doi:10.3772/j.issn.1002-0470.2015.12.007

面向安全事件新闻的时间抽取与转换①

李明月②* 王树鹏③* 王海平* 付 戈**

(*中国科学院大学 信息工程研究所 北京 100093)

(**国家计算机网络应急技术处理协调中心 北京 100029)

摘 要 阐述了事件新闻文本的时间信息抽取与处理对事件研究的重要性,研究了安全事件新闻的时间抽取与转换。考虑到目前采用的基于时间抽取规范 TIMEX2/3 和机器学习的抽取处理方法得到的时间信息缺少完全统一的形式,在安全事件的舆情发现及分析等场景下很难直接利用的问题,提出了针对安全事件新闻中的时间信息抽取与转换方法。该方法首先对安全事件的新闻根据时间的分类分别对不同形式的时间进行抽取,然后利用六大时间转换算子及时间冲突处理算子输出其时间的年月日时分秒的统一格式。试验表明,采用该方法的抽取结果与使用条件随机场(CRF)的方式进行抽取的结果相差不大,并且在时间转换上的正确率达到 90% 以上。

关键词 新闻,事件,时间抽取,时间转换,舆情分析

0 引 言

在自然语言中,时间信息是一种重要的信息,它是一个事件的重要组成部分,是关键信息的载体。时间信息的抽取和处理是当前自然语言处理中的一个重要研究方向。在信息抽取、主题发现、知识问答系统及 Web 分析中,时间抽取都起到至关重要的作用。特别是在新闻媒体类型的文本中,时间信息是其重要的组成部分。人们阅读新闻报道时总是要把其中的内容和时间信息联系起来,通过时间信息了解到一个事件的开始、结束以及事件发生的频率,把握一个事件的全过程,这对于研究公共安全事件意义重大。我们可以通过特定的时间抽取算法找到新闻中危害公共安全事件的时间属性,从而了解危害公共安全事件在时间维度上的发展状态,找到事件的传播机制,帮助有关部门及时做好防护及预测。传统的新闻文本时间抽取只是单纯的基于时间属性在文本中找到有用的时间信息。但是新闻文本描述的种类多样化,时间也具有多样性,除了在文本中找到时间表达式(时间表达式包含日期、时间、一段时间、指代时间的表达式以及事件描述的时间词等)之外,很难根据先验条件将文本中的时间进行统一表示。而在现实中,特别是在安全事件的新闻文本中,找到事件发生的时间尤为重要,因此时间的抽取应包含两个任务,分别称为时间的抽取和时间的转换。本文研究了安全事件新闻文本的时间表达式的抽取与转换。之所以研究描述安全事件的新闻文本,一是由于近几年安全事件频发,找到安全事件的时间信息对于安全事件的舆情分析至关重要;二是新闻作为事件传播的重要媒体及主要传播媒介,对于事件的描述准确度和可信度较高;三是新闻文本在对安全事件的时间进行描述时会使用特殊的且比较正式固定的时间描述词,在时间的抽取转换上可以最终对时间进行更精确的表示,为后续的事件传

① 国家自然科学基金(61271275,61202067),863 计划(2012AA013001, 2013AA013205,2013AA013204)和北京市科技计划基金(Z131100001113034, Z131100001113034612020 67)资助项目。

② 女,1991 年生,硕士;研究方向:大数据处理,数据分析;E-mail: limingyue@iie.ac.cn

③ 通讯作者,E-mail: wangshupeng@iie.ac.cn

(收稿日期:2015-04-22)

播机制研究奠定基础。本研究主要包含两大步骤,即时间的抽取和时间的转换,时间的转换又包含六大时间转换算子以及时间冲突处理算子,后者至关重要。

1 相关工作

国外对于时间抽取的研究主要针对英文文本语料。早在 1974 年,Reichenbach 首次把时间系统引入到自然语言处理中。MUC-6(1995)(1995 年举行的第 6 次信息理解会议(Message Understanding Conference))正式将时间表达式的标注方案作为一个独立的问题提出,MUC-7(1998)则将该问题做出了扩展,提出了完整的时间表达式标注方案,并用 TIMEX[1] 标签对该方案进行了定义。2001 年,Ferro 等人[2,3] 在"潮汐计划"中将 ISO 的标准日期格式规范引入时间表达标注,并定义为 TIMEX2 标签。2007 年,数据评测开发测试集 SemEval2007 将时间表达式识别与规范化(time expression recognition and normalization, TERN)任务加入其中,本次测评对 TIMEX2 进行深化并提出 TIMEX3[4]。

Mani 于 2003 年采用决策树识别方法在 TimeML 语料库上进行时间事件关系识别,并达到 75.4% 的准确率[5]。2006 年,Mani 采用了最大熵分类器使识别效果进一步提升[6]。Lapata 等人通过时间词语子句概率关系建立 Bayes 统计分类模型,在 TimeBank 语料上进行训练,进而识别出句子的时间关系[7]。Chambers 提出了一种基于纯文本的时间事件关系识别方法,该方法首先采用机器学习来自动标注事件属性,然后进行时间事件的关系识别[8]。Abe 等[9] 和 Chklovski 等[10] 都采用模板和标注联合训练分类器处理时间关系的识别问题。2009 年,Yoshikawa 实现了基于马尔科夫模型的时间事件关系识别[11]。2011 年,Mirroshandel 等利用支持向量机作为分类器识别时间事件关系[12]。

中文时间事件关系识别的研究起步比英文晚,研究也相对较少。其原因主要是专业语料库的匮乏,到目前为止还没有专业的语料库专门用于中文时间事件的关系识别。和英文相比,中文有自己独有的语言特性。更多的语言信息需要通过上下文语境或者助词体现。中文语言结构的不规律性增加了中文时间信息的处理难度。

2004 年,Gerber 等人制定了 TIMEX2 的中文标注草案[13]。2005 年自动内容抽取(automatic content extraction, ACE)会议对中文 TIMEX2 标注做了更详细的定义和说明。由于 TIMEX2 对中文表达能力不足,2008 年,清华大学苑春法等人在此基础上修改完善了 TIMEX2 的中文标注规范[14],并开发了一个中文的自动标注系统(CTAT),其标注结果也具有一定程度的错误,这主要是由于规则方法的不完善及语义知识的缺乏引起的,有待进一步改进和完善。中文时间表达式识别多使用机器学习方法来识别时间范围,然后使用规则的方法对时间信息规范化,但是特征的选择和规则的制定都很不健全。并且很多研究是只针对文本的时间抽取,即根据 TIMEX3 等规范将文本中的时间表示出来,然后时间仍划分为很多类别(TIMEX3 对时间表达式的分类如表 1 所示)。

表 1 TIMEX3 对时间表达式的分类

时间	举例
时间名词	今天,星期一等
时间名词短语	今天早上,星期二晚上等
时间形容词	现在,目前等
时间副词	最近等
时间形容词或副词短语	一小时前,两天后等

对于危害公共安全事件的舆情分析等应用而言,即使在新闻文本中将所有的时间类型已全部找到,在后续的事件舆情分析中,对时间比较方面的处理仍然比较棘手,因此本文进行时间抽取之后加入了六大时间处理算子和时间冲突处理算子,保证了事件新闻中的时间最终不会出现冗余,找到与事件发生最可能相关的单个时间,最终用年月日时分秒的日期格式来进行表示。

2 时间表达式抽取与转换

2.1 整体处理框架

面向新闻安全事件的时间抽取和转换的整体处理流程如图 1 所示。输入是安全事件的新闻文本,

首先采用正则表达式抽取方法和基于时间词典抽取方法来得到提取的时间表达式,再结合新闻的发布时间作为辅助时间,经过时间转换算子集得到时间的统一表示:年月日时分秒格式。

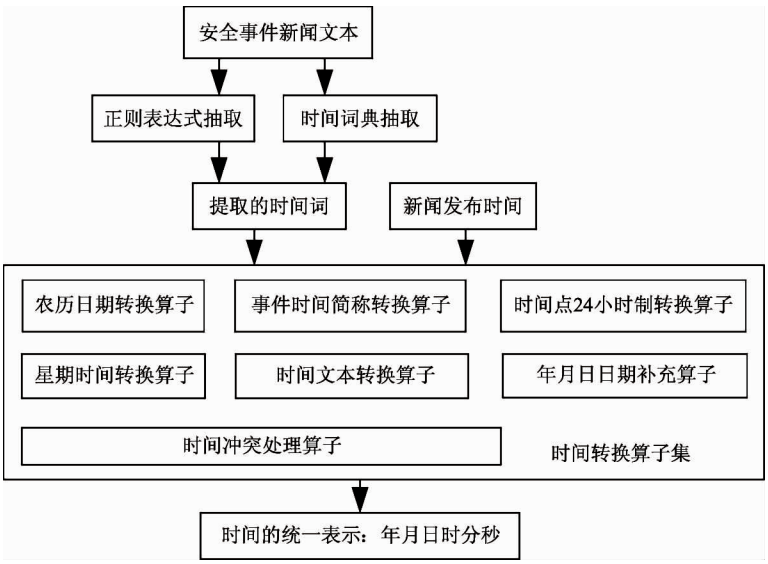

图 1 时间抽取与转换的整体处理流程

2.2 时间表达式抽取

我们处理的文本数据主要是针对公共安全事件的新闻数据集,由于新闻数据中时间属性的重要性,新闻在描述某个事件时一般都会加上时间这一重要信息。对于新闻表示事件时间的词语可以分为两类,如表 2 所示。

表 2 中,安全事件新闻的时间表示形式分为直接时间与间接时间,直接时间是指新闻中的时间本身就符合年月日时分秒的表示形式,因此对此类时间的处理较为简单,直接在文本上利用正则表达式进行抽取即可,并且对此类时间的准确性要求较高,若采用机器学习的方式可能会带来较大的误差,需要引入训练集处理,操作较为复杂。

表 2 安全事件新闻的时间表示形式

时间的分类	举例
直接时间	事件的年月日表示形式又可以分为较多的表示形式如 2013 年 05 月 06 日 08 时 23 分 12 秒,或者 2013/03/08 23:08,或者 2013-03-06,等等 农历的表示如正月十五 事件的时间简称如“3.8”事件
间接时间	用文本来表示的时间如清晨、今天早上、下午、今日晚上、不久前,等等

另外一类是比较复杂的间接时间表示,如“去年春天”、“今年夏天”、“前天”、“昨天晚上”,等等,类似于这样的词是无法通过正则表达式进行获取的,一般机器学习的方法[15,16]需要大量的训练集才能提高准确性,因此本文采用基于时间词典的方式,时间词典的构造首先基于先验知识及专家知识构造,然后利用 Google word2vec[17]进行时间词的同义词发现,得到同义词只取相似程度大于 0.8 的词语加入时间词典,然后再对不符合的词语进行筛选来保证时间词典的准确性。

2.3 时间表达式转换

通过以上方式得到提取的时间后,因为时间不具有统一的形式,在后续事件聚类与事件合并中会遇到无法比较的问题,因此需要利用数据中的参考时间即新闻的发布时间将时间统一表示成年月日时分秒形式。本文使用了六大时间转换算子以及关键的时间冲突处理算子来进行时间的转换,具体的转换流程如图 2 所示。下面说明各个算子的功能。

(1)农历日期转换算子

将抽取的农历日期表示形式转换为公历日期表示形式,具体到年月日。

(2)星期时间转换算子

将抽取的星期的时间表示形式,如周三或者星

期三等,转换为当天的年月日形式。

(3)事件简称转换算子

将抽取的类似“3.8”这样的事件简称转换为3月8日。

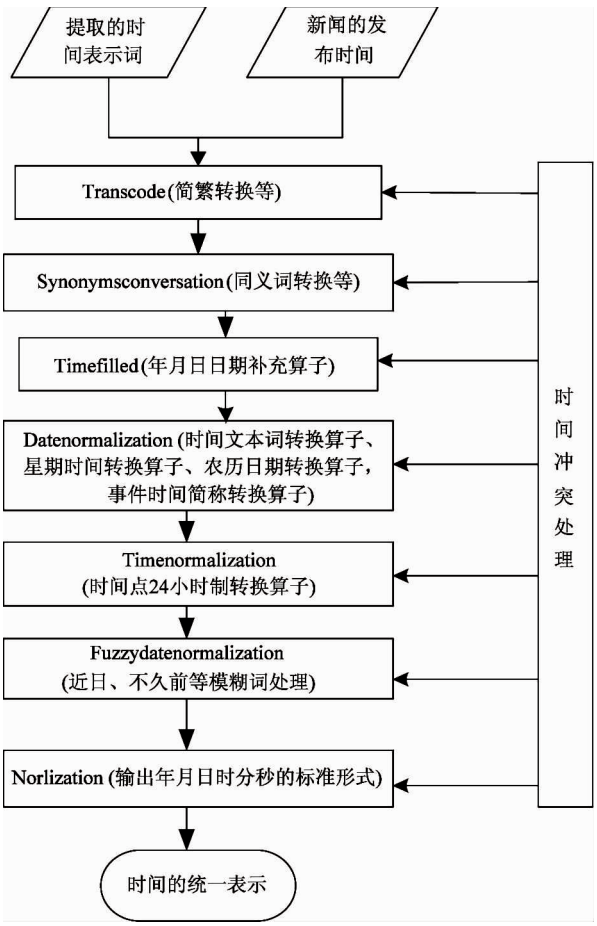

图2　时间转换的处理流程图

(4)时间文本转换算子

将时间文本词,类似于“今天上午”、“今天晚上”、“前天晚上”等,转化为对应的年月日表示。不同的时间文本词转换方式不同,需设定一定的规则。

(i)对“今天早上”、“今天上午”、“今日上午”、“清晨”等时间词,在新闻发布时间中直接取出年月日进行替换即可。

(ii)对“今天下午”、“今晚”、“今天晚上”、“下午”等时间词,除了在新闻发布时间中取出年月日替换外还要观察时间词表述时间是否是24小时制,如果不是需要转化成24小时制,如下午3时需转换成15时。

(iii)对“昨天上午”、“昨天”、“昨天早上”等时间词,需要在新闻发布时间中取出年月日之后进行天数减1处理,并且需要利用日期加减函数。如2013年08月01日减1天应该是2013年07月31日。

(iv)对“昨天晚上”、“昨晚”、“昨天下午”等时间词在新闻发布时间中取出年月日之后需要进行天数减1处理,并且要比较时间是否是24小时制。

(5)时间点24小时制转换算子

将“晚上7点”、“下午5点”等类似的时间点转换为19:00以及17:00等24小时制并且将“晚上”和“下午”等时间文本词去掉,直接去掉会导致晚上7点和上午7点成为一个时间点,而实际上这两个时间点相差24小时。

(6)年月日日期补充算子

(i)对只出现月日的时间通过在新闻发布时间取出年进行补充。

(ii)对只出现日的时间通过在新闻发布时间取出年和月进行补充,并且如果日大于新闻发布时间的日时,需要对月进行减1处理。

(7)时间冲突处理算子

提取转换后的时间会包含大于、等于两个时间的情况。时间冲突处理需要在时间的转换之前的每一步都进行检测,这样保证最后冲突处理结果的准确性。设定时间的优先级,优先级从高到低进行排序如下所示:

“月.日”;

具体到秒的日期;

具体到分的日期;

具体到时的日期;

具体到日的日期;

具体到月的日期;

具体到年的日期;

带有A的近似时间。

根据上述设定的时间优先级来指定冲突处理规则:

(1)若抽取得到的时间属于不同优先级,首先判断时间是否需要合并,然后再判断选择优先级高的时间作为该条新闻关联的事件的发生时间;

(2)时间的合并情况为时间之间可以互相补充或者可以判定属于同一个时间则需要进行合并;

(3)具体到相同优先级的时间,需要进一步比较新闻发布时间与各个时间的差别,如果差别小于1min 且是最开始的时间,则可判定该时间为新闻报道事件时间,然后选择与新闻发布时间第二接近的时间作为该新闻关联的事件时间;

3 试验结果

试验数据集采用第二届中国大数据创新大赛中危害公共安全事件的关联关系挖掘及预测赛题的数据集,该数据集是海量公司提供的来自互联网媒体报道和 UGC 数据的新闻和微博数据。

由先验知识和 word2vec 构造的时间词典如表 3 所示(展示部分)。

表 3 时间词典(部分)

昨天上午	昨天下午	今晨
昨日上午	昨日下午	清晨
昨日早上	下午	早上
⋮	⋮	⋮

本实验只采用了新闻数据,其数据内容包括公交车爆炸事件、暴恐事件及校园砍伤事件三类系列公共安全事件数据。分别在三类数据集上进行时间的抽取与转换。由于并不是每条新闻都会包含时间表达式。利用条件随机场(CRF)的方式对时间表达式进行抽取,条件随机场被认为是序列标注领域较好的机器学习算法,在数据集中抽取 5000 条记录,条件随机场的特征模板如表 4 所示。

进行人工标注后 CRF 算法得到的结果与本文提出的时间抽取与转换算法进行比较得到结果如表 5所示。

利用 CRF 抽取的结果与本文提出的时间抽取算法得到的结果在抽取率上相差在 1% 左右,差别不大。在准确率上,采用随机抽样的方式人工查看准确程度,每次抽取 20 条,循环 10 次,平均准确率 CRF 为91% 左右,本文提出的时间抽取与转换算法的时间抽取准确率在 92% 左右,而在时间复杂度上 CRF 训练时间是该方法的 3 倍。

表 4 条件随机场特征模板

特征编号	特征
1	当前词
2	当前词的标注
3	当前词的前面第一个词
4	当前词的前面第一个词的标注
5	当前词的前面第二个词
6	当前词的前面第二个词的标注
7	当前词的后面第一个词
8	当前词的后面第一个词的标注
9	当前词的后面第二个词
10	当前词的后面第二个词的标注

在转换后的时间记录中随机抽样,每次抽出 20 条记录,循环多次,抽样得到的结果的正确率在 90% 以上。抽取结果如表 6 所示。

4 结 论

本文提出了面向安全事件新闻文本的时间表达式抽取和转换方法。时间表达式的抽取采用的是基于正则表达式和时间词典的方法,采用该方式可以为后续时间转换节省时间,可以基于 TIMEX3 的时间规范,采用机器学习的方式提高准确率。得到时间表达式后采用六大时间转换算子和时间冲突处理算子对时间表达式进行转换,从而得到统一的时间表达形式即时间的年月日时分秒格式。在接下来的工作中可以在时间抽取阶段基于 TIMEX3 规范利用机器学习算法来提高准确率,从而更大程度地抽取时间表达式。

在危害公共安全事件的关联分析挖掘与预测中,对安全事件新闻文本中的时间抽取准确度较高,保证了后续新闻文本分类以及事件聚类的效果。在新闻文本中描述安全事件时会有特有的形式,如“X. X”事件。本文时间词典构造的先验条件是安全事件时间类型。在非安全事件新闻文本中出现的更加多样化的时间表达方式在时间词典中涵盖不完全,在后续的研究工作中会将本文的时间抽取与转换方式扩展到其他领域,使之更具普适性。

表 5　时间抽取结果对比

事件类型	总记录数	CRF 抽取出的时间记录数	抽取出的时间记录数	转换后的时间记录数
公交车爆炸	77370	55087(71.2%)	54701(70.7%)	54657(70.6%)
暴恐	223469	163356(73.1%)	165814(74.2%)	165369(74%)
校园砍伤	64533	43753(67.8%)	42850(66.4%)	42823(66.3%)

表 6　时间抽取与转换结果示例表

新闻文本	新闻发布时间	最终时间表示
4 月 7 日下午,湘潭县易俗河镇发生一起大巴车爆炸事故,三名行人被冲击波击碎的玻璃所伤。	2013-04-09 08:57:24	2013 年 04 月 07 日
昨天早上 8 时 20 分,一辆从铜厂开往北京站东方向的公交车在行驶到东侧路车站时发生爆炸	2013-08-01 15:47:00	2013 年 07 月 31 日 08 时 20 分
31 日记者从厦门市招办获知,在“6·7”纵火案中受伤的 7 名高考学生今年都直升大学	2013-08-01 12:14:17	2013 年 06 月 07 日

参考文献

[1] Mani I, Wilson G. Robust temporal processing of news. In: Proceedings of the 38th Annual Meeting on ACL. Morristown. 2000. 69-76

[2] Ferro L, Mani I, Sundheim B, et al. TIDES temporal annotation guildelines, version 1.0.2: [Technical Report]. The MITRE Corporation, McLean, Virginia. Report MTR 01 W0000041

[3] Ferro L, Gerber L, Mani I, et al. TIDES 2003 standard for the annotation of temporal expressions [EB/OL]. (2003-09) http://timex2.mitre.org, 2015

[4] Pustejovsky J, Castano J, Ingria R, et al. TimeBank 1.2 Documentation [EB/OL]. (2006-4) http://timeml.org/site/timebank/documentation-1.2.html, 2015

[5] Mani I, Schiffman B, Zhang J P. Inferring temporal ordering of events in news. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT′03), Stroudsburg, USA, 2003, 2: 55-57

[6] Mani I, Marc V, Wellner B, et al. Pustejovsky. Machine leaning of temporal relations. In: Proceedings of the 21th International Coference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, Australia, 2006, 44:753-760

[7] Lapata M, Lascarides A. Learning Sentence-International Temporal Relations. *Journal of Artifical Intelligence Research*, 2006,27:85-117

[8] Chambers N, Wang S, Jurafsky D. Clasifying Temporal Relations between Events. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, Czech, 2007, 45:173-176

[9] Abe S, Inui K, Mastsumoto Y. Two-phased event relation acquisition coupling the relation-oriented and argument-oriented approaches. In: Proceedings of 22nd International Conference on Computational Linguistics, Manchester, UK, 2008,1:1-8

[10] Chklovski T, Panel P. Global path-based refinement of noisy graphs applied to verb semantics. *Lecture Notes in Computer Science*, 2005,3651:792-803

[11] Yoshikawa K, Riedel S, Asahara M, et al. Jointly identifying temporal relations with Markov logic. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2009,1:405-413

[12] Mirroshandel S A, Khayyamian M, Ghassem-Sani G. Syntactic tree kernels for event-time temporal relation learning. *Lecure Notes in computer Science*, 2011,6562: 213-223

[13] Gerber L, Huang S, Wang X. Standard for the annotation of temporal expressions, Chinese supplement draft [EB/OL]. (2004-04). http://timex2.mitre.org, 2015

[14] 林静,曹德芳,苑春法. 中文时间信息的 TIMEX2 自动标注. 清华大学学报(自然科学版),2008, 48(1): 117-120

[15] Zhang X Y, Wang S P, Yun X C. Bidirectional active learning: a two-way exploration into unlabeled and labeled dataset. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(12): 3034-3044

[16] Zhang X Y, Xu C S, Cheng J, et al. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia*, 2009, 11(2): 272-285

[17] word2vec Tool for computing continuous distributed representations of words, https://code. google. com/p/word2vec/, 2015

Extraction and normalization of temporal expressions for news reports on security events

Li Mingyue*, Wang Shupeng*, Wang Haiping*, Fu Ge**

(*Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093)

(**National Computer Network Emergency Response Technical / Coordination Center of China, Beijing 100029)

Abstract

The importance of event news reports' time extraction and processing to event research was interpreted, and the extraction and normalization of temporal expressions for the news reports on security events were studied. Considering that now temporal expressions extraction is mainly based on the established norms TIMEX2 or TIMEX3 and machine learning, thus the temporal expressions acquired, are not in an unified form and are not directly applied to discovery and analysis of the public opinions about security incidents, a method for temporal expressions extraction and normalization for news reports on security events news was proposed. This method extracts different kinds of temporal expressions respectively according to the classification of them. And then it uses six temporal expressions normalization operators and the time conflict processing operator to give the unified form for representation of time using year, month, day, hour, minute and second. The proposed method was tested by experiment, and the results indicated that its time extraction effect was similar to the approach using the form of condition random field (CRF). What's more, its correctness of temporal expressions normalization was above 90%.

Key words: news, events, temporal expression extraction, temporal expression normalization, public opinion analysis