基于词对向量空间模型的新事件检测方法

樊旭琴[1],张永奎[1,2]
FAN Xu-qin[1],ZHANG Yong-kui[1,2]

1.山西大学 计算机与信息技术学院,太原 030006
2.山西大学 计算智能与中文信息处理省部共建教育部重点实验室,太原 030006
1.School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2.Key Laboratory of MOE for Computation Intelligence and Chinese Information Processing,Shanxi University,Taiyuan 030006,China
E-mail:fxq514@126.com

FAN Xu-qin,ZHANG Yong-kui.New event detection method based on word pairs vector space model.Computer Engineering and Applications,2010,46(12):123-125.

Abstract: New Event Detection(NED) aims at detecting the first news item on one topic from one or more news reports.The traditional vector space model adopts single word to represent the text features,considering the information of word position and other information of expressing content,this paper proposes an approach using word pairs to express text content.Combined with the HowNet,the extracted word pairs are normalized.Then the different weight parameters of different part of speech pairs are given according to different types of news reports.Experiments on emergency news corpus show that the word-pair method can significantly improve the representation results.

Key words: vector space model;word pair feature;new event detection

摘 要:新事件检测(NED)的目标是从一个或多个新闻源中检测出报道一个新闻话题的第一个新闻。传统向量空间模型采用单个词来表示文本特征,考虑到词的位置信息以及其他的表示内容的信息,提出了词对表示文本的方法,并结合 HowNet 资源对所抽取的词对进行归一化处理,最后对不同类别新闻中不同词性对的权重参数进行优化。通过在已有的突发性新闻语料上进行实验,表明这种改进方法的效果比较明显,性能也有一定的提高。

关键词:向量空间模型;词对特征;新事件检测
DOI:10.3778/j.issn.1002-8331.2010.12.036 **文章编号**:1002-8331(2010)12-0123-03 **文献标识码**:A **中图分类号**:TP391

新事件检测是话题检测与跟踪研究[1](TDT)课题中的 5 项任务之一,新事件检测的目标是从一个或多个新闻源中检测出报道一个新闻话题的第一个新闻。在当今时代信息爆炸的情况下,新话题往往淹没于每日海量的信息流中,这一现象极大限制了人们及时掌握重要的新闻动态。新事件检测的结果可以让用户尽早全面掌握国内外各种突发事件的发生情况和发展趋势,为国家和各级地方政府有关部门及时采取应急措施和制定防范计划等提供参考依据。

1 相关研究

近几年来,很多研究者都将工作集中在对新闻的表示模型与新闻间的相似度模型的改进上。Stokes[2]等人将新闻的表示分为两个部分:第一部分为普通的文本特征向量;第二部分通过 WordNet 中的词汇链扩展而成。两种表示通过线性方式进行组合,但文中实验结果的改善并不十分明显。Yang[3]等人在 tf-idf 模型基础上直接对地点名称给予 4 倍的加权。DOREMI[4]研究组计算人名、地名、时间的语义相似度并结合文本相似度得出最终的相似度。

TDT 领域的研究在国内也逐渐受到重视,洪宇等人将话题和报道划分为不同子话题[5],根据相关子话题的比例关系和分布关系建立新话题识别模型,提出了基于子话题分治匹配的新事件检测。张阔等人提出了一种基于词元再评估的新事件检测模型[6]。

2 基本模型

新事件检测系统包括三个部分:新闻描述,新闻间的相似度计算,新事件的检测过程。新闻描述部分通过预处理建立新闻文档的向量表示。相似度计算部分根据新闻描述计算新闻间相似度。新事件检测过程为对按时间顺序排列的新闻进行比较,判断新事件新闻的过程。该文就是针对基本新事件检测系

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60475022);山西省自然科学基金(the Natural Science Foundation of Shanxi Province of China under Grant No.20041041);山西省回国留学人员基金(No.2002004)。
作者简介:樊旭琴(1985-),女,硕士研究生,主要研究领域:中文信息处理;张永奎(1945-),男,教授,博士生导师,主要研究领域:人工智能、中文信息处理。
收稿日期:2009-04-08 **修回日期**:2009-06-08

统中的新闻描述部分进行了改进和扩展。

2.1 新闻描述

对于新事件检测任务,要判断某个事件是否是新事件。首先要解决用什么模型表示它们的问题,即新闻描述。不论是话题还是报道,都要表示成计算机所能识别的形式,目前常用的模型有向量空间模型。

向量空间模型是目前最简便高效的文本表示模型之一。为了把文本表示成向量形式,首先要做的就是进行特征项提取,将文本表示为项的集合,然后根据项的权重把文本表示成向量。在传统的向量空间模型中,使用单个词作为特征项进行提取。

2.2 新闻间的相似度计算

使用 Hellinger 距离计算新闻之间的相似度,对两个新闻 d 和 d',它们之间的相似度表示为:

$$sim(d,d',t)=\sum_{w\in d,d'}\sqrt{weight(d,t,w)*weight(d',t,w)} \quad (1)$$

2.3 新事件的检测过程

对于在时间 t 加入的新闻 d,将与之前获得的所有新闻文档进行比较,根据其中的最大相似度得到 d 为新事件报道的支持度:

$$n(d)=1-\max(sim(d,d',t)) \quad (2)$$

其中,d'为 d 之前出现的文档。若支持度大于一个阈值 θ,则认为 d 报道了一个新事件,反之则认为描述的是已报道事件。同时支持度与阈值 θ 的差越大,表明决策的自信度越高。

3 改进模型

基本向量空间模型采用单个词来表示文本特征,虽然用来表示文本在很多方面的应用都取得了很好的成绩,但是它仍存在很多问题。它只使用了词和词频的信息,而忽略了大量的其他信息,这样原文本中的大量信息都被丢弃了,比如:段落、句子、词序信息以及句法结构都没有体现出来,词与词之间的关系也没有得到利用。为了能够更有效地表示文本,提出了词对向量空间模型,使用词对信息进行新闻描述。另外,考虑到时间信息不能进行词对表示,将时间要素抽取出单独进行描述。

3.1 基于词对特征的新闻描述

为了充分利用新闻事件中的信息,选用词对特征模型来进行聚类的特征提取。其基本思想是:考虑文本中出现的词汇组合,利用它们的词性和位置信息来构造词对,使用这些词对特征来表示文本。

3.1.1 词对的定义与选择

词对[7]:文本中特定词性,特定距离内包含词序的一对词。

设 wp 表示词对,w_i、w_j 为构成词对的两个词,c_i 和 c_j 为 w_i、w_j 的词性,POS_i、POS_j 为 c_i 和 c_j 的允许词性组合,D_w 为 w_i、w_j 允许的文本跨度。则 $wp=\{(w_i/c_i,w_j/c_j),|w_i,w_j|\in D_w,c_i\in POS_i,c_j\in POS_j\}$。

这里限定词性只考虑动词和名词,并且 w_i 和 w_j 要在同一句段里。即设置 D_w 为以逗号、句号或其他符号终结的句段,$POS_i=\{n\}\&\&POS_j=\{v\}$或者 $POS_i=\{v\}\&\&POS_j=\{n\}$。即:只考虑主-动,动-宾这样的词对集合。

词对生成举例:

输入文本:南方网讯:截至 2 日晚 20 山西省介休市连福镇发生爆炸的金山坡煤矿已发现 26 人死亡,还有 2 名矿工被困井下,下落不明。

输出的词对集合:山西省-发生,介休市-发生,连福镇-发生,发生-爆炸,金山坡-发现,煤矿-发现,发现-死亡,矿工-困。

由此可以看出,输出的词对集大都是体现了某些行为发生的主谓或动宾结构,较为符合新闻事件的特点。

3.1.2 词对归一化

通过观察得出,有些词对之间存在同义词、近义词,因此引入 Hownet 语义资源[8],得出词之间的语义相似度从而识别同义词对、近义词对,在计算词对的权重时,将它们当作一个词对来考虑。

3.1.3 词对的权重计算

由词对的选择可得,词对可以看作是一对关键字的组合,计算文章中词对的权重则是通过这两个关键字的权重计算得来。

通过分析新闻文本可以看出,对于不同类别的话题,不同种类的词性对于话题的区分有着不同程度的作用[6]。其中自然灾害、军事冲突、金融三个类别对于地名较为敏感。而选举、犯罪、科学发现三个类别对于人名较为敏感。

采用普通的权重公式来计算词对中单个关键词在文本集中的权重,最后进行加权平均得到词对的权重。不同类别的话题对不同要素的敏感度不同,因此两个词将赋予不同的加权值。

$$weight(wp)=\alpha*weight(w_1)+\beta*weight(w_2) \quad \alpha+\beta=1 \quad (3)$$

其中,wp 为词对,w_1、w_2 为构成词对 wp 的两个词。

以上述新闻为例,如"山西省-发生"这一个词对,因为该新闻类别为爆炸类事件,则对地名较为敏感。在使用公式计算词对权重时,设定 $\alpha=0.7$,$\beta=0.3$。

3.2 基于词对特征的新闻间相似度计算

3.2.1 词对相似度计算

词对相似度采用一般新闻报道的相似度计算公式进行计算。

新闻报道内容特征向量间的相似度计算公式为:

$$sim(\boldsymbol{m}_i,\boldsymbol{m}_j)=\frac{\sum_{k=1}^{M}w_{ik}\times w_{jk}}{\sqrt{(\sum_{k=1}^{M}w_{ik}^2)(\sum_{k=1}^{M}w_{jk}^2)}} \quad (4)$$

其中,$\boldsymbol{m}_i$ 为要检测新闻报道 d_i 的特征向量,$\boldsymbol{m}_j$ 为已经检测过的第 j 个新闻报道 d_j 的特征向量,M 为特征向量的维数,w_k 为向量的第 K 维。

3.2.2 时间相似度计算

在一篇新闻报道中,时间信息可以很容易地抽取出来,但是一些时间的表示并不规范,如:7 月 5 日下午,上周三,去年等。为此必须将时间进行规范化,规范后的格式如:2009 年 03 月 17 日 07 时 30 分。

时间相似度是用时间点对点的匹配来衡量的,考察时间段对应的开始点和结束点。时间段重合得越多则它们的相似度越大。利用下面公式来计算时间相似度:

$$sim_t([t_i,t_j],[t_k,t_l])=\frac{2\Delta[t_i,t_j]\cap[t_k,t_l]}{\Delta(t_i,t_j)+\Delta(t_k,t_l)} \quad (5)$$

3.2.3 相似度加权

在分别计算了词对相似度和时间相似度后,将对所得到的相似度进行加权,从而得到两篇新闻报道最终相似度,在进行两个相似度整合时,采用支持向量机学习器,训练得到事件间的最终相似度,利用该值的正负属性对新事件进行判断。

4 实验准备

4.1 数据集

目前已从互联网上收集了从 2000 年到 2007 年 4 000 多篇突发性新闻语料,文本语料库的规模约 700 万字。从该语料库中挑选了 30 个样本集,约 900 多篇报道。然后在剩余的语料中随机抽取 600 多篇报道作为测试集,进行了小规模的前期实验。

表 1 样本事件

事件编号	事件标题	文档数
1	尼加拉瓜飞机失事 16 人死亡	15
2	山西省介休市连福镇发生爆炸	98
…	…	…
30	广东海丰发生特大交通 事故 4 死 7 伤	65
Total		920

4.2 实验设计

为了测试该文的改进效果,实现并测试了如下实验:

实验 1 此系统为基线系统,采用第 2 部分介绍的基本模型,即使用词进行文本特征描述。

实验 2 采用第 3.1.1 节介绍的改进方法,即使用词对特征模型。

实验 3 采用第 3.1.2 节介绍的方法,即使用 HowNet 资源对词对进行归一化处理。

实验 4 采用第 3.1.3 节介绍的改进方法,针对新闻类别对词对进行简单的权重更新。

实验 5 综合采用第 3.1 节中提出的改进模型,即首先提取特征词对,再结合 HowNet 进行词对归一化,最后针对所属新闻类别对新闻中的词对进行简单的权重更新。

表 2 5 种实验采用的方法比较

实验	新闻描述方法
实验 1	词
实验 2	词对
实验 3	词对+词对归一化
实验 4	词对+更新权重
实验 5	词对+词对归一化+更新权重

4.3 评价标准

使用失报率(Miss)、误报率(FA)和 C_{Det} 代价函数对结果进行评价,公式如下:

$$C_{Det}=C_{Miss}*P_{Miss}*P_{Target}+C_{FA}*P_{FA}*P_{Nontarget} \quad (6)$$

其中,P_{Miss} 为对一篇新事件新闻失报的概率,P_{Target} 表示新事件新闻出现的概率,P_{FA} 为将一篇新闻误报为新事件新闻的概率,$P_{Nontarget}$ 表示非新事件新闻出现的概率。C_{Miss} 与 C_{FA} 分别是漏报和误报的代价,它们的值通常根据应用预先给定,目前在大多数 TDT 评测任务中它们分别取 10 和 1。

一般使用标准化的代价函数作为最终评价标准:

$$Norm(C_{Det})=\frac{C_{Det}}{\min(C_{Miss}*P_{Target},C_{FA}*P_{Nontarget})} \quad (7)$$

5 实验结果及分析

表 3 列出了 5 个实验的 NED 结果比较。

表 3 新事件检测结果比较

实验	*Miss*/(%)	*FA*/(%)	*Norm*(C_{det})
实验 1	41.43	4.87	0.596 8
实验 2	41.04	4.54	0.589 2
实验 3	41.25	4.38	0.579 4
实验 4	40.56	4.19	0.576 2
实验 5	40.23	4.29	0.571 4

实验分析:

(1)从表 3 中得出,采用词对进行新闻描述的方法不论失报率还是误报率都低于一般的使用单个词的描述方法,而且实验 2 的代价也比实验 1 低。

(2)引入归一化后,实验 3 的误报率比实验 2 的误报率降低了 0.16 个百分点,但系统的失报率却提高了 0.21 个百分点。分析原因,归一化将同义近义词合并,虽然降低了向量的维数,但同时也可能将一些有用的信息剔除。

(3)实验 4 引入更新权重的方法,实验系统有了明显的改进,这与更新权重的依据有很大关系。

(4)实验 5 中将词对特征、归一化、更新权重进行结合,系统的性能有了一定改进,但效果并不如预期想象的好。如何将三者进行结合,将会在以后的研究中继续探讨。

6 结语

对传统向量空间模型进行了改进,提出了词对表示文本的方法,并对所提取的词对特征进行了简单的归一化处理,最后通过分析不同词在不同类型新闻中的重要程度,对词对的权重参数进行了优化。实验结果表明,提出的词对向量空间模型对新事件检测的结果得到了一定程度的改善。

参考文献:

[1] 李保利,俞士汶.话题识别与跟踪研究[J].计算机工程与应用,2003,39(17):7-10.

[2] Nicola S,Joe C.Combining semantic and syntactic document classifiers to improve first story detection[C]//Proceedings of the 24th Annual International ACM SIGIR Conference.New York,NY,USA:ACM Press,2001:424-425.

[3] Yang Y,Pierce T,Carbonell J.A study on retrospective and on-line event detection[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.CMU,USA:ACM,1998:28-36.

[4] Juha M,Helena A M,Marko S.Simple semantics in topic detection and tracking[J].Information Retrieval,2004,7(3/4):347-368.

[5] 洪宇,张宇,范基礼,等.基于子话题分治匹配的新事件检测[J].计算机学报,2008,31(4):687-689.

[6] 张阔,李涓子,吴刚,等.基于词元再评估的新事件检测模型[J].软件学报,2008,19(4):817-825.

[7] 王会珍.面向话题追踪的特征选取与文本表示技术的研究[D].沈阳:东北大学,2004.

[8] 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.