文章编号：1003-0077(2013)06-0075-07

基于双语信息和标签传播算法的中文情感词典构建方法

李寿山[1,2]，李逸薇[2]，黄居仁[2]，苏 艳[1]

（1. 苏州大学 计算机科学与技术学院，江苏 苏州 215006；
2. 香港理工大学 中文及双语学系，香港）

摘 要：文本情感分析是目前自然语言处理领域的一个热点研究问题，具有广泛的实用价值和理论研究意义。情感词典构建则是文本情感分析的一项基础任务，即将词语按照情感倾向分为褒义、中性或者贬义。然而，中文情感词典构建存在两个主要问题：1)许多情感词存在多义、歧义的现象，即一个词语在不同语境中它的语义倾向也不尽相同，这给词语的情感计算带来困难；2)由国内外相关研究现状可知，中文情感字典建设的可用资源相对较少。考虑到英文情感分析研究中存在大量语料和词典，该文借助机器翻译系统，结合双语言资源的约束信息，利用标签传播算法(LP)计算词语的情感信息。在四个领域的实验结果显示我们的方法能获得一个分类精度高、覆盖领域语境的中文情感词典。

关键词：情感分析；双语信息；情感字典；标签传播

中图分类号：TP391 **文献标识码**：A

Construction of Chinese Sentiment Lexicon using Bilingual Information and Label Propagation Algorithm

LI Shoushan[1,2], LEE Sophia Yat Mei[2], HUANG Chu-Ren[2], SU Yan[1]

(1. School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu 215006, China;
2. Department of Chinese and Bilingual Studies, the Hong Kong Polytechnic University, Hong Kong, China)

Abstract: Currently, sentiment analysis has become a hot research topic in the natural language processing (NLP) field as it is highly valuable for many practice usages and theory studies. One basic task in sentiment analysis, named the construction of sentiment lexicon, aims to classify one word into positive, neutral or negative according to its sentimental orientation. However, there are two major challenges: 1) Chinese words are very ambiguities, which makes it hard to compute the sentimental orientation of a word; 2) Given the related research on sentiment analysis, available resource for constructing Chinese sentiment lexicons remains few. Note that there are several corpus and lexicons in English sentiment analysis. In this study, we first use machine translation system with bilingual resources, i.e., English and Chinese information, then get the sentiment orientation of Chinese words by the label propagation algorithm. Experiment results across four domains demonstrate that the lexicon generated with our approach reach an excellent precision and could cover domain information effectively.

Key words: sentiment analysis; bilingual; sentiment lexicon; label propagation algorithm

1 引言

随着互联网的迅猛发展，网络上出现了大量对于人物、事件、产品等进行评论的文本信息。为了处理和分析这些海量的评论信息，情感分析(Sentiment Analysis)正渐渐发展成为自然语言处理中一项越来越受关注的研究课题[1-3]。其中情感词典资

收稿日期：2013-06-15 定稿日期：2013-07-30

基金项目：香港 GRF 项目(543810)；国家自然科学基金资助项目(61003155,61273320)

作者简介：李寿山(1980—)，男，副教授，主要研究方向为自然语言处理，模式识别；李逸薇，女，助理教授，主要研究方向为语言学；黄居仁(1958—)，男，讲座教授，国际计算语言学委员会(ICCL)会士，主要研究方向为自然语言处理，语言学。

源建设是情感分析研究中的一个基础任务。情感词典构建的主要任务为词语情感倾向计算,目标是将词语按照情感倾向分为褒义、中性或者贬义。该任务的研究对于情感分析研究有着非常重要的意义。例如,情感词典的构建能够给大粒度文本(如句子级、篇章级文本)的分类任务提供基础,例如,大量无监督篇章级文本情感分类方法都是基于一些已知情感类别的词语集合,这个词集合一般被称为种子情感词(Seed Words)[4]。此外,进行词语级情感分析对于词语语义理解和消歧具有重要的意义。例如,文献[5]指出情感倾向性可以与词义定义相关联的,并实验验证了情感倾向性计算有助于传统的词义消歧(Word Sense Disambiguation)任务。

然而,中文情感分析研究一直以来缺乏一个涵盖广的语义倾向的情感词典。构建这样的词典非常困难。首先,很多词语在不同语境中它的语义倾向不尽相同。词有多义,歧义的现象,同一观点词在不同领域,不同的语境中甚至会有相反的倾向性表达。例如,词语"圆滑"在句(1)电子领域表现为褒义,而在句(2)的语境中则表现为贬义。其次,传统的词语倾向性分析方法是由已有的电子词典或词语知识库扩展生成词典,但是从国内外的相关研究现状可知,中文情感分析研究起步晚,中文情感字典建设的可用资源相对薄弱。因此,设计高效的中文情感词典构建算法是一个相当具有挑战性的工作。

(1) 该笔记本电脑外形边角处理十分圆滑。

(2) 他变得圆滑,只能选择一再躲避现实。

相对而言,英文的情感分析研究起步较早,已经存在大量相关语料及一些比较成熟的情感词典。本文基于已有的英文资源提出一种新的中文词语情感计算方法,即利用英文种子词典,结合双语言约束的方法构建一个中文情感词典。我们的方法借助机器翻译系统来消除中英文两种语言之间的障碍,把一篇源语言的评论和对应的翻译语言评论作为一整篇文档,通过计算中文词语与英文正向基准词及负向基准词之间的逐点互信息(Point-wise Mutual Information)。在此基础上,我们利用这些互信息值作为词语之间的连接权重,进而构建词语网络,借助标签传播(Label Propagation)算法将英文的情感词极性信息传播到中文的词语上。该词典的构建既利用了英文词典中的情感词极性信息,同时也充分利用了双语言资源对情感词极性在语境环境中的约束信息。在四个领域上的实验表明,我们的方法能获得一个分类精度高,并且能覆盖领域中词语的语境信息的中文情感词典。

本文结构组织如下:第2节介绍情感词典资源构建已有的相关研究工作;第3节介绍我们提出的基于双语信息的情感词典构建方法;第4节给出实验结果及分析;第5节给出结论,并展望下一步工作。

2 相关工作

文本情感分析研究按照所关注的文本粒度大致可以分为三个主要方向:词语级[6-7]、句子级[8]、篇章级[2,9]。其中,词语级的情感分析研究一直备受关注。词语情感计算方法主要分为三类:(1)由已有的电子词典或词语知识库扩展生成倾向词典。在英文中利用WordNet资源[6,10-11],在中文中利用HowNet资源[12]。这种方法的主要思想是:给定一组已知极性的词语集合作为种子,对于一个情感倾向未知的新词,在电子词典中找到与该词语义相近、并且在种子集合中出现的若干个词,根据这几个种子词的极性,对未知词的情感倾向进行推断。但这种方法不能覆盖词语的语境信息;(2)无监督机器学习的方法。这种方法是根据词语在语料库中的同现情况判断其与某种情感类别的联系紧密程度,代表性的文章包括文献[1,13-14]。这些方法中初始的种子情感词对算法的成败起到关键作用。(3)基于人工标注语料库的学习方法。首先对情感倾向分析语料库进行手工标注。标注的级别包括篇章级的标注(即只判断文档的情感倾向性)、短语级标注和句子级标注。在这些语料的基础上,利用词语的共现关系、搭配关系或者语义关系,判断词语的情感倾向性。这种方法需要大量的人工标注语料库,代表性的工作见文献[7,15-17]。

本文提出的方法不同于上面提到的任何一种方法,我们在构建中文情感词典时不需要依赖中文的种子情感词,而是充分利用双语言语料及英文种子词典,这些资源很容易得到。同时,我们利用标签传播算法,结合英文种子词和双语语料统计的方法构建出的情感词典能有较好的领域语境信息。

3 基于标签传播算法的中文情感词典构建方法

3.1 结合英文种子词和双语言约束信息

构建一个中文情感词典,依照传统的词语倾向

性分析方法有两种，一种是由已有的电子词典或词语知识库扩展生成倾向词典，这种方法对中文种子词数量的依赖比较明显。另一种方法是基于非标注语料库的学习方法，该方法也需要大规模中文语料。众所周知，中文情感分析起步晚，可用的中文语料和中文种子词资源有限。相对而言，英文情感分析相关资源比较充分。因此，本文搜集了四个领域的中英文评论语料，结合英文种子词，利用双语信息来帮助构建中文情感词典。具体来讲，我们借助机器翻译系统 Google Translate[①] 把英文评论翻译成中文评论，同时把中文评论翻译成英文评论。因此，每篇评论的内容由源语言文本和对应的翻译文本共同表示，表示同一意思的中文情感词和英文情感词会出现在同一篇文档中。通过计算每个词语的逐点互信息(PMI)，再利用标签传播算法(LP)将种子词的极性在词语间传递，从而得到一个带倾向权值的中文情感词词典。我们提出的方法能充分利用双语言资源的约束信息，整个算法的框架如图 1 所示。

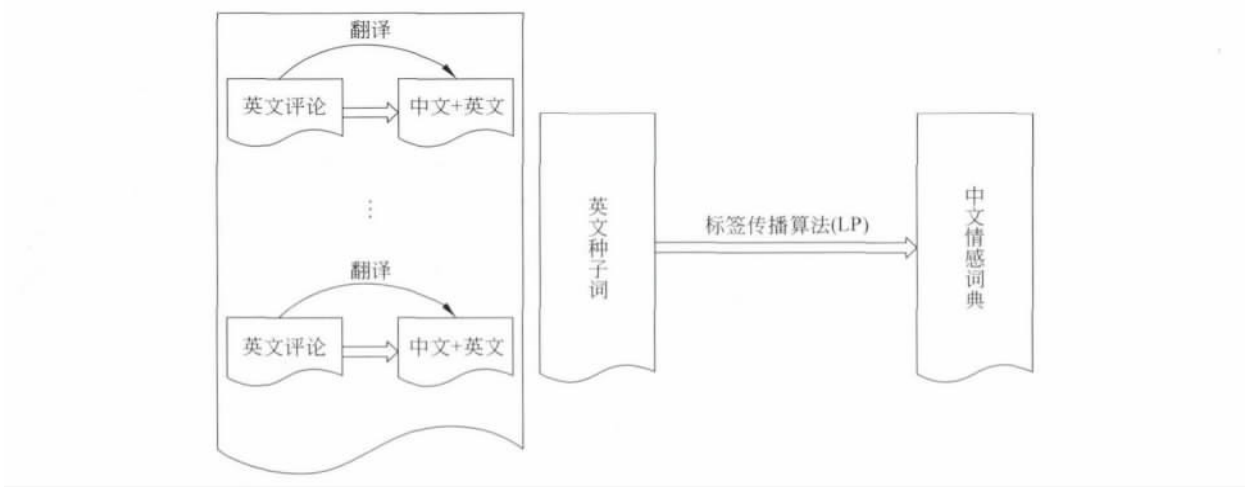

图 1　基于标签传播算法的中文情感词典构建框架图

3.2　结合双语信息构建相似度矩阵

首先，将双语评论中的每一个词语和种子词语都抽象成一个结点；其次，如果词语 A 跟另一个词语 B 共同出现在同一篇文本里，则存在一条有向边从 A 链接到 B，同时存在一条有向边从 B 链接到 A；有向边的权重为词语 A 和 B 的语义相似度的计算值 PMI(A,B)，计算方法如式(1)所示。可以将任意两个词语的相似度计算出来，然后构建一个相似度矩阵 PMI，其中 PMI[i][j]的值表示单词 j 与单词 i 的相似度值。

$$\mathrm{PMI}[i][j]=\log_2\left(\frac{p(w_i\&w_j)}{p(w_i)p(w_j)}\right) \tag{1}$$

3.3　构建相似度概率转移矩阵及 LP 计算

按以上方法，整个双语语料里的词语被抽象为一张有向图。假设一共有 n 个节点，然后可组建一个 n 维的相似度概率转移矩阵。其中 $T[i][j]$的值表示单词 j 与单词 i 的相似度转移概率，计算如式(2)所示。

$$T[i][j]=\mathrm{PMI}(w_i,w_j)\Big/\sum_{j=0}^{j<n}\mathrm{PMI}(w_i,w_j) \tag{2}$$

假设有向图有 6 个结点 A、B、C、D、E、F，其中 A 是褒义种子词，极性为+1；B 是贬义种子词，极性是−1；其余四个词语的极性未知，设为 0。用向量 V 表示结点 A 到 F 的初始情感极性 rank。

$$V=[+1,-1,0,0,0,0]^T \tag{3}$$

如图 2 中节点 A 链向 C、D、F，在这里假设节点间相似度都为 1，所以 A 跳向 C，D，F 的概率各为 1/3。相似度概率转移矩阵如下：

$$T=\begin{bmatrix}0 & 0 & 1/2 & 1/2 & 0 & 1/2\\ 0 & 0 & 0 & 0 & 1/2 & 0\\ 1/3 & 1/2 & 0 & 0 & 1/2 & 0\\ 1/3 & 0 & 0 & 0 & 0 & 1/2\\ 0 & 1/2 & 1/2 & 0 & 0 & 0\\ 1/3 & 0 & 0 & 1/2 & 0 & 0\end{bmatrix} \tag{4}$$

图中结点 A 和 B 是种子词，C、D、F 与种子词 A 直接联系，通过基于 PMI 的算法可以直接求出 C、

① www.google.com

D、F的情感极性。然而从图中我们可以发现，结点E通过C也与种子词A间接联系，因此通过LP算法可以在整个有向图中更精确地计算出E的极性。

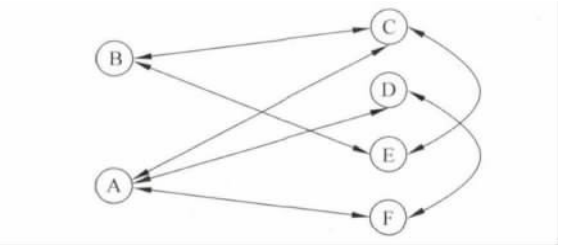

图2 结点间抽象结构图

$$SO[i] = \sum_{i=0}^{i<n} T[j][i] * V[j] \quad (5)$$

其中，$T[j][i]$的值表示结点j到结点i的相似度矩阵转移概率。$V[j]$表示迭代前结点j的初始情感极性，$SO[i]$表示经过迭代计算后的结点i的情感极性。经过一次迭代计算后，每个结点的极性值都会发生改变，得到新的情感极性向量SO。

$$SO=[1,-1,1/2,1/2,-1/2,0]^T \quad (6)$$

在下一次迭代前将种子词的极性恢复初始值，并且将$V=SO$。经过多次迭代计算后，得到情感词典的准确率更高，而且能覆盖领域内上下文语境信息。

3.4 算法流程

基于双语的语义相似度和LP算法的中文情感词典构建算法流程如下：

算法流程：

输入：

英文评论U_{en}，中文评论U_{cn}；

英文种子词典L_{en}；

输出：

一个带倾向权值的中文情感词典L_{cn}；

程序：

1. 双语言评论$U=\varnothing$；

2. 将U_{en}中的每一篇英文评论翻译成中文，组成一篇双语评论并添加到U中；

3. 将U_{cn}中的每一篇中文评论翻译成英文，也组成一篇双语评论添加到U中；

4. 在U的特征向量集上，计算每个词语之间的相似度值矩阵PMI，每一项的值计算方法按照式(1)；

5. 按照式(2)计算出词语之间的相似度概率转移矩阵T；

迭代N次：

6. 按照式(5)在整个有向图中计算得到词语的新的情感极性向量SO；

7. 将SO中种子词的极性恢复初始值，+1或者−1；

8. $V=SO$；

9. 重复步骤6−9，直至循环结束；

10. 最终得到的向量SO即每个词语的情感倾向性，正负号就可以表示词语的极性，而绝对值就代表了该词具有的情感倾向强度；

11. 按照词语情感倾向性的强度排序。

4 实验

4.1 实验设置

本实验中，我们收集了四个领域的中英文语料库，这四个领域分别为：电脑、照相机、化妆品、软件。该语料库的主要来源是亚马逊网站①，每个领域包含正类和负类的评论篇数如表1所示。英文种子词来自文献[18]，包括2 000个褒义的种子词和4 000个贬义的种子词。在进行计算之前，首先我们采用中国科学院计算技术研究所的分词软件ICTCLAS对中文文本进行分词操作。给定分好词的文本后，我们分别对文本的所有中文词语进行了情感倾向计算。

表1 四个领域的中英文正负评论篇数

	电脑	照相机	化妆品	软件
英文正类	1 000	1 000	1 000	1 000
英文负类	1 000	1 000	1 000	1 000
中文正类	1 000	1 000	1 000	1 000
中文负类	1 000	850	1 000	700

4.2 实验结果及分析

4.2.1 LP算法构建中文情感词典的实验结果

表2给出了用LP方法四个领域内打分最高的前100词中分类的精确度，打分最高的前200词中分类的精确度及褒义词实例、贬义词实例。从表中可以看出，打分最高的100词中，电脑领域判断错误的词语有9个，照相机判断错误的有10个，化妆品领域判断错误的仅为8个，软件领域为15个。而四个领域前200个词的分类正确率仍然能达到87.5%、88.5%、87.0%、87%。从实验结果可以看出，四个领域前200个词语的正确率相比基于PMI的方法平均提高了8.6%。这是因为在实验中不仅

① http://www.amazon.cn/

结合了双语的信息，同时根据种子词的极性和相似度概率转移矩阵，用 LP 方法计算出所有词语的情感极性。在此基础上，继续根据词语之间的相似度转移概率在整个有向图中学习词语的情感极性，直至迭代结束，使得构建的中文情感词典精确度更高。

表 2　LP 方法的四个领域内打分最高的 100 及 200 个词中情感分类精确度

	电脑	照相机	化妆品	软件
精确度(100)/%	91.0	90.0	92.0	85.0
精确度(200)/%	87.5	88.5	87.0	87.0
褒义词实例	亮丽、精致	划算、轻巧	豪华、优雅	优、实惠
贬义词实例	郁闷、费劲	气愤、气死	气愤、凌乱	郁闷、棘手

4.2.2　不同情感词典构建方法的比较试验

在本节中，还实现了 PMI 方法构建的情感词典进行比较研究。这些方法包括：

1) **PMI**：直接计算了词语与英文种子词之间的相似度，然后将每个中文词语与正向种子词及负向种子词之间的逐点互信息之和的差值作为该词语的情感倾向，从而得到情感词典；

2) **LP**：计算每个词语的逐点互信息(PMI)，构建相似度概率转移矩阵后，再利用标签传播算法(LP)将种子词的极性在词语间传递，从而得到一个带倾向权值的中文情感词词典。该方法就是本文提出的方法。

图 3 统计出了 PMI 和 LP 方法构建的情感词典中 Top 100、200、300 的正确率。从图可以看出，两种方法构建得到的四个领域情感词典 Top100 正确率都比较理想，电脑和化妆品领域的正确率分别达到了 91%和 92%。前三个领域 LP 方法相比 PMI 方法分别提高 1%、5%、1%，软件领域降低 2%。

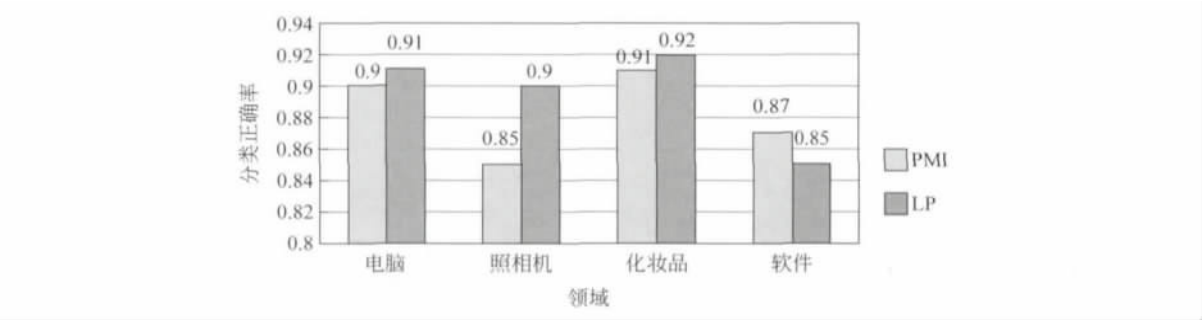

图 3　PMI 和 LP 方法构建的情感词典 Top 100 正确率比较

图 4 和图 5 分别统计出了 PMI 和 LP 方法构建的情感词典中 Top 200、300 的正确率比较。四个领域情感词典 Top200 的正确率上，用 LP 方法明显好于 PMI 方法，平均超过 8.6%。其中，电脑、照相机、化妆品领域分别提高 7%、8%、9%。软件领域甚至提高了 10.5%。在 Top300 的实验结果看，PMI 方法渐渐失去了高精确度的优势，而 LP 方法仍然能获得一个高质量的情感词典。四个领域在前 300 词语极性正确率分别为 85%、85%、83%、83.3%。相比 PMI 方法四个领域平均提高 10.25%。

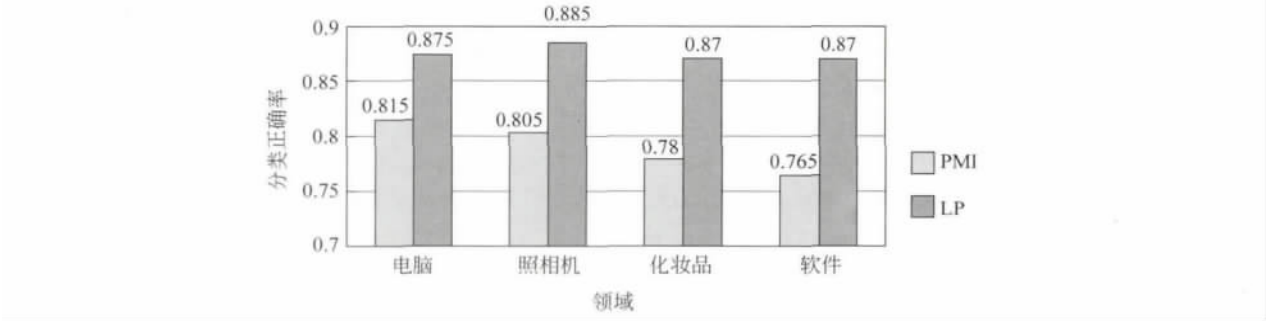

图 4　PMI 和 LP 方法构建的情感词典 Top 200 正确率比较

与 PMI 方法相比，基于 LP 算法构建中文情感词典有更大优势的原因在于：首先，PMI 方法只计算了与种子词直接相连接的词语的相似度来判别自身的情感极性，而 LP 方法可以将种子词的情感极

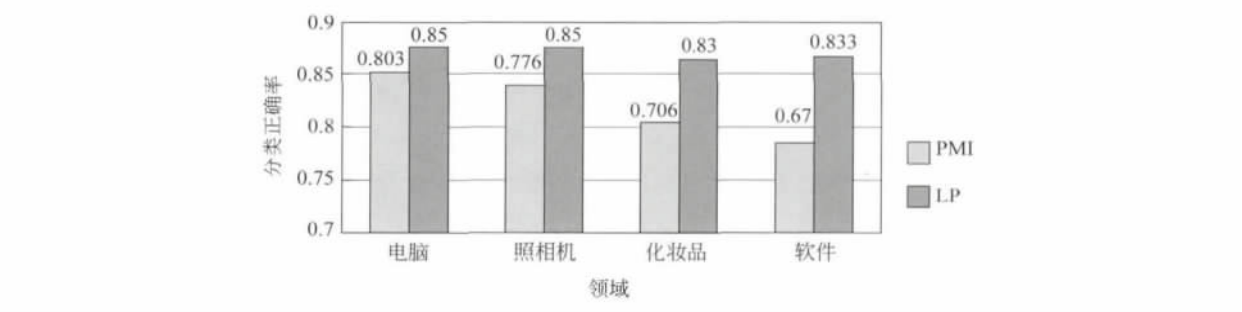

图 5 PMI 和 LP 方法构建的情感词典 Top 300 正确率比较

性沿着边向相邻的结点传播，从而充分利用词语情感极性在所有词语空间的全局上面进行考虑。通过多次迭代，可以提升标签传播的效率及性能。其次，LP 算法中词语的极性不仅受到临近种子词的影响还受到临近非种子词的影响，使得距离近的词语倾向于拥有相同的标签，可对不正确的词语极性及时地重新计算和标注。因此，使用 LP 算法构建得到的中文情感词典质量更高，在正确率和性能上具有更大的优势。

5 结语

本文利用已有的英文情感词典资源及网络中现存的大量中英文评论语料，设计了基于双语信息的情感词典构建方法，通过计算词语间相似度，基于标签传播算法(LP)给出一个带倾向权值的情感词典。实验结果表明我们的方法对于不同领域内情感词的褒贬分类具有较好的分类精确度，获取较多的领域相关的中文情感词。

下一步工作中，我们将考虑已标注语料(有人工打分标签)，来帮助提高领域里面的情感词典计算性能。同时，也将考虑语料中存在大量的情感极性反转的情况，如否定、转折等语言现象。

参考文献

[1] Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[C]//Proceedings of ACL-02, 2002: 417-424.

[2] Pang B, L Lee, S Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques [C]//Proceedings of EMNLP-02, 2002: 79-86.

[3] 宗成庆. 统计自然语言处理[M]. 清华大学出版社: 北京, 2008, 5.

[4] Kennedy A, D Inkpen. Sentiment Classification of Movie Reviews using Contextual Valence Shifters[J]. Computational Intelligence, 2006, 22(2): 110-125.

[5] Wiebe J, R Mihalcea. Word Sense and Subjectivity [C]//Proceeding of ACL-COLING-06, 2006: 1065-1072.

[6] Hatzivassiloglou V, K McKeown. Predicting the Semantic Orientation of Adjectives[C]//Proceedings of ACL-97, 1997: 174-181.

[7] Wiebe J. Learning Subjective Adjectives from Corpora [C]//Proceedings of AAAI-2000, 2000: 735-740.

[8] Pang B, L Lee. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts[C]//Proceedings of ACL-04, 2004: 271-278.

[9] Cui H, V Mittal, M Datar. Comparative Experiments on Sentiment Classification for Online Product Reviews [C]//Proceedings of AAAI-06, 2006: 1265-1270.

[10] Andrea E. Determining the Semantic Orientation of Terms through Gloss Classification[C]//Proceedings of CIKM-05, 2005: 617-624.

[11] Hassan A, D Radev. Identifying Text Polarity Using Random Walks[C]//Proceedings of ACL-10, 2010: 395-403.

[12] 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(01): 14-20.

[13] Hatzivassiloglou V, J M Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity [C]//Proceedings of ACL-2000, 2000: 299-304.

[14] Popescu A, O Etzioni. Extracting Product Features and Opinions from Reviews [C]//Proceedings of HLT/EMNLP, 2005: 339-346.

[15] Akkaya C, J Wiebe, R Mihalcea. Subjectivity Word Sense Disambiguation[C]//Proceeding of EMNLP-09, 2009: 190-199.

[16] 李寿山, 黄居仁. 基于特征提取方法的词语情感倾向计算[C]//第十一届汉语词汇语义学研讨会, 2010.

[17] Lu Y, M Castellanos, U Dayal, et al. Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach [C]//Proceedings of

WWW-11，2011：347-356.

[18] Wilson T，J Wiebe，P Hoffmann. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis [C]//Proceedings of HLT/EMNLP，2005：347-354.

（上接第 15 页）

[7] 李佐丰. 古代汉语语法学[M]. 北京：商务印书馆，2004.

[8] Fei Xia. The Segmentation Guidelines for the Penn Chinese Treebank (3.0)，IRCS Report 00-06[R]，University of Pennsylvania，Oct，2000.

[9] Fei Xia. The Part-of-Speech Guidelines for the Penn Chinese Treebank (3.0)，IRCS Report 00-07[R]，University of Pennsylvania，Oct，2000.

[10] Yan Song，Fei Xia. Using a goodness measurement for domain adaptation：A case study on Chinese word segmentation[C]//Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)，Istanbul，Turkey，2012：3853-3860.

[11] Chunyu Kit，YorickWilks. Unsupervised learning of word boundary with description length gain[C]//Proceedings of CoNLL-99，1999：1-6.

[12] Chunyu Kit. Unsupervised lexical learning as inductive inference via compression[C]//J. W. Minett and W. S. Y. Wang，editors，Language Acquisition，Change and Emergence. Hong Kong：City University of Hong Kong Press，2005：251-296.

[13] Hal Daume III. Frustratingly easy domain adaptation [C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)，2007：256-263.

[14] Hai Zhao，Chunyu Kit. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition [C]//Proceedings of The Sixth SIGHAN Workshop on Chinese Language Processing，Hyderabad，India，2008：106-111.

[15] Hai Zhao，Chunyu Kit. Integrating unsupervised and supervised word segmentation：The role of goodness measures[J]. Information Sciences，2011，181(1)：163-183.

[16] 程湘清.《论衡》双音词研究[C]//程湘清. 两汉汉语研究. 济南：山东教育出版社，1992：262-340.

[17] 方一新. 东汉语料与词汇史研究刍议[J]. 中国语文，1996，(2)：140-144.