文章编号：1003-0077（2011）00-0000-00

一种基于连接关系的中文情感词典构建方法

王科　　夏睿

（南京理工大学 计算机科学与工程学院，江苏 南京 210094）

摘 要：社交媒体及电子商务网站评论的兴起促进了文本情感分析的发展。其中，情感词典构建是文本情感分析的重要内容，然而现有的通用情感词典和词典构建方法，具有领域适用性问题，且不能处理一词多情感。本文提出的方法，利用转折词和否定词对文本极性造成的翻转，将语料中的情感词进行极性分类。实验结果显示，我们的方法能够有效构造领域特定的中文情感词典，与现有的通用情感词典和常见的情感词典构建方法相比，本文方法在篇章级、属性级文本情感分析上表现出了更好的性能。

关键词：情感分析；情感词典；连接关系

中图分类号：TP391　　**文献标识码：**A

An Approach to Chinese Sentiment Lexicon Construction Based on Conjunction Relation

Ke Wang　　Rui Xia

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094 Jiangsu)

Abstract: The rise of Internet reviews in social media and e-commerce sites promotes the development of text sentiment analysis.Sentiment lexicon is an important part of sentiment analysis. However, the existing common sentiment lexicons and construction methods have applicability problems in the area, and they still can't handle the words with one more sentiment. This paper presents a construction method that utilize transition words and negative words to classify the sentiment words in corpus, as these words can reverse the polarity of text. Experimental results show that our method can effectively construct domain specific Chinese sentiment lexicon, and has a better performance compared to existing common sentiment lexicon and construction methods in the chapter level, attribute-level text sentiment analysis.

Key words: sentiment analysis; sentiment lexicon; conjunction relation

1.引言

随着互联网的迅速发展，尤其是以微博、论坛、电子商务网站为代表的网络平台的发展，越来越多的用户倾向于在网上发表自己对产品或热门事件的评论。这些评论信息通常包含用户的主观情感，整理并分析这些信息，有助于提高各方整体效率，作出最合理的决策。评论数据的激增，导致人工整理、分析这些海量数据异常困难，文本情感分析技术应运而生。基于情感词典的方法是文本情感分析的一种有效途径，其构建方法也逐渐成为研究热点[1-3]。

情感词典构建的主要任务是识别情感词并确定情感词的褒贬倾向。然而，中文缺少类似 WordNet 的大型语言词典，情感词典构建的可用资源也相对较少，现有的通用情感词典和构建方法存在诸多问题。首先，单词的情感极性受诸多因素影响，同一个词的不同释义可能具有相反的情感极性，比如“*我为你骄傲*”和“*你太骄傲了*”中的两个“*骄傲*”，前者带有褒义情感，而后者带有明显的贬义色彩，通用词典 NTUSD 将所有“*骄傲*”都视为贬义，若用它来判断相关评论，会造成很多错误。另外，同一单词同一释义在描述不同的商品属性时，也会具有不同的极性，比如“*手表走时偏快*”和“*物流速度快*”，前者为褒义评价，后者则是贬义的。其次，评论的情感还与转折词、否定词有关，两者都是造成情感极性转变的关键因素，而现有的情感词典构建方法，多数只考虑了词间的并列关系而并未考虑转折关系，从而造成单词情感极性被错误判断。本文提出的情感判断方法，依据连接关系对情感极性的影响，提出使用转折词、否定词以及

收稿日期：　　**定稿日期：**

基金项目：国家自然科学基金（61305090）；江苏省自然科学基金（BK2012396）

并列连词来判断语料中所有形容词的情感倾向，并用情感表达明确的词集，校正判断结果。在两组语料七个领域上的实验结果表明，我们的方法能获得较好的结果。

本文组织结构如下：第 2 节介绍情感词典构建已有的相关工作；第 3 节介绍了我们提出的基于连接关系的情感词典构建方法；第 4 节给出实验结果并进行分析；第 5 节给出结论，并展望下一步研究工作。

2. 相关工作

目前已有较多情感词典构建方法研究见诸报道。文本情感词典构建方法主要可以分为包括三种：基于词典的方法，基于语料的方法以及词典与语料结合的方法。

基于词典的方法主要通过构造少量已知极性的单词集，利用语言词典（如英文的 WordNet[1]）中的语义关系来扩大种子词。Hu & Liu[4]先构建少量褒义、贬义形容词，作为种子词；然后在 WordNet 中查找他们的同义词与反义词来扩大这两个集合，将同义词放入种子词所在集合，反义词放入另一集合；通过循环迭代，构建一个具有一定规模的形容词词库。Kamp[5]等认为，意思越是相似的单词，他们通过同义词迭代所需的次数就越少，于是使用两个单词相互迭代所需的次数来表示这两个词的情感相似程度，从而确定未知单词的情感极性。Baccianella[6]等还使用同义词集的释义而非同义词本身作为训练集，构造一个三分类器(褒义，贬义，客观)来判断其它未知情感的释义。柳位平等[7]以 HowNet[2]为基准，先挑选出一部分常用的情感词构成基础情感词集，然后计算词语的相似度，得到其它单词的情感倾向值，最终构成一个基础情感词典。这类方法很依赖词典，能较为快速地构造一份通用情感词典，但无法覆盖语料中所有情感词。

基于语料的方法，能总结语料中的情感词并判断其情感极性，构建出具有领域适用性的情感词典。该方法通常利用统计学知识，挖掘语料中未知情感词与已知情感词的关系，形成领域情感词典。Turney 等[8]通过计算语料中其它词与正负情感种子词之间的点互信息[9](Pointwise Mutual Information，PMI)，来判断单词的极性。PMI 能衡量词与种子词间的共现程度，用每个词分别与各个正向种子词、负向种子词的 PMI 值累加并作差，能得到单词的情感倾向(Sentiment Orientation, SO)，值越大，则该词为褒义词的概率越大。阳爱民等[10]借用 Turney 的思想，利用种子词与其它词的百度搜索返回结果，计算单词的 SO-PMI，来判断单词的情感极性。除了统计方法外，基于语料的方法还使用语句中的连词来判断两句话中情感词的极性关系。Hatzivassiloglou & McKeow[11]首先提取出连词连接的形容词，选取部分词标注极性，作为种子词；再结合线性模型来确定连起来的两个词具有相同还是相反的极性；然后用聚类算法产生两个单词集并调整，最后形成情感词典。

目前，很多构建方法都把词典和语料结合起来。两者都提供了词与词之间的关系，词典主要提供单词间标准的语义关系(同义、反义、上位、下位关系等)，而语料则主要提供两个单词在文本中的关系，包括位置(单词距离)、情感关系(极性相同还是相反)等。李寿山等[12]利用英文种子词典，借助机器翻译，把原评论和对应的翻译评论作为一篇文档，计算其它词与种子词的 PMI；然后利用词之间的 PMI 值，构建连接矩阵，借助标签传播算法将英文的情感词极性传播到中文词上，克服中文在现有资源上的一些劣势，构建情感词典。Xu 等[13]利用人民日报 1997~2004、哈工大同义词词林扩展版、现代汉语词典以及拥有公共字的词相似度比较大的思想，构建四个相似度矩阵；接着利用选出的一部分种子词并做标注，迭代推导未知词的情感极性，并人工纠正迭代过程中产生的错误。

本文提出的情感词典构建方法，是利用大量语料中的连接关系，依据转折词和否定词改变文本情感极性的思想，将单词划分成两个集合并确定情感极性。对部分有歧义的情感词，我们将其与所描述的对象相结合。该方法无需标注数据，流程简单，并且具有良好的效果。

3. 基于连接关系的情感词典构建方法

3.1 情感连接关系分析

情感分析的主要对象是主观文本，而评论本身就是用户的主观表达。本文认为，情感词大部分是形容词，在评论中形容词起修饰性作用。正是这些修饰性的词，表达了用户的情感色彩。因此，本文将语料中

[1] http://wordnet.princeton.edu/wordnet/download/
[2] http://www.keenage.com/html/e_index.html

所有形容词视为情感词。同时，由于情感受多方面的影响，故本文方法做出以下分析：

1) 情感极性反转

一句话中情感表达的转变，通常是由转折词和否定词造成的。否定词的作用域是当前分句，即只改变当前分句中否定词后的情感词的倾向，而不影响下一句。分句是指两个标点符号中的句子，分句中不含有其它标点符号。“*宝贝一般，店家不怎么爽快*”，否定词“*不*”只改变当前分句“*店家不怎么爽快*”中“*爽快*”的情感倾向，使其情感与上一句中的“*一般*”相反。

转折词有别于否定词，它会改变评论中，转折词之后所有分句的情感倾向，直到再次遇到转折词。比如：“*宝贝不错，就是物流有点慢，等了好久！*”，转折词“*就是*”改变了转折词之后，整体的情感倾向，使得转折词之后的分句中的形容词“*慢*”、“*久*”与转折词之前的形容词“不错”的情感倾向相反。

2) 情感极性保持

分句间无连词，或存在并列连词“*和*”、“*而且*”等，不会影响文本情感极性。“*卖家服务很好，发货快，而且宝贝也很棒！*”，“好”和“快”间没有连词，与下一句中情感词“*棒*”用并列连词“*而且*”连接，情感都不发生改变。故“*好*”、“*快*”、“*棒*”三者具有相同的情感倾向。

3) 情感极性虚拟

虚拟语气中包含的情感词不是对客观事实的描述，不做考虑。其中，虚拟词是虚拟语气的一个明显标志，虚拟句中包含的情感词，并不是对商品的真正描述。比如“*物流太慢了，但愿东西别让我失望。*”，“*失望*”并不是对商品真正的描述。所以我们并不考虑虚拟句中包含的形容词。

4) 情感极性多样

Ding[14]指出，许多单词在同一个领域的不同文本中，可能会有不同的情感指向。如相机评论中，单词“*长*”在描述电池续航时间和聚焦时间上的情感倾向是相反的，我们称这类词为情感极性多样词。当遇到这类词时，我们把情感词所修饰的属性词一起写入词典，如“*续航时间 长*”、“*聚焦时间 长*”等。

3.2 算法示例

情感极性的转移通常会伴随着转折词的出现，而普通的连词不改变情感极性。否定词可以改变分句中否定词后单词的情感极性。举个例子：

“**(1)***总体/NN 不错/VA ,/PU* **(2)***就是/AD 有点/AD 贵/VA ,/PU* **(3)***而且/AD 物流/NN 不/AD 是/VC 很/AD 快/VA ,/PU* **(4)***不过/AD 还是/AD 很/AD 满意/VA 的/DEC 一次/CD 网购/NN ,/PU* **(5)***希望/VV 用/VV 久/VA 一点/AD 。/PU*”

形容词共有“*不错*”、“*贵*”、“*快*”、“*满意*”、“*久*”这 5 个，其中褒义种子词为“*不错*”、“*满意*”，贬义种子词为“*贵*”，待确定情感词为“*快*”、“*久*”。由于**(2)**中转折词“*就是*”的出现，使得**(1)**和**(2)**后的句子，整体情感倾向发生了改变，且**(2)**中不存在否定词，所以“*不错*”和“*贵*”具有相反的情感极性。**(3)**中连词为“*而且*”，故整体极性和**(2)**保持相同，但**(3)**中有否定词“*不*”，使得“*快*”和“*贵*”具有相反的极性。**(4)**中有转折词“*不过*”，故**(4)**的整体情感极性和**(2)**、**(3)**相反，**(4)**中没有否定词，而**(3)**中有否定词，所以“*满意*”和**(3)**中的“*快*”具有相同情感极性，与**(2)**中的“*贵*”具有相反的情感极性。由于**(5)**为虚拟语气，故不考虑。最后得到两个单词集：[“*不错*”，“*快*”，“*满意*”]和[“*贵*”]。最后利用种子词集便可以判断情感极性，前者包含 2 个褒义种子词，故其中所有单词均为褒义词，所以可以确定“*快*”是褒义的，且“*快*”属于情感极性多样词，查找该分句前后的名词，确定“*物流*”为属性词，故“*物流 快*”被确定为褒义情感词。

3.3 预设词典

本文提出的构建方法，将现有词典资源和语料结合使用，其中词典资源主要包括：否定词词典、转折词词典、虚拟词词典、情感极性多样词词典、标准情感词典。

标准情感词典是将通用词典——HowNet 情感倾向词集和 NTUSD[3]合并去重，并筛除情感极性多样词和其它歧义词，使得情感表达明确。情感表达明确要求只表达一种情感，如“*漂亮*”、“*优秀*”只表达褒

[3] http://www.datatang.com/data/44317

义；而“*垃圾*”除表示中性外，还有贬义成分，“*骄傲*”同时具有褒义和贬义成分，因此，这类词不被用作标准词。标准词典用于对可能存在的因语法不规范导致情感词错分的情况，进行强制纠正，提高当前判断的准确率，以降低对后续判断的影响。

我们对评论中常见的否定词、转折词、虚拟词、情感多样词做了总结归纳，如表 1 所示。

表 1. 预设词典

否定词	不 没 没有 不如 不及 没什么 不是 非 应该 不会 不至于
转折词	就是 但 但是 只是 不过 美中不足 可惜 遗憾 虽然 其它 只怪 却 然而
虚拟词	要是 希望 但愿 如果 换成 假如 原来以为 除非 若是 若 即使 倘
情感多样词	大 小 高 低 多 少 强 弱 冷 热 快 慢

3.4 算法流程

由于本文方法主要依据连词判断，因此我们将其命名为连接关系法(Conjunction Relation Method，CRM)，算法框架如图 1 所示。在对所有评论进行分词、词性标注后，统计形容词词频，取高频词的前 5%~10%，标注其情感极性，构成种子词典。将评论按照标点进行断句，得到多个分句后，按照本文极性转移规则将形容词划分为两个集合，用种子词集确定集合中单词的褒贬性，并用标准情感词典检查结合中是否有错分的词，若有，则纠正。最后将结果添加到种子词中，并更新单词被判断为该极性的频数。

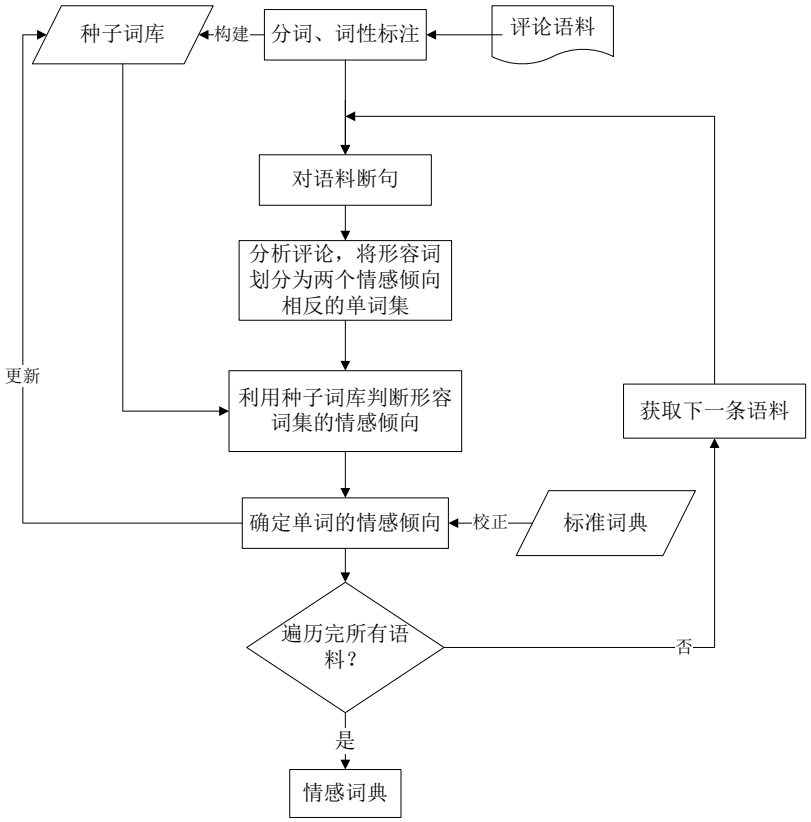

图 1. 基于连词的情感词典构建方法框架

我们总结了 CRM 的核心过程——极性转移规则进行了总结。设当前情感极性为 P，初始为 1，若出现转折词，则取负，如公式(1)所示。当 P=1 时，说明此前极性未发生翻转或经过偶数次翻转；若 P=-1，说明此前极性发生奇数次翻转。

$$P=\begin{cases}1, & \text{初始值}\\ -P, & \text{转折词}\end{cases} \qquad (1)$$

两个列表 L_1 和 L_2，分别存放两组形容词，组内单词情感极性视为一致，组间情感极性视为相反，每组的情感倾向未知，其中 L_1 中存放与初始情感极性相同的词，L_2 中存放与初始情感极性相反的词。若该句第一个词不是虚拟词，查找并定位一个形容词，若该形容词为情感极性多样词，则将属性词与形容词结合，视为情感词；判断当前分句中形容词前是否有否定词，以及该分句第一个词是否是转折词或虚拟词，根据结果判断，按照极性转移规则将形容词归入相应列表中。极性转移规则如下，流程图如图 2 所示。

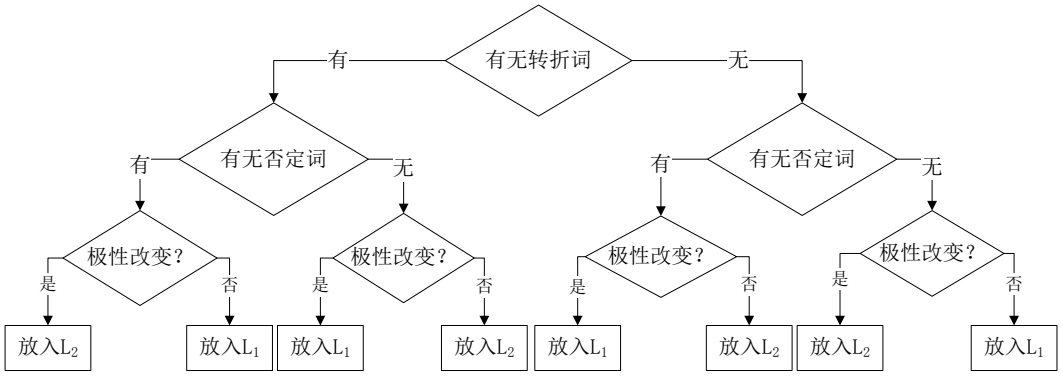

图 2. 极性转移规则

设当前评论分句为 R={s_1,s_2,s_3,...,.s_i}， s_m 为 R 中情感词，$m \in [1, i]$:

1) 若 s_1 为转折词，且任意 s_k 不是否定词，$k \in [2, m-1]$，

$$s_m \ belong \ to \begin{cases} L2, & \text{P=1} \\ L1, & \text{P=}-1 \end{cases}$$

同时，P = -P；

2) 若 s_1 为转折词，且存在 s_k 是否定词，$k \in [2, m-1]$，

$$s_m \ belong \ to \begin{cases} L1, & \text{P=1} \\ L2, & \text{P=}-1 \end{cases}$$

同时，P = -P；

3) 若 s_1 不是转折词，且存在 s_k 是否定词，$k \in [1, m-1]$，

$$s_m \ \ belong \ \ to \begin{cases} L2, & \text{P=1} \\ L1, & \text{P=}-1 \end{cases}$$

4) 若 s_1 不是转折词，且任意 s_k 不是否定词，$k \in [1, m-1]$，

$$s_m \ \ belong \ \ to \begin{cases} L1, & \text{P=1} \\ L2, & \text{P=}-1 \end{cases}$$

按照上述流程，将语料中的形容词分别放入 L_1 和 L_2 中。

根据极性转移规则，我们对每条评论进行分析，将其中的形容词分别放入 L_1 和 L_2 中，但 L_1 和 L_2 的极性未知。我们需要使用种子词集来确定它们的极性：统计两个词集中褒义种子词和贬义种子词的数量，若 L_1 中褒义词占多数，则 L_1 中所有单词视为褒义词；反之 L_2 中所有词视为褒义词。由于每个人的表达能力和习惯有所差别，在评论过程中难免会存在不规范的情况，如：“*宝贝不错，客服反应迟钝。*”，此时我们得到的结果是“*不错*”和“*迟钝*”具有相同的情感倾向，这显然是错误的。CRM 在构造情感词典过程中，加入

了纠错机制，利用标准词典，对错分的结果进行强制纠正。纠正方法是：在标准词典中查找评论包含的情感词，并检查其极性结果是否与标准词典一致，若不一致，则将其纠正。之后，语料中各个词被划分为褒义、贬义的频数。遍历完所以语料后，针对结果中可能存在部分情感词同时出现在褒义词集和贬义词集中，

我们对贬义词词频加权后与褒义词词频比较，因为大多评论中，好的方面的评论要比坏的方面多[15]，导致分类过程中，结果会偏向褒义，所以需要对贬义词频加权。最后以拥有词频高的情感极性确定为单词的最终极性。

4. 实验

4.1 实验设置

在实验过程中，我们使用了两组语料：中文情感挖掘语料-ChnSentiCorp[4]与 COAE2008[5]任务 3 的语料。

ChnSentiCorp 包含酒店、电脑(笔记本)、书籍三个领域的评论，每个领域包含正类和负类评论各 2000 篇，我们使用线性支持向量机(LibSVM[6])，根据其默认参数，对语料进行监督文本情感分类[7]。此外，利用 COAE2008 任务 3 的语料，我们进行了属性级情感文本分析，该语料包含数码相机、汽车、笔记本、手机四个领域的商品评论信息，每个领域的评论数如表 2 所示。在进行实验之前，我们首先采用东北大学的 NiuParser[7]对中文评论语料进行分词和词性标注操作。

表 2. COAE 2008 任务三语料规模

产品	相机	笔记本	汽车	手机
评论数量	137	56	157	123

4.2 实验结果

表 3 给出了使用 CRM 方法，在三个领域上的实验结果，包含情感词典大小、前 100 个高频词的分类精确度[12]，我们得到的褒义词有“干净”、“漂亮”、“生动”等；贬义词有“陈旧”、“郁闷”、“粗糙”等。同时，由于我们对文中的情感极性多样的词进行了特殊处理，于是在褒义词词典中还得到如：“性价比 高”，“ 实用性 强”，“声音 小”这样的<属性，情感词>模式，在贬义词词典中得到“价格 高”，“ 内存 小”，“ 清晰度 低”等。

表 3. CRM 在 ChnSentiCorp 上实验结果一览

	酒店	电脑	书籍
褒义词总数	451	337	542
贬义词总数	232	97	293
褒义词精确度	90.0%	91.0%	94.0%
贬义词精确度	94.0%	94.4%	98.0%

从表中我们可以看出，对贬义词的分类准确率通常要高于褒义词。三个领域上，两者的精度分别相差 4%、3.4%、4%。原因在于，贬义词的出现，通常会有明显的转折，且我们对贬义词词频进行了加权，降低了单词的错误分类对最终结果产生的影响。

4.3 与其它情感词典性能比较

4.3.1 篇章级情感分析结果

我们将情感词典运用于篇章级情感分类，从而对情感词典性能进行评测。表 4 是基于 ChnSentiCorp 语料，分别利用 CRM 生成的情感词典、HowNet 情感倾向词库(以下简称 HowNet)、NTUSD 以及大连理工情感词汇本体库(DLUT)作分类特征，进行监督文本分类，得到的准确率。由于本文阐述的是情感词典构造方法，所以在文本级情感分类任务的性能比较上，仅和其它词典进行平行比较，而未与特征选择方法进行比较。从表 4 可以看出，使用 CRM，在三个领域上分类的准确率，比 HowNet 分别有 9.4%、13.33%、4.91%

[4] http://www.searchforum.org.cn/tansongbo/corpus-senti.htm
[5] http://www.ir-china.org.cn/coae2008.html
[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html
[7] http://www.niuparser.com/

的提升，比 NTUSD 分别有 0.39%、0.55%、2.43%的提升，比 DLUT 分别有 3.56%、4.08%、7.7%的提升。原因在于本文构造的情感词典是针对语料中的情感词构造领域词典，因此应用在本领域中，情感词的覆盖率比通用情感词高，当用作文本分类时，CRM 方法能提供更多的特征，使得文本能够更好地被表达。

表 4. 篇章级情感分类准确率

词典	酒店	电脑	书籍
CRM	**74.28%**	**76.27%**	**78.52%**
HowNet	64.88%	62.94%	73.61%
NTUSD	73.89%	75.72%	76.09%
DLUT	70.72%	72.19%	70.82%

4.3.2 属性级情感分析结果

相比篇章级的文本情感分析，属性级的情感分析不仅能衡量词典中情感词在语料中的召回率，还能衡量情感词判断的精确率。针对同一属性集，我们以修饰属性词的形容词为该属性的情感词，通过对情感词的褒贬分析完成对属性的情感分析。COAE2008 任务 3 是一个属性级的情感分析任务，我们在该任务上，基于相同的特征抽取方法，将 CRM 产生的情感词典分别与 HowNet、NTUSD 以及 DLUT 进行了对比。在四个单独领域以及全部语料上，属性级情感分析的精确率、召回率、F_1 值的结果如表 5 所示。从表 5 中可以看出，CRM 相比其它情感词典，在各个领域上，各项性能都有明显提升。原因除了 CRM 能根据语料构造领域情感词典，覆盖语料中的大部分情感词，使得召回率大大提升外，CRM 还对情感极性多样的词进行了单独分析，解决了同一情感词在描述不同的特征时可能会具有不同的情感的问题，而现有的通用情感词典则不具有该能力。

此外，标准词典的加入，使得情感词的分类结果更加精确。在 COAE2008 任务 3 的语料上，我们对加入校正前后，属性级情感分析性能作了对比，从表 5 可以看出，使用标准词典校正后，属性级情感分析的性能有了较大提高。

表 5. COAE2008 任务 3 评测对比

语料	评测指标	HowNet	NTUSD	DLUT	SO-PMI	CRM（未校正）	CRM
相机	Precesion	0.2460	0.2452	0.1198	0.2548	0.1501	**0.2931**
	Recall	0.1556	0.1551	0.0758	0.1526	0.1572	**0.1854**
	F_1-score	0.1907	0.1900	0.0928	0.1909	0.1535	**0.2271**
汽车	Precesion	0.1193	0.1193	0.0632	0.1288	0.0907	**0.1490**
	Recall	0.0674	0.0674	0.0357	0.0718	0.0792	**0.0841**
	F_1-score	0.0861	0.0861	0.0457	0.0922	0.0845	**0.1075**
手机	Precesion	0.2958	0.2934	0.1408	0.2851	0.1660	**0.3482**
	Recall	0.1565	0.1552	0.0745	0.1411	0.1300	**0.1841**
	F_1-score	0.2047	0.2030	0.0975	0.1888	0.1458	**0.2409**
笔记本电脑	Precesion	0.2198	0.2179	0.1282	0.2203	0.1523	**0.2784**
	Recall	0.1159	0.1150	0.0676	0.1217	0.1179	**0.1469**
	F_1-score	0.1518	0.1505	0.08855	0.1568	0.1329	**0.1923**
全部	Precesion	0.2041	0.2018	0.1021	0.2064	0.1291	**0.2424**
	Recall	0.1300	0.1285	0.0650	0.1268	0.1262	**0.1555**
	F_1-score	0.1588	0.1570	0.0795	0.1571	0.1276	**0.1898**

4.3.3 不同情感词典构建方法比较

现有的情感词构建方法有很多，本节中我们将对比 CRM 方法与 SO-PMI 方法生成的情感词典在属性

级情感分析任务上的性能。我们把 SO 值，按照正数、零、负数把形容词分为褒义词、中性词、贬义词，从而构建情感词典。

表 5 给出了 SO-PMI 和 CMR 方法构成的情感词典在 COAE2008 语料上的评测结果。可以看出，CRM 方法在四个领域上的结果都优于 SO-PMI 方法得到的情感词典。分析 PMI 方法发现，其计算过程中并未考虑转折词和否定词产生的影响，比如：“*宝贝 不错 ， 就是 贵*”，若经常出现这样的评论，则得到的结果是“*贵*”和“*不错*”具有相同的情感极性。若以转折词切分句子，则分句中同时出现两个情感词的概率较低，影响实验结果。而 CRM 方法针对转折词和否定词的影响，有效地将评论中的情感词进行了分类。

5. 结论与展望

本文利用连接关系对文本情感产生的影响，主要考虑了否定词和转折词，来确定其中形容词的情感倾向，构造情感词典。实验表明，本文的方法相比现有的通用情感词典，在篇章级和属性级的文本情感分析任务上具有更优的性能，并且该方法能适应不同领域的情感分析任务。

由于情感词不仅有形容词，还有名词、动词，甚至部分副词，且文中不是所有的形容词都具有情感，在今后的工作中，我们将扩展情感词至其它词性，结合更加精确的情感词发现任务，构造一份较为完备的情感词典。另外，我们也将完善词典校正机制，以减少因错分而引起的对后续情感词极性判断的影响。

参考文献

[1] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques[C]// IN PROCEEDINGS OF EMNLP2002:79--86.

[2] XIA R, ZONG C, LI S. Ensemble of feature sets and classification algorithms for sentiment classification [J]. Information Sciences, 2011, 181(6): 1138-1152.

[3] LIU B. Sentiment analysis and opinion mining [J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1-167.

[4] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.

[5] KAMPS J, MARX M, MOKKEN R J, et al. Using wordnet to measure semantic orientations of adjectives [J]. 2004,

[6] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]//LREC. 2010, 10: 2200-2204.

[7] 柳位平, 朱艳辉, 栗春亮, et al. 中文基础情感词词典构建方法研究 [J]. 计算机应用, 2009, 29(10): 2875-2877.

[8] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 417-424.

[9] CHURCH K W, HANKS P. Word association norms, mutual information, and lexicography [J]. Computational linguistics, 1990, 16(1): 22-29.

[10] 阳爱民, 林江豪, 周咏梅. 中文文本情感词典构建方法 [J]. 计算机科学与探索, 2013, 7(11): 1033-1039.

[11] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives[C]//Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics. Association for Computational Linguistics, 1997: 174-181.

[12] 李寿山, 李逸薇, 黄居仁, et al. 基于双语信息和标签传播算法的中文情感词典构建方法 [J]. 中文信息学报, 2013, 27(6): 75-81.

[13] Xu G, Meng X, Wang H. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010: 1209-1217.

[14] Ding X, Liu B, Yu P S. A holistic lexicon-based approach to opinion mining[C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008: 231-240.

[15] TABOADA M, BROOKE J, TOFILOSKI M, et al. Lexicon-based methods for sentiment analysis [J]. Computational linguistics, 2011, 37(2): 267-307.

作者简介：

王科（1990——），男，硕士研究生，主要研究领域为自然语言处理，情感分析。
Email:wangkk998@gmail.com；

夏睿（1981——），男，2011 年 6 月博士毕业于中科院自动化研究所，现为南京理工大学计算机学院副教授、硕士研究生导师。主要研究领域为自然语言处理、机器学习、情感分析与观点挖掘。本文通讯作者。
Email: rxia@njust.edu.cn。