从 Twitter 提取开放域事件

艾伦·里特
华盛顿大学计算机科学系。
&Eng。
华盛顿州西雅图市
AdoTr.C.W.

四季倾城
华盛顿大学计算机科学系。&
Eng。
华盛顿州西雅图市 MUSAMWC.W.

山姆·克拉克*
决定公司
华盛顿州西雅图市 斯克拉尔
K.UW@ Gmail 网站

华盛顿大学计算机科学
与技术学院的 Oren
Etzioni。&Eng。
华盛顿州西雅图市
EZIONIIA.C.W.

摘要

推文是有关当前事件的最新信息和最具包容性的信息流，但它们又零散且嘈杂，激发了对可以提取，汇总和分类重要事件的系统的需求。以前提取事件的结构化表示的工作主要集中在新闻专线文本上。Twitter 的独特特征为开放域事件提取带来了新的挑战和机遇。本文介绍了 TwiCal，这是 Twitter 的第一个开放域事件提取和分类系统。我们证明，从 Twitter 准确提取重要事件的开放域日历确实是可行的。此外，我们提出了一种新颖的方法，用于根据潜在变量模型发现重要事件类别并对提取的事件进行分类。通过利用大量未标记的数据，我们的方法可使最大 F1 超出监督基准的 14%。可以在 http://statuscalendar.com 上查看我们系统的不断更新的演示。我们的 NLP 工具可在以下位置获得：http://github.com/aritter/ twitter_nlp。

类别和主题描述符

I.2.7［自然语言处理］：语言解析和理解；H.2.8［数据库管理］：数据库应用程序－数据挖掘

一般条款

算法，实验

1. 介绍

诸如 Facebook 和 Twitter 之类的社交网站提供有关当前信息的最新信息和动态

实体	活动词组	日期	类型
史蒂夫●乔布	死亡	10/6/11	死亡
苹果手机	公告	10/4/11	产品发布会
共和党	辩论	9/7/11	政治事件
阿曼达●诺克	判决书	10/3/11	审讯

表 1：TwiCal 提取的事件示例。

事件。但是，最近每天发布的推文数量已超过 2 亿，其中许多要么多余[57]，要么兴趣有限，导致信息过载。[1]显然，我们可以从事件的更结构化表示中受益。由各个推文合成。

事件提取的先前工作[21、1、54、18、43、11、7]主要集中在新闻报道上，因为从历史上看，这种类型的文本一直是有关当前事件的最佳信息来源。同时，诸如 Facebook 和 Twitter 之类的社交网站已成为此类信息的重要补充来源。尽管状态消息包含大量有用的信息，但它们却杂乱无章，激发了对自动提取，聚合和分类的需求。尽管人们对跟踪社交媒体中的趋势或模因有很大兴趣[26、29]，但很少有工作解决因从简短或非正式文本中提取事件的结构化表示而引起的挑战。

从杂乱无章的嘈杂文本语料库中提取事件的有用结构化表示形式是一个具有挑战性的问题。另一方面，个别推文简短且自成一体，因此不像包含叙事文本的情况那样由复杂的话语结构组成。在本文中，我们证明了从 Twitter 进行开放域事件提取的确是可行的，例如，如§8 所示，我们提取的最高置信度的未来事件的准确性为 90%。

Twitter 具有几个特征，这些特征为开放域事件提取任务带来了独特的挑战和机遇。

挑战：Twitter 用户经常在日常生活中提及平凡的事件（例如午餐吃的东西），而这仅是其直接社交网络感兴趣的事件。相反，如果新闻专线文字中提到某个事件，则该事件

*这项工作是在华盛顿大学进行的

[1]http://blog.twitter.com/2011/06/200 亿鸣叫每 day.html

可以肯定地说这是非常重要的。单个推文也非常简洁，通常缺乏足够的上下文来将它们分类为感兴趣的主题（例如，体育，政治，产品发布等）。此外，由于 Twitter 用户可以谈论他们选择的任何内容，因此尚不清楚哪种事件类型合适。最后，以非正式形式编写推文，导致为编辑文本而设计的 NLP 工具的效果极差。

机会：推文的简短和自成一体的性质意味着它们具有非常简单的论述和实用的结构，这些问题仍然挑战着最先进的 NLP 系统。例如，在新闻专线中，通常需要对事件之间的关系（例如，之前和之后）进行复杂的推理，才能将事件与时间表达式准确关联起来[32、8]。Tweets 的数量也比新闻数量大得多，因此可以更轻松地利用信息冗余。

为了解决 Twitter 的嘈杂风格，我们在嘈杂的文本[46，31，19]中跟踪了 NLP 的最新工作，为带有事件的推文注解，然后将其用作序列标签模型的训练数据，以识别数以百万计的事件提及消息。

由于推文的简洁性，有时是平凡的但高度冗余的性质，我们被激励着重于提取事件的聚合表示形式，这为诸如事件分类之类的任务提供了额外的上下文，并且还通过利用信息的冗余性来过滤掉了世俗的事件。我们建议将重要事件识别为那些提到的事件与对唯一日期的引用紧密相关，而不是在整个日历中均匀分布的日期。

Twitter 用户讨论了各种各样的主题，因此事先无法确定哪种事件类型适合分类。为了解决 Twitter 上讨论的事件的多样性，我们引入了一种新颖的方法来发现重要事件类型并在新域内对汇总事件进行分类。

监督或半监督的事件分类方法将需要首先设计注释准则（包括选择一组适当的类型进行注释），然后对 Twitter 中发现的大量事件进行注释。这种方法有几个缺点，因为事先不清楚应该注释哪些类型的类型。在同时完善注释标准的同时，需要大量的精力来手动注释事件集。

我们提出了一种基于潜在变量模型的开放域事件分类方法，该方法揭示了一组与数据匹配的适当类型。随后检查自动发现的类型，以过滤掉任何不连贯的类型，其余的使用信息性标签标注;[2]发现的类型的示例

图 3 中列出了使用我们的方法的结果。然后，将结果集类型应用于数亿个提取事件的分类，而无需使用任何手动注释的示例。通过利用大量未标记的数据，我们的方法可使 F1 得分比使用相同类型的一组监督基线提高 14%。

[2] 此注释和过滤工作量最小。之一作者花了大约 30 分钟的时间检查并取消了处理自动发现的事件类型。

	P	R	F_1	F1 公司
斯坦福●纳尔	0.62	0.35	0.44	-
T-seg	0.73	0.61	0.67	52%

表 2：通过对域内数据的训练，在推文中对实体进行细分时，与斯坦福命名实体识别器相比，我们的 F1 分数提高了 52%。

2. 系统总览

TwiCal 提取事件的 4 元组表示形式，其中包括命名实体，事件短语，日历日期和事件类型（请参见表 1）。选择这种表示方式是为了与重要事件在 Twitter 中通常提到的方式紧密匹配。

图中概述了我们用于从 Twitter 提取事件的系统的各个组件。

1. 给定原始的推文流，我们的系统提取与重要事件中涉及的事件短语和明确日期相关联的命名实体。首先，对推文进行 POS 标记，然后提取命名实体和事件短语，解析时间表达式，然后将提取的事件分类为类型。最后，我们根据每个命名实体共发的推文数量来衡量每个命名实体与日期之间的关联强度，以确定事件是否重大。

NLP 工具（如命名实体分段器和语音标记器的一部分）设计用于处理已编辑的文本（例如，新闻文章），由于其嘈杂且独特的样式，在应用于 Twitter 文本时，其效果非常差。为了解决这些问题，我们利用命名实体标记器和部分语音标记器，这些标记器是根据先前工作中提出的域内 Twitter 数据进行训练的[46]。我们还开发了一个事件标记器，该标记器针对第 4 节中所述的域内批注数据进行了培训。

3. 命名实体细分

NLP 工具（如命名实体分段器和语音标记器的一部分）设计用于处理已编辑的文本（例如，新闻文章），由于其嘈杂且独特的样式，在应用于 Twitter 文本时，其效果非常差。

例如，大写字母是新闻中命名实体提取的一项关键功能，但此功能在推文中高度不可靠。单词通常只是为了强调而大写，而命名实体常常全部小写。此外，由于 Twitter 的 140 个字符限制和用户的创造性拼写，tweets 包含的语音词汇比例更高。

为了解决这些问题，我们使用了一个经过命名的实体标记器，该标记器是根据先前的域内 Twitter 数据进行训练工作[46]。[3]

对推文进行培训可以极大地提高对命名实体进行细分时的性能。例如，表 2 列出了与经过最新新闻训练的斯坦福命名实体识别器[17]相比的性能。在对命名实体进行细分时，我们的系统比 Stanford Tagger 的 F1 得分提高了 52%。

4. 提取事件提及

为了从 Twitter 的嘈杂文本中提取事件提及，我们首先注释一系列推文，然后

[3] 可在 http://github.com/aritter/twitter_nlp.

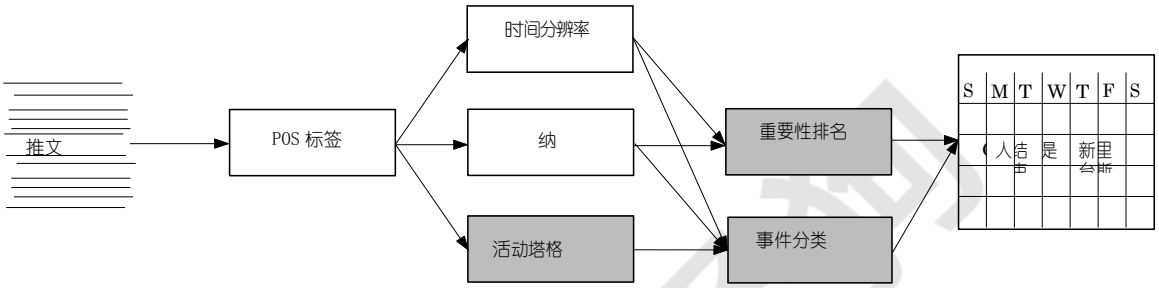

图 1：用于从 Twitter 提取事件的处理管道。在这项工作中开发的新组件以灰色阴影表示。

用于训练序列模型以提取事件。尽管我们对嘈杂的文本中的序列标记任务应用了既定方法[46，31，19]，但这是在 Twitter 中提取事件引用短语的第一项工作。

事件短语可以包含语音的许多不同部分，如以下示例所示：

- 动词：　　苹果将于 10 月 4 日发布 iPhone 5？是！
- 名词：iPhone 5 将于 10 月 4 日发布
- 形容词：今天 WOOOHOO 新 IPHONE！无法等待！

这些短语提供了重要的上下文，例如，提取实体 Steve Jobs，以及与 10 月 5 日相关的事件短语死亡，比仅提取 Steve Jobs 具有更多的信息。此外，如第 6 节所述，事件提及在上游任务中很有用，例如将事件分类为类型。

为了构建用于识别事件的标记器，我们使用与事件库中为事件标记开发的注释准则相似的注释准则为事件消息添加了 1,000 条推文（19,484 个标记）。我们使用条件随机场进行学习和推理，将事件触发识别为序列标记任务[24]。线性链 CRF 对相邻单词的预测标签之间的相关性进行建模，这对于提取多单词事件短语很有帮助。我们使用上下文，字典和正字法功能，还包括基于 Twitter 调整的 POS 标记器[46]的功能，以及 Sauri 等从 WordNet 收集的事件术语词典。[50]。

表 3 中报告了分割事件短语的精度和召回率。我们的分类器 TwiCal-Event 的 F 得分为 0.64。为了说明对域内训练数据的需求，我们将其与在 Timebank 语料库上训练系统的基线进行了比较。

5. 提取和解析时间表达

除了提取事件和相关的命名实体之外，我们还需要在事件发生时进行提取。通常，用户可以使用多种不同的方式来引用相同的日历日期，例如“ textet”，“ August 12th”，“明天”或“昨天”都可以引用同一天，具体取决于写 tweet 的时间. .为了解析时间表达式，我们使用 TempEx [33]，它需要

	精确	召回	F1
TwiCal-事件	0.56	0.74	0.64
没有 POS	0.48	0.70	0.57
时间银行	0.24	0.11	0.15

表 3：事件短语提取时的精度和召回率。通过对 1,000 个手动注释的推文（约 19K 令牌）进行 4 倍交叉验证来报告所有结果。我们将其与不使用基于 Twitter 训练的 POS Tagger 生成的功能的系统进行比较，此外，还使用了在 Timebank 语料库上训练的，使用相同功能集的系统。

作为输入的参考日期，一些文本和词性（来自我们的 Twitter 训练的 POS 标记器），并使用明确的日历参考标记时间表达。尽管这个主要基于规则的系统旨在用于新闻专线文本，但我们发现其在 Tweets 上的精度（94%-估计为 268 个提取示例）足够高，可以用于我们的目的。TempEx 在推文上的高精度可以由以下事实解释：某些时间表达相对明确。尽管通过处理嘈杂的时间表达似乎有改善 Twitter 上时间提取的回忆的空间（例如，有关“明天”一词的 50 多种拼写变化的列表，请参见 Ritter 等人[46]），我们离开将时间提取应用于 Twitter 作为潜在的未来工作。

6. 事件类型分类

为了将提取的事件分类为类型，我们提出了一种基于潜在变量模型的方法，该方法可以推断一组适当的事件类型以匹配我们的数据，还可以通过利用大量未标记的数据将事件分类为各种类型。

由于多种原因，事件类别的受监督或半监督分类存在问题。首先，先验不清楚哪种类别适合 Twitter。其次，需要大量的人工来注释带有事件类型的推文。第三，重要类别（和实体）的集合可能会随时间或在关注的用户人口统计范围内转移。最后，许多重要的类别相对很少见，因此，即使是大型的带注释的数据集也可能仅包含这些类别的几个示例，从而使分类变得困难。

由于这些原因，我们有动机进行调查，

体育	7.45%	冲突	0.69%
派对	3.66%	奖	0.68%
电视	3.04%	法律	0.67%
政治	2.92%	死亡	0.66%
名人	2.38%	拍卖	0.66%
音乐	1.96%	视频游戏	0.65%
电影	1.92%	毕业	0.63%
餐饮	1.87%	赛跑	0.61%
音乐会	1.53%	筹款/驱动器	0.60%
性能	1.42%	展示	0.60%
身体素质	1.11%	庆典	0.60%
访问	1.01%	图书	0.58%
的	0.95%	电影	0.50%
会议	0.88%	打开/关闭	0.49%
时尚	0.87%	婚礼	0.46%
金融	0.85%	假日	0.45%
学校	0.85%	医	0.42%
白蛋白释放	0.78%	摔角	0.41%
宗教	0.71%	其他	53.45%

图 2：自动发现的事件类型的完整列表，其中包含数据百分比。代表重要事件的可解释类型覆盖了大约一半的数据。

受监督的方法将自动引发与数据匹配的事件类型。我们采用基于潜在变量模型的方法，该模型的灵感来自对选择偏好进行建模的最新工作[47、39、22、52、48]，并且不受监督信息提取[4，55，7]。

我们数据（e）中的每个事件指示器短语均建模为类型的混合。例如，事件短语“欢呼”可能会出现在“政治事件”或“体育事件”中。每种类型除了对应于发生类型事件的日期 d 上的分布外，还对应于该类型的特定实例所涉及的命名实体 n 上的分布。在我们的模型中包括日历日期，具有鼓励（尽管不是必须）在同一日期发生的事件被分配为相同类型的作用。这有助于指导推理，因为对同一事件的不同引用也应具有相同的类型。

我们数据的生成故事基于 LinkLDA [15]，并作为算法 1 表示。此方法的优点是，在提及的事件中共享有关事件短语的类型分布的信息，同时也自然保留了歧义。另外，由于该方法基于生成的概率模型，因此直接执行关于数据的许多不同的概率查询。例如，在对聚合事件进行分类时，这很有用。

为了进行推断，我们使用折叠的 Gibbs 采样[20]，其中依次对每个隐藏变量 zi 进行采样，并对参数进行积分。示例类型显示在图 3 中。为了估计给定事件在类型上的分布，在充分烧入后，从 Gibbs markov 链中获取相应隐藏变量的样本。使用流方法进行新数据的预测[56]。

6.1 评估

为了评估模型对重大事件进行分类的能力，我们收集了 6,500 万个以下形式的提取事件

标签	前 5 个活动短语	前 5 名实体
体育	后挡板-混战-拖尾-回家-常规赛	espn-ncaa-老虎-ea-高兴-大学
音乐会	音乐会-预售-每个表格-音乐会-门票	泰勒●斯威夫特-多伦多-布兰妮斯皮尔斯-蕾哈娜-岩石
演出	日场音乐-邪恶的眼看-普里西	史瑞克-les mis-lee埃文斯-邪恶-百老汇
电视	新赛季-赛季 FI-nale-完成季-剧集-新剧集	泽西海岸-真血-欢乐合唱团-dvr-hbo
电影	看爱-对话主题-盗版-大厅通行证-电影	netflix-黑天鹅-in-阴险-特隆-斯科特朝圣者
体育	一局局-投手本本垒打-垒打	毫升-红袜-洋基-双胞胎-dl
政治	总统辩论-大阪-总统候选人-共和党辩论-辩论表演	奥巴马-主席奥巴马-中国-中国-美国
电视	网络新闻广泛演员-播出-黄金时段戏剧-频道-流	美国广播公司-美国广播公司-美国广播公司-福克斯-音乐电视
产品	揭幕-揭幕-an-名词-发射-结束	苹果-Google-mi-crosoft-英国-索尼
会议	显示交易-大厅-mtg-分区-简介	市政厅-市政厅-俱乐部-商业-白宫
金融	股票-下跌-交易-报告-高开-下跌	路透社-纽约-美国-中国-欧元
学校	数学-英语测试-考试-修订-物理	英语-数学-ger-男人-生物-Twitter
专辑	在商店中-专辑发行-首张专辑-下降-热门商店	itunes-ep-英国-亚马逊-光盘
电视	投票-偶像-scotty-偶像季-派息	Lady Gaga-美国偶像-美国-碧昂斯-欢乐
宗教	讲道说教-宣讲崇拜-	教堂-耶稣-牧师-信仰-神
冲突	宣战-战争-炮击-开火-受伤	利比亚-阿富汗-＃叙利亚-叙利亚-北约
政治	参议院-立法-重新呼吁-预算-选举	参议院-众议院-国会-奥巴马-天哪
奖	中奖者-乐透结果-输入-获胜者-竞赛	iPad-奖项-Facebook-祝你好运-获胜者
法律	保释请求-谋杀案审判-被判刑-认罪-被定罪	凯西●安东尼-法院-印度-新德里-最高法院
电影	电影节-放映-主演-电影-小鹅	好莱坞-纽约-洛杉矶-洛杉矶
死亡	永远活着-通过离开-悲伤的消息-慰问-埋葬	迈克尔杰克逊-阿富汗-约翰●列侬-年轻-和平
拍卖	加-50%减-加-运送-节省	groupon-早起的鸟儿-脸书-@etsy-etsy
驾驶	捐赠-龙卷风救济-relief 灾-捐赠-筹集资金	日本-红十字会-乔普林-6 月-非洲

图 3：我们的模型发现的示例事件类型。对于每种类型 t，我们列出给定 t 的概率最高的前 5 个实体，以及将 t 的概率最高的 5 个事件短语。

算法 1 我们的数据的生成故事，涉及事件类型作为隐藏变量。贝叶斯推断技术被用于反转生成过程并推断出一组适当的类型来描述观察到的事件。

对于每个事件类型 t = 1 ... T do
  产生 β^n_t 根据对称 Dirichlet 分布 Dir（ηn）。
  产生 β^d_t 根据对称 Dirichlet 分布 Dir（ηd）。
结束于
对于每个唯一事件短语 e = 1 ... | E |做
  根据 Dirichlet 分布 Dir（α）生成 θe。
  对于每个与 e 共同出现的实体，i = 1 ... Ne do
    产生 $z^n_{e,i}$ 来自多项式（θe）。
    从多项式（βzn）生成实体 $ne_{e,i}$，i。
  结束于
  对于与 e 共同出现的每个日期，i = 1 ... Nd
    生成 $z^d_{e,i}$ 来自多项式（θe）。
    从多项式（βzn）生成日期 $de_{d,i}$，i。
  结束于
结束于

在图 1 中列出（不包括类型）。然后，我们运行了 100 种类型的 Gibbs 采样，进行了 1000 次 Burnin 迭代，并保留了在上一个示例中找到的隐藏变量分配。[4]

一位作者手动检查了产生的类型，并根据它们在实体上的分布以及为该类型分配最高概率的事件词，为它们分配了标签，例如体育，政治，音乐发行等。在这 100 种类型中，我们发现 52 种对应于涉及重大事件的连贯事件类型；[5]其他类型要么是不连贯的，要么是没有广泛关注的涵盖事件类型，例如存在与用户讨论事件相对应的短语，例如，应用，呼叫，联系，工作面试等。

与找工作有关。此类事件类型与普遍关注的重大事件不对应，仅标记为“其他”。图 2 列出了用于注释自动发现的事件类型的标签的完整列表以及每种类型的覆盖范围。请注意，标签对类型的分配只需要执行一次，并为任意数量的标签生成标签。事件实例。另外，可以使用流推论技术轻松地使用同一组类型对新事件实例进行分类[56]。未来工作的一个有趣方向是自动标记和自动发现的事件类型的一致性评估，类似于主题模型的最新工作[38，25]。

为了评估我们的模型对聚合事件进行分类的能力，我们将出现的 20 倍或更多倍的所有（实体，日期）对分组在一起，然后使用发现的事件类型对关联性最高的 500（请参阅§7）进行注释。根据我们的模型。

为了帮助证明利用大量未标记数据进行事件分类的好处，我们将其与监督的最大熵基线进行了比较，该基线利用 10 倍交叉验证利用了 500 个带注释事件。对于功能，我们将处理事件短语集

[4] 为了扩展到更大的数据集，我们使用类似于 Newmann 等人提出的 Gibbs 采样程序的近似值，对 40 个核并行执行了推理。人。[37]。

[5] 标记后，将某些类型组合在一起，得到 37 个不同的标记。

	精确	召回	F_1
TwiCal-分类	0.85	0.55	0.67
监督基准	0.61	0.57	0.59

表 4：事件类型类别的精确度和召回率最大 F1 时的化 得分。

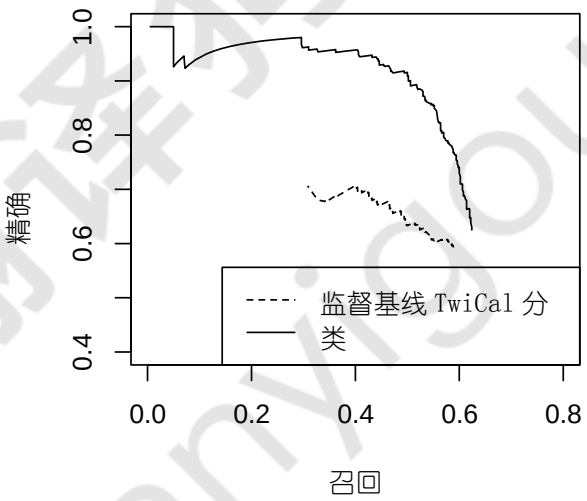

图 4：精确度和召回率预测事件类型。

与每个（实体，日期）对同时出现的单词袋，并且还包括关联的实体。由于许多事件类别很少见，因此某个类别的培训示例很少或没有，导致性能低下。

图 4 通过改变最可能类型概率的阈值获得的精确召回曲线，比较了我们的无监督方法与监督基线的性能。另外，表 4 比较了最大 F 分数时的精度和召回率。我们对事件进行分类的无监督方法使最大 F1 得分比受监督的基线提高了 14%。图 5 绘制了随着基线使用的训练数据量的变化而变化的最大 F1 分数。似乎有了更多的数据，性能将达到不使用任何带注释事件的方法的性能，但是我们的方法会自动发现合适的事件类型集，并以最小的努力提供初始分类器，从而使其在第一步-注释数据不立即可用的位置。

7. 排名事件

仅使用频率来确定哪些事件是重要的是不够的，因为许多推文涉及用户日常生活中的常见事件。例如，用户经常提到午餐时吃什么，因此，像麦当劳这样的实体与大多数日历日的引用相对频繁地出现。重要事件可以与那些与唯一日期有强烈关联的事件区分开来，而不是在日历的各天之间平均分配。为了从 Twitter 提取人们普遍关注的重大事件，因此，我们需要某种方法来衡量实体与日期之间的关联强度。

为了测量之间的关联强度

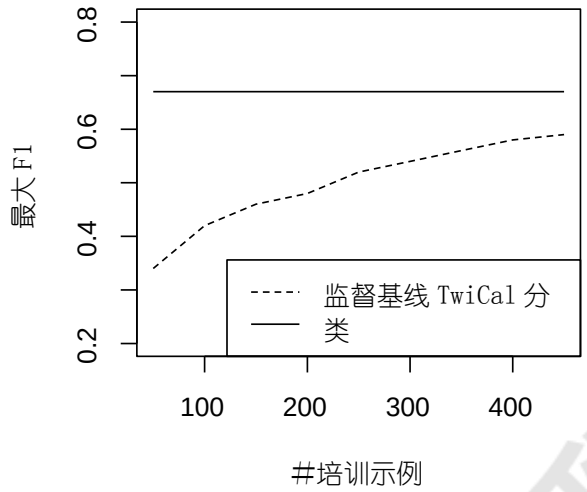

图 5：随着训练数据量的变化，监督基线的最大 F1 分数。

实体和特定日期，我们利用 G^2 对数似然比统计量。有人认为 G^2 比 χ^2[12]更适合于文本分析任务。尽管 Fisher 精确检验会产生更准确的 p 值[34]，但鉴于我们正在使用的数据量（样本大小大于 10^{11}），事实证明，计算 Fisher 精确检验统计信息十分困难，这会导致浮动点溢出，即使使用 64 位操作也是如此。G^2 测试在我们的环境中可以很好地工作，但是，作为计算关联
实体和日期之间产生的稀疏偶然性较少

表格，而不是使用成对的实体（或单词）。G^2检验基于实体以日期为条件的模型与实体与日期参考之间的独立性模型之间的似然比。
对于给定的实体 e 和日期 d，该统计量可以按以下方式计算：

$$G^2 = \sum_{x\in\{e,\neg e\},y\in\{d,\neg d\}} \text{氧}, y \times ln \frac{\text{牛年}, Y}{\text{当然}, Y}$$

其中 Oe，d 是包含
e 和 d 均为 Oe，¬d，是包含 e 但不包含 d 的 tweet 的观察分数，依此类推。类似地，Ee，d 是包含 e 和 d 的推文的预期分数，假设具有独立性模型。

8. 实验

为了估算使用我们的方法生成的日历项的质量，我们手动评估了 11 月 3 日未来两周内出现的前 100、500 和 1,000 个日历项的样本。

8.1 数据

出于评估目的，我们在 2011 年 11 月 3 日收集了大约 1 亿条最新推文（使用 Twitter Streaming API[6] 收集），并跟踪了一系列临时关键词，包括“今天”，“明天”，工作日名称，个月等）。

我们从每个 100M 的文本中提取了事件短语和时态表达式之外的命名实体

[6]https://dev.twitter.com/docs/streaming-api

鸣叫。然后，我们将提取的三元组添加到第 6 节中所述的用于推断事件类型的数据集中，并执行了 50 次 Gibbs 采样迭代，以预测新数据上的事件类型，同时使原始数据中的隐藏变量保持不变。这种流式推论方法与 Yao 等人提出的方法类似。[56]。

然后，按照§7 中的描述对提取的事件进行排名，并从排名最高的 100、500 和 1,000 中随机抽取 50 个事件。我们用 4 个单独的标准注释了事件：

1. 是否有涉及提取实体的重大事件会在提取日期发生？
2. 最频繁提取的事件短语是否有用？
3. 事件的类型是否正确分类？
4. （1-3）中的每一个都正确吗？也就是说，事件是否包含正确的实体，日期，事件短语和类型？

请注意，如果（1）对于特定事件被标记为不正确，则后续条件始终被标记为不正确。

8.2 底线

为了证明自然语言处理和信息提取技术在提取信息事件中的重要性，我们将其与不使用 Ritter 等人的简单基准进行比较。人。命名实体识别器或我们的事件识别器；而是将每条推文中的 1-4 克全部视为候选日历条目，
2

依靠 G 测试以过滤掉短语数量较少的词组与每个日期关联。

8.3 结果

评估结果显示在表 5 中。该表显示了不同产量级别（聚集事件的数量）下系统的精度。这些是通过更改 G^2 统计信息中的阈值获得的。请注意
基线仅可与第三列进行比较，即
（实体，日期）对的精度，因为基线未执行事件识别和分类。虽然

在某些情况下，ngram 确实与信息日历条目相对应，与我们的系统相比，ngram 基线的精度非常低。

在许多情况下，ngram 不对应于与事件相关的显着实体。它们通常由难以解释的单个词组成，例如 11 月 18 日上映的电影《暮光之城：破晓》中的“突破”一词，尽管“突破”一词与 11 月 18 日有着很强的联系呈现给用户的信息不是很丰富。[7]

我们的高可信度日历条目具有出乎意料的高质量。如果我们将数据限制为将来两周内的日期范围内排名最高的 100 个日历条目，则提取的（实体，日期）对的精度非常好（90%）-比 ngram 基线增加 80%。随着预期的精度下降，显示更多的日历条目，但是

[7] 此外，我们注意到 ngram 基线倾向于生成许多几乎重复的日历条目，例如：“暮光之城”，“突破黎明”和“暮光之城黎明”。尽管这些条目中的每一个都被注释为正确，但是将如此多的条目描述给用户却是有问题的。

2011 年 11 月						
11 月 7 日星期	11 月 8 日，	11 月 9 日	11 月 10 日，	星期五十一月	星期六 11 月 12	星期日十一月 13
贾斯汀 *遇到* *其他*	巴黎 *爱* *其他*	大通 *测试* *其他*	罗伯特·帕丁森 *节目*	苹果手机 *登场* *产品发布*	悉尼 *演出* *其他*	游戏机 *答案* *产品发布*
摩托罗拉 Pro + *踢*	苹果手机 *保持* *产品发布*	美联储 *隔断* *其他*	詹姆斯·默多克 *提供证据* *其他*	纪念日 *打开* *性能*	铂尔曼夏会厅 *提拔* *其他*	三星 Galaxy Tab *发射* *产品发布*
角落颜色 2 *发射* *产品发布*	选举日 *投票* *政治事件*	托卡里维拉 *提拔* *性能*	RTL TVI *岗位* *电视节目*	法国 *玩* *其他*	狐狸 *斗争* *其他*	索尼 *答案* *产品发布*
开斋节 UL-阿扎 *著名* *性能*	蓝色滑梯公园 *听* *音乐发行*	警报系统 *测试* *其他*	哥蒂生活 *工作* *其他*	退伍军人节 *关闭* *其他*	广场 *派对* *派对*	《 K.O.小拳王》 Chibi Burger
MW3 *午夜发布* *其他*	赫德利 *专辑* *音乐发行*	最长天数 *给* *其他*	小鹿斑比奖 *演出* *性能*	天际 *到达* *产品发布*	红地毯 *邀请* *派对*	杰克斯波·克马约兰 *提拔*

图 6：我们的系统在 11 月 7 日当周提取的未来日历条目示例。数据收集至 11 月 5 日。对于每一天，我们都会列出前 5 个事件，包括实体，事件短语和事件类型。尽管存在一些错误，但大多数日历条目都是有益的，例如：穆斯林假期 eid-ul-azha，发行了一些视频游戏：Modern Warfare 3（MW3）和 Skyrim，以及发行了新的 Playstation。 11 月 13 日推出 3D 显示屏，11 月 11 日在香港推出新款 iPhone 4S。

	精确				
#个日历条目	ngram 基线	实体+日期	事件词组	事件类型	实体+日期+事件+类型
100	0.50	0.90	0.86	0.72	0.70
500	0.46	0.66	0.56	0.54	0.42
1,000	0.44	0.52	0.42	0.40	0.32

表 5：在不同召回水平下的准确性评估（通过更改 G^2 统计数据的阈值生成）。我们评估前 100、500 和 1,000（实体，日期）对。另外，我们评估最频繁提取的事件短语的精度，以及与这些日历条目关联的预测事件类型。还列出了所有预测（“实体+日期+事件+类型”）正确的情况下所占的比例。我们还将比较不使用我们的 NLP 工具的简单 ngram 基线的精度。请注意，ngram 基线仅与实体+日期精度（第 3 列）相当，因为它不包含事件短语或类型。

保持足够高以显示给用户（在排名列表中）。除了不太可能来自提取错误外，排名较高的实体/日期对也更可能与流行事件或重要事件相关，因此用户更加感兴趣。

另外，我们在图 6 中的日历上提供了提取的未来事件的样本，以给出如何将其呈现给用户的示例。除了最频繁提取的事件短语和最高概率事件类型之外，我们还提供与每个日期关联的前 5 个实体。

8.4 错误分析

我们发现实体/日期对在日历上显示时信息不灵通的两个主要原因：

分割错误某些提取的“实体”或 ngram 不对应于命名实体，或者由于它们的分割错误而通常无用。示例包括“ RSVP”，“突破”和“ Yikes”。

实体与日期之间的关联性较弱在某些情况下，实体被适当地分割了，但是却没有提供信息，因为它们与关联日期上的特定事件没有强烈关联，或者涉及当天发生的许多不同事件。示例包括“纽约”之类的位置，以及“ Twitter”之类的经常提及的实体。

9. 相关工作

虽然我们是第一个研究 Twitter 内部开放域事件提取的公司，但是有两个关键的相关研究领域：从 Twitter 提取特定类型的事件，以及从新闻中提取开放域事件[43]。

最近，人们对 Twitter 中的信息提取和事件识别非常感兴趣。本森等。[5]使用远程监督来训练关系提取器，该提取器识别在将其位置列为纽约市的用户的推文中提到的艺术家和地点。Sakaki 等。[49]训练分类器来识别报告日本地震的推文；它们证明了其系统能够识别日本气象厅报告的几乎所有地震。另外，最近在 Twitter 中进行了有关检测事件或跟踪主题的工作[29]，该工作不提取结构化表示，但具有的优点是它不限于狭窄的域。Petrovi´c 等。研究了一种识别推文的流方法，该方法是第一个使用本地敏感哈希函数报告突发新闻的方法[40]。贝克尔等。[3]，Popescu 等。[42，41]和 Lin 等。[28]研究发现与正在发生的事件相对应的相关单词或推文的群集。与先前有关 Twitter 事件识别的工作相比，我们的方法与事件类型或域无关，因此可以更广泛地应用。此外，我们的工作重点是提取事件日历（包括将来发生的事件），

引用事件引用表达式并将事件分类为类型。

与事件识别相关的工作[23、10、6]，以及从新闻文章中提取时间表[30]也很相关。[8] 与新闻文章相比，Twitter 状态消息既带来了独特的挑战，也带来了机遇。Twitter 的嘈杂文本对 NLP 工具提出了严峻挑战。另一方面，它包含对当前日期和将来日期的更高比例的引用。推文不需要将事件之间的关系进行复杂的推理就可以将它们放置在时间线上，这在包含叙述的长文本中通常是必需的[51]。此外，与新闻不同，推文通常讨论的是普通事件，而这些事件并不是人们普遍关注的，因此利用信息冗余来评估事件是否很重要至关重要。

以前有关开放域信息提取的工作[2、53、16]主要集中于从 Web 语料库中提取关系（相对于事件），并且还基于动词提取关系。相比之下，这项工作使用适合 Twitter 嘈杂文本的工具提取事件，并提取通常是形容词或名词的事件短语，例如：2 月 5 日的超级碗派对。

最后，我们注意到，最近越来越有兴趣将 NLP 技术应用于非正式的短消息，例如 Twitter 上的消息。例如，最近的工作探索了词性标注[19]，在 Twitter 上发现的语言地域差异[13、14]，对非正式对话进行建模[44、45、9]，还应用了 NLP 技术来帮助危机工作者自然灾害后的信息泛滥[35，27，36]。

10. 结论

我们提出了一种可扩展的开放域方法，用于从状态消息中提取事件并对其进行分类。我们在手动评估中评估了这些事件的质量，显示出与 ngram 基线相比，性能有了明显改善

我们提出了一种新颖的方法来对类型未知的开放域文本类型中的事件进行分类。我们基于潜在变量模型的方法首先发现与数据匹配的事件类型，然后将其用于分类汇总事件，而无需任何带注释的示例。由于此方法能够利用大量未标记的数据，因此其性能比监督基准高出 14%。

将来工作的可能途径是提取更多更丰富的事件表示，同时保持域独立性。例如：将相关实体分组在一起，根据事件中实体的角色对实体进行分类，从而提取事件的基于帧的表示形式。

可以在 http://statuscalendar.com 上查看我们系统的不断更新的演示。我们的 NLP 工具可在以下位置获得：http://github.com/aritter/twitter_nlp.

[8] http://newstimeline.googlelabs.com/

11. 确认

作者要感谢 Luke Zettlemoyer 和匿名审阅者对先前的草稿提供了有益的反馈。这项研究得到了 NSF IIS-0803481 和 ONR N00014-08-1-0431 的部分资助，并在华盛顿大学图灵中心进行了研究。

12. 引用

[1] J. Allan，R。Papka 和 V. Lavrenko。在线新事件检测和跟踪。在 SIGIR，1998 年。

[2] M. Banko，MJ Cafarella，S。Soderl，M。Broadhead 和 O. Etzioni。从网上打开信息提取。在 In IJCAI 中，2007 年。

[3] H. Becker，M。Naaman 和 L. Gravano。超越热门话题：Twitter 上的真实事件识别。在 ICWSM 中，2011 年。

[4] C. Bejan，M。Titsworth，A。Hickl 和 S. Harabagiu。用于非监督事件共参考解析的非参数贝叶斯模型。在 NIPS 中。2009 年。

[5] E. Benson，A。Haghighi 和 R. Barzilay。社交媒体源中的事件发现。在 ACL 中，2011 年。

[6] S. Bethard 和 JH Martin。识别事件提及及其语义类别。在 EMNLP 中，2006 年。

[7]钱伯斯（N. Chambers）和尤拉夫斯基（D. Jurafsky）。没有模板的基于模板的信息提取。在 ACL 会议录中，2011 年。

[8]钱伯斯（N. Chambers），王（S. Wang）和尤拉夫斯基（D. Jurafsky）。分类事件之间的时间关系。在 ACL 中，2007 年。

[9] C. Danescu-Niculescu-Mizil，M。Gamon 和 S.杜迈斯记住我的话！社交媒体中的语言风格调节。在 WWW 诉讼中，第 745-754 页，2011 年。

[10] A. Das Sarma，A。Jain 和 C. Yu。动态关系和事件发现。在 WSDM 中，2011 年。

[11] G. Doddington，A。Mitchell，M。Przybocki，L. Ramshaw，S。Strassel 和 R. Weischedel。自动内容提取（ACE）程序-任务，数据和评估。LREC，2004 年。

[12] T. Dunning。统计意外和巧合的准确方法。COMPUT。语言学家，1993 年。

[13] J. Eisenstein，B。O'Connor，NA Smith 和 EP Xing。地理词汇变化的潜在变量模型。在 EMNLP 中，2010 年。

[14] J. Eisenstein，NA Smith 和 EP Xing。发现具有结构性稀疏性的社会语言联系。在 ACL-HLT 中，2011 年。

[15] E. Erosheva，S。Fienberg 和 J. Lafferty。科学出版物的混合成员模型。*p*nas，2004 年。

[16] A. Fader，S。Soderland 和 O. Etzioni。识别开放信息提取的关系。在 EMNLP 中，2011 年。

[17] JR Finkel，T。Grenager 和 C. Manning。通过 gibbs 采样将非本地信息纳入信息提取系统。在 ACL 中，2005 年。

[18] E. Gabrilovich，S。Dumais 和 E. Horvitz。Newsjunkie：通过分析信息新颖性来提供个性化的新闻源。在 WWW 中，2004 年。

[19] K. Gimpel，N。Schneider，B。O'Connor，D。Das，D. 米尔斯，J.艾森斯坦，M.海尔曼，D.瑜伽塔马，J. Flanigan 和 NA Smith。词性标记

对于 Twitter：注释，功能和实验。在 *ACL*，2011 年。

[20] TL Griffiths 和 M. Steyvers。寻找科学课题。美国国家科学院院刊，2004 年 1 月 101 日增刊。

[21] R. Grishman 和 B. Sundheim。消息理解会议-6：简要历史。在《国际计算语言学会议论文集》（1996 年）中。

[22] Z. Kozareva 和 E. Hovy。使用递归模式学习语义关系的论点和超类型。在 ACL 中，2010 年。

[23] G. Kumaran 和 J. Allan。用于新事件检测的文本分类和命名实体。在 SIGIR，2004 年。

[24] JD Lafferty，A。McCallum 和 FCN Pereira。条件随机字段：用于分割和标记序列数据的概率模型。在 ICML 中，2001 年。

[25]刘建勋，K。格里瑟，D。纽曼和 T.鲍德温。自动标记主题模型。在 ACL 中，2011 年。

[26] J. Leskovec，L。Backstrom 和 J. Kleinberg。模因跟踪和新闻周期动态。在 *k*dd，2009 年。

[27] W. Lewis，R。Munro 和 S. Vogel。危机 mt：为危机情况下的 mt 开发食谱。在第六届统计机器翻译研讨会论文集，2011 年。

[28] CX Lin，B。Zhao，Q。Mei 和 J. Han。PET：用于跟踪社交社区中流行事件的统计模型。在 KDD 中，2010 年。

[29] J. Lin，R。Snow 和 W. Morgan。自适应在线语言模型的平滑技术：tweet 流中的主题跟踪。在 KDD 中，2011 年。

[30] X. Ling 和 DS Weld。时间信息提取。在 AAAI，2010 年。

[31] X. Liu，S。Zhang，F。Wei 和 M. Zhou。识别推文中的命名实体。在 ACL 中，2011 年。

[32] I. Mani，M。Verhagen，B.Wellner，CM CM Lee 和 **J.** Pustejovsky。时间关系的机器学习。在 ACL 中，2006 年。

[33]马尼（I. Mani）和威尔逊（G. Wilson）。对新闻进行鲁棒的临时处理。在 ACL 中，2000。

[34] RC 摩尔。关于对数似然比和罕见事件的意义。在 EMNLP 中，2004 年。

[35] R. Munro。子词和时空模型，用于识别海地 Kreyol 中的可行信息。在 CoNLL，2011 年。

[36] G. Neubig，Y。Matsubayashi，M。Hagiwara 和 **K.**村上。安全信息挖掘-NLP 在灾难中可以做什么-在 IJCNLP 中，2011 年。

[37] D. Newman，非盟亚松森，P。Smyth 和 M.威灵。潜在狄利克雷分配的分布式推理。在 NIPS 中，2007 年。

[38] D. Newman，JJ Lau，K。Grieser 和 T. Baldwin。自动评估主题的连贯性。在 *hlt-naacl，2010 年。*

[39] D. Ó´S´eaghdha。选择偏好的潜在变量模型。在 ACL 中，ACL '10，2010。

[40] S. Petrovic，M。Osborne 和 V. Lavrenko。流第一个故事检测，并将其应用到 Twitter。在 HLT-NAACL 中，2010 年。

[41] A.-M.Popescu 和 M. Pennacchiotti。与明星共舞，nba 游戏，政治：Twitter 用户对事件的响应的探索。在 ICWSM 中，2011 年。

[42] A.-M.Popescu，M。Pennacchiotti 和 DA Paranjpe。从 Twitter 提取事件和事件描述。在 WWW 中，2011 年。

[43] J. Pustejovsky，P。Hanks，R。Sauri，A。参见，R. Gaizauskas，A。Setzer，D。Radev，B.Sundheim，D. Day，L。Ferro 和 M. Lazo。TIMEBANK 语料库。在《语料库语言学进展》2003 年，2003 年中。

[44] A. Ritter，C。Cherry 和 B. Dolan。Twitter 对话的无监督建模。在 HLT-NAACL 中，2010 年。

[45] A. Ritter，C。Cherry 和 WB Dolan。社交媒体中数据驱动的响应生成。在 EMNLP 中，2011 年。

[46] A. Ritter，S。Clark，Mausam 和 O. Etzioni。推文中的命名实体识别：一项实验研究。EMNLP，2011 年。

[47] A. Ritter，Mausam 和 O. Etzioni。一种针对选择偏好的潜在狄利克雷分配方法。在 ACL 中，2010 年。

[48] K. Roberts 和 SM Harabagiu。选择限制的无监督学习和论点强制的检测。在 EMNLP 中，2011 年。

[49] T. Sakaki，M。Okazaki 和 Y. Matsuo。地震动摇了 Twitter 用户：通过社交传感器进行实时事件检测。在 2010 年的 WWW 中。

[50] R. Saur´ı，R。Knippen，M。Verhagen 和 **J.** Pustejovsky。Evita：用于 qa 系统的强大事件识别器。在 HLT-EMNLP 中，2005 年。

[51] Song 和 R. Cohen。叙事语境中的时态解释。在第九届全国人工智能会议论文集-第 1 卷，AAAI'91，1991 年。

[52] B. Van Durme 和 D. Gildea。用于以语料库为中心的知识概括的主题模型。2009 年，罗彻斯特罗彻斯特大学计算机科学系在技术报告 TR-946 中。

[53] DS Weld，R。Hoffmann 和 F. Wu。使用维基百科来引导开放信息提取。SIGMOD 建议，2009 年。

[54] Y. Yang，T。Pierce 和 J. Carbonell。回顾性和在线事件检测的研究。在第 21 届国际 ACM SIGIR 信息检索研究与开发会议记录中，SIGIR '98，1998 年。

[55] L. Yao，A。Haghighi，S。Riedel 和 A. McCallum。使用生成模型的结构化关系发现。在 EMNLP 中，2011 年。

[56] L. Yao，D。Mimno 和 A. McCallum。用于流文档集合的主题模型推断的有效方法。在 KDD 中，2009 年。

[57] FM Zanzotto，M。Pennaccchiotti 和 **K.**Tsioutsiouliklis。Twitter 中的语言冗余。在 EMNLP 中，2011 年。