过 95%。②若过分强调触发词的准确率，则召回率一定会大大降低，但由于两个或两个以上的词组合召回事件句的准确率可能大于单个词召回事件句的准确率，故有时可适当通过使用词的组合形成触发词的方式在保证准确率的同时提高查全率。例如，同一子句中含有“据”和“报道”的句子为事件句的可能性大于含有“据”或“报道”的句子为事件句的可能性，含有“据”和“报道”的事件句可通过两词的组合召回。③将训练语料分为事件句和非事件句，选择召回事件句但不召回非事件句的词或词组合是保证事件句提取的准确率的有效途径。

因此，设计基于时间元素优先的事件触发词字典构建方法。算法描述如下：

Step 1 收集各种表示近期的时间类触发词，如：近日、年…月…日、今日、今天、刚刚、日前、前日、昨日、本周、上周、明天、昨天、正在、下周、周一、周二、周三、周四、周五、周六、周日、近期、最近、前不久、不久前、本月、上月、下月、下个月、上个月、今年等；

Step 2 收集一定数量的行业新闻语料；以“。”、“？”、“！”为分隔符将语料切分为句子；人工提取新闻事件句；

Step 3 将包含 Step 1 中词列表的事件句过滤掉；

Step 4 从 Step 3 中的新闻事件句中发现未列入 Step 1 中的时间类词汇，若该词召回新闻事件句的准确率大于给定的阈值 P，则添加到 Step 1 的列表中，并进行同义词扩充；

Step 5 重复 Step 4，直至无法提取新的时间类触发词；

Step 6 对前 6 句进行分词并统计词频（每句出现计 1 次）；

Step 7 选择词长大于 2，仅属于新闻事件句词表且频次大于等于 3（非同一事件）的动词或动名词列入非时间类触发词表，并进行同义词扩充;

Step 8 当上述触发词或触发词组合的召回率大于 R，算法终止，否则进入 Step 9;

Step 9 将事件句中不包含在非事件句中的 2 个词的共现对（即两词不同时出现在非事件句，但同时出现在同一事件句中），且共现频次大于等于 3 的 2 个词共现对列入非时间类触发词组合列表，直至召回率大于 R。

由于时间类触发词对于行业新闻事件句的提取具有较高的召回率和准确率，一般而言，应优先抽取含有时间类触发词的事件句，且每段新闻语料仅需提取一句即可。但对于一些含指代对象的语料，如语料 3:

复旦大学信息科学与工程学院副教授仇志军……取得突破性进展……。相关论文近日在《自然—通讯》上发表。

该语料所描述的事件与语料 2 描述的事件相同，若以“近日”为触发词，则提取到的事件句为“相关论文近日在《自然—通讯》上发表。”，该句主语为指代词“相关论文”，因此，更详细的信息需要使用其前句进一步补充。因此，设计以下行业新闻事件句提取方案:

Step 1 首先构建含有各种指代词的列表，如: 这、相关、他、她、该、我、上述等;

Step 2 抽取各语料中含有新闻事件触发词的句子各一句（记为句子 1，依触发词的次序而行，每段语料仅抽取一句）; 若基于“年…月…日”所得事件句的发生时间与当前系统日期的差大于 N 年，则继续以其后的触发词抽取事件句; 若所得句子的句首字为 Step 1 中的指代词，则将句子 1 的前句与句子 1 合并为 1 句。

其中 N 为参数，可根据实际需要进行设置。

2 实验结果与分析

2.1 新闻事件触发词提取

本次实验关注导电塑料制备行业的新闻事件，训练语料描述详见 1.1，触发词算法中的 R = P，均设为 95%，所抽取的时间类触发词及相应的准确率详见表 1。

表 1 时间类触发词准确率

序号	触发词	句子数	非新闻事件句	准确率	例 外
1	某月某日	242	5	97.94	无实质事件，仅有报道日期
2	某年某月	200	1	99.5	
3	今 年	105	2	98.1	
4	近日，不含最近日本	130	0	100	
5	最近，不含“年”、“月”、“天”	83	0	100	
6	日 前	64	0	100	
7	昨 日	25	0	100	
8	今 日	10	0	100	
9	刚刚，不含“刚刚开始”及“刚刚起步”	10	0	100	
10	最近，且与“年”间间隔超过 3 个汉字	7	0	100	
11	近 期	42	2	95.24	近期比较热的、近期发展

上述触发词从 658 篇语料中共抽取到 701 条事件句，其中: 37 句事件句主语部分存在“行业新闻事件句提取方案”Step 2 中的指代词，需要补充前 1 句; 不具新闻性的仅 5 句，这 5 句均含有事件发生的明确时间，故可根据系统时间进行过滤。进一步可以对上述触发词进行同义词扩充，如“周一”、“周四”可以扩展为: 周二、周三、周五、周日等。

为进一步获得语料中新闻事件句的非时间类触发词，抽取各篇的前 6 句共1 322句中的新闻事件句 114 句。依据触发词提取算法 Step 6 ~ 7，提取训练语料新闻事件句中准确率为 100% 的非时间类触发词及召回的句子数见表 2。

计算这些时间类和非时间类触发词对于前述训练语料的事件句抽取召回率和准确率分别为 93.48% 和 99.34%，故继续采用触发词提取算法 Step 9，提取训练语料新闻事件句中准确率为 100% 的组合类非时间类触发词组合召回的句子数见表 3。

表 2 非时间类触发词召回的句子数

触发词	召回句子数	触发词	召回句子数
推出 v	50	带领 v	16
开拓 v	4	披露 v	7
面对 v	3	来自 v	16
上涨 vn	8	开工 v	11
申报 v	4	获悉 v	8
展出 v	6	透露 v	6
找到 v	3	做出 v	6
凝聚 v	5	宣布 v	6
通过 v	69	正在 d	13
入选	5	调研 vn	5
商业化	5	改进 vn	5
增发 v	5	募集 v	5
扩建 vn	5		

表 3 非时间类触发词组合召回的句子数

触发词组合	召回句子数	触发词组合	召回句子数
开发 v 出 v	42	已 d 获得 v	10
将 d 会 v	10	制作 v 了 ul	10
将 d 研发 v	4	将 d + 变化 v	4
项目 n 名称 n	4	成果 n + 发表 v	23
创造 v 出 v	6	合作 vn 生产 vn	9
创造 v 了 ul	11	据 p + 报道 v	46
发明 v 了 ul	17	研究 n + 发表 v	32
研发 v 了 ul	9	研究 vn + 获得	17
介绍 v 了 ul	5	成功 v + 解决 v	13
研制 v 了 ul	11	研制 v 出 v	28

注：表 3 中不含“+”的词组合表示两个词组成的一个词，含有“+”的词组合表示两个词分别出现在同一句的同一部分（即不含“;”等子句分隔符）。

至此，上述时间类触发词表、非时间类触发词表的事件句召回率为 96.2%、准确率为 99.34%，达到算法终止条件。虽然本文选择了导电行业新闻语料作为提取事件句触发词词表的语料，但从表 3 可以看出，所得触发词均不具行业相关性，因此可以应用于不同行业的语料的事件句识别。

2.2 开放测试实验

随机收集了 2015 年以来导电塑料行业新闻语料 20 篇（简称开放语料 1）、太阳能电池行业新闻语料 20 篇（简称开放语料 2），共含 58 条新闻事件句，其中含时间类和非时间类触发词的事件句分别为 33 条、41 条，测试 3.1 中提取的触发词表提取事件句的效果。

2.2.1 时间类触发词事件抽取效果

使用前文提取的时间类触发词表 A 召回的句子数见表 4，准确率均为 100%。

表 4 时间类触发词召回句子数

序号	触发词	句子数
1	近 日	6
2	某月某日	6
3	某年某月	2
4	某年某月某日	3
5	某 日	3
6	昨 日	2
7	最 近	3
8	日 前	6
9	不久前	1
10	近 期	1
	合 计	33

2.2.2 非时间类触发词事件抽取效果

对 41 篇语料通过非时间类触发词表提取语料中的新闻事件句，召回率达到 98.27%，各触发词的抽取准确率均超过 95%，召回句子数见表 5。

表 5 非时间类触发词召回句子数及准确率

触发词	召回句子数（非事件句数）	触发词	召回句子数
通过 v	7（2）	发明 v 了 ul	2
宣布 v	7	取得 v + 进展 n	3
正在	2	成果 n + 发表 v	2
开发 v 出 v	9	研究 v/vn + 发表 v	2
创造 v 出 v	1	据 p + 报道 v	1
已 d 获得 v	1	成功 v + 解决 v	1
找到 v	3		

3 结束语

本文提出的事件触发词抽取技术与已有的触发词提取技术不同，在选择触发词时，更强调“专指性”。首先充分利用事件句必备的时间元素及事件触发词相对固定的特点，将触发词分为时间类和非时间类。利用一定数量的语料事件句抽取准确率很高的两类触发词或词组合字典，优先使用时间类触发词提取出多数事件句后，再以非时间类触发词或词组合提取余下的事件句，开放测试效果良好。

参 考 文 献

[1] 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.

[2] 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.

[3] 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.

[4] Pei－Feng Li，Qiao－Ming Zhu，Guo－Dong Zhou. Using compositional semantics and discourse consistency to improve Chinese trigger identification [J]. Information Processing & Management，2014，50（2）：399－415.

[5] 魏小梅，黄钰，陈波，等. 生物事件触发词识别方法研究 [J]. 计算机科学，2015，（10）：239－243.

[6] Fu Jianfeng，Liu Zongtian，Zhong Zhaoman，et al. Chinese event extraction based on feature weighting [J]. Asian Network for Scientific Information，2010，9（1）：184－187.

[7] 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.

[8] 陈锦辉. 导电塑料产业新闻事件抽取技术应用研究 [D]. 上海：华东理工大学，2015.

[9] 孙晓彦. 新闻写作技巧与范例 [M]. 北京：蓝天出版社，2011.

（本文责任编辑：马 卓）