基于事件实例驱动的新闻文本事件抽取

许旭阳 李弼程 张先飞 韩永峰
(解放军信息工程大学信息工程学院 郑州 450002)

摘 要 目前,事件抽取的流行方法是以事件元素或触发词进行驱动,但该方法容易导致正反例不平衡,且在语料库规模较小时存在一定的数据稀疏问题。提出了一种基于事件实例驱动的事件抽取方法。首先,从文档句子中抽取出刻画一个事件发生有代表性的特征,构成候选事件实例表示;其次,通过二元分类器对新闻文本中的事件实例与非事件实例进行分类;最后,对事件实例采用基于层次聚类的k-medoids算法完成事件抽取。该方法不仅克服了正反例失衡以及数据稀疏问题,而且解决了预先定义事件类别的局限性。实验结果验证了该方法的有效性,对比传统方法,事件抽取的准确率与召回率均获得了显著的提高。

关键词 事件实例,分类,新闻文本,聚类,事件抽取
中图法分类号 TP391 **文献标识码** A

News Text Event Extraction Driven by Event Sample

XU Xu-yang LI Bi-cheng ZHANG Xian-fei HAN Yong-feng
(Information Engineering Institute, PLA Information Engineering University, Zhengzhou 450002, China)

Abstract At present, popular methods of event extraction regard event arguments or triggers as drivers, but they may cause positive and negative samples imbalance. Furthermore, there will be data sparseness problem when the corpus is small. This paper proposed an event extraction method driven by event sample. Firstly, features of event samples were extracted from news text sentences to compose the description of candidate event. Secondly, event samples and non-event samples of news text were classified through binary classification. Finally, event samples were clustered by hierarchical and k-medoids clustering algorithm to complete event extraction. The method not only overcomes positive and negative samples imbalance and data sparseness problem, but also resolves the limit of pre-defined event types. Experimental results indicate that the proposed method is effective, improves precision and recall of event extraction compared to traditional methods.

Keywords Event sample, Classification, News text, Clustering, Event extraction

1 引言

事件抽取(Event Extraction)隶属于信息抽取领域,主要研究如何把含有事件信息的非结构化文本以结构化的形式呈现出来。它涉及自然语言处理、数据挖掘、机器学习等多个学科的技术和方法,在自动摘要[1]、信息检索[2]等领域均有着广泛的应用。因此,事件抽取技术的研究具有重要的现实意义。

目前,事件抽取的主要方法有两种:模式匹配方法和机器学习方法。

模式匹配方法采用模式匹配算法将待抽取的事件与固定模板匹配,如上海交通大学的冯礼[3]提出的基于事件框架的突发事件信息抽取。这种方法准确率较高,且接近人的思维方式,但依赖于具体领域及文本格式,可移植性较差,性价比不高。

机器学习方法是一种基于统计的方法,将事件抽取看作分类问题,主要包括事件类别识别和事件元素识别两个步骤。事件类别由预先定义的事件模板决定,自动内容抽取(Automatic Content Extraction, ACE)评测会议[4]定义了8个事件大类及34个子类,每种事件类别对应唯一的事件模板。该类方法克服了模式匹配方法的缺陷,比较客观,不需要太多的人工干预和领域知识。因此,机器学习成为当前事件抽取多采用的研究方法。2002年,Chieu[5]首次在事件抽取中引入最大熵分类器,用于事件元素的识别。2006年,Ahn[2]结合MegaM和Timbl两种机器学习方法识别文本中的事件触发词,然后进行分类,完成事件抽取。国内事件抽取的代表有:2008年,赵妍妍[6]基于触发词扩展和二元分类相结合的事件抽取方法;2010年,张先飞[7]基于触发词指导的自相似度聚类的事件抽取方法。然而,无论以事件元素还是以触发词作为实例进行驱动,均会引入大量的反例,造成正反例失衡和数据稀疏。

到稿日期:2010-09-06 返修日期:2010-12-28 本文受国家社科重大基金项目(09&ZD014)和国家863项目(2007AA01Z439)资助。
许旭阳(1985—),男,硕士生,主要研究方向为事件抽取、自动文本摘要,E-mail:xuxuyang_88@163.com;李弼程(1970—),男,教授,主要研究方向为智能信息处理及语音信号处理;张先飞(1981—),男,博士生,主要研究方向为文本信息处理、自动内容抽取与数据挖掘;韩永峰(1984—),男,硕士生,主要研究方向为信息抽取。

针对以上问题，本文避开以事件元素和触发词来驱动进行事件抽取，提出一种基于事件实例驱动的事件抽取方法。首先，将新闻文本中的每个句子作为一个候选事件，从句子中抽取出刻画一个事件发生的有代表性的特征，将其构成候选事件实例表示；其次，利用支持向量机（Support Vector Machine，SVM）[8]对新闻文本中的事件实例与非事件实例进行区分，过滤非事件实例；最后，用基于层次聚类 k-medoids 算法对事件实例聚类，实现新闻文本中的事件抽取。实验结果验证了本文方法的有效性，为事件抽取提供了一种新的思路。

2 事件实例识别

统计表明，新闻文本中包含大量非事件实例，降低了事件抽取的准确率，因此需要尽可能地过滤掉非事件实例。

2.1 事件定义

“事件”（Event）起源于认知科学。认知科学家认为，以“事件”为单位来体验和认识世界符合人们正常的认知规律。但目前对“事件”还没有统一的定义，不同领域对“事件”的理解不同。

在 ACE 评测会议中，“事件”[4]被描述为一个动作的发生或状态的变化。

美国佛罗里达州大学的 Zwaan[9]将每个单句等同为一个“事件”。

本文研究的“事件”也属于句子级，但不是每个句子都是事件实例。只有当一个句子含有事件特征时才构成事件实例，否则为非事件实例。

2.2 事件实例识别算法

本文将新闻文本中的每个句子作为一个候选事件，从句子中抽取出刻画一个事件发生的有代表性的特征，构成候选事件实例表示，构造二元分类器对事件实例与非事件实例进行自动识别，如图 1 所示。

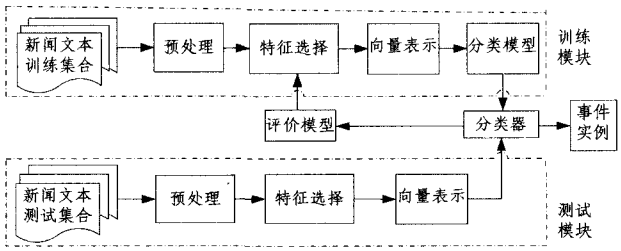

图 1 事件实例识别流程图

新闻文本事件实例识别过程主要由训练、测试和评价模型这 3 部分构成，具体步骤如下：

（1）文本预处理，主要包括中文分词、词性标注、句子切分等；

（2）由于将事件实例的识别看作分类问题，特征的选择和发现尤为关键，在步骤（1）的基础上，主要选取了以下几个特征：句子的长度、位置、词语的个数、命名实体的个数、时间的个数、数值的个数、停用词的频率以及相应的词语等；

（3）在完成特征提取之后，利用向量空间模型（Vector Space Model，VSM）对所有候选的事件进行向量表示；

（4）完成对候选事件的向量表示后，利用 SVM 分类器进行分类。由于 SVM 分类器通用性好、分类精度高、分类速度快、分类速度与训练样本个数无关，在召回率和准确率方面都优于传统的分类器；

（5）训练时，对训练文档集进行预处理、特征提取以及向量表示，然后对 SVM 分类器进行训练，得到分类模型；

（6）分类评价模型是对分类器性能进行评价，同时给出反馈信息进行学习，从而对分类特征进行不断修正；

（7）测试时，先对测试文档集进行预处理、特征提取以及向量表示，再输入到已经训练好的分类器中，完成事件实例的识别。

3 基于事件实例驱动的事件抽取

3.1 事件实例相似度计算

事件实例相似度的计算是事件抽取的一个重要环节。通常采用基于 VSM 的计算方法，它将一个事件实例 s 看作 n 维空间的一个向量 $s(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$，其中 t_i 为事件实例中互不相同的词，w_i 为 t_i 的 $tf*idf$ 值。设两个事件实例分别为 $s_1(t_{11},w_{11};t_{12},w_{12};\cdots;t_{1m},w_{1m})$ 和 $s_2(t_{21},w_{21};t_{22},w_{22};\cdots;t_{2m},w_{2m})$，则两者的相似度为

$$\mathrm{Sim}(s_1,s_2)=\frac{\sum_{i=1}^{m} w_{1i}\cdot w_{2i}}{\sum_{i=1}^{m} w_{1i}^2\cdot\sum_{i=1}^{m} w_{2i}^2} \tag{1}$$

但与通常的文本相似度计算相比，事件相似度计算的特殊性在于：所依据的信息粒度相对更小，仅是一个事件实例。基于 VSM 的方法对于这种粒度较小的信息单元，得到的特征向量是非常稀疏的。在实验中发现，事件实例的特征向量中非 0 的值平均只有 2.8%。因此，基于 VSM 的方法在事件实例相似度计算上难以得到很好的相似度计算结果。

基于词语语义相似度的方法在事件表示形式和相似度计算方法上都和基于 VSM 的方法有着本质的不同。它是将事件实例看作词的序列，通过比较两个事件中包含的词相似度，并进行一定的加权而获得事件间的相似度。该方法主要是利用《知网》[10]等词典首先计算事件实例中词语的语义相似度，然后根据词语的语义相似度进行事件之间的相似度计算。

知网（HowNet）是一个以汉语和英语词汇所代表的概念为描述对象的语义资源，它对于词汇的语义描述具有明显的结构化特征。在知网中，词汇语义的描述被定义为义项（概念），每一个词可以表达为几个义项。义项又是由一种知识表示语言来描述的，这种知识表示语言所用的词汇称作义原。义原是从所有汉语词汇中提炼出的可以用来描述其他词汇的不可再分的基本元素。

本文利用《知网》计算词语语义相似度，但单纯依靠语义词典无法计算词典中不包含的词的相似度，所以必须考虑语义词典中未收录词语的处理。

对于知网收录的词语，采用刘群[11]提出的利用《知网》计算词汇语义相似度方法。利用知网的结构特征将词汇语义相似度计算层层分解，先将两个词汇语义相似度计算分解为义项相似度计算，再将义项相似度计算分解为义原相似度计算。

对于知网未收录的词语，无法用上述方法对其进行语义相似度计算，这将很大程度影响该方法的性能。为此，本文制定一些规则来计算未收录词的相似度。

规则 1 人名的处理。如果一个为姓氏加职务，另一个为姓氏加名字，则认为两者的相似度很大，如“温总理”和“温

家宝”。

规则 2　地名的处理。知网中一般只收录行政区域名称，并不附加诸如市、县、区、村等后缀名。对此，本文自动对地名的后缀进行识别，完成地名的语义相似度计算。如知网中收录了“玉树”，但未收录“玉树州”，本文认为这两者的相似度为 1。

规则 3　机构名的处理。如果组织机构名中存在全称和简称，则认为两者的相似度较大，如“青海师大”和“青海师范大学”。

通过以上的计算得到了事件实例中词语的语义相似度，在此基础上计算事件之间的相似度，具体方法如下：

设两个事件 s_1 和 s_2，s_1 中包含的词语为 $s_{11},s_{12},\cdots,s_{1m}$，$s_2$ 中包含的词语为 $s_{21},s_{22},\cdots,s_{2n}$，利用上述方法计算词语 $s_{1i}(1\leqslant i\leqslant m)$ 和 $s_{2j}(1\leqslant j\leqslant n)$ 之间的语义相似度 $s(s_{1i},s_{2j})$。其相似度 $\mathrm{Sim}(s_1,s_2)$ 为

$$\mathrm{Sim}(s_1,s_2)=\frac{1}{2}\left[\sum_{i=1}^{m}a_i/m+\sum_{j=1}^{n}b_j/n\right] \tag{2}$$

式中，$a_i=\max(s(s_{1i},s_{21}),s(s_{1i},s_{22}),\cdots,s(s_{1i},s_{2n}))$，$b_j=\max(s(s_{2j},s_{11}),s(s_{2j},s_{12}),\cdots,s(s_{2j},s_{1m}))$。

为了验证词语语义相似度计算方法（语义方法）的有效性，在同一环境下与基于 VSM 的计算方法（VSM 方法）进行了对比实验。选择包含 522 个事件实例的新闻文档集，分别用两种算法计算文档中事件实例的相似度，为每个事件实例选出与该实例相似度最大的一个事件实例。仅当选出的事件实例与人工评判的最相似的事件实例一致时，才认为该事件实例的相似度计算结果正确。因此，可以计算出两种方法的准确率，实验结果如表 1 所列。

表 1　事件相似度计算准确率比较

	测试事件实例数	结果正确的数目	准确率(%)
VSM 方法	522	380	72.79
语义方法	522	423	81.03

由实验结果可见，语义方法的准确率明显高于基于 VSM 方法。

3.2　基于层次聚类的 k-medoids 算法

目前，聚类算法的研究比较成熟，大体可分为两类：层次聚类算法和基于划分聚类算法。

层次聚类算法应用较广泛，它不需要预先设定聚类最后的目标类别数，通过停止阈值就可以确定聚类是否结束。停止阈值定义为

$$D\_Threshold=a\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}\mathrm{Sim}(c_i,c_j)}{(N^2-N)/2} \tag{3}$$

式中，a 为常数，可以通过实验对它取值。a 的值越大，得到的聚类类别越多。N 为新闻文本中的事件实例数。$\mathrm{Sim}(c_i,c_j)$ 为两个类间的相似度。但是，层次聚类也存在明显的缺点——不可回溯，一个点一旦被归为某个类不能再改变。

基于划分的聚类算法如 k-medoids 可以满足不断调整聚类结果的要求，但需要预先设定目标类别的 k 值，且初始质心也是随机选取。

为解决单一聚类算法的不足，本文采用基于层次聚类的 k-medoids 算法，先使用层次聚类对事件实例进行初始聚类，以得到的聚类类别数作为基于划分聚类方法的 k 值，然后进行基于划分聚类得到最终的聚类结果。

给定的事件实例集合 $S=\{s_1,s_2,\cdots,s_n\}$，基于层次聚类的 k-medoids 算法基本过程如下

(1)将 S 中的每个事件实例 s_i 看作一个具有单个成员的类 $c_i=\{s_i\}$，这些类构成 S 的一个聚类 $C=\{c_1,c_1,\cdots,c_n\}$；

(2)计算 C 中各个类两两之间的相似度 $\mathrm{Sim}(c_i,c_j)=\max\mathrm{Sim}(s_i,s_j)$，其中 $s_i\in c_i,s_j\in c_j$，$\mathrm{Sim}(s_i,s_j)$ 按照式(2)计算；

(3)选择具有最大相似度的类对 $\{c_i,c_j\}$，将 c_i 和 c_j 合并成一个新的类 $c_k=c_i\cup c_j$，把 c_k 添加到 C 中并从 C 中删除 c_i 和 c_j，更新 C 中类间的相似度；

(4)由式(3)计算层次聚类的停止阈值，当任意的两个事件实例集合间相似度都小于 $D\_Threshold$ 时聚类结束，层次聚类的结果为 k 个类 $C_1,C_1,\cdots,C_k$；

(5)从每个 $C_i(1\leqslant i\leqslant k)$ 中任意选取一个事件实例作为 k-medoids 聚类的 k 个初始中心点，并计算出事件实例之间的距离：

$$D(s_i,s_j)=1-\mathrm{Sim}(s_i,s_j) \tag{4}$$

(6)根据 $D(s_i,s_j)$ 将余下的事件实例按照最相近的原则分到 k 个类中去，然后重新选取类的中心点。循环进行，当各类别中的事件实例不再移动时聚类结束。

3.3　事件抽取算法描述

基于事件实例驱动的事件抽取主要思想是：首先，从句子中抽取刻画一个事件发生的有代表性的特征，构成候选事件实例表示；其次，对文本中的事件实例通过分类器进行区分；最后，对事件实例进行聚类，得到同一主题文档集合中所包含的不同事件集，完成事件抽取，如图 2 所示。

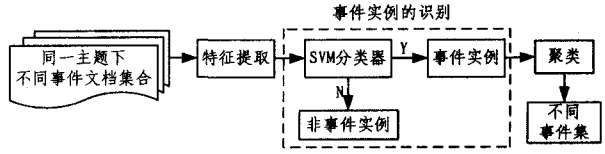

图 2　基于事件实例驱动的事件抽取流程图

基于事件实例驱动的新闻文本事件抽取算法描述如下：

(1)对同一主题下不同事件的新闻文档中所有的句子按照 2.2 节中的(2)完成特征抽取，构成候选事件表示；

(2)对所有候选事件按照 2.2 节进行事件实例的识别，过滤掉非事件实例；

(3)完成事件实例识别后，对事件实例按照 3.2 节中基于层次聚类的 k-medoids 算法完成事件实例的聚类。算法最后得到 k 个类 $C_1',C_2',\cdots,C_k'$，其中 $C_i'(1\leqslant i\leqslant k)$ 代表了新闻文档集合中同一主题下的不同的事件集合，其中的事件实例具有相似的语义。

4　实验及性能分析

4.1　实验数据

本文实验数据是从网上下载的 500 篇同一主题下的新闻文本，去除广告链接等无关内容，只保留标题和正文内容。为了实验验证算法，对 500 篇文档进行了数据统计，包括 500 篇文档包含的句子数、事件实例数以及非事件实例数，结果如表 2 所列。

表 2　实验数据统计结果(个)

句子数	事件实例数	非事件实例数
7368	6949	419

4.2　评价指标

本文采用信息抽取中的评价标准:准确率 P(Precision)、召回率 R(Recall)和 $F1$ 值来进行性能评价。各评价指标定义如下

$$P=\frac{\text{抽取准确的事件个数}}{\text{抽取事件总数}}\times 100\% \tag{5}$$

$$R=\frac{\text{抽取准确的事件个数}}{\text{标注事件总数}}\times 100\% \tag{6}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{7}$$

4.3　实验结果对比及分析

为了验证本文事件抽取方法的有效性,做了两组对比实验。

4.3.1　与基于事件元素和触发词驱动的事件抽取结果对比

将基于事件实例驱动的事件抽取方法(本文方法)与基于事件元素驱动的事件抽取方法[5](事件元素方法)以及基于触发词指导的自相似度聚类方法[7](触发词方法)结果进行比较。为使对比结果具有可信性,在同一个实验环境中对文献[5,7]中的方法进行了实现,结果如表 3 所列。

表 3　3 种方法的对比实验结果(%)

方法 \ 指标	准确率(P)	召回率(R)	F1 值
事件元素方法	59.96	61.28	60.61
触发词方法	62.84	66.73	64.72
本文方法	69.56	71.34	70.44

本文以事件实例进行驱动,避开传统的以事件元素或触发词构建抽取模型。由表 3 可见,无论抽取的准确率还是召回率均获得了提高。主要因为文本中非事件元素和触发词所占的比例太大,目前还没有高效的算法对其进行过滤,因此以它们为实例进行驱动,不仅增加了计算的负担,更重要的是引入太多反例,导致正反例严重失衡,且在语料规模较小的时候存在着一定的数据稀疏问题。

4.3.2　与分类法事件抽取结果对比

将本文方法与基于特征选择的事件检测和分类[12](分类法)结果进行比较。文献[12]采用 ACE2005 语料进行实验,但官方不公布此语料。为了不影响结果的比较,在本文的实验数据上对文献[12]中的方法进行了实现,结果如表 4 所列。

表 4　两种方法对比实验结果(%)

方法 \ 指标	准确率(P)	召回率(R)	F1 值
分类法	67.73	64.48	66.07
本文方法	69.56	71.34	70.44

分类法认为事件抽取就是判断一个句子是否为一个事件句。如果是事件句,则再根据一定的特征判断其所归属的类别,从而完成事件抽取。但此方法仍然局限于预先定义事件的类别,对于没有定义的事件则无法抽取,因此导致了其召回率不高。本文方法突破预先定义事件类别的局限,根据事件实例对新闻文本中的事件进行自动聚类,使得所抽取结果的召回率有显著提高。同时,本文通过事件实例的识别对非事件实例进行了过滤,使得所抽取结果的准确率也有了进一步的提高。

结束语　本文提出了一种基于事件实例驱动的新闻文本事件抽取模型,不仅克服了传统基于触发词或事件元素驱动的事件抽取方法中所带来的正反例失衡和数据稀疏的问题,而且通过对事件实例的聚类很好地解决了 ACE 评测会议中预先定义事件类别的局限性,对以事件实例驱动的事件抽取进行了初步的研究和有意义的探索。实验结果表明,本文方法是一种有效的事件抽取方法。

但是,中文事件抽取仍处于起步的阶段,还有很广阔的研究空间。本文只是对领域相关的事件抽取进行了研究,如何对非受限领域进行事件抽取,以及进一步提高事件抽取的效果,均有待于进一步的研究与探讨。

参 考 文 献

[1] Li Wen-jie, Wu Ming-li, Lu Qin. Extractive Summarization Using Inter- and Intra- event Relevance [A]// Proceedings of the 44th Annual Meeting of the Association for Computational Liguistics[C]. 2006:369-376

[2] Ahn D. The stages of event extraction[A]// Proceedings of the Workshop on Annotations and Reasoning About Time and Events[C]. Sydney, 2006:1-8

[3] 冯礼. 基于事件框架的突发事件信息抽取[D]. 上海:上海交通大学,2008

[4] ACE(Automatic Content Extraction) Chinese Annotation Guidelines for Events [M]. National Institute of Standards and Technology, 2005

[5] Leong C H, Tou N H. A Maximum Entropy Approach to Information Extraction from Semi-structured and Free Text[A]// Proceedings of the 18th National Conference on Artificial Intelligence[C]. 2002:786-791

[6] 赵妍妍,秦兵,车万翔,等. 中文事件抽取技术研究[J]. 中文信息学报,2008,22(1):3-8

[7] 张先飞,郭志刚,刘嵩,等. 基于触发词指导的自相似度聚类事件检测[J]. 计算机科学,2010,27(3):212-214

[8] Vapnik V. Nature of Statistical Learning Theory [M]. New York:Springer Press, 2000

[9] Zwaan R A, Radvansky G A. Situation models in language comprehension and memory [J]. Psychological Bulletin, 1998, 123(2):162-185

[10] 董振东,董强. 知网[OL]. http://www.keenage.com, 2009-06-08

[11] 刘群,李素建. 基于《知网》的词汇语义相似度的计算[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2):59-76

[12] 谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨:哈尔滨工业大学,2008