一种跨语言基于语义关系的中文情感词典自动构建方法**[[1]](#footnote-1)\***

摘 要：情感分析资源是进行情感分析研究的基础，具有十分重要的意义。本文提出了基于语义关系的情感词典自动构建方法，并以HowNet、WordNet语义词典和SentiWordNet情感词典为基础进行了实现。实验结果表明，本文设计实现的方法取得了较好的评测性能。设计生成的SentiHowNet情感词典记录了词语的语义关系、情感极性值等多种信息，提供了丰富的情感资源和应用访问接口，对推动情感分析技术不断发展，促进舆情分析等情感分析应用提供了基础资源。

关键词：情感分析；情感词典；语料资源；语义关系

An Automatic Approach of Sentiment Lexicon Construction Based on Semantic Relations

**Abstract:** It is a fundamental and important task to construct sentiment lexicon for the research of sentiment analysis. This paper proposes an automatic construction method of sentiment lexicon based on semantic relations. The method was implemented with the WordNet, HowNet semantic resources and SentiWordNet lexicon. The experimental results show that the method achieved good performance with evaluation of recall and precision. Finally, this paper introduces the developed sentiment lexicon (SentiHowNet) and its application interface based on the research results of the proposed constructing method. Its entries records the semantic relationships of words, sentimental polarity value and other information, provide a wealth of sentiment resources and application interface for sentiment analysis, and also provide fundamental resources for applications such as the public opinion analysis.

**Keywords:** Sentiment analysis, Sentiment lexicon, Corpus resource, Semantic relationship

1 导言

情感分析是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程 [1][2]。情感分析涉及情感信息的抽取、情感信息的分类以及情感信息的检索与归纳等多项非常具有挑战性的任务。情感分析资源建设是进行情感分析研究的基础，具有十分重要的意义。情感资源主要表现为由人工判断或计算机自动整理的人类语言中具有明确情感倾向的词典知识，往往是是进行情感分析的基础。

情感分析研究在国外已经有了几十年的发展史，积累了许多情感资源和研究成果，尤其是英文情感词典（如SentiWordNet[3]等）已经作为情感分析领域的资源建设标准。将情感资源丰富的英文词典跨语言向资源相对贫乏的语言进行适应性的转化，以产生其他语言情感分析词典是近年来的研究发展趋势。国内关于中文情感资源建设和扩展等相关研究还十分匮乏，这为我们进行汉语情感信息的倾向性分析带来了许多困难。如何能够充分利用国外研究成果，结合中文处理的特点，设计实现有效的自动情感资源构建方法，并将研究成果运用到中文情感分析领域，一直是值得深入探讨和研究的问题。

基于构建和扩展情感资源进行网络信息挖掘和舆情监控等应用的迫切需求，本文主要研究利用WordNet[4]和HowNet[5]、SentiWordNet等资源，抽取描述情感信息的词汇，构建情感词典。目前基于语义词典构建情感词典的研究主要集中于英文资源。对于中文情感词典构建和实现的研究十分匮乏，由于中文的情感资源并不像英文资源那样丰富，使得构建中文情感词典存在许多困难，为我们进行中文情感信息的倾向性分析带来了许多不便。因此，设计实现有效的中文情感词典构建方法具有重要的意义。本文就是利用了资源丰富的英文情感词典SentiWordNet，提出了一种基于HowNet和WordNet语义关系的中文情感词典自动构建方法。

2 国内外研究现状

情感词语是带有情感倾向性的词语，在情感分析相关应用中有着重要的作用。对于情感分析研究来说，设计实现高效的情感词典构建和扩展方法是一项非常基础而且十分重要的工作。目前，设计和实现情感词典主要有两类方法：基于语料的方法和基于语义词典的方法。

基于语料的情感词典构建方法是利用语料库的统计特征，观察一些现象来挖掘语料库中的情感词并判断其情感倾向。如Hatzivassiloglou [6] 等人从华尔街日报的语料库中挖掘出大量的形容性的情感词语，Wiebe[7]等人使用一种类似相似度分布的词聚类方法在大语料库中完成了形容词性的情感词语的获取。Turney[8]等人提出使用点互信息（Point Mutual Information）的方法来识别各种词性的情感词语。基于语料库的方法构建情感词典的优点在于简单易行，缺陷在于可利用的语料库有限，情感词语在大语料库中的分布等现象不容易归纳。

基于语义词典的情感词典构建方法通常是采用语义词典来判断词语的相似度。常用的词典包括WordNet和HowNet。Kamps[9]等人利用点互信息的方法，通过计算WordNet中的所有形容词与种子评价词“good”和“bad”之间的关联度值来识别出评价词语。Stefano Baccianella[10]等人提出基于WordNet的标签传播算法并构建了情感词典SentiWordNet。在国内，比较有代表性的是刘群和李素建[11]提出的基于HowNet的词汇语义相似度计算方法，采用该方法可以构建中文情感词典。

构建情感词典从开始就注重机器学习或自动学习方法的应用。在早期，有学者通过手工收集种子情感词语（基准词）、并进行人工标注或采用半自动学习标注进行情感词典的扩展与构建的相关研究[12-15]。而近期的研究在于基于已有的情感词典和语义词典采用机器学习进行情感词典的构建和扩展，并尝试使用迁移学习方法进行跨领域或跨语言的情感词典构建。比如，情感词典构建时引入基于图的方法[16-19]。基于图的方法是一种新颖的方法，它可以灵活的将词语间的各种联系作为特征融入图中继而进行迭代计算。这种方法的有效性受限于选取的词语特征和图算法。寻找有效的词语间特征以及如何选取图算法是值得深入研究的问题。

情感词典构建还存在领域问题。不同领域使用不同的情感词语及表达方式，同一情感词语或表达方式在不同应用领域表现出不同的语义倾向。这导致通用情感词典或某一领域获得的情感词典在应用到新的领域时存在较大的问题。研究情感词典的领域移植问题对于情感分析的领域细化，发挥出情感分析技术巨大潜力具有重要意义。已有不少学者采用迁移学习等方法进行情感词典知识的领域化和跨语言研究，并且取得了一定的成果[20-24]。

3 词典资源简介

3.1 HowNet语义词典

HowNet是描述概念与概念之间的关系，以及概念的属性与属性之间的关系的知识库。它支持中英文两种语言，有着自己独特的知识表示方法。HowNet把客观世界看作是由很多的概念构成。概念之间相互交织构成一个网。这是HowNet区别于其他语义词典的本质特征[5]。HowNet的网状知识体系如图1所示。HowNet所着力要反映的是概念的共性和个性。例如，对于“医生”和“患者”，“人”是它们的共性；HowNet在主要特征文件中描述了“人”所具有的共性，“医生”的个性是他是“医治”的施事，而“患者”的个性是他是“罹患”的经验者。如图2所示，HowNet采用KDML（Knowledge Dictionary Mark-up Language）其中W\_X表示词语，G\_X表示词语词性，E\_X表示词语例子，DEF表示概念定义，X为C时表示中文，X为E时表示英文。

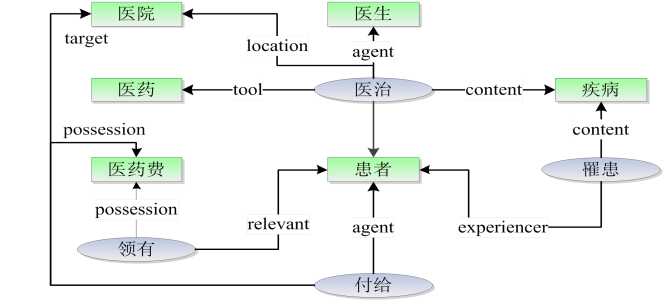
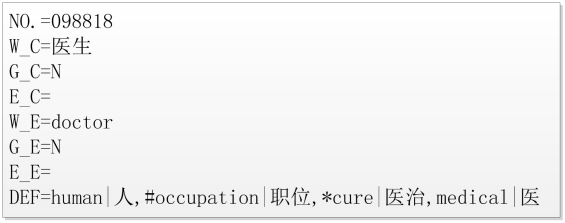
 

图1 HowNet中的网状知识体系结构 图2 HowNet中概念的定义方式

在HowNet中，每一个概念的定义项都是由义原和动态角色来描述的。出现在定义项中由双语标注的是义原。义原是HowNet中最小语义单元，用于定义和描述概念的属性和概念间的相互关系。HowNet中描述了概念之间的多种类型的关系，将图1所示的网状知识体系明确的教给了计算机进而使得知识对计算机而言是可操作的。

3.3 WordNet语义词典

WordNet是由Princeton 大学的心理学家，语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典。WordNet是根据词义而不是词形来组织词汇信息。WordNet使用同义词集合（Synset）代表概念（Concept），词汇关系在词语之间体现，语义关系在概念之间体现。WordNet将英语的名词、动词、形容词和副词组织为Synsets，每一个Synset表示一个基本的词汇概念，并在这些概念之间建立了包括同义关系（synonymy）、反义关系（antonymy）、上下位关系（hypernymy & hyponymy）、部分关系（meronymy）等多种语义关系。其中，WordNet最重要的关系就是词的同义反义关系，因为判断词之间这种关系的能力是在词汇矩阵中表达词义的先决条件。

4 基于语义关系的情感词典构建方法

构建情感词典包括情感词语抽取、情感词语倾向计算、情感词典生成三个基本过程。通过前面对HowNet、WordNet词典的介绍，根据构建情感词典的基本过程和语义词典的基本特点，将构建情感词典时需要解决的问题描述如下：

* 词语抽取和义原抽取

从HowNet中抽取中文情感词语W\_C和相应的英文词语W\_E以及词语的属性定义DEF。HowNet中的W\_C是通过DEF进行定义的，DEF是由多个义原构成。

* 词语倾向计算

在WordNet中抽取W\_E和DEF相应的英文词语Synset以及该英文词语所在的Sense；从SentiWordNet抽取Synset和Sense相关极性值PosScore和NegScore。分别分析HowNet和WordNet中的语义关系，设计词语倾向计算方法，计算W\_C的极性值PosScore(W\_C)和NegScore(W\_C)。

4.1 词语抽取和义原抽取

词语和义原抽取处理流程如图3所示。

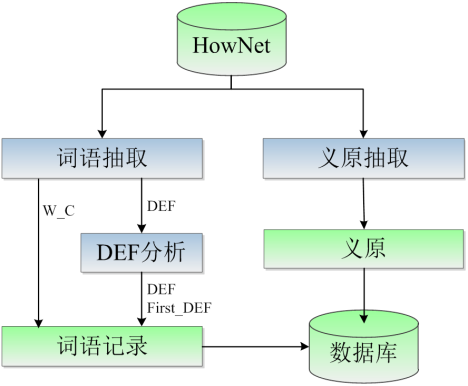


图3 词语和义原抽取处理流程

4.1.1 词语抽取

从HowNet中抽取出的词语，定义其记录格式如图4所示。在抽取得到的词语记录中，主要关注的内容有词语编号（No.）、中文词语（W\_C）、中文词性（G\_C）、英文词语（W\_E）、英文词性（G\_E）、属性（DEF）、第一属性（First\_DEF）等。其中第一属性是指位于属性DEF第一位置的义原。通过第一属性可以分析出该词语所属的特征类[24]。

图4 抽取的词语记录格式 图5 抽取的义原记录格式

4.1.2 义原抽取

由于HowNet中的词语是由义原和语义关系描述等构成的。在进行词语倾向计算时，需要根据义原进行词语的语义分析和倾向计算。本文定义从HowNet中抽取得到的义原的记录格式如图5所示。在抽取得到的义原的记录中，主要关注的内容有词语编号（No.）、特征类别（Category）、中文词语（W\_C）、英文词语（W\_E）、属性（DEF）、层次（Layer）、父亲节点编号（Father）等。根据记录中的层次（Layer）和父亲节点编号（Father）可以得到义原之间的语义关系，如编号为33的义原“依靠”位于“事件类（Event）”的第五层，其父亲节点编号为32，通过查询编号为32的义原，得到其父亲节点义原为“有关（relate）”。

4.2词语倾向计算

HowNet中词语是由其属性DEF定义的，DEF是由多个义原组成的，词语的倾向性是由义原的倾向性按照一定的规律组合而成的。词语的倾向性值可以通过对义原的倾向性值根据语义关系计算获得。

4.2.1 义原倾向性值计算

义原倾向性值计算主要包括义原特征分析、基于WordNet和SentiWordNet倾向性值计算和基于HowNet语义倾向性值计算三个主要步骤。

1.义原特征分析

HowNet中义原共分为七类，具体分类情况如表1所示。通过对各个类别义原进行分析，我们得到以下结论：

表1 HowNet义原分类表

|  |  |
| --- | --- |
| **特征类别** | **义原数目** |
| 事件（Event） | 813 |
| 事物（Entity） | 142 |
| 属性（Attribute） | 117 |
| 属性值（Attribute Value） | 433 |
| 数量（aValue） | 3 |
| 数量值（qValue） | 13 |
| Secondary Feature | 100 |

* 属性类和数量类的义原是所有属性概念词语或数量概念词语的唯一主要特征，其倾向通常为中性。例如，属性类中的“length|长度”和数量类中的“rate|比率”均为中性词语.
* 属性值类和数量值类中用于表示具体属性值和数量值的义原通常带有情感倾向，需要进行计算。例如，“good|好”和“sufficient|足”均带有一定的情感倾向。
* 事件类中的多数义原带有情感倾向。例如，事件类中的“blame|埋怨”等。
* 事物类中的义原用于标识事物的主要特征，通常为中性词语。
* Secondary Feature类中部分义原带有情感倾向。例如“desired|良”和“undesired|莠”等带有情感倾向。

2.基于WordNet和SentiWordNet的义原倾向计算

基于WordNet和SentiWordNet的义原倾向计算过程如图6所示。在HowNet中获取义原后，首先对义原进行特征分析；然后将义原的英文词语（如“good”）映射到WordNet中，得到该词语所在的Sense（如“good”的Sense共有27个）；将这些Sense映射到SentiWordNet中得到英文词语对应Sense的极性值；通过倾向性值加权计算得到义原的情感倾向值（如“good”的倾向值为PosScore=0.597，NegScore=0.004）。

为了计算义原的情感极性值（积极、消极、中性），提出了义原倾向计算公式（1）[27]。

（1）

公式中表示极性类型（积极、消极、中性，“P、N、O”），为与义原相对应的Sense的总数，表示义原，表示义原的极性值，表示义原在编号为的Sense中的类型极性值。

3.基于HowNet语义关系的倾向性值计算

并不是所有的义原都能映射到WordNet中，此外，有些义原通过WordNet和SentiWordNet计算得到的极性值与义原的真实语义倾向相反。因此，需要利用HowNet中的语义关系对上述计算方法进行补充和校正。本文设计实现了基于HowNet中对义和反义等语义关系进行义原倾向性值计算的方法，具体计算过程如图7所示，若义原尚未标注情感倾向，则利用对义和反义义原进行情感倾向计算（即待标注义原情感倾向性值与对义或反义义原情感倾向性值相反）；若义原已标注情感倾向性值，则利用对义和反义进行情感倾向性值校正（通过投票的方式决定义原的情感倾向性值）。

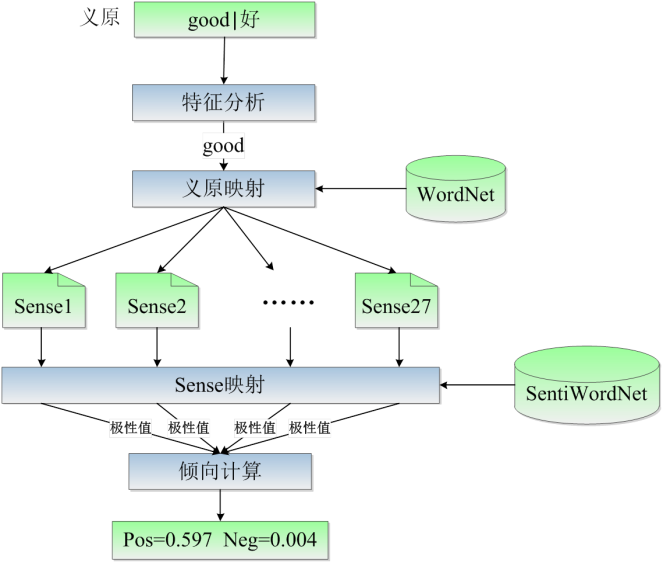
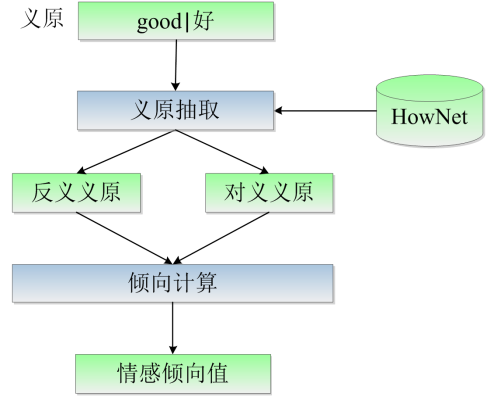
 

图6义原倾向性值计算 图7 基于对义和反义关系的义原倾向计算

4.4.2 词语倾向性值计算

本文的词语倾向性值计算方法主要是在前述义原情感倾向性值基础上进行设计实现的。通过对HowNet中词语属性DEF进行分析，根据DEF中义原以及语义关系，所设计词语倾向性值计算过程如图8所示，其中针对语义关系的不同提出如下定义：

**定义1 情感倾向值取反。**词语的极性值取反运算是，将的积极倾向值和消极倾向值互换，运算过程如公式（2）。

（2）

**定义2 乘法运算。**因子与词语的极性值的乘法运算定义为乘法运算，运算过程如公式（3）。

（3）

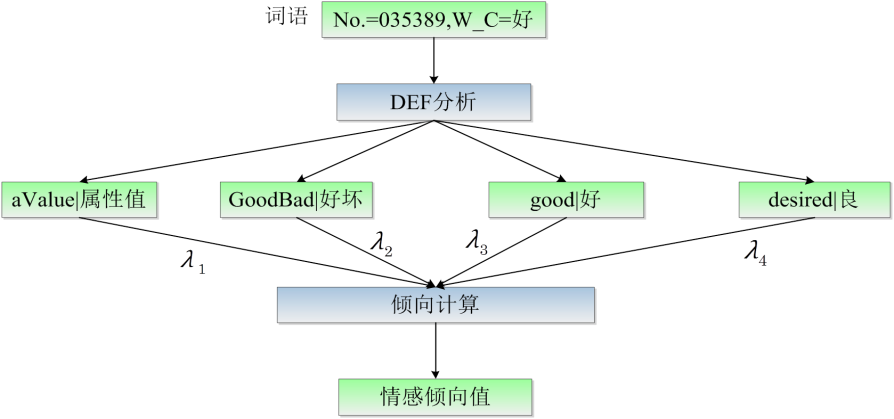


图8 词语倾向计算过程

取值的确定需要根据义原的类别特征、词语DEF的组成特征和义原间的语义关系进行确定。如图8中所示的词语“好”的DEF中每个义原的可以均取值为1。再如，词语“扭亏为盈”的DEF为“DEF=alter|改变，StateIni=InDebt|亏损，StateFin=earn|赚”，义原“InDebte|亏损”为初始状态，“earn|赚”为最终状态，经过分析后，义原“InDebte|亏损”的取值为0，义原“earn|赚”的取值为1。经过以上分析计算，可以得到词语情感极性值。本文将词语倾向计算过程总结为公式（4）。公式（4）通过对组成词语DEF的义原的情感极性值进行加权求均值得到词语的情感极性值。其中表示词语的极性值，表示词语DEF中第个义原，为词语DEF中义原总数。

（4）

5 实验及结果分析

实验使用的词典资源及相关信息如表2所示。

表2 实验使用的词典资源

|  |  |  |
| --- | --- | --- |
| **词典** | **版本** | **备注** |
| HowNet | 2000年Beta版 |  |
| WordNet | 3.0 | 开源API（JAWS[37]） |
| SentiWordNet | 3.0.0 |  |
| HowNet评价词词典 | 2007年版 | 用于评测算法性能 |

5.1 义原倾向性值计算结果

经过基于WordNet和SentiWordNet的方法进行倾向性值计算后，有1349项义原被标注了情感倾向性值，经过基于语义关系的方法进行倾向性值计算后，所有义原均被标注了情感倾向性值，其中情感倾向性值被校正的义原有349项。在所有被标注的义原中，有1026项带有情感倾向，595项为中性义原。

5.2 词语倾向性值计算结果

经过词语情感倾向性值计算后，HowNet中所有中文词语均被标注了情感极性值（PosScore和NegScore）。在构建情感词典时，我们给出了词语唯一的的情感倾向。情感倾向判别参考SentiWordNet的倾向判别方法，情感倾向值计算如公式（5）所示，

（5）

通过计算词语的情感倾向值Score来判断词语的极性。公式（5）中T表示极性阈值，即Score(s)处于-T和T之间的词语为中性词。词语的情感倾向的强度分级如表3所示。

在使用公式（1）、（2）和（3）对本章实验进行评测时，采用HowNet发布的评价词词典[25]作为人工标注的评测标准。由于HowNet评价词词典是2007年发布的，而本文采用的HowNet词典是2000年版本，因此，HowNet评价词词典中的部分词语在HowNet词典中找不到对应项。经过将两者比对筛选，HowNet评价词词典中可以用于评测的情感词共有6497项，其中积极情感倾向词语3436项，消极情感倾向词语3061项。

1. 阈值T=0时的性能

评测性能如表4所示。可以看出在T=0时，召回率较高达88.58%，但准确率较低，仅有24.4%。

表4 T=0时的性能评测

|  |  |  |  |
| --- | --- | --- | --- |
|  | **准确率（）** | **召回率（）** | **值** |
| **积极倾向** | 24.55% | 88.85% | 38.48% |
| **消极倾向** | 24.23% | 88.27% | 38.03% |
| **总体性能** | 24.4% | 88.58% | 38.27% |

1. 阈值T=0.05时的性能

T=0.05时评测性能如表3-9所示可以看出在T=0.05时，准确率与T=0相比由24.2%提高到60.75%，有较大提高。召回率与T=0相比由88.58%下降到87.61%，下降幅度较小，结果仍然比较理想。值与T=0相比也有较大提高，由38.27%提高到71.75%。

通过设置阈值T=0和T=0.05，对比两种情况下的评测性能发现，合理设置阈值可以提高算法的性能指标。

表5 T=0.05时的性能评测

|  |  |  |  |
| --- | --- | --- | --- |
|  | **准确率（）** | **召回率（）** | **值** |
| **积极倾向** | 70.91% | 87.83% | 78.47% |
| **消极倾向** | 52.30% | 87.36% | 65.43% |
| **总体性能** | 60.75% | 87.61% | 71.75% |

6 总结

本文对情感资源构建相关研究进行了分析，设计了基于语义词典的情感词典自动构建方法，并以HowNet、WordNet语义词典和SentiWordNet情感词典为基础进行了实现。实验结果表明，本文设计实现的方法取得了较好的评测性能。本文设计生成的SentiHowNet情感词典记录了词语的语义关系、情感极性值等多种信息，提供了丰富的情感资源和应用访问接口，对促进舆情分析等情感分析应用具有重要意义。

情感资源构建技术的研究和实现对推动情感分析技术的不断进步发展具有重要意义。构建和扩展情感词典的关键是情感词语的选择和倾向计算方法的设计和实现。本文虽然设计实现了基于语义关系的情感词典构建方法，但仍有一些不足。在下一步工作中将重点从以下方面进行研究：研究如何将基于语料的情感词语选择和倾向性计算方法，并扩展至多种词性，而不是局限于形容词性和副词性的词语；研究如何利用语义词典对扩展的情感词语进行自动语义标注，丰富情感词典的语义资源，提高情感词典的实用性。

参 考 文 献

1. 吴琼,谭松波,程学旗.中文情感倾向性分析的相关研究进展[J].信息技术快报.2010,Vol.8 No.4:16-18.
2. 赵妍妍,秦兵,刘挺.文本情感分析综述[J].软件学报.2010,第8期.
3. Andrea Esuli, Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining[C]. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC’06). 2006. 417-422.
4. Miller G,Reckwith R, Fellbaum C, et al. Introduction to WordNet: An On-line Lexical Database[EB]. Cognitive Science Laboratory Princeton University, 1993.
5. 知网.HowNet[EB/OL]. http://www.keenage.com. 2000.3.10.
6. Hatzivassiloglou V, McKeown KR. Predicting the semantic orientation of adjectives[C]. In: Proceedings of EACL-1997. Morristown: ACL, 1997. 174-181.
7. Wiebe J. Learning Subjective Adjectives from Corpora[C]. In Proceeding of AAAI. 2000.
8. Turney P, Littman ML. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems (TOIS). 2003. 21(4):315-346.
9. Kamps J, Marx M, Mokken RJ. Using WordNet to Measure Semantic Orientation of Adjectives[C]. In: Proceedings of the LREC. 2004.
10. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. http://swn.isti.cnr.it/. 2010.5.
11. 刘群,李素建.基于《知网》的词汇语义相似度计算[A].第三届汉语词汇语义学研讨会. 2002.5.
12. Kim SM, Hovy E. Automatic Detection of Opinion Bearing Words and Sentences[C]. In: Proceedings of IJCNLP-2005. 2005. 61-66.
13. Kim SM, Hovy E. Identifying and Analyzing Judgment Opinions[C]. In: Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). 2006. 200-207.
14. Zhu YL, Min J, Zhou YQ, Huang XJ, Wu LD. Semantic Orientation Computing Based on HowNet[J]. Journal of Chinese information processing. 2006. 20(1):14–20.
15. Rao D, Ravichandran D. Semi-supervised Polarity Lexicon Induction[C]. In: Proceedings of EACL-2009. 2009. 675-682.
16. Takamura H, Inui T, Okumura M. Extracting Semantic Orientation of Words Using Spin Model[C]. In: Proceedings of the Association for Computational Linguistics (ACL). 2005. 133-140.
17. B. Pang, L. Lee. Opinion Mining and Sentiment Analysis[M]. Foundations and Threads in Information Retrieval. 2008, Vol. 2:1-135.
18. Huang XJ, Zhao J. Sentiment Analysis for Chinese Text[J]. Communications of CCF. 2008, 4(2).
19. Riloff E, Wiebe J. Learning Extraction Patterns for Subjective Expressions[C]. In: Proceedings of EMNLP-2003. 2003. 105-112.
20. Fangtao Li, Sinno Jialin Pan, etc. Cross-Domain Co-Extraction of Sentiment and Topic Lexicons. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics ([ACL 2012](http://www.acl2012.org/)). Korea, July 2012.
21. Blitzer, John, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2007). 2007.
22. Du, Weifu, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. Adapting information bottleneck method for automatic construction of domainoriented sentiment lexicon. in Proceedings of ACM International Confernece on Web search and data mining (WSDM-2010). 2010.
23. Duh, Kevin, Akinori Fujino, and Masaaki Nagata. Is machine translation ripe for cross-lingual sentiment classification? in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011). 2011.
24. Hatzivassiloglou V, Wiebe J. Effects of Adjective Orientation and Gradability on Sentence Subjectivity[C]. In: Proceedings of the International Conference on Computational Linguistics (COLING). 2000. 299-305.
25. 知网.HowNet评价词词典[EB/OL]. http://www.keenage.com/html/c\_index.html, 2007.
26. Turney P D, Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), 417-424, Philadelphia, PA, USA, July 6-12, 2002.
27. 张猛,彭一凡,樊扬,李丹,林小俊,吴玺宏.中文倾向性分析的研究[J].第一届中文倾向性分析评测. 2008.11

1. \*本课题得到中国国家自然科学基金（61170156）资助。 [↑](#footnote-ref-1)