中文词语倾向性分析处理

李 娟[1,2],张 全[2],贾 宁[1,2]
LI Juan[1,2],ZHANG Quan[2],JIA Ning[1,2]

1.中国科学院 研究生院,北京 100039
2.中国科学院 声学研究所,北京 100190
1.Graduate University of Chinese Academy of Sciences,Beijing 100039,China
2.Institute of Acoustics,Chinese Academy of Sciences,Beijing 100190,China

LI Juan,ZHANG Quan,JIA Ning.Semantic orientation identification for Chinese opinion terms.Computer Engineering and Applications,2009,45(2):131-133.

Abstract: Opinion mining is a new hotspot in the area of natural language processing.Determining the opinion orientation of the glossary is a foundation and very important component in an opinion mining system.An experiment is carried out on opinion orientation identifying for Chinese opinion terms.In the experiment,the authors take the words which are in COMTEMPORARY CHINESE LANGUAGE ORIENTATION USAGE DICTIONARY as the seed words,and extend them by synonyms dictionary.Further more,Bigram theory is adopted to disambiguate the multi-orientation for one word.The F-score of the experiment reaches 79.31% for positive words and 78.18% for negative words.

Key words: opinion mining;semantic orientation;2-Gram

摘 要:意见挖掘是自然语言处理研究领域的一个新热点。词语倾向性的判定是意见挖掘的基础和重要环节。该文进行了中文词语倾向性的自动判定实验。实验中采用了《现代汉语褒贬用法词典》中的词语做为褒贬判定的核心词汇,以同义词词典扩展了褒贬义词典的词语,并使用二元语法模型来判定多倾向性词语的倾向。实验结果褒义词的 F-Score 为 79.31%,贬义词的 F-Score 为 78.18%。

关键词:意见挖掘;词语倾向;二元语法
DOI:10.3778/j.issn.1002-8331.2009.02.038 **文章编号**:1002-8331(2009)02-0131-03 **文献标识码**:A **中图分类号**:H087

1 引言

意见挖掘是近年来自然语言处理领域研究中发展起来的一个新方向，它是信息检索与计算语言学交叉产生的学科,意见挖掘研究的不是文档所谈论的话题,而是它所表达的倾向性观点,即肯定/否定或者褒扬/贬损性意见。意见挖掘的应用范围很广泛,包括提取用户对产品的评价、对公众人物的评价以及客户关系管理等等。

意见挖掘任务可分为几个步骤,一是识别文档中的主观性词语或短语(Opinion Terms),二是对主观性词语或短语的倾向性进行判定,三是结合主观性词语或短语的倾向性和句子结构进行分析,获得句子的倾向性,四是获取段落或篇章的倾向性,如图 1 所示。作为意见挖掘的一个基础环节,主观性词语的倾向性进行自动判定即判定词语的褒贬性研究越来越受到研究人员的关注。本文的主要内容就是关于这方面的研究。

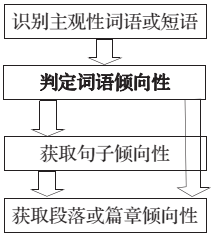

图 1 意见挖掘步骤

目前已有的中文词语语义倾向性分析方法主要有以下两类:基于 HowNet 的词汇语义倾向性分析法和基于同义词词林的方法，例如复旦大学的朱嫣岚等通过手工选定少量的基准词,然后利用 HowNet 的语义相似度和语义相关场功能来计算新词与基准词之间的相似度,从而得到新词的语义倾向性[1];上海交通大学的娄德成等通过手工对 HowNet 包含的所有词条进行倾向性标注,并从网络上选取一定量的极性词语作为种子集合,得到了数量可观的褒贬词语词典,词典中不包含的新词,通过计算新词与种子集合中词语的互信息而得到新词的语义倾向性[2];北京大学的路斌等使用同义词词林,把种子词汇扩展

基金项目:国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318104);中科院声学所知识创新工程项目(No.0654091431);中国科学院声学研究所"所长择优基金"(No.GS13SJJ04);中国科学院青年人才领域前沿项目(No.O754021432)。

作者简介:李娟(1981-),女,硕士生,主要研究方向为自然语言理解处理技术;张全(1968-),男,博士,研究员,博士生导师,主要研究方向为 HNC 自然语言理解处理技术、计算语言学;贾宁(1981-),男,博士生,主要研究方向为自然语言理解处理技术。

收稿日期:2008-07-03 **修回日期**:2008-10-17

得到更大的褒贬义词集合。除此之外,也有使用机器学习等统计的方法来获取词语的语义倾向性的研究。

现有的相关研究都取得了值得肯定的成果,但是值得注意的是,虽然大部分的词语都具有唯一的倾向性,即只有褒义或只有贬义。但是也有部分词语,具有多倾向性,即同时具有褒义和贬义,需要根据上下文来判断。例如词语"骄傲",在"我们为你感到骄傲"中是褒义,在"骄傲使人落后"中是贬义。这种多倾向性词语的上下文褒贬判定是意见挖掘中一个不可忽视的部分,也是一个难点。以上的研究工作中都没有对这一问题进行研究,这也对词语的语义倾向性判定以及在此基础上进行的后续工作—语句乃至篇章的倾向性判定造成了不利的影响。

为了解决上述问题,本文使用了《现代汉语褒贬用法词典》等[3-5]已有的褒贬义词典,构成褒贬义词库,并使用了同义词词库,结合二元语法来判定词语的褒贬性。褒贬义词库包含词语6 619条,其中具有多个倾向性的词语为220条,同义词词库来自网络资源[6],包含词语5 223条,涵盖了大部分常用词语。褒贬义词典的使用,,减少了人工标注的误差。对于褒贬义词库中不包含的新词语,本文采用一个假设:词语和它的同义词具有同样的倾向性。基于这个假设,对新词语,查找同义词词库,再根据其同义词的倾向性确定该词语的倾向性。对于褒贬性词典和同义词词典都不包含的词语,利用SO-PMI算法[6]计算其倾向性。而对于褒贬义词典中具有多倾向性的词语,本文采取的办法是利用同义词典,结合二元语法来进行判断。一般而言,这种多倾向性词语会分布在多个同义词群中,用二元语法计算每个群内的词语在上下文中出现的概率,找出概率最大的群,再查询该群内的词语倾向性,作为该词语的倾向性。这种方法有效地解决了多倾向性词语的倾向性判定问题。

2 方案设计和算法介绍

通过查找褒贬义词典可以获得一部分词语的倾向性,但是单纯使用褒贬义词典存在两个问题:由于褒贬义词典规模的限制,无法处理没有在褒贬义词典中出现的词语;如果一个词语在不同的情况下可以是褒义词,也可以是贬义词,如何判定其倾向性。

对于第一个问题,本文采用同义词词典来对褒贬义词典进行扩充。同义词词典按照不同的语义将词语分为相应的同义词群,确保同一群内的词语其倾向性是相同的。如昂扬、奋发、高昂、振奋为一个同义词群,倾向性为褒义。败北、铩羽、失败、失利、战败是一个同义词群,倾向性为贬义。当处理的词语不在褒贬义词典内时,通过同义词词典查找其同义词群,以同义词群的倾向性作为该词的倾向性。

对于第二个问题,同样可以通过同义词词典来确定其倾向性。由于一个同义词群内的词语倾向性是一致的,因此一个同义词群只有一个唯一的倾向性。如果一个词 w 在不同的上下文环境中可以表达褒义也可以表达贬义或无倾向性,那么词 w 会同时出现在不同的同义词群中。通过确定词 w 在当前上下文环境中应属于哪一个同义词群,就可以判定 w 的倾向性。词 w 的倾向性判定问题转化为词 w 的同义词群归属判定问题,本文采用 N 元语法模型来确定 w 的同义词群归属。

句子 $S=w_1,w_2,\cdots,w_{i-1},w_i,w_{i+1},\cdots,w_n$,词 w_i 为具有多个倾向性可能的词。为简单计,设 w_i 有两种倾向性的可能。

$G_1=\{g_{1-1},g_{1-2},\cdots,g_1-n_1\}$ 和 $G_2=\{g_{2-1},g_{2-2},\cdots,g_2-n_2\}$ 分别是 w_i 在不同倾向性下所属的同义词群,g_{i-j} 为群内的词,i 为群的标号。由于同一词群的词语为同义词,因此考虑将 w_i 替换为同义词群内的词语。替换后生成的句子为

$$W=w_1,w_2,\cdots,w_{i-1},g_{j-k},w_{i+1},\cdots,w_n$$

求 g_{j-k} 使 $P(W/S)$ 最大,并认为此时的 g_{j-k} 所属的群即为 w_i 所属的群,该同义词群的倾向性即为 w_i 的倾向性。

$$P(W/S)=P(W)P(S/W)/P(S)$$

$P(S)$ 是未替换时句子的概率,是一个常数,可以不用考虑。$P(S/W)$ 是从替换的词串恢复到替换前词串的概率,替换前词串是唯一的,因此 $P(S/W)=1$。所以有:

$$\max(P(W/S))=\max(P(W))$$

即求解使替换后词串 W 概率最大的词 g_{j-k} 即可。

采用二元语法模型,有:

$$P(W)=\prod p(w_i/w_{i-1})=p(w_1/BOS)*p(w_2/w_1)*\cdots*$$
$$p(g_{j-k}/w_{i-1})*p(w_{i+1}/g_{j-k})*\cdots*p(EOS/w_n)$$

其中 $p(w_1/BOS)*\cdots*p(w_{i-2}/w_{i-1})$ 和 $p(w_{i+1}/w_{i+2})*\cdots*p(EOS/w_n)$ 是常数,不必计算,因此有:

$$\max(P(W/S))=\max(p(g_{j-k}/w_{i-1})\times p(w_{i+1}/g_{j-k})) \quad (1)$$

$$P(g_{j-k}/w_{i-1})=\frac{p(g_{j-k},w_{i-1})}{p(w_{i-1})}$$

$$P(w_{i+1}/g_{j-k})=\frac{p(w_{i+1},g_{j-k})}{p(g_{j-k})}$$

使用大规模语料库(本文使用了1998年1月的人民日报语料库)进行训练,根据大数定理,可知 $p(w_i)$ 的极大似然估计值等于词频,即有:

$$p(w_i,w_j)\approx\frac{k_{i,i}}{\sum_{x,y}k_{x,y}} \quad (2)$$

$k_{i,j}$ 是词对 (w_i,w_j) 在训练语料库中出现的次数,$\sum_{x,y}k_{x,y}$ 是训练语料库中所有词对出现的次数之和。

$$p(w_i)\approx\frac{k_i}{\sum_x k_x} \quad (3)$$

k_i 是词 w_i 在训练语料库中出现的次数,$\sum_x k_x$ 是训练语料库中所有词出现的次数之和。

根据式(2)、式(3)求得满足式(1)的 g_{j-k},g_{j-k} 所属的同义词群即可认定为词 w_i 所属的同义词群,该群的倾向性即为词 w_i 的倾向性。

对于褒贬性词典和同义词词典都不包含的词语,利用SO-PMI算法[7]计算其倾向性。即计算新词语与褒贬性词典中词语的互信息,取与其互信息较大的词语的倾向性作为该词的倾向性。词语 w_i 与 w_j 的互信息PMI为:

$$PMI(w_i,w_j)\approx\log(\#(w_i,w_j)/\#w_i\#w_j) \quad (4)$$

其中 $\#(w_i,w_j)$ 为词语 w_i 和 w_j 在训练语料中相邻出现的次数,$\#w_i$ 和 $\#w_j$ 分别为 w_i 和 w_j 在训练语料中出现的次数。

从而得到新词语的倾向性如下:

$$SO(w)=\sum_{w_i\in POS}PMI(w,w_i)-\sum_{w_i\in Neg}PMI(w,w_i)$$

其中 *POS* 表示褒义词集合,*Neg* 表示贬义词集合。

若 $SO(w)$为正值,则 w 的倾向性为褒义,若 $SO(w)$为负值,则 w 的倾向性为贬义。

3　实验结果与分析

实验所用的测试集为 114 个含有主观性评价的语句,这些语句都是从网络文章里面随机抽取,文章类型包括:当前时事的新闻报道与媒体评论、政治选举中候选人形势分析与预测、历史人物评析等。这些语句都是手工获取,选取时尽量做到相关度比较小,目的是使测试集的覆盖范围更广泛,使测试结果更准确,更能体现实验方法的普遍性。测试语句中所包含的主观性词语既包括经典常用词语,也包括网络新词语;既包括书面语,也包括口语。

本文的测试实验是这样进行的,对于测试集中的每个句子,挑选出其中的主观性词语,对其倾向性进行人工标注,结果保存待用。然后依次将每个句子输入系统,对其中的主观性词语进行自动标注。最后比较人工标注结果和系统自动标注结果,比较结果如表 1 所示。

表 1　倾向性判定结果

倾向性	测试语句数量	判定为该倾向的数量	正确的数量	召回率/(%)	正确率/(%)	F-Score/(%)
贬义	56	54	43	76.79	79.63	78.18
褒义	58	60	46	79.31	76.67	77.97

实验的结果显示,贬义的词语判定召回率较低,是由于部分贬义词语被判定为了褒义,如:

她长得很好看

否则,我让你好看

第一个“好看”是褒义词,第二个“好看”是贬义词,“好看”在这里呈现两种倾向性。“好看”虽然在褒贬义词典中出现了,但是在同义词词典中只出现了一个倾向性,即只在表示褒义的同义词群中出现,贬义的那一方面没有在同义词词典中出现。因此根据本文建立的模型无法正确地识别出贬义的倾向性。

小姑娘的眼神清澈而且单纯

他感觉自己的想法太幼稚单纯

第一个“单纯”是褒义词,第二个“单纯”是贬义词,意为头脑简单。“单纯”的两个倾向性义项在褒贬义词典和同义词词典中都出现了。根据本文的模型,判定“单纯”属于哪个义项分别需要计算两个义项的同义词同“单纯”前后词的条件概率 $p(g_{j-k}/w_{i-1})$和 $p(w_{i+1}/g_{i-k})$。但是在训练语料中查找后,发现“单纯”的两个同义词群中的词没有和“单纯”的上下文搭配的情况出现,因此计算出的结果无法正确区分两种倾向性。

由上面的分析可见,词典和语料库对判定结果有明显的影响,改进并扩充词典和使用更好的语料库能够提升判定的效果。

4　结束语

本文采用了基于褒贬义词典和同义词词典的方法来自动判定主观性词语的倾向性,该方法把同义词词典与褒贬义词典结合起来,扩大了褒贬义词典的范围,并使用 *N* 元语法模型有效地解决了多倾向性词语的倾向性判定问题。经过测试,验证了该思路的可行性和有效性。进一步的分析可以发现,影响本算法正确率提高的原因有两个,首先,由于本算法在解决一词多倾向性问题时依赖于同义词词典,然而本实验所使用的同义词词典收录的词语有限,从而影响了实验的正确率。将来使用本算法时,可以增加同义词词典的规模,这样可以提高算法的正确率。其次,适合于本实验的训练语料应当是数量较大且包括大量评论性语句的已经做过分词和词性标注的语料。本实验所使用的训练语料是人民日报的新闻语料,与所需要的语料吻合度不高,也影响了本实验的正确率。将来的工作应该包括构建一定规模的适用于倾向性分析或意见挖掘的专用语料。

下一步考虑在这两方面进行进一步的工作,通过建设更充分更有效的资源来实现性能的改进。并且尝试其它统计模型以获得比 *N* 元语法模型更好的效果。

参考文献:

[1] 朱嫣岚,闵锦,周雅倩,等.基于 HowNet 的词汇语义倾向计算[J].中文信息学报,2006,20(1):14-20.

[2] 娄德成,姚天昉.汉语句子语义极性分析和观点抽取方法的研究[J].计算机应用,2006,26(11):2622-2625.

[3] 张家太,徐彻.现代汉语褒贬用法词典[M].沈阳:辽宁人民出版社,1991.

[4] 王国璋.汉语褒贬义词语用法词典[M].北京:华语教学出版社,2001.

[5] 张伟,刘缙,郭先珍.学生褒贬义词典[M].北京:中国大百科全书出版社,2003.

[6] 同义词词典[EB/OL].(2008-03).http://www.365zn.com.

[7] Esuli A.Opinion mining[EB/OL].(2008-03).http://medialab.di.unipi.it.

(上接 73 页)

本文利用粗糙模糊集理论提出了一种模糊粗传递模型,利用模糊信息在传递过程中保持不变、发生损失与信息传递者知识之间的关系,得到提高模糊信息传递精确性的方法。给出了该模型在风险投资管理系统中的应用。模糊粗传递是粗糙模糊集的一种新的应用。

参考文献:

[1] 耿志强,朱群雄,李芳.知识粗糙性的粒度原理及其约简[J].系统工程与电子技术,2004,26(8):1112-1116.

[2] Mousavi A,Jabedar-Maralani P.Double-faced rough sets and rough communication[J].Information Sciences,2002(148):41-53.

[3] Pawlak Z.Rough sets[J].International Journal of Computer and Information Sciences,1982(11):341-356.

[4] Pawlak Z.Rough set theory and its applications to data analysis[J].Cybernetics and System,1998,29(7):661-668.

[5] Wang Hong-kai,Li Xiu-hong,Shi Kai-quan.Information measure in rough communication[J].An International Journal:Advances in Systems Science and Applications,2005,5(4):638-643.

[6] 刘纪芹.基于粗集的信息粗传递[J].系统工程与电子技术,2007,29(3):437-442.

[7] Liu Ji-qin.A fuzzy rough communication model and its properties[C]//Proceedings of 2007 International Conference on Machine Learning and Cybernetics,2007,7(7):3699-3703.

[8] Dubois D,Prade H.Rough fuzzy sets and fuzzy rough sets[J].International Journal of General Systems,1990(17):191-208.

[9] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001.