基于 Word2Vec 的情感词典自动构建与优化

杨小平 张中夏 王 良 张永俊 马奇凤 吴佳楠 张 悦
(中国人民大学信息学院 北京 100872)

摘 要 情感词典的构建是文本挖掘领域中重要的基础性工作。近几年,情感词典的极性标注从二元褒贬标注向多元情绪标注发展,词典的领域特性也日趋明显。但是情感类别的手工标注不但费时费力,而且情感强度难以得到准确量化,同时对领域性的过分关注也大大限制了情感词典的适用性[1]。通过神经网络语言模型对大规模中文语料进行统计训练,并在此基础上提出了基于转换约束集的多维情感词典自动构建方法;然后研究了基于词分布密度的感情色彩消歧方法,对兼具褒贬意味词语的感情极性进行区分和识别,并分别计算两种感情色彩下的情感类别与强度;最后提出基于多个语义资源的全局优化方案,得到包含10种情绪标注的多维汉语情感词典 SentiRuc。实验证实该词典[1)]在类别标注检验、强度标注检验、情感消歧效果及情感分类任务中均具有良好的效果,其中的情感强度检验证实该词典具有极强的情感语义描述力。

关键词 情感分析,多元情感分类,神经网络语言模型,情感消歧,情感强度优化框架
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.008

Automatic Construction and Optimization of Sentiment Lexicon Based on Word2Vec

YANG Xiao-ping ZHANG Zhong-xia WANG Liang ZHANG Yong-jun MA Qi-feng WU Jia-nan ZHANG Yue
(School of Information, Renmin University of China, Beijing 100872, China)

Abstract The construction of sentiment lexicon plays an important role in text mining. In recent years, the lexicon annotating format gradually evolves from binary annotation to multiple annotation, and sentiment lexicons of a single specific domain have caught more and more attentions of researchers. However, manual annotation costs too much labor work and time, and it is also difficult to get accurate quantification of emotional intensity. Besides, the excessive emphasis on one specific field has greatly limited the applicability of domain sentiment lexicons[1]. This paper implemented statistical training for large-scale Chinese corpus through neural network language model, and proposed an automatic method of constructing a multidimensional sentiment lexicon based on constraints of Euclidean distance group. In order to distinguish the sentiment polarities of those words which may express either positive or negative meanings in different contexts, we further presented a sentiment disambiguation algorithm to increase the flexibility of our lexicon. Lastly, we presented a global optimization framework that provides a unified way to combine several human-annotated resources for learning our 10-dimensional sentiment lexicon SentiRuc. Experiments show the superior performance of SentiRuc lexicon in category labeling test, intensity labeling test and sentiment classification tasks. It is worth mentioning that in intensity label test, SentiRuc outperforms the second place by 23%.

Keywords Sentiment analysis, Multivariate sentiment classification, Neural network language model, Sentiment disambiguation, Optimization framework of sentiment intensity

基于互联网上文本大数据的意见挖掘和情感计算等任务都依赖于全面而高质量的情感词典作为有监督分类的语义特征[2],因此情感词典的构建是一项重要的基础性研究工作。

情感词典的基础形态是褒贬二元词典,代表成果有意大利信息科技研究所的 Sentiwordnet[3,4],台湾大学的通用中文情感词典 NTUSD[5],中科院的中文情感程度词典,以及英文的 Xsimilarity 等。近几年出现了如大连理工情感本体[6]等多元情感词典,它们采用连续数值标注来定量表示某情感极性上的强度。受具体分类任务的驱动,近年研究都通过要素抽取技术形成〈评价对象,评价词〉的二元组,然后通过二元组之间的语义关联计算出情感倾向与强度,但其领域局限性太大且二元组规模极难控制,从而导致情感特征易稀疏化。面对互联网上复杂的数据源和庞大的数据信息,同时随着情感分析任务范围的扩大,领域情感词典的局限性日益显现,研究

1) 该词典已发布在中国人民大学信息系统工程实验室网站(isel.ruc.edu.cn)。

到稿日期:2015-09-13 返修日期:2015-11-13 本文受国家自然科学基金(71271209),北京市自然科学基金(4132067),教育部人文社会科学青年基金(11YJC630268),数字出版技术国家重点实验室开放课题资助。

杨小平(1956-),男,博士,教授,博士生导师,主要研究方向为数据分析、管理信息系统,E-mail:yang@ruc.edu.cn;王 良(1963-),男,博士,副教授,主要研究方向为系统软件、数据库工程,E-mail:wangliang@ruc.edu.cn(通信作者)。

通用的情感挖掘方法来构建与丰富情感语义资源，从而提高情感词典的适应性和自动化程度，是现在亟需解决的问题。

本文旨在研究如何通过海量语料的统计分析实现多维情感词典的自动构建与优化。本文主要工作如下：首先利用 Word2Vec 工具从海量语料中提取词向量；然后研究情感类别划分并选取种子词；接着提取情感词集合并计算情感词到种子词的语义距离；随后定义将语义距离转化为词语相似度应遵循的约束条件，将距离值转化为情感强度；最后借助多个语义资源，从多个角度出发利用构建的全局误差函数对情感词典进行评价，以此为标准对参数组进行优化。

本文第 1 节总结了相关研究成果；第 2 节提出了基于 Word2Vec 的情感词典自动构建模型；第 3 节介绍了基于多黄金标准的情感强度全局优化策略；第 4 节通过多组实验验证了情感词典构建模型和优化策略的有效性；最后对研究工作进行了总结，并对可改进之处进行了探讨。

1　相关工作

情感标注的前提是划分情感类别。人的情感类别复杂多样，分类标准也各不相同。早在 1957 年，心理学家 Osgood[8] 就将人类情感总结为强与弱、好与坏、主动与被动 3 个方面。徐琳宏等人于 2008 年发布了大连理工情感本体[6]，该本体将情感分为乐、好、怒、哀、恶、惧、惊 7 类。全昌勤等人构建了博客情感语料库[9]，其中共提取了 8 类情绪，并通过矩阵空间的方式运用支持向量机实现情感分类。以上研究对情感词典的构建起到了巨大的推动作用，但现有的情感类别划分普遍存在着类别不对称现象，例如“惧怕”、“愤怒”等类别没有与之直接相对的情感，这在有监督分类中对特征提取和选择过程造成了不便。另外，情感类别之间也存在着一定程度的耦合现象，因此需要研究更适合计算语言学的情感类别划分方法。

情感词典不仅要对词语进行定性情感标注，还要定量标示出情感强度，主要方法有专家标注法和扩展法。专家标注法是一种最直接的方法，即由专家来标注每个情感词的情感极性，如已有的 WordNet[10]、General Inquirer(GI)词典[11]和知网情感分析用词语集[12]等。这种方法需要大量的人工标注工作，效率较低且易受到主观性的影响，强度标注的细粒度与精确度也得不到保证。因此，基于词典的扩展法得到了广泛采用，其核心思路是在专家标注的基础上定义基准词，借助语义距离、词典标注等计算情感强度。何凤英等[13]以 HowNet 为基础，按搜索引擎返回结果数确定种子词，并利用语义相似度公式计算情感词与种子集中每个词的语义倾向相似度，得出情感词的情感权值。李荣军等[14]引入 PageRank 模型来研究词汇的情感极性，利用 HowNet 计算到情感词之间的连接权重从而解决待测情感词的极性判别问题，但是基准情感词的选取具有较强的主观性或应用局限性，对此可通过无监督算法从其他语义资源中提取出情感词典。Francesco[15]从在线评论抽取出(侧面，评价)词对构成词条融合图，利用领域知识将词对整合为评论的总体情感倾向。Raghava[16]把情感表达概括为一个四元模糊集，通过计算各模糊集之间的隶属度关系得出这些情感词间模糊关系的强弱来确定情感强度。Turney[17]提出了基于情感词组的语义分类方法，通过制定好的一些模板来提取符合模式的主观词组模板，并计算词语之间的点互信息，从而确定词语的情感倾向。以上无监督方法的思想为后人提供了很多经验和帮助，但对词间关系的选择、识别与抽取方法都有较大依赖，标注准确率仍有进一步提高的空间。

因此，对情感词典条目集合以及强度标注的优化成为重要的补充研究点。Lu Chen 等人[18]通过构建两个词或词组间的极性平方误差函数来判断它们是否同为情感词，进而实现对情感词典的迭代扩充。H. Wang[1]以及 Y. Jo[19]等人都将基于领域的情感极性标注作为情感计算任务的副产品，但并没有对这些标注的质量进行评估。Turney[17]比对词条与种子词间的共现参数寻求合理的强度标注方法。还有一些工作尝试将同义关系或反义关系引入评估框架中来对强度进行优化[20,21]。和文中工作相比，以上研究依赖的语义资源较为单一，且未考虑到有些词语在不同情境下可能具有不同的感情色彩。

基于以上几点考虑，本文提出了无监督的基于 Word2Vec 的多维情感词典自动构建模型与全局优化框架。主要贡献在于：

1)提出了“五对十维”情感类别划分法，使得情感语言学特征更适合于计算；

2)定义了词向量空间中词语距离到情感相似度的转换约束集，提出基于 Word2Vec 的情感词典自动构建模型；

3)提出了基于多个语义资源的情感强度全局优化方案，使 SentiRuc 在同义关系、反义关系、句子级情感等多方面具有更强的语义描述力。

2　SentiRuc 情感词典的自动构建模型

本章提出了情感类别的“五对十维”划分法；并定义了高维空间中多个欧氏距离同时作用时计算相似度的约束条件集合，自动标注了多维情感词典 SentiRuc；最后针对词典中的多情感词汇研究了基于支持向量机的情感消歧。

2.1　情感类别划分及统计语言模型工具 Word2Vec

传统的褒贬二元情感标注已经不能应对情感计算任务种类的发展与应用范围的扩大，因此近几年出现了多情绪标注的多维情感词典，首要工作就是对情感类别进行划分。第 1 节已讨论过心理学和社会学等领域对人类情感划分的研究。在前人成果的基础上，本文提出了同时适用于心理学成果、语言学理论和计算学特征的“五对十维”情感划分方法，将人类情绪划分为快乐-悲哀、喜爱-厌恶、信心-意外、褒扬-贬斥、感激-愤怒 10 种基本情绪，其中前 4 对较易于理解。文献[26]的实验证明了感激和愤怒的对立，该文指出这两种情绪“除了他人负责性认知评价差异不显著，其他认知维度上均存在显著性差异”。在之后的研究中，这 10 个情感词作为情感词典的初始种子词参与运算。

词汇蕴含的内在语义极其丰富，以至于很难用几个简单特征值将其描述得完全且准确。统计语言模型可以较好地解决这一问题。借助于互联网上的海量真实语料，可以通过神经网络语言模型把词汇映射到高维连续空间中。Word2Vec[23]就是这样的开源工具，它可以从海量语料库中

学习出一套高维的词向量，有实验证实这些词向量之间具有极为良好的线性语义关系[24]。因此，本文通过 Word2Vec 将海量中文语料(搜狗新闻语料，3.17GB)中的词语映射到高维连续向量空间中，研究了词间距离对词语相似度的影响。

2.2 距离转换为相似度的约束条件集

在词向量的基础上，本文将以上 10 种情绪作为基本情感，通过计算词语 W 与这 10 种情感的距离来确定 W 在每一维上的情感强度。但是“快乐”类中有很多表达喜悦之情的词语，很难从中确定出唯一的种子词。为了最大限度降低由种子词选取的主观性造成的计算偏差，本文在每个情感类上都选取了与初始种子词欧氏距离最近的 50 个词，并人工选择出实际含义与初始种子词最相近的一些词，与初始种子词一起作为该类情感的种子集。例如，在计算“悲喜交加”的“快乐”强度时，就是计算“悲喜交加”和“快乐”类种子词在词向量空间中的欧氏距离，将这些距离的均值作为“悲喜交加”到“快乐”的距离。对于任一词语 W，都可以按这种方法得到如下的距离向量 $Dis(W)$，各个维度分别表示 W 到快乐、信心、感激、褒扬、喜爱、悲哀、意外、愤怒、贬斥、厌恶 10 个类别的距离。

$Dis(W)=(6.64,19.55,24.90,14.18,22.30,7.95,24.65,14.67,31.91,15.01)$

由于一个词语一般只具有一种或少数几种较强的情绪意味[6]，因此只保留 $Dis(W)$ 中距离最小的部分维度作为有效距离，较大的距离值则会被抛弃，这可以解释为相似度较低的情绪类别被剔除。例如，假设 T 取 16，则将只保留 $Dis(W)$ 的快乐 6.64 和悲哀 7.95 两个有效距离，这两维距离的累加和为 14.6，尚未超过阈值 T，这两种情感就是 W 蕴含的主要情感。

高维词向量具有优良的语义空间聚类特性[24]，因此 W 到某一情感类的欧氏距离越远，它在这一个情感维度上的强度值就越低，据此可以计算出 W 在各个情感类别上的强度值。但是有效距离可能多于一个，在多种情感类别上的有效欧氏距离同时存在并作用的情况下，应研究这些有效距离的分布对相似度产生的实际影响。在将有效距离转换为相似度的过程中，本文定义了 3 种计算约束。

约束 1 分散度约束。一个词语的各维情感强度 $Senti(W)[i]$ 与该词的有效距离个数 $Count(Dis(W))$ 负相关；

约束 2 自体约束。某维情感强度 $Senti(W)[i]$ 应与该维距离 $Dis(W)[i]$ 负相关；

约束 3 全局对比度约束。某维情感强度 $Senti(W)[i]$ 应与该维距离与全部距离均值之比 $Dis(W)[i]/Avg(Dis(W)[i])$ 负相关。

综合以上约束，得出由词语 W 的距离向量 $Dis(W)$ 生成情感向量 $Senti(W)$ 的公式，如式(1)所示：

$$Senti(W)[i]=Diverge\cdot Self\cdot Contrast \tag{1}$$

式(1)表示，$Senti(W)[i]$ 由 3 个因子共同参与运算得出，分别是分散度约束因子 $Diverge$、自体约束因子 $Self$、全局对比度约束因子 $Contrast$，这 3 个因子可分别表述为式(2)—式(4)：

$$Diverge(Dis(W))=\frac{C_0}{\alpha*Count(Dis(W)+C_1)} \tag{2}$$

$$Self(Dis(W))=(\frac{C_2}{Dis(W)[i]+C_2})^{\beta} \tag{3}$$

$$Contrast(Dis(W))=(\frac{Avg(Dis(W))}{Dis(W)[i]})^{\gamma} \tag{4}$$

在式(2)—式(4)中，α,β,γ 这 3 个参数决定了每种约束在实际语言系统中的实际作用强度。其中 $Count(Dis(W))$ 表示有效距离个数，$Avg(Dis(W))$ 表示有效距离均值。C_0，C_1 和 C_2 是常数，可根据具体的语料和词向量进行适当调整。通过第 3 节的优化框架可以训练出最优参数组。

综上，在 3.17GB 的真实语料上采用 Word2Vec 工具训练出一套词向量，再根据词语到 10 种情感类别的距离，即可初步计算得到各词的情感向量，初步生成 SentiRuc 情感词典。

2.3 基于词分布密度的感情色彩消歧

一些情感词在不同的上下文中可能表达出不同的情感极性，因此有必要对常见的多情感词汇进行情感消歧。林鸿飞等[25]指出“情感消歧与词义消歧存在区别”，因为某词语的义项与其感情色彩并无直接关联。

本文将台湾大学的 NTUSD 词典、知网情感分析用词语集(HowNet)、大连理工大学开发的情感本体(DUT 本体)等语义资源中蕴含的条目做了综合和筛选，构成 SentiRuc 的情感词库，其中共包含 14250 个情感词。使用机器与人工混合的方法从中筛选多情感词汇，对包含多情感词汇的句子中的词语情感倾向进行手工标注后，研究了基于支持向量机的感情色彩消歧模型。

首先，目前并没有较好的多情感词汇的自动筛选方法。本文试图自动抽取出《同义词词林》和《小学生同义词词典》中出现在多组同义词中的词汇构成多情感词汇候选集 S。虽然 S 的覆盖准确率较高，但是由于实际语料中的修辞手法和引申含义繁多，导致很多词语的感情色彩出现迁移，因此 S 的覆盖全面性很差，例如“幼稚”、“龙飞凤舞”等词语的褒贬两种使用方式均频繁出现，但却未包含在 S 中。因此，用手工方式从上述两个同义词典中甄选出了 148 个褒义与贬义均较常出现的多情感词语构成多情感词集合 $S_{multiSenti}$。

然后从真实语料中筛选出包含 $S_{multiSenti}$ 中元素的全部句子，并标注了这些句子中多情感词汇的感情色彩类别，共对 148 个词汇的 113694 个样例进行了标注。最后从上下文中提取出每个词语在两类句子中的词分布密度作为特征，采用 SVM 模型进行了感情色彩的二元分类实验。某词语在褒贬两类文本的词分布密度可由式(5)和式(6)计算得出：

$$\rho c_{W^{positive}}=\frac{Count(C_W positive)}{Count(C_W)} \tag{5}$$

$$\rho c_{W^{negative}}=\frac{Count(C_W negative)}{Count(C_W)} \tag{6}$$

其中，$Count(C_W positive)$ 表示上下文词语 C_W 在 W 褒义样例中的出现次数，$Count(C_W negative)$ 表示 C_W 在 W 贬义样例中的出现次数，$Count(C_W)$ 表示 C_W 在所有 W 样例中的出现次数。

这样，采用基于词分布密度的感情色彩消歧方法将某多情感词 W 消歧后作为两种互不干扰的独立的情形对待，按照

$W_{positive}$ 和 $W_{negative}$ 两个词语重新分词、再次训练词向量并根据转换约束集生成情感标注，进而分别计算出 $W_{positive}$ 和 $W_{negative}$ 的情感类别和情感强度。

3　全局优化框架

针对第 2 节提出的 3 种面向距离的约束，本节建立了统一形式的评估函数来进一步研究各种约束所起作用的不同。本文收集了哈工大同义词词林、小学生同义词词典、小学生反义词词典以及 NLPCC2013 与 NLPCC2014 的情感句标注数据集，这些资源都是由人工标注构建的，可视为黄金标准。将 SentiRuc 与这些标注资源之间的偏差作为评估函数，我们的目标就是寻找使得偏差最小的一组情感强度生成参数。

3.1　同义关系优化

设 W_1 和 W_2 是同义词词典中标注的同义词对，那么在 SentiRuc 中，W_1 和 W_2 所拥有的情感维度应一致，且在这些维度上的情感强度也应尽量一致。

依照该标准，通过计算同义词对在多维情感词典中对应维度上的情感强度差值构建同义误差函数 $f_1(D)$。具体函数表示如式(7)所示：

$$f_1(D)=\frac{1}{|D_1|}\sum_{D_1,i}\frac{Senti(W_1)[i]-Senti(W_2)[i]}{Senti(W_1)[i]+Senti(W_2)[i]} \tag{7}$$

其中，$W_1,W_2\in D_1$，且为同义词词对，$|D_1|$ 表示同义词词对的个数，W_1,W_2 表示同义词词对中的两个词，$f_1(D)$ 表示同义词词对在词典中对应情感维度上的情感强度之差与情感强度之和的比值的平均值。

3.2　反义关系优化

设 W_1,W_2 是反义词词典中标注的反义词对，那么在 SentiRuc 中，W_1,W_2 所拥有的情感维度应反向对应，且在这些维度上的情感强度也应尽量一致。

依照该标准，通过计算反义词对在多维情感词典中对应维度上的情感强度差值构建反义误差函数 $f_2(D)$，如式(8)所示。

$$f_2(D)=\frac{1}{|D_2|}\sum_{D_2,i,j}\frac{Senti(W_1)[i]-Senti(W_2)[j]}{Senti(W_1)[i]+Senti(W_2)[j]} \tag{8}$$

其中，$W_1,W_2\in D_2$，且为反义词词对，i,j 为相对维度，其中 $|D_2|$ 表示反义词词组的个数，W_1,W_2 表示反义词对中的两个词，$f_2(D)$ 表示反义词词组在词典中相对情感维度上的情感强度之差与情感强度之和的比值的平均值。

3.3　句子级描述力的优化

若 SentiRuc 具有足够准确的情感强度表示，那么使用该情感词典进行情感结果的计算应当比使用其他词典得出更加符合人类判断的计算结果。

依照该标准，从 NLPCC2013 与 NLPCC2014 情感分类任务提供的数据集中选取了 6000 条句子进行了十维情感的主情感和次情感标注，只包含一种情感的，将次情感强度标注为 0。然后使用多特征融合的情感分类模型，采用不同参数组生成的 SentiRuc 进行情感分类，计算分类结果与情感标注的杰卡德相似系数，构建分类误差函数 $f_3(D)$，如式(9)所示：

$$f_3(D)=\frac{1}{D_3}\sum_{D_3}Jaccard(Label(d_3),Sentence(d_3)) \tag{9}$$

其中，$|D_3|$ 表示数据集中的例句个数，$Label(d_i)$ 表示某句子的标注情感，$Sentence(d_i)$ 表示采用 SentiRuc 计算得到的某句子的情感，$f_3(D)$ 表示使用 SentiRuc 对各例句进行十元情感分类结果与标注结果的杰卡德相似系数的平均值。

3.4　全局误差函数

将以上 3 个基于黄金标准的评估部分联合起来，即得到基于多黄金标准的全局优化框架和全局误差函数，如图 1 所示。

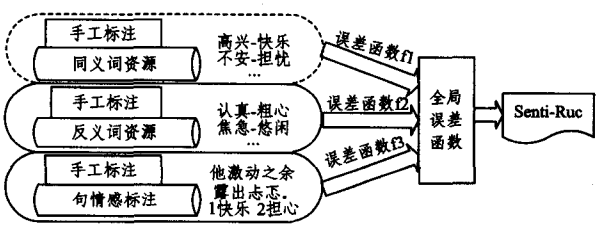

图 1　基于多个黄金标准的全局优化框架

全局误差函数如式(10)所示：

$$f(D)=f_1(D)+f_2(D)+f_3(D) \tag{10}$$

通过固定变量法调节 α,β,γ 参数组，找出全局最小误差，这样就能通过观察全局误差而逐步筛选出最优的参数组合 arg Min($f(D)$)，从而使 SentiRuc 情感词典具有良好的情感语义描述能力。考虑到多情感词汇仅占情感词典容量的 1.04%(148 个)，以上各项评价指标并未包含这些多情感词汇。

4　实验验证

主要从两个角度进行实验评估。一种是词典生成过程的合理性评估，其中包括约束集合理性评估(实验 1)以及情感消歧实用性评估(实验 2)；另一种是 SentiRuc 的标注质量评估，其中包括类别标注与强度标注的比较式评估(实验 3)以及在情感分类任务中的表现(实验 4)。在这两个角度上的分析中，都将 SentiRuc 与台湾大学 NTUSD 词典、HowNet 情感库、大连理工情感本体等主流汉语情感词典进行了对比。

在本节所有实验中，都将有效距离累加阈值 T 设为距离向量中各维距离的均值，这样绝大多数词语最后被保留一种或两种主要情感类别。将式(2)中的参数 C_0 设为情感类别个数 10，参数 C_1 设为 8，表示除了绝大多数词语最大剩余距离个数 2 之外的维数；将式(3)中的 C_2 设为 20.08，表示 SentiRuc 所收录词条在词向量空间中两两之间的欧氏距离均值。词向量维数设为 60，实验表明更多或更少维数都会使词典强度标注评估结果更差。每个情感类别的种子词数量选为 10 个。

使用搜狗互联网新闻语料作为训练集，语料大小为3.17GB，使用中科院分词工具 ICTCLAS 5.0 共识别出约 8.3 亿个词，其中互不重复的词有 1104914 个。由于语言模型需要基于词汇的真实使用情境做出计算，因此除了进行分词外，没有对该语料进行去停用词、删标点符号等任何预处理，以保证每一个语言模型案例都是绝对真实的中文信息表达序列。

4.1　生成参数寻优实验

2.2 节叙述了基于转换约束集的情感强度生成方法；第 3 节构造了全局误差函数来对情感强度生成中的相关参数进行

优化。在实验过程中，首先将各参数初始值设为 1，将其作为基准实验；然后分别去除一个参数，观察此时误差函数的变化，以验证各个参数是否会对情感强度的计算起到正向调节作用；最后通过固定变量法逐步得到各个参数的最优值。

各参数组对应的误差函数值以及最优参数组如表 1 所列。

表 1 生成参数调优实验

	α	β	γ	f(D)
基准实验	1	1	1	2.190
逐个去除参数	0	1	1	2.297
	1	0	1	2.594
	1	1	0	2.534
参数寻优实验	1.875	1	1	1.910
	1.875	1.075	1	1.882
	1.875	1.075	1.145	1.839

通过表 1 可以看出，去掉任一约束条件后，情感强度的全局误差均有显著增加，这说明这些约束对情感词典的语义描述均产生积极影响。通过多组对比实验对各约束条件的影响系数进行调节，得到最优生成参数组。

4.2 情感消歧效果的评估

在进行感情色彩消歧时，从 3.17GB 的搜狗新闻语料中收集了包含多情感词汇的 113694 条语句，并对这些句子中出现的多情感词的感情色彩进行了人工标注：当多情感词汇 W 在该句中的情感意味为褒义时将其标注为 1，否则标注为 2。为了获得高质量的标注，8 位本专业研究生分别对 5 万多条语句做出了独立标注结果，每条句子均由 4 位标注者做出标注。对于有冲突的标注结果，经过小组讨论达成共识，最终对全部 113694 条句子中的多情感词做出了褒贬标注。

采用基于词分布密度的感情色彩消歧算法进行了实验。基于 SVM 分类的十折交叉验证的实验结果如表 2 所列。

表 2 感情色彩消歧实验

	实验词语	案例个数	消歧准确率(%)
平均	全部 148 个词	113694	95.52
最高准确率	滋生	3095	98.71
	幼稚	1924	98.70
最低准确率	息事宁人	281	87.90
	萧规曹随	130	61.54

从表 2 中可以看到，在包含多情感词汇的全部 113694 条例句中，词语感情色彩的整体识别准确率达到 95.52%。识别准确率较低的词是“萧规曹随”和“息事宁人”，主要是由于这两个词语在语料中的出现频率较低，筛选出的训练样例个数有限导致的。从整体准确率可以看出，该消歧方法可有效区分不同感情色彩的使用情形。

4.3 SentiRuc 标注质量的评估

由于 SentiRuc 的情感极性分布和情感强度值是从 GB 量级的真实语料中统计得出的客观结果，因此其语义描述相比于人工标注更贴近现实语义。将现有的情感词典与同义词词典、反义词词典的契合度作了分析评估，从情感类别一致性(定性评价)和情感强度一致性(定量评估)上对几个现有情感词典进行了对比分析。情感类别一致性是指选定测试集中两个词语在 SentiRuc 的褒贬倾向标注符合同义词词林和反义词词典标注的百分比，情感强度一致性是指在符合同义/反义标注的情感极性上的情感强度差异，如式(11)、式(12)所示。其中 D_{same} 与 D_{oppo} 分别代表同义/反义词对里同时具有情感类别标注的对应维数之和。

$$Value_{syn}=1-\frac{1}{|D_{same}|}\sum\frac{Senti(W_1)[d_{same}]-Senti(W_2)[d_{same}]}{Senti(W_1)[d_{same}]+Senti(W_2)[d_{same}]}\times 100\% \tag{11}$$

$$Value_{ant}=1-\frac{1}{|D_{oppo}|}\sum\frac{Senti(W_1)[d_{oppo}]-Senti(W_2)[d_{oppo}]}{Senti(W_1)[d_{oppo}]+Senti(W_2)[d_{oppo}]}\times 100\% \tag{12}$$

从作为测试标准的《同义词词林》和《小学生同义词词典》中抽取出 55265 组同义词对，从中选取出 2500 个不包含多情感词的同义词对作为测试集合 $S_{syn2500}$，从反义词词典中抽取出不包含多情感词汇的 1774 个反义词对作为测试集合 $S_{anti774}$。评价结果如表 3 和表 4 所列。

表 3 各词典标注的同义描述评估

词典列表	同义词对个数	类别一致数	类别一致率(%)	强度一致性(%)
NTUSD	2179	1902	87.29	无
HowNet	2500	2226	89.04	无
DUT 本体	2500	2211	88.44	70.89
SentiRuc	2500	2292	91.68	93.58

表 4 各词典标注的反义描述评估

词典列表	反义词对个数	类别一致数	类别一致率(%)	强度一致性(%)
NTUSD	1450	1218	84.00	无
HowNet	1772	1533	86.51	无
DUT 本体	1774	1515	85.40	67.55
SentiRuc	1774	1557	87.77	92.10

如表 3 和表 4 所列，SentiRuc 在同义和反义上的类别标注更接近人工语义资源，且在各词语情感强度为独立计算出的前提下，同义与反义词对在对应维度上的情感强度数值的一致性高达 93%和 92%，该数值远超预期，且高出 DUT 本体的标注一致性 20 余个百分点，这说明了转换约束集与式(1)的有效性，也说明与人工标注情感强度相比，SentiRuc 的情感强度自动标注具有更强的语义描述能力。

4.4 SentiRuc 在情感分类任务中的应用

首先，对选自 NLPCC 的 3100 条句子依照 SentiRuc 中的 10 种情感类别进行了主情感和次情感标注，如果某句子只包含一种情绪，则将该句的次情感强度标注为 0。通过抽取这些标注语句中的情感词特征、二元词性特征、三元词性特征，采用 SVM 进行了 10 类情感的多元分类实验。实验显示，将人工标注结果作为测试基准，情感类别细化到 10 种的十元分类实验准确率可达到 61.5%。由于实验选用的几个词典中情感类别的划分不尽相同，为了能对比验证 SentiRuc 情感强度标注的准确性和情感消歧在情感分类中的重要性，用褒贬二元分类将几个主流情感词典进行了比较。

从 NLPCC2013 与 NLPCC2014 情感分类任务提供的数据集中选取了 3100 条语句，另外也通过微博抓取了包含 148 个多情感词汇的 3700 条句子。4 位本专业研究生分别对这 6800 条句子做出了褒、贬标注，对于有冲突的标注结果，经过小组讨论最终达成共识。这些句子和另外 3200 条客观句一起，构成测试数据集。分别使用这几个通用情感词典，采用多

特征融合的 SVM 分类模型进行了二元分类实验，同时评估了消歧前后 SentiRuc 的表现。

本文使用准确率(Precision)、召回率(Recall)以及 F1 值(F1-measure)作为评价标准。计算方式为：

$$Precision=\frac{Result\_Correct}{Result\_Proposed}\times 100\% \tag{13}$$

$$Recall=\frac{Result\_Correct}{Result\_Labeled}\times 100\% \tag{14}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall}\times 100\% \tag{15}$$

Result_Correct 表示分类结果中与人工标注匹配的数目，*Result_Proposed* 表示分类结果中包含该情感的句子数，*Result_Labeled* 是人工标注的包含该情感的句子数。实验结果如表 5 所列。

表 5　基于不同词典的情感分类实验对比

褒义文本的分类结果			
词典	准确率	召回率	F1 值
NTUSD 词典	0.603	0.375	0.462
HowNet 情感词典	0.728	0.540	0.620
DUT 情感本体	0.721	0.552	0.593
消歧前 SentiRuc	0.737	0.586	0.653
消歧后 SentiRuc	0.774	0.673	0.720
贬义文本的分类结果			
词典	准确率	召回率	F1 值
NTUSD 词典	0.480	0.319	0.383
HowNet 情感词典	0.611	0.451	0.519
DUT 情感本体	0.572	0.445	0.501
消歧前 SentiRuc	0.628	0.466	0.535
消歧后 SentiRuc	0.663	0.586	0.622

从表 5 可以看出，在通用领域文本上的情感分类任务中，SentiRuc 的表现要明显优于其他词典。另外，由于实验数据中包含多情感词汇的句子在所有主观句中所占比例高达 54.41%，因此可以很明显地从消歧前后的 F 值观察出，消歧后 SentiRuc 的实验结果得到了显著提升，这证明了感情色彩消歧对情感词典标注的重要性。也正是由于多情感词汇的例句占比较高，使得情感分类的整体实验结果受到了一定影响，消歧后的 SentiRuc 在褒、贬两类文本上的二元分类 F 值分别为 0.720 和 0.622。而经过统计，实际上在不包含多情感词汇的 6300 条数据中，基于 SVM 的分类方法在褒、贬文本上的 F 值分别达到了 0.808 和 0.734。

结束语　本研究基于目前词典构建的主观性强、领域单一、二元组规模不易控制等缺点，研究了基于统计语言模型的多维情感词典自动构建框架。主要研究内容包括情感类别的划分、基于 Word2Vec 与转换约束集的情感强度计算、基于多黄金标准的全局优化框架以及多情感词汇的情感消歧等。4.4节表明该通用情感词典可以在通用领域数据集上取得较好的实验结果。另外在 4.3 节中，SentiRuc 的强度一致性评估与人工语义资源的一致性高达 92%，比人工标注词典高出 20 多个百分点，充分说明统计语言模型和词向量的语义特征描述方式是非常适用于情感语义表示的。

由于各个多维情感词典的情感类别划分不尽相同，且测试文本的多元情感标注需要消耗大量人力，因此本文通过将多维情感压缩映射到二元情感的方式比较了各词典在实际任务中的表现。未来我们拟针对该问题，研究多元标注与分类的质量评估方法。

此外，虽然 4.3 节验证了词向量在情感词典构建中的优异表现，但词向量的独有特点也对语义距离的抽取产生了一定影响。首先，统计语言模型高度依赖于内在语义与外在语法的契合度，因此如何理解、区分并利用词向量空间中的“相似词”和“相关词”及其与词向量生成模型的关系都是非常值得深入研究的问题。其次，词向量维数高达数十甚至上百，相似词语的向量仅在少数维度有显著差异，如何处理这种定性区分强但定量差异弱的表征方式有待进一步研究。针对该问题，我们也将进一步研究距离加权的统计语言模型，并尝试研究将各种向量运算用于语义距离估计的效果和可行性分析。

参 考 文 献

[1] WANG Hong-ning, LU Yue, ZHAI Cheng-xiang. Latent aspect rating analysis on review text data: a rating regression approach [C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010. Washington, DC, USA, 2010: 783-792.

[2] CHOI Y, CARDIE C. Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification[C]// Conference on Empirical Methods in Natural Language Processing. 2009: 590-598.

[3] ESULI A, SEBASTOAMO F. Sentiwordnet: a publicly available lexical resource for opinion mining[C]// Proceedings of LREC. Genoa-Italy: LREC, 2006: 417-422.

[4] BACCIANELLA S, ESULI A, SEBASTIANI F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]// International Conference on Language Resources and Evaluation, Lrec 2010. Valletta, Malta, 2010: 83-90.

[5] TANG Da-ta. National Taiwan University: simplified Chinese emotional dictionary [EB/OL]. [2013- 03-05]. http://www.datatang.com/data/11837.

[6] XU Lin-hong, LIN Hong-fei, PAN Yu, et al. Constructing the affective lexicon ontology [J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185. (in Chinese)
徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008，27(2)：180-185.

[7] NEVIAROUSKAYA A, PRENDINGER H, ISHIZUKA M. SentiFul: A Lexicon for Sentiment Analysis [J]. IEEE Transactions on Affective Computing, 2011, 2(1): 22-36.

[8] OSGOOD C E. The nature and measurement of meaning [J]. Psychological Bulletin, 1952, 49(3): 197-237.

[9] QUAN Chang-qin, REN Fu-ji. Construction of a blog emotion corpus for Chinese emotional expression analysi[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3, Association for Computational Linguistics. 2009: 1446-1454.

[10] FELLBAUM C, MILLER G. WordNet: An Electronic Lexical Database[M]. Bradford Book, 1998.

[11] General Inquirer (GI). Harvard University. [EB/OL]. [2012-04-25]. http://www.wjh.harvard.edu/~inquirer.

[12] 董振东. 知网情感分析用词语集[CP/OL]. [2012-04-25]. http://www.keenage.com.

(下转第 74 页)

tems. Springer Berlin Heidelberg, 2008: 67-75.

[2] CHEN W, WANG C, WANG Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 1029-1038.

[3] LI Dong, XU Zhi-ming, LI Sheng, et al. A survey on information diffusion in online social networks [J]. Chinese Journal of Computers, 2014, 37(1): 189-206. (in Chinese)
李栋, 徐志明, 李生, 等. 在线社会网络中信息扩散[J]. 计算机学报, 2014, 37(1): 189-206.

[4] YANG J, COUNTS S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. ICWSM, 2010, 10: 355-358.

[5] WANG F, WANG H, XU K. Diffusive logistic model towards predicting information diffusion in online social networks[C]// 2012 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW). IEEE, 2012: 133-139.

[6] GUILLE A, HACID H. A predictive model for the temporal dynamics of information diffusion in online social networks[C]// Proceedings of the 21st International Conference Companion on World Wide Web. ACM, 2012: 1145-1152.

[7] BOURIGAULT S, LAGNIER C, LAMPRIER S, et al. Learning social network embeddings for predicting information diffusion [C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. ACM, 2014: 393-402.

[8] WANG Y, XIANG G, CHANG S K. Sparse Multi - Task Learning for Detecting Influential Nodes in an Implicit Diffusion Network[C]// AAAI. 2013.

[9] LIN Y, RAZA A A, LEE J Y, et al. Influence propagation: patterns, model and a case study[M]// Advances in Knowledge Discovery and Data Mining. Springer International Publishing, 2014: 386-397.

[10] WANG F, WANG H, XU K, et al. Characterizing information diffusion in online social networks with linear diffusive model [C]// 2013 IEEE 33rd International Conference on Distributed Computing Systems (ICDCS). IEEE, 2013: 307-316.

[11] YANG J, LESKOVEC J. Modeling information diffusion in implicit networks[C]// 2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE, 2010: 599-608.

[12] MA H, ZHOU D, LIU C, et al. Recommender systems with social regularization[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 287-296.

(上接第 47 页)

[13] HE Feng-ying. Orientation analysis for Chinese blog text based on semantic comprehension [J]. Journal of Computer Applications, 2011, 31(8): 2130-2133. (in Chinese)
何凤英. 基于语义理解的中文博文倾向性分析[J]. 计算机应用, 2011, 31(8): 2130-2133.

[14] LI Rong-jun, WANG Xiao-jie, ZHOU Yan-quan. Semantic Orientation Computing Using PageRank Model [J]. Journal of Beijing University of Posts and Telecommunications, 2010, 5(5): 141-144. (in Chinese)
李荣军, 王小捷, 周延泉. PageRank 模型在中文情感词极性判别中的应用[J]. 北京邮电大学学报, 2010, 5(5): 141-144.

[15] COLACE F, SANTO M D, GRECO L. SAFE: A Sentiment Analysis Framework for E-Learning[J]. International Journal of Emerging Technologies in Learning, 2014, 9(6): 37-41.

[16] MUKKAMALA R R, HUSSAIN A, VATRAPU R. Fuzzy-Set Based Sentiment Analysis of Big Social Data[C]// IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC), 2014. IEEE, 2014: 71-80.

[17] TURNEY P D, LITTMAN M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[18] CHEN Lu, WANG Wen-bo, NAGARAJAN M, et al. Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter[C]// The Sixth International AAAI Conference on Weblogs and Social Media (ICWSM). 2012.

[19] JO Y, OH A H. Aspect and sentiment unification model for online review analysis[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 815-824.

[20] NEVIAROUSKAYA A, PRENDINGER H, ISHIZUKA M. SentiFul: Generating a reliable lexicon for sentiment analysis[C]// 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009 (ACII 2009). IEEE, 2009: 1-6.

[21] SAIF M, CODY D, Bonnie D. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus [C]// Proc. of 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09). 2009: 599-608.

[22] CONTE H R, PLUTCHIK R. A circumplex model for interpersonal personality traits[J]. Journal of Personality & Social Psychology, 1981(4): 701-711.

[23] TOMÀŠ M. Statistical Language Models based on Neural Networks[D]. Brno University of Technology, 2012.

[24] TOMÀŠ M, KARAFIÁ T M, BURGET L, et al. Recurrent neural network based language model[C]// Conference of the International Speech Communication Association, 2010. Makuhari, Chiba, Japan, 2010: 1045-1048.

[25] CHEN Jian-mei, LIN Hong-fei, YANG Zhi-hao. Word Emotion Disambiguation Based on Bayesian Model[C]// The Ninth China National Conference on Computational Linguistics, 2007. (in Chinese)
陈建美, 林鸿飞, 杨志豪. 基于贝叶斯模型的词汇情感消歧[C]// 内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集. 2007.

[26] DING Ru-yi, ZHOU Hui, LIN Ma. Cognitive Appraisal Basis of Gratitude. [J]. Acta Psychologica Sinica, 2014, 46(10): 1463-1475. (in Chinese)
丁如一, 周晖, 林玛. 感激情绪的认知评估体系[J]. 心理学报, 2014, 46(10): 1463-1475.