

doi:10.3772/j.issn.1000-0135.2012.10.008

基于 CRF 模型的半监督学习迭代观点句识别研究¹⁾

丁晟春 文能 蒋婷 孟美任

(南京理工大学经济管理学院信息管理系, 南京 210094)

摘要 本文从句子级的角度进行了中文文本的情感倾向分析,提出以 HowNet 中的情感词表为种子情感词集,采用基于 CRF 模型的半监督学习迭代方法获取大量评价词,然后依据中文词间的语义规则判断句子的极性的方法。将该方法应用于 COAE2011 中任务 2-观点句识别,在评价词的识别和观点句极性判断都取得了很好的结果。

关键词 CRF 观点句 半监督 情感倾向性

Sentiment Sentence Recognition through Semi-supervised Bootstrapping Based on CRF

Ding Shengchun, Wen Neng, Jiang Ting and Meng Meiren

(Department of Information and Management of Nanjing University of Science & Technology, Nanjing 210094)

Abstract During recent years, sentiment analysis about text in Chinese is becoming more and more popular in academic research. In this paper, sentiment analysis is processed on sentence level. Sentiment words published by HowNet is used as the original evaluated-word set, a large amount of evaluated-words are obtained by semi-supervised bootstrapping based on CRF model. Then sentiment sentence can be recognized by evaluated-words, and the polarity of sentiment sentence can be judged by the designed semantic rules.

Keywords CRF, sentiment sentence, semi-supervised, sentiment analysis

1 引言

随着网络上各类贴吧、论坛、博客、评论网站等的兴起,网络成为了一个供大众评论交流的大平台,出现了大量的主观性文本。这类文本中包含了大量有价值的观点信息,如社会舆论导向、网民焦点话题、经济发展趋势、用户体验及消费经验等。由此,中文文本的情感倾向性分析成为观点挖掘领域学术

研究热点,涵盖了自然语言处理、数据挖掘、信息抽取、机器学习等多个领域,获得多领域学者的关注。另外,由于这些大量有价值的观点信息具有非常重要的现实意义,由此也成为众多互联网企业关注的焦点。这类主观文本中包含了表达者对评论对象的褒贬态度,在一定程度上还表达了态度的强度。国内外许多研究者进行了相关的研究。刘鸿宇等^[1]把商品评论句子划分为四类:具有明显褒贬极性;含有褒贬评价词的数目相同,句子不带有明显褒贬极

收稿日期:2012年2月10日

作者简介:丁晟春,女,1971年生,南京理工大学信息管理系,副教授,主要研究方向:Web数据挖掘、信息检索、信息系统开发。E-mail: todingding@163.com。文能,女,1987年生,情报学硕士在读,主要研究方向:信息抽取与文本挖掘,情感倾向性分析评价。蒋婷,女,1988年生,情报学硕士在读,主要研究方向:数据挖掘、知识工程。孟美任,女,1988年生,情报学硕士在读,主要研究方向:文本挖掘、可信度分析。

1) 本文受国家自然科学基金项目“基于文本语义挖掘的商品评论信息可信度分析研究”(71103085)、教育部人文社会科学规划基金项目“基于语义的电子商务产品主/客观信息提取研究”(09YJA870015)、江苏省研究生科研创新计划“基于领域本体_CRF的商品主观”(CX10S-001R)的资助。

性;句子没有明显褒贬极性,但上下文句子有明显褒贬极性;中性句和客观句,即句子没有表达褒贬极性。王波等^[2]选择采用半监督-交叉训练方法,以迭代的方式挖掘评论对象和评价词,通过朴素贝叶斯分类器在只有很少标注语料的前提下取得的结果与大量标注语料情况下取得的结果非常接近。左维松等^[3]将规则和统计相结合进行了篇章级商品评论极性判别,实验中获得了较好的效果,采用机器学习的方法构建情感词典获取句子的褒贬极性,其中设计搭配规则以整句规则集优先的规则很好的处理了组合极性词和动态极性词的抽取,对于不满足搭配规则的则采用了判定表进行极性分析,避免了重要信息的缺失。Turney 等^[4]利用词汇“共现”(PMI)的关系,通过搜索引擎的返回结果数目确定词语与基准词的共现程度来计算词语的情感极性。张紫琼等^[5]采用了一种基于网络的无监督方法对情感短语的极性进行了判断,筛选 PMI-IR 方法中搜索引擎返回的结果,认为只有两个词在文本中出现在既定距离内才具有共现特征,避免了由于返回结果中实际不存在任何联系词汇产生的噪声。复旦大学的朱嫣岚等^[6]提出了基于 HowNet 的语义相关场中文情感倾向性计算方法,取得了很好的结果。另外,研究者还通过构建情感词典的方法判断词语的极性,姚天昉等^[7]以人工标注的方式构建了汽车领域的极性词汇表作为判断依据,Hatzivassiloglou 和 Wiebe^[8]构建了 73 个程度副词表用以计算词语的极性强度。

本文从句子级的角度进行了中文文本的情感倾向分析,提出以 HowNet^[9]中的情感词表为种子评价词集,并将该评价词集作为模型的情感特征,采用基于 CRF 模型半监督学习迭代方法获取大量评价词。最后利用评价词集,采用模式匹配的方法自动识别测试集中的观点句,并依据中文词间的语义规则判断观点句表达的总体极性。课题组将该方法用于 COAE2011 中任务 2-观点句的抽取^[10],要求从每个领域的测试集中自动识别出所有观点句并判断其表达的总体极性(褒义,贬义,或混合观点)。任务 2 中的观点句是指包含了对其他对象的评价的句子,不包括内心自我情感(如“心情”),因此,本文以是否含有这类用来对其他对象进行评价的词为判断句子是否为观点句的依据,这类词在本文中被定义为评价词,考虑否定词修饰后观点的情感倾向并不完全是反面的,所以本文进一步讨论涉及否定观点情感倾向的五分类划分(褒义,否定修饰后褒义,中性,否定修饰后贬义,贬义)。

2 基于 CRF 模型半监督学习方法构建评价词集

条件随机场(Conditional Random Fields, CRFs)^[11]最早由 Lafferty 等于 2001 年提出,它结合了最大熵模型和隐马尔可夫模型的特点,可以看成是一个无向图模型或马尔可夫随机场,它是一种用来标记和切分序列化数据的统计模型。

CRF 模型不同于产生式模型,它可以使用丰富的、彼此重叠的观察序列的特征,而且不需要很严格的前提假设;同时,不同于最大熵和马尔可夫等概率模型,不对单个标记归一化,而是在整个观测序列求解一个最优的标记序列,避免了标记偏置问题。

本文采用灵玖软件^[12]对测试集进行批量处理,获取测试集的分词及词性结果。本节将介绍以 HowNet 中情感词表为种子评价词集,采用基于 CRF 模型的半监督迭代学习方法构建评价词集。由于某些评价词具有领域特征,如金融领域的评价词具有很强领域特征,如“股指”,因此本文针对三个领域测试集(数码、金融、娱乐)分别构建评价词集。

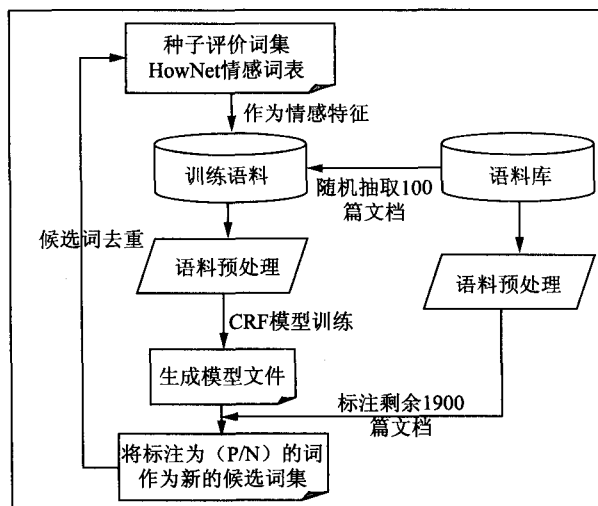


图1 基于 CRF 模型的半监督迭代的评价词集构建流程

针对分词后的各领域测试集,本文提出如下所示的基于 CRF 模型的半监督迭代基本流程:

(1) 每次随机选择 100 篇语料作为训练语料,剩余的 1900 篇语料作为标注语料。

(2) 对语料预处理,通过对语料的文本特征分析,发现单句中存在很多与判断是否为观点句不相关的词(停用词、量词等),并且这些词的存在会影响

模型对上下文内容语义特征的判断的准确性。因此本文借助相关词表对经分词后的语料中这类词进行剔除,作为模型的训练语料和模型标注语料。

(3) 将 HowNet 中情感词表作为种子评价词集,并将评价词集作为情感特征、结合词、词性构造 CRF 模型的特征集,评价词集的特征描述如表 1,特征模板如表 2。

表 1 特征集描述

特征类别	特征标记	特征描述
词特征(W)	W	词
词性特征(P)	Pos	词性
情感特征(S)	RP	当前词为 Hownet 中正面评价词语
	RN	当前词为 Hownet 中负面评价词语
	SP	当前词为 Hownet 中正面情感词语
	SN	当前词为 Hownet 中负面情感词语
	OP	当前词为新褒义评价词
	ON	当前词为新贬义评价词
	O	当前词为中性词

(4) 使用外部工具包 CRF++ -0.53 训练 CRF 模型,将生成新的模型文件对标注语料进行情感特征分类,分为三类:褒义(P)、贬义(N)、客观(O)。

(5) 将标注为(P/N)词作为候选评价词,在评价集中查找,如果该词不存在与评价集中,则作为新的种子评价词添加到评价词集中,并记录其情感倾向。

(6) 重复上述过程,直至不再产生新的评价词,最终获得评价集。

3 基于评价词集的观点句的识别

本文提出将单句中是否存在评价词作为判断单句是否为观点句的依据,深入研究了单句中的否定词和连词对整个单句观点的影响。对话料的文本特征进行分析,发现单句存在下述几个语义特征:

表 2 特征模板

特征类型	特征表示	描述
原子特征	$P_n P_{n+1}, W_n W_{n+1}, S_n S_{n+1} (n = -2, -1, 0, 1)$	单个特征
复合特征	$W_n P_n S_n, W_n S_n, P_n S_n (n = -2, -1, 0, 1)$	多个特征组合

注: n 表示在文本中观察词相对当前词的位置, n 的值表示观察值与当前词之间间隔词的个数。当 $n > 0$,表示观察词的位置在当前词的前面;当 $n < 0$,表示观察词的位置在当前词的后面。

3.1 句子中评价词的处理

1. 单个评价词的处理

因为单句中只有一个评价词,所以该句子被判定为观点句,并且该单句的情感倾向与句中唯一的评价词的情感倾向一致。

2. 多个评价词处理

单句中存在多个评价词,则该句子被判定为观点句。如果这些评价词都表示同类情感倾向(褒\贬),则这类单句情感倾向与评价词的极性(褒\贬)一致;但如果存在多个不同情感倾向,这类单句的情感倾向为“或褒或贬”。

3.2 句子中否定词处理

评论者在评论文本中经常使用否定词(不、没有等)与评价词一起表达对评论对象的评价。因此,本文进一步深入分析了否定词的相关内容,并且研究了其在评论文本中对句子观点倾向的影响。王力先生的在《中国现代语法》^[13]中对否定词的相关概念进行了定义,并提出了四种否定词的分法。

定义 1 否定作用:凡能在消极方面否定某一事情者。

定义 2 否定词:即为表示否定作用的词。包括无,非,没有,没,不,未,不曾,从未,从不,别,不要,甭等。

定义 3 双重否定:否定词的重复,和完全不用否定词,其效用并不完全相同。这有两种情形:一是如果是“不无”“未免不”“未必不”之类,则比普通正面的话显得弱些,同时也就委婉些;如果是普通的两个否定词,无论是否被别的词隔开,都带着若干情绪,使所说的话显得更有力。

对否定词的上述特征,本文以《中国现代语法》中提及的否定词为基础,即:无、非、没有、没、不、未、不曾、从未、从不、别、不要、甭,进行扩展得到常用的 29 个否定词。

否定词的否定范围包括单词和整个短语。短语里的修饰词越多,则否定的范围越小。例如,“不大”

并不完全等于“小”，“不亮”并不完全等于“暗”，“不老”并不完全等于“年轻”。一般而言，否定语总比肯定语的分量要轻些，因此否定语在表示贬义时比较委婉，而表示褒义时不够铺张。

考虑到当对褒贬评价词加以否定修饰时，否定组合后的极性并不是完全相反，而是比褒贬评价词的反义词的情感程度委婉一些，所以本文参照调查问卷中里克特量表，将倾向性情感极性分为五类（褒义，否定修饰后褒义，中性，否定修饰后贬义，贬义），如表3所示。

结合上述分析，经否定词修饰后，该否定评价词组合的情感倾向如表3所示为否定修饰后贬义和否定修饰后褒义，本文在判定含有否定词和评价词的单句的组合情感倾向时定义了如下规则：

(1) 如果单句含有否定词，则将其与紧邻其后的评价词合并，组合评价词情感倾向为评价词否定修饰后的贬义或褒义。

(2) 如果单句含有多个否定词且否定词间没有评价词，将多个否定词组合，根据“否定之否定为肯定”判断否定词组合的情感倾向。然后将否定词组合与紧邻其后的评价词合并，组合评价词情感倾向为评价词否定修饰后的褒义或贬义，如“不是不好”的情感倾向为“否定修饰后的褒义”。

(3) 只含有一个组合评价词的单句情感倾向与组合评价词的情感倾向一致，而含有多个组合评价词的单句情感倾向判断与含有多个评价词的单句（见3.1节中多个评价词处理）处理方法相同。

3.3 句子中连词处理

评论文本存在着大量的连词，连词对句子观点的情感倾向变化有非常重要的作用。孙汝建先生在《现代汉语》^[14]中对连词定义为：能够连接词、词组、句子和段落以表示两者关系的词，这种关系可以分为

联合和偏正关系。本文对影响情感倾向的连词进行了整理，见表4，并按照对情感倾向的影响对这些连词进行了三分类：并列连词、递进连词以及转折连词。

由于连词对观点句的情感倾向有重要的作用，对含有上述三类连词的观点句情感倾向判断规则如下：

(1) 如果连词两端是并列（如“而且”，“既……又……”等），这类连词两端的子句情感倾向是相同，所以含有这类连词的观点句的情感倾向为观点句中所包含评价词的情感倾向。

(2) 如果连词是递进（如“不但……而且……”，“尚且……何况……”），这类连词用来加强观点句情感倾向的程度，观点句的情感倾向为其所包含评价词的情感。

(3) 如果是转折（如“但是”，“即使”），这类连词一般被用来表示子句情感的转折。含有这类连词的观点句中有多个评价词，如果评价词情感倾向不一致，有褒有贬，则这个观点句情感倾向为“或褒或贬”；如果评价词的情感倾向一致，则这个观点句情感倾向与评价词一致。

4 评测与分析

本文对 COAE2011 任务 2 观点句抽取提供的 2000 篇 digital 领域语料进行了观点句抽取研究。

4.1 评价词获取分析

本文进行了两次评价词获取的迭代实验，每次从 2000 篇 digital 语料中随机抽取 100 篇文档作为训练语料，用训练所得模型标注其余 1900 篇语料，将模型标注结果为 P/N 的词作为新的褒/贬义评价词的初始来源，进行迭代构建评价词集。首先对模型识别出的褒/贬义评价词进行正确性判断，然后进行与候选评价词集进行比较，获取本次迭代识别出

表3 褒贬义情感五分类

褒义(1)	否定修饰后褒义(0.5)	中性(0)	否定修饰后贬义(-0.5)	贬义(-1)
非常同意	同意	不确定	不同意	非常不同意
好	不坏(不是不好)	一般	不好(不是不坏)	坏
喜欢	不讨厌		不喜欢	讨厌
和蔼可亲				面目可憎
趣味	不枯燥		没趣味	枯燥
爱	不恨		不爱	恨
美	不丑		不美	丑

的新的褒/贬义评价词。

在第一次迭代中,为了验证本方法的可行性,进行了预实验。在预实验中,随机抽取了 80 篇作为训练语料,20 篇作为测试语料,其中训练语料中人工标注褒义评价词 334 个,贬义评价词 67 个。预实验中,在 20 篇测试语料中识别出 36 个新的并且正确的褒义评价词,说明了该方法的是可行的。因此,用本次迭代实验训练所得模型,对语料集中其余 1900 篇文档进行标注。

标注结果中也出现了一些标注错误的词,如失望,差强人意,差,不,吃亏,顾名思义,冷峻,沉重,居高不下等。因此,对于这些标注错误的词进行正确性判断,即结合词的上下文通过人工判断的方式确定是否为评价词以及评价词的情感极性,但不作为本次迭代识别出的新词添加到评价词集中。经进行新评价词比较后,测试语料和待标注语料中共获取新的褒/贬义评价词个数如表 5。

在第二次迭代中,随机抽取了 100 篇文档作为训练语料,其中训练语料中人工标注新褒义评价词 191 个,新贬义评价词 42 个(即 HowNet 情感词表和一次迭代新词中没有的)。经新词正确修正和与候选词集比较后,二次迭代获取的新的褒/贬义评价词如表 6 所示,共标注了 164 个新褒义评价词和 16 个贬义评价词,正确标注褒义评价词为 105 个,贬义评价词为 8 个。一次迭代中产生的 192 个新褒义评价词有效地辅助完成了二次迭代的标注任务。

从两次迭代的实验结果可以看出,识别出的贬义评价词的数量非常少,两次迭代共识别出 16 个贬义评价词,其中标注错误为 8 个。这是由两个原因导致的,一是语料的不平衡导致,在 2000 篇实验语料中,贬义评价词相对较少;二是否定作用,有一部分贬义观点是通过否定词的否定作用来表示的。因此,对贬义评价词的获取可通过调整训练语料中褒贬观点的句子来提高识别的准确性。

表 4 影响句子情感倾向的连词

情感倾向变化	连词	连词类型
相同	和,跟,同,连同,与	联合
	及,以及	联合
	既……又……、一边……一边	联合
	于是,而,从而,继而	联合
	或,或者,要么……要么……	偏正
转折	否则,不然	
	不是……而是……	
	……虽然(固然)……但是……,而,不过,其实	
	宁可,与其	
递进	不但……而且,并,并且,且,而且,甚至	联合

表 5 digital 领域第一次迭代获取新的褒/贬义评价词

Data	褒义评价词	贬义评价词	新的标注正确的褒义评价词	新的标注正确的贬义评价词	评价词识别的准确率
测试语料	80	1	30	0	83.33%
1900 篇语料	1017	91	198	0	84.97%

表 6 digital 领域二次迭代获取的新的褒/贬义评价词

Data	褒义评价词	贬义评价词	新的标注正确的褒义评价词	新的标注正确的贬义评价词	评价词识别的准确率
1900 篇语料	1642	207	105	8	67.2%

4.2 观点句识别分析

本文采用了基于 CRF 模型的半监督迭代方法,并进一步结合语义规则的方法针对 digital 领域的语料集进行了观点句识别。

表7 修正后 digital 测试集观点句识别结果

参评系统标识	Precision	Recall	F1
8-NJUST_D	0.5839	0.55859	0.576245
Average	0.4050466	0.4768499	0.429604
Best	0.729751	0.798097	0.693304

按照 COAE2011 任务 2 的观点句句子的提交标准,本实验将 digital 领域的观点句识别结果提交给测评委员会,得到了表 7 中的测评结果,其中“参评系统标识”列的“8-NJUST_D”表示本方法参与评测获取的结果,“Average”表示所有参与评测系统提交结果的平均值,“Best”表示所有参与评测系统提交结果的最好结果。在 digital 领域,COAE2011 中正确观点句为 5675 个,本系统共识别出观点句 5429 个,其中与 COAE2011 给出的正确答案一致的(注:句子识别和极性的判断都一致)的结果有 3170 个观点句。表 7 中的结果可以看出在 digital 领域对观点句的识别明显高出平均值,说明了本文采用基于 CRF 模型的半监督迭代获取评价词集,并在此基础上进行观点句极性判断的规则设计是有效的。

另外,评论语料集中存在一些特殊的句式结构,例如比较句、假设性的句子等,这类句式结构会直接影响观点句提取的准确性。由于本方法没有考虑此类句式结构的影响,在一定程度上影响了观点句提取的效果。

5 结论

本文论述了观点句识别所采用的方法,主要探讨了基于 CRF 模型的迭代构建评价词集的方法,以及用于观点句极性判别的规则的设计。经过两次迭代,共获取新的并且情感极性判断正确的褒义评价词为 341 个,贬义评价词为 8 个,说明了本文提出的评价词集构建方法的有效性。将本文提出的观点句识别的方法应用于 COAE2011 任务 2 中观点句句子的提取,实验结果说明了该方法在观点句识别上的有效性。课题组将在以后的研究中,深入特殊句式结

构对观点情感倾向的影响,以提高对观点句的识别以及情感极性的判断。另外,将加入与其他算法的比较实验(如 SVM 等),进一步验证本方法的有效性。

参 考 文 献

- [1] 刘鸿宇,赵妍妍,秦兵,等. 评价对象抽取及其倾向性分析[J]. 中文信息学报,2010,(01):84-88,122.
- [2] Wang B, Wang H F. Bootstrapping both product properties and opinion words from chinese reviews with cross-training [C]//Proceedings of 2007 IEEE/WIC/ACM International Conference on Web Intelligence, 2007, 259:62.
- [3] 左维松,红英,张坤丽,等. 规则和统计相结合的情感分析研究[C]. 第五届信息检索会议学术论文集, 2009:280-287.
- [4] Turney P D, Littman M L. Measuring Praise and Criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.
- [5] Zhang Z Q, Li Y J, Ye Q, et al. Sentiment Classification for Chinese Product Reviews Using an Unsupervised Internet-based Method [C]//Proceedings of 2008 International Conference on Management Science & Engineering (15th), 2008, 1 and 2:3-9.
- [6] 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.
- [7] 姚天昉,聂青阳,李建超,等. 一个用于汉语汽车评论的意见挖掘系统[C]. 中国中文信息学会二十五周年学术会议论文集,2006:260-281.
- [8] Hatzivassiloglou V, Wiebe J M. Effects of adjective orientation and gradability on sentence subjectivity [C]//Proceedings of the 18th conference on Computational linguistics, 2000:299-305.
- [9] 知网. [EB/OL] [2011-08-16]. http://www.keenage.com/html/c_index.html.
- [10] 许洪波,孙乐,姚天昉,等. 第三届中文倾向性分析评测(COAE2011)总结报告[C]. COAE2011.
- [11] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. ICML, 2001:282-289.
- [12] 灵玖软件. [EB/OL] [2011-05-09]. <http://www.ling-join.com>.
- [13] 王力. 中国现代语法[M]. 北京:商务印书馆,1954:123.
- [14] 孙汝建. 现代汉语[M]. 南京:南京大学出版社, 2003:300.

(责任编辑 化柏林)