金融领域的事件句抽取*

李江龙[1], 吕学强[1], 周建设[2], 刘秀磊[1]

(1. 北京信息科技大学 网络文化与数字传播北京市重点实验室, 北京 100101; 2. 首都师范大学 北京成像技术高精尖创新中心, 北京 100048)

摘 要: 事件句抽取是事件抽取中的核心环节,在金融领域中,公司名识别则是事件句抽取中的重点和难点。针对金融领域的事件句抽取,首先充分利用互联网搜索和上市公司名信息进行公司名识别,如果一个 N 元组是公司名,则进行互联网搜索的结果中包含"公司""集团"等字词多,同时与公司名库中部分公司名有较高的匹配度;其次,综合考虑句子位置信息、包含公司名信息、包含领域动词信息、与标题相似度四个方面特征,构造权值表达式;最终从句子集中选出金融事件句。在数据集上测试,实验结果证明提出的金融领域事件句抽取方法是可行的,公司名识别方法的正确率可达 82.28%,召回率达 68.93%,事件句抽取的正确率可达 66.83%。

关键词: 公司名识别; 事件句; 简称; 事件抽取

中图分类号: TP391.1 **文献标志码**: A **文章编号**: 1001-3695(2017)10-2915-04

doi:10.3969/j.issn.1001-3695.2017.10.008

Event sentence extraction in financial field

Li Jianglong[1], Lyu Xueqiang[1], Zhou Jianshe[2], Liu Xiulei[1]

(1. *Beijing Key Laboratory of Internet Culture & Digital Dissemination Research, Beijing Information Science & Technology University, Beijing* 100101, *China*; 2. *Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing* 100048, *China*)

Abstract: Event sentence recognition is an important part of the event extraction, and in the financial field, the identification of the company's name is an essential as well as a difficult part of the event sentence recognition. For the event sentence identification in the financial field, this paper first made full use of the Internet search information to identify the company's name. Secondly, it considered four factors to construct the value of multi-factor expression: the position of the sentence, the information of the company name, the domain verb information and the similarity between sentence and title. Finally, it chosed the financial event sentences from the sentence sets. The experimental results prove the method's feasibility that the correct rate of the company name recognition method is 82.28%, and the recall rate is 68.93%. And the correct rate of event sentence recognition is 66.83%.

Key words: company name identification; event sentence; abbreviation; event extraction

作为信息抽取的一个重要分支,事件抽取是从非结构化的文本中抽取出用户感兴趣的事件信息,并以结构化的形式保存起来以供后续的分析应用。其在自动摘要[1~3]、自动问答[4]、信息检索[4]等领域有着广泛的应用。

随着国内市场经济不断发展,特别是股市经济,对金融事件越来越敏感。研究面向金融领域的事件抽取对于深入分析金融领域的文本信息、为投资决策提供支持具有重要意义。在当下,面对海量的互联网金融信息,单纯依靠人工的分析很难达到实际的要求。相对于一般的事件抽取,在对金融文本进行事件抽取时,一个比较突出的问题是公司名识别。据统计[5],在公司名的使用上,仅有 7% 使用的是公司全称,而更多的是根据口语习惯使用公司简称。公司简称的使用给金融事件抽取带来了很大的难度。

本文分析了现有公司名识别和事件句抽取工作后,首先提出了基于互联网信息的公司名识别方法;其次,综合考虑语句所在位置、公司名信息、领域动词信息、语句与标题相似度四个方面的特征,构造权值表达;最终从一篇文本的句子集中选择出金融事件句。

1 相关工作

1.1 公司名识别

公司名识别是金融事件抽取中的一个重点,同时也是一个难点。首先,公司名属于未登录词,现在的主流分词平台在进行公司名识别方面还不成熟。如表 1 所示,以常用的三个分词平台为例,对句子"金瑞矿业跨界恐遭叫停,重组标的业绩未达标已停牌"进行分词,句子是新浪财经新闻文本中随机抽取的一句话。显然,"金瑞矿业"这个公司简称实体不能够很好地被识别出。其次,在金融文本中还存在一个很重要的现象[5],即公司简称比公司全称的使用频率要高得多。对于公司全称,还有些命名规律可以依赖,简称更倾向口语化,加大了公司名识别的难度。

收稿日期: 2016-07-05; **修回日期**: 2016-08-25 **基金项目**: 2014 年度国家社会科学基金委托课题(14@ZH036);北京成像技术高精尖创新中心资助项目(BAICIT-2016003);国家自然科学基金资助项目(61271304,61671070)

作者简介:李江龙(1990-),男,河北大名人,硕士研究生,主要研究方向为自然语言处理(lijianglong8@163.com);吕学强(1970-),男,教授,博士,主要研究方向为中文与多媒体信息处理;周建设,教授,博导,博士,主要研究方向为语言学及应用语言学;刘秀磊,讲师,博士,主要研究方向为本体、语义搜索.

表 1 三种分词平台分词结果

分词平台	分词效果
哈工大语言技术平台	金瑞/nh 矿业/n 跨界恐/v 遭/v 叫/v 停/v,/w 重组/v 标/v 的/u 业绩/n 未/d 达标/v 已/d 停牌/v。/w
NLPIR/ICTCLAS2016 分词系统	金/b 瑞/b 矿业/n 跨/v 界/k 恐/d 遭/v 叫/v 停/v,/w 重组/v 标的/n 业绩/n 未/d 达标/v 已/d 停/v 牌/n。/w
HanLP	金瑞/nr 矿业/nis 跨界/nz 恐/vg 遭/v 叫停/nz,/w 重组/vn 标的/n 业绩/n 未/d 达标/vi 已/d 停牌/nz。/w

在命名实体识别方面,研究人员对公司全称和其他命名实体的识别已经提出了很多算法并达到了很高的识别效率。但针对公司简称的识别研究,目前不多而且效果也有待改进。文献[6]以人工总结规则、公司名构成特征及其上下文信息等六个知识库为基础,通过二次扫描进行公司全称及简称的识别。但是对规则知识库的依赖性太强,针对公司简称的识别方法也不完善。在开放测试过程中,文献[6]的正确率和召回率分别为 62.8% 和 62.1% 。文献[7]利用简称在文本中第一次出现时伴随的全称信息,提出了基于规则的算法用于识别简称。该算法取得了比较好的效果,在开放集测试中,准确率和召回率分别达 76.8% 和 73.8% 。但在财经新闻、网络评论等金融文本中,简称的首次出现很少伴随全称,这使得文献[7]的应用范围大大受限。文献[5]首先从文本中提取 N 元组,然后建立每个 N 元组与公司全称表的最优对齐关系,最后对每组对齐关系进行评价和筛选以确定 N 元组是不是公司简称。虽然这种方法的准确率达到了 83.62% ,但是其严重依赖于公司全称表。

1.2 事件句识别

事件抽取属于信息抽取领域,在事件抽取技术的发展过程中,MUC(message understanding conference) 会议和 ACE(automatic content extraction) 会议起了很大的推动作用。根据 ACE[8] 中的定义,事件由事件触发词(trigger) 和描述事件结构的元素(argument) 构成。事件抽取的很多相关研究也就是围绕着触发词和事件元素来进行的。相应地,事件抽取的任务可分解为两步进行[9~12]:a) 从一篇文本的句子集中抽取出事件句;b) 再从事件句中抽取出事件元素。因此,事件句抽取是事件抽取的一个关键环节,其抽取效果对后续的事件类型识别、事件元素识别有很大的影响。在文献[13,14]中也称事件句为主题句,描述一个事件信息或文章主题信息的句子。现有检测事件句的方法主要是基于触发词。赵妍妍、Ji 等人[9,15] 都是采用这种方法来发现文本中的事件句。在这类方法中,针对触发词构建,利用机器学习模型将每个词作为一个实例来训练并判断是否为触发词。基于触发词的事件句抽取效果严重依赖触发词表的长度,特别是未知事件触发词是影响事件抽取效果的主要原因之一[16,17]。在触发词的扩展方面也有不少研究,目前通过同义词或聚类方法进行触发词扩展是常用的方法。Chen 等人[18] 采用自举方法分别在英文和中文语料上进行事件抽取的联合训练,从而提高中文和英文的事件抽取性能;Ji[19] 从中英平行语料库入手,从英文语料中扩展新的中文触发词;Qin 等人[20] 则用同义词词林来扩展中文事件。

除了基于触发词,也有基于特征的事件句识别研究。文献[13]基于主题模型来进行事件句抽取,文献[14]综合考虑了文本标题、文本句子长度、位置等信息来进行事件句抽取。如何利用两类事件句抽取方法的优点又避免缺点是本文方法的出发点,基于触发词的事件句抽取方法利用词的信息直接、简单,但缺点是对词表严重依赖;基于特征的方法利用了文本结构、句子特征,但是对领域词的利用不直接、充分。

2 基于互联网信息的公司名识别

综合第 1 章对现有公司名识别工作的分析总结,复杂的规则库和人工构建公司全称表是影响各种方法适用性的最大问题。同时经过对网上金融新闻文本进行特征分析,本文提出了一种基于互联网信息的公司名识别方法。该方法简单、易理解,基本流程如图 1 所示。

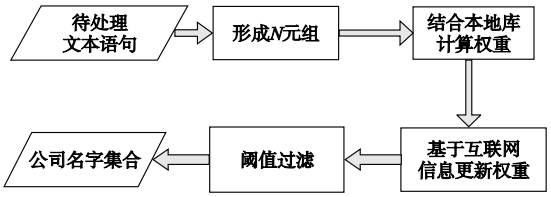

图1 公司名识别基本流程

对于待处理文本,首先提取文本句子中的每个 N 元组(N-gram) 形成 N 元组集体,以此集合作为公司名候选集合;其次,结合公司名库为每个 N 元组进行初步的权重计算;然后,对每个 N 元组进行互联网查询,结合返回的搜索信息对 N 元组进行权重更新计算;最后在 N 元组集合中,将得分高于阈值 β 的 N 元组作为公司名,否则,作为非公司名。

2.1 结合公司名库进行权重计算

同文献[5,6]做法一样,本文也构建了公司名库。但与其人工方式构建的做法不同,本文以国内上市公司名作为库内容,用计算机程序从新浪财经接口通过股票代码可以获得。比如由代码 sh600130 可以获得公司名“波导股份”。此种构建公司名库的方法排除了人工构建过程中主观因素的干扰,通用性更强。

对金融文本进行分析,公司名的简称多是从全称里摘取部分字词,以全称的开头或结尾更为常见。比如“中国石油天然气集团公司”简称“中石油”或“中国石油”,“神州泰岳软件股份有限公司”简称“神州泰岳”。根据此特点,初步对每个 N 元组进行权重计算。针对作为候选公司名的 N 元组,首先计算 N 元组与库中每一个公司名的相似度值,然后选择最大的相似度值作为此 N 元组的权重得分。一个 N 元组 A 与一个公司名 C 的相似度值计算方法如式(1)所示。

$$\mathrm{sim}(A,C)=\sum_{w\in A\cap C}1+\mathrm{len}(A)\times(\mathrm{start}(A,C)\parallel\mathrm{end}(A,C)) \quad (1)$$

2.2 基于互联网信息更新权重

百度搜索是全球最大的中文搜索引擎,拥有全球最大的中文网页库,早在 2010 年收录中文网页已超过 200 亿,而且还在不断更新。对于每个关键词的搜索,百度搜索引擎将在首页给出 10 条搜索结果的简介。经过分析,如果一个 N 元组是公司名全称或者简称,那么利用其作为关键词来进行互联网搜索,在搜索结果中,伴随此 N 元组经常出现的有公司、企业、集团或者股票代码。例如,表 2 是搜索词“中石油”的部分搜索返回条目。基于此,本文主要利用百度搜索结果对 2.1 节中的候选公司名集合进行权重更新。

结合 N 元组的搜索结果对其进行权重更新,具体计算方法如下:

a) 若此搜索结果包含此 N 元组,并且在其后的位置出现“公司”“集团”或“企业”,则此 N 元组权值得分加 1;

b) 若此搜索结果包含此 N 元组,并且在其后的位置内出

现 8 位字符加数字的字符串,即“sh＊＊＊＊＊＊＊”或“sz＊＊＊＊＊＊＊＊”,则此 N 元组权值得分加 2。

经过此步,在公司名识别过程中,充分利用了具有一定实时性的互联网语料库。

表 2　网络搜索返回条目

编号	标题	内容简介
1	中国石油天然气集团公司	中国石油天然气集团公司是以油气业务、工程技术服务、石油工程建设、石油装备制造、金融服务、新能源开发等为主营业务的综合性国际能源公司……
2	中石油_百度百科	中国石油天然气集团公司(简称中国石油集团、中石油)是一家集油气勘探开发、炼油化工、油品销售、油气储运、石……
3	中石油的最新相关信息	中石油甩卖不良资产昆仑系加速整合或为天然气价改(图)
…	…	…

3　金融事件句识别

综合现有的事件句抽取方法:以文献[9]为代表,基于触发词方法对词表依赖性强同时没有很好利用句子位置、与标题相似度等特征信息;文献[14]基于特征进行事件句抽取,其中只是泛泛地利用命名实体,没有充分利用领域词信息。基于此,本文提出了基于语句权值体系的事件句抽取方法,综合公司名信息、领域动词信息、与标题相似度和语句位置四个方面的特征,兼顾各个因素,同时又有所侧重。

3.1　基本定义

定义 1　金融事件句。在金融事件报道中,一个句子包含事件的主体(subject)、谓词(predicate)两个核心要素,并能够代表文章主旨,则称此句子为该篇报道的金融事件句。

定义 2　领域动词集。它是指一组能够代表描述事件核心内容的动词组合。本文主要是进行金融方面领域动词集的研究与构建。

3.2　构建领域动词集

动词往往包含较多的事件信息,领域动词是事件句的重要特征。本文采用半监督的方式来构建金融领域动词表,充分考虑一个动词的上下文信息和在句子中的语义角色,利用最大熵模型计算一个词属于金融领域动词的概率。关键步骤如下:

a)人工从语料集中选出一些金融领域动词;

b)结合人工选出的领域动词,从训练语料中构建所有动词的特征窗口,特征窗口包含上下文信息和语义角色信息两部分;

c)在扩展语料集中构建所有动词的特征窗口;

d)训练阶段,利用最大熵模型对步骤 b)中特征窗口进行训练;

e)概率计算阶段,利用步骤 d)训练得到的模型对步骤 c)中的特征窗口进行概率运算,得到一个动词属于金融领域动词和非金融领域动词的概率。

其中动词的上下文和语义角色特征窗口如表 3 所示。

根据上述特征模板表,构建训练特征模板。例如训练语料中经过分词后的一个小句子片段“华神/nz 集团/n 闪电/v 停牌/v 谋/v 重组/v 。/wp”,显然这里“停牌”是本次金融事件的关键动词。经依存句法分析后,“停牌”标注角色为“HED”,则此关键词的特征窗口为“集团/n 闪电/v 停牌/v 谋/v 重组/v HED 1”。

本文依存句法分析器采用哈尔滨工业大学信息检索研究中心的依存句法分析模块 GParser。在 1 000 篇文章中,经过人工标注 200 个领域动词后,再选择机器标注,最终形成包含 679 个动词的金融领域动词表。

表 3　特征模板表

特征标记	特征释义
$W-2$	当前词条的前第二个词
$T-2$	当前词条的前第二个词的词类标注
$W-1$	当前词条的前第一个词
$T-1$	当前词条的前第一个词的词类标注
W_0	当前词条
T_0	当前词条的词类标注
$W+1$	当前词条的后一个词条
$T+1$	当前词条的后一个词条的词类标注
$W+2$	当前词条的后第二个词条
$T+2$	当前词条的后第二个词条的词类标注
SV	当前词条的句法角色
keyFlag	是否为关键动词,0 或 1

3.3　金融事件句抽取

分析一个句子是否为一篇报道的事件句,主要考虑四个特征:公司名信息、领域动词信息、与标题相似度和语句位置。

3.3.1　事件句特征

1)公司名信息　通过对新闻文本分析,金融事件的重要主体为公司,所以将公司名作为事件句的一个重要特征。计算如式(2)所示,$\mathrm{count}(S_i)$ 表示句子 S_i 包含的公司名数量。

$$\mathrm{score}_{\mathrm{company}}(S_i)=\mathrm{count}(S_i) \tag{2}$$

2)金融领域动词信息　动词一般作为一个事件的核心,本文在 3.2 节中已经构建了金融领域动词表。计算领域动词信息的权值方法如式(3)所示,一个句子中包含金融领域动词,那么这个句子是事件句的可能性更高。

$$\mathrm{score}_{\mathrm{keyVerb}}(S_i)=\begin{cases}1 & \text{句子中包含领域动词}\\ 0 & \text{句子中不包含领域动词}\end{cases} \tag{3}$$

3)句子位置　句子位置信息跟文本类型相关。在新闻中,信息含量高的句子通常出现在前几句,所以本文将句子位置作为一个特征,权值计算如式(4)所示。

$$\mathrm{score}_{\mathrm{location}}(S_i)=1/i \tag{4}$$

4)句子与标题相似度　文本的标题一般含有较多的信息量。通过式(5)计算句子与标题的相似度,可以评估句子作为该篇报道事件句的可能性。其中,动词和名词包含更多的信息量,单个词条的权重通过式(6)来计算。

$$\mathrm{score}_{\mathrm{title}}(S_i)=\frac{\sum_{w\in\{\mathrm{title},S_i\}}\mathrm{Weight}(w)}{\sum_{w\{\mathrm{title}\}}\mathrm{weight}(w)} \tag{5}$$

$$\mathrm{weiht}(w)=\begin{cases}2 & w\text{ 为动词或名词}\\ 1 & \text{其他}\end{cases} \tag{6}$$

3.3.2　事件句提取

设新闻文本中有 n 个句子,每个句子的得分是四个特征分量的线性组合,如式(7)所示。

$$\mathrm{score}(S_i)=w_k\mathrm{score}_k(S_i) \tag{7}$$

其中:$k\in\{\mathrm{company},\mathrm{keyVer},\mathrm{location},\mathrm{title}\}$,各个特征分量的权重 w_k 在数据集上通过训练之后会得到最优组合。

综合考虑句子包含的公司名信息、领域动词信息、与标题相似度、句子位置四部分特征,可以减少某一个特征缺陷带来

的影响。公司名、动词信息属于金融领域相关的,但一篇报道中往往有多个句子满足这一条件。利用标题提取事件句的做法依赖于标题的质量,如果标题没有意义,这种方法就失去了意义;根据句子位置判断易受新闻报道写作手法的影响。

4 实验结果与分析

4.1 公司名识别结果分析

实验数据是从新浪财经网上下载的5 000篇财经新闻,从中随机选出1 000句进行公司名识别测试,按基本均等原则将1 000句分为三组数据。在实验中,调整阈值β,设定β值为16时,在第一组数据上可达到最好的效果。以此阈值在其他两组数据上测试,从表4可以看出达到了同等的识别效果。

表4 公司名识别结果

实验编号	公司数量	正确率/%	召回率/%
第1组	208	82.23	67.45
第2组	221	84.52	64.25
第3组	217	80.37	75.09
综合	646	82.28	68.93

综合三组数据测试结果,本文的公司名识别方法的正确率、召回率达到82.28%、68.93%。同时对公司名识别结果中的错误进行分析,发现错误的主要类型在以下两个方面:

a)公司名子串的影响。比如公司名“唐德影视”,其子串“唐德”“唐德影”,无论是互联网搜索返回结果,还是对比本地公司名库,其都有成为公司名的特征。两个子串的得分都与正确公司名“唐德影视”接近,如果提高阈值,则会影响召回率,而阈值太小则正确率大大降低。本文曾尝试对子串进行合并的策略,但较长的N元组会更易被识别为公司名,正确的公司名反而被错误合并。

b)“公司”带来的错误。本文的公司名识别方法基于互联网搜索信息,故一些带“公司”的热点N元组会被误认为公司名,比如“上市公司”“市公司”“家公司”。将进一步考虑对包含“公司”的N元组进行长度和频率分析以减少此类错误。

4.2 事件句识别结果分析

4.2.1 参数学习

对于式(7),需要确定w_k的值。本文将人工标注的216篇财经新闻文本随机抽取100篇作为参数学习语料,另116篇作为测试。对于w_k在满足$0 < w_i < 1$和$\sum W_i = 1$的条件下进行遍历,精确到0.1。通过对结果的比较,最后确定w_{company}、w_{keyverb}、w_{location}、w_{title}分别为0.1、0.2、0.6和0.1。

4.2.2 参考方法

a)首句法(FS)。基于对新闻报道特征的分析,新闻首句通常包含最重要的事件信息,直接选择新闻首句作为事件句。这种方法简单,很多情况下效果也不错,但显然会受新闻写法的影响。本文将它作为参考方法。

b)去公司名(NCompany)。在去除公司名特征因素后看事件句抽取效果,以此检验公司名对事件句抽取的影响。

c)五特征法(FiveC)。FiveC法是文献[14]从相对词频、句子位置、句子长度、命名实体、句子与标题重合度五个特征出发,对一个句子成为主题事件句进行可能性计算。对比文献[14]的抽取方法,本文同样考虑了句子位置、与标题相似度,但不同之处在于:a)文献[14]针对的是广泛的新闻事件,而本文只针对金融领域事件,不考虑相对词频;b)基于触发词思想,本文加入了领域动词和公司名两部分特征信息。

4.2.3 结果分析

在两组实验数据上,四种方法的事件句抽取结果如表5所示。

表5 金融事件句识别结果 /%

实验方法	第1组	第2组
FS	47	48.26
NCompany	55	53.45
FiveC	54	56.03
本文方法	69	64.66

从实验结果可得到如下结论:

a)句子位置是一个重要的特征。不用其他特征或参数调整,只考虑句子是不是第一句,则可以很高的准确率抽取出新闻中的事件句。这是由新闻的特点决定的,新闻报道是新闻事件的载体,为了吸引读者,在叙述方式上大多采用倒叙,即首先呈现事件关键信息以引起兴趣。

b)公司名对金融事件句的抽取效果具有重要的影响。对比NCompany和本文方法,在利用了公司名后,金融事件句抽取的准确率可以提高13%左右。相应地,FiveC方法泛泛地考虑命名实体。在针对金融事件上,并不是所有类型的实体都有意义,另外现有的分词工具对公司名的识别效果并不是很好。从实验结果对比上看,本文事件句抽取方法比FiveC方法的正确率要高近12%。

同时,也进一步分析了影响事件句抽取效果的因素。首先,一些新闻文本中不存在完整的事件句,或事件的完整信息存在于两句话中。比如,一篇关于“金瑞矿业跨界恐遭叫停,重组标的业绩未达标已停牌”的金融新闻,此事件完整信息意思为:公司因重组标的业绩没有达标而遭到停牌。对应到文本中,应该是“根据金瑞矿业的公告,本次重组标的公司的一季度业绩未达标,未实现重组报告中的业绩承诺。对此,公司昨日已经停牌”。显然,此报道的事件句可以说是由两句话组成。而本文中事件句抽取的对象是一句话,因此在处理这类文本时,存在方法盲点。其次,若报道文本中有多个句子都描述事件而各个句子的信息量又基本相同,本文方法也很难抽取出确切的事件句。实际上在这种情况下,人工标注也难以确定哪个是最好的金融事件句。

5 结束语

针对公司名识别,特别是简称使用频繁、口语化现象严重带来的问题,本文提出了基于互联网信息的公司名识别方法。这种方法利用的规则少,不受训练语料限制,充分为事件句的提取及事件元素的识别做好了准备。同时,本文充分结合基于特征和基于触发词的二类事件句抽取方法,从公司名信息、领域动词信息、与标题相似度、语句位置四个方面对句子进行综合权重计算,最终选出金融事件句。在实际数据集上的实验结果证明了该方法能够很好地识别和提取金融事件句。

本文的公司名识别方法和事件句抽取方法可以支持主题事件抽取和事件级金融新闻浏览服务。下一步要进行的工作包括:在公司名的识别上,研究公司名子串问题的解决方案;深度方面,在完成事件句的抽取后进行事件元素的抽取工作。

(下转第2945页)

得到酒店在各个特征的特征—观点对及好评率。该方法不仅可以使用户更好地了解人们对酒店类产品各种特征的情感倾向分布,并优化用户对酒店类产品的购买决策;还可以使酒店更清晰地了解消费者对自己服务和设施的反馈信息,为酒店对各种特征的改进提供了更加准确的参考。

实验结果表明,本文方法的准确率较高,同时召回率也保持了较高的水平,说明本文的方法是有效的。本文将汉语组块分析应用到了产品特征和情感词的提取中,明显提高了提取产品特征—观点对的准确率。但是在产品特征—观点对的情感分析的过程中,其准确率和召回率有待提高。今后的研究方向将对产品特征—观点对的情感分析方法进行改进,提高其准确率和召回率。

参考文献:

[1] Kim S M, Hov Y. Determining the sentiment of opinions[C]//Proc of the 20th International Conference on Computational Linguistics. [S. l.]:Association for Computational Linguistics, 2004:1367-1374.

[2] Hu Minqing, Liu Bing. Mining and summarizing customer reviews [C]//Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004:168-177.

[3] Popescu A M, Etzioni O. Extracting product features and opinions from reviews[M]//Natural Language Processing and Text Mining. London: Springer,2007:9-28.

[4] 李实,叶强,李一军. 中文网络客户评论的产品特征挖掘方法研究[J]. 管理科学学报,2009,12(2):142-152.

[5] 李实,李秋实. 中文评论中产品特征挖掘的剪枝算法研究[J]. 计算机工程,2011,37(23):43-45.

[6] Li Xin, Xie Haoran, Rao Yanghui, *et al*. Weighted multi-label classification model for sentiment analysis of online news[C]//Proc of International Conference on Big Data and Smart Computing. 2016:215-222.

[7] 伍星,何中市,黄永文. 基于弱监督学习的产品特征抽取[J]. 计算机工程,2009,35(13):199-201.

[8] Kamal A, Abulaish M, Anwar T. Mining feature-opinion pairs and their reliability scores from Web opinion sources[C]//Proc of the 2nd International Conference on Web Intelligence, Mining and Semantics. 2012:1-7.

[9] 孙晓,唐陈意. 基于层叠模型细粒度情感要素抽取及倾向分析[J]. 模式识别与人工智能,2015,28(6):531-520.

[10] 李业刚,黄河燕. 汉语组块分析研究综述[J]. 中文信息学报,2013,27(3):1-9.

[11] Kudo T, Matsumoto Y. Chunking with support vector machines [C]//Proc of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies. [S. l.]:Association for Computational Linguistics,2001:1-8.

[12] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proc of Meeting on Association for Computational Linguistics. [S. l.]:Association for Computational Linguistics,2002:417-424.

[13] 李婷婷,姬东鸿. 基于 SVM 和 CRF 多特征组合的微博情感分析[J]. 计算机应用研究,2015,32(4):978-981.

[14] 蒋宗礼,金益斌. 结合点评情感分析的推荐算法研究[J]. 计算机应用研究,2016,33(5):1312-1314,1326.

[15] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications[J]. Knowledge-Based Systems,2015,89(C):14-46.

[16] Li Qiudan, Jin Zhipeng, Wang Can, *et al*. Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems[J]. Knowledge-Based Systems, 2016, 107(C): 289-300.

[17] 尹裴,王洪伟. 面向产品特征的中文在线评论情感分类:以本体建模为方法[J]. 系统管理学报,2016,25(1):103-114.

[18] 王洪伟,郑丽娟,尹裴,等. 基于句子级情感的中文网络评论的情感极性分类[J]. 管理科学学报,2013,16(9):64-74.

(上接第 2918 页)

参考文献:

[1] 韩永峰,许旭阳,李弼程,等. 基于事件抽取的网络新闻多文档自动摘要[J]. 中文信息学报,2012,26(1):58-66.

[2] Lahari E P, Kumar D V N S, Ubale M. A comprehensive survey on feature extraction in text summarization[J]. International Journal of Computer Technology and Applications,2014,5(1):248.

[3] 熊娇,王明文,李茂西,等. 基于词项—句子—文档三层图模型的多文档自动摘要[J]. 中文信息学报,2014,28(6):201-207.

[4] 钱强,庞林斌,高尚. 一种基于词共现图的受限领域自动问答系统[J]. 计算机应用研究,2013,30(3):841-843.

[5] 陈超,朱洪波,王亚强,等. 中文财经文本中公司名简称的自动识别[J]. 四川大学学报:自然科学版,2011,48(2):308-314.

[6] 王宁,葛瑞芳,苑春法,等. 中文金融新闻中公司名的识别[J]. 中文信息学报,2002,16(2):1-6.

[7] 张占英,王中立. 中文文本中公司名简称的识别[J]. 许昌学院学报,2003,22(2):99-101.

[8] ACE (automatic content extraction) Chinese annotation guidelines for events, version 5.5.1[R/OL]. (2005-07-01). http://www.ldc.upenn.edu/Projects/ACE/.

[9] 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.

[10] 许旭阳,韩永峰,宋文政. 事件抽取技术的回顾与展望[J]. 信息工程大学学报,2011,12(1):113-118.

[11] 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.

[12] 吴平博,陈群秀,马亮. 基于事件框架的事件相关文档的智能检索研究[J]. 中文信息学报,2003,17(6):25-30,59.

[13] 王力,李培峰,朱巧明. 一种基于 LDA 模型的主题句抽取方法[J]. 计算机工程与应用,2013,49(2):160-164,257.

[14] 王伟,赵东岩,赵伟. 中文新闻关键事件的主题句识别[J]. 北京大学学报:自然科学版,2011,47(5):789-796.

[15] Ji Heng, Grishman R. Refining event extraction through unsupervised cross-document inference[C]//Proc of the 46th Annual Meeting of the Association for Computational Linguistics. 2008:254-262.

[16] Li Peifeng, Zhou Guodong, Zhu Qiaoming, *et al*. Employing compositional semantics and discourse consistency in Chinese event extraction[C]//Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics. 2012: 1006-1016.

[17] 李培峰,周国栋,朱巧明. 基于语义的中文事件触发词抽取联合模型[J]. 软件学报,2016,27(2):280-294.

[18] Chen Zheng, Ji Heng. Can one language bootstrap the other: a case study on event extraction[C]//Proc of Workshop on Semi-Supervised Learning for Natural Language Processing. Stroudsburg: Association for Computational Linguistics,2009:66-74.

[19] Ji Heng. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning[C]//Proc of Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics. 2009:27-35.

[20] Qin Bing, Zhao Yanyan, Ding Xiao, *et al*. Event type recognition based on trigger expansion[J]. Tsinghua Science and Technology,2010,15(3):251-258.

[21] 赵军,刘康,周光有,等. 开放式文本信息抽取[J]. 中文信息学报,2011,25(6):98-110.