加权维吾尔语句子倾向性分析①

年 梅[1], 刘若兰[1], 玛尔哈巴·艾赛提[2], 范祖奎[3]

[1](新疆师范大学 计算机科学技术学院, 乌鲁木齐 830054)
[2](新疆师范大学 文学院, 乌鲁木齐 830054)
[3](新疆警察学院 语言系, 乌鲁木齐 830011)

摘 要: 准确可靠的文本倾向性分析是网络舆情分析与网络内容安全的前提. 本文提出了利用中文极性情感词典 HowNet、NTUSD 以及大连理工大学发布的褒贬情感词词典进行并交运算, 选择并翻译为维吾尔语词汇, 借助于维吾尔语同义近义词词典, 扩展构建了维吾尔语极性情感词典; 然后分析总结了否定词、程度副词以及句中的转折连词等情感修饰成分对维吾尔语句子情感极性的影响, 并量化为情感词权值; 最后设计了基于维吾尔语极性情感词和权值相结合的加权句子情感极性判定算法. 利用自建语料库进行测试, 并与汉语倾向性判定实验结果比较, 证明了本算法进行维吾尔语句子褒贬情感性分析基本是有效地.

关键词: 极性情感词; 情感修饰成分; 加权算法; 情感倾向性分析; 维吾尔语

Analysis of the Sentence Tendency in Uighur Language

NIAN-Mei[1], LIU Ruo-Lan[1], MARHABA Asat[2], FAN Zu-Kui[3]

[1](College of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China)
[2](School of Literature, Xinjiang Normal University, Urumqi 830054, China)
[3](School of Languages, Xinjiang Police Academy, Urumqi 830011, China)

Abstract: The precondition of network public opinion analysis and network content security is based on the accurate and reliable text tendency analysis. In this paper the Chinese emotion dictionary which included HowNet, NTUSD and emotion word dictionary released by Dalian university is implemented intersection operation and union Operation. By selecting, translating for Uighur vocabulary, with the synonymous dictionary, the Uighur emotional polarity dictionary is constructed. At the same time the impact is analyzed and summarized about the Emotional Modifier such as the effect of negative word, degree adverbs and the Sentence's adversative conjunction etc. Then the effect degree is converted into emotional weight value. In the end the weighted algorithm based on the emotion words and the modifier words is designed to decide a sentence emotional polarity. The experiment result proves that this algorithm is basic validity compared to Chinese emotional tendentiousness algorithm by testing on self-built corpus.

Key words: polar emotion words; emotional modifier; weighted algorithm;sentiment orientation analysis; Uyghur

1 引言

随着移动通信技术和各种社交软件的迅猛发展, 网络成为人们购物、交流以及抒发情感的重要平台. 越来越多的人通过网络了解、评论社会热点事件, 网络评论影响到成千上万网民的观点, 从而导致一些群体性事件的发生, 如网络中的暴力搜索、利用网络传播攻击政府的反动言论等. 掌握网络舆情的主动性是保证社会安定团结的前提, 据此, 政府能够迅速掌握民众的关注热点, 激发社会正能量, 快速化解社会矛盾. 国家对网络舆情分析工作越来越重视, 而网络文本倾向性分析是支持舆情分析的前提.

句子情感倾向分析是网络文本倾向性分析的基础,

① 基金项目:国家自然科学基金(61163064);教育部人文社会科学工程科技人才培养专项(15JDGC022);新疆"十一五"规划项目(070708);新疆师范大学校级教学改革研究一般项目(SDJGY2014-01)

收稿时间:2015-11-20;收到修改稿时间:2016-01-14 [doi:10.15888/j.cnki.csa.005277]

所谓句子的情感分析是指识别出一个句子中作者对评价对象所持的态度是肯定还是否定, 或者支持还是反对. 现阶段句子情感分析技术主要有两类, 一类是基于语言情感知识的方法, 另一类是机器学习的方法.

基于情感知识的方法是指利用情感词典确定词语的极性值, 通过分析修饰词对情感词极性的影响、文本的语义距离、否定词修饰等得出句子倾向性. 如: Hu 和 Liu 等基于 WordNet 知识库中的同义、反义关系, 构建情感极性词典, 然后根据句中词汇情感倾向判定句子的极性[1]. 文献[2]提出了基于句法路径的自动识别情感评价单元并获取评价对象及评价词语间的关系. 文献[3]利用文献[4]的成果构建了情感极性词库, 通过句法分析, 借助于语义和语法距离衡量句子的褒贬性. 文献[5]设计了 SBV 极性传递算法, 通过主题与情感描述项的关系计算主题极性. 文献[6]则提出采用否定模式匹配的方法, 即通过词语情感极性受到的否定词共享影响, 研究句子的倾向性. 文献[7]提出了汉语中基于情感词加权的句子倾向性识别算法等.

机器学习的方法是将情感分类看作一种特殊的文本分类, 通过机器学习算法训练标注好的训练集得到分类模型, 再由分类模型确定测试文本的倾向性. 如文献[8]将极性情感词作为句子的特征值, 通过 SVM 分类器分析褒贬性. 文献[9]则比较了不同的特征提取方法、文本分类方法, 结论是: 使用 BiGrams 特征表示方法、信息增益特征选择方法和 SVM 分类方法情感分类效果较好. 文献[10]基于中心词对短语的极性进行研究, 通过计算中心词的极性值和短语内两个词的 MI 计算短语的极性, 获取文本倾向识别.

从现有研究成果可知, 情感倾向分析一般建立在极性情感词基础上, 利用上下文关系或者机器训练确定文本的倾向性. 目前维吾尔语文本倾向性分析还处于起步阶段, 共享的语义资源还未见发布, 缺乏大规模标注的情感语料库, 基于机器学习算法的维吾尔语句子倾向性识别在现阶段难以实现. 为此, 本文采用了基于情感知识的句子倾向性分析算法, 借鉴中文情感语义资源成果, 筛选翻译建设了维吾尔语极性情感词典, 以句中情感词为核心, 结合情感词修饰成分的作用, 设计了加权算法计算维吾尔语句子的极性, 为维吾尔语文本倾向性分析研究提供技术支持.

2 维吾尔语情感词表的建立

目前, 中英文情感词典的研究已有了一定的成果, 如英文情感词典 General Inquirer(GI)(http://www.wjh.harvard.edu/~inquirer/), 中文情感词典 HowNet、台湾发布的 NTUSD、大连理工大学发布的情感词表以及《褒义词词典》和《贬义词词典》等. 本文充分调研分析了中文情感语义研究的成果, 对 HowNet、NTUSD 和大连理工大学发布的情感词典通过筛选、翻译, 建立了维吾尔语极性情感词典.

具体步骤是: 从以上词典中筛选出同时存在于两个以上词库中的词汇, 也就是将上述三个情感词典两两相交, 再对三个交集进行并运算, 最终的集合中包含了褒义词 2727 个, 贬义词 3055 个. 组织维吾尔语计算语言学的研究生, 选出了维吾尔语中使用且具有情感极性的词汇, 利用电子版《维软大辞典》软件, 将其翻译为维吾尔语词汇.

经过上述处理, 获取词汇 4540 个, 其中褒义词 2003 个, 贬义词 2537 个. 由于词汇总量不足, 考虑到同义词和近义词的情感极性相同, 使用维吾尔语同义词典和近义词典对以上词汇进行了扩展, 形成了本文的极性情感词库, 总词汇数为 9342 个, 其中褒义词 4325 个, 贬义词 5017 个.

3 基于情感词语义加权的维吾尔语句子倾向性识别方法

3.1 情感词的语义加权倾向

极性情感词对句子的情感倾向判定具有举足轻重的作用, 情感词如同句子中的基因, 修饰该基因的其它成分与其共同决定了句子的情感倾向. 极性情感词的修饰词包括了否定词、程度副词和连词等. 为了设计句子倾向性判定算法, 分析总结维吾尔语句中情感词的各种修饰成分及其对句子极性的影响, 转换为相应的权值, 与情感词相结合定量句子的倾向性.

极性情感词经常被程度副词修饰, 程度副词能够增强或者减弱句子情感倾向的程度. 例如“بەك ناچار(太糟糕)”, “ناھايىتى چىرايلىق(非常美)”等, 句子中的情感词“ناچار(糟糕)”和“چىرايلىق(美)”已经分别具有贬义和褒义, 但在其前加了“بەك(太)”以及“ناھايىتى(非常)”等程度副词后, 情感倾向变得更加强烈. 本文搜集整理了维吾尔语中使用的 57 个程度副词, 组织计算语言学的研究生根据维吾尔语中程度副词对句子情感倾向的褒贬影响, 并借鉴中文程度副词对情感词的权值定义[11], 最终将其划分为最、很、稍、欠 4 个级别, 定义了每

个级别的权值. 表 1 给出了本文使用的维吾尔文中的程度副词的分级以及权值.

表 1 维吾尔文程度副词表

程度级别	程度副词	权重	个数
最	ئاجايىپ(十分)، ھەددىدىن زىيادە(格外)، ئالاھىدە(分外)، ئەڭ(最)، ئالىي(极)، پەۋقۇلئاددە(极)، تازا(顶)، تامامەن(完全)، نەقەدەر(多么)، زور(极)، شەكسىز(无疑)، قەتئىي(坚决)، مۇتلەق(绝对)، نېمىدېگەن(何等)، ھەقىقەتەن(的确)، ھەددىدىن ئارتۇق(过)، پۈتۈنلەي(完全)، ئۇزۇل-كېسىل(彻底)، ھەرگىز(绝不)، كارامەت(棒)	1.5	20
很	بارغانسېرى(越)، بولۇپمۇ(尤其是)، بەك(很)، بەكمۇ(太)، تېخىمۇ(更)، يەنىمۇ(更)، ئىلگىرلىگەن ھالدا(进一步)، تولىمۇ(更)، تولا(太)، خوپمۇ(更)، زەپمۇ(不过)، قانچىكى(越发)، ناھايىتى(最)، ئەنتايىن(非常)، ئارتۇقچە(过于)	1.25	15
稍	بىر قەدەر(比较)، خېلى(挺)، قارىغاندا(比)، نىسبەتەن(相对)، ھەرھالدا(还可以)، ھەممىدىن(比…最)، يەنە تېخى(还)	0.75	7
欠	ئازراق(有点儿)، ئانچىكى(不怎么)، ئانچە(不那么)، ئازغىنا(些许)، بىرئازمۇ(一点儿也不)، ئاز بىر(略微)، بىرئازمۇ(丝毫)، سەل(微)، سەل-پەل(稍)، قىلچە(毫不)، كىچىككىنە(稍微)	0.5	11

否定词也经常修饰情感词, 当否定词直接修饰情感词时, 会改变情感词的极性, 但倾向程度不变. 与汉语不同的是, 维语中否定词修饰情感词时, 位于情感词之后, 例如, “رازى(满意)”为褒义, “رازى ئەمەس(不满意)”则为贬义(ئەمەس为表示“不”的否定词). 维吾尔语中的否定词数量较少, 主要使用了“يوق(没/没有)”, “ئەمەس(不/不是)”两个否定词. 此外, 在维吾尔语中, 有种特殊方式表达否定, 即使用词缀表示否定. 如“ ما , مە”是构成动词否定式的词缀, 用在部分情感词干后面, 表示否定的意义. 基于以上分析, 本研究参考汉语否定词对情感词的权值定义方法[12], 将修饰情感的否定词和包含否定词缀的修饰权值均定义为-1.

除了以上在各种语言中均使用的情感词修饰成分外, 维吾尔语中还有一些特殊的词缀影响着情感词的极性. 具有“ -پەرۋەر ، -ئانە ”后缀的词汇在维吾尔语中, 均表示褒义. 具有“-نا, -بى،” 两个前缀和“-سىز”后缀的词汇在维吾尔语中, 均表示贬义. 本文将具有这些词缀的情感词作为褒贬义词处理, 定义包含这些成分的情感词的权值分别为 1 和-1.

关联词会影响句子的情感倾向程度, 尤其是具有递进关系的关联词, 例如句子“她不但美丽而且大方”, 该句中包含了两个褒义词“美丽”和“大方”, 但句子后半部分的褒义程度更加强烈. 为此, 本文选择了维吾尔语中最常用的两个递进关联词“(不仅……还)بۇلۇپلا....قالماستىن يەنە”和“(不但……而且)بۇلۇپلا قالماستىن....بەلكى”, 参考文献[12]中定义递进连词对句子情感值的影响, 以递进关联词作为句子的分割点, 将句子分为前后两部分, 前半部分修饰权值定义为 1, 后半部分修饰权值定义为 1.5.

此外, 在汉语中还经常出现否定词和程度副词结合在一起使用的情况, 如“程度副词+否定词+情感词”或者“否定词+程度副词+情感词”等. 但维吾尔语中只有“程度副词+情感词+否定词”一种情况, 且使用频率低. 例如句子“ناھايىتى ياخشى ئەمەس(非常不好)”中, “ياخشى(好)”是褒义词, 后面加了否定词“ئەمەس(不)”之后, 首先倾向性发生了翻转, 由于情感词前增加了程度副词“ناھايىتى(非常)”, 则倾向性程度比“不好”更为强烈. 考虑到“程度副词+情感词+否定词”这种修饰成分在维语中不常用, 故不考虑不同级别程度副词的影响, 定义“程度副词+情感词+否定词”的修饰权值为-1.5.

3.2 基于维吾尔语情感词的语义加权倾向算法

基于情感词加权的极性判定算法首先识别句中的极性情感词, 以句子中的情感词为中心将句子分为几个词群块; 用情感词的极值与修饰成分的权值乘积运算得到词群块的极性值, 如公式(1)所示; 然后对词群块极性值加和获取句子的情感极性值, 算法如公式(2)所示.

$$OS(WordGroup) = W(DW) * V(EW) \tag{1}$$

$$\begin{aligned} OS(Sentence) &= \sum_{i=1}^{n} OS(WordGroup_i) \\ &= \sum_{i=1}^{n} W(DW_i) * V(EW_i) \end{aligned} \tag{2}$$

$$V(EW) = \begin{cases} 1 & EW \in \{褒义词\} \\ 1 & EW \in \{含有两个褒义词缀的词\} \\ -1 & EW \in \{贬义词\} \\ -1 & EW \in \{含有三个贬义词缀的词\} \end{cases} \tag{3}$$

$$W(DW)=\begin{cases}-1 & DW\in\{否定词缀\}\\ -1 & DW\in\{否定词\}\\ 1.5 & DW\in\{极性增强副词集\}\\ 1 & DW\in\{极性保持副词集\}\\ 0.5 & DW\in\{极性减弱副词集\}\\ -1.5 & DW\in\{程度副词+否定词\}\end{cases} \quad (4)$$

其中, OS(WordGroup) 是词群的情感极性值, OS(Sentence)指句子的倾向值. 算法中, EW 指极性情感词, V(EW)指情感词的极性值, V(EW)的赋值如公式(3)所示. DW 指情感词的修饰成分, W(DW)指情感词修饰成分的权值, 并按照公式(4)赋值. n 表示句中词群的数量. 公式(2)的计算结果作为句子极性判别的依据, 结果大于 0, 判别该句子为正向; 结果小于 0, 则为负向; 结果等于 0, 则判定句子无极性.

3.3 加权倾向算法的描述

为了进行情感倾向的判别, 本文使用了 5 个词表. 第一个词表是维吾尔语情感词典, 词典中共包括词汇 9342 个, 其中褒义词 4325 个, 贬义词 5017 个. 第二个词库是程度副词词库, 一共包括了 57 个词汇. 第三个是否定词库, 包括了否定词缀和否定词各两个. 第四个是包含两个递进关联词的关联词表. 最后一个是情感词词缀表, 包括了表示贬义的两个后缀和一个前缀及表示褒义的两个后缀. 以情感词为句子核心, 设计了维吾尔语句子的情感极性判别算法, 具体算法描述如下:

输入: 维吾尔语句子 *Sentence*

输出: 句子的情感极性值 *OS*(*Sentence*)

步骤 1: 扫描整个句子, 搜索句中是否有递进关联词, 如果有, 则以关联词为分割点, 将句子划分为前后两部分 , 对前后两部分分别按照步骤 2 进行处理, 处理完毕后转步骤 4; 如果没有关联词, 按照步骤 2 处理;

步骤 2: 扫描句中的情感词, 以情感词为中心, 将句子分为词群块, 对每个情感词词群块进行以下处理:

① 以情感词为中心, 根据维吾尔语情感词典确定情感词的极值(包括含三个贬义词缀和两个褒义词缀的状况).

② 向后搜索, 如果情感词无否定词缀或情感词的后面无否定词, 转③; 如果情感词有否定词缀或情感词的后面是否定词, 再对情感词的前面部分进行搜索, 判别情感词前是否有程度副词, 如果有, 则将情感词群修饰成分的权值定义为-1.5, 转④; 如果无程度副词, 该词群权值为-1, 转④;

③向前搜索, 如果情感词前面是程度副词, 检索程度副词表, 按照表 1 中程度副词所属级别根据公式(4)定义其权值, 转④; 否则, 该词的修饰成分的权值默认为 1.

④ 根据公式(1)计算该情感词所在词群的情感极值

步骤 3: 按照公式(2), 对每个情感词群的情感极值进行加和运算, 转步骤 5.

步骤 4: 按照公式(2), 分别将关联词前和关联词后的每个情感词词群极值加和, 然后将关联词后加和的结果乘以 1.5 再与关联词前加和结果再进行加和运算. 然后转步骤 5.

步骤 5: 根据计算结果判断句子极性, 如果结果大于 0, 则判定该句子为褒义; 若小于 0, 判定该句子为贬义; 若结果等于 0, 判定该句子无倾向.

4 实验结果及分析

4.1 测试参数

对本文倾向性判别算法的性能, 使用了准确度、召回度以及 F1 三个参数值衡量. 其定义如下:

准确度(Accuracy):

$$P=\frac{正确分类的句子数}{测试集中分为该类的句子总数}\times 100\%$$

召回率(Recall):

$$R=\frac{正确分为某类的句子数}{测试集中属于该类的句子总数}\times 100\%$$

F1 测试值:

$$F1=\frac{2\times P\times R}{P+R}\times 100\%$$

4.2 实验测试

为了验证算法的有效性, 自建了语料进行实验测试. 搜集了 2013 至 2015 年期间的维吾尔语新闻语料、论坛评价语料, 总计 5000 句, 其中正面句子 1280 句, 反面句子 2373 句. 情感词表使用了本文构建的维吾尔语极性情感词典. 首先使用了基于情感词极性加和的句子倾向性判别算法[13]进行句子的分类, 得到的实验结果如表 2 所示.

表2　基于极性情感词典的维吾尔语句子极性判别结果

情感类别	评价指标		
	准确率(%)	召回率(%)	F 值(%)
正面	75.7	92.4	83.22
负面	78.5	52.3	62.78
无倾向	73.2	76.7	74.91

由表 2 可以得知, 利用本文建立的维吾尔语极性倾向词进行极性判别的算法, 与参考文献[13]中汉语词汇判别算法相比, 其准确率和召回率相差不大, 从而证明本文建立的维吾尔语极性情感词典是有效的. 性能参数略低于汉语测试结果的原因在于, 本文选用的情感词典多数基于汉语而来, 而且通用词居多, 网络极性情感词相对较少, 本文测试语料主要基于网络文本, 故性能参数略低于汉语极性情感词判别结果. 从表中还可以发现, 对负面句子的判别性能略低于正面句子, 原因在于, 负面句经常使用了“褒义词+否定成分+”的方式进行表达, 而仅使用极性情感词不考虑句子的修饰成分, 导致该词群的极性成分判断错误, 从而影响到整个句子的分类结果.

为了比较本文加权判别算法的性能, 使用了相同的语料, 按照2.3的算法进行了实验. 结果如表3所示.

表 3 基于维吾尔语情感词的语义加权倾向算法

情感类别	评价指标		
	准确率(%)	召回率(%)	F 值(%)
正面	86.5	94.5	90.32
负面	85.7	67.5	75.51
无倾向	85.3	87.7	86.48

从表 3 结果可以看出, 在考虑了修饰成分的影响后, 句子倾向性判别的各项指标有了明显的提高, 尤其是负面句子正确识别率和召回率提高的较为明显. 从而证明了在句子倾向性判别中同时考虑到情感词极性及其修饰成分的影响, 对判别结果的准确率有明显提高.

5 结束语

维吾尔语文本倾向性识别技术的研究目前处于起步阶段, 还没有太多有价值的资源和成果, 但维吾尔语文本倾向性分析对网络舆情分析和网络内容过滤等有重要的意义. 维吾尔语句子倾向性识别是维吾尔语文本倾向性识别的一个重要的基础性工作. 本文首先借用了中文极性情感词以及否定词等的成果, 经过统计、分析和选择建立了维吾尔语的基础情感词典、否定词典、程度副词词典等; 通过对维吾尔语极性句子的分析, 提出了维吾尔语句子情感倾向判定的语言模型, 以句中的情感词为核心, 定义了情感词修饰成分的情感权值; 最后设计了基于句中的情感词以及情感词语义加权的维吾尔语句子极性识别方法. 实验结果表明, 利用本文方法获得的句子倾向性识别的判全率、判准率和 F 值等指标和汉语句子倾向性判别成果的性能基本相同, 从而显示了本文算法的合理性和有效性. 今后, 研究组将会深入研究网络文本的极性情感词以及表达方式, 为进一步改善网络文本倾向性分类性能不断努力.

参考文献

1 Hu MQ, Liu B. Ming and summarizing customer reviews. Proc. of the Tenth ACM. SIGKDD. 2004. 168−177.

2 赵妍妍,秦兵,车万翔,等.基于句法路径的情感评价单元识别.软件学报,2011,22(5):887−898.

3 熊德兰,程菊明,田胜利.基于 HowNet 的句子褒贬倾向性研究.计算机工程与应用,2008,44(22):143−145.

4 朱嫣岚,闵锦,周雅倩,等.基于 HowNet 的词汇语义倾向计算.中文信息学报,2006,20(1):14−20.

5 姚天昉,娄德成.汉语语句主题语义倾向分析方法的研究.中文信息学报,2007,21(5):75−78.

6 党蕾,张蕾.一种基于知网的中文句子情感倾向判别方法.计算机应用研究,2010,27(4):1370−1372.

7 赵鹏,赵志伟,卓景文.一种情感词语义加权的句子倾向性识别方法.计算机工程与应用,2011,47(35):161−163.

8 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.中文信息学报,2007,21(1):96−100.

9 唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研究.中文信息学报,2007,21(6):88−94.

10 李钝,曹付元,曹元大,等.基于短语模式的文本情感分类研究.计算机科学,2008,35(41):132−134.

11 邸鹏.基于句子情感权值合成算法的篇章情感分析[学位论文].太原:太原理工大学,2015.

12 刘玉娇,琚生根,伍少梅,苏狮. 基于情感字典与连词结合的中文文本情感分类.四川大学学报(自然科学版),2015,01:57−62.

13 王素格,杨安娜,李德玉.基于汉语情感词表的句子情感倾向分类研究.计算机工程与应用,2009,45(2)4:153−155,161.