**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)\***

谢松县，王挺

（国防科技大学 计算机学院， 湖南 长沙 410073）

**摘 要：**情感分析是人工智能的重要研究领域，情感词典资源是进行情感分析研究的基础。构建英文情感词典研究相对成熟，形成了可靠的丰富词典资源。而针对中文的研究时间短，中文情感分析词典资源较少。借鉴现有可靠的英文资源，提出了跨语言的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。生成的情感词典记录了词语的语义关系、情感极性值等多种信息，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；HowNet；语义关系

**中图分类号：**TP316 **文献标志码：**A　　　**文章编号**：1001-2486

A Cross-language Automatic Constructing Approach For Sentiment Lexicon Based on Semantic Relations

**Abstract:** Sentiment analysis is an important research field of artificial intelligence, and sentiment lexicon resources are basic for sentiment analysis research. Research on Constructing English sentiment lexicon is relatively mature, and there are several reliable and abundant dictionary resources. While for Chinese studies, the time is short, and there are fewer Chinese sentiment lexicon resources. Using a reliable English sentiment lexicon as source of reference, a cross-language automatic constructing approach has been proposed, which is based on semantic relationships of semanic dictionary. The mehod was implemented with the HowNet, WordNet semantic dictionary and SentiWordNet sentiment lexicon. Experimental results show that the proposed approach achieved good performance. The resulted sentiment lexicon records such information of Chinese words as semantic relationships and sentimental strength values,etc. It can improve the development of Chinese sentiment analysis technology and provide basic resources for such applications as public opinion analysis.

**Keywords:** Sentiment analysis, Sentiment lexicon, HowNet, Semantic relationship

1 导言

随着互联网的发展，尤其是社交网络的发展，以微博为代表的用户发布内容平台中出现了海量含有用户主观情感色彩的文本数据。针对文本的信息处理开始由获得关键词[1]、事件[2]、话题[3]等事实信息，逐渐开始向情感观点等主观信息深入，情感分析便是近年来迅速发展的信息处理技术[4]。从这些数据中提炼出用户的主观信息对于商业情报、舆情分析等具有重要意义。由于数据量巨大，人工分析变得不可能，需要借助计算机技术进行自动分析处理。情感分析技术就是对带有情感色彩的主观性文本进行自动推理、分析、归纳的过程，涉及自然语言处理、机器学习、认知科学以及社会心理学等方面的研究[5, 6]。情感分析包括情感信息的抽取、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。语言的情感表达往往使用具有明确情感色彩的词汇，因此构建带有情感色彩的词典资源是进行情感分析研究的基础。情感资源可以由人工判断整理，或者使用计算机自动整理的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。

情感分析研究在英文上发展迅速，积累了许多情感词典资源，比如：General Inquirer(GI)[7]，Opinion Finder (OF)[8]，Appraisal Lexicon (AL)[9]，SentiWordNet[10]以及Q-WordNet[11]。英文情感词典事实上已经作为情感分析领域的词典资源建设标准。相对于其它语言，能够将情感词典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，以产生其相应情感分析词典，既可以省去耗费大量人力的人工标注，又可以克服自动或半自动方法的可靠性问题。关于中文情感词典资源建设和扩展等相关研究还刚刚起步，词典资源相对较少，并且汉语词语的复杂语义使得词典的可靠性受到影响，对汉语的情感倾向性分析带来了困难。如何能够充分利用国外研究成果，结合中英文文语言上的语义对应关系，设计实现有效的自动情感词典资源构建方法，是值得深入探讨和研究的问题。

基于构建和扩展情感词典资源进行网络信息挖掘和舆情监控等应用的迫切需求，主要研究如何利用英文中比较成熟的情感词典SentiWordNet，根据中文语义词典HowNet的中英文双语语义对应关系，自动构建中文情感词典。

2 研究现状

情感词语是带有情感色彩的词语，在情感分析中通常指带有褒义和贬义情感倾向性的词语。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。基于语料统计的情感词典构建方法通过词语间的共现信息计算它们之间的相似度，并计算其倾向性。而基于语义知识库的情感词典构建方法则通过语义词典来判断词语义相似度，常用的英语语义词典主要为WordNet，中文主要包括HowNet和同义词辞林等。

基于语料的情感词典构建方法是利用语料库的统计特征挖掘语料库中的情感词并判断其情感倾向。如Hatzivassiloglou[12]等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[13]等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情感词语的获取。Turney[14]等人提出使用点互信息（Point Mutual Information，PMI）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。Kamps[15]等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]等提出基于WordNet的标签传播算法并构建了情感词典SentiWordNet。在国内，刘群和李素建[16]首先提出了基于HowNet的词汇语义相似度计算方法，后续很多工作基于此方法构建了中文情感词典[17, 18]。

构建情感词典从开始就注重机器学习方法的应用。比如，情感词典构建时引入基于图的方法[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和语言问题。不同领域和语言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和语言中表现出不同的语义倾向。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, 21]。

由于中文情感分析研究起步较晚，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[22]，NTUSD情感词典[23]以及大连理工大学的情感词汇本体词库[24]。这些词典主要是以手工或半自动方式编辑而成，可靠性和领域适应性受到限制，并且情感词以主要褒贬二值区分，缺少情感强度值的细粒度划分。本文提出基于可靠的英文情感词典资源的构建方法，可以根据语义关系将英文词与情感值转化得到中文词语的情感强度值，并且完全是自动的，可靠性和适应性更高。

3 词典资源简介

本章对要用到的HowNet中文词典、WordNet和SentiWordNet英文词典进行简单介绍。

3.1 HowNet语义词典

HowNet是一个以中英文词语所代表的概念为描述对象，揭示概念与概念之间以及概念的属性与属性之间的关系的知识库。义原是HowNet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，义原通过一个树状的层次结构组织构成上下位关系。概念是对词汇语义的一种描述，每一个词可以表达为几个概念。HowNet所着力要反映的是概念的共性和个性，这是HowNet区别于其他语义词典的本质特征[16]。HowNet的网状知识体系如图1所示，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）语言描述概念，其中W\_X表示词语，G\_X表示词语词性，E\_X表示词语例子，X为C时表示中文，X为E时表示英文。DEF是对于该概念的定义项，称之为一个语义表达式，其中中英文标注的是义原，“#\*”等标示符号来对概念属性之间关系进行描述，DEF中还可以包含概念，概念之间相互交织构成一个网。

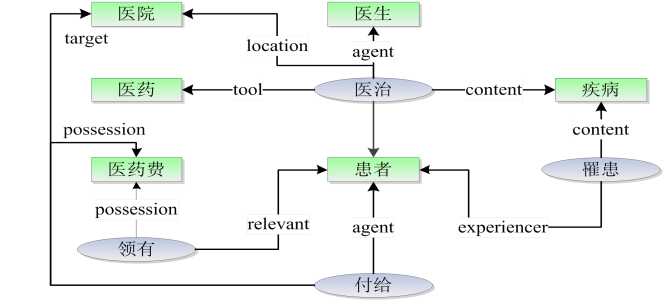
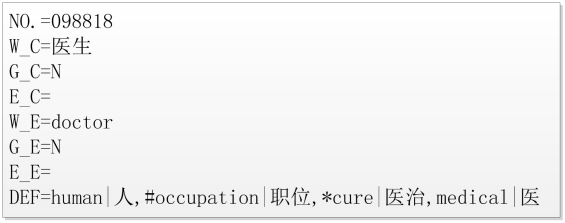
 

图1 HowNet中的网状知识体系结构 图2 HowNet中概念的定义方式

HowNet一共有超过1500义原，收录了近15万条概念记录，涵盖了绝大部分中文常用词语，本文将基于HowNet的词语进行情感词典的构建。

3.2 WordNet语义词典

WordNet[25]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文词典。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，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。

3.3 SentiWordNet情感词典

SentimentWordNet是Baccianella[10]等在语义词典WordNet基础上使用随机游走的图算法模型得到的情感词典。词典的每条记录都是一个WordNet的Synset，并且每个Synset都计算出了褒义（Pos(s)）贬义（Neg(s)）以及中性（Obj(s)）强度值，本文就是利用SentimentWordNet的情感强度值以及HowNet概念语义关系进行计算得到中文词语的情感强度值。SentimentWordNet共有117,000多 Synsets，192,493英文单词。

4 基于语义关系的情感词典构建方法

为了能够从语言中自动识别出观点和情感等主观性信息，需要建立计算机可以识别的从语言到情感值的映射关系。如果能够将某些语义相对稳定常用语言表达的褒贬情感色彩固化下来，可极大方便对其他复杂语言现象的情感分析。最基本的语言单位是词语，词语的语义相对较固定，情感倾向也容易确定，因此建立情感词典就成了情感分析基础性的工作。情感分析在各种语言上发展是不均衡的，目前主要的工作都是在英文开展的，中文的研究目前还处于起步阶段，并且常常要借助于英文已取得的研究成果基础上进行。如果能够将英文情感词典的研究成果转化为中文资源，不但可以利用语言之间的语义对应关系减少词典的歧义，使情感词典更加可靠，还可以直接将英文中对情感强度的计算直接转化为中文词语的情感强度计算，减少了计算开支。本研究正是基于这种动机而展开的。

HowNet对义原和概念进行了英汉双语标注，具有了英汉双语词典的功能，但是英文词语和中文词语都存在一词多义现象，不同语义所表达的情感倾向也不同，仅仅根据英汉之间对应关系从英文情感词典直接得到中文词语情感值也会存在歧义现象。HowNet中概念是靠义原进行描述的，具有语义关系，因此可以利用这种语义关系对中文词语的情感值进行“消歧”。总体来说，解决方案如图3框架所示。

构建中文情感词典框架包括义原和词语抽取及语义分析、义原和词语情感值查询与计算以及词语的情感值计算三个基本过程。通过前面对HowNet、WordNet和SentiWordNet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，将需要解决的问题描述如下：



图2基于语义关系的情感词典解决方案

1. 词语和义原抽取及语义分析：从HowNet中抽取中文词语、义原W\_C和相应的英文词语W\_E，获取词语的属性定义DEF。HowNet中的W\_C是通过DEF进行定义的，DEF是由多个义原构成，通过分析DEF的语义可以得到每个义原在DEF中的语义角色。根据义原的语义角色分析义原在情感倾向中作用并设定计算系数λ。
2. 情感值的查询与计算：将义原与词语对应的英文词语W\_E在WordNet中查询得到所有的Synset，从SentiWordNet查询Synset对应极性值PosScore和NegScore，将Synset集合并计算义原和词语的情感值。
3. 词语情感值计算：根据DEF语义关系，使用义原的英文情感值以及系数结合词语的英文情感值，计算中文词语W\_C的极性值PosScore(W\_C)和NegScore(W\_C)。

4.1 词语抽取和义原抽取及语义分析

词语抽取主要是从HowNet词典中抽取词语（W\_C）和属性定义（DEF）并对DEF进行分析。DEF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。情感词语抽取处理流程如图3-2所示。

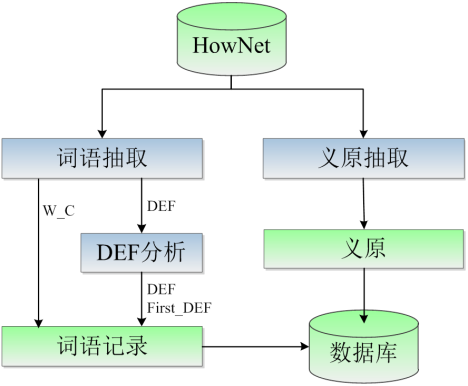


图3 词语和义原抽取处理流程

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。

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。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）”，表示DEF中包含，因此我们抽取的记录中包含了义原及其在词语中的语义关系。

4.2情感值的查询与计算

HowNet中词语附带了对应的英文，因此有的可以直接将抽取到的英文词语（W\_E）、英文词性（G\_E）直接送入英文情感词典查询其情感值，但是也要注意到，大部分的英文不是一个单词，因此无法直接得到情感值，而且一个单词的情况下由于词语的多义性，也无法获得唯一的情感值，因此需要进行“消歧”；其实HowNet中词语是由其属性DEF定义的，DEF是由多个义原按照一定的语义关系组合而成的，词语的倾向性可以看作是由义原的倾向性按照一定的规律组合而成的。因此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，一方面可以获得直接查询无法获得情感值的词语，另外一方面也可以通过DEF情感值的结算进行修正并消歧。

4.2.1 词语倾向性值查询与计算

从HowNet中抽取到的词语记录中包含有词语对应的英文词语（W\_E）以及英文词性（G\_E），因此可以直接映射到英文词典中进行查询，因为WordNet英文词典是以词义（sense）来记录的，并且sense的是以同一词义的词集Synset表示。因此通过查询可以得到词语W\_E所有的sense，将每个sense映射到SentiWordNet就可以得到对应的情感值。

WordNet中Sense索引文件中记录了词典中所有的Sense。在索引文件中，每一行记录表示一个Sense。每一行索引的格式如图4所示。



图4 Sense索引文件组织格式

在将WordNet中的Synset与SentiWordNet映射时，需要关注Sense索引结构中的synset\_offset字段，该字段由8个10进制整数组成，表示WordNet中包含于相应Sense的Synset的编号。

SentiWordNet中记录项组织结构如图5所示，其中POS表示Sense中Synset的词语类型，ID表示Synset的编号，PosScore和NegScore表示Sense中的Synset的极性值，SynsetTerms表示Sense中包含的Synset条目，Gloss，表示该Sense中Synset的注释。



图附录B-2　SentiWordNet的记录格式

WordNet与SentiWordNet中词语的映射主要是通过获取WordNet中词语的Synset编号（即synset\_offset），并将Synset编号与SentiWordNet中词语的ID编号进行比对，如果两者相等，则完成映射，获取相应词语的情感极性值等信息。

4.2.2 义原倾向性值查询与计算

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。

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） | 819 |
| 事物（Entity） | 154 |
| 属性（Attribute） | 248 |
| 属性值（Attribute Value） | 892 |
| 数量（aValue） | 3 |
| 数量值（qValue） | 13 |
| Secondary Feature | 121 |

* 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，并且事件类中的义原有很多在DEF描述中可以引起情感倾向值的变化，比如“DoNot|不做，lose|失去”等，因此我们标注了819个事件类义原的在情感值计算中的语义角色，并用系数λ来表示。
* 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。

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中进行查询，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中查询得到英文词语对应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）。

（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。

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。因此设计实现了基于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）。

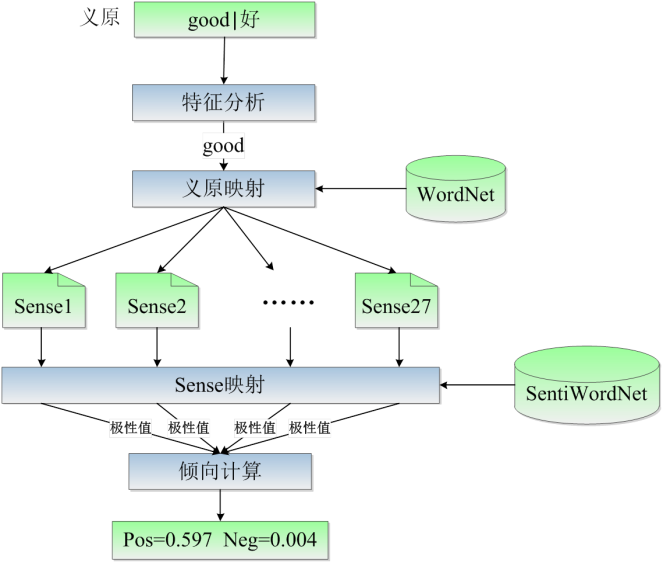
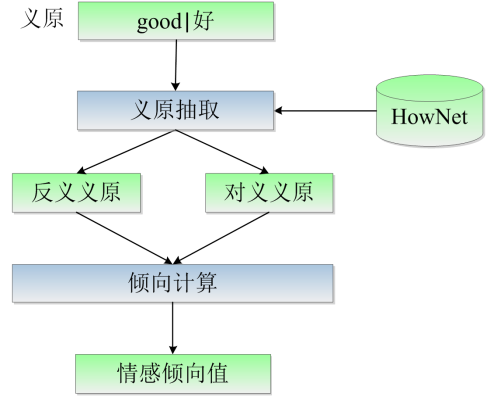
 

图6义原倾向性值计算 图7 基于对义和反义关系的义原倾向计算

4.3 词语倾向性值计算

在4.2.1通过查询可以获得部分词语的情感倾向值，有些词语由于是多义的，情感值可能有几个，因此需要根据词语DEF描述中义原情感倾向性值进行计算。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）

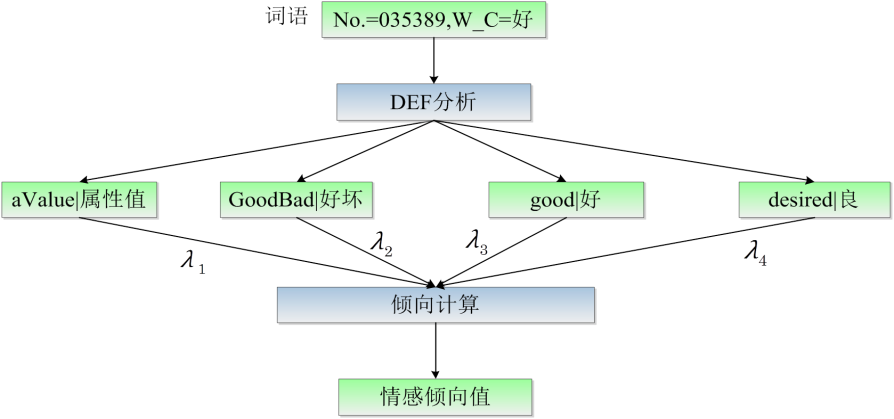


图8 词语倾向计算过程

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，这些都已经在抽取部分和义原情感值计算部分记录下来。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。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）

而对于已经通过查询得到情感值的词语而言，可以在多个英文词义sense对应的情感值中取最接近DEF分析计算得到的情感值的，然后加和平均，计算公式为：

(5)

5 实验及结果分析

情感词典的实验评测有两种，一种是与人工编辑的或者其他可靠性较高的词典进行对比，另外就是将词典应用到情感分析的其他任务上观察性能的提升，本实验使用第一种方法，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 | 用于评测本文方法性能 |

情感分析（包括情感词典构建）研究通常采用准确率（）、召回率（）以及值作为评测标准。在许多应用中，通常会为了提高准确率而适当牺牲一些召回率。

设表示分类器判断为正向，且与人工标注结果一致的样本数；表示分类器判断为负向，且与人工标注结果一致的样本数；表示分类器判断为正向的样本数；表示分类器判断为负向的样本数；表示人工标注为正向的样本数；表示人工标注为负向的样本数。

准确率计算公式为：

 　　 （6）

召回率计算公式为：

 　　 （7）

值计算公式为：

 　　 （8）

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449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）。在实验评测时，由于基准词典只是具有褒贬二值倾向的，因此需要将计算出的情感值转换为词语的褒贬倾向，如公式（5）所示，并且对于具有多个词性的词语，为了便于评测，只取情感值最大的作为该词的情感值，这是很简单的处理，不能完整反映出所构造词典的优点。

（9）

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。

在使用公式（1）、（2）和（3）进行评测时，采用HowNet发布的评价词词典作为人工标注的评测标准。

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，设计了跨语言的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。本文设计生成的情感词典记录了词语的语义关系、情感极性值等多种信息，对促进舆情分析等情感分析应用具有重要意义。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

**参考文献**

[1] Yuan S, Wang J, van der Meer M. Adaptive Keywords Extraction with Contextual Bandits for Advertising on Parked Domains[J]. CoRR. 2013, abs/1307.3573.

[2] 张辉，李国辉，贾立，等. 一种基于TF·IEF模型的在线新闻事件探测方法[J].国防科技大学学报. 2013, 35(3): 55-60.

[3] 健刘，绮李，刘宝宏，等. 基于话题模型的专家发现方法[J]. 国防科技大学学报. 2013, 35(2): 127-131.

[4] Liu B. Sentiment Analysis and Opinion Mining[M]. Morgan & Claypool Publishers, 2012.

[5] 黄萱菁，张奇，吴苑斌. 文本情感倾向分析[J]. 中文信息学报. 2011, 25(6): 118-126.

[6] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报. 2010, 21(8): 1834-1848.

[7] Stone P J, Dunphy D C, Smith M S. The General Inquirer: A Computer Approach to Content Analysis.[M]. 1966.

[8] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[C]. Association for Computational Linguistics, 2005.

[9] Taboada M, Grieve J. Analyzing appraisal automatically[C]. 2004.

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

[11] Agerri R, Garc I A-Serrano A. Q-WordNet: Extracting Polarity from WordNet Senses.[C]. 2010.

[12] Hatzivassiloglou V, Mckeown K. Predicting the Semantic Orientation of Adjectives[C]. 1997.

[13] Wiebe J. Learning Subjective Adjectives from Corpora[C]. AAAI Press, 2000.

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

[15] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientations of Adjectives[C]. 2004.

[16] 刘群，李素建. 基于《知网》的词汇语义相似度计算[Z]. 中国台北: 2002.

[17] 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(1): 14-20.

[18] 黄硕，周延泉. 基于知网和同义词词林的词汇语义倾向计算[J]. 软件. 2013, 34(2): 73-74, 94.

[19] Takamura H, Inui T, Okumura M. Extracting semantic orientations of words using spin model[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005.

[20] Duh K, Fujino A, Nagata M. Is machine translation ripe for cross-lingual sentiment classification?[C]. 2011.

[21] Li F, Pan S J, Jin O, et al. Cross-domain co-extraction of sentiment and topic lexicons[C]. Association for Computational Linguistics, 2012.

[22] 知网HowNet评价词词典2013/2013-07-25[2013-08-15].

[23] 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.

[24] 情感词汇本体库2013/2013-07-30.

[25] Fellbaum C. WordNet: An Electronic Lexical Database[M]. MIT Press, 1998.

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