一种非监督的事件触发词检测和分类方法*

陈自岩[1,2] 黄 宇[2] 王 洋[2] 傅兴玉[2] 付 琨[2]
(1. 中国科学院大学 北京 100049;2. 中国科学院空间信息处理与应用系统重点实验室 北京 100190)

摘 要:事件触发词检测和分类是事件抽取中至关重要的第一步。传统的抽取和分类方法往往倾向于监督学习方法,如条件随机场、SVM等,但由于这类方法需要繁重的人工标注且受限于预先定义好的类别,因此很难在开放领域中得到应用。提出了一种非监督的事件触发词检测和分类方法,利用主题模型获取候选触发词在主题上的分布,然后利用二值状态自动机模型捕获高概率的主题,从而筛选出真正的事件触发词和相应的分类。在大规模的未标注新浪新闻数据集上的实验结果充分验证了本文方法的有效性。

关键词:事件触发词检测和分类;主题模型;二值状态自动机模型
中图分类号: TP3 **文献标识码**:A **国家标准学科分类代码**: 520.2020

Unsupervised method for event trigger identification and classification

Chen Ziyan[1,2] Huang Yu[2] Wang Yang[2] Fu Xingyu[2] Fu Kun[2]
(1. University of Chinese Academy of Sciences, Beijing 100049, China; 2. Key Laboratory of Technology in Geospatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The identification and classification of event trigger plays a decisive role in event extraction. Usually, the trigger words are extracted based on supervised machine learning methods such as CRF. However, since these methods rely on expensive manual annotation and require predefined event types, they are not sufficient for open domain application. In this paper, we present an unsupervised method for event trigger identification and classification. First, we run a topic model to obtain the topic distribution over each candidate trigger word. Then, an improved two-state automaton is proposed to detect the real trigger word and capture the corresponding topics. The experiment on a large unlabeled corpus shows our unsupervised model is very inspiring.

Keywords: event trigger identification and classification; topic model; two-state automaton

1 引 言

在互联网信息极大丰富并保持爆炸式增长的当今时代,如何高效地从海量、繁杂、冗余的信息中提炼出有价值的信息,已经成为急需解决的问题。信息抽取技术旨在从非结构化的文本中抽取出结构化的信息,以便于数据库的储存,并为问答系统、自动文摘和推荐系统等领域打下基础。事件抽取是信息抽取3大任务之一,而事件触发词检测又是事件抽取中至关重要的第一步。

事件触发词就是最能表达事件发生或事件类型的词,在词性上一般表现为动词或少量名词。早期的事件触发词检测往往采用基于统计的方法[1-2]或基于规则的方法[3]。但基于统计方法往往过度依赖丰富的标注语料,而基于规则的方法往往需要很多的语言专家去定义繁多的规则。随着机器学习的快速发展,研究者们提出了许多机器监督学习方法,如 k 近邻[4]、支持向量机(SVM)[5-6]和条件随机场(CRF)[7]等,并在很多领域得到广泛应用,其中包括事件触发词检测和分类。2006年,AHN D[8]基于ACE采用了TiMBL(基于记忆的学习器)和MegaM(最大熵学习器)进行对比实验,比较了两种方法在事件触发词抽取与事件类型分类中的效果;2012年,RITTER A等人[7]把事件触发词抽取当成一个序列标注任务,并采用Link_LDA对事件触发词进行分类;2012年,LI P F等人[9]提出了一种基于ILP的推导框架,把事件触发词检测和分类联合起

收稿日期:2016-03
* 基金项目:国家自然科学基金(61331017)项目资助

来学习;2014年,王健等人[10]采用间接的句法信息模式,利用深层句法信息来对生物事件触发词进行检测。这些监督机器学习方法往往依赖繁重的人工标注的数据集,且事件类型都是预先定义好的,使其很难在开放领域得到有效应用。许多研究者也试图把基于统计、规则和机器学习的方法综合起来进行抽取[11-12],但是这些工作仍停留在事件触发词检测上,还不能对事件触发词进行分类。2012年,丁效等人[13]针对音乐领域提出了一种非监督方案,利用领域事件词抽取算法抽取事件触发词,并利用事件类型发现算法实现了事件词的聚类,但这种方法只适合应用在特定领域上。

针对以上问题,本文提出了一种事件触发词检测和分类的非监督解决方案。由于事件触发词绝大部分为动词和动名词,因此在预处理中,把动词和动名词作为候选触发词。首先利用基于滑动窗口的方法获得事件触发词的共现词,然后采用主题模型获取候选事件触发词在主题上的分布。由于事件触发词表征了事件的发生或事件的类型,因此真正的事件触发词在主题上的分布会出现高概率的峰值,即主题分布的不均匀性。基于此,本文利用二值状态自动机来捕获高概率的峰值,从而检测出真正的触发词和其对应的类别。在12万多条的新浪新闻数据集上的实验,充分证明了本文方法的有效性。最终训练获得的事件触发词及其对应的主题分布可以直接服务于开放领域的事件抽取、摘要生成等上层应用。

2 方 法

在自然语言处理领域,主题模型是用来发现文档集中蕴含的抽象主题的一种统计模型。传统的主题模型一般自动分析每个文档与文档中的词语,根据共现来实现词的聚类,从而发现文档的主题。本文工作针对的不是文档集,而是候选事件触发词词典。因此,在直接应用主题模型前,必须对数据进行预处理,获得与事件触发词共现的词集合。

2.1 预处理

在文本处理领域,中文的分词与词性标注是文本预处理过程中不可或缺的关键步骤。ICTCLAS采用了层叠隐马尔可夫模型,将汉语的分词、词性标注与命名实体识别等环节统一到一个完整的框架体系,其在实际应用和实验评测中都得到了很好的效果,本文基于此工具包对数据集进行预处理,标注出每篇文档中词的词性,并把动词和动名词当成候选事件触发词。

为了获取与候选事件触发词共现的词,本文设计了一种基于滑动窗口的方法。预先定义一个固定大小为 N 的窗口,并用此窗口滑动遍历整个数据集,一旦捕获到候选事件触发词,就把此窗口内的所用词加入到此候选事件触发词的共现列表中。

2.2 潜在狄利克雷分配

潜在狄利克雷分配(Latent Direchlet allocation,LDA)[14]是一种建模文本数据产生过程的概率有向图模型,是第一种被显式提出的概率主题模型,由于其具有良好的数学基础和灵活扩展性,因此自提出以来,很快得到了各个领域研究者的关注,包括文本挖掘[15]和图像标注[16]等。本文将采用LDA获取候选事件触发词在主题上的分布,其基本思想是认为每个事件触发词是若干隐含主题的混合分布,而每个主题由一些语义相关的词组成。具体地,假设经过数据集的预处理得到事件触发词词典 $E=\{e_1,e_2,\cdots,e_E\}$,每个事件触发词共现的词集合 $W=\{w_1,w_2,\cdots,w_N\}$,整个数据集中包含 V 个不同的词,则LDA的概率图模型表示如图1所示,该图描述了如下生成过程:

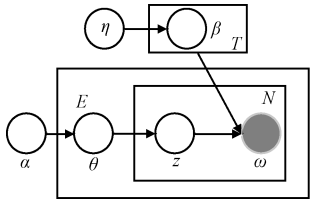

图1 LDA的概率图模型

1)对每个事件类型 $t=1,2,\cdots,T$ 有:

由狄利克雷分布 $Dir(\eta)$ 生成 β_t

2)对每个唯一的事件触发词 $e=1,\cdots,E$ 有:

①由狄利克雷分布 $Dir(\alpha)$ 生成 θ_e

②对与每个事件触发词 e 共现的词 $i=1,\cdots,N_e$ 有:

a)由多项式分布 $multinomial(\theta_e)$ 生成 $z_{e,i}$

b)由多项式分布 $multinomial(\beta_{z_{e,i}})$ 生成 $W_{e,i}$

假设参数 α、η 已知,则每个事件触发词的主题分布 θ,与事件触发词共现的所有词的主题分配 z 以及数据集中所有的词 W 的联合概率为:

$$p(\theta,Z,W\mid\alpha,\beta)=p(\theta\mid\alpha)p(\beta\mid\eta)\cdot\prod_{n=1}^{N}p(Z_n\mid\theta)p(W_n\mid\beta,Z_n)$$

在LDA模型训练过程中,只有观测变量 W 是已知的,隐变量 θ、Z 和 β 均是未知,本文通过Gibbs采样方法求解这些隐变量。从LDA概率图模型和联合概率中,可以看出 $p(\theta\mid\alpha)$ 服从狄利克雷分布,$p(z\mid\theta)$ 服从多项式分布,前者是后者的共轭先验,因此可以在Gibbs采样中将 θ 略去,仅对 z 进行采样,从而由 z 的统计量求出 θ,此类采样方法称为Collapsed Gibbs[17]采样。同理可以求出 β,所得结果为:

$$\theta_{e,z}=\frac{n_{e,z}+\alpha_z-1}{\sum_{z=1}^{T}(n_{e,z}+\alpha_z)-1}$$

$$\beta_{z,v}=\frac{n_{z,v}+\eta_v-1}{\sum_{v=1}^{V}(n_{z,v}+\eta_v)-1}$$

式中：θ 为本文所需的候选事件触发词在主题上的分布，下文将根据此分布对事件触发词进行检测和分类。

2.3 改进的二值状态自动机

在获得候选事件触发词在主题上的分布后，任务是捕获高概率的峰值，并将此峰值对应的主题作为事件触发词的事件类型。捕获高概率的峰值最常用的方法就是阈值法，但阈值法的最大缺点是难以给定合适的阈值。有限状态自动机常用在时间流数据的建模中，其基本思想是用指数分布建模时间间隔，或用泊松分布建模时间点上的统计数量[18-19]。但事件触发词的主题序列服从狄利克雷分布，因此无需再用指数分布或泊松分布进行建模。

v_t 表示当前主题的状态，其中 $v_t=0$ 表示候选事件触发词在当前主题没有出现峰值，$v_t=1$ 则表示候选事件触发词在当前主题出现了峰值。状态之间的转移关系如图 2 所示。

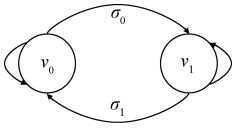

图 2 状态转移示意

候选事件触发词在主题上的分布状态序列（v_0，v_1，…，v_T）构成了一个马尔科夫链，其具有固定的状态转移概率 σ_l（l 为二值状态 0 或 1）和发射概率 $\theta_{e,t}$。基于以上的假设，定义如下的代价函数：

$$c(\boldsymbol{q}\mid\boldsymbol{x})=\ln(\sigma_0{}^a(1-\sigma_0)^b\sigma_1{}^c(1-\sigma_1)^d)+\left(\sum_{t=1}^{T}-\ln\theta_{e,t}\right) \tag{3}$$

式中：a、b、c、d 分别代表各状态转移之间的次数，并满足 $a+b+c+d=T$。目标是求最优状态序列 q 使得损失函数最小，但上式直接求解比较困难，因此本文将采用维特比算法求解最优状态序列。

维特比算法的本质用动态规划求概率最大路径，即最优路径。其基本原理是从最初的主题 $t=1$ 开始，递推地计算在主题 t 状态为 l 的各条路径的最大概率，直到主题 $t=T$，此时的最大概率即为最优路径的概率，再由此开始追溯最优路径上的结点，便得到最优状态序列 $\boldsymbol{q}$。

3 实 验

本节将在大规模的新浪新闻数据集上测试本文的非监督方法，该方法主要包含 3 个步骤：1）预处理获得候选事件触发词；2）通过主题模型获得候选事件触发词在主题上的分布；3）利用二值状态自动机捕获高概率峰值。实验流程如图 3 所示。

3.1 实验数据集与实验设置

本文采用的数据集是 2015 年 3 月 1 日至 2015 年 8 月 31 日之间的所有新浪新闻数据，一共 121 157 篇，涵盖了经济、政治、体育、军事等各种类别的主题。由于事件触发词绝大部分为动词或动名词，因此在预处理中只选择动词和动名词词性作为候选事件触发词，共有 8 538 个。另外

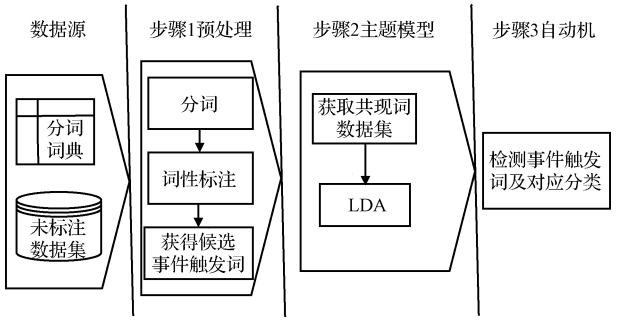

图 3 实验流程

由于未标注数据的庞大，只选取名词词性作为候选事件触发词的共现词，共有 5 276 个。

在 LDA 训练过程中，设置 Gibbs 采样的迭代次数为 1 000，主题个数为 50，固定参数 $\alpha=0.5$，$\eta=0.1$。在二值状态自动机中，取经验值 $\sigma_0=0.9$，$\sigma_1=0.6$。

3.2 候选事件触发词的主题实例

基于以上的实验数据和设置，采用 Gibbs 采样训练数据，如表 1 所示为本文模型生成的其中 10 个典型主题实例，以及其对应的前 5 个事件触发词和前 5 个共现词。表中展示了 10 个典型的事件类型实例，纵观本文模型发现的所有 50 个事件类型，发现金融、经济和法治占了绝大多数，这也符合了数据集的特点。

表 1 本文模型发现的事件类型实例

序号	类型	前 5 个事件触发词	前 5 个共现词
1	股票	下跌，收盘，上涨，反弹，走低	指数，市场，价格，股市，跌幅
2	法治	羁押，强奸，审理，犯罪，逼供	案件，警方，法院，责任，嫌疑人
3	体育	决赛，晋级，参赛，夺冠，卫冕	冠军，选手，锦标赛，满贯，成绩
4	食品药品	抽检，添加，配制，检测，含有	食品，企业，国家，总局，药品
5	军事	演习，发射，侦察，作战，武装	海军，军事，导弹，战略，部队
6	金融	浮动，报价，贬值，稳，积聚	人民币，汇率，央行，货币，外汇
7	经济政策	改革，推进，出台，试点，审批	市场，经济，国企，政策，方案
8	政治	反省，访华，特赦，抗战，参拜	国家，政府，战争，国际，关系
9	人事	巡视，提拔，免职，升任，辞去	企业，单位，人员，情况，领导
10	三农	产能，补贴，存栏，支农，增收	经济，政策，国家，生猪，价格

3.3 候选事件触发词的主题分布及二值化结果分析

本文实验的任务是获得候选事件触发词在主题上的分布 θ，然后经过改进的自动机模型进行二值化。如图 4 所示为 4 种候选事件触发词在主题上的分布。

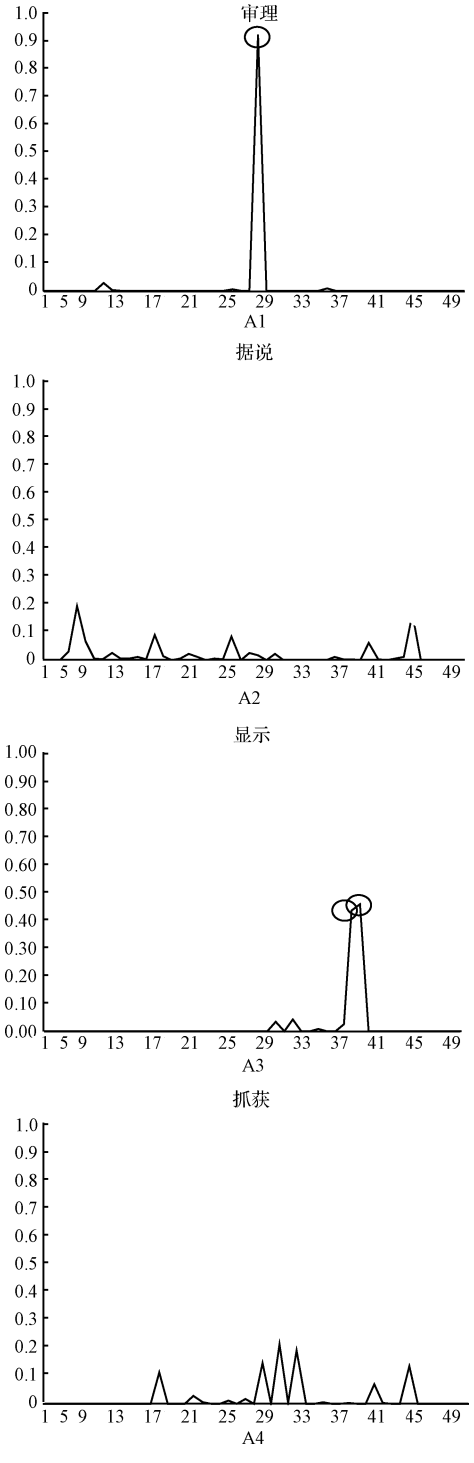

图4 候选事件触发词在主题上的分布

图中,折线代表候选事件触发词在主题上的分布,虚线圆圈代表高概率峰值出现情况。A1和A2展示了期望并实际获得的结果。真正的事件触发词在主题上分布是不均匀的,而本文的模型能捕获到高概率的主题值,即事件类型。A3和A4代表了结果中的两种特例。A3中的"显示"一词代表了37和38事件类型,从上节的事件类型分析中,发现这两个主题都是关于经济指标等的统计,其属于特定主题下的惯用词,此类的词还有"表现"、"共有"、"为期"等。A4中的"抓获"一词本应属于"法律"事件类型,但是由于法律类的主题占有比例较高,且"抓获"一词在法律类的主题上普遍存在,因此造成漏检。为了评测每类结果占有比例,引入了一种人工标注评测方法。标注者对每个候选事件触发词进行两类标注:事件触发词和非事件触发词,从而可以得到如表2所示的混淆矩阵。进而得到事件触发词词典的正确率为84.75%,如果将惯用词也纳入到事件触发词词典中,则广义上的正确率为94.76%。

表2 人工评测结果

	预测的事件触发词	预测的非事件触发词
人工标注的事件触发词	真正的事件触发词 $TP=6\ 022$	漏检 $FN=447$
人工标注的非事件触发词	惯用词 $FP=855$	真正的非事件触发词 $TN=1\ 214$

为了与前人监督学习方法进行对比,本文设计了基于条件随机场(CRF)的事件触发词检测的对比方法,所用的特征有上下文特征、词性特征等。对比结果如表3所示。正确率和$F1$值都充分表明本文模型能取得与基于CRF的监督学习方法媲美的结果。

表3 实验对比结果 (%)

比对方法	A	P	R	$F1$
条件随机场	86.81	89.00	92.09	90.52
词典方法	84.76	87.18	90.53	88.82

4 结 论

本文提出了一种非监督的方法来检测事件触发词,并对事件触发词进行分类。本文的主要贡献有3个方面:1)提出一种滑动窗口的方法来获取候选事件触发词的共现词;2)通过主题模型获取候选事件触发词在主题上的分布,其中建模的是事件触发词而不是文档;3)结合主题模型得到的狄利克雷分布结果,利用二值状态自动机模型捕获高概率峰值,从而实现事件触发词的检测和分类。在大规模的未标注数据上的实验证明了本文方法的有效性,同时也分析了实验带来的惯用词和漏检问题及原因。接下来的研究将集中在以下3个方面:1)本文方法比较适合大规模的主题涵盖广的非标注数据,但在特定领域数据上可能会面临比较高的漏检率;2)本文关注的是动词和动名词,但也有一小部分的名词和形容词也能触发事件的发生,未来将在事件性名词和形容词上进行深入研究;3)本文的事件触发词分类问题其实就是聚类,需要人工判读完成事件类型的标注,未来的工作将会重点研究事件类型的自动标注。

参考文献

[1] BUYKO E, FAESSLER E, WERMTER J, et al. Event extraction from trimmed dependency graphs [C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009: 19-27.

[2] VLACHOS A, BUTTERY P, S? AGHDHA D O, et al. Biomedical event extraction without training data[C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009: 37-40.

[3] LE MINH Q, TRUONG S N, BAO Q H. A pattern approach for biomedical event annotation[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 149-150.

[4] 郑新元,严军,范浩,等.线性不稳定环境下的WIFI室内定位系统[J].电子测量技术,2015,38(12):121-124

[5] 王道明,鲁昌华,蒋薇薇,等.基于粒子群算法的决策树SVM多分类方法研究[J].电子测量与仪器学报,2015,29(4):611-615

[6] 徐超,高梦珠,查宇锋,等.基于HOG和SVM的公交乘客人流量统计算法[J].仪器仪表学报,2015,36(2):446-452

[7] RITTER A, ETZIONI O, CLARK S. Open domain event extraction from twitter[C]. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012: 1104-1112.

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

[9] LI P, ZHU Q, DIAO H, et al. Joint modeling of trigger identification and event type determination in chinese event extraction[C]. Proceedings of COLING 2012, 2012: 1635-1652.

[10] 王健,吴雨,林鸿飞,等. 基于深层句法分析的生物事件触发词抽取[J]. 计算机工程,2014,40(1):25-30.

[11] TIAN L, MA W, ZHOU W. Automatic event trigger word extraction in chinese event[J]. Journal of Software Engineering & Applications, 2012, 5(12):208-212.

[12] 轩小星,廖涛,高贝贝. 中文事件触发词的自动抽取研究[J]. 计算机与数字工程,2015,43(3):457-461.

[13] 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.

[14] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3):993-1022.

[15] 宋俊.基于概率主题模型的话题演化与摘要生成方法研究[D].北京:中国科学院大学,2015:2-11.

[16] 田璟,郭智,黄宇,等. 一种基于多模态主题模型的图像自动标注方法[J]. 国外电子测量技术,2015,34(5):22-26.

[17] CHEN M H, SHAO Q M, IBRAHIM J G. Monte Carlo methods in Bayesian computation[M]. Springer Science & Business Media, 2012:19-66.

[18] IHLER A, HUTCHINS J, SMYTH P. Adaptive event detection with time-varying poisson processes[C]. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006: 207-216.

[19] DIAO Q, JIANG J, ZHU F, et al. Finding bursty topics from microblogs[C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 536-544.

作者简介

陈自岩,1988年出生,博士研究生,主要研究方向为数据挖掘、关键要素检测等。
E-mail:zychen0207@163.com