藏文情感词典的构建及微博情感计算研究

孙本旺[1],田 芳[2]
(1. 青海大学 计算机技术与应用系,青海 西宁 810016;
2. 青海大学 信息化技术中心,青海 西宁 810016)

摘 要:针对国内尚缺乏系统的藏文情感词典,提出借助中文情感词典资源自动构建藏文情感词典的方法,并基于构建的藏文情感词典对藏文微博进行情感分析研究。首先,通过合并去重算法、字符串匹配算法等自动地构建了藏汉情感词典;然后,通过去重算法得到藏文情感词典和藏文停用词词典;最后,通过加权叠加微博中的情感词或情感短语相应的权值来研究藏文微博的情感倾向。实验自动构建了藏文情感词典,包含基础情感词、程度词、否定词、转折词、双重否定词、藏文停用词。基于实验构建的藏文情感词典,与其他藏文情感词典相比,有效地提高了藏文微博情感倾向分类的准确率。实验结果表明,该词典达到了良好的实用性。

关键词:中文情感词典;藏汉情感词典;藏文情感词典;藏文微博;权值;情感分类
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2018)11-0212-05
doi:10.3969/j.issn.1673-629X.2018.11.046

Research on Construction of Tibetan Emotional Dictionary and Emotional Computing of Micro-blog

SUN Ben-wang[1], TIAN Fang[2]
(1. Department of Computer Technology and Applications, Qinghai University, Xining 810016, China;
2. Information Technology Center, Qinghai University, Xining 810016, China)

Abstract: In view of the lack of a systematic sentiment dictionary of Tibetan in China, we propose a method to automatically construct a Tibetan emotion dictionary by using Chinese emotion dictionary resources, and conduct an emotional analysis of Tibetan micro-blog based on the constructed Tibetan emotion dictionary. First, Tibetan-Chinese sentiment lexicon are automatically constructed by using the merge to weight algorithm and the string matching algorithm. Then, the Tibetan emotional dictionary and the Tibetan dictionary of words which are never used again are obtained through the de-weight algorithm. Finally, corresponding the weight of the emotional words and emotional phrases respectively to study the emotional bias inclination of Tibetan micro-blog. The experiment automatically builds the Tibetan emotional dictionary, including the basic emotional words, degree words, negative words, turning words, double negative words and words which are never used again. This Tibetan sentiment dictionary constructed, compared with other Tibetan sentiment dictionaries, can effectively improve the accuracy of sentiment classification of Tibetan micro-blog. The experiment shows that the dictionary has achieved strong practicability.

Key words: Chinese emotion dictionary; Tibetan-Chinese sentiment lexicon; Tibetan emotion dictionary; Tibetan micro-blog; weight; emotional analysis

0 引 言

藏文情感词典的构建研究是自然语言处理的重要部分,也是藏文文本情感分析的基础。基于藏文情感词典的情感计算,主要是通过藏文基础情感词、藏文程度词、藏文否定词等来实现,因此藏文情感词典构建的好坏直接影响情感分类的结果。利用已有的中文情感词典资源自动构建藏文情感词典,不但能解决藏文词典构建费时费力的问题,还能保证藏文情感词典拥有足够多的词汇量。藏文情感词典构建方法的研究将有利于推动藏文词典的构建研究、藏文文本的情感倾向

收稿日期:2017-12-12 **修回日期**:2018-04-18 **网络出版时间**:2018-06-29
基金项目:国家自然科学基金(61461045);青海省科技计划项目(2016-ZJ-743)
作者简介:孙本旺(1990-),男,硕士研究生,研究方向为自然语言处理;田 芳,博士,教授,通信作者,研究方向为自然语言处理、语义关系抽取、本体自动构建等。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20180629.1704.028.html
万方数据

分析研究。

1　相关研究

1.1　藏文情感词典构建研究现状

祁坤钰构造了一个英藏机器翻译的藏语语义分类体系,并提出了藏语语义词典设计的理论框架、语义分类思想和属性描述原则[1];邱莉榕等研究了藏文语义本体中的上下位关系模式匹配算法来构建藏文语义本体[2];柔特基于 Word Net 对比英文和藏文词之间的语义关系、构建双语大型数据库和制定映射过程中词汇空缺等方法,构建了基于半自动匹配的藏文语义词典[3];巴桑卓玛等人工收集和整理了一部藏文基准情感词典,在此基础上,基于词向量扩充情感词典,并利用 KNN 扩充法自动扩建藏文情感词,最终建立了一部比较实用的藏文情感词典[4];杜雪峰提出利用藏汉双语词典和 Hownet 相结合的办法来构建藏文词典,构建了包含基础情感词、否定词、转折词和程度副词等的藏文词典[5];Zhang Zhen 等利用英文情感词典和藏英词典相结合的方式来构建藏文基本情感词典[6];Yan X 等通过人工手段建立了一个全面的、高效的极性词典,其中包括基本词典、程度副词词典等藏文词典[7]。

1.2　藏文微博情感分析研究现状

麻省理工学院的 Picard 教授最早提出了情感分析的概念。Picard 教授在 1995 年发表了论文《Affective Computing》[8],并在两年后在此论文的基础上撰写了有关情感计算的最早同名论著[9]。在藏文情感分析方面,扎西本等基于情感词典的褒贬义词、转折词及否定词来计算藏文句子的情感倾向[10];张俊等通过借鉴中文微博情感分析中比较常见的基于统计的方法和基于词典的方法对藏文微博进行情感分析,实验结果是基于藏文词典的藏文微博情感分析的准确率明显高于基于 TF-IDF 的藏文微博情感分析的准确率[11];普次仁等将藏文分词后,用词向量表示词语,把藏文语句变为由词向量组成的矩阵,利用无监督递归自编码算法对该矩阵向量化,有监督地训练输出层分类器以预测藏文语句的情感倾向[12];袁斌针对藏文微博中存在的藏汉混排问题,提出了一种基于语义空间的藏文微博情感表示方法,该方法通过句法树实现了语义向量化,提高了情感特征中的语义成分,并解决了多语言混合文本处理问题[13];李苗苗提出了藏文文本情感分析的词语级、句子级、篇章级三层框架,提出了利用情感词典和规则集分析藏文句子情感的一种方法,采用 SVM 算法对篇章级进行情感分析[14];江涛等提出了基于多特征的情感倾向性分析算法,算法以情感词、词性序列、句式信息和表情符号作为特征,并针对藏文微博常出现中文表述的情况,将中文的情感信息也作为特征进行情感计算,利用双语情感特征有效提高了情感倾向性分析的效果[15]。

2　藏文情感词典的构建

研究藏文情感词典的构建,将有利于藏文微博的情感倾向分析,推动藏文的网络舆情分析、机器人情感识别等应用研究。

2.1　现有情感词典的基本信息

基于现有的中文情感词典构建情感基本信息库。收集了清华大学褒贬义词典(包含情感正反极性,褒义词 4 468 个,贬义词 5 567 个),台湾大学 NTUSD 词典[16](包含情感正反极性,褒义词 2 810 个,贬义词 8 276个),Hownet 词典[17](包含情感正反极性,褒义词 4 766 个,贬义词 4 370 个,程度词分为最(most)、很(very)、较(more)、稍(ish)、欠(insufficiently)、超(over)、以及否定词表),以及哈工大停用词词典(包含停用词 767 个)。《藏汉大辞典》[18]中所有单词表现格式如下所示:

"ཀུན་རྟོག(名)1. 动机。2. 等起(佛);起因。3. 普遍发动。4. 观念;念头;发心;思想"。经过预处理得到"ཀུན་རྟོག动机 等起 起因 普遍发动 观念 念头 发心 思想"。收集了《藏汉大辞典》中所有单词经过预处理得到的 27 402 条数据。

2.2　TSD 的构建

按先行经验,通过基于 Hownet 词典和《藏汉大辞典》来构建藏文情感词典(Tibetan sentiment dictionary,TSD)。TSD 词典包含基础情感词、程度词、否定词、转折词。构建流程如图 1 所示。

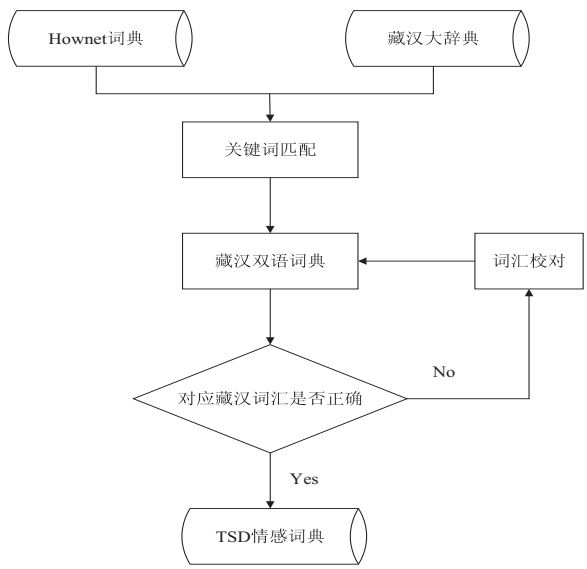


图 1　TSD 藏文情感词典的构建

2.3　SSTSD 的构建

主要采用合并去重算法,首先将 Hownet 词典和清华大学褒贬义词典合并,然后将合并结果与台湾大学

万方数据

NTUSD 词典合并去重,最后将合并结果与《藏汉大辞典》通过匹配算法来构建藏文情感词典(Sanjiang source Tibetan sentiment dictionary,SSTSD),SSTSD 情感词典用 Hownet 词典表达的 TXT 表示。利用哈工大停用词典与《藏汉大辞典》通过匹配算法来构建藏文停用词词典。首先,通过匹配算法查找相应词汇自动构建藏汉双语词典;然后校对检验对应词汇的正确性;最后提取藏文词汇,并利用去重算法得到正确的 SSTSD 藏文情感词典和停用词词典。SSTSD 词典除了转折词、双重否定词等小部分通过人工翻译和校对得到,绝大部分是自动构建。SSTSD 词典包含基础情感词、程度词、否定词、转折词、双重否定词、藏文停用词。构建步骤如图 2 所示。

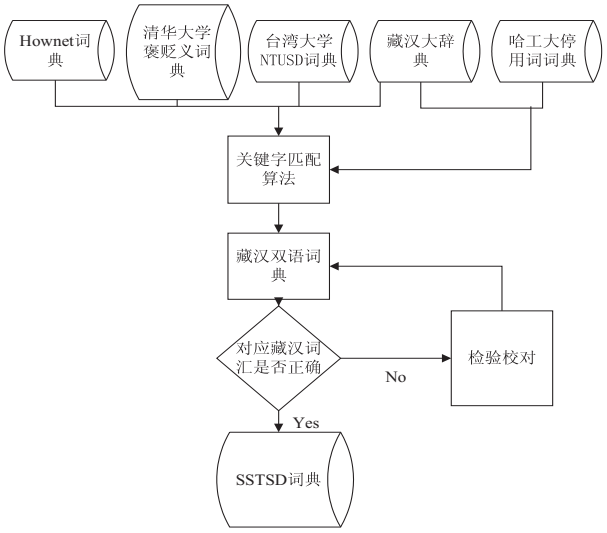


图 2 SSTSD 词典的构建

3 藏文微博的情感倾向分析

3.1 情感倾向分析方法

基于情感词典的藏文微博情感分析的方法一般是通过微博中情感词或情感短语的权值叠加计算来判断某条微博的情感倾向。在进行微博情感计算时首先要考虑去停用词,然后考虑藏文微博是否含有转折词。微博中的转折词可以改变整条微博的情感倾向,因此对藏文微博情感倾向的判断分两种情况处理,即微博包含转折词的情况和微博未出现转折词的情况。

3.1.1 未出现转折词

如果微博中包含褒义情感词或情感短语,则按照式 1 来计算:

$$W\_Emotion\_Result(i) = \sum_{i=1}^{\infty} W\_Emotion(x) \quad (1)$$

其中, $W\_Emotion\_Result(i)$ 表示褒义情感类在微博中通过累加计算的情感值; $W\_Emotion(x)$ 表示微博中第 x 个褒义情感词的情感权重值,最终通过叠加微博中的褒义情感值得到某条微博在褒义情感类下的情感值。

如果微博中包含贬义情感词或情感短语,则按照式 2 来计算:

$$W\_Emotion\_Result(j) = \sum_{j=1}^{\infty} W\_Emotion(x) \quad (2)$$

其中, $W\_Emotion\_Result(j)$ 表示贬义情感类在微博中通过累加计算的情感值; $W\_Emotion(x)$ 表示微博中第 x 个贬义情感词的情感权重值,最终通过叠加微博中的贬义情感值得到某条微博在贬义情感类下的情感值。

如果微博中未包含情感词或情感短语,则判定此条微博的情感倾向是中性的。

3.1.2 出现转折词

在微博中一般只包含一个转折词,如果出现了转折词,直接取转折词后面的微博部分。如果转折词后面的微博部分中包含褒义情感词或情感短语,则按照式 1 来计算;如果转折词后面的微博部分中包含贬义情感词或情感短语,则按照式 2 来计算;如果转折词后面的微博部分中未包含情感词或情感短语,则判定此条微博的情感倾向是中性的。

3.2 情感倾向计算

因为藏文语料较少且没有开放,文中采用人工切分好的藏文微博语料进行实验。在微博的情感倾向计算中,首先将藏文微博分词,去停用词。将褒贬义情感词的初始权值都赋值为 1;然后查找对应情感词的程度词级别,不同级别的程度词将会给出不同的程度级别。如果微博中含有奇数的否定词则情感倾向取反,含有双重否定词则情感值要适当增强;最后微博得分是整条微博的褒贬义倾向权值的差值。如果差值大于零则微博为正向倾向,如果差值小于零则微博为负向倾向,否则微博为中性倾向。

算法步骤如下:

```
Input:
Set micro-blog positivescore:positive
Set micro-blognegative score:negative
Set degree word Wi:={most,very,more,-ish,insufficiently,over};
Set degree level W_Degree(Wa):={2,1.75,1.25,0.75,0.5,1.5};
Set micro-blog number M:={M1,M2,…,Mk};
Output:
Micro-blog scores: Scores(k);
for each Mk do
for word in Mk
if(word in stop_word)
Delete word;
else if(word in turn_word)
Mk=word.next;
```

万方数据

```
else if(word in pos_word)
positive=W_Emotion_REsuit(i);
if(word.pre==Wi)
positive=W_Emotion_REsuit(i)*W_Degree(Wa);
else if(word in neg_word)
negative=W_Emotion_REsuit(j);
if(word.pre==Wi)
negative=W_Emotion_REsuit(j)*W_Degree(Wa);
else if(word.pre==deny_word)
positive=(-1)*positive;
negative=(-1)*negative;
else if(word.pre==double-deny_word)
positive=2*positive;
negative=2*negative;
for each Scores(k) do
Scores=positive-negative;
end
end
end
```

算法最后对比了基于不同藏文情感词典的情感倾向性分析结果,验证词典的正确性。

4　实验结果与分析

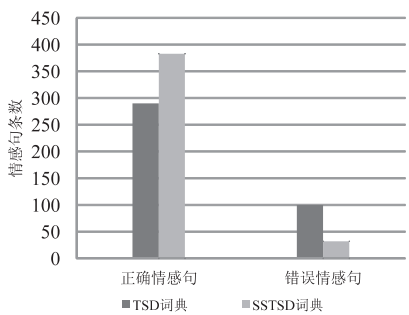
4.1　SSTSD 词典与现有词典的比较

SSTSD 词典得到了 4 183 个褒义情感词,4 823 个贬义情感词,1 427 个中性词,192 个程度词,17 个否定词,11 个转折词,13 个双重否定词。巴桑卓玛等构建的藏文情感包括 2 000 个正向情感词,2 000 个负向情感词,1 739 个中性情感词。杜雪峰构建的情感词典包含了 5 070 个藏文基础情感词,否定词的总个数为 26,双重否定词的总个数为 11,程度副词的个数为 71,转折词的个数为 2。

从以上信息可以得出,SSTSD 词典更加系统全面,具有良好的参考性。

4.2　情感计算结果分析

利用 500 条藏文微博语料,通过不同藏文情感词典计算得出微博情感倾向,将基于 TSD 词典与 SSTSD 词典得到的正确微博情感句和错误微博情感句作对比,如图 3 所示。



万方数据 3　实验结果对比

通过图 3 可以看出,在微博语料相同的条件下,基于 SSTSD 词典的正确微博情感句要明显高于基于 TSD 词典的正确微博情感句,基于 SSTSD 词典的错误微博情感句又明显低于 TSD 词典的错误微博情感句。

基于两个词典的实验结果通过评价指标进行对比,如表 1 所示。

表 1　TSD 词典与 SSTSD 词典的对比　%

指标	TSD 词典	SSTSD 词典
准确率	58	76.6
召回率	72.5	81.7
F 值	64	79.1

从表 1 可以得出,基于 SSTSD 情感词典的准确率、召回率、*F* 值相对于 TSD 情感词典,分别提高了 8.6%、9.2%、15.1%。

将基于 TSD 词典和 SSTSD 词典的准确率、召回率和 *F* 值进行对比,如图 4 所示。

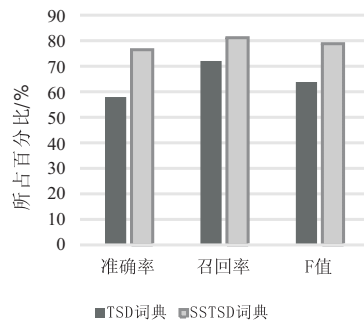


图 4　实验结果计算对比

通过图 4 可以看出,基于 SSTSD 词典的情感分类准确率、召回率和 *F* 值都大于基于 TSD 词典情感分类。TSD 词典含有情感词 4 093 个,SSTSD 词典含有情感词 10 433 个。通过分析可以看出,基于情感词典的情感倾向性分析很大程度上依赖情感词,情感词的好坏直接影响情感分类的准确性。基于情感词典的情感分析方法简单容易实现,后期要实现新情感词不断得自动添加。

5　结束语

文中构建的藏文情感词典更加系统全面,其总词量达到 10 666 个,含有基础情感词、停用词、否定词、转折词等,在以后的工作中还会不断更新完善。此外人工半自动地构建了藏文语义情感词典,情感分类更加细化,词汇量达到了 2 万多。基于 SSTSD 词典的藏文微博情感倾向分析的准确率达到了 76.6%,实验结果表明,该词典达到了实用性和参考性的价值。

参考文献:

[1]　祁坤钰.《机器翻译用现代藏语语义词典》的设计研究

[J]. 西北民族大学学报:自然科学版,2004,25(3):33-37.
[2] 邱莉榕,翁 彧,赵小兵. 藏文语义本体中的上下位关系模式匹配算法[J]. 中文信息学报,2011,25(4):45-49.
[3] 柔 特. 基于 WordNet 的藏文语义词典半自动构建方法研究[J]. 西藏大学学报,2014,29(2):48-53.
[4] 巴桑卓玛,李苗苗,高定国. 基于词向量的藏文情感词典的构建方法研究[J]. 电子技术与软件工程,2017(20):132-134.
[5] 杜雪峰. 藏文句子倾向性分析研究[D]. 北京:中央民族大学,2015.
[6] ZHANG Zhen, QIU Lirong. A sentiment calculation method based on tibetan semantic relations[J]. International Journal of Database Theory and Application, 2016, 9(9): 149-156.
[7] YAN Xiaodong, HUANG Tao. Research on construction of tibetan emotion dictionary[C]//International conference on network-based information systems. Taipei, Taiwan: IEEE, 2015: 570-572.
[8] LISETTI C L. Affective computing[J]. Pattern Analysis & Applications, 1998, 1(1): 71-73.
[9] PICARD R W. Affective computing: challenges[J]. International Journal of Human-Computer Studies, 2003, 59(1-2): 55-64.
[10] 扎西本,安见才让. 藏文句子的情感倾向研究[J]. 电脑知识与技术,2016,12(6):201.
[11] 张 俊,李应兴. 基于情感词典的藏文微博情感分析研究[J]. 硅谷,2014(20):220.
[12] 普次仁,侯佳林,刘 月,等. 深度学习算法在藏文情感分析中的应用研究[J]. 计算机科学与探索,2017,11(7):1122-1130.
[13] 袁 斌. 藏文微博情感分类研究与实现[D]. 兰州:西北民族大学,2016.
[14] 李苗苗. 藏文文本情感分析方法研究[D]. 拉萨:西藏大学,2017.
[15] 江 涛,袁 斌,于洪志,等. 基于多特征的藏文微博情感倾向性分析[J]. 中文信息学报,2017,31(3):163-169.
[16] KU L W, LO Y S, CHEN H H. Using polarity scores of words for sentence-level opinion extraction[C]//Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access. [s. l.]: [s. n.], 2007.
[17] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用,1998(3):79-85.
[18] 孙怡荪. 藏汉大辞典[M]. 北京:民族出版社,1985.

(上接第 211 页)

[10] 赵 宏,尹 磊,曹 洁,等. 多媒体终端的设计与实现[J]. 科学技术与工程,2011,10(22):5420-5425.
[11] 李 廷. 计算机信息技术存储平台的开发与应用[J]. 电子技术与软件工程,2017(23):146.
[12] 眷剑秋,包冰莹. 一种小型嵌入式 GUI 系统的设计[J]. 盐城工学院学报:自然科学版,2012,25(4):48-52.
[13] ZHANG Junfeng, ZHAO Jichun, WANG Guojie. Study on intelligent terminal system based on Andriod for distance learning[C]//Proceedings of joint international information technology, mechanical and electronic engineering conference. [s. l.]: [s. n.], 2016.
[14] KONG Bo, XIE Zhidong, BIAN Dongming, et al. A novel hybrid distributed storage strategy for space information network[C]//IEEE advanced information management, communicates, electronic and automation control conference. Xi'an, China: IEEE, 2016.
[15] LIAO Xiaofei, LI He, JIN Hai, et al. VMStore: distributed storage system for multiple virtual machines[J]. Science China: Information Sciences, 2011, 54(6): 1104-1118.
[16] 谭 霜,贾 焰,韩伟红. 云存储中的数据完整性证明研究及进展[J]. 计算机学报,2015,38(1):164-177.
[17] XU Xiaolong, ZHOU Jinglan, WANG Xinheng, et al. Multi-authority proxy re-encryption based on CPABE for cloud storage systems[J]. Journal of Systems Engineering and Electronics, 2016, 27(1): 211-223.
[18] 李 晖,孙文海,李凤华,等. 公共云存储服务数据安全及隐私保护技术综述[J]. 计算机研究与发展,2014,51(7):1397-1409.

万方数据