ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.10, No.1, January 2014
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

微博观点句识别的话题影响研究

罗凌,陈毅东,曹茂元

(厦门大学 信息科学与技术学院智能科学与技术系,福建 厦门 361005)

摘要:为了从海量的网络信息中迅速准确地获取评价信息,观点句识别已经成了自然语言处理的一个研究热点。现在观点句识别系统大都是基于机器学习的方法,一般使用机器学习的方法来进行分类会受到领域差异性影响。针对这个问题,该文对微博观点句识别系统是否会受到微博话题影响做了经验性研究,同时为了弥补训练数据的不足,该文通过规则方法自动标注网络数据进行了训练集的扩充。实验结果表明,微博话题间存在差异,进行分话题模型训练可以提升微博观点句识别系统的性能。

关键词:观点句识别;机器学习;话题;规则

中图分类号:TP18 文献标识码:A 文章编号:1009-3044(2014)01-0123-05

A Study on the Effects of Topics on an Opinion Sentences Identification System for Micro-blog Data

LUO Ling, CHEN Yi-dong, CAO Mao-yuan

(Department of Cognitive Science, Xiamen University, Xiamen 361005, China)

Abstract: As an important stage for information extraction, the problem of Opinion Sentence Identification (OSI) has attracted more and more attentions from NLP researchers in the past decade. Similar to other areas in NLP, most current OSI systems are built based on machine learning (ML) technologies, which often suffer from the problem of domain/topic adaptation. In this paper, an empirical study was conducted to test whether the topic difference among the micro-blog data effects on the performance of an ML-based OSI system, which used rule-based automatic annotation methods to expand the training set. The experimental results indicated that by introducing a topic classifier and performing the training based on the sub topics, the performance of the OSI system for micro-blog data could be improved significantly.

Key words: opinion sentences identification; machine learning; topic; rule-based

1 概述

随着网络信息量的日益增长,人们想要从巨大的冗余信息中准确、迅速地获取对一个事物或对象的评价,这就需要快速的识别出语段中的观点句。目前,观点句识别已经成为自然语言处理领域中的一个研究热点,对于观点句这种不受语言表达约束的非规范文本,很难使用规则方法将观点句全面地识别出来,机器学习的方法在这方面体现出了一定优势,所以现今的观点句识别系统大多是基于机器学习的方法来进行二元分类[1]。但是,缺乏标注训练数据和话题间差异性一直都是机器学习分类的研究难点。基于机器学习的观点句识别系统也同样存在着这样的问题,网络上并没有这种大量用于观点句识别的标注数据集,若要进行人工标注,这需要花费大量的人力和物力。而且由于不同话题间的差异性,使用同一个分类器对不同话题去进行观点句识别,识别效果会有所影响。针对这些问题,我们首先通过一些人工规则对网络上获取的资源进行自动标注,然后将这部分自动标注的语料加入到原有的少量训练语料中,以扩充训练语料,再进行分类器分类,并做了一些常用分类器的性能比较。同时为了验证话题会影响观点句的识别,我们针对话题做了经验研究,对比了通用分类模型和分话题分类模型的性能。该文中的实验使用NLP&CC 2012中文微博情感分析评测中的数据集,该数据集来自于20个微博话题,实验中定义的观点句只限定于对特定事物或对象的评价,不包括内心自我情感、意愿或心情。实验结果表明,加入基于规则的自动标注数据,对机器学习分类模型的训练是有帮助的,微博话题间也存在着差异性,分话题模型比通用模型有更好的效果。

文章其他部分安排如下:第二节将进行相关工作的介绍,对观点句识别进行概述,介绍观点句的概念和观点句识别的研究现状;第三节将介绍规则与机器学习相结合的观点句识别方法;第四节,针对微博话题差异性做了经验研究,话题会影响观点句的识别;第五节给出在NLP&CC 2012中文微博情感分析评测数据集上的实验数据,并进行分析讨论;第六节是进行总结和展望。

收稿日期:2013-11-15
基金项目:国家自然科学基金项目(61005052);国家科技支撑计划课题(2012BAH14F03);中央高校基本科研业务费专项资金(2010121068);福建省自然科学基金项目(2011J01369)
作者简介:罗凌(1988-),男,广西桂林人,理学硕士,主要研究方向为自然语言处理、机器翻译;陈毅东(1977-),男,福建厦门人,副教授,工学博士,主要研究方向为自然语言处理、机器翻译等。

2 相关工作

观点句,即在表达的过程中带有某种情感和观点的句子,它是对特定事物或对象的评价,这种观点可以是作者本人的、引用于他人的、或是某群体、组织发表的[1]。国外对观点句的研究起步较早,较有代表性的工作有:Wiebe[2]选择某些词类(代词、形容词、序数词、情态动词和副词)、标点和句子位置作为特征,实现对观点句识别。Riloff[3]等人利用boot-strapping算法学习得到主观性名词,单独使用主观性名词为特征,采用朴素贝叶斯分类器对观点句识别。Wiebe和Riloff[4]他们依靠先前研究中确定的主观特征,分别建立了主观分类器和客观分类器,自动从未标注的文本中获得大量主观句和客观句,再从这些句子中得到更多主观性词语搭配,再用准确性很高的词语搭配更新原始的主观特征。Yu和Hatzivassiloglou[5]利用相似性方法、朴素贝叶斯分类和多重朴素贝叶斯分类等三种统计方法进行观点句识别研究。近几年,由于微博的兴起,针对微博数据,Alexander Pak等人[6]选取n-gram和微博中的词性标注作为特征,利用朴素贝叶斯分类器对微博中的观点句进行识别研究,Luciano Barbosa等人[7]采用微博中的词性信息、词本身的主观性、词的情感极性以及否定词作为特征,训练分类器,对微博主客观性进行分类。D. Davidiv等人[8]提取Twitter中的标签和表情符号作为训练集,训练了一个类似KNN的分类器,对微博情感极性进行分类。

国内较早开始该工作的是姚天昉和彭思威[9]使用了机器学习的方法进行分类识别。叶强等[10]提出了一种根据连续双词词类组合模式(2-POS)自动判断句子主观性程度的方法。王根和赵军[11]提出了一种基于多重冗余标记的CRFs进行观点句识别。蒙新泛和王厚峰[12]通过对比试验,分析了上下文信息对于主客观分类的影响。张博[9]使用模块串行的方法进行观点句识别。宋乐等人[13]在2009年的第二届COAE评测中文观点句抽取的任务中使用了一种类似最小图个的方法。在2011年第三届COAE评测中,徐瑞峰等人[14]提出一种基于图的句子排序算法SentenceRank。

3 观点句识别系统框架

3.1 方法概述

对于基于机器学习的观点句识别系统,需要一定量的标注数据进行训练,如果标注数据很少,这将会大大降低分类器的性能。针对没有标注训练数据这个问题,张文文和王挺[15]通过基于词典和基于规则的方法自动构造训练样例,再使用SVM分类器进行观点句识别。我们借鉴了这篇文章的工作,通过一些人工规则,先对未标注的网络数据进行自动标注,加入到原始的训练集中以扩充训练集,提高分类器的效果。此外,考虑到不同话题的数据在分类特征方面可能存在差异,除了通用的分类模型外,该文引入了分话题模型进行对比,我们对分类器是否受话题差异性影响做了经验研究,实验结果表明话题会影响观点句识别,分话题模型比通用模型有更好的效果。该文实验训练和分类流程如图1所示:

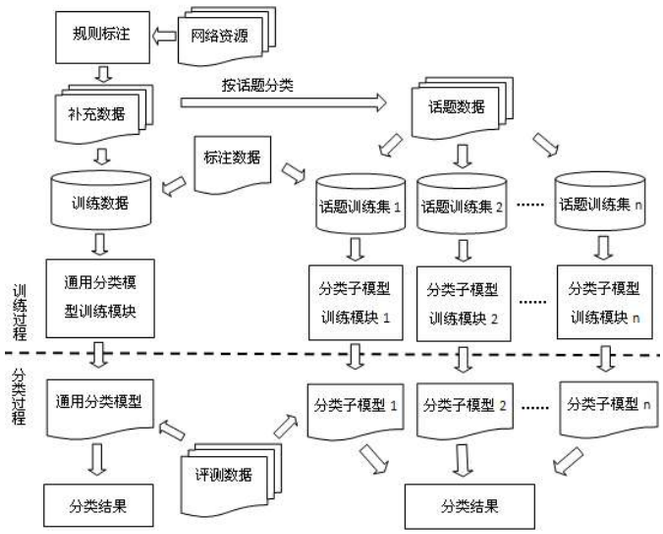

图1 规则与机器学习相结合的观点句识别方法流程图

可以看到,系统的主体采用了机器学习的方法,但为了弥补分类器训练集大小的不足,在训练前,我们利用规则模块对从网络中自动挖掘的微博数据进行标注以扩充训练集。在通用分类模型中,我们将规则自动标注的补充数据和原来的标注数据融合在一起作为通用分类模型的训练数据,再由分类训练模块训练分类模型,再对评测数据进行分类;在分话题模型中,我们将规则自动标注的补充数据按话题分类,在各自加上原始的标注集去分别作为相应分类子模型的训练数据,由分类训练模块训练出分类子模型,然后把测试集也按话题分类,再使用相应的分类子模型进行分类,得出分类结果。

3.2 基于规则的自动训练集标注

如前所述,在本系统中,使用了规则方法对从网络中挖掘的微博数据进行自动观点句识别。通过对标注数据的分析,我们制定了如下的规则来进行观点句识别。在规则中需要用到情感词典,该词典来源于HowNet情感词典和清华褒贬义词典去重合并而得,共16007个词。

观点句规则:

●…+(代词|人名|地名|专有名词)+….+是+名词+….
●…+(代词|人名|地名|专有名词)+….+副词+形容词+….
●...+副词+情感词+(代词|人名|地名|专有名词)
●...+比较词+(代词|人名|地名|专有名词)+情感词
●(代词|人名|地名|专有名词)+比较动词+(代词|人名|地名|专有名词)
●(代词|人名|地名|专有名词)+指示性动词+(代词|人名|地名|专有名词)+...+情感词
非观点句规则:
●仅包含hashtag,表情符合,标点符号的句子判定为非观点句。
●仅包含网址,无实际信息。
●不满足观点句规则且以动词开头的短句。
●只包含愿望词。
●在单句中不含网络新闻,且不是反问句式判定为非观点句。
我们对从网络上挖掘的微博数据进行规则匹配,凡是满足上面规则的句子我们将其抽取出来进行标注,作为训练语料的补充。

3.3基于机器学习方法的观点句识别

观点句识别可以看成一个二元分类问题,我们使用成熟的机器学习分类算法进行分类。我们在进行分类训练时采用了通用和分话题两种训练方法。通用模型是通过将所有话题的训练数据全部拿给分类器训练出一个通用模型;话题模型是通过该话题相关的训练数据给分类器分别训练出20个话题模型。这两种方法我们采用的特征都是在姚天防[5],张博[9]使用的特征基础上,加入了主题相关的人名特征,并进行了特征组合实验,最终选取了如下的特征:

1.情感词,我们整合了知网和清华的情感词典,总共约16000个词。
2.指示性动词,我们使用了张博论文[9]中的指示性动词表和根据数据集自己添加的一些动词,总共约100个词。
3.人称代词、专有名词、人名、地名。
4.叹词和语气词。
5.副词。
6.主题中的对象名。
7.标点符号。
8.N-POS,N-POS是指语句中N个连续词性的顺序组合,系统中我们采用了1-pos和2-pos。

我们进行了不同分类器的效果对比实验,实验中使用了以下5种分类器进行了结果分类:(1)朴素贝叶斯分类算法(Naïve Bayes)(2)支持向量机分类算法(SVM)(3)用于支持向量分类的连续最小优化算法(SMO)(4)随机森林算法(Random Forest)(5)分类与回归树算法(Classification Via Regression)。

4 话题差异性

领域适应性问题一直是自然语言处理领域的一个研究重点,在文本分类,问答系统,自动文摘,机器翻译,文本情感分析等都存在领域适应问题。因为不同的领域数据会有不同的特点,使用同一个模型去处理不同领域的同一个问题,效果也并不理想。对于领域适应性问题,在不同的方向已经有了很多相关研究。在文本情感分类研究中,相同的词语在不同的领域中可能指示着不同的情感倾向,已经有许多研究证明了情感文本分类在分类的精确率上是会受到领域的影响,研究者们也提出了一些方法来解决此类问题[16]。观点句识别作为文本情感分类的基础工作,我们认为也是存在领域差异性的。

本次实验使用的测试数据来自于20个微博话题,我们根据分类器提取的特征对数据进行观察和对比,发现不同话题间的数据是存在着差异性的,下面我们通过对微博话题数据的举例分析来说明这个问题。

1)在不同的话题中,情感倾向偏向不同,导致情感词在不同的话题中分布是不一致的。比如,在话题“90后当教授”里面,总共有观点句123句,其中110句是正面的情感,13句是负面的情感,里面“聪明”、“佩服”、“崇拜”等正面的情感词出现的比较多。而在话题“90后暴打老人”里面,总共有观点句97句,其中3句是正面的情感,94句是负面的情感,里面“畜牲”、“失败”、“流氓”等负面的情感词出现的比较多。由于话题的情感倾向性有差异,有的话题偏向正面情感,有的话题偏向负面情感,那么对于情感词的分布就有所不同。

2)在不同的话题中,与主题相关的人名、地名、专有名词和人称代词有明显的差异。观点句是对一个对象的评价,所以与主题相关的人名、人称代词作为观点句分类系统中的特征是有比较大意义的,但是不同的话题,围绕的对象是不同的,比如在话题“疯狂的大葱”里,“大葱”,“物价局”等出现得比较频繁,而在话题“名古屋市长否认南京大屠杀”里,“名古屋市长”、“日本”等出席得比较频繁。不同的话题评价的对象是不同的。

3)在不同的话题中,使用的句式是有比较大的区别的,所以N-POS在不同话题中是存在着差异的。比如在话题“90后当教授”中,观点句的句式大多是对这个90后的赞扬,“人才!”,“像刘路学习。”,“牛人!”等多是些名词性的短句。而在“彭宇承认撞了南京老太”话题中,多是“说实话,我不太信。”,“这件事绝对不是这样,很可能就是南京市政府搞的鬼!”,“面对政治压力,我觉他是不得已才这样做。”等对这件事的一个看法和评论,基本都是多词性的复合句式。由于在不同话题中表达的句式不一样,抽取出来的N-POS也就存在着很大的差异。

根据上面对测试集数据的分析,可以看出观点句分类器要抽取的文本特征,在不同的话题中,数据分布是存在着差异的,如果

我们把所有标注训练数据一起用来训练一个通用的分类器，然后对所有话题进行观点句识别，可能会由于这些数据差异，导致特征稀疏，影响分类器的精确度。针对该问题，我们根据不同的话题，使用相应的话题训练数据去训练话题子模型，对相应的测试集进行观点句识别，以解决话题间差异性的问题，后面的实验结果也表明话题间是存在差异的，我们的分话题训练也是对观点句识别有帮助的。

5 实验结果及讨论

5.1 实验设置

本文实验使用了由中国计算机学会主办的NLP&CC 2012中文微博情感分析评测中任务一的数据集，还有我们从网络上爬取的与评测数据相关主题的微博数据，并与测试集去重后作为补充数据。具体数据信息如下：

1.标注数据：NLP&CC 2012中文微博情感分析评测提供的标注数据。共包含已标注毁容案话题约240句和Ipad话题约220句。

2.测试数据：NLP&CC 2012中文微博情感分析评测提供的测试数据，共包含菲军舰恶意撞击、疯狂的大葱等20个话题，每个话题约200句。

3.补充数据：从腾讯微博上爬取的与评测数据相关主题的微博。共包含菲军舰恶意撞击、疯狂的大葱等20个话题，每个话题约2000句。接着使用基于规则的方法对其进行了自动标注，标注后每个话题约600句。

本文使用了weka平台中的机器学习分类算法来进行实验[17]。

本文的实验设置如下：

1. 规则与机器学习实验。在标注数据集中，使用毁容案话题数据集作为训练集，Ipad话题数据集作为测试集，进行只使用毁容案直接分类和加入补充数据后再进行分类的对比实验。以验证本文提出的基于规则对机器学习数据集补充的有效性。

2. 通用模型和分话题模型实验。使用标注数据和补充数据一起作为训练集，测试数据作为测试集，进行实验比较通用模型和分话题模型的性能。

3. 分类器性能实验。使用不同的分类器进行前面2个实验，对比不同分类器在该问题上的性能。

5.2 实验结果

本文进行了多个分类器比较，为了方便下面用标号来表示各个分类器：(1)标准概率朴素贝叶斯分类算法(NB)(2)支持向量机分类算法(SVM)(3)用于支持向量分类的连续最小优化算法(SMO)(4)随机森林算法(RF)(5)分类与回归树算法(CVR)

在进行分类器训练时，由于提供的标注训练语料过少，这会影响到分类结果，我们通过上面提出的规则方法自动标注了从网络中挖掘的微博数据，并将这部分数据作为扩充语料加入到原来的标注集里作为训练集进行分类器的训练。为了证明我们加入这些规则方法自动标注的语料对分类器训练是有帮助的，我们按照实验设置1做了下面的实验。我们用原来标注集中的毁容案话题数据作为训练，和加上了自动标注的扩充数据作为训练，对同样的Ipad话题测试集进行测试，得到了如下各个分类器的对比结果，见表1：

表1 加入扩充数据后对比结果

标号	正确率	召回率	F值	+/-
NB	0.645	0.396	0.491	
NB+Extra	0.578	0.515	0.545	+0.084
SVM	0.560	0.782	0.653	
SVM+Extra	0.575	0.762	0.655	+0.002
SMO	0.578	0.515	0.545	
SMO+Extra	0.583	0.733	0.649	+0.104
CVR	0.538	0.624	0.578	
CVR+Extra	0.570	0.802	0.667	+0.089
RF	0.560	0.644	0.599	
RF+Extra	0.549	0.782	0.645	+0.046

没有“Extra”表示训练集中只包含了毁容案的标注数据，“+Extra”表示在原来毁容案的标注数据上，还加入了使用规则自动标注的Ipad话题补充数据。

从表1的结果我们可以看出加入了自动标注的扩充数据进行训练后，基本每个分类器都有或多或少的提升，其中SMO分类器提高的最多，提高了0.104，而CVR分类器在所有分类器中表现最好，F值达到0.667，这表明我们加入的这部分自动标注数据，对训练集数据缺乏的分类器训练是有很大帮助的。

为了实验话题间是否存在差异性，比较通用模型和分话题模型的性能差异。我们按照实验设置2做了下面的实验，这次实验使用标注数据和补充数据一起作为训练集，测试数据作为测试集，对于通用模型，我们直接使用训练集训练出1个通用模型，然后对所有测试集直接进行分类，得出结果；对于分话题模型，我们将补充数据按照20个话题进行分类，每个话题补充集加上原来的标注集作为改话题的训练集，分别训练20个话题子模型，然后测试集也分成同样的20个话题，分别使用相对应的子模型进行分类，得出结果在合并起来进行评测。为了以示区分，我们在分类器简写前加ALL-表示通用模型结果，加Topic-的表示分话题模型的结果，实验结果如表2：

表2 通用模型和话题模型对比结果

标号	正确率	召回率	F值
ALL-NB	0.742	0.376	0.499
Topic-NB	0.744	0.432	0.547
ALL-SVM	0.735	0.675	0.704
Topic-SVM	0.735	0.682	0.708
ALL-SMO	0.737	0.609	0.667
Topic-SMO	0.747	0.623	0.679
ALL-RF	0.727	0.657	0.690
Topic-RF	0.728	0.684	0.705
ALL-CVR	0.720	0.657	0.687
Topic-CVR	0.725	0.720	0.722

从表2结果可以看出分话题进行训练得到的分类结果都比通用模型的分类结果要好,最高的是NB分类器,高出了0.048个点,但是和其他分类器相比,NB比其他分类器低了很多,可能是由于特征选择的问题,导致了NB分类器的性能比较差。所以分类器中CVR分类器性能最好,分话题模型的F值为0.722比通用的高出了0.035。这些实验数据说明领域间存在着话题差异,使用分话题的训练模型比通用模型更能体现出话题的差异,在性能上也有更好的表现。

6 总结与展望

本文针对基于机器学习的观点句识别系统存在训练语料不足的问题,引人了基于规则的方法,通过使用规则的方法对从网络上挖掘的数据进行了自动标注来扩充训练数据,经过实验证明,加入使用我们规则自动标注的数据对训练分类模型有很大帮助,这解决了在机器学习训练过程中语料不足的问题。实验中使用的数据分了20个话题,我们针对话题进行了分话题模型的训练,5种分类算法结果都表明分话题模型比通用模型分类的结果要理想,这说明了话题间的分类特征是存在差异的,使用分话题模型比通用模型效果更好。

本次实验使用的数据来自于NLP&CC 2012中文微博情感分析评测,处理的数据都是来自于微博,微博的最大特点是简短,不规范,里面不仅包含了大量的网络术语,表情,还有很多错别字,病句,这对我们进行分词,提取特征都有很大的影响。如今,由于网络的迅速发展,微博等形式的网络数据大量出现,对微博这种网络文本如何进行更有效的处理,需要我们更深入的研究。通过多个分类器的性能比较,发现各个分类器有各自的特点,如何利用他们自己的特点,进行融合以提高观点句识别的效果,也是我们未来的工作。

参考文献:

[1] 张博. 基于SVM的中文观点句抽取[D]. 北京:北京邮电大学计算机学院,2011.
[2] Wiebe J, Bruce R, Bell M, et al. A corpus study of evaluative and speculative language[C]. acm, 2001.
[3] Riloff E, Wiebe J, Wilson T. Learning Subjective Nouns using Extraction Pattern Bootstrapping[C]. CoNLL-03, 2003:25-32.
[4] Riloff E, Wiebe J. Learning Extraction Patterns for Subjective Expressions[C]. EMNLP-03, 2003:105-112.
[5] Hong Yu, Hatzivassiloglou V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[C]. EMNLP, 2003.
[6] Alexander P, Patrick P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]. Proceedings of International Conference on Language Resource and Evaluation. Lisbon, 2010:1320-1326.
[7] Barbosa Luciano, Feng Junlan. Robust Sentiment Detection on Twitter from Biased and Noisy Data[C]. Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, 2010:36-44.
[8] Davidiv D,Tsur O,Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys[C]. Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, 2010:241-249.
[9] 姚天昉,彭思威. 汉语主客观文本分类方法的研究[C]. 第三届全国信息检索与内容安全学术会议论文集, 2007:117-123.
[10] 叶强,张紫琼,罗振雄. 面向互联网评沦情感分析的中文主观性自动判别方法研究[J]. 信息系统学报, 2007,1(1):79-91.
[11] 王根,赵军. 基于多重冗余标记CRFs的句子情感分析研究[J]. 中文信息学报, 2007,21(5):51-55
[12] 蒙新泛,王厚峰.主客观识别中的上下文因素的研究[C]. 中国计算机语言学研究前沿进展(2007-2009), 2009:594-599
[13] 徐睿峰,王亚伟,徐军,等. 基于多知识源融合和多分类器表决的中文观点分析[C]. 第三届中文倾向性分析评测会议(COAE), 济南, 2011:77-87.
[14] 宋乐,何婷婷,王倩,等. 中文情感词句识别及文本观点抽取研究[C]. 第二届中文倾向性分析评测会议(COAE). 上海, 2009:30-37.
[15] 张文文,王挺. 不规范文本的无监督观点句抽取[J]. 计算机与数字工程, 2013,41(1):64-68.
[16] 任德斌. 主观性文本的情感极性分析研究[D]. 东北大学信息科学与工程学院, 2009.
[17] 李德有,李凌霞,郭瑞波. 基于Weka平台的机器学习方法探究[J]. 电脑知识与技术,2012,8(10):2334-2337.