融合直推式学习和语义理解的词语倾向性识别

闻 彬,饶 彬,赵君喆,焦翠珍,戴文华
(湖北科技学院 计算机科学与技术学院,湖北 咸宁 437100)

摘 要:目前词语情感倾向性识别研究主要分为机器学习和语义理解,机器学习不能很好地识别通用领域词语,语义理解又存在准确率和召回率不够高的问题,因此文中提出了一种融合直推式学习和语义理解的词语倾向性识别方法。首先对 HowNet 知识库体系进行改进,在已有的四种义原的基础上,提出第五义原—情感义原;然后将第五义原手工融入到 HowNet 知识库中,再在此基础上提出词语情感相似度计算方法计算词语的情感值;最后将该方法融合直推式学习以判定词语情感倾向性。通过实验结果表明,与支持向量机和原语义理解方法相比,该方法在识别情感词上取得了较好的效果。

关键词:词语倾向性识别;机器学习;语义理解;意见挖掘;情感义原;HowNet
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2016)01-0074-04
doi:10.3969/j.issn.1673-629X.2016.01.015

Identifying of Word Sentiment Orientation of Transductive Learning and Semantic Comprehension

WEN Bin, RAO Bin, ZHAO Jun-zhe, JIAO Cui-zhen, DAI Wen-hua
(College of Computer Science and Technology, Hubei University of Science and Technology, Xianning 437100, China)

Abstract: At present, the research on word sentiment orientation identification is mainly divided into machine learning and semantic comprehension, but machine learning cannot handle general field words effectively, semantic comprehension also cannot get high scores at precision and recall, therefore, a new fusion method between transductive learning and semantic comprehension for judging word polarity was put forward in this paper. Firstly the HowNet knowledge base system is improved, on the basis of four primitive, the fifth primitive—sentimental primitive was proposed, which was integrated into HowNet manually, on the basis of this, then a new word sentimental similarity calculation method was proposed to compute word's sentimental value. At last, combine this way with transductive learning for identifying word's sentimental orientation. The performance of experiment shows that compared with SVM or traditional semantic comprehension, it can get better results.

Key words: word sentiment orientation; machine learning; semantic comprehension; opinion mining; sentimental primitive; HowNet

1 概 述

由于越来越多用户乐于在互联网上分享自己的观点和意见,使得互联网中这类信息迅速膨胀,仅靠传统的人工方法难以有效及时地获取网上的海量信息,更难以提供准确的分析和处理,因此,迫切需要相关的自然语言处理技术来处理这些相关的评价信息。意见挖掘技术在此背景下应运而生,并引起了广泛的关注。

意见挖掘的目的是发现文本中作者所持有的主观态度,为产品推荐、舆情监控和观点抽取等提供支持。现有的意见挖掘技术主要分为基于语义理解的和基于机器学习的。其中基于机器学习的方法典型的有:朴素贝叶斯(Naïve Bayes, NB)、支持向量机(Support Vector Machine, SVM)、最大信息熵(Maximum Entropy, ME)等。

机器学习方法在处理特定领域语料时有着较高的准确率,但是分类器设计复杂,训练语料标注工作繁琐,同时,当涉及到通用语料时,机器学习往往不能得到较好的效果。而基于语义理解的方法则可以解决这类问题。语义理解的方法从情感词出发,构建文本的情感模型,从而判断出文本的情感倾向性,因此,如何

收稿日期:2015-04-20 **修回日期**:2015-07-22 **网络出版时间**:2016-01-04
基金项目:国家自然科学基金面上项目(61373108);湖北省教育厅科研项目(Q20112809, B20082803);湖北省教育厅人文社会科学研究项目(13g389)
作者简介:闻 彬(1982-),男,讲师,硕士,研究方向为自然语言处理、机器学习。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160104.1453.016.html

识别情感词是语义理解方法的核心。

目前国内外研究词语倾向性的方法主要分为两种—基于统计学的方法和基于语义理解的方法。基于统计学的方法主要是利用机器学习来获取词语的情感倾向性。

在英文方面,Hatzivassiloglou 和 McKeown[1] 使用监督学习的方法对词语进行情感语义倾向性判别;Turney 等[2] 利用点互信息(PMI-IR)方法搜索引擎的"NEAR"操作来计算待定词与具有强烈倾向性的种子词集合的关联程度;Yu 等[3] 挑选出若干极性较强的形容词(情感词)构建一个种子词集合,通过计算新词和种子词的共现概率来判断新词的语义倾向性。在文本情感分类方面,Pang 等[4] 利用人工标注语料,分别使用朴素贝叶斯、最大熵和支持向量机三种分类模型对影视文本进行分类,Sinno Jialin Pan[5]、Xavier Glorot[6] 和 Blitzer[7] 等众多学者利用领域适应算法分析文本的情感倾向性;Wan[8] 利用已有的英文情感语料库完成中文文本的情感分类。基于语义理解的方法主要有基于现存的本体知识库,例如中文的 HowNet 和英文的 Wordnet。在英文处理方面,Jaap 等[9] 利用 WordNet 的同义词关系确定形容词的褒贬;Baccianella 等[10] 基于 WordNet 构建了认可度最高的 SentiWordNet;Maks 和 Vossen[11] 基于词典模型进行情感分析和意见挖掘;在中文处理方面,具有代表性的是朱嫣岚等[12] 采用基于 HowNet 的语义相似度和语义相关场两种方法计算词语的倾向性。同时国内很多学者[13-14] 研究建立情感词典来处理观点挖掘等问题,但是到目前为止还没有一部权威的情感词典可供借鉴。

因此文中首先在 HowNet 知识库定义的四个义原的基础上,人工添加 HowNet 第五义原—情感义原[15],然后利用改进的 HowNet 知识库计算词语之间的情感相似度,再融合直推式学习判定情感词极性。

2　融合直推式学习和语义理解的词语倾向性识别

2.1　基于 HowNet 的情感词判别方法

HowNet 语义相似度的方法反映词语语义的相似程度,也即两个词语在不同上下文环境中在词语替换的情况下不改变文本句法语义结构的程度。因此,利用词语的语义相似度概念计算词语的情感值。

HowNet 中若词语有多种表达含义,则词语有多个义项,每个义项又由多个义原组成。那么词语的语义相似度计算实际上是义原的相似度计算[16]。

对于两个词语 Word_1 和 Word_2,假设词语 Word_1 有 n 个义项 $Y_1, Y_2, \cdots, Y_n$,词语 Word_2 有 l 个义项 $Z_1, Z_2, \cdots, Z_l$,则词语的相似度计算如式(1)所示:

$$\mathrm{Sim}(\mathrm{Word}_1, \mathrm{Word}_2) = \max_{i=1,2,\cdots,n, j=1,2,\cdots,l} \mathrm{Sim}(Y_i, Z_j) \tag{1}$$

将词语相似度的计算转换成概念之间的相似度计算。

2.2　HowNet 义原相似度计算

在 HowNet 中用义原表示词语概念,所以概念相似度计算就是义原相似度计算。

由于所有义原构成了一个树状义原层次体系,因此可以使用公式(2)计算两个义原 p_1, p_2 之间的语义距离。

$$\mathrm{Sim}(p_1, p_2) = \frac{\alpha}{d + \alpha} \tag{2}$$

其中,d 是 p_1 和 p_2 在树状义原层次体系中的路径距离;α 是一个可调节的参数。

2.3　概念情感相似度计算

在 HowNet 知识库中概念分成四个义原:"第一基本义原"、"其他基本义原"、"关系义原"和"符号义原"。但是 HowNet 中的这四种义原的相似度计算没有考虑词语的情感语义。词语概念 S_1, S_2 之间的相似度计算如式(3)所示。

$$\mathrm{Sim}(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} \mathrm{Sim}_j(S_1, S_2) \tag{3}$$

文中在计算情感相似度时引入了情感义原作为词语概念的第五义原,并人工挑选 HowNet 中的情感词加入第五义原:"desired/良"、"undesired/莠"。

词语 S_1', S_2' 的情感相似度 $\mathrm{Sem}(S_1', S_2')$ 定义如式(4)所示。

$$\mathrm{Sem}(S_1', S_2') = \begin{cases} 1 & S_1' \text{ 第五义原} = S_2' \text{ 第五义原} \\ 0 & \text{其他} \end{cases} \tag{4}$$

引入情感义原后,词语 S_1', S_2' 的相似度 $\mathrm{Sim}(S_1', S_2')$ 计算如式(5)所示。

$$\mathrm{Sim}(S_1', S_2') = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} (\theta_1 \mathrm{Sim}_j(S_1', S_2') + \theta_2 \mathrm{Sem}(S_1', S_2')) \tag{5}$$

其中,$\mathrm{Sim}_1(S_1', S_2')$ 表示概念 S_1' 与 S_2' 的"第一基本义原"的相似度;$\mathrm{Sim}_2(S_1', S_2')$ 表示概念 S_1' 与 S_2' 的"其他基本义原"的相似度;$\mathrm{Sim}_3(S_1', S_2')$ 表示概念 S_1' 与 S_2' 的"关系义原"的相似度;$\mathrm{Sim}_4(S_1', S_2')$ 表示概念 S_1' 与 S_2' 的"符号义原"的相似度;$\mathrm{Sem}(S_1', S_2')$ 表示概念 S_1' 与 S_2' 的"情感义原"的相似度;β_i 是可调节参数,其中 $i = 1, 2, 3, 4$,$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$,从 Sim_1 到 Sim_4 对于总体相似度所起的作用依次递减,因此设置的参数满足 $\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \beta_4$;$\theta_1, \theta_2$ 用来设置知网义原与情感义原的权重,$\theta_1 + \theta_2 = 1$。

2.4　基于概念情感相似度的词语情感语义值

计算出词语概念情感相似度之后,结合训练集对

测试集中的词语计算情感值。计算方法如式(6)。

$$\text{Sentiment}(\text{word}) = \frac{1}{n}\sum_{i=1}^{n}\text{Sim}(\text{word},\text{Set\_}P_i) - \frac{1}{m}\sum_{j=1}^{m}\text{Sim}(\text{word},\text{Set\_}N_j) \quad (6)$$

其中,Sentiment(word) 表示测试集中词语 word 的情感值; Sim(word,Set_P_i) 表示词语 word 与褒义训练集 Set_P_i 的相似性; Sim(word,Set_N_j) 表示词语 word 与贬义训练集 Set_N_j 的相似性。

2.5 直推式学习

通过上面的基于 HowNet 的情感词计算方法,可以得到每个词语的情感值。文献[15]中实验证明,该方法可以取得较好的实验效果,因此在此方法的基础上进行进一步研究,将直推式方法融入其中。将每次判定出来的情感词加入到训练集中,如果判定该词语为褒义情感词,则加入到褒义测试集中;如果判定为贬义情感词,则加入到贬义训练集中;若属于中性词,则放回待测测试集中。然后用新的训练集和测试集重复该工作,直到所有词的极性不再改变为止,显而易见,该过程必然是收敛的,直推式算法详细过程如下所示。

Step1:建立训练集和测试集;

Step2:对测试词集利用文中提出的方法计算词语情感值,并判定词语的情感倾向性;

Step3:若待判定词语判定为正面情感词,则从测试集中移动到正面训练集中;若为负面情感词,则从测试集移动到负面训练集中;若为中性词,则将该词放回测试集中等待下一次判定;

Step4:重复 Step2-3 直到测试集和训练集中的词语不再改变。

3 实验结果及分析

首先构造出初始训练集和测试集。为了达到更好的实验效果,尽量选择极性较强的中文词语作为训练集,具体训练集组成如表 1 所示,其中褒义贬义各包含 20 个情感词。

表 1 训练集

极性	词语
褒义	好,健康,快乐,优秀,美丽,善良,勤劳,孝顺,富有,高大,优雅,时尚,聪明,坚强,乐观,赞美,贤惠,完美,矜持,温柔
贬义	坏,猥琐,萎靡,奸诈,歹毒,丑陋,愚蠢,阴暗,白痴,懒惰,不良,虚假,腐败,缺陷,寒酸,恶心,弊病,废物,下贱,脆弱

为了能够达到较好的通用性,文中从新浪、网易、百度三大平台下载新闻语料 12 854 篇,然后利用中科院分词工具 ICTCLAS 对文本进行分词处理;再根据停用词表删除停用词;由于词语中只有名词、形容词和动词才存在情感,因此抽取出所有的名词、形容词和动词,最后进行人工调整得到测试集词语共 6 961 个,其中褒义情感词 1 989,贬义情感词 2 056,中性词 2 916。

对于知网知识库中的词语,人工标注“desired/良”和“undesired/莠”,标注数据如表 2 所示。

表 2 良莠标注情况

极性	词性			总数
	形容词	名词	动词	
desired/良	845	225	192	1 262
undesired/莠	635	341	365	1 341

对于 2.3 中的参数,文献[15]中对 θ_1,θ_2 设置进行了实验,并根据实验结果发现当设置为 0.7 和 0.3 时可以达到最好的实验效果。在 HowNet 中对参数 $\beta_1,\beta_2,\beta_3,\beta_4$ 分别设置为:0.5,0.3,0.15,0.05。对 2.4 中的计算词语情感值的阈值,文献[15]也进行了讲解,并将其设置如式(7)所示。

$$\text{Sentiment}(\text{word}) = \begin{cases} \geqslant 4.1 \text{ 褒义} \\ \text{其他 中性} \\ \leqslant -4.1 \text{ 贬义} \end{cases} \quad (7)$$

实验利用三种方法进行验证:支持向量机(Support Vector Machine,SVM)、原语义理解方法(Semantic Comprehension,SC)以及融合直推式学习和语义理解(Transductive Learning & Semantic Comprehension,TL&SC)。利用准确率(Precision)、召回率(Recall)和 F(F-measure)值作为判定准则。其中 SVM 方法中将褒义词、贬义词和中性词平均分成三部分,然后以其中一部分作为训练集,另外两部分作为测试集,依次替换三部分角色。基于篇幅限制,表 3 列出的 SVM 结果是循环三次后所取得的平均值。SC 和 TL&SC 方法的实验结果见表 4 和表 5。

表 3 SVM 实验结果

极性	Precision	Recall	F-measure
褒义	0.668	0.650	0.659
贬义	0.690	0.710	0.700
全部	0.612	0.677	0.643

表 4 SC 实验结果

极性	Precision	Recall	F-measure
褒义	0.829	0.742	0.783
贬义	0.876	0.632	0.739
全部	0.853	0.680	0.757

表 5 TL&SC 实验结果

极性	Precision	Recall	F-measure
褒义	0.910	0.880	0.895
贬义	0.944	0.831	0.884
全部	0.927	0.854	0.889

三个实验数据对比如图 1 所示。

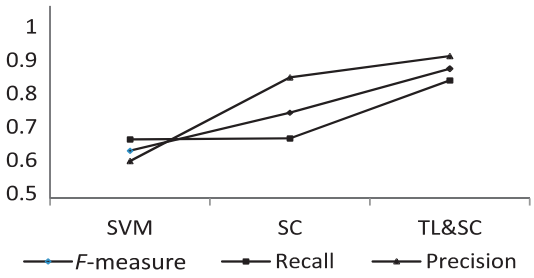

图 1　三种方法结果比较

从图中可以很容易看出,在处理通用文本时,SVM 方法得分都不是很高;当使用文中提出的 SC 方法时,准确率有明显提升,但是不足的是召回率不能达到较高效果;最后使用 TL&SC 时,可以看到,不管是准确率、召回率还是 *F* 值,相对于其他两种方法,都达到了较为理想的效果。

4　结束语

文中所提方法利用了 HowNet 知识库计算词语的情感相似度,然后根据计算得到的词语情感值结合阈值来判断词语的情感倾向性,再将该方法融入直推式学习中。文中针对支持向量机、原语义理解方法和融合语义理解和直推式学习三种方法分别进行了实验,结果表明,针对通用领域获取的词语,第三种方法不论在准确率、召回率还是在 *F* 值上都有明显的性能提升。

当然,文中方法也存在不足之处:由于针对单个词语判定情感倾向性,这样势必忽略了特定语义环境下词语的情感倾向性,如何获取这些情感词是未来的研究方向之一;同时文中利用 ICTCLAS 进行分词、词性标注处理,这样会忽略掉许多网络(非常态)用语,而这些网络用语却表达了极强的极性,如果能结合网络环境判定出这些词语也是未来的重要研究方向。

参考文献:

[1] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives[C]//Proceedings of the 35th annual meeting of association for computational linguistics and the 8th conference of the European chapter of the ACL. [s. l.]:[s. n.],1997:174-181.

[2] Peter T, Michael L. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[3] Yu Hong, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences[C]//Proc of EMNLP-03. Sapporo, Japan: [s. n.], 2003: 129-136.

[4] Pang Bo, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[C]//Proceedings of the 2002 conference on empirical methods in natural language processing. Philadelphia: Association for Computation Linguistics, 2002: 79-86.

[5] Pan S J, Ni Xiaochuan, Sun Jiantao, et al. Cross-domain sentiment classification via spectral feature alignment[C]//Proceedings of the 19th international conference on World Wide Web. [s. l.]: [s. n.], 2010: 751-760.

[6] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: a deep learning approach[C]// Proc of 28th international conference on machine learning. Bellevue, WA, USA: [s. n.], 2011.

[7] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boomboxes and blenders: domain adaptation for sentiment classification[C]//Proc of ACL. [s. l.]: [s. n.], 2007: 187-205.

[8] Wan Xiaojun. Co-training for cross-lingual sentiment classification[C]//Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP of the AFNLP. [s. l.]: [s. n.], 2009: 235-243.

[9] Kamps J, Marx M, Mokken R J, et al. Using WordNet to measure semantic orientation of adjectives[C]//Proceedings of the 4th international conference on language resources and evaluation. Lisbon, Portugal: [s. n.], 2004: 1115-1118.

[10] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of the 7th conference on international language resources and evaluation. Valletta, Malta: [s. n.], 2010: 2200-2204.

[11] Maks I, Vossen P. A lexicon model for deep sentiment analysis and opinion mining applications[J]. Decision Support Systems, 2012, 53(4): 680-688.

[12] 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.

[13] 徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报,2007,21(1):96-100.

[14] Wen Bin, Dai Wenhua, Zhao Junzhe. Sentence sentimental classification based on semantic comprehension[C]//Proc of fifth international symposium on computational intelligence and design. [s. l.]: [s. n.], 2012: 458-461.

[15] 闻　彬,何婷婷,罗　乐,等. 基于语义理解的文本情感分类方法研究[J]. 计算机科学,2010,37(6):261-264.

[16] 刘　群,李素建. 基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会. 台北:出版者不详,2002.

融合直推式学习和语义理解的词语倾向性识别

作者： 闻彬， 饶彬， 赵君喆， 焦翠珍， 戴文华， WEN Bin， RAO Bin， ZHAO Jun-zhe， JIAO Cui-zhen， DAI Wen-hua

作者单位： 湖北科技学院 计算机科学与技术学院,湖北 咸宁,437100

刊名： 计算机技术与发展 ISTIC

英文刊名：

年，卷(期)： 2016(1)