中文褒贬义词语倾向性的分析

王根[1]，赵军[2]

（中科院自动化研究所模式识别国家重点实验室，北京 100080）

摘 要： 倾向性语言是观点表达的重要形式，识别出文本中的倾向性语言是挖掘文本中观点信息的关键工作之一。本文分析了褒贬义词语在句子中所起的主观作用，旨在以此判定文本句子中是否带有倾向性。本文认为褒贬义词语在句中所起的作用应该由词语在句中位置和自身意义共同决定。为了证明这两点我们首先对语料库进行了统计和分析，然后分别从这两方面分析了词语在句子中表现的倾向性：一方面在同时考虑倾向性强弱和极性的测度下，提出用极坐标来表示一个词语的倾向性，并使用一种均衡化互信息的方法，探讨了词语在统计意义下，独立于具体上下文体现出的自身倾向性；另一方面以包含上下文的2000形容词标注语料为例，使用最大熵方法测试了30个上下文特征的对词语倾向性的作用。最后给出了评测的方法和结果。

关键词： 主观性；倾向性；上下文；观点分析

Orientation Analysis of Chinese Word

Gen WANG, Jun ZHAO

(National Lab. of Pattern Recognition, Beijing 100080)

Abstract: Subjective language contains lots of information about personal opinion and thoughts. This paper focus on the role of orientional words in sentence context, instead of individual word only. For this purpose, an general analyze on the corpus precedes to prove both prior orientation and the context impact are indispensable. Then we proposed two method to calculate the two aspects respectively, one is a revised mutual information, the other is machine learning. And the evaluation of them follows in the experiment.

Keywords: Subjective; orientation; context; opinion analysis

1 简介

与客观实体信息相同，主观性的语言表达了说话人的情感和态度[1]，是文本信息的重要组成部分。识别出主观性语言中带有倾向性的语句，可以帮助理解篇章的写作目的和作者的态度立场。能够为面向国家安全的网络社会态势分析、信息过滤，面向新闻情报的收集管理，面向商务需求的市场调查、产品反馈，面向金融需求的预测和分析、面向政府管理的民意分析等提供信息技术的支持。

判断句子中词语的倾向性也缺少一种客观的准则。为此，本文给倾向性语言下了一个模糊的定义：如果一个词语将说话人对某事物、人物的某种态度表达出来了，那么这个词语在句子中就表现出了倾向性。例如：

“主力巡山队组织大型打击盗猎活动，多是精壮的藏族小伙子，特别能吃苦。”

“美国已成为达赖集团分裂活动的总后台，达赖则沦为美国反华的忠实工具。”

“李达是我国传播马克思主义的先驱者和中国共产党的创始人之一，他为中国共产党的 建设特别是为党的

作者简介：王根(1981-)，男，贵州贵阳，在读硕士生。Email：gwang@nlpr.ia.ac.cn

思想理论建设做出了重要的贡献。”

分析倾向性语言一个重要的特征来源于句子中的倾向性词语，而褒贬义词语是我们能够比较容易获得，并且是十分重要的一部分。但很遗憾并不是所有褒贬义词在句子中都表现出倾向性：一方面，很大一部分褒贬义词语本身并不带有很强的主观性，例如“明显”“人才”；另一方面，在特定的上下文中它们常常只体现出一些客观的意义，只是表达指称，限定，或者作为一个部分构成其他词语等，对于带有客观义项的褒贬义词语还可能只是取了它的客观义项。例如：

“大部分人不会选择与网友见面，没必要破坏自己美好的想像。”

“房价的高低主要取决于市场需求的强劲以及外界的环境。”

“建立开放、流动、竞争、协作的科学研究机制。”

为了给判定句子的主观性倾向性选择合理的特征，就必须首先排除这些对句子倾向性没有贡献的褒贬义词语。本文认为来自于以下两个方面的因素对分析词语在句子中的倾向性有很大作用：

1. 词语本身的倾向性强弱和倾向性的极性偏向[2]。

2. 词语在句中的位置、句子的句型和语气。

本文的主要工作就是围绕以上两点展开。第 2 节介绍了国内外的相关工作，第 3.1 节是语料库分析和标注时的一些感性认识，第 3.2 节介绍词语本身先验倾向性的计算，第 3.3 节介绍判定上下文影响的选用的特征。第 4 节给出了实验方法和实验分析的结果。最后做了一个简短的结论。

2 相关工作

对应于上文提到的第一方面，判定词语先验倾向性，目前国内外主要有三种方法：

1.基于语义词典的扩展和抽取。比较具有代表性的工作是 Hatzivassiloglou[3]在 WordNet 上扩展，Hiroya[4]利用电子自旋模型在普通字典上扩展。复旦大学朱嫣岚、黄萱菁[5]在中文 HowNet 上利用词语间的相关性和相似性为测度分析了词语的倾向性。

2.基于标注语料库的抽取。利用带有主观性的语料进行统计，抽取其中的主观性表达，这种方法不仅仅可以找到词语，还可以找到词组和搭配，但是不能计算倾向性的极性。相关工作有 Weibe[6][7]

3.基于大规模语料库信息的判定和抽取。Turney[8]从大规模语料中利用词语和起始倾向性词语组的同现统计信息判定该词语的倾向性强弱并以此扩展倾向性词语的集合。

本文的判定倾向性的方法是在 Turney[8]采用的互信息的方法下延伸的，我们目的不是找到并扩展倾向性词语，而是判定倾向性词语的极性强弱和极性偏向。

对于第二方面，最早期的工作包括 Wiebe[9]仅利用倾向性词语和标点符号作为特征的贝叶斯分类。

2003 年 Yu 和 Hatzivassiloglou[10]选用带极性和词性标注的 tri-gram 作为特征的单层朴素贝叶斯分类器对句子进行主客观分类，但是我们的方法关注在句子中词语是否真正表达出倾向性，而不是直接判定句子是否是主观性的语言。

Pang 和 Lee 在其 2004 年的文章[11]中提出一种利用图论中最小割集方法判别方法，识别电影评论中的客观句，最小割集的方法更关注和邻近句子的判别结果，对于本地句子 Pang 和 Lee 的方法还是最简单的 unigram 的朴素贝叶斯方法，没有考虑词语对句子倾向性的贡献。

与本文最相近应该是 Wilson、Wiebe 和 Hoffmann 在 2005 年的工作[12]，他们探讨的也是在上下文环境中判定词汇对句子整体倾向性的作用，同样选用了大量的特征，使用 AdaBoost 的统计学习方法。但和他们工作不同，我们的这部分工作特点表现在：以中文为出发点，选用的特征大都依据中文的语言特点，用语义类别替换词语本身，没有采用句法分析的特征，使用的是最大熵模型来分类。

其他的相关工作，还包括在判定文档整体的主客观性或极性的目标上[13][14]。

3 计算

3.1 先验倾向性的计算

本文采用互信息来计算褒贬义词语的先验倾向性。这种方法是基于假设：一个词如果与其他褒贬义词语的同现互信息较大，说明它的主观性较强，如果和褒义词集的同现互信息大于贬义词集的那么，该词偏向于正倾向性。为同时表达主观性的强弱和倾向性的偏向，本文采用极坐标来表示词语的先验倾向性，模表示词语主观性的强弱，而夹角则反应了倾向性的方向。

互信息的计算：如果 A，B 是两个事件，具体地说，就是一个词的出现。$p(A,B)$ 表示 A、B 两个事件同时发生的概率，具体的说就是两个词同时出现；$p(A), p(B)$ 分别表示 A 事件单独发生的概率和 B 事件单独发生的概率。采用拉普拉斯平滑后我们这样估计：

$$I(A,B)=\log\frac{p(A,B)}{p(A)\cdot p(B)}=\log\frac{\frac{(n_{win}(A,B)+\lambda)}{\sum_{A,B}(n_{win}(A,B)+\lambda)}}{\frac{n(A)}{N}\cdot\frac{n(B)}{N}} \qquad (1)$$

极坐标的表示：X 轴表示倾向为正的极性，Y 轴表示倾向为负的极性，一个词的主观性表示为（x，y）的一个向量，强弱由该向量的模表示，偏向由向量与 x 轴的夹角 0 表示。

连续倾向性的计算：将词表中两两词之间的互信息计算出来后，对于其中一个词，将与它相关的词按初始的倾向性相加：

$$y=P(a)=\sum_{b\in Pset}I(a,b);x=N(a)=\sum_{b\in Nset}I(a,b) \qquad (2)$$

其中 *Pset* 是所有词典里面极性为褒义的词的集合，*Nset* 是贬义词集合。对于 $I(a,a)$ 我们是这样计算的：

$I(a,a)=\log\frac{1}{p(a)}$ 这里借用了 I 的符号，其意义是单纯的熵的含义。

归一化均衡化：归一化将计算出来向量的模限定在 (0,1) 之间，并且尽量拉开数值的分布，采用了图像处理中常用到的均衡化的方法。

3.2 判定上下文的影响

我们在自己给出的定义下，标注了形容词在 2000 个句子中是否真正包含了倾向性。在学习时为了逼近句法上的结构和语义层上的信息选择了以下相关特征：

1	nom	当前指示词的小义类	11	nom	前方“的地得”	21	bin	句中有无疑问词
2	nom	当前指示词词性	12	bin	指示词前面有没有系动词	22	bin	是否在书名号间
3	nom	后一个词词性	13	bin	被动	23	bin	是否在时间地点状语之中
4	nom	前一个词词性	14	bin	指示词前方有无动词	24	bin	是否包含“应该”语气
5	nom	句中所有副词小义类	15	bin	后方有无动词	25	nom	当前词的倾向性
6	nom	后名语义大类	16	bin	可以、能够的语气	26	num	整个句里的倾向词有多少个
7	nom	前名语义大类	17	bin	有无否定	27	num	整个句里的正倾向词个数

8	nom	前动语义小类	18	bin	句末有没有问号	28	num	整个句里的负倾向词个数
9	nom	后动语义小类	19	bin	前方有无转折连词	29	bin	是否有限定性代词和命名实体
10	nom	后方“的地得”	20	bin	前方有无假设连词	30	nom	前面的一个代词的小义类

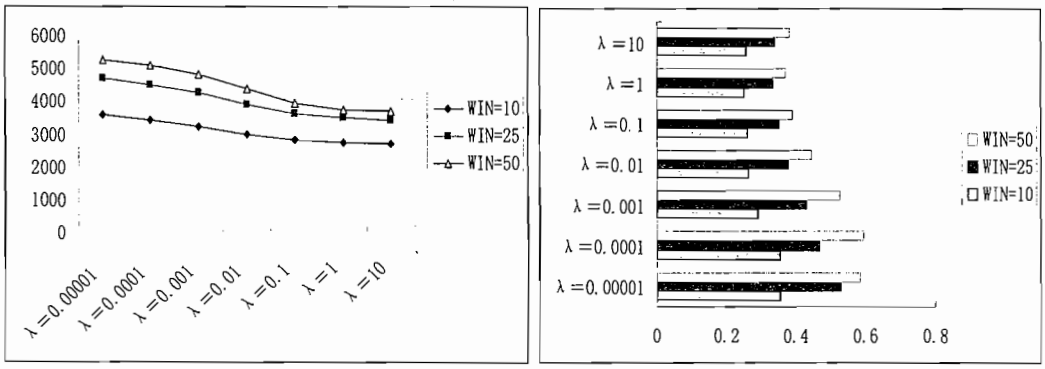
特征分为二值特征，枚举特征和计数特征三类。小义类是用《现代汉语分类词典》的第三级类别，语义大类是《现代汉语分类字典》的第二级类别。

4 实验及分析

词语先验倾向性计算的评测：由于词汇本身并没有一个比较准确的连续强弱数值，所以采用了以下的方法来进行评测和参数选择：首先选用了《学生褒贬义词典》中的词作为作为标准词，假设标准词集的倾向性比较强，据此提出以下两种评价方法：1. 倾向性向量的模在对数词频加权下的打分，即 $score(\lambda, win) = \sum_{w \in O} M_w \cdot \log N_w$

其中，O 是标准词的集合，λ, win 分别是拉普拉斯平滑参数和统计同现时选取的窗口大小，M 是窗口为 w 下计算出来的词语倾向性向量的模值，N_w 是词语在语料库中的词频。

2.计算倾向性模的强度大于 0.5 的词占标准词集的比例。下图是在新浪网 1G 的语料下计算的结果：



从结果中可以看出，随着平滑参数的减小和窗口的扩大 *score* 值会变好，但是从第二项评价中可以看到，当平滑参数更小时，标准词集中倾向性强度大于 0.5 的词比例会变小，说明平滑因子取小会偏向于提高词频大的词语的强度。

上下文特征的影响：这里使用的语料是从 6047 篇人民网新闻中抽取出的句子，包含形容词 2000 个，抽取方法是随机的但是用概率地方法控制了相同词语出现的次数和相同句子出现的次数。使用 maxent 工具包，在 5 交叉验证的比较了两组特征的准确率，直接采用未做特征选择的训练方法；实验在不同的特征集下作了测试：

Part1	**2、3、4、5、8、9、12、14、15、16、20、29、30**
Part2	**2、3、4、5、8、9、10、11、12、15、16、17、18、19、20、24、29、30**

Part1	**69.14**
Part2	**70.617**
All	**70.4177**

从上面的数据可以看出，句子中倾向性词语的数量对判定有作用，在看数据的同时应当还考虑到，由于标注时采用主观的标准，很有可能带来标注准则不易把握带来的不一致，这同样是影响结果的一大原因。

5 结论与展望

主观性语言的分析有着的重要的应用和研究意义，但目前的研究尚处于起步阶段，有很多新的问题有待研究

和解决，本文以研究词汇在句子中表达表达的倾向性为入口，仅仅是一次尝试。文本的更加细致的情感分析和分类，情感主体、客体的识别等也都是主观性语言分析的重要方面。

从微观上看，主观性语言涉及了大量的语义信息，依赖句法分析，为 NLP 提供了很好的研究内容。从宏观上看，主观性语言大量存在，其不依赖语义和句法的统计规律，依然可以为应用提供部分有效的支持，所以今后主观性语言的分析可能会朝着这两个方向发展。

参考文献：

[1] 沈家煊. 语言的“主观性”和“主观化”.外语教学与研究(外国语文双月刊).2001 年 7 月第 33 卷第 4 期

[2] HATZIVASSILOGLOU, V. AND WIEBE, J. M. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of 18th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2000 New Brunswick, N.J.

[3] Hatzivassiloglou, McKeown.Predicting the semantic orientation of adjectives. In Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, pages 174 - 181, Madrid, ES, 1997. Association for Computational Linguistics.

[4] Hiroya Takamura Extracting Emotional Polarity of Words using Spin Model ACL05

[5] 朱嫣岚，闵锦，周雅倩，黄萱菁，吴立德. 基于 HowNet 的词汇语义倾向计算. 中文信息学报 2006 Vol20 No.1

[6] WIEBE, J. M..Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, Calif.

[7] J.Weibe Theresa Wilson.Learning Subjective Language .Computational Linguistics Volume 30,No.3.

[8] Turney Peter2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems.2003

[9] Rebecca Bruce and Janyce Wiebe. 1999. Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, 5(2).

[10] Hong Yu, Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*.

[11] Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the ACL, 2004*.

[12] TheresaWilson, JanyceWiebe, Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language.Processing (HLT/EMNLP), pages 347–354, Vancouver, October 2005.*

[13] Peter Turney.. Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002 Philadelphia, Pennsylvania.

[14] Pang B., Lee L., Vaithyanathan, S.: Thumbs up? Sentiment Classi_cation Using Machine Learning Techniques. In: Proceedings of the Conference on EMNLP （EMNLP-2002). (2002) 79{86}