基于语义相似度的情感特征向量提取方法

林江豪[1] 周咏梅[1,2] 阳爱民[1,2] 陈 锦[1,3]
(广东外语外贸大学语言工程与计算实验室 广州 510006)[1]
(广东外语外贸大学思科信息学院 广州 510006)[2] (广东外语外贸大学国际学院 广州 510420)[3]

摘 要 针对现有情感特征在语义表达和领域拓展等方面的不足,提出了一种基于语义相似度的情感特征向量提取方法。利用25万篇sogou新闻语料和50万条微博语料,训练得到Word2vec模型;选择80个情感明显、内容丰富、词性多样化的情感词作为种子词集;通过计算候选情感词与种子词的词向量之间的语义相似度,将情感词映射到高维向量空间,实现了情感词的特征向量表示(Senti2vec)。将Senti2vec应用于情感近义词和反义词相似度分析、情感词极性分类和文本情感分析任务中,实验结果表明Senti2vec能实现情感词的语义表示和情感表示。基于大规模语料的语义相似计算,使得提取的情感特征更具有领域拓展性。

关键词 情感特征向量,语义相似度,情感词,Word2vec
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.053

Extraction Method of Sentimental Feature Vector Based on Semantic Similarity

LIN Jiang-hao[1] ZHOU Yong-mei[1,2] YANG Ai-min[1,2] CHENG Jin[1,3]
(Laboratory for Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006, China)[1]
(Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006, China)[2]
(International College, Guangdong University of Foreign Studies, Guangzhou 510420, China)[3]

Abstract In order to fill the gap of the semantic representation and domain expansion on sentimental features, in this paper, an extraction method of sentimental feature vector based on semantic similarity was proposed. First of all, the Word2vec model is trained based on 250 thousand sogou news texts and 500 thousand micro-blog texts. Eighty sentimental words, which are obvious sentiment, rich content and diverse POS, are chosen as a set of seed words. Then, the semantic similarity between the candidate sentimental words and the seed words are calculated based on their word vectors. The sentimental words are mapped to the high dimensional vector space and the feature vector representation (Senti2vec) is extracted. Senti2vec is applied into the similarity analysis of sentimental synonyms and antonyms, polarity classification of sentimental words and sentimental text analysis. The experimental results show that Senti2vec can represent the meaning and sentiment of the sentimental words. Senti2vec is based on semantic similarity calculation from large scale of data, which enables this method more adaptable into different domains.

Keywords Sentimental feature vector, Semantic similarity, Sentiment word, Word2vec

1 引言

互联网上的海量文本情感挖掘有利于产品推荐、观点抽取和舆情监控等方面的研究。词语是组成文本的语义单元,表达了用户的观点和情感。以词汇为情感单元构建情感词典,可简单、快速地应用于海量文本情感分析[1]。在情感词典中,情感词的情感特征可用情感倾向和权重来表示。如图1所示,国外知名的情感词典SentiwordNet[2]将词汇映射到正向(P)、中性(ne)和负向(N)3个极性,并分别赋予权重,则每一个情感词可用{P,ne,N}三维情感特征来表示。国内权威的知识库HowNet[3]则将正向情感词映射到{+1,0},将负向情感词映射到{0,−1}的二维特征向量。

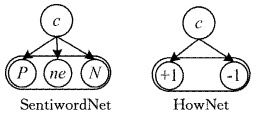

图1 SentiwordNet和HowNet的情感向量表示

针对情感特征的提取,国内外学者提出了多种研究方法,如利用SentiwordNet和HowNet等情感词典的方法[4-9]、基于种子词和互信息的方法[10-12]、基于机器学习的方法[11-15]

到稿日期:2016-10-03 返修日期:2016-12-22 本文受国家社科基金项目(12BYY045)资助。
林江豪(1985−),男,硕士,助理研究员,主要研究方向为数据挖掘、文本情感分析,E-mail:lin_hao@foxmail.com;周咏梅(1971−),女,硕士,教授,主要研究方向为文本情感分析、数据挖掘;阳爱民(1970−),男,博士,教授,主要研究方向为模式分类、文本情感分析、机器学习;陈 锦(1989−),女,硕士,讲师,主要研究方向为语言学。

等。文献[4]利用情感词典作为情感特征提取的基础，实现了文本情感分析。鉴于同一词汇在不同情感词典中的极性不同，文献[5]提出了一种自动的领域情感极性特征映射方法。文献[6]基于神经网络和情感词典，对短文本进行情感特征提取。朱嫣岚等人提出了基于语义相似度和语义相关场的两种词汇语义倾向性计算方法，通过计算目标词汇与 HowNet 中已标注褒贬性词汇间的相似度得到目标词汇的倾向性[7]。文献[8]在 HowNet 情感词语集的基础上，利用其提供的义原计算两个词的相似度，根据词与正向种子词和负向种子词的平均相似度的差来提取词语的情感倾向。文献[9]基于 HowNet 和 SentiWordNet 将词语自动分解为多个义原后计算其情感倾向强度值。利用 SO-PMI 方法，选择具有情感代表性的种子词集，在已标注语料中或基于搜索引擎计算词语在正向和负向上的情感权重，可将情感词语有效地映射到情感特征向量$\{P,N\}$[10-12]。

利用机器学习的方法，通过对语料的统计和计算，也可实现情感特征的提取。例如，彭丽针和吴扬扬提出将单词之间的语义特征映射到页面、页面社区和页面社区所属类别上[13]。文献[14]根据评论和新闻的对比分析获得候选情感特征，然后经过相关的扩充和验证操作得到通用的情感特征。文献[15]在评论中提取普通分类特征和情感特征，普通分类特征可以用来训练一个情感分类器；然后使用 spectral 聚类算法把这些情感特征映射成扩展特征。文献[16]结合 TF-IDF 方法与方差统计方法，提出了一种实现多分类特征抽取的计算方法，其采用先极性判断再细粒度情感判断的处理方法，构建细粒度情感特征提取的过程。文献[17]基于词语情感权重与文本情感倾向的相关假设，结合二元分类的特性改进了信息增益（Information Gain，IG）和卡方统计量（Chi-square，CHI），将特征选择技术应用于情感词权重特征的计算。Alaa Hamouda 等基于机器学习方法，构建了 Machine Learning Based Senti-word Lexicon（MLBSL），取得了较 SentiWordNet 更高的微平均值[21]。

现有的情感特征提取方法更注重将词语的情感映射到二维或三维特征向量，也有研究人员尝试对情感特征进行拓展[13-17]。低维的情感特征向量虽然表示简单，使用方便，但只表达了情感，缺乏语义表示。其他情感特征拓展方法主要针对特定领域的语料，使用专门的方法进行拓展，容易受限于语料环境，存在特征通用性差的缺陷。随着深度学习研究的发展，对词汇语义和情感的抽象表示也有了新的思路与方法[18-20]。例如，Mnih 和 Kavukcuoglu 利用噪声估计对比（noise-contrastive estimation）方法获得了更简单、快速和效果更优的词向量表示[18]。在 ACL 的 2016 年 Workshop 中，Camacho-Collados 和 Navigli 提出了 QVEC-CCA 模型，其主要基于人工标注的语言资源构建具有语言特性的矩阵；在词语类别分类相似性的任务上（如 fish 属于 NN. ANIMAL）表现出较好的效果[19]。文献[20]提出了一种词向量表示的评价框架，通过孤立点检测任务进行验证，将实验结果表示与人工标注的结果进行对比，对比结果表明该框架仍然存在较大的改进空间。这些研究都为词语情感特征向量表示研究提供了新的方法。Google 于 2013 年利用深度学习的思想，通过训练大量文本，把词语映射到 k 维向量空间中的向量运算，而向量空间上的相似度可以用来表示词语在语义上的相似度。因此，本文利用 Word2vec 对大量语料进行学习，选择情感种子词作为基准，利用 Word2vec 的语义计算优势将情感特征进行高维映射，并通过实验验证情感特征在语义表示和情感表示上的有效性。

本文第 2 节介绍情感特征提取框架和具体算法；第 3 节对提出的方法进行验证和分析；最后对全文进行总结。

2　基于语义相似的情感特征向量提取研究

2.1　词向量模型

利用机器学习算法完成自然语言处理任务的首要工作就是特征符号的数学表示，通常会用词向量来表示一个词语。本文通过在大量数据中训练神经网络来获得词向量（Distributed Representation），基本思想是通过训练将每个词映射成 k 维实数向量，并设置上下文的深度，可获得文本数据更加深层次的特征表示。这种词语的表示方式优于传统的 One-hot Representation，k 维向量不但包含了词语间的潜在语义关系，同时也避免了维数灾难。

Google 词向量开源工具 Word2vec 的核心架构主要基于 CBOW（Continuous Bags-of-Words）和 Skip-gram 两个模型。本文选用 CBOW 模型，如图 2 所示，CBOW 模型的原理与神经网络概率语言模型（Neural Netword Language Model，NNLM）类似，都是利用当前词 w_t 的上下文 $w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+k}$ 预测当前词 w_t 的概率 $P(w_t \mid w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+k})$。与 NNLM 不同的是，CBOW 模型去掉了最耗时的非线性隐藏层，并让输入层的所有词语共享映射层。图 1 中 CBOW 模型使用单词 w_t 周边的词语作为输入，在映射层做加权处理，然后输出单词 w_t。

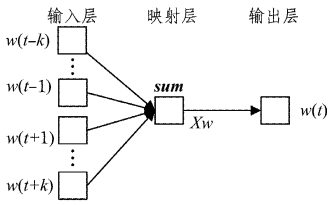

图 2　CBOW 模型的框架

Word2vec 考虑了当前词的上下文信息，因此学习到的词向量包含了丰富的语义和语法关系。本文的训练语料集为 sogou 全网新闻数据（SogouCA），随机取其中 25 万篇新闻，涉及到国内、国际、体育等 18 个频道的新闻数据。采集新浪、腾讯微博共 50 万条微博文本。将 sogou 新闻语料和微博语料共 75 万篇作为 Word2vec 的训练数据集 Data_w2v。训练模型选用 CBOW，得到模型 M_w2v。

2.2　情感特征向量生成模型的概述

本文采用如图 3 所示的情感特征生成算法。首先，设置 Word2vec 的模型参数，将数据集 Data_w2v 作为训练集输入到 Word2vec 进行训练，保存 Word2vec 模型为 M_w2v；种子词集 $S=\{(s_1,wt_1),(s_2,wt_2),\cdots,(s_k,wt_k)\}$，其中 $s_j(j=1,2,\cdots,k)$ 为种子词，$wt_j(j=1,2,\cdots,k)$ 为 s_j 对应的情感权重。S 表示具有明显情感极性的情感词集合，如“美丽”、“漂亮”等为正向情感种子词；“邪恶”、“悲哀”为负向情感种子词。候选

情感词集 $C=\{(c_1),(c_2),\cdots,(c_m)\}$，表示情感极性未知的词汇。该算法的主要原理是通过计算词集 C 中每一个词 c_i 与 S 中各个词的相似度来实现 c_i 的情感分布。如“靓丽”与种子词集{“美丽”,“漂亮”,“快乐”,…,“邪恶”,“悲哀”,“愚蠢”,…}中各个词汇的相似度{0.75,0.68,0.32,…,0.08,0.11,0.09,…}，表示“靓丽”在这些情感表达上的权重，实现对 c_i 情感表达的抽象表示，以便计算机实现情感计算。具体方法是：通过相似度 similarity 函数 f_s 分别计算情感词 $c_i(i=1,2,\cdots,m)$ 与种子词集 S 中各个种子词的相似度，得到相似度向量 $Sim=\{Sim_{i1}, Sim_{i2},\cdots,Sim_{ij},\cdots,Sim_{ik}\}$；进而将 Sim 与 $\{wt_1, wt_2,\cdots,wt_k\}$ 进行点乘计算，获得情感词 c_i 的情感特征向量分布 e_i；最终获得情感向量特征集合 $Senti2vec=\{(c_1,[e_1]^{1\times k}),(c_2,[e_2]^{1\times k}),\cdots,(c_m,[e_m]^{1\times k})\}$。

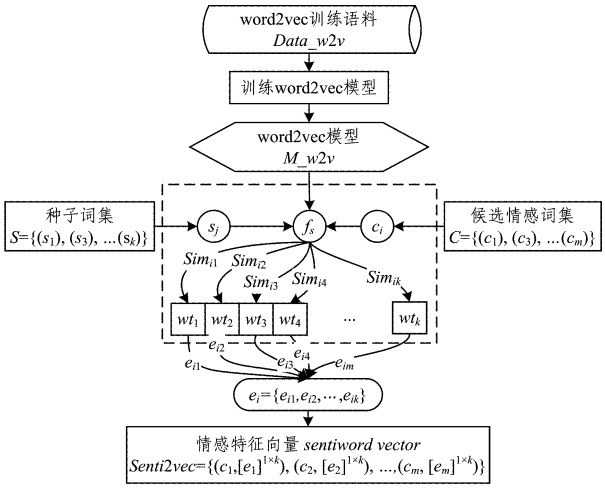

图 3 情感特征向量生成模型

根据图 3，利用 Word2vec 的语义相似度和文献[9]中的情感种子词的情感强度，将候选情感词 c 映射到高维向量表示，实现了情感词在语义空间和情感空间上的映射。具体算法如下所示。

算法 情感特征向量提取算法

输入：Word2vec 模型参数{向量维数(-size)，上下文窗口大(-window)，高频词亚采样的阈值(-sample)，cbow/skip-gram(-isCbow)}，Word2vec 训练数据集 Data_w2v，种子词集 S[9]，情感词集 C

输出：Senti2vec

步骤 1 初始化 Word2vec 模型：-size=200，-window=5，-sample=le^{-3}，-isCbow=true；

步骤 2 将 Data_w2v 进行分词、去标点符号等预处理之后输入模型，训练得到 Word2vec 的模型 M_w2v；

步骤 3 对于每一个 $c_i(c_i\in C)$：

在 M_w2v 中获得 c_i 的词向量 v_{ci}；

对于每一个 $s_j(s_j\in S)$：

在 M_w2v 中获得 s_j 的词向量 v_{sj}；

计算 V_{ci} 和 V_{sj} 的余弦相似度 $Sim_{ij}(Sim_{ij}\in Sim)$，如式(1)所示：

$$Sim_{ij}\leftarrow f_s(c_i, s_j)=(v_{ci}\cdot v_{sj})/(|v_{ci}|+|v_{sj}|) \quad (1)$$

步骤 4 将 Sim 与种子词情感权重向量 $\{wt_1, wt_2,\cdots,wt_k\}$ 进行点乘计算，得到词语的情感特征向量 $[e_i]^{1\times k}$，如式(2)所示：

$$[e_i]^{1\times k}\leftarrow Sim\odot\{wt_1, wt_2,\cdots,wt_k\} \quad (2)$$

步骤 5 输出 $Senti2vec=\{(c_1,[e_1]^{1\times k}),\cdots,(c_m,[e_m]^{1\times k})\}$。

模型的输出为 *Senti2vec*，融合了 Word2vec 模型的语义和种子词的情感两个层面，实现了情感词的高维向量空间映射。同时，该算法应用于词汇的情感特征向量生成时不受情感词典内容范围的约束。

3 实验结果及分析

3.1 实验设计

情感特征向量将情感映射到高维的空间，对情感表达的有效性主要体现在情感和语义相近的词汇的空间距离小，而情感和语义相反的词汇的空间距离大。例如，“漂亮-美丽”、“开心-快乐”等情感近义词组具有较强的情感和语义相似性，其空间距离应该小，即相似度应该高；如“美丽-丑陋”、“欢喜-悲伤”等情感反义词组具有相反的情感，在语义上是反向的，其空间距离应该大，相似度应该低。另外，通过情感特征向量能准确获得情感词的极性(“正向”或“负向”)，这表明情感特征向量是区分词语情感的有效验证方法。因此，将所得的情感特征向量应用到如表 1 所列的任务中。

表 1 情感特征向量验证任务

编号	任务名称	方法	数据
任务 1	近义情感词组和反义情感词组的相似性分析	Word2vec 模型中的 similarity 算法	50 对情感近义词组 50 对情感反义词组
任务 2	情感词极性分类实验	SVM	取 HowNet 中的 10000 个情感词
任务 3	新闻评论文本情感分类实验	SVM	NLP&CC2012[1)]微博语料

对以上任务进行评价时选用总体准确率作为性能评价指标：$acc=cor/all$，$acc\in[0,1.0]$，其中 acc 代表总体准确率，cor 是分析正确的数量，all 是全部分析对象的数量。

3.2 情感特征向量提取结果

将 HowNet 的情感词汇作为候选情感词，选择语料中出现的情感词形成情感词集 C。最终，利用情感特征向量提取算法获得情感特征向量 *Senti2vec*，共有情感词 12695 个，情感特征向量与种子词集的词汇量一致，即将每个情感词映射到 80 维的向量空间，采用基于 python 的数据可视化工具 t-sen[22] 进行可视化，结果如图 4 所示。

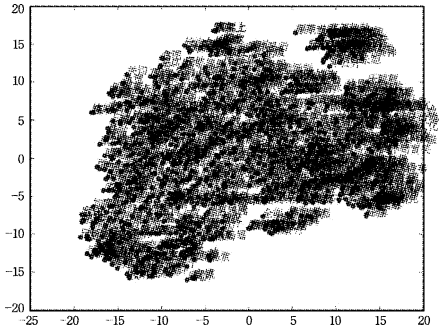

图 4 t-sen 情感词及其特征向量可视化结果

为更好地观察情感词的相似性，对图 4 进行局部放大显

1) http://tcci.ccf.org.cn/conference/2012/

示，如图 5 所示。从图 5 中可见，如“匠心独具”、“别具匠心”、“独具一格”、“妙趣”等情感词在语义和情感上都有一定的相似性，其在情感向量空间上相对应的距离也比较近，说这明情感特征向量对语义和情感的聚合效果比较好。

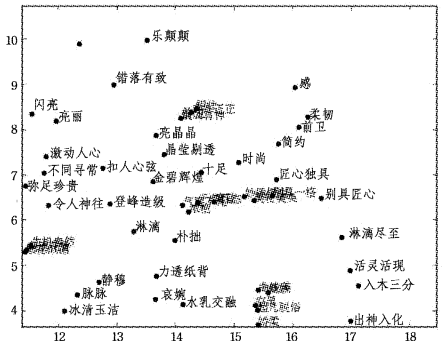

图 5　t-sen 情感词及其特征向量的局部可视化效果

对于基于 Word2vec 模型输出的 200 维词向量和 Senti2vec 的 80 维情感特征向量，利用 similarity 函数获取情感词最相似的 5 个词汇，结果如表 2 所列。

表 2　相似词及相似度示例(top 5)

情感词	Word2vec	Senti2vec
悲哀	可悲 0.7031468749046326	耻辱 0.9783877730369568
	可笑 0.6274901032447815	可悲 0.9780582189559937
	可怕 0.5995386242866516	渺小 0.9722716808319092
	不可理喻 0.5955908298492432	无耻 0.9661660194396973
	耻辱 0.5938669443130493	憎恨 0.9655710458755493
漂亮	可爱 0.7082474231719971	可爱 0.9688816070556641
	好看 0.697650671005249	讨人喜欢 0.9682539701461792
	讨人喜欢 0.6583585739135742	乖巧 0.9661190509796143
	洋气 0.6495774388313293	机灵 0.9638960361480713
	小巧玲珑 0.6462069749832153	好看 0.958671510219574
不错	不好 0.6326699256896973	要好 0.9078071117401123
	不俗 0.6292985677719116	好 0.9053456783294678
	差强人意 0.6165729761123657	令人满意 0.9027157425880432
	出色 0.6126718521118164	可观 0.8967028260231018
	好 0.6089481115341187	得心应手 0.890740156173706

表 2 所列为情感词“悲哀”、“漂亮”和“不错”的 top 5 相似词汇及其对应的相似度。观察表 2 可知，Word2vec 和 Senti2vec 在语义上都有较好的表现，能获得情感和语义均比较相似的词汇，但 Senti2vec 的词汇的相似性更强；另外，以词汇“不错”为例，利用 Word2vec 获得的最相似词为“不好”，两者在语义和情感上的距离均比较大，而利用 Senti2vec 则在语义和情感上均有一致性的表现。从数据量角度分析，Senti2vec 将 Word2vec 从 200 维的向量降到了 80 维，表明 Senti2vec 也是一种有效的降维方法。这些都说明了情感特征向量的优越性。

3.3　情感特征向量的验证实验及结果分析

3.3.1　近义情感词组和反义情感词组的相似性分析

选择表 3 所列的近义词组和反义词组各 50 对作为研究对象来对词组的相似性进行分析。

近义词组的相似度分析结果如图 6 所示。由图 6 可见，Word2vec 和 Senti2vec 在近义词的相似性分析上均有较好的效果，语义上的相似度都大于 0.65。在情感近义词的分析上，基于 Senti2vec 的相似度总体上比 Word2vec 的高，这说明 Senti2vec 能够更有效地进行表达。

表 3　情感特征向量验证任务

近义词组(50 对)
夸耀—炫耀，惬意-舒服，惊讶-惊奇，轻蔑-轻视，闻名-著名，慈祥-慈爱，满意-满足，精彩-出色，挺秀-挺拔，洒脱-潇洒，奇丽-秀丽，柔美-优美，贵重-珍贵，漂亮-美丽，恐惊-恐惧，简朴-简单，恬静-舒适，著名-闻名，高兴-兴奋，温暖-暖和，恐惧-惧怕，强盛-强大，舒服-舒适，焦急-焦虑，侮辱-欺侮，崎岖-坎坷，诚实-老实，幽静-清幽，劫难-灾难，笑容-笑脸，温和-温顺，暴躁-急躁，刚强-坚强，挖苦-讥讽，糟蹋-糟践，情谊-友谊，安谧-静谧，痛快-愉快，新颖-新奇，智慧-聪明，用心-专心，气愤-生气，雄伟-宏伟，垂头丧气-没精打采，纤尘不染—一尘不染，全神贯注-聚精会神，迷迷糊糊-模模糊糊，情不自禁-不由自主，赏心悦目-心旷神怡，斗志昂扬-意气风发
反义词组(50 对)
好-坏，对-错，爱-恨，乐-悲，恶劣-良好，尊重-侮辱，伟大-渺小，开心-苦闷，扫兴-高兴，刚强-软弱，痛快-难受，喜欢-厌恶，坚强-软弱，失信-守信，美丽-丑陋，强盛-衰败，犹豫-坚定，勤劳-懒惰，喜欢-讨厌，胜利-失败，整齐-纷乱，舒畅-苦闷，可爱-可恶，高兴-难过，气愤-欢喜，光明-黑暗，惩罚-奖励，好心-恶意，诚实-撒谎，表扬-批评，善良-凶恶，精致-粗糙，糟糕-精彩，危险-安全，繁荣-衰败，聪明-愚笨，脆弱-坚强，活泼-呆板，心惊肉跳-镇定自如，毫不犹豫-犹豫不决，风平浪静-狂风恶浪，全神贯注-心不在焉，力倦神疲-精力充沛，悔过自新-执迷不悟，唇枪舌剑-心平气和，安居乐业-颠沛流离，粗制滥造-精雕细刻，一丝不苟-粗心大意，雪中送炭-雪上加霜，赏心悦目-触目惊心

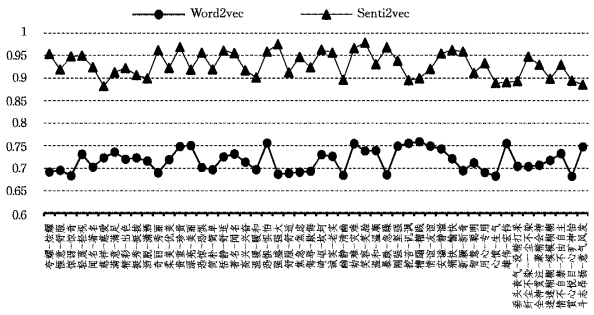

图 6　近义词组的相似度分析结果

反义词组的相似度分析结果如图 7 所示，Word2vec 和 Senti2vec 对反义词进行分析时均存在一定的误差。Word2vec 对反义词组“犹豫-坚定”、“活泼-呆板”、“心惊肉跳-镇定自如”、“粗制滥造-精雕细刻”、“唇枪舌剑-心平气和”、“舒畅-苦闷”、“诚实-撒谎”、“善良-凶恶”的相似度大于 0.65；Senti2vec 对反义词组“一丝不苟-粗心大意”、“心惊肉跳-镇定自如”、“美丽-丑陋”的相似度也大于 0.65。定义反义词组相似度大于 0.6 为误差，则 Word2vec 和 Senti2vec 的准确率分别为 0.84 和 0.94。

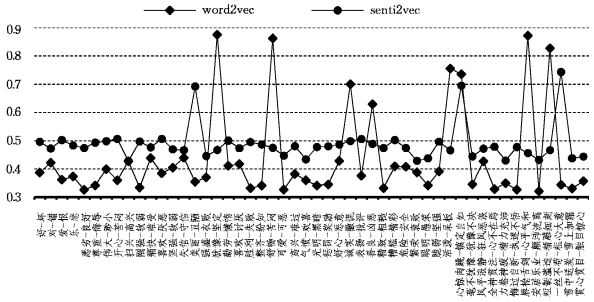

图 7　反义词组的相似度分析结果

3.3.2　情感词极性分类实验

情感词的极性判断和权重计量是情感词典构建的核心研究。柳位平等提出了基于 HowNet 和 SO-IR 算法的情感词极性判断方法，利用褒贬义种子词与 HowNet 中词语义原之间的相似度，将正向义原的语义相似度平均值与负向义原的

平均值之差作为 SO-IR 值，该值大于零表示正向，小于零表示负向，绝对值为权重[6]。阳爱民等选用 80 个情感种子词，利用搜索引擎返回共现数，通过改进的 PMI 算法计算情感词的情感权值[10]。周咏梅等通过新闻评论语料和基础情感词典获得评论情感词集和种子词，利用基于 PageRank 算法的方法判定评论情感词集的极性并计算其强度[9]，提出了基于 HowNet 和 SentiWordNet 的中文情感词典方法。将词语自动分解为多个义原后计算其情感倾向强度，并且使用词典校对方法对词语情感倾向强度进行优化[23]。

为了验证情感特征的有效性，本文选取 HowNet 中的 10000 个情感词及其极性作为标准词集，其中 2000 个作为训练集，8000 个作为测试集。对词语进行情感极性分类时，情感词 c_i 的 Word2vec 向量表示为 $[v_{ci}]^{1\times200}$，情感特征向量 Senti2vec 表示为 $[e_i]^{1\times80}$。用于训练 SVM 分类器的情感词集合向量矩阵为 $[v]^{2000\times200}$ 和 $[e]^{2000\times80}$，训练得到 SVM^{w2c} 和 SVM^{s2c} 两个分类器。将测试情感词集 $[v]^{8000\times200}$ 和 $[e]^{8000\times80}$ 分别作为 SVM^{w2c} 和 SVM^{s2c} 的输入，对测试集进行二分类，得到情感词极性的分类结果。同时，进行了相关情感词极性判断研究，实验结果如表 4 所列。

表 4 情感词极性分类结果

情感词典构建方法	acc/%
SO-IR[6]	82.54
搜索引擎＋PMI[10]	69.87
HowNet＋SentiWordNet＋语义相似度[7]	76.34
PageRank[9]	78.76
Word2vec＋SVM	76.15
Senti2vec＋SVM(本文方法)	82.15

从实验结果可以看出，Senti2vec＋SVM 方法可获得较高的准确率 82.15%，略低于 SO-IR 方法。两种方法都基于语义相似度进行计算，HowNet 是权威的知识库，利用 HowNet 的相似度可获得良好的效果，但是 HowNet 受限于词典的规模，比如网络用语“么么哒”、“累觉不爱”等，因此无法利用 HowNet 的方法来进行相似度的计算。而利用 Senti2vec 的方法是基于海量语料的知识统计获得的，能有效获得这些网络用语的情感特征向量，具有比基于 HowNet 的方法更强的可拓展性。文献[10]基于搜索引擎的方法的主要问题是存在太多的垃圾网页信息，检索出的页面数量存在很多不相关因素，导致准确率不高。文献[7,23]的方法在实验中的准确率略低于 Senti2vec，同时也容易受限于情感词典的规模。结果表明，基于 Senti2vec 的方法在词汇拓展和情感极性判断方面均可获得更优的效果。

3.3.3 基于情感特征向量的微博文本情感分类结果

将提取的情感特征向量应用于中文微博文本情感分析实验。实验数据来源于 NLP&CC2012 中文微博情感分析评测的样例数据集，共 2173 条微博。抽取其中包含情感词的 1000 条微博作为验证数据，并随机抽取 2/3 的数据用于模型训练，剩下的数据用于测试。

采用支持向量机(SVM)构建文本情感分类器。SVM 是一种监督式学习的方法，属于一般化线性分类器，它能够同时最小化经验误差并最大化几何边缘区，因此也被称为最大边缘区分类器。

利用 SVM 进行分类时，将微博进行分词、去停用词等预处理后，选择其中的情感词作为微博的情感特征，每条微博可表示为 $doc=\{c_1,c_2,\cdots,c_n\}$，情感词 c_i 的 Word2vec 向量表示为 $[v_{ci}]^{1\times200}$，情感特征向量 Senti2vec 表示为 $[e_i]^{1\times80}$。则微博的 Word2vec 特征向量 $V_{w2v}=\sum v_{ci}/n$，向量长度为 200 维；微博的情感特征向量 Senti2vec 可表示为 $V_{s2c}=\sum e_i/n$，向量长度为 80 维。利用 HowNet 对微博进行向量表示时，本文基于传统的 One-hot Representation，微博的向量长度为情感词典的长度。微博文本的情感分类结果如表 5 所列。

表 5 文本情感分类结果

文本情感分类方法	acc/%
SOSL[10]	58.62
情感词典＋SVM[23]	62.84
情感词典＋NB[24]	57.13
HowNet＋SVM	57.78
Word2vec＋SVM	59.63
Senti2vec＋SVM	63.65

观察实验结果可知，采用 Senti2vec＋SVM 可获得较高的分类准确率。总体的准确率偏低，主要是因为只提取了微博中的情感词汇作为特征，没有考虑转折词、程度副词等对情感的影响。由于本文主要是验证提取的情感特征向量的有效性，因此仅提取其中的情感词汇作为情感特征。对比基于情感词典[10]和基于机器学习[21-22]的微博情感分类方法，Senti2vec 将情感进行高维映射的效果明显优于传统的二维或三维情感映射，这说明 Senti2vec 能更好地表示情感。同时，Senti2vec 取得了比 Word2vec 更好的效果，说明 Senti2vec 能实现语义表示。实验结果表明，提出的 Senti2vec 能更好地实现语义和情感的有效表示。

结束语 本文针对情感词的情感特征向量表示，提出基于语义相似度的情感特征向量(Senti2vec)提取方法。这种方法融合了 Word2vec 模型的语义优势和情感种子词集的情感表达优势，将情感词汇映射到高维度向量空间，实现了情感词在语义空间和向量空间的有效表示。同时，利用大量语料的语义计算，使得情感词汇的特征向量领域可移植性更强。通过实验和结果分析，验证了所提方法的可行性和 Senti2vec 的有效性。通过设定不同领域的种子词集，也可将本文方法用于特定领域情感词汇的特征向量表示。在下一步研究工作中，将进一步拓展训练数据集，加入维基百科等语料，训练更广域的语义表示，进一步提升 Senti2vec 的领域可用性。

参 考 文 献

[1] XU G, MENG X F, WANG H F. Build Chinese Emotion Lexicons Using A Graph-based Algorithm and Multiple Resources [C]// Proceedings of the 23rd International Conference on Computational Linguistics. 2010: 1209-1217.

[2] 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.

[3] DAI L L，XIA Y N，LIU B，et al. Measuring Semantic Similarity between Words Using HowNet[C]// Proceedings of the 2008 International Conference on Computer Science and Information Technology. 2008：601-605.

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

[5] DRAGUT E C，WANG H，SISTLA P，et al. Polarity Consistency Checking for Domain Independent Sentiment Dictionaries[J]. IEEE Transactions on Knowledge and Data Engineering，2015，27(3)：838-851.

[6] VO D T，ZHANG Y. Don't Count，Predict! An Automatic Approach to Learning Sentiment Lexicons for Short Text[C]// The 54th Annual Meeting of the Association for Computational Linguistics. 2016：219.

[7] ZHU Y L，MIN J，ZHOU Y Q，et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing，2006，20(1)：14-20. (in Chinese)
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.

[8] LIU W P，ZHU Y H，LI C L，et al. Research on building Chinese basic semantic lexicon[J]. Journal of Computer Applications，2009，29(11)：2882-2884. (in Chinese)
柳位平，朱艳辉，栗春亮，等. 中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29(11)：2882-2884.

[9] ZHOU Y M，YANG A M，YANG J N. Construction Method of Sentiment Lexicon for News Reviews[J]. Computer Science，2014，41(8)：67-69. (in Chinese)
周咏梅，阳爱民，杨佳能. 一种新闻评论情感词典的构建方法[J]. 计算机科学，2014，41(8)：67-69.

[10] YANG A M，LIN J H，ZHON Y M，et al. Research on Building a Chinese Sentiment Lexicon Based on SO-PMI[J]. Applied Mechanics and Materials，2013，263-266：1688-1693.

[11] ZHOU Y M，YANG A M，LIN J H. A method of building Chinese microblog sentiment lexicon[J]. Journal of Shandong University (Engineering Science)，2014，44(3)：36-40. (in Chinese)
周咏梅，阳爱民，林江豪. 中文微博情感词典构建方法[J]. 山东大学学报(工学版)，2014，44(3)：36-40.

[12] WANG G W，ARAKI K. Modifying SO-PMI for Japanese Weblog Opinion Mining by Using a Balancing Factor and Detecting Neutral Expressions[C]// Proceedings of NAACL HLT. 2007：189-192.

[13] PENG L Z，WU Y Y. Semantic Similarity Computing Based on Community Mining of Wikipedia[J]. Computer Science，2016，43(4)：45-49. (in Chinese)
彭丽针，吴扬扬. 基于维基百科社区挖掘的词语语义相似度计算[J]. 计算机科学，2016，43(4)：45-49.

[14] TAO F M，GAO J，WANG T J，et al. Topic Oriented Sentimental Feature Selection Method for News Comments[J]. Journal of Chinese Information Processing，2010，24(3)：37-43. (in Chinese)
陶富民，高军，王腾蛟，等. 面向话题的新闻评论的情感特征选取[J]. 中文信息学报，2010，24(3)：37-43.

[15] LI S K，JIANG Y B. Semi-Supervised Sentiment Classification Based on Sentiment Feature Clustering[J]. Journal of Computer Research and Development，2013，50(12)：2570-2577. (in Chinese)
李素科，蒋严冰. 基于情感特征聚类的半监督情感分类[J]. 计算机研究与发展，2013，50(12)：2570-2577.

[16] HE F Y，HE Y X，LIU N，et al. A Microblog Short Text Oriented Multi-class Feature Extraction Method of Fine-Grained Sentiment Analysis [J]. Acta Scientiarum Naturalium Universitatis Pekinensis，2014，50(1)：48-54. (in Chinese)
贺飞艳，何炎祥，刘楠，等. 面向微博短文本的细粒度情感特征抽取方法[J]. 北京大学学报(自然科学版)，2014，50(1)：48-54.

[17] WU J Y，JI J Z，ZHAO X W，et al. Weight Calculation of Emotional Word Based on Feature Selection Technique[J]. Journal of Beijing University of Technology，2016，42(1)：142-151. (in Chinese)
吴金源，冀俊忠，赵学武，等. 基于特征选择技术的情感词权重计算[J]. 北京工业大学学报，2016，42(1)：142-151.

[18] PENNINGTON J，SOCHER R，MANNING C. Glove：Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014：1532-1543.

[19] TSVETKOV Y，FARUQUI M，DYER C. Correlation-based Intrinsic Evaluation of Word Vector Representations[C]// The Workshop on Evaluating Vector-Space Representations for Nlp. 2016：111-115.

[20] CAMACHO-COLLADOS J，NAVIGLI R. Find the word that does not belong：A Framework for an Intrinsic Evaluation of Word Vector Representations[C]// The Workshop on Evaluating Vector-Space Representations for Nlp. 2016：43-50.

[21] HAMOUDA A，MAREI M，ROHAIM M. Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis[J]. Journal of Advances in Information Technology，2011，2(4)：199-203.

[22] VAN DER MAATEN L J P. Accelerating t-SNE using Tree-Based Algorithms[J]. Journal of Machine Learning Research，2014，15(1)：3221-3245.

[23] ZHOU Y M，YANG J N，YANG A M. A method on building Chinese sentiment lexicon for text sentiment analysis[J]. Journal of Shandong University (Engineering Science)，2013，43(6)：27-33. (in Chinese)
周咏梅，杨佳能，阳爱民. 面向文本情感分析的中文情感词典构建方法[J]. 山东大学学报(工学版)，2013，43(6)：27-33.

[24] YANG D，YANG A M. Classification approach of Chinese texts sentiment based on semantic lexicon and naïve Bayesian [J]. Application Research of Computers，2010，27(10)：3737-3739，3743. (in Chinese)
杨鼎，阳爱民. 一种基于情感词典和朴素贝叶斯的中文文本情感分类方法[J]. 计算机应用研究，2010，27(10)：3737-3739，3743.