汉语主观性词典创建方法

张靖[1]，金浩[2]

(1. 攀枝花学院 网络中心，四川 攀枝花 617000; 2. 南京大学 计算机科学与技术系，江苏 南京 210093)

摘 要：根据汉语情感分析现状和需求，分析和研究了从目标语料库自动获取汉语主观性词典，提出了一种主观性词典创建方法，定义了主观性词典和语言模型，设计了自适应主观性自举算法和主观性属性特征模型，实现了主观性词条中情感倾向、主观性强度和词汇主客观自动判别。采用机器学习方法证明，提出的汉语主观性词典自动创建方法高效，性能优良。

关键词：情感分析；主观性词典；创建方法；机器学习；模型；算法

中图分类号：TP391 **文献标识码**：B **文章编号**：1000-436X(2010)8A-0172-05

Creation method of Chinese subjective lexicon

ZHANG Jing[1]，JIN Hao[2]

(1. Computer Network Center, Panzhihua University, Panzhihua 617000, China;

2. Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

Abstract: Based on the current situation of and demand for Chinese sentiment analysis, the method of automatically obtaining subjective lexicon from the target corpus was studied. A creation method of subjective lexicon was presented, the subjective lexicon and language model were defined, a self-adaptive subjectivity bootstrapping algorithm and the characteristic model of subjectivity attribute were designed, and all these lead to the realization of the automatic judgment of sentiment polarity, subjectivity intensity and the subjectivity and objectivity of a word in the subjectivity entry. Experiments prove that by using machine learning the proposed method of automatic creation of Chinese subjective lexicon is highly efficient and with excellent performance.

Key words: sentiment analysis; subjective lexicon; creation method; machine learning; models; algorithms

1 引言

主观性分析工作主要集中在情感分析上，识别出情感、评价和判断的正负面情况。文本情感分析已成为自然语言处理热点问题[1,2]，但在国内尤其是针对汉语情感分析和主观性分析研究还刚开始。由于缺乏基础数据集（主观性词典、情感标注语料）[3,4]，进一步的汉语情感和主观性分析受到制约。

由于主观性的表达与语言本身的关系比较紧密，在处理不同的语种时需要考虑：①语言建模的差异以及主观性表达方式的差异（不同语种语言学上情感表达的方式不一样），因此主观性研究工作在跨语种时不能全部复用，如有关汉语[1~4]、日语[5]、德文[6]的主观性等；②主观性基础数据集是整个主观性分析和情感分析的基础，大部分为手工创建、手工标注的主观性词典和标注了主观观点的标注语料库，人工方法不但非常消耗时间，而且也难保证个

收稿日期：2010-06-02

基金项目：四川省科技基金资助项目(2009zr0159)

Foundation Item: The Research of Science and Technology Department Program of Sichuan Province (2009zr0159)

人经验水平差异导致准确率的差异。在整个情感分析之中，由于主观性词典作为最小粒度的主观性知识的基础地位，主观性词典的自动快速生成创建方法研究成为了有重要价值的研究内容[7,8]。

近年来，结合自然语言处理和人工智能 2 种方法来进行情感分析研究成为了主流，一般在细粒度（词语、短语、句子）级别上主要依赖自然语言处理的方法、结合机器学习，在较大粒度（句子、文档、多文档）级别上多使用人工智能、机器学习的方法。基于 COAE2008 目标语料库，从语言学角度结合机器学习方法尝试情感分析的自动主观性词典创建方法的研究。

2　主观性词典条目和语言学模型定义

2.1　主观性词典条目定义

根据自然语言情感主观性特征属性[9]，进行主观性词典条目定义。

定义 1　Entry := Word POS Polarity Intensity Context

Word := {词语}；POS := {词性}；Polarity := {positive, negative, other}；Intensity: = {strong, weak}；Context: = {SEPs}。

Word 表示该情感词汇的文本特征；POS 表示该情感词汇的词性；Polarity 表示该情感词汇的倾向性；Intensity 表示该情感词汇的情感强度；Context 表示该情感词汇具有以上情感属性的上下文特征。

2.2　PTBLD 语言学模型建立

语法结构由词条组成，词条之间的链接非对称关系称为依赖，根据词性信息及二元词语语法依赖关系[10]语法语义信息。定义 PTBLD（POS, tagged bigram lexical dependency）作为基础语言学模型。

定义 2　PTBLD := Relation (GovernWord GovernPOS GovernPosition DepWord DepPOS DepPosition)

Relation 为词语语法依赖关系，表示是何种修饰作用；GovernWord 为支配词；GovernPOS 为支配词的词性；GovernPosition 为支配词在句子中的位置；DepWord 为依赖词；DepPOS 为依赖词的词性；DepPosition 为依赖词在句子中的位置。

3　主观性提取模式以及算法设计

3.1　主观性提取模式

根据 PTBLD 的特性，定义主观性提取模式。

定义 3　主观性提取模式

SEP: = GovHook GovPOS Relation DepPOS DepHook

GovHook := {govword, *, null}

GovPOS := {主导词词性, *}

Relation := {依赖关系}

DepPOS := {依赖词词性, *}

DepHook := {depword, *, null}

3.2　主观性词汇的获取

在词汇发现、词典的扩展过程中自举是一个常用的方法[3,4]，设计主观性自举方法，用中文主观性种子词典直接从目标语料库中提取出主观性词汇。

采用 Score(SEP) 和 Score(Word) 2 个评价函数分别对 SEP 和候选主观性词汇进行评分。

$$\mathrm{Score}(SEP_i)=\frac{F_{i,\mathrm{FIMOD}}}{N_{i,\mathrm{FIMOD}}}\times \mathrm{lb}(F_{i,\mathrm{FIMOD}}) \tag{1}$$

$F_{i,\mathrm{FIMOD}}$ 是通过主观性提取模式 i 在 FIMOD 下提取的主观性词汇的数量；$N_{i,\mathrm{FIMOD}}$ 是这个主观性提取模式提取的所有词汇的总数。

$$\mathrm{Score}(\mathrm{Word}_j)=\frac{\sum_{i=1}^{P_j}\mathrm{lb}(F_{i,\mathrm{FIMOD}}+1)}{P_{j,\mathrm{FIMOD}}} \tag{2}$$

其中，$P_{j,\mathrm{FIMOD}}$ 是在 FIMOD 下提取出词语 word_j 的提取模式的数量。

3.3　自适应主观性自举算法设计

设计自适应主观性自举（Bootstrapping）算法。

算法的输入：汉语主观性种子词典；原始语料库的 PTBLD 库。

算法的输出：面向该语料库的汉语主观性词典。

参数：k, m, n, FIMOD。

其中，k 为每次从排序的所有 SEP 中返回的前 K 个 SEP；m 为每次候选 SEP 的叠加量；n 为 Bootstrapping 迭代过程中，从候选主观性词汇列表中提取的主观性词汇的数量；FIMOD 为使用主导词位置来提取主观性词汇还是从依赖词位置来提取主观性词汇，值域为{*GOV*, *DEP*, *BOTH*}。

初始化阶段：

1) SEPs = {所有的主观性提取模式}；

2) ValidSEPs = { } #有效 SEP；

3) CandidateWords = { } #候选主观性词汇；

4) lexicon = {seed words} #为主观性种子词典中包含的所有词汇。

BOOTSTRAPPING 阶段：

1) 对 SEPs 使用式(1)进行评分，并从高到低排序；

2) 提取前 k 个 SEP 作为 ValidSEPs；

3) 使用 ValidSEPs 依据 FIMOD 确定的提取位置，提取出来的所有词汇作为 CandidateWords；

4) 使用式(2)对 CandidateWords 进行评分，并从高到低进行排序；

5) 把 CandidateWords 中的前 n 个不存在 lexicon 的，并且主观性判别为真的词汇加入到 lexicon 中，如果没有提取出主观性词汇，则主观性 Bootstrapping 过程结束，否则进行第 1)步。

4 主观性词条属性自动判别

情感分析关注主观观点中的感情倾向以及该倾向相关的表达者（源）、观点针对的人或者事物（目标）、观点的强度等信息，与之对应的主观性词典中词条的 5 个属性 Word、POS、Polarity、Intensity 和 Context，其中 Word 和 POS 信息在词法解析词条提取时就可获取，余下属性在主观性自举时尚没有确定。

为了获得最佳的实验结果，尝试了多种包括 Na ve Bayes、SVM（Support Vector Machine）、Basic Rule, KNN（K-Nearest Neighbor）和 Decision Tree 在内的机器学习算法及特征选择算法，利用定义的 PTBLD 语言学模型对情感倾向、主观性强度以及主观性进行特征的建模。特征模型的基础上，使用汉语主观性种子词典作为标注数据，利用机器学习算法构造分类器进行自动属性判断。

4.1 情感倾向特征建模

设计汉语主观性种子词典倾向性特征向量 ***PFeature*** 为

$$\boldsymbol{PFeatureSet}(w)_{P-S}=\begin{vmatrix} PFeature_{\text{positive}}(w) \\ PFeature_{\text{negative}}(w) \\ PFeature_{\text{other}}(w) \end{vmatrix} \tag{3}$$

以及特征矩阵

$$\boldsymbol{PFeatureMatrix}(w)=\begin{vmatrix} PFeatureSet_{\text{Gov-Same}}(w)^{\mathrm{T}} \\ PFeatureSet_{\text{Gov-Diff}}(w)^{\mathrm{T}} \\ PFeatureSet_{\text{Dep-Same}}(w)^{\mathrm{T}} \\ PFeatureSet_{\text{Dep-Diff}}(w)^{\mathrm{T}} \end{vmatrix} \tag{4}$$

在此特征矩阵上用机器学习方法来构造情感倾向分类器。

4.2 主观性强度特征模型

使用情感倾向特征以及主观性强度特征的联合特征，扩展情感倾向互信息特征模型，添加主观性强度互信息特征。

$$IFeature_{p-s}(w)=\frac{\sum_{i=1}^{k} IPMI(w,w_k)}{k} \tag{5}$$

$$\text{where } w_k \in \{\text{words extracted from p-s}\}$$

其中，主观性强度互信息特征：

$$IPMI(w,w_j)=\text{lb}\left(\frac{Intensity(w_j)\,p(w,w_j)^2}{p(w)p(w_j)}+1\right) \tag{6}$$

$$\text{where } w \neq w_j$$

构造主观性强度特征矩阵：

$$\boldsymbol{IFeatureSet}(w)=\begin{vmatrix} PFeature_{\text{Gov-Same}}(w) \\ PFeature_{\text{Gov-Diff}}(w) \\ PFeature_{\text{Dep-Same}}(w) \\ PFeature_{\text{Dep-Diff}}(w) \\ IFeature_{\text{Gov-Same}}(w) \\ IFeature_{\text{Gov-Diff}}(w) \\ IFeature_{\text{Dep-Same}}(w) \\ IFeature_{\text{Dep-Diff}}(w) \end{vmatrix} \tag{7}$$

其中，$PFeature_{\text{Gov-Same}}(w)$为通过同一句子主导词位置提取的情感倾向特征值，$PFeature_{\text{Gov-Diff}}(w)$为不同句子主导词位置提取的情感倾向特征值，$PFeature_{\text{Dep-Same}}(w)$为同一句子依赖词位置提取的情感倾向特征值，$PFeature_{\text{Gov-Diff}}(w)$为不同句子依赖词位置提取的情感倾向特征值；$IFeature_{\text{Gov-Same}}(w)$为通过同一句子主导词位置提取的主观性强度特征值，$IFeature_{\text{Gov-Diff}}(w)$为不同句子主导词位置提取的主观性强度特征值，$IFeature_{\text{Dep-Same}}(w)$为同一句子依赖词位置提取的主观性强度特征值，$IFeature_{\text{Dep-Diff}}(w)$为不同句子依赖词位置提取的主观性强度特征值。

使用汉语主观性种子词典进行特征选择得到有效特征集合，结果特征权重为[0 0 1 0 0 0 1 0]，选择其中有效的特征（权重值为 1）得到主观性强度有效特征集合。

$$IFeatureSet(w)=\begin{vmatrix} PFeature_{\text{Dep-Same}}(w) \\ IFeature_{\text{Dep-Same}}(w) \end{vmatrix} \tag{8}$$

其中，$PFeature_{\text{Dep-Same}}(w)$为同一句子依赖词位置提取的情感倾向互信息特征，$IFeature_{\text{Dep-Same}}(w)$为同一句

子依赖词位置提取的主观性强度互信息特征值。

4.3　词汇的主客观特征模型

利用情感倾向和主观性强度特征来进行词语的主观性特征的建模，产生符合主客观判别的主观性特征集合。利用式（7）的情感倾向特征集和主观性强度特征集的联合特征作为特征选择的依据，使用主观性种子词典手动标注时的标注数据，进行特征选择，结果特征权重为[0 0 1 1 0 0 1 1]，选择其中有效的特征（权重值为 1）得到主观性强度有效特征集合。

$$SFeatureSet(w)=\begin{vmatrix} PFeature_{\text{Dep-Same}}(w) \\ PFeature_{\text{Dep-Diff}}(w) \\ IFeature_{\text{Dep-Same}}(w) \\ IFeature_{\text{Dep-Diff}}(w) \end{vmatrix} \tag{9}$$

其中，$PFeature_{\text{Dep-Same}}(w)$为通过同一句子主导词位置提取的情感倾向特征值，$PFeature_{\text{Dep-Diff}}(w)$为不同句子主导词位置提取的情感倾向特征值，$IFeature_{\text{Dep-Same}}(w)$为同一句子依赖词位置提取主观性强度的特征值，$IFeature_{\text{Dep-Diff}}(w)$为不同句子依赖词位置提取的主观性强度特征值。

5　实验及结果分析

5.1　情感倾向自动判断

5 种机器学习算法全部分类器性能曲线如图 1 所示。

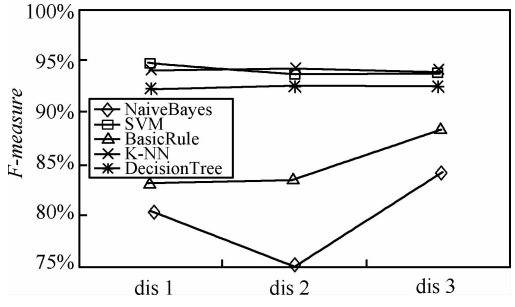

图 1　情感倾向自动判断的全部分类器 *F* 值性能曲线

实验表明，基于二元语法依赖关系的双层情感倾向互信息特征模型，所有分类器对情感倾向性的分类准确率（Accuracy）和 F 值（F-measure）都比较好，特别是 SVM（Accuracy: 95.47%, F-measure: 93.90% ）、 KNN(Accuracy: 94.98%, F-measure: 93.75% ） 和 Decision Tree （ Accuracy:92.62%, F-measure:92.62%）效果较理想。Peter Turney 对形容词的倾向判断[11]准确率在 78%～92%之间，Kanayama 等提出了一种非监督学习方法来获得领域相关的倾向性词典[5]，其准确率达到 94%，本文中性能表现最好的 SVM 准确率达到 95.47%，*F* 值达到 93.90%，性能有了较大提高。

5.2　主观性强度自动判断

主观性强度全部分类器性能曲线如图 2 所示。实验表明在本文的强度互信息特征模型下所有分类器对主观性强度的分类准确率都比较好，Accuracy 和 *F* 值都高于 90%，特别是 KNN(95.37%)、SVM (93.75%)和 Decision Tree(93.51%)效果比较理想。

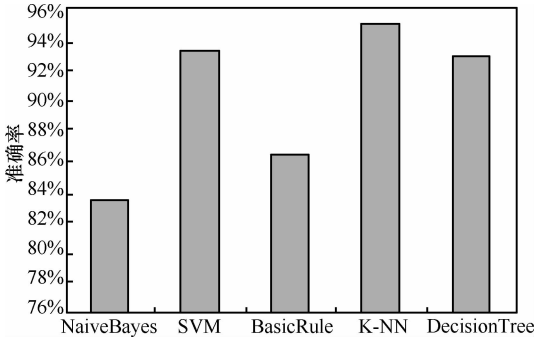

图 2　主观性强度自动判别全部分类器 *F* 值性能曲线

5.3　词语主客观性自动判别

词语主客观全部分类器性能曲线如图 3 所示。

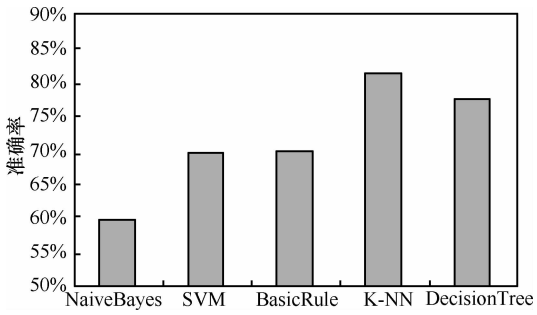

图 3　词语主客观性自动判别全部分类器 *F* 值性能曲线

在特征选择后的主观性特征集合上进行机器学习，获得了主客观性分类器，表现最好 KNN 的 F-measure 为 81.52%，accuracy 为 81.27%，似乎并不是特别理想。Valentin Jijkoun 等使用了德语 WordNet 语义词典[8]，词语的主客观判别平均准确率不到 55%。Banea C 等创建罗马尼亚语的主观性词典[7]*F* 值低于 70%。在本文中主要从语言学角度利用互信息进行特征建模来进行词语主观性的判断，没有利用现有的词典资源，方法的独立性更好，KNN 的 F-measure 为 81.52%，accuracy 为 81.27%，Decision Tree 分类器方法 accuracy 也达到了 77.59，

F 值为 77.65%，在优化过程中，发现 *k*=17 时准确率最高，准确率 82.55%、*F* 值 82.60%，相比同类型的研究[7,8]性能上有了超过 10%的提升。

6　结束语

本文分析和研究了面向情感分析的主观性词典的自动化创建方法，定义了语言学模型，对主观性词典的词条、情感倾向性、主观性强度和主客观的自动获取进行了研究，使用设计的自适应主观性自举方法自动提取出目标语料库中的主观性词汇、构造情感倾向性特征并利用机器学习方法进行情感倾向的自动判别、使用特征选择方法构造主观性强度特征并利用机器学习方法进行主观性强度的自动判别、使用特征选择方法构造词语主观性特征并利用机器学习方法进行词语主客观性的自动判别。实验表明，提出的模型、方法以及设计的算法有效，并取得了较好的性能，效果令人满意。

参考文献：

[1] 夏云庆, 杨莹, 张鹏洲. 基于情感向量空间模型的歌词情感分析[J]. 中文信息学报, 2010,24(1):99-103.
XIA Y Q, YANG Y, ZHANG P Z. Lyric-based song sentiment analysis by sentiment vector space model[J]. Journal of Chinese Information Processing, 2010, 24(1):99-103.

[2] 杨超, 冯时, 王大玲. 基于情感词典扩展技术的网络舆情倾向性分析[J].小型微型计算机系统, 2010,31(4): 691-695.
YANG C, FENG S, WANG D L. Analysis on Web public opinion orientation based on extending sentiment lexicon[J]. Journal of Chinese Computer Systems, 2010,31(4):691-695.

[3] 徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报,2008,22(1):116-122.
XU L H, LIN H F, ZHAO J. Construction and analysis of emotional corpus[J]. Journal of Chinese Information Processing, 2008, 22(1):116- 122.

[4] 宋鸿彦, 刘军, 姚天昉. 汉语意见型主观性文本标注语料库的构建[J]. 计算机应用,2009,23(2):123-128.
SONG H Y, LIU J, YAO T F. Construction of an annotated corpus for Chinese opinioned-subjective texts[J]. Journal of Chinese Information Processing, 2009,23(2):123-128.

[5] KANAYAMA H, NASUKAWA T. Fully automatic lexicon expansion for domain-oriented sentiment analysis[A]. Proceedings of the Conference on Empirical Methods in Natural Language Processing[C]. Sydney, Australia, 2006. 355-363.

[6] KIM S M, HOVY E. Identifying and analyzing judgment opinions[A]. Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics[C]. New York City, USA, 2006. 200-207.

[7] BANEA C, MIHALCEA R, WIEBE J. A bootstrapping method for building subjectivity lexicons for languages with scarce resources[A]. Proceedings of the Learning Resources Evaluation Conference[C]. Marrakech, Morocco ,2008. 2764-2767.

[8] VALENTIA J, KATJA H. Task-based Evaluation Report: Building a Dutch Subjectivity Lexicon[R]. ILPS-ISLA, University of Amsterdam, 2008.

[9] WILSON T A. Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States[D]. Dissertation, School of Arts and Sciences, University of Pittsburgh. 2008.

[10] MICHAEL C. Head-driven statistical models for natural language parsing[J]. Computational Linguistics, 2003, 29(4):589-637.

[11] PETER D T. Thumps up or thumbs down? semantic orientation applied to unsupervised classification of reviews[A].Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C]. Philadelphia, Pennsylvania, 2002. 417-424.

作者简介：

张靖（1972-），男，四川广元人，攀枝花学院副教授，主要研究方向为图像处理算法和自然语言处理。

金浩（1978-），男，江苏南京人，南京大学博士，主要研究方向为自然语言处理、文本挖掘、情感分析。