文章编号:1671-9352(2014)11-0051-08 DOI:10.6040/j.issn.1671-9352.3.2014.255

基于词向量的情感新词发现方法

杨阳,刘龙飞,魏现辉,林鸿飞
(大连理工大学信息检索研究室,辽宁 大连 116023)

摘要:词语级的情感倾向性分析一直是文本情感计算领域的热点研究方向,如何自动识别情感新词,并判断其情感倾向性已经成为当前亟待解决的问题。首先用基于统计量的方法识别微博语料中的新词,然后利用神经网络去训练语料中词语的词向量,从语料自身挖掘出词与词之间的相关性,最后提出了基于词向量的情感新词发现方法。实验表明该方法可以有效应用于情感新词发现。

关键词:情感词;神经网络;词向量
中图分类号:TP391 **文献标志码**:A

New methods for extracting emotional words based on distributed representations of words

YANG Yang, LIU Long-fei, WEI Xian-hui, LIN Hong-fei
(Information Retrieval Laboratory, Dalian University of Technology, Dalian 116023, Liaoning, China)

Abstract: Word-level sentiment analysis is a hot research interest in the field of affective computing. How to recognize and analyze these new emotional words automatically becomes an urgent problem. Firstly, statistics-based approach was used to identify the new words in Micro-blog corpus and then distributed representation of new words was trained by using neural network in order to get the correlation between words in corpus. Finally three vector-based methods to find new emotional words were introduced. The experimental results indicate that the proposed methods in this paper can be effectively used in discovery of new emotional words.

Key words: emotional words; neural network; distributed representations of words

0 引言

随着互联网迅速的发展,人人网、微博、微信等这些新型社交媒体逐渐改变着人们的生活习惯。人们倾向于从微博上获取资讯、新闻、观点、评论、娱乐等信息,每日的热门话题和各地的新闻事件往往会第一时间出现在微博中。不知不觉间,微博对网络舆情的传播施加了越来越重要的影响,已经成为网络上新词创造和传播的主要平台之一,每天都会有富含情感的新词出现。这种"新"不仅包括真正意义上的新词,如"屌丝","高富帅","切糕","傲娇"等,也包括大量带有错别字的"新词"。拿"尼玛"这个新词来说,其源于"你妈"这一词语,但不仅限于"尼玛"这种表达,在微博中常会发现"尼马","泥马"等一些另类表达。正因为微博这种言论自由、口语化、毫无约束、不严肃表达的特性,使其与传统的文本写作表达有很大的不同,常会出现

收稿日期:2014-08-28; 网络出版时间:2014-10-24 14:10
网络出版地址:http://www.cnki.net/kcms/doi/10.6040/j.issn.1671.9352.3.2014.255.html
基金项目:国家自然科学基金资助项目(60673039,60973068);国家高技术研究发展计划("八六三"计划)项目(2006AA01Z151);教育部留学回国人员科研启动基金资助项目(20090041110002);高等学校博士学科点专项科研基金资助项目(20110041110034)
作者简介:杨阳(1989-),男,硕士研究生,主要研究方向为情感计算. E-mail:yangyang0477@mail.dlut.edu.cn
* 通讯作者:林鸿飞(1962-),男,教授,主要研究方向为搜索引擎、文本挖掘、情感计算和自然语言处理. E-mail: hflin@dlut.edu.cn

缺失主语或含有错别字等情况,给情感计算及观点挖掘带来了很大的困难。而人工构建和维护一部完善的微博情感词典是非常不现实的。针对以上情况,本文提出了基于词向量的情感新词发现方法,主要使用Google开源词向量工具word2vec训练微博语料中的词语向量,以获得语料中词语与词语之间的潜在语义关系,然后利用这些词向量之间的潜在联系,自动识别和发现微博中最新用于情感表达的词汇。

1 相关工作

目前,针对于词语情感倾向性判断(情感词识别),已有大量的相关工作。如朱嫣岚等人[1]使用HowNet提供的语义相似度和语义场的计算功能,计算词与词之间的相似度,对词语的褒贬倾向性按照一定的计算法则进行赋值,从而判别其褒贬倾向。实验表明,使用少量的基准词,在常用的词语集合中也可以达到80%的准确率。王素格等人[2]认为采用目标词与基准词的相关性来确定目标词的情感倾向这种方法,并没有考虑到目标词与同义词之间的关系,于是提出了基于同义词的词汇情感倾向判别的方法,不仅考虑了目标词语基准词的相关性,而且也考虑了目标词的同义词与基准词的相关性。唐都钰[3]描述了领域自适应中文情感分析词典构建的三个步骤,分别是种子词获取、语义图的构建以及计算情感得分,其中语义图的构建依赖于哈工大的同义词词林。以上相关研究都使用外部语义资源来获取词与词之间的语义关系。但对于情感新词发现而言,词语本身为新词,外部语义资源如果没有相应的更新与拓展,很难去获取新词的语义信息,这样就无法计算新词与其他词汇之间的相关度。针对这种情况,本文提出了利用神经网络训练出来的词向量来获取词与词之间的相关度,从而判断其情感倾向性。这种基于神经网络训练出来的词向量,从语料自身挖掘词与词之间的相关性,在一定程度上摆脱了对外部语义资源的依赖。

2 新词发现

情感新词发现,首先必须发现语料中的新词,其次是判断新词的情感倾向性。本文中提到的“新词”,是指不在COAE 2014给定的旧词词典之内的词语,所使用的语料为COAE 2014提供的大规模微博句子集合(千万级规模),其中并未有标注数据,所以主要采取无监督的方法进行新词发现。

2.1 语料处理

本文对大规模的微博语料进行了如下预处理:按中文标点的逗号(,)、叹号(!)、句号(。)、分号(;)、顿号(、)对微博句子进行短句切分;计算每个短句的前缀与后缀,方便后续统计量的统计。表1是对“保险资金进场”这句短语提取的后缀与前缀。

表1 后缀与前缀的示例

Table 1 Example of the suffix and prefix

后缀	前缀
保险资金进场	保险资金进场
险资金进场	保险资金进
资金进场	保险资金
金进场	保险资
进场	保险
场	保

由于部分分割短语字串相对较长,本文在建后缀与前缀的过程中,进行了字串长度的限定。长度大于10的字符串,在建立后缀的过程中会被“截尾”,在建立前缀的过程中会被“去头”。

2.2 统计量介绍

在新词发现过程中,主要参考词频、左邻接熵、右邻接熵、互信息这4个统计量来计算字符串的成词概率。

1)词频。统计语料中长度(本文中定义的为6,长度定义越长,计算时间越长)小于等于某个值的所有候选词串的频次。

2)左右邻接熵。邻接熵是Huang等[4]提出的判断词串是否成词的一个很重要的统计量。邻接熵统计量利用信息熵来衡量候选新词t的左邻字符和右邻字符的不确定性。不确定性越高,表明候选新词t的前后字符串越混乱,越不稳定,所以其成词的可能性就越高。如“肯德基”这个词语,在语料中,“肯德”的左邻接熵会比较高,因为“肯德”的左边会相邻很多不同的字符,如“吃肯德基”,“去肯德基”,“喜欢肯德基”等,而“肯德”的右边相邻的字符很少,大部分都为“基”,所以右邻接熵会比较低。因此“肯德”无法成为一个词

语,但“肯德基”的左右邻接熵都比较高,所以其成词的可能性要高于“肯德”。本文使用节 2.1 中的字符串后缀信息统计计算备选字符的右邻接熵,使用字符串的前缀信息统计计算备选字符的左邻接熵。假设用字符 x 和字符 y 表示候选新词 t 的左邻字符和右邻字符,则 t 的左邻接熵 $\mathrm{HL}(t)$ 和右邻接熵 $\mathrm{HR}(t)$ 的计算方法如下:

$$\mathrm{HL}(t) = -\sum_{x} p(x|t)\log p(x|t), \tag{1}$$

$$\mathrm{HR}(t) = -\sum_{y} p(y|t)\log p(y|t)。 \tag{2}$$

其中,$P(x|t)$ 表示字符 x 是备选词 t 的左邻字符的概率,$P(y|t)$ 表示字符 y 是备选词 t 的右邻字符的概率。

3) 互信息。互信息是新词发现中常用的统计量[5]。假设对于候选词语“周杰伦(t)”我们想要知道“周杰(x)”和“伦(y)”这两个字符在语料中连接的紧密程度,按照公式(3)计算之后,值越大代表两者之间的相关性越高,两个字符串连接之后成词的可能性也就越高。

$$\mathrm{MI}(t) = \log\left(\frac{p(t)}{p(x)p(y)}\right) = \log\left(N \cdot \frac{n_t}{n_a \cdot n_b}\right)。 \tag{3}$$

采用简单的归一化频率形式来估计概率:$P(t) = n_t/N$, $p(x) = n_x/N$, $p(y) = n_y/N$,其中 n_t,n_x,n_y 分别表示 t、x、y 字符在语料中出现的频次,N 是集合中所有长度满足阈值(本文中为小于等于 6)的字符串总数。

2.3　阈值选取

由于语料中未有标注数据,如果采取监督学习的方法,把 4 个统计量作为特征来训练分类器,会带来大量的人工标注工作。所以本文采取了无监督的方法,对每一个统计量设定了一个阈值,一个备选字符串满足这 4 个阈值,就会被判断为一个词语,如公式(4)所示:

$$\mathrm{wordset} = \{t|\mathrm{Fre}(t) > a_1 \wedge \mathrm{MI}(t) > a_2 \wedge \mathrm{HL}(t) > a_3 \wedge \mathrm{HR}(t) > a_4\}, \tag{4}$$

其中,a_1、a_2、a_3、a_4 分别为词频、互信息、左邻接熵、右邻接熵的阈值。

如表 2 所示,系统自动识别的新词按互信息排序 Top 10。值得注意的是由于只需要对最后结果进行排序,所以本文对互信息的计算并没有进行公式(3)中的对数运算。

表 2　新词 Top 10

Table 2　Top 10 new words

词语	互信息	词频	左邻接熵	右邻接熵
≥‿≤	0.001 140 870	355	4.685 104 850	4.254 826 247
桀骜	0.000 715 372	406	2.605 456 894	4.046 636 711
蘆薈	0.000 701 579	1 121	3.813 146 574	4.508 211 613
涅槃	0.000 494 604	555	4.163 172 662	4.519 149 197
奮鬥	0.000 432 311	421	4.194 576 893	3.970 243 046
討厭	0.000 396 102	1 313	4.869 316 838	3.991 196 373
惺惺	0.000 360 844	417	2.519 200 053	2.563 723 244
嚯嚯	0.000 350 936	441	2.893 203 773	2.886 910 264
習慣	0.000 338 519	973	4.529 916 125	4.155 318 673
聖誕	0.000 333 174	780	3.921 305 888	4.686 654 537

2.4　结果分析

通过阈值的选取,本文将生成的词语进行了数量上的控制,初步识别出 24 136 个词语,再通过与 COAE 2014 提供的原始词典进行对比筛选,最终留下 16 597 个新词。

对这 16 597 个新词进行大体上的观测,可以发现新词中包含了以下几类词语:

- 网络表情符号,如“(￣▽￣)”、“≥‿≤”、“≥∇≤”等,这些符号表情在微博中对于个人情感的表达有着很重要的作用。
- 命名实体,如“佟丽娅”,“邹恒甫”等各行业名人,也有如“泸州老窖”、“奥迪 A4l”、“肯德基”、“麦当劳”等品牌产品名。
- 网络用语,如“Orz”、“屌丝”、“白富美”、“高富帅”、“尼玛”等大量网络新词。
- 英文单词,如“fall”、“room”、“know”、“root”等。
- 缩写词,如“TMD”、“SB”、“CNM”等。

• 繁体词语,如“奮鬥”、“聖誕”、“煩惱”等。

• 长词,如“和妈妈一起”、“整形美容”、“忽然好想吃”、“很大程度上”、“化妆很简单”等。

• 不准确词语,如“了不少”、“它采用”、“给我来一”、“? 我说:”等。

从上面列出的词语可以看出,新词发现的效果还是比较不错的。本文对识别的新词进行了5次抽样,随机抽取500个词语,手动人工标注评价,成词平均准确率为82.7%,如图1所示。

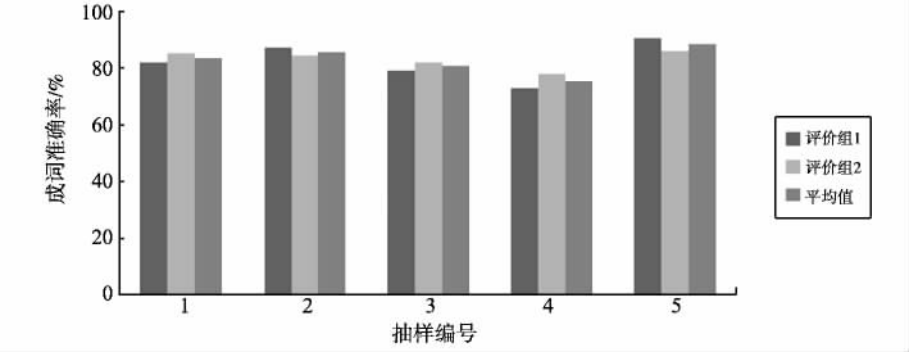

图1 新词成词准确率

Fig. 1 Precision rate of new words

对结果进行分析会发现,诸如“给我来一”这类词语本身左邻接熵满足阈值,由于右邻字符有“碗”、“个”、“瓶”、“箱”、“桶”等,所以所计算的右邻接熵也会比较高,达到阈值,根据规则也会被误判为一个词语。同样的道理,诸如“了不少”这类词语,左邻接熵也会比较高,所以也会误判为是一个词语。但为了在更大程度上保留新词,阈值选取人为地调低,误判在所难免。

3 词向量

在自然语言处理中,如果要使用机器学习的方法去解决一个问题,第一步就是要找一种方法将特征符号数学化,其中词语一般用词向量去表示。本文所描述的词向量(distributed representation)是指采用神经网络训练出来的词语向量,其基本的思想是通过训练将语料中的词语映射到N维实数向量。这种词语的表示方式优于One-hot Representation(建立一个词库,向量维度等于词表大小,词语表示为对应维度为1),N维向量不但包含了词语间的潜藏语义关系,同时也避免了维数灾难①。

本文采用Google开源词向量工具word2vec(https://code.google.com/p/word2vec/)进行词向量的训练,其中模型选择的是Skip-Gram模型,如图2所示。假设语料中有一组w_1、w_2、w_3、…、w_t词语序列,Skip-Gram模型最大化的目标函数[6-7]如公式(5)所示:

$$F=\frac{1}{T}\sum_{t=1}^{T}\sum_{-b\leqslant i\leqslant b,i\neq 0}\log p(w_{t+i}\mid w_t)。\tag{5}$$

其中,b是决定上下文窗口大小的常数,b越大训练时间会增加,同时精确度也会提高。本文窗口大小选择的是6,同时选择了Hierarchical Softmax[7]方法去训练Skip-Gram模型,最终训练出的词语向量维度为100维。

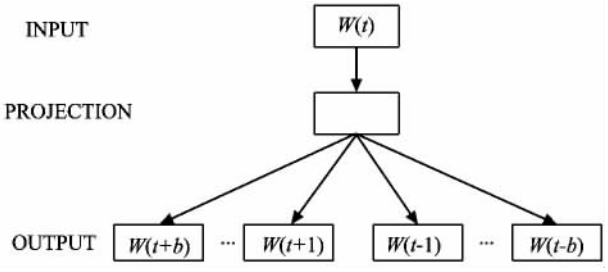

图2 word2vec中的Skip-Gram模型

Fig. 2 Skip-Gram model in word2vec

① 维数灾难(curse of dimensionality):通常是指在涉及到向量的计算的问题中,随着维数的增加,计算量呈指数倍增长的一种现象。

4　情感新词

为了更多地发现和保留语料中的新词,本文将所识别的 16 597 个新词作为分词系统的自定义词典,使用结巴分词系统(https://github.com/fxsjy/jieba)对语料进行了分词处理。这样可以弥补采用 HMM 方法[8]进行分词的过程中新词发现能力不足的问题。对语料进行分词处理之后,采用节 3 介绍的方法训练词向量,然后利用词语向量中所蕴含的语义关系,挖掘情感新词。

4.1　倾向性判断

本文采用词语相似度计算的方法,使用大连理工大学 DUTIR 情感词典本体[9]中已经标记的情感词作为种子词语,获取其 100 维向量,按公式(6)计算语料中与其 Cosine 相似度(http://en.wikipedia.org/wiki/Cosine_similarity)大于 0.8 的词语,加入备选词词典,然后通过节 4.1.1 和节 4.1.2 所描述的方法,发现情感新词。

$$\text{similarity}(\boldsymbol{A},\boldsymbol{B}) = \cos(\boldsymbol{A},\boldsymbol{B}) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\| \times \|\boldsymbol{B}\|}。\tag{6}$$

通过计算词向量 Cosine 相似度得到的与“尼玛”最相似的新词如表 3 所示。从结果中可以看出,利用神经网络训练出来的词语向量包含了语料潜在的语义信息,通过词语之间的相关度计算结果表明是有效的。

表 3　语料中与“尼玛”最相近的词语

Table 3　The words close to “nima” in corpus

词语	相似度	词语	相似度
泥马	0.924 357 53	气死我了	0.862 207 65
次奥	0.903 864 30	劳资	0.859 399 20
我勒个去	0.881 188 33	你妹	0.847 774 74
特么	0.874 074 50	我操	0.846 065 34
坑爹啊	0.862 696 65	老纸	0.845 525 40
坑爹呀	0.862 365 25	nnd	0.839 742 66

4.1.1　权重递增法

算法 1 在使用种子词典寻找备选词的同时,对词语进行倾向性的判断。

算法 1　权重递增法

输入　大连理工大学 DUTIR 已标注情感词作为种子词集合 $S1$,备选词集合 $S2$。

输出　情感新词及其情感倾向性

步骤

(1) 对于 $S1$ 中的每个词 T,如果 T 为正面情感词 T.pos 初值设置为 3,T.neg 设置为 1。若 T 为负面情感词,初值设置相反;

(2) 循环遍历 $S1$ 中每个种子词 T,选取语料中与 T 相似度大于 0.8 的词语 W,若 W 不在 $S2$ 中,设置 W.pos = 0,W.neg = 0,将 W 加入 $S2$ 中;

(3) 待(2)结束之后,循环遍历 $S2$ 中每个备选词 W,计算 W 与 $S1$ 中每个种子的相似度 P = Similarity(W, T),同时更新 W.neg = W.neg + P* T.neg,W.pos = W.pos + P* T.pos 直到循环结束;

(4) 比较 T.neg 与 T.pos 的大小,设置阈值,判定是正面还是负面情感词。

4.1.2　SVM 分类法

从已标记的情感词中获取 100 维词向量作为特征输入,然后进行 SVM 分类器的训练。本文随机选取了大连理工大学情感本体中 7 000 情感词进行训练,其中正向情感词、负向情感词、中性词的比例为 1∶1.5∶1,采用五倍交叉验证进行训练。

如图 3 所示,在 SVM 参数训练的过程中,当横坐标为 3、纵坐标为 -5 的时候,准确率达到最大,为 79.581 4%。将节 4.1.1 所得到的备选词集合 $S2$ 中每个备选词 T 的 100 维向量作为输入,用训练出的分类模型,预测备选词的情感分类。

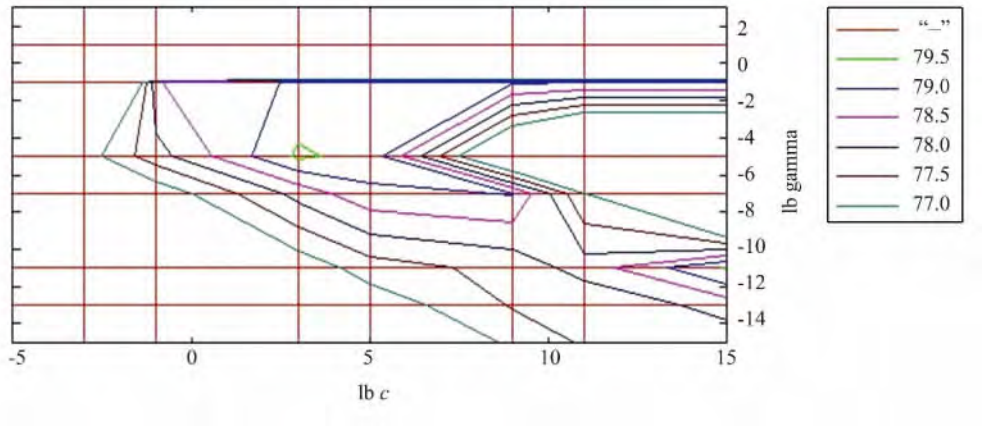

图 3 SVM 参数训练
Fig. 3 Parameters of the SVM training

4.1.3 中心向量法

人工选取语料中的情感词语,并标注其情感倾向性。将标注为同一类的情感词,对应词语向量各维度相加然后取平均,从而计算出此类别情感词的中心向量。如公式(7)所示,v_c 表示中心向量,x 为 C 类中的词向量,$|C|$ 表示 C 类中词向量的个数。然后通过此中心向量与语料中未标注词语的词向量进行 Cosine 相似度计算:

$$v_c = \frac{\sum_{x \in c} x}{|C|}。 \tag{7}$$

本文手动选取"屌丝"、"弱爆"、"坑爹"、"吐槽"、"尼玛"、"BT"、"猥琐"、"郁闷"、"废柴"、"闷骚"、"打酱油"、"忧桑"、"艾玛"、"重口味"这 15 个情感词作为负面情感词的代表,所得结果如图 4(a)所示。以"高兴"、"开心"、"快乐"、"漂亮"、"给力"、"欢喜"、"可爱"、"满足"作为正面情感词的代表,所得结果如图 4(b)所示。图中与向量中心节点越近的节点表示其与中心向量相似度得分越高。

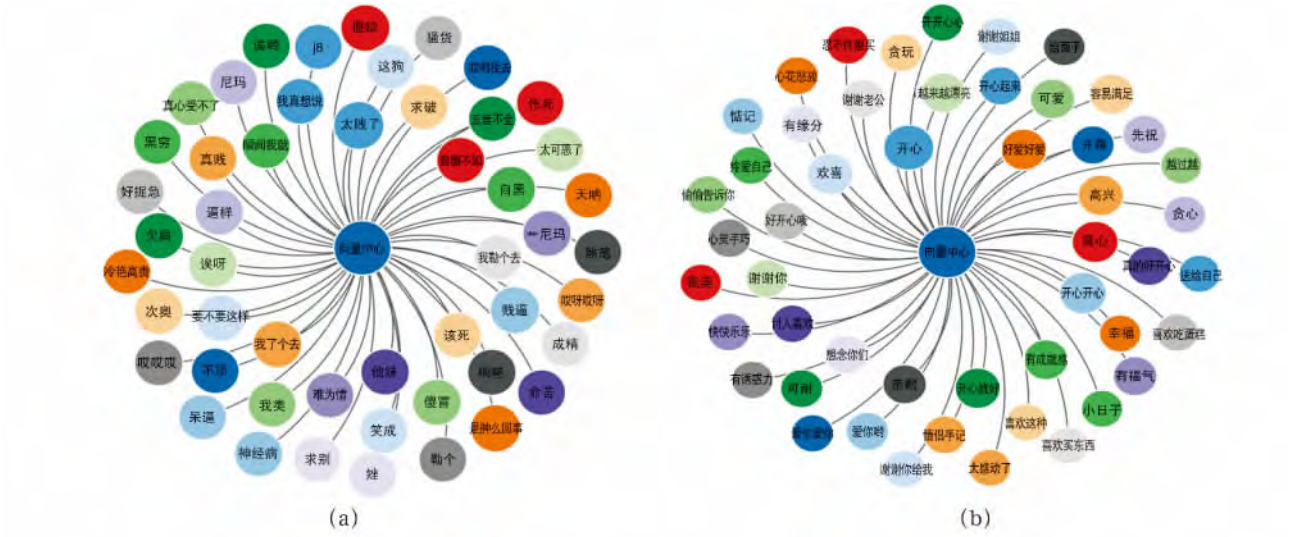

图 4 与中心向量最相似词语 Top 50
Fig. 4 Top 50 words close to central vector

4.2 结果分析

采用节 4.1.1 和 4.1.2 的方法,通过与 COAE 2014 发布的评测结果进行对比,整体实验结果并不理想(表 4),主要原因有如下几点:

(1) 实验以大连理工大学情感词汇本体中的所有情感词作为种子词,由于情感词汇本体中存在一些情感倾向性不明显的词汇,同时情感本体中大部分词语都为常用的情感词,所扩展出的备选词也多为常用情感词以及一些与情感词搭配较多的专有名词,在一定程度上影响了实验的结果,所扩展出的备选新词不太具有代表性。由于节 4.1.1 和 4.1.2 的方法都采用了相同的备选新词,所以导致 W_P(情感词正确率)、W_R

(情感词召回率)、W_F(情感词 F 值)结果相同,而针对于情感词极性的判断从结果中可以看出节 4.1.1 的方法,B_P(情感词极性正确率)、在 B_R(情感词极性召回率)、在 B_F(情感词极性 F 值)方面都要优于节 4.1.2 的方法。

(2) 通过对 COAE 2014 所提供的标注答案进行分析,部分标注与本文实验结果有偏差,例如:①针对微博"移动的服务态度很温馨,让人有亲切的感觉,价格也越来越优惠了!",评测答案中对其标注的情感新词为"温馨"。②针对微博"鲁迅和卡夫看都是享受悖谬的高手",评测答案中对其标注的情感新词为"悖谬"。但是"温馨"和"悖谬"这两个词都在给定的旧词典当中,不能定义为新词。这也从一定程度上影响了本文方法的评测结果。

表 4　COAE 2014 评价结果

Table 4　The evaluation results of COAE 2014 task3

方法	W_P/%	W_R/%	W_F	B_P/%	B_R/%	B_F
DUTIR1	4.846	5.266	0.050	3.344	3.634	0.034 829 739
DUTIR2	4.846	5.266	0.050	3.117	3.387	0.032 463 958
MEDIAN	16.743	10.070	0.118	10.870	5.579	0.067 934 096
MAX	49.702	20.961	0.207	21.793	16.834	0.166 141 662

虽然节 4.1.1 和 4.1.2 的方法并未达到评测的预期效果,但是作为一种情感新词识别的新方法,为情感新词发现以及在此基础上的文本情感分析提供了新的研究思路。

针对节 4.1.3 的方法,本文使用 $P@N$ 来评价此结果,$P@N$ 表示按相似度得分排序,前 N 个实验结果中,负(正)面情感词的准确率。由于评价标注过程中,个人对词语的辨识能力会带来评价结果的偏差,所以本实验分 5 组人员去评价,每组 2 人,最终取平均值。负面情感词结果如表 5 所示,正面情感词结果如表 6 所示。

表 5　负面情感词评价结果

Table 5　The evaluation results of negative emotional words

评价组	$P@50$	$P@100$	$P@200$	$P@500$	$P@1\,000$
评价组 1	0.74	0.76	0.750	0.628	0.498
评价组 2	0.76	0.75	0.735	0.656	0.620
评价组 3	0.80	0.78	0.730	0.682	0.663
评价组 4	0.70	0.72	0.690	0.670	0.564
评价组 5	0.62	0.69	0.715	0.690	0.648
平均值	0.72	0.74	0.724	0.665	0.599

表 6　正面情感词评价结果

Table 6　The evaluation results of positive emotional words

评价组	$P@50$	$P@100$	$P@200$	$P@500$	$P@1\,000$
评价组 1	0.70	0.64	0.550	0.522	0.493
评价组 2	0.56	0.54	0.460	0.372	0.287
评价组 3	0.73	0.66	0.610	0.542	0.435
评价组 4	0.64	0.61	0.570	0.474	0.442
评价组 5	0.68	0.65	0.565	0.432	0.415
平均值	0.66	0.62	0.551	0.468	0.414

可以看出,评价组不同,在结果的辨别上会有或多或少的差异。但整体趋势是随着 N 的增大,无论是正面情感词还是负面情感词,部分非相关词都会逐渐增加,而准确率 P 明显降低,如图 5 所示。关于 $P@N$ 会随着 N 增大而降低的原因分析如下:

(1) 从实验结果看,随着 N 的增加,非相关词中不成词的词语如:"个屁啊"、"可惜俺"等也在增加从而降低了准确率。从各组标注的结果中可以明显地看出,正面情感词的非相关词数量要大于负面情感词的非相关词数量,所以本文节 4.1.3 所提出的中心向量法对负面情感词语的识别要优于正面情感词。

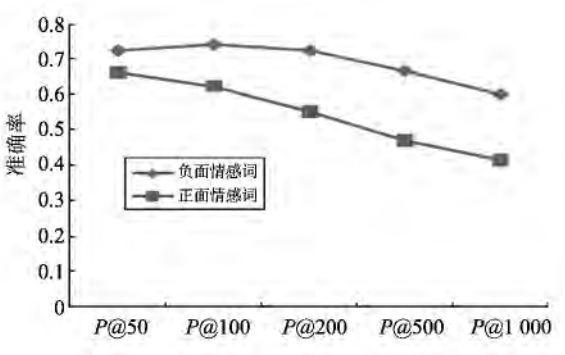

图 5　情感词评价结果

Fig. 5　The evaluation results of emotional words

(2)对于千万级别的微博语料而言,主题的非单一性,不同主题下词语用法的不同,都会给词语向量的训练带来干扰,也影响到了后续的相似度计算。一些跨主题的专有名词如人名、地名、品牌名等随着 N 的增大,也会有所增长,所以情感词的准确率降低。

5 结论

本文所提出的方法是在词向量应用方面的一个初步尝试。基于词向量的情感新词发现,只需要利用语料本身信息进行词向量的训练,无需外部语义资源,比较容易实现。同时从实验结果来看,也有一定的实用价值。文中所使用的种子词对结果影响比较严重,所以如何准确挑选种子词语,以及在实验中如何对各个主题下的微博进行情感新词的单独领域发现,各个主题下的情感中心向量如何确定,这是进一步研究的方向。

参考文献:

[1] 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.
ZHU Yanlan, MIN Jin, ZHOU Yaqian, et al. Semantic orientation computing based on HowNet [J]. Journal of Chinese Information Processing, 2006, 20(1):14-20.

[2] 王素格,李德玉,魏英杰,等. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报,2009,23(5):68-74.
WANG Suge, LI Deyu, WEI Yingjie, et al. A synonyms based word sentiment orientation discriminating [J]. Journal of Chinese Information Processing, 2009, 23(5):68-74.

[3] 唐都钰. 领域自适应的中文情感分析词典构建研究[D]. 哈尔滨:哈尔滨工业大学,2012.
TANG Duyu. Research on domain adaptive Chinese sentiment lexicon construction [D]. Harbin: Harbin Institute of Technology, 2012.

[4] HUANG J H, POWERS D. Chinese word segmentation based on contextual entropy [C] // Proceedings of the 17th Asian Pacific Conference on Language, Information and Computation. Singapore, 2003:152-158.

[5] YE Yunming, WU Qingyao, LI Yan, et al. Unknown Chinese word extraction based on variety of overlapping strings [J]. Information Processing & Management, 2013, 49(2): 497-512.

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013-10-23) [2014-02-23]. http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

[7] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013:3111-3119.

[8] 黄昌宁,赵海. 中文分词十年回顾[J]. 中文信息学报,2007,21(3):8-19.
HUANG Changning, ZHAO Hai. Chinese word segmentation: a decade review [J]. Journal of Chinese Information Processing, 2007, 21(3):8-19.

[9] 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.
XU Linhong, LIN Hongfei, 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.

(编辑:许力琴)

(上接第36页)

[7] RENDLE S. Factorization machines [C] // Proceedings of the 10th IEEE International Conference on Data Mining (ICDM2010). Los Alamitos: IEEE Computer Society, 2010:995-1000.

[8] RENDLE S. Factorization machines with libFM [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3(3):57.1-57.22.

[9] BLEI D M, NG A Y, JORDAN M I, et al. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3:993-1022.

[10] WANG Yi, BAI Hongjie, STANTON M, et al. PLDA: parallel latent dirichlet allocation for large-scale applications [C] // Proceedings of Algorithmic Applications in Management (AAIM). Berlin, Heidelberg: Springer, 2009: 301-314.

[11] FAN R E, CHANG K W, HSIEH C J, et al. LIBLINEAR: a library for large linear classification [J]. The Journal of Machine Learning Research, 2008(9):1871-1874.

[12] JOACHIMS T. Training linear SVMs in linear time [C] // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06). New York: ACM, 2006: 217-226.

(编辑:许力琴)