基于语言特征的舆情事件抽取

景悦诚， 黄 征

（上海交通大学 信息安全与工程学院，上海 200240）

[摘 要]随着社交媒体技术的快速发展，人们越来越喜欢在微博这个社交平台上发布信息。在这仅仅140个字的消息当中，蕴藏着大量嘈杂而有价值的文本信息。寻找一个有效的舆情事件抽取方法也越来越受到人们的关注，事件抽取也成为一个热门的研究领域。本文采用了一系列的方法用于事件抽取。主要是采用新浪微博作为语料数据，选取金融舆论事件作为事件语料，使用条件随机场对事件元素生成模型。并在预测结果中加入参数，使得抽取结果的准确率有所提高。

[关键词]事件抽取；内容安全；中文微博；自然语言处理；条件随机场

[中图分类号]TP181 [文献标志码]A [文章编号]1009-8054(2015)04-0096-05

PublicOpinions Event Extraction based on Language Feature

JINGYue-cheng, HUANG Zheng

(School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

[**Abstract**] With the rapid development of social media technology, people are increasingly eager to publish their messages by Chinese microblog. Multitudes of noisy but valuable test messages are contained in this mere 140-character microblog. Therefore, the way how to extract events from Chinese microblog receives much attention from the folks, and event extraction also becomes a hot research field. A series of methods on events extraction are described in this paper. It uses Sina Weibo as corpus data and financial event as the topic of public opinions, CRFs are applied to generate prediction model. In particular, the parameter is added to the prediction results, thus to increase the precision of this extraction result.

[**Key words**] event extraction; content security; Chinese microblog; NLP; CRFs

0 引言

社交网络在最近几年得到了飞速发展。如美国的Facebook、Twitter，中国的人人、新浪微博，人们通过这些社交平台发布他们的想法，共享他们的信息。成立于2009年8月的新浪微博是目前中国最大的微博平台，目前其拥有的注册用户数超5亿，月活跃用户数1.67亿[1]。庞大的用户群背后是每天数以亿记的微博信息。微博的操作便利性、低门槛性使得其成为人们日常发布信息的一个途径。作为现实社会与网络社会的重要媒介节点，微博中蕴含着海量的信息，分析微博的语义信息、挖掘微博热点话题、研究微博信息处理技术具有重要的理论意义。在政府管理领域、政策风险评估、网络舆情分析、商业广告应用等领域中都有着重要的实用价值。这一领域的研究成果在政府舆情分析、事件监控及企业商业智能系统等诸多领域有着广阔的应用空间和发展前景[2]。最近几年来，国内外在微博事件挖掘领域的研究有：Jui-Yu Weng提出了一个智能微博分析系统[3]，Alan Ritter等人提出了一个基于Twitter的事件抽取和分类系统[4]，郑斐然等人提出了一套微博数据处理方法和新闻话题的检测算法[5]，张晨逸等人提出了基于MB-LDA模型的微博主题挖掘方法[6]。

收稿日期：2014-11-19

1 微博事件抽取系统框架

目前大部分的事件抽取主要有两种方法：一种是模式匹配的方法，另一种是机器学习的方法。前者是指对于某类事件的识别和抽取是在一些模式的指导下进行的，采用各种模式匹配的算法将待抽取的句子和已经抽取的模板匹配。这类方法的优点是准确率较高，缺点是依赖于具体的领域知识。后者是把事件抽取看成分类问题，把主要的精力放在分类器的构建和特征的发现、选择上。这类方法的优点是不需要太多的人工干预和领域知识。因此目前的事件抽取研究大多采用机器学习的方法。

本文采用机器学习的方法，是一种基于概率模型的方法。这个概率模型是在最大熵模型和隐马尔科夫模型的基础上提出的条件随机场模型，它是一种判别式概率无向图学习模型，主要用于标注和切分有序数据，非常适合事件抽取这个场景。整个事件

抽取系统由微博预处理,语言特征标注,CRF 模型学习,事件抽取展示。微博预处理主要是微博抓取,微博标准化,写入语料数据库。语言特征标注包括:分词、词性标注、命名实体识别以及人工标注事件元素。将合并了语言特征的文本作为 CRF 模型学习的输入,得到事件抽取的模型。最终使用 CRF 模型对新抓取的微博进行事件抽取、展示。图 1 为整个系统的框架。

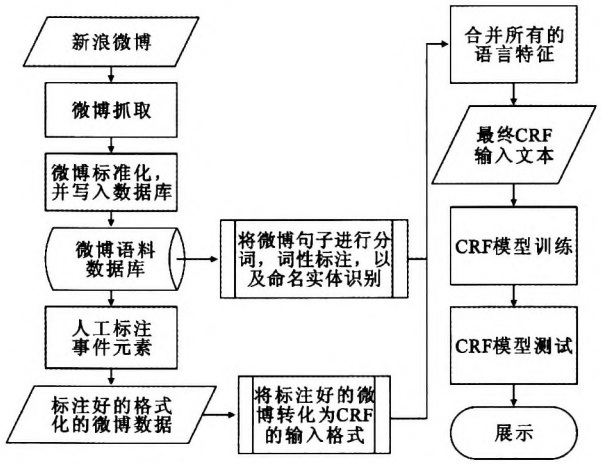

图 1 微博事件抽取框架图

2 具体实现

2.1 抓取微博

新浪微博提供了微博开放平台,它向开发者提供了 200 多个开放的公共 API,用户注册开发者账号后可以在一定的限制内调用这些 API,来获得与微博有关的信息,例如微博内容、评论、用户、话题、关系等。本文使用新浪微博官方的 API,实现信息读取。为了使语料库能够快速的建立,微博账号关注的都是金融类的微博,能够在短期建立足够多的专业语料。

在本文中,希望获取的是当前登录用户及其所关注用户的最新微博,对应的 API 为 statuses/friends_timeline[7]。该 API 支持格式为 JSON,HTTP 请求方式为 GET,访问级别为普通接口,需要登录授权,有频次限制。本文中使用的是具有测试授权的账号,调用总限制为 150 次/小时[8]。主要获取其"text"、"crated_at"等属性,即微博信息内容、微博创建时间等。

2.2 微博语料数据库

将抓取的微博信息写入数据库中,为将来标注事件元素所用。一条微博中可能包含多个句子。在多个句子中进行指定成分的挖掘是一件复杂且正确率较低的事情。所以针对每条微博,程序提取其中的一句话作为微博的代表。在当前程序中,如果微博具有形如"[xxx]"的格式,且方括号中的内容超过 9 个字,则认为方括号中为微博内容的提要,存入数据库。如果没有该格式的内容或内容字数太少,则舍弃方括号中的文字,直接搜索微博内容中以句号分隔的句子中数字最多的一句话作为微博内容的提要,将其保存在数据库中。此外还将微博文本进行哈希运算,将结果存入数据库。如果新获取的微博进行哈希运算之后有相同的值,则说明已经写入过数据库,不再进行重复保存。

2.3 人工事件元素标注

事件标注是本文的重要环节,指在微博条目中抽取出感兴趣的信息。本文中获得的微博为金融类微博,所以设置兴趣点为时间(When)、地点(Where)、事件(Trigger)、主体(Who)和受事体(Whom)。事件标注的初始集合为之前提及的微博语料数据库。训练集为人工标注获得。

其中,"s.."和"e.."分别表示相应关键字为兴趣点的开始和结束。需要说明的是,为了以后进一步的研究使用,在标注中保留了 eventtype、trigger、who、whom、how、where、whyhow、whywho 等多个关键字。在本文中,标注时只使用了 who、trigger、whom、where、when 等关键字,其他关键字并未进行标注。

2.4 中文分词

如今中文自然语言处理的工具越来越多了。其中三种比较广泛使用的分词工具,中科院中文分词系统 ICTCLAS[9],哈工大社会计算与信息检索研究中心研制的语言技术平台(LTP)[10],斯坦福分词[11]。在中文分词的结果上,哈工大的 LTP 更为准确[12]。此外哈工大的 LTP 还提供了更为丰富的词性标注以及命名实体识别,以前这些也是需要人工标注的。使用了哈工大语言技术平台云之后,可以得到微博文本分词、词性标注以及命名实体识别的结果。使用哈工大 LTP 接口后,返回的结果形式如图 2。

```
0  福田  _  _  nz  O  1  ATT_  _  _  (AO*(AO*
1  汽车  _  _  n   O  2  SBV_  _  _  *)  *)
2  拟定  _  _  v   O  -1 HED_  _  拟定  (v*)*
3  增募资 _  _  v   O  2  VOB_  _  _  *  *
4  不    _  _  d   O  5  ADV_  _  _  *  (ADV*)
5  超过  _  _  v   O  2  COO_  _  超过  *  (v*)
6  40亿  _  _  m   O  7  ATT_  _  _  *  (A1*
7  元    _  _  q   O  9  ATT_  _  _  *  *
8  发力  _  _  n   O  9  ATT_  _  _  *  *
9  主业  _  _  n   O  5  VOB_  _  _  *  *)
10 。    _  _  wp  O  2  WP _  _  _  *  *
```

图 2 哈工大 LTP 接口返回的结果图

结果是 CoNLL 格式[13],它是一种表示语言分析结果的通用格式。在哈工大 LTP 的 CoNLL 格式中,分析结果的每一行代表句子中每个词的信息,词标号从 0 开始。分析结果的基础列有 10 列,之后的每一列代表文本中的语义信息,每列之间用 Tab 分割。此列值为空用"_"占位。表 1 为 CoNLL 每列的含义。

表1 CoNLL格式的含义

列号	含义
1	单词在句子中的标号,从0开始
2	单词本身
3	空
4	空
5	单词词性标注信息
6	命名实体识别
7	依存句法关系类型
8	空
9	空
10	如果单词是语义角色标注中的谓词,则为单词本身,否则为空
11及以后	每个谓词占一列,每一列为该谓词的语义角色标注信息

本文只使用5,6列的信息作为语言特征输入CRF进行模型学习。其他的语言特征可以在日后的研究中添加到CRF的输入中,并进行实验验证。

2.5 合并语言特征

得到微博的分词、词性标注以及命名实体识别结果之后,需要和之前人工标注的结果合并,变成CRF标准输入格式。格式如下:

万科	nh	S-Nh	swho
一季度	nt	O	swhen
净利	n	O	strigger
15.3亿	m	O	show
同比	j	O	how
降	v	O	how
5%	m	O	how
。	wp	O	ehow

2.6 CRF模型建立

2.6.1 条件随机场(CRF)

CRF是一种概率模型,是随机场的一种,常用于标注或分析序列资料,如自然语言文字或是生物序列。CRF能够很好的结合分类法和图模型法,能对复杂数据的建模和对大规模输入的特征进行预测。

在给出CRF定义之前,先要设标记的序列集用随机变量X来表示,用随机变量Y表示其对应的标记序列集,并且对所有$Y_i \in Y$都在一个大小为N的有限字符集内。随机变量X和Y是联合分布,p(Y|X)表示观察序列和标记序列的条件概率模型,p(X)表示隐含的边缘概率模型。

CRF的定义:在无向图G(V,E)中,有$Y=(Y_v)_{v\in V}$,且随机变量Y中的元素与无向图G中的顶点一一对应。在条件X下,随机变量Y_v的条件概率分布服从图的马尔科夫属性:

$$P(Y_v|X,Y_\omega,\omega\neq v)=P(Y_v|X,Y_\omega,\omega\sim v) \quad (1)$$

其中$\omega\sim v$表示无向图G的边,这时我们称(X,Y)是一个条件随机场。

CRF是一个无向图模型。它的节点分为两个分离的集合X和Y,X代表了被观察的变量,Y代表了输出变量。那么条件分布$P(Y|X)$就是被建立的模型。CRF的目的就是找出标注的序列$y\in Y$,使得序列X的条件概率最大。

$$y=\text{argmax } P(Y|X) \quad (2)$$

因此选择CRF作为学习模型的工具。

2.6.2 CRF工具pcrf

pcrf[14]是一个简单的,标准的,使用python实现的CRF。pcrf的主要是设计用来进行各种自然语言处理的,如命名实体识别,事件发掘,文本分组。相比其他CRF工具,提供了一些很重要的特性。如可以重新定义特征集、基于LBFGS的快速训练、与热门CRF工具CRF++通用、多线程工作等。使用之前得到的CRF标准格式的文本即可作为pcrf的训练输入。

此外还需要给CRF模型写一个特别的模板,表明哪些特征将会被用来训练和测试。pcrf的模板定义如下:模板文件中,每一行代表一个模板,定义宏%x[row,col]来作为输入数据中的块。其中,row表示相对于关注块的行数,col表示相对于关注块的列数。即以关注块为坐标(0,0),横向为y轴,向右为y轴正方向,纵向为x轴,向下为x轴正方向的坐标(x,y)[15]。本文中将词语本身,词性,命名实体识别,事件元素作为关注的块,一共四列作为输入,生成模型。

2.7 带参数筛选的事件元素抽取

建立好CRF模型之后,将新抓取的微博进行分词、词性标注以及命名实体识别,然后把结果转换为标准的CRF输入,一共三列,得出事件抽取结果。最后将其存入语料数据库,用于展示。

使用新微博的三列语言特征作为输入之后,可以得出第四列特征的概率分布。在本文中就会得出一个分词是,When、Where、Trigger、Who以及Whom的概率。如果微博是金融类的,并且结构化比较明显的时候,这些概率会比较显著,某一个的概率会大于80%,其他加起来不到20%,这个时候正确率会很高。但是一些非相关主题的事件,一些分词的概率就会不显著,比如最大的一个是30%,第二的25%,这种情况其实不应该采取任何标注。但是CRF会选出概率最大的一个作为输出,这样会导致正确率下降。所以我们以第一的概率值与第二的概率值的比作为参数,若值不大于某个阈值时,抽取结果为O,即非事件元素。这样可以有效的提高准确率,但是会降低一点召回率。

3 实验

3.1 评测指标

在自然语言处理中,召回率 R、准确率 P 和正确率 A 通常被用来衡量系统性能的常用指标。召回率指系统能正确识别出的个体数目占标准结果中个体总数的比例。准确率是指系统正确识别出的个体数目占系统总共识别出的个体数目的比例。正确率指系统正确识别的个体数目占所有个体数目的比例。将召回率和准确率进行几何加权平均,得到 F1 指数,作为系统的总体评价。F_λ 的计算公式如下:

$$F_\lambda=\frac{(\lambda^2+1)(P\times R)}{(\lambda^2\times P)+R} \tag{3}$$

其中 P 是准确率,R 是召回率,λ 是二者的相对权重,在本文中取 1,即两者权重相同,这个 F 值就是 F1 指数。

3.2 交叉验证

交叉验证(Cross-validation)主要用于建模应用中。在给定的建模样本中,拿出大部分样本进行模型建立,留小部分样本用于刚建立模型的预报,并求这小部分样本的预报误差,记录它们的平方加和。这个过程一直进行,直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和,称为 PRESS(predicted Error Sum of Squares)。

10 折交叉验证(10-fold cross validation)是比较常用的交叉验证,将数据集分成十份,轮流将其中 9 份做训练 1 份做测试,10 次的结果的均值作为对算法精度的估计,一般还需要进行多次 10 折交叉验证求均值。

本文的研究对象主要是微博上的文本信息,以新浪微博作为实验数据。通过新浪微博 API 将关注的金融类的热门微博抓取下来,并人工从中筛选出 1000 条标注成为实验数据。通过对这 1000 条微博进行 10 折交叉得到实验结果。

表 2 不带参数筛选的事件抽取结果

实验序号	召回率(%)	准确率(%)	正确率(%)	F1 指数(%)
1	40.74%	70.40%	54.42%	51.61%
2	40.25%	62.75%	60.77%	49.04%
3	56.46%	64.84%	68.31%	60.36%
4	57.31%	75.97%	70.62%	65.33%
5	47.10%	66.97%	66.76%	55.30%
6	36.61%	82.83%	54.05%	50.77%
7	39.38%	82.41%	54.78%	53.29%
8	31.56%	84.52%	51.73%	45.95%
9	41.63%	82.88%	55.69%	55.42%
10	35.92%	88.00%	51.01%	51.01%
平均	42.70%	76.16%	58.81%	53.81%

表 3 是带参数筛选的 10 折交叉验证结果,数据为同一参数下的 10 次实验平均值。

表 3 带参数筛选的事件抽取结果

参数	召回率(%)	准确率(%)	正确率(%)	F1 指数(%)
1.0	42.70%	76.16%	58.81%	53.81%
2.0	34.36%	81.81%	57.47%	47.75%
3.0	29.99%	85.46%	56.40%	43.81%
4.0	26.87%	86.88%	55.26%	40.54%
5.0	24.41%	88.72%	54.41%	37.84%
6.0	22.08%	90.22%	53.55%	35.12%
7.0	20.84%	91.52%	53.15%	33.60%
8.0	19.55%	91.89%	52.58%	31.90%
9.0	18.17%	92.40%	51.91%	30.03%
10.0	17.07%	92.57%	51.33%	28.46%

图 3 是带参数筛选的事件抽取结果的曲线图,横轴是参数值,纵轴是四个指标的百分率。参数为 1 时,是默认的 CRF 抽取结果,即选择概率最大的标出。从 2 开始,概率最大的元素的概率值必须是第二大的 2 倍才会标注出,以此类推。可以发现,在 1~2 之间,准确率提升的最快,而召回率下降的也是最快。也就是说在这一区间内,参数对抽取结果影响最大。这个参数可以根据实际的事件内容进行调整,在召回率下降和准确率上升中做一定的取舍。

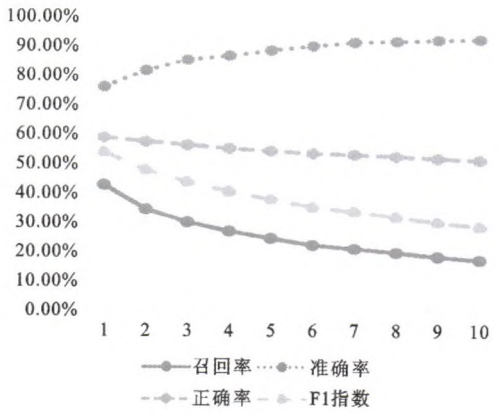

图 3 带参数筛选的事件抽取结果曲线图

4 结语

本文给出了一个基于语言特征的微博事件抽取系统,用于对舆情事件进行发掘。主要使用了哈工大的 LTP 语言技术平台云做分词、词性标注以及命名实体识别的工作。然后与人工标注的事件元素进行合并,生成机器学习文本。使用随机条件场(CRF)作为学习模型,将规范的学习文本作为输入,得出预测模型。使用交叉验证得出系统的评价,并使用参数对系统进行调整。最后使用模型对抓取过来的新微博进行事件抽取。实验表明:通过使

用语言特征作为CRF的输入,并通过参数调节,可以得出比较高准确率的事件抽取结果。目前,在中文微博领域上的事件抽取研究还处于起步阶段。在这个方向上的研究还有很大的空间。例如如何找到一种更有效的方式去提高事件抽取的召回率,因为从实验结果中可以看出,提高准确率必须降低召回率作为牺牲,那么如果想要抽取出更多更准确的信息一定需要提高召回率。

参考文献:

[1] 新浪微博.新浪微博2014年第三季度财报[EB/OL].美国:新浪微博,2014(2014-11-14)[2014-11-18].http://tech.sina.com.cn/i/2014-11-14/05309789970.shtml? sina-fr=bd.ala.cb.

[2] 文坤梅,徐帅,李瑞轩等.微博及中文微博信息处理研究综述[J].中文信息学报,2012,26(6):27-37.

[3] WENG J Y, YANG C L, CHEN B N, et al. IMASS: An Intelligent Microblog Analysis and Summarization System [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. Portland, Oregon. Association for Computational Linguistics, 2011: 133-138.

[4] RITTER Alan, Mausam, ETZIONI Oren. Open Domain Event Extraction from Twitter [C]//Proceedings of KDD. [s.l.]: ACM,2012:1104-1112.

[5] 郑斐然,苗夺谦,张志飞等.一种中文微博新闻话题检测的方法[J].计算机科学,2012,39(1):138-141.

[6] 张晨逸, 孙建伶, 丁轶群. 基于MB-LDA模型的微博主题挖掘[J]. 计算机研究与发展, 2011(10):1795 - 1803.

[7] 新浪微博.新浪微博API说明文档[EB/OL].中国:新浪微博,2013(2013-09-10) [2014-11-18].http://open.weibo.com/wiki/2/statuses/home_timeline.

[8] 新浪微博.新浪微博API接口访问频次权限[EB/OL]. 中国:新浪微博,2014(2014-05-21) [2014-11-18].http://open.weibo.com/wiki/接口访问频次权限.

[9] 中国科学院计算技术研究.ICTCLAS汉语分词系统[EB/OL].中国: 中国科学院计算技术研究所,2010(2010-12-21)[2014-11-18].http://www.ictclas.org/.

[10] CHE W, LI Z, LIU T. Ltp: A Chinese Language Technology Platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Beijing, China. Association for Computational Linguistics, 2010: 13-16.

[11] The Stanford Natural Language Processing Group. Stanford Word Segmenter [EB/OL].USA: The Stanford Natural Language Processing Group (2014-10-26)[2014-11-18].http://nlp.stanford.edu/software/segmenter.shtml.

[12] XIONG J, HAO Y, HUANG Z. Civil Transportation Event Extraction from Chinese Microblog [C]//Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on. Fuzhou, China. IEEE, 2013: 577-582.

[13] 哈工大社会计算与信息检索研究中心.哈工大语言云新版API使用文档[EB/OL].中国:哈工大社会计算与信息检索研究中心 2014(2014-9-26)[2014-11-18].http://www.ltp-cloud.com/document/new/.

[14] 黄征.Python Linear CRF[EB/OL].中国:黄征,2014(2014-4-3)[2014-11-18].https://github.com/huangzhengsjtu/pcrf.

[15] Taku Kudo.CRF++: Yet Another CRF toolkit[EB/OL].Japan: Taku Kudo,2013(2013-2-13)[2014-11-18].http://crfpp.googlecode.com/svn/trunk/doc/index.html.

作者简介:

景悦诚(1990—),男,硕士研究生,研究方向为机器学习、事件发掘、自然语言处理;

黄 征(1975—),男,研究生导师,研究方向为人工智能、安全多方计算和数字取证信息的收集与分析。■

(上接第95页)

参考文献:

[1] 信息系统审计和控制联合会(ISACA).COBIT5, Control Objectives for Information and Related Technology Fifth Edition [S].ISACA,2013.

[2] 国际标准化组织(ISO).ISO/IEC 27001:2013, Information Technology—Security Techniques—Information Security Management Systems—Requirements[S].ISO,2013.

[3] 金融标准化技术委员会.JR/0071-2012,金融行业信息系统信息安全等级保护实施指引[S].中国人民银行,2012.

[4] 黄作明.信息系统审计[M].胡记兵 余小兵.东北财经大学出版社,2012.

作者简介:

袁慧萍(1969—),女,博士,高级工程师,CISP,主要研究方向为网络安全、信息安全、涉密管理、信息技术审计;

董贞良(1983—),女,硕士,高级工程师,主要研究方向为网络安全、信息安全、信息技术审计、操作风险管理。■