基于语料资源的中文情感词典扩展方法 *

谢松县，刘博，王挺

（国防科学技术大学 计算机学院，湖南 长沙 410073 ）

摘要：设计实现了基于语言特征和统计特征的中文情感词典扩展算法，提出了基于混合特征的方法。首先，通过选择具有并列、递进以及转折关系的连接词实现了基于语言特征的算法，然后设计了上下文向量统计特征并实现了基于上下文向量的情感词极性计算算法，最后针对两种方法的特点提出基于混合特征的情感词极性计算方法。实验结果表明，基于混合特征的情感词极性计算方法达到了性能最优。

关键词：情感分析；情感词典；语言特征；统计特征；混合特征

中图分类号：TP391　　**文献标志码：**A　　**文章编号：**

Extension methods of Chinese sentiment lexicon based on corpus

XIE Song-xian, LIU Bo, WANG Ting

(School of Computer Science, National University of Defense Technology, Changsha 410073, China)

Abstract: Extension algorithms for Chinese lexicon are designed and implemented based on linguistic characteristics and statistical features, and a novel method based on mixing features is put forward. Firstly, the algorithm based on linguistic characteristics is implemented by choosing coordinating, progressive and adversative Chinese conjunctions. Secondly, a statistical context vector is designed, and the method of polarity calculation based on the context vector is implemented. Finally, a novel method based on mixing features is put forward and implemented according to the characteristics of the above methods. The results of experiments show that the mixing features method reaches best performance.

Keywords: Sentiment Analysis; sentiment lexicon; linguistic characteristics; statistical features; mixing features

1 引言

自然语言中，一个词的语义极性（semantic polarity）或语义倾向（orientation）表示其对于它的语义组（semantic group）或词汇场（lexical field）规范的偏离方向[1][1]。在自然语言处理领域，情感分析（Sentiment Analysis）能够使用计算手段自动从自然语言中发现观点和情感等主观信息[2, 3]，通常会使用一些标注了极性（积极或消极）的词汇构成的情感词典资源。情感词典的构建方法研究一直得到计算语言学和自然语言处理研究人员关注。Wilson等[4, 5]对一些英文单词进行了积极和消极类别的标注形成了OpinionFinder词典；Bradley等[6]标注了并发布了情感范式的英文词典ANEW，后来Nielsen等[7]在Twitter语言上应用并扩展了ANEW，形成AFINN词典。Esuli和Sebastiani[8]以及后来Baccianella等[9]在著名的英文词典Wordnet基础上采用自动计算的方式开发出了情感词典SentiWordnet。Thelwall等[10]设计实现了能对词语的情感强度进行估计的词典。情感也可以通过创建情绪词典来进行计算，Plutchik情绪轮[11]，提出了四对对立的情绪状态：joy-trust，sadness-anger，surprise-fear和anticipation-disgust。Mohammad和Turney[12]根据Plutchik情绪轮分类方法使用情绪分值标注了一些词语形成NRC情绪词典。在2013和2014年举办的SemEval（Semantic Evaluation）评测中，NRC-Canada队利用NRC词典并扩展出两种新的词典，取得了最好成绩[13, 14]。为了克服以上语法层面建立的词典的上下文语境以及领域适应性问题，一些学者提出了基于概念（concept-based）构建情感词典[15]，其中SenticNet是使用常识知识库建立的公开可用的基于概念的情感词典[16]。

*收稿日期：2014-XX-XX

基金项目：国家自然科学基金资助项目（61170156，61202337）

通信地址：410073 湖南长沙国防科学技术大学计算机学院学员七队

Address：Team seven, School of Computer Science, National University of Defense Technology, Changsha 410073, China

中文情感分析研究起步较晚，想对于英文丰富的情感词典资源，缺乏普遍认可的可靠的中文情感词典。目前研究使用主要有HowNet情感词典[17]，NTUSD情感词典[18]以及大连理工大学的情感词汇本体词库[19]。这些词典主要是以手工或半自动方式编辑而成。我们的前期工作提出了根据语义词典HowNet语义关系将英文情感词典跨语言转换为中文情感词典的方法，并构建了比较全面的情感词典[20]。

基于语义词典的情感词典构建方法是一种常用的情感词典构建方法。采用这种方法的优势在于可以比较容易获取情感词语，基于词语的语义关系也易于进行情感极性计算。但是，基于语义词典的情感词典构建方法受限于语义词典的规模和语义关系的定义，而且对于专业领域中不断涌现的新词语，对情感词典的覆盖度提出了严峻的挑战。随着互联网应用，尤其是社交媒体的不断涌现，越来越多的用户在各种网络平台上发布信息，网络上的用户产生内容（User Generated Content，UGC）不断涌现，研究如何利用网络语料对情感词典进行自动扩展具有十分重要的意义。

本文提出的基于语料资源的无监督的情感词典扩展方法，可以使用于无需标注的网络数据语料对中文情感词典进行自动扩充。

2 问题描述

基于语料资源的情感词语选择与极性计算，在英文中相关研究通常有两种实现思路：一是基于语言特征的方法。例如，Hatzivassiloglou[1]等人采用并列或转折连词来判断新的情感词并计算其极性。二是基于统计特征的方法。例如，Turney等[21]采用点互信息统计学方法从语料中抽取新的情感词并计算其极性。

基于以上情感词语选择与极性计算技术的相关分析，本文将基于中文语料资源扩展情感词典时需要解决的问题描述如下：

- 研究根据中文语言中的并列、递进以及转折关系对情感词的选取以及极性计算的作用；
- 根据中文特点，基于统计特征相关知识设计情感词语选择和极性计算方法；
- 研究采用基于语言特征和统计特征相混合的方式进行情感词语选择和极性计算。

2 数据集及预处理

本文使用的数据资源如表1所示，选取的语料资源是谭松波博士提供的酒店评论、书籍评论和电子商品评论三个领域的语料文本各4000篇[22]。

Table 1　Datesets and resources

表 1　数据集及资源

词典	SentiHowNet	基于前期工作[20]
语料	Hotel 评论	4000 篇
	Book 评论	4000 篇
	NoteBook 评论	4000 篇

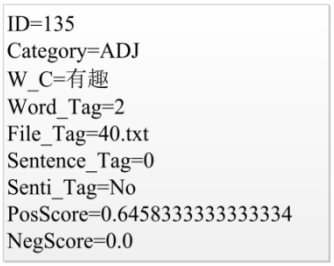

Figure 1　Recording format of preprocessing

图 1 语料预处理记录格式

其中对语料进行预处理需要将中文文本进行分词并进行词性标注。中文分词处理是对语料进行

进一步处理的基础，采用的是中科院设计实现的ICTCLAS分词软件[23]；然后将词性标注为形容词（ADJ）和副词（ADV）的，在SentiHowNet中出现的进行极性和极性值标注；生成的结构化语料预处理记录格式如图1所示，主要有词语编号（ID）、词性（Category）、中文词语（W_C）、词语在句子中的编号（Word_Tag）、词语所在语料文件编号（File_Tag）、词语所在句子编号（Sentence_Tag）、情感标号（Senti_Tag）、积极极性值（PosScore）和消极极性值（NegScore）。值得说明的是，情感标号的取值为Yes和No，分别表示已标注和未标注。

3 基于语言特征的情感词典扩展

早期关于英文的一些研究[1]发现，由连词（如and或but）连接的两个形容词的极性往往存在一定的关联性，如“and”连接的形容词（如“nice and good”）极性相同，而“but”连接的形容词（如“nice but unnatural”）极性相反。而对于中文来说，基于语言特征的中文情感词是否会遵循想通的规律，需要通过实验进行研究。

3.1 连词选择

连词是用来连接词与词、词组与词组或句子与句子、表示某种逻辑关系的虚词。连词可以表示并列、承接、转折、因果等关系。本文主要研究基于表达并列、转折和递进三种关系的连词如何影响情感词的极性计算，选择的连词为：

- 并列关系连词：和、跟、与、既、同、及、况、况且、乃至、并、也、又；
- 转折关系连词：却、虽然、但是、然而、偏偏、只是、不过、至于、致、不料、岂知；
- 递进关系连词：不但、不仅、何况、并、且、而且。

3.2 基于连词的极性计算

基于连词的情感词语选取和极性计算如算法1所示，主要包括获取待标注词语、连词结构分析和极性计算三个过程。待标注词采用基于连词的方法能够计算出情感极性值，其情感极性可以根据计算得到的情感极性值判别。

Algorithm 1　　Poarity calculation based on Conjunction

算法 1　　基于连词的极性计算

已知：待标注词语集$\{w_1\}$，连词集合$\{c\}$，极性已知词语集合$\{w_2\}$

1. **for** 每一待标注词语 $w_1 \in \{w_1\}$：
2. **for** 每一与 w_1 在同句子中已标注词 $w_2 \in \{w_2\}$
3. **if** w_1 和 w_2 在 c 同侧 **then**
4. $\begin{cases} PosScore(w_1) += PosScore(w_2) \\ NegScore(w_1) += NegScore(w_2) \end{cases}$;
5. **else**
6. **if** c 为并列或递进连词 **then**
7. $\begin{cases} PosScore(w_1) += PosScore(w_2) \\ NegScore(w_1) += NegScore(w_2) \end{cases}$;
8. **end if**
9. **if** c 为转折连词 **then**
10. $\begin{cases} PosScore(w_1) -= NegScore(w_2) \\ NegScore(w_1) -= PosScore(w_2) \end{cases}$;
11. **end if**
12. **end for**
13. **end for**

14. 计算极性均值 $\begin{cases} PosScore(w_1) = \dfrac{PosScore(w_1)}{i} \\ NegScore(w_1) = \dfrac{NegScore(w_1)}{i} \end{cases}$

15. 根据情感值 $PosScore(w_1)$ 与 $NegScore(w_1)$ 判断极性

16. 将 w_1 加入到集合 $\{w_2\}$

17. **end for**

3.3 实验

实验中用于评测参考的极性标注标准是基于人工标注和网络注释（百度百科等）等多种途径综合得到的。

评价指标采用正确率、召回率以及F值作为评测标准。a_1 表示积极极性判断正确词数；a_2 表示消极极性判断正确词数；b_1 表示判断为积极极性的词数；b_2 表示判断为消极极性词数；c_1 表示标准积极极性词数；c_2 表示标准消极极性词数。正确率计算公式为：

$$P = \frac{a_1 + a_2}{b_1 + b_2} \times 100\% \quad (1)$$

召回率计算公式为：

$$P = \frac{a_1 + a_2}{c_1 + c_2} \times 100\% \quad (2)$$

F值计算公式为：

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (3)$$

针对三个领域的情感词典扩展实验结果如表2所示，对于三个语料，其召回率均达到67%以上。其中对于Hotel语料，其正确率最低，为43.69%，而其召回率最高为88.24%。其余语料正确率较高。经分析，Hotel语料中可以用于计算的连词结构的语句所占的比例小于其他语料。从平均值上可以看出，基于连接词的词语极性计算同样适用于中文。

Table 2 Performance of different domain corpus

表2 各个领域性能评测结果

	正确率	召回率	F值
Hotel语料	43.69%	88.24%	58.44%
Book语料	67.47%	67.47%	67.47%
NoteBook语料	67.21%	67.21%	67.21%
平均值	59.46%	74.31%	64.37%

4 基于统计特征的情感词典扩展

词语的上下文是词语在实际应用中的语言环境，它在自然语言处理中的价值体现在两个方面：一方面，在自然语言知识获取的过程中，上下文是知识获取的来源；另一方面，在自然语言处理的应用问题解决过程中，上下文扮演着解决所需信息和资源提供者的重要角色。特别是在语料库语言学中，各种机器学习方法的引入使词语的上下文成为计算语言学知识获取和问题求解过程中最为重要的资源，在无监督学习方法中更是如此[24]。本文设计实现的基于统计特征的情感词典扩展方法主要是采用基于上下文的方法进行情感词语极性计算，因为出现在相似上下文环境中的词语具有相似

的极性。

上下文的选取时基于核心词左右一定范围进行的，这个固定的范围被称为“窗口”。选择合适的窗口，可以使得上下文的计算提供的信息量足够大，产生的噪声足够小。在英文中，核心词左右5个词的范围可以为词语搭配提供95%的信息，上下文±2是最好的选择，范围进一步扩大后提供的信息量不会有明显的增加且会带来不必要的计算开销。

本文的方法首先是对待标注词语，分析其上下文词语的词性，获取其特征向量；其次，根据其上下文特征向量实现情感词语选取和极性计算方法。

5.1 统计特征选择

定义4-1 词语 w 的特征向量 $V(w)$ 和窗口 W :词语 w 的特征向量 $V(w)$ 是指由词语 w 与其相邻上下文词语的词性组成的向量，具体形式见公式（4），其中，C_0 表示词语 w 的词性，$C_i(\mathrm{i}\neq 0)$ 表示与 w 相邻的词语的词性， i 表示与词语 w 的相对距离， W 表示窗口，即特征向量中与词语 w 相对距离的最大值。

$$V(\mathrm{w})=<C_{-W},C_{-W+1},\cdots C_{-1},C_0,C_1,\cdots C_{W-1}C_W> \quad （4）$$

5.2 基于上下文的情感词极性计算

基于上下文的情感词极性计算算法2所示，待标注词采用基于上下文的方法能够计算出情感极性值，其情感极性可以根据计算得到的情感极性值判别。

Algorithm 2　　Poarity calculation based on Statistics

算法 2　　基于统计特征的极性计算

已知：待标注词语集 $\{w_1\}$，极性已知词语集合 $\{w_2\}$，每个词特征向量集合 $\{V(w)\,|\,w\in\{w_1\}\cup\{w_2\}\}$

1. **for** 每一待标注词语 $w_1\in\{w_1\}$：
2. **for** w_1 每一特征向量 $V(\mathrm{w}_1)$：
3. **for** 每一与 $V(\mathrm{w}_1)$ 形同的特征向量 $\{V(w_1)=V(w_2)\,|\,w_2\in\{w_2\}$：
4. **if** $Senti\_Tag(\mathrm{w}_2)=positive$ **then**
5. $\begin{cases}PosScore(w_1)+=PosScore(w_2)\\ NegScore(w_1)+=NegScore(w_2)\end{cases}$
6. **else** $\begin{cases}PosScore(w_1)-=PosScore(w_2)\\ NegScore(w_1)-=NegScore(w_2)\end{cases}$
7. **end if**
8. **end for**
9. 计算特征向量 $V(\mathrm{w}_1)$ 下极性均值

$$\begin{cases}PosScore(\mathrm{w}_1)=\left|\dfrac{PosScore(\mathrm{w}_1)}{j}\right|\\ NegScore(\mathrm{w}_1)=\left|\dfrac{NegScore(\mathrm{w}_1)}{j}\right|\end{cases}$$

10. 对各个特征向量下的情感值累加
11. $\begin{cases}PosScore(w_1)+=PosScore(w_1)\\ NegScore(w_1)+=NegScore(w_1)\end{cases}$
12. 计算极性均值 $\begin{cases}PosScore(\mathrm{w}_1)=\dfrac{PosScore(\mathrm{w}_1)}{i}\\ NegScore(\mathrm{w}_1)=\dfrac{NegScore(\mathrm{w}_1)}{i}\end{cases}$

13. 根据情感值 $PosScore(w_1)$ 与 $NegScore(w_1)$ 判断极性
14. 将 w_1 加入到集合 $\{w_2\}$
15. **end for**

5.3 实验

针对三个领域（Hotel、Book、NoteBook）的情感词典扩展实验结果如图2、图3和图4所示。对于三个语料，当窗口 $W=1$ 时，准确率最高，分别为67.65%、72.89%和72.13%；当窗口 $W=2$ 时，召回率有所上升，准确率略有下降；当窗口 $W=3$ 时，召回率最高，准确率和F值下降较多。通过对评测结果进行分析，本文发现在设计基于统计特征的情感词典扩展方法时，采用窗口 $W=1$ 进行情感词语选择，采用窗口 $W=2$ 进行情感词语极性计算，可以获得较好的性能。

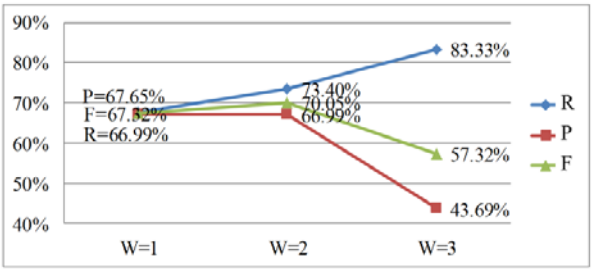

Figure 2 Performance for Hotel corpus

图 2 Hotel语料评测结果

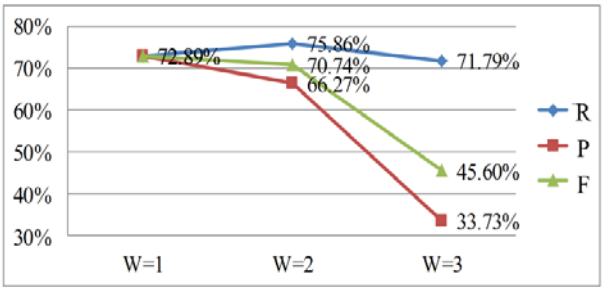

Figure 3 Performance for Book corpus

图 3 Book语料评测结果

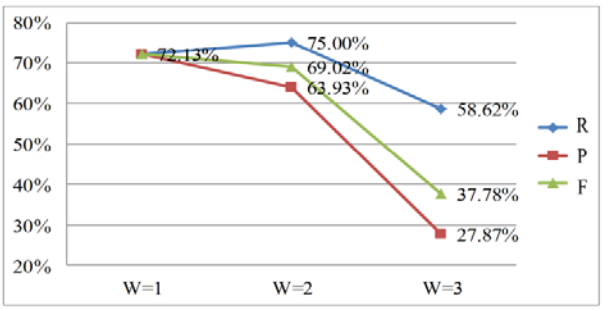

Figure 4 Performance for NoteBook corpus

图 4 NoteBook语料评测结果

5 基于混合特征的情感词典扩展

本文对设计实现的基于语言特征的情感词典扩展方法和基于统计特征的情感词典扩展方法的实验结果进行仔细分析发现，采用语言特征无法进行选择和情感极性计算的词语，可以采用统计特征进行处理；同样的，采用统计特征无法进行选择和情感极性计算的词语，可以采用语言特征进行处理；两种方法可以相互补充。因此本文提出基于混合特征的情感词典扩展方法。

5.1 基于混合特征的情感词极性计算

基于混合特征的情感词语选取和极性计算如算法3。待标注词语采用基于混合特征的方法能够计算出情感极性值的选取为情感词语，其情感极性可以根据计算得到的情感极性值判别。

将选取的情感词语集合分别采用两种方法进行极性计算，在将两种方法计算的极性值合成时，遵循以下原则：

- 优先采用基于统计特征的方法计算出的情感极性值作为待标注词语的情感极性值。
- 当采用基于统计特征的方法进行计算时，优先设置窗口大小为2，其次为1。
- 当采用基于统计特征的方法无法对待评价词语进行情感计算时，采用基于语言特征的方法进行计算。

Algorithm 3　　Poarity calculation based on combination

算法 3　　基于混合特征的极性计算

已知：待标注词语集 $\{w_1\}$，连词集合 $\{c\}$，极性已知词语集合 $\{w_2\}$，

每个词特征向量集合 $\{V(w) \mid w \in \{w_1\} \cup \{w_2\}\}$

1. **for** 每一待标注词语 $w_1 \in \{w_1\}$：
2. 依据算法 2 计算情感极性值
3. **if** $\begin{cases} PosScore(w_1) = 0 \\ NegScore(w_1) = 0 \end{cases}$ **then**
4. 依据算法 1 计算情感极性值
5. 根据情感值 $PosScore(\mathrm{w}_1)$ 与 $NegScore(\mathrm{w}_1)$ 判断极性
6. 将 w_1 加入到集合 $\{w_2\}$
7. **end for**

5.2 实验

针对三个领域（Hotel、Book、NoteBook）的情感词典扩展实验结果如表3所示。

Table 3　　Performance of evaluation

表 3 评测结果

	正确率	召回率	F值
Hotel语料	75.49%	74.76%	75.12%
Book语料	77.11%	77.11%	77.11%
NoteBook语料	78.69%	78.69%	78.69%

基于语言特征的情感词典扩展、基于统计特征的情感词典扩展和基于混合特征的情感词典扩展的实验评测结果对比情况如图5、图6和图7所示，通过分析发现，基于混合特征的情感词典扩展方法的评测性能是在各个领域语料中均是最优的。

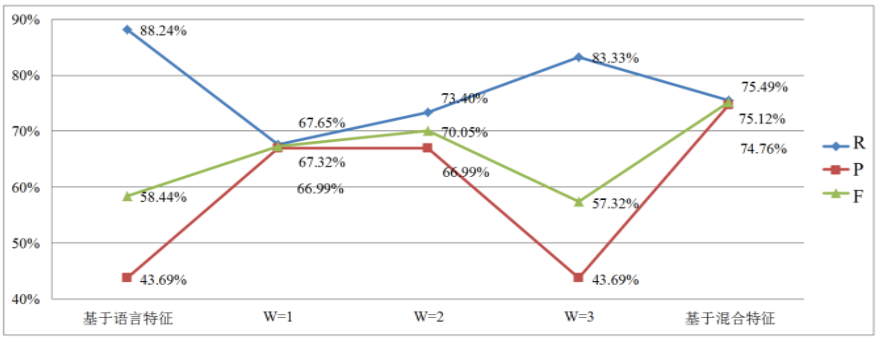

Figure 5 Performance comparison for Hotel corpus

图 5 Hotel语料评测结果综合比较

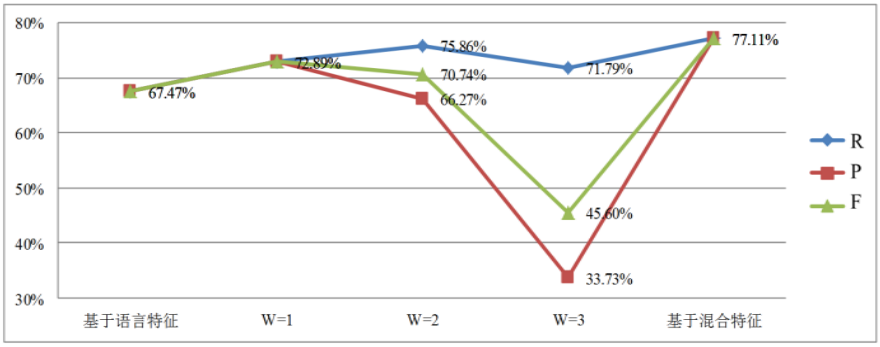

Figure 5 Performance comparison for Book corpus

图 6 Book语料评测结果综合比较

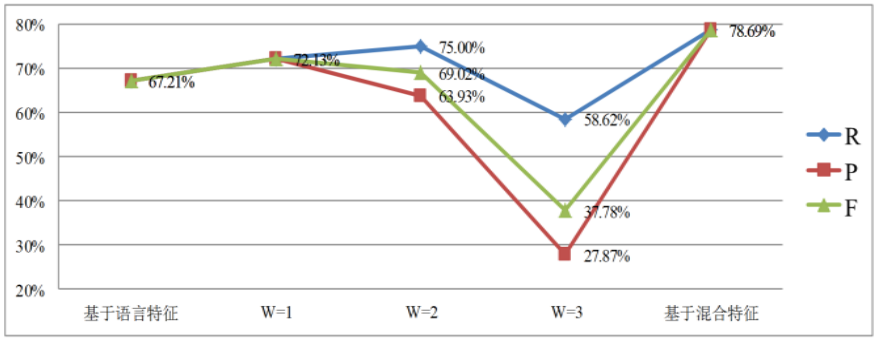

Figure 5 Performance comparison for NoteBookcorpus

图 7 NoteBook语料评测结果综合比较

6 结论

本文详细讨论了基于语料资源的中文情感词典扩展问题描述和方法设计，对基于语言特征的情感词典扩展和基于统计特征的情感词典扩展的关键技术分别进行了研究和算法实现，并提出了了基于混合特征的无监督的情感词典扩展方法。通过分析每个方法的实验结果，发现基于混合特征方法能够达到最好性能。

参考文献(References)

[1] 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. 1997: 174-181.

[2] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval. 2008, 2(1-2): 1-135.

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

[4] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[C] //Proceedings of the conference on human language technology and empirical methods in natural language processing. 2005: 347-354.

[5] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis[J]. Computational linguistics. 2009, 35(3): 399-433.

[6] Bradley M M, Lang P J. Affective norms for English words (ANEW): Instruction manual and affective ratings[R]. Citeseer, 1999.

[7] Nielsen F A R. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs[J]. arXiv preprint arXiv:1103.2903. 2011

[8] Esuli A, Sebastiani F. Sentiwordnet: A publicly available lexical resource for opinion mining[C] //Proceedings of LREC. 2006: 417-422.

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

[10] Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web[J]. Journal of the American Society for Information Science and Technology. 2012, 63(1): 163-173.

[11] Plutchik R. The nature of emotions[J]. American Scientist. 2001, 89(4): 344-350.

[12] Mohammad S M, Turney P D. Crowdsourcing a word--emotion association lexicon[J]. Computational Intelligence. 2013, 29(3): 436-465.

[13] Mohammad S M, Kiritchenko S, Zhu X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets[C] //Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013). Atlanta, Georgia, USA: 2013.

[14] Kiritchenko S, Zhu X, Cherry C, et al. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews[C] //Proceedings of the International Workshop on Semantic Evaluation, SemEval. 2014.

[15] Tsai A C, Wu C, Tsai R T, et al. Building a concept-level sentiment dictionary based on commonsense knowledge[J]. IEEE Intelligent Systems. 2013, 28(2): 22-30.

[16] Cambria E, Olsher D, Rajagopal D. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis[C] //Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

[17] 知网 HowNet 评价词词典 [EB/OL]. 2013/2013-07-25[2013-08-15]. http://www.keenage.com/html/c_index.html.

[18] Ku L W, Chen H H. Mining opinions from the Web: Beyond relevance retrieval[J]. Journal of the American Society for Information Science and Technology. 2007, 58(12): 1838-1850. Wiley Online Library.

[19] 情感词汇本体库 [EB/OL]. 2013/2013-07-30[2013-08-15]. http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx.

[20] 谢松县，刘博，王挺. 应用语义关系自动构建情感词典[J]. 国防科技大学学报. 2014, 36(3): 111-115.

[21] 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. 2002: 417-424.

[22] Lin Z, Tan S, Cheng X, et al. Effective and efficient?: bilingual sentiment lexicon extraction using collocation alignment[C] //Proceedings of the 21st ACM international conference on Information and knowledge management. New York, NY, USA: ACM, 2012: 1542-1546.

[23] 张华平. NLPIR/ICTCLAS2014 分词系统 [EB/OL]. 2014/2014-06-18http://ictclas.nlpir.org/newsdownloads?DocId=389#.

[24] 鲁松，白硕. 自然语言处理中词语上下文有效范围的定量描述[J]. 计算机学报. 2001, 24(7): 742-747.