面向评论信息的跨领域词汇情感倾向判别方法

吴 斐 张玉红 胡学钢
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 词汇的情感倾向判别对文本情感分类具有重要意义。已有方法多假设存在基准词,根据目标词与基准词的关联度来判别目标词的情感倾向。实际应用中,尤其是评论语料库中基准词往往存在情感歧义问题,从而影响判别结果的准确性。基于上述分析,面向给定语料库,提出一种基准词的提取和消歧方法,并在此基础上实现跨领域的词汇情感倾向判别。首先在任一标记语料库中自动提取候选基准词;然后基于共现矩阵评估并过滤部分具有情感歧义的基准词;最后通过计算基准词与目标词的相似性,实现目标词的情感倾向判别。实验结果表明了方法的有效性和可行性。

关键词 基准词,PMI,跨领域,情感倾向,情感歧义
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.046

Approach of Cross-domain Word Sentiment Orientation Identification on Reviews

WU Fei ZHANG Yu-hong HU Xue-gang
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract The sentiment orientation identification for word is important for text sentiment classification. The existing efforts identify the sentiment orientation of target words according to the similarity of the target words and paradigm words with the assumption that there are some paradigm words. However, the ambiguity of emotion of paradigm words in different corpus affects the results of sentiment classification. We proposed a novel method for cross-domain word sentiment orientation identification based on extracting paradigm words and eliminating the ambiguity of paradigm words in the given corpus. More specifically, we first extracted the candidate paradigm words automatically in a labeled corpus, and then filtered the paradigm words those involve in ambiguity of emotion based on the co-occurrence matrix of paradigm words and target words in the target domain. Finally through computing the similarity of paradigm words and target words, we identified the sentiment orientation of target words. The experimental results demonstrate the effectiveness of our method.

Keywords Paradigm word, PMI, Cross-domain, Sentiment orientation, Ambiguity of emotion

1 引言

网络购物、聊天以及社交平台的涌现产生了大量的评论信息,如何自动分析评论文本的情感倾向,已成为文本分类的一个重要研究领域。词汇是语句以及文本的基本组成单位,因此词汇情感倾向的判别也成为文本情感分类的研究重点。

词汇的情感倾向,是指词汇所表达的情感是褒义或者贬义。当前的研究方法多数是根据目标词与人工选取的基准词(已知情感倾向)的关联度来判别词汇的情感倾向[1-9]。其中,基准词的选取对结果具有重要的影响。所谓基准词,是指具有非常明显的褒贬义倾向的代表性词汇[1]。然而实际应用中,评论数据由于受用户表达方式、商品特征等因素影响,导致基准词在不同商品的语料库中存在情感歧义,进而导致其在数量和质量上难以同时满足要求,从而给词汇情感倾向标注带来一定的困难。

情感歧义是指一个词语在两个语料库中的情感倾向不同,如"long"在 kitchen-application 评论中表示厨具的使用寿命长,是褒义词;而在 computer 评论中表示程序运行时间长,是贬义词。显然,基准词的情感歧义问题会导致分类精度下降。

为此,本文提出一种基准词的提取和消歧方法,并实现跨领域词汇情感倾向判别。首先在源语料库中找出具有高情感倾向的基准词;为避免词汇在不同语料库中的情感歧义问题,基于目标语料库中的共现矩阵对候选基准词进行评估筛选;最后根据目标词与基准词间的相似性来判断目标词的情感倾向。

本文的主要贡献:1)基于任意语料库自动提取较多数量的基准词,一定程度上减小了数据共现的稀疏性;2)识别候选基准词在目标语料库是否存在歧义,并剔除具有歧义的基准

到稿日期:2014-06-24 返修日期:2014-10-21 本文受国家高技术研究发展计划(863计划)(2012AA011005),国家自然科学基金(61273292,61305063),安徽省自然科学基金(1208085QF122)资助。

吴 斐(1989-),男,硕士生,主要研究方向为智能计算与理论;张玉红(1979-),女,副教授,主要研究方向为数据挖掘、机器学习,E-mail:zhangyh@hfut.edu.cn;胡学钢(1961-),男,教授,博士生导师,主要研究方向为智能计算,E-mail:jsjxhuxg@hfut.edu.cn。

词，从而提高判别准确率。

2 相关工作

目前，词汇情感倾向判别主要分为两类：一种是借助外部词典，例如 HowNet、WordNet 等[2,3,5]，该类方法所获取的词语的规模非常可观，但也含有较多的歧义词[4]；另一种是通过语法或者词汇在语料库中的共现建立模型[1,6-12]，对目标词汇的情感倾向进行判别，该类方法可分为基于基准词和完全无监督方法两类。

2.1 基于基准词的方法

Turney[6] 利用词汇与正负基准词的 PMI-IR 差来获得词汇的情感倾向，其中 PMI-IR 是依据搜索引擎的返回结果计算的。之后，Turney[7] 提出利用 LSA 方法来计算词汇和基准词的关联度，结果表明 PMI-IR 优于 LSA。王素格[1] 和宋晓雷[8] 首先求解目标词的同义词组，通过计算同义词组与基准词的相关性求解目标词汇的情感倾向，结果表明，与计算单个词汇与基准词相关性方法相比，该方法的分类精度有一定改善。Takamura[9,10] 分别利用“自旋模型”(Spin Model)和“波特模型”(Potts Model)建立词汇的无向图，并基于图思想计算词汇的相关性，得到了不错的效果。

以上方法都需要人工选取一定数量的基准词[1-10]，且并未对基准词的歧义问题进行讨论。

2.2 无监督方法

有一些学者提出了无监督的词汇标注方法。杜伟夫[11] 等人利用词语之间的相似性建立了词汇间的无向图，使用“最小切分”法对无向图进行划分，最后使用模拟退火算法对问题进行求解。实验证明该方法具有较好的可扩展性及鲁棒性。

此外，Yoshida[12] 探讨了同一词汇在不同语料库中的情感歧义问题，将词语的情感倾向作为 LDA 模型的潜在变量，利用 LDA 模型求得词汇的情感倾向，实现了面向多个领域的通用词汇提取。但这类方法往往比较复杂，算法的时间开销较大。

3 本文方法

本文的任务是：对于给定的标签已知的源语料库 $D_S=\{x_{si},y_{si}\}_{i=1}^{N_s}$，从中提取基准词，并对目标语料库 $D_T=\{x_{ti}\}_{i=1}^{N_t}$ 中的词汇进行情感倾向标注。其中，N_s，N_t 分别为源语料库和目标语料库的文档数，$x_{si}=\{w_{s_1},w_{s_2},\cdots,w_{s_n}\}$，$y_{si}\in\{+1,-1\}$，$x_{T_i}=\{w_{T_1},w_{T_2},\cdots,w_{T_n}\}$。

该方法主要分为 3 个步骤：1)从源语料库中提取候选基准词；2)筛选候选基准词，消除情感歧义问题；3)基于基准词对目标词汇进行情感倾向判别。

3.1 提取候选基准词

基准词选取须具有 2 个条件：在语料库中出现；具有明显的情感倾向。为此，首先选取在源和目标语料库中共现的词汇作为基准词候选集，表示为 $CPW=\{w_i\mid x_s\bigcap x_t\}$。

计算这些词汇在源领域的情感倾向 $P_{w_i}^S$，选取其中情感倾向明显的词汇作为候选基准词。本文采用优势率[13] 等指标来度量词汇情感倾向。优势率(Odds Ratio，OR)的计算如式(1)：

$$P_{w_i}^S=\log\frac{P(w_i\mid pos)(1-P(w_i\mid neg))}{P(w_i\mid neg)(1-P(w_i\mid pos))} \tag{1}$$

其中，$P(w_i\mid pos)$ 和 $P(w_i\mid neg)$ 分别表示词语 w_i 在正、负面文档中出现的概率。$P_{w_i}^S\in\{+\infty,-\infty\}$，$P_{w_i}^S>0$ 表示 w_i 的情感倾向为褒义，$P_{w_i}^S<0$ 为贬义，$|P_{w_i}^S|$ 越大表示情感倾向越强，将 CPW 按照 $P_{w_i}^S$ 降序排列，最终选取前 k 个作为候选词。

这样，通过 OR 方法在源语料库中自动选取一定数量的具有较强情感倾向的词作为候选的基准词，降低了人工选取的开销。

3.2 基准词消歧

由于不同语料库的特征不同，导致了同一基准词在不同语料库中可能存在情感歧义，为此，需进行基准词的筛选。通过评估 CPW 中词汇在两个语料库中的情感差异，提出了一种基准词的消歧方法。对候选基准词在目标语料库中情感倾向重新进行评估，若在两个语料库中的情感倾向一致，则保留，否则删除。

基于这样的假设：若基准词 w_i 与正(负)向词的共现度大于与负(正)向词的共现度，则其具有正(负)的语义倾向。

利用候选基准词 CPW 对目标语料库词汇进行评估，构建目标语料库的词汇共现矩阵，基于 Jaccard 相似度来度量目标词汇和 CPW 中正向词和负向词之间的距离，具体定义如下：

$$\tau_{w_i}^T=\frac{\sum\limits_{P_{w_j}^S>0}(w_i\bigcap w_j)}{\sum\limits_{P_{w_j}^S>0}(w_i\bigcup w_j)}-\frac{\sum\limits_{P_{w_j}^S<0}(w_i\bigcap w_j)}{\sum\limits_{P_{w_j}^S<0}(w_i\bigcup w_j)} \tag{2}$$

$\tau_{w_i}^T\geqslant 0$ 表示 w_i 在目标语料库中情感倾向为正，反之为负。将在 D_S 和 D_T 中情感倾向未发生改变的候选基准词作为最终的跨领域基准词，记作：

$$PW=\{w_i\mid P_{w_i}^S*\tau_{w_i}^T>0\} \tag{3}$$

3.3 判别目标词的情感倾向

在得到基准词后，通过计算目标词与基准词的相关性矩阵，来计算目标词的情感倾向。

本文使用优化的点互信息(PMI)在目标语料库中计算目标词与基准词的相关性。考虑到数据稀疏性导致的统计偏差[14]，使用 Pantel 和 Ravichandran[15] 提出的一种基于衰减因子的方法来计算词汇相关性，如式(4)所示。式中前一项是标准 PMI，后一项是衰减因子，其中 $h(w_i,w_j)$ 表示 w_i 和 w_j 在文档中共同出现的次数，$h(w_i)$ 表示 w_i 在文档中出现的次数。

$$R(w_i,w_j)=PMI(w_i,w_j)*\frac{h(w_i,w_j)*\min\{h(w_i),h(w_j)\}}{[h(w_i,w_j)+1]*\{\min\{h(w_i),h(w_j)\}+1\}} \tag{4}$$

通过式(4)得到目标词与基准词的相关性矩阵 R^{T*k}，其中 T 是目标词数。将 U^{T*k} 与基准词的情感倾向 $P_{w_i}^S$ 相乘，得到目标词的情感倾向，$P_{w_T}\geqslant 0$，则目标词 w_T 为正向词，否则为负向词。

$$P_{w_T}^T=R^{T*k}*P_{w_i}^S \tag{5}$$

4 实验设计及结果分析

4.1 实验数据集

本文采用了 1 个模拟数据集和 2 个实际数据集进行实验。模拟数据集采用 Blitzer[16] 的亚马逊产品英文评论数据(DataSet_Bli)，一共有 4 个领域：books(B)、dvd(D)、electronics(E)、kitchen(K)，每个领域有 2000 条评论(正、负各 1000 条)，任选两个作为源语料库和目标语料库。

两个实际数据集为中文数据集，分别是 DataSet_Tan 和

DataSet_Ours。DataSet_Tan是谭松波整理的评论数据集，共有书籍、笔记本电脑和酒店3个领域，每个领域4000条评论（正、负各2000条）。DataSet_Ours是作者从亚马逊中国网站（http://www.amazon.cn/）上爬取的产品评论数据集，共有食品、手机和玩具3个领域，每个领域2000条评论（正、负各1000条）。

4.2 参数设置

基准词个数k对算法结果具有重要的影响，为此对这一参数k进行了实验讨论。图1展示了k取不同值时的实验结果，纵坐标的准确率表示12个跨领域任务的平均分类精度。由图1可见，基准词的数目在[400－900]范围时，分类精度较平稳，最大的精度差异只有3.97%。k过小，则容易导致目标语料库的相关矩阵稀疏；k过大，则基准词的情感倾向不强，易产生随机共现，从而影响分类精度。本文中选定k=700。

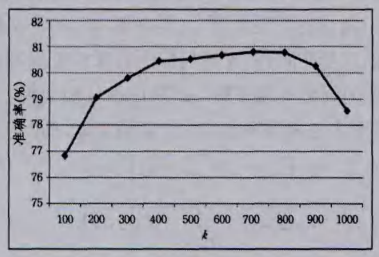

图1 DataSet_Bli数据集上不同k值的分类结果曲线

4.3 基准词消歧对比

基准词的情感歧义筛选是本文的一个主要工作。为了验证该步骤的有效性，对比了候选基准词消歧（OURS_a）和不消歧（OURS_b）两种方法的实验结果，如图2所示。其中，B表示以B为目标语料库，D、E、K分别为源语料库的平均分类精度。

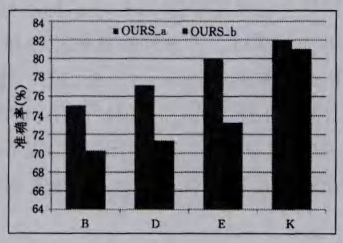

图2 DataSet_Bli数据集上基准词消歧与否的实验结果

由图2可知，进行基准词消歧的方法在分类精度上要明显好于未经消歧的方法，这是由于部分候选基准词在不同的语料库的情感歧义导致的。可见，基准词的消歧有利于分类精度的提高。

4.4 与经典方法对比

选用文献[7]中Turney方法和文献[8]中PLSA方法作为对比算法。人工选取14个情感极性较强的情感词作为文献[7,8]的基准词。对比实验结果如图3所示。

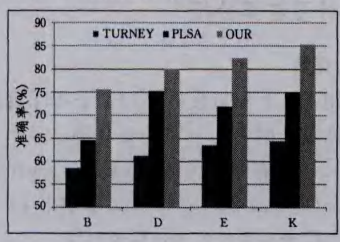

图3 DataSet_Bli数据集上3种方法的分类精度对比

由图3可以看出，本文方法整体上优于另两种方法，另两种方法使用的都是人工选取的通用的基准词集，这不仅使得基准词的数量较少，而且对领域的针对性不强。因此，对特定领域的情感词分类，面向特定领域的基准词效果优于采用通用基准词。

4.5 在实际数据集上的应用

为了进一步验证方法的有效性，在两个实际数据集DataSet_Tan和DataSet_Ours上对算法效果进行了验证，如图4和图5所示。

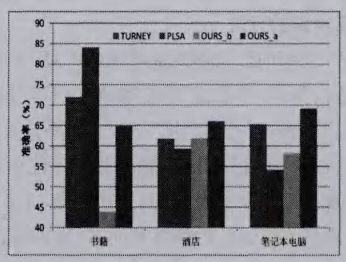

图4 数据集DataSet_Tan上的实验结果

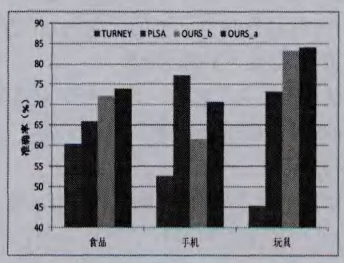

图5 数据集DataSet_Ours上的实验结果

由图4和图5可以看出，经过基准词筛选的方法的效果要明显好于未经过筛选的方法，再次验证了基准词消歧筛选的有效性。此外，从两个数据集的总体上看，本文方法的分类效果优于Turney和PLSA方法，在以笔记本电脑和玩具为目标语料库时，其分类优势较为明显。但同时也注意到，在以书籍和手机为目标语料库时，本文方法的精度要低于基准方法，其主要原因是由于领域分布差异导致自动提取的基准词在目标语料库中出现频率较低，进而导致分类结果不好。如书籍和酒店的领域差异比较大，书籍中提取的基准词在酒店中与目标词的共现较少，从而使计算不准确（书籍→酒店：55.72%，酒店→书籍：63.09%），而PLSA的方法将特征提升为主题层面，一定程度上解决了数据稀疏、共现率较低的问题。

结束语 词汇是语句或文本的基本组成单位，词汇的情感倾向直接影响到对文本的情感判别。本文对于给定的语料库，在自动提取基准词的基础上进行了基准词的消歧，实现了面向领域的词汇情感标注方法。然而，在跨领域分类任务中，数据稀疏一直是一个棘手的问题，这将是我们下一步的研究重点。

参 考 文 献

[1] 王素格，李德玉，魏英杰，等. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报，2009，23(5)：68-74
Wang Su-ge，Li De-yu，Wei Ying-jie，et al. A Synonyms Based Word Sentiment Orientation Discriminating [J]. Journal of Chinese Information Processing，2009，23(5)：68-74

（下转第238页）

framework for combining multiple partitions[J]. Journal of Machine Learning Research,2002(12):583-617

[14] 马继涌,高文. 基于最大交叉熵估计高斯混合模型参数的方法[J]. 软件学报,1999,10(9):974-978
Ma Ji-yong,Gao Wen. An Approach for Estimating Parameters in Gaussian Mixture Model Based on Maximum Cross Entropy [J]. Journal of Software,1999,10(9):974-978

[15] 洪飞,吴志美. 基于小波的 Hurst 指数自适应估计方法[J]. 软件学报,2005,16(9):1685-1689
Hong Fei,Wu Zhi-mei. Adaptive Hurst Index Estimator Based on Wavelet[J]. Journal of Software,2005,16(9):1685-1689

[16] 商琳,王金根,姚望舒,等. 一种基于多进化神经网络的分类方法[J]. 软件学报,2005,16(9):1577-1583
Shang Lin,Wang Jin-gen,Yao Wang-shu,et al. A Classification Approach Based on Evolutionary Neural Networks[J]. Journal of Software,2005,16(9):1577-1583

[17] Dy J G,Brodley C E. Feature subset selection and order identification for unsupervised learning[C]// ICML. 2000:247-254

[18] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph[J]. Journal of mathematical psychology,1975,12(4):387-415

[19] 阳琳赟,周海京,卓晴,等. 基于属性重要性的加权聚类融合[J]. 计算机科学,2009,36(4):243-245
Yang Lin-yun,Zhou Hai-jing,Zhuo Qing,et al. Weighted Cluster Ensemble Based on Significance of Attribute[J]. Computer Science,2009,36(4):243-245

[20] Asur S, Parthasarathy S, Ucar D. An ensemble approach for clustering scale-free graphs[C]// Proc. of ACM KDD. 2006

[21] Li T,Ding C,Jordan M I. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization[C]// Seventh IEEE International Conference on Data Mining, 2007 (ICDM 2007). IEEE,2007:577-582

[22] Topchy A P,Jain A K,Punch W F. A Mixture Model for Clustering Ensembles[C]// SDM. 2004

[23] Berikov V. Weighted ensemble of algorithms for complex data clustering[J]. Pattern Recognition Letters,2014,38:99-106

[24] Minaei-Bidgoli B,Parvin H,Alinejad-Rokny H,et al. Effects of resampling method and adaptation on clustering ensemble efficacy[J]. Artificial Intelligence Review,2014,41(1):27-48

[25] 王红军,李志蜀,成飚,等. 基于隐含变量的聚类集成模型[J]. 软件学报,2009,20(4):825-833
Wang Hong-jun,Li Zhi-shu,Cheng Yang,et al. A Latent Variable Model for Cluster Ensemble[J]. Journal of Software,2009, 20(4):825-833

[26] Zhou Z H,Tang W. Clusterer ensemble. Knowledge-Based Systems,2006,19(1):77-83

(上接第 222 页)

[2] Dai Liu-ling,Liu Bin,Xia Yu-ning. Measuring Semantic Similarity between Words Using HowNet[C]// International Conference on Computer Science and Information Technology. 2008: 601-605

[3] 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,2009,20(1):14-20
Zhu Yan-lan,Min Jin,Zhou Ya-qian,et al. Semantic Orientation Computing Based on HowNet [J]. Journal of Chinese Information Processing,2009,20(1):14-20

[4] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. Journal of Software, 2010,21(8):1834-1848
Zhao Yan-yan, Qin Bing, Liu Ting. Sentiment Analysis [J]. Journal of Software,2010,21(8):1834-1848

[5] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientations of Adjectives[C]// Proc of LREC-04, 4th Int Conf on Language Resources and Evaluation. Lisbon,2004:1115-1118

[6] Turney P D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[C]// Proceedings of the 40th Annual Meeting of the Association Computational Linguistics(ACL). 2002:417-424

[7] Turney P D, Littman, Michael L. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactions on Information Systems,2003,21(4):315-346

[8] 宋晓雷,王素格,李红霞,等. 基于概率浅语义分析的词汇情感倾向判别[J]. 中文信息学报,2011,25(2):89-93
Song Xiao-lei,Wang Su-ge,Li Hong-xia,et al. Word Sentiment Orientation Discrimination Based on PLSA [J]. Journal of Chinese Information Processing,2011,25(2):89-93

[9] Takamura H, Inui T, Okumura M. Extracting Semantic Orientation of Words Using Spin Model[C]// Proc. Ann. Meeting of the Assoc. Computational Linguistics. 2005:133-140

[10] Takamura H,Inui T,Okumura M. Extracting Semantic Orientation of Phrases from Dictionary[C]// Proc. Conf. North Am. Ch. Assoc. for Computational Linguistics. 2007:292-299

[11] 杜伟夫,谭松波,云晓春,等. 一种新的情感词汇语义倾向计算方法[J]. 计算机研究与发展,2009,46(10):1713-1720
Du Wei-fu,Tan Song-bo,Yun Xiao-chun,et al. A New Method to Compute Semantic Orientation [J]. Journal of Computer Research and Development,2009,46(10):1713-1720

[12] Yoshida Y,Hirao T,Iwata T,et al. Transfer Learning for Multiple-Domain Sentiment Analysis-Identifying Domain Dependent/ Independent Word Polarity[C]// Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011: 1286-1291

[13] Mladenic D, Grobelnik M. Feature selection for classification based on text hierarchy[C]// Proceedings of the Conference on Automated Learning and Discovery(CONALD-98). 1998

[14] Bollegala D,Weir D,Carroll J. Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(8):1719-1731

[15] Pantel P, Ravichandran D. Automatically Labeling Semantic Classes[C]// Proc. Conf. North Am. Ch. Assoc. for Computational Linguistics: Human Language Technologies. 2004: 321-328

[16] Blitzer J,Dredze M,Pereira F. Biographies,bollywood,boom-boxes and blenders:Domain adaptation for sentiment classification [C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:432-439