

基于网络知识百科的情感语义抽取研究

田 芳¹, 孙 晓², 孙本旺³

- (1. 青海大学 信息化技术中心, 青海 西宁 810016;
2. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009;
3. 青海大学 计算机技术与应用系, 青海 西宁 810016)

摘 要:针对情感词汇语义关系抽取缺乏问题,提出一种简单地利用网络知识百科抽取情感词汇语义关系的方法。情感语义关系抽取采用的是递归算法,选用网络百度百科数据源为百度汉语,抽取内容包括情感词汇、情感词汇的同义词和反义词两种情感语义关系。其次,利用抽取出的情感词汇语义关系和现有倾向词典自动扩展标注情感词汇的情感倾向。该方法有效地构建了中文情感词汇语义关系,抽取结果和现有情感词典相比提高了情感词汇数量。同时,基于现有情感词典和抽取的情感词汇间语义关系,实现了快速地扩展情感词语的倾向标注。实验结果表明,抽取获得了 2 万多个中文情感词汇及其语义关系,并通过情感词汇语义关系实现对抽取词汇的情感倾向扩展标注,准确率达到 78.1%。

关键词:情感词汇;情感语义关系;百度百科;关系抽取;情感倾向

中图分类号:TP391.1

文献标识码:A

文章编号:1673-629X(2019)06-0052-05

doi:10.3969/j.issn.1673-629X.2019.06.011

Research on Extraction of Affective Lexical Semantic Using Encyclopedia of Network Knowledge

TIAN Fang¹, SUN Xiao², SUN Ben-wang³

- (1. Information Technology Center, Qinghai University, Xining 810016, China;
2. School of Computer and Information, Hefei University of Technology, Hefei, 230009, China;
3. Department of Computer Technology and Applications, Qinghai University, Xining 810016, China)

Abstract: Aiming at the lack of semantic extraction of affective lexicon, a simple and effective method to extract the semantic relationship of affective lexicon by using knowledge encyclopedia is proposed. Recursion algorithm is adopted to extract the emotional semantic relationship. Baidu encyclopedia online data source is selected as Baidu Chinese, and the extracted content includes two kinds of emotional semantic relationship: emotional vocabulary, synonyms and antonyms of emotional vocabulary. The extracted affective lexicon is labeled the affective tendency based on the extracted semantic relationships and existing tendency dictionaries. The method effectively constructs the semantic relationships of Chinese affective lexicon, improving the number of affective lexicon compared with the existing affective lexicon. At the same time, based on the semantic relationship between the existing sentiment lexicon and the extracted emotional vocabulary, the tendency to quickly expand the emotional words is realized. The experiment shows that more than 2 thousands Chinese affective semantic words and their semantic relationships are extracted, and the emotive tendency of the extracted words is extended and labeled through the semantic relations of emotive words, with an accuracy rate of 78.1%.

Key words: affective lexicon; affective semantic; Baidu encyclopedia; relational extraction; affective tendency

0 引言

为了使计算机理解人类语言,自然语言处理研究越来越得到研究者的重视。自然语言处理技术的基础是语料知识库,如词汇语义。语义信息概念最早由 Bar-Hillel 教授和 Carnap 教授在 1953 年提出^[1]。随

后,1992 年 John F. Sowa 教授明确了语义网络 (semantic network)^[2]。从二十世纪九十年代中期,世界各国研究者研究开发了语义词典^[3],包括美国普林斯顿大学的 WordNet^[4]、美国微软的 MindNet^[5]、意大利信息研究所情感词典 SentiWordNet^[6]等。中文语义

收稿日期:2018-07-03

修回日期:2018-11-14

网络出版时间:2019-03-06

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:田 芳(1971-),女,博士,教授,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.20190306.0901.016.html>

词典比较著名的有中科院董振东先生提出的知网 HowNet^[7]、哈尔滨工业大学的同义词词林^[8]、北京大学计算语言研究所的中文概念词典^[9]、现代汉语语义词典^[10]、中国科学院声学研究所黄曾阳教授提出的 HNC 概念层次网络等^[11]。1997 年情感计算的概念由 MIT 媒体实验室 Picard 教授提出^[12],情感词典的构建研究起步。情感语义词典是文本情感分析研究的基础。情感语义词典注重情感词的义元和各类语义关系。基于情感语义词典的词汇间关系,可以更好地分析文本情感信息。但现有的情感语义词典缺乏情感语义关系,如何自动构建情感语义关系显得非常重要。

1 相关研究

国外情感词方面获取研究主要集中在情感倾向词汇获取和极性判断。WIEBE 等基于少量标注的词汇种子,根据种子分布相似性对主观形容词进行聚类,实现对未标注主观形容词的分类提取^[13]。RILOF 等采用基于模式匹配利用步步为营算法实现主观性名词抽取^[14]。BARONI 和 KAJI 等基于网络概念共现互信息识别主观性形容词^[15]。MOILANEN 等基于语素进行情感新词发现并标注等^[16]。文献[17]使用英文连词关系,抽取形容词情感倾向。Turney 等利用情感种子在网络搜索引擎查询片段,基于情感词汇的互信息 PMI 判别词汇的情感倾向的极性^[18]。

中文情感词典构建主要的研究思路为基于语料统计以及语义词典等方法。基于语料的情感词典构建方法通过统计词语间的共现频率信息或语义情感词典利用词语相似度来计算词语的语义倾向^[19]。主要使用的中文语义词典包括 HowNet 和同义词词林等。朱嫣岚等提出基于语义相似度和语义相关的计算词汇语义倾向性方法,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^[20]。李军等采用机器学习方法进行语义分类^[21]。大连理工大学信息检索研究室采用人工标注和自动分类的方法构建情感词汇本体^[22]。柳位平等采用人工挑选情感词汇结合 HowNet 语义相似度计算的方法构建了中文基础情感词词典^[23]。台湾大学整理构建了中文情感词典 NTUSD。张成功等整理了包含基础情感词典及领域词典、网络词词典及修饰词词典的综合词典^[24]。周咏梅等考虑情感词在不同语义环境的情感倾向,基于 HowNet 和 Senti-WordNet 建立中文情感词典 SLHS^[25]。林江豪等利用 SO-PMI 技术构建中文情感词典^[26]。陈建美等基于情感词汇语法特征和 CRF 自动获取情感词^[27]。金宇等提出基于直推式学习的中文情感词极性判别方法,情感词的词源来自《现代汉语大辞典》^[28]。很多学者对现有语义词典构建中文

情感词典的研究正说明了现有情感词典的不足^[29-30]。

互联网汇集了很多人的智慧和知识积累,包括各类网站和知识百科(维基百科、互动百科、百度百科等等)。利用互联网的海量知识获取概念及语义关系,已得到很多学者的共识^[31-32]。研究者主要利用互联网的海量信息,基于情感种子词汇或现有词典,利用搜索引擎返回的共现抽取情感词汇,计算情感词倾向和权值等。如阳爱民等在利用 NTUSD 和 HowNet 词典构建基础词典的基础上,选用情感种子词,基于搜索引擎构建情感词典^[33]。搜索引擎是面向全网搜索,需要经过分词、规则、互信息等过滤词汇,算法复杂度高。

综上所述,目前情感词典的自动构建研究中,主要是面向词汇的发现、情感分类和情感倾向标注,缺乏情感词汇的语义关系抽取。相对全网数据源的数据抽取,尚没有利用知识百科针对情感词汇语义关系抽取的相关研究。由此提出基于网络知识百科,获取情感词汇和词汇间语义关系的方法,同时利用抽取的情感同义词语义关系,自动扩展标注情感词汇倾向。

2 情感语义自动抽取

人类情感有复杂性,描述同类情感可选择很多相近词汇来表达,由此提出一个假设:每个情感概念一般都有同义词和反义词。基于这个假设,提出利用情感种子词汇基于同义词和反义词关系抽取情感词汇,并递归抽取新的情感词汇同义词和反义词关系的算法。

2.1 抽取流程

情感语义关系抽取采用的是递归算法,选用网络百科数据源为百度汉语。网络百科是志愿者填写,缺乏审核,选择两类标签抽取,目的是验证数据的准确性。数据源标签具体第一个是情感种子词汇的近义词和反义词标签,第二个是情感词汇的释义标签。

词汇的同义词集合和反义词集合不同,所以这两个集合分别利用递归算法抽取,即每个新抽取词汇被视为新的种子进入递归抽取其同义词或反义词。算法的输入为情感种子词汇 Seed(x),用 Seed(x)作为同义词种子 Syn-Seed(x)和反义词种子 Ant-Seed(x)。抽取结果为新的情感词汇 R(x),及其同义词集合 Syn(x)和反义词集合 Ant(x)。抽取算法对于每一个情感词汇种子,首先在过滤规则 1 和规则 2(过滤规则见 2.2)的条件下,利用数据源百度汉语中抽取其近义词 Syn(x)和反义词 Ant(x),同时抽取在数据源百度汉语中的释义 Exp(x)。然后,对于抽取获得的近义词 Syn(x)和反义词 Ant(x)经过规则 3 过滤,并分别和百度汉语抽取 Exp(x)结果进行合并去重,获得的结果添加到 Syn-Seed(x)和 Ant-Seed(x)。最后,算法返回到第一步实现递归,直到种子集合的所有都使用过,将

情感词汇种子 $R(x)$ 、同义词集合 $Syn(x)$ 和反义词集合 $Ant(x)$ 输出为抽取结果。

2.2 抽取规则

抽取规则设计主要是为了减少噪音数据被抽取,提高抽取精度和效率。抽取方法采用递归算法,错误的语义和词汇可视为噪音数据。噪音数据不仅消耗时间,且影响抽取精度。由于数据源百度汉语由非专业人士填写,其中不乏词汇的语义、拼写等错误。如检索

词汇“惊讶”,其近义词里发现拼写错误“惊呀”,“惊讶”又出现为自己的近义词。经过多次测试,提出的抽取过滤规则如表 1 所示。

一个词汇自身不可以定为同义词或反义词,由此设计规则 1;一个词不可能既是一个词的同义词又是反义词,由此设计规则 2;根据假设,非感情词汇相对的同义词和反义词较少,此外,错误拼写词和错误语义在所有的数据源里出现可能性低,由此设计规则 3。

表 1 抽取算法过滤规则

序号	规 则	目 的
1	过滤关系语义与种子一样的词	过滤错误语义词汇
2	过滤一个词同时出现在同义词和反义词	过滤错误语义词汇
3	词的同义词和反义词数少于指定阈值,且词在百度汉语里没有释义	过滤非情感词和错误拼写词汇

2.3 抽取测试与分析

抽取算法测试阶段,任选五个种子词汇抽取了情感语义词汇。表 2 是从 1 532 个词中连续取出了 30 个词示例,表中第 2 列为不通过规则 3 抽取出来的词汇,第 3 列为词汇拼写正确结果(不标注的为抽取结果书写正确的词汇),第 4 列标注了错误类型,包括 5 类错

误拼写、地方话、非词、不常用词和其他,表中标注“1”表示是该类型,不标表示不是该类型。拼写错误词基本都是情感词汇,这些词对情感词汇抽取没有任何意义;地方话基本上是情感词汇如“焦急”;非词主要是由单字和词汇组合构成,这些结果基本上也不是情感词汇;不常用词基本是情感词汇;其他主要包括非情感词如

表 2 无过滤规则 3 算法抽取的词汇示例

序号	抽取结果	正确	出错种类/个				
			错误拼写	地方话	不是词	不常用词	其他
1	但心	担心	1				
2	忧伤	忧伤	1				
3	追到	追悼	1				
4	困扰	困扰	1				
5	掉				1		
6	恶运	厄运	1				
7	凄婉	凄婉	1				
8	妨碍	妨碍	1				
9	忙碌	忙碌	1				
10	忙				1		
11	解体						1
12	吵杂	嘈杂	1				
13	焦急			1			
14	喧闹	喧闹	1				
15	思疑					1	
16	倦怠						1
17	仙游						1
18	凋射	凋谢	1				
19	默寞					1	
20	畏惧	畏惧	1				
21	走丢						1
22	消弱	削弱	1				
23	大祸						1
24	词语同情				1		
25	沉着冷静						1
26	微博						1
27	贫饕					1	
28	繁杂词语				1		
29	所有这个词				1		
30	糜烂	糜烂	1				

“仙游”、“微博”,专业用词如“解脬”是医学专用词,还有少量情感词汇。经过统计,被过滤的1 532 词汇中近48.7%为拼写错误,3%为地方话,16.9%不是常规词汇,4%为不常用词,其他为27.4%。

利用40个情感词汇种子,抽取算法结果经过合并去重,最终抽取结果记作中文情感语义词汇集合CASL(Chinese affective semantic lexicon)。CASL总计22 068个词汇。CASL实现包括中文情感词汇以及这些词汇的同义和反义两个语义关系。

对于CASL中22 068个词汇,通过和现在常用的情感词典做了比较(详见表3),结果说明抽取算法有效地抽取了情感词汇。选择4个词典,包括1个语义词典HowNet(使用情感倾向词汇)和3个常用情感词典:清华大学褒贬义词典、台湾大学NTUSD和大连理工大学的情感本体。表3中“覆盖词量”指CASL与比较词典重合的词汇数量,最高覆盖数量为10 829个词汇。结果表明,CASL有效地抽取了中文情感词汇,及其同义词和反义词等两种语义关系;此外,也表明现有的中文情感词典对情感词汇的认定不同。

表3 CASL对现有情感词典覆盖的词汇量

比较词典	覆盖词量数量/个
HowNet	3 962
清华大学褒贬义词典	4 711
台湾大学 NTUSD(简体版)	2 711
大连理工大学情感本体	829

3 情感倾向标注及结果分析

情感词汇的倾向标注词典对情感计算和分析十分重要,标注词汇的完整性和情感倾向分析精度直接相

关。对CASL词汇的情感倾向标注方法是基于现有的情感词典和CASL中词汇的语义关系。抽取算法获得的CASL包括大量的情感词汇、词汇的同义词和反义词关系。基于情感语义关系,近义词的褒贬性一致,反义词的褒贬性相反。基于同义词和反义词关系,利用现有的情感词典(前面4个词典)标注及扩展标注抽取词汇的情感倾向。基于情感语义关系的CASL词汇的情感倾向标注结果如表4所示。CASL的情感倾向标注方法为:首先标注CASL覆盖现有词典的情感词汇,结果为表4中“词典标注词汇数量”。然后,基于CASL中情感词汇的同义词集合,对于CASL中的未标注的词汇 W_i 进行标注。扩展标注方法是从前向后循环检索 W_i 的同义词 S_i ,如果发现有情感标注的词汇,则设置 W_i 的情感倾向和 S_i 一致;然后,基于CASL中情感词汇的反义词集合,对于CASL中的未标注的词汇 W_i ,从前向后循环检索 W_i 的反义词 A_i ,如果发现有情感标注的词汇,则设置 W_i 的情感倾向和 A_i 一致;在倾向扩展标注中,如果在 S_i 和 A_i 都没有找到,就不标注 W_i 的情感倾向。

基于4个基本词典,CASL的情感倾向扩展标注实验结果如表4所示,结果说明基于情感语义关系的情感词汇的倾向标注方法有效。表中第3列为基于语义关系标注词汇数量和标注正确的数量;表中第4列说明方法对CASL扩展标注了178.7%、167.3%、261.2%和59.1%的词汇情感倾向。情感标注的准确率分别是88.1%、86.9%、86.2%和79.1%。基于4个现有词典,实现CASL词汇平均扩展166.6%情感倾向标注,78.1%准确标注。

表4 基于情感语义关系的CASL词汇情感倾向标注结果

基础词典	基于词典标注 词汇数量/个	基于语义关系标注词汇 数量/标注正确的数量/个	准确率 /%	扩展增 加率/%
HowNet	3 962	12 540/11 044	88.1	178.7
清华大学褒贬义词典	4 711	14 494/12 596	86.9	167.3
台湾大学 NTUSD(简体版)	2 771	11 608/10 011	86.2	261.2
大连理工大学情感本体	10 829	21 776/17 232	79.1	159.1

4 结束语

基于情感的复杂性,提出了一个情感词汇的假设,实验结果证明这个假设是可靠的。利用中文网络知识百科,提出了一种简单、高效的方法抽取中文情感词汇,并成功地抽取了词汇的两个重要语义关系即同义词关系和反义词关系。通过和现有情感词典的比较,该方法抽取结果基本覆盖现有情感词典的词汇数量较高。同时,基于现有情感词典和抽取的情感词汇间语义关系,实现了快速地扩展情感词语的倾向标注。

该方法的局限性是过于依赖网络词汇的准确度。虽然可以通过规则去过滤,但规则过滤会减少情感词汇的抽取。抽取数据源的语义错误、错别字等问题影响了数据的抽取结果。情感语义关系词汇CASL的构建,更加方便情感语义词典的构建和文本的情感分析。研究中已经基于情感语义关系,进行了情感词汇倾向自动扩展标注。今后的研究,还可以进行情感词汇情感权值自动扩展标注和计算、情感词语语义相关度计算等等。此外,基于规则抽取算法过滤出来的词汇有很大一部分是错误拼写,这一部分词汇将被考虑生成

一个中文的错误拼写词典。

参考文献:

- [1] BAR-HILLEL Y, CARNAP R. Semantic information[J]. British Journal for the Philosophy of Science, 1953, 4(14): 147-157.
- [2] SOWA J F. Semantic networks[M]//Encyclopedia of cognitive science. [s. l.]: John Wiley & Sons, Ltd, 2006: 1-50.
- [3] 崔艳菊, 严灿灿, 刘慧敏. 从语义关系的复杂性看语义词典建设[J]. 解放军外国语学院学报, 2011, 34(4): 13-17.
- [4] FELLBAUM C, MILLER G. WordNet: an electronic lexical database[M]. Cambridge, MA: MIT Press, 1998.
- [5] RICHARDSON S D, DOLAN W B, VANDERWENDE L. MindNet: acquiring and structuring semantic information from text[C]//Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics. Montreal, Quebec, Canada: ACL, 1998: 1098-1102.
- [6] ESULI A, SEBASTIANI F. Sentiwordnet: a publicly available lexical resource for opinion mining[C]//Proceedings of LREC. Genoa, Italy: LREC, 2006: 417-422.
- [7] 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.
- [8] 梅家驹. 同义词词林[M]. 上海: 上海辞书出版社, 1983.
- [9] 于江生, 俞士汶. 中文概念词典的结构[J]. 中文信息学报, 2002, 16(4): 12-20.
- [10] 王惠, 詹卫东, 俞士汶. “现代汉语语义词典”的结构及应用[J]. 语言文字应用, 2006(2): 134-141.
- [11] 黄曾阳. HNC 理论概要[J]. 中文信息学报, 1997, 11(4): 11-20.
- [12] PICARD R W. Affective computing[M]. Cambridge, MA: MIT Press, 1997.
- [13] WIEBE J M. Learning subjective adjectives from corpora[C]//Proceedings of the national conference on artificial intelligence. Austin: AAAI Press, 2000: 735-740.
- [14] RILOFF E, WIEBE J M, WILSON T. Learning subjective nouns using extraction pattern bootstrapping[C]//Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003. Edmonton, Canada: ACL, 2003: 25-32.
- [15] KAJI N, KITSUREGAWA M. Building lexicon for sentiment analysis from massive collection of HTML documents[C]//Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning. Prague: ACL, 2007: 1075-1083.
- [16] MOILANEN K, PULMAN S. The good, the bad, and the unknown: morph syllabic sentiment tagging of unseen words[C]//Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: short papers. Stroudsburg: ACL, 2008: 109-112.
- [17] HATZIVASSILGLOU V, MCKEOWN K R. Predicting the semantic orientation of adjectives[C]//Proceedings of the 35th annual meeting of the association for computational linguistics. Madrid, ES: ACL, 1997: 174-181.
- [18] TURNEY P D, LITTMAN M L. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.
- [19] LI Sujian, ZHANG Jian, HUANG Xiong, et al. Semantic computation in a Chinese question-answering system[J]. Journal of Computer Science and Technology, 2002, 17(6): 933-939.
- [20] 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.
- [21] LI J, SUN M. Experimental study on sentiment classification of Chinese review using machine learning techniques[C]//Natural language processing and knowledge engineering. Beijing: IEEE, 2007: 393-400.
- [22] 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.
- [23] 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): 2875-2877.
- [24] 张成功, 刘培玉, 朱振方, 等. 一种基于极性词典的情感分析方法[J]. 山东大学学报: 理学版, 2012, 47(3): 47-50.
- [25] 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.
- [26] YANG A M, LIN J H, ZHOU 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.
- [27] 陈建美, 林鸿飞, 杨志豪. 基于语法的情感词汇自动获取[J]. 智能系统学报, 2009, 4(2): 100-106.
- [28] 金宇, 朱洪波, 王亚强, 等. 基于直推式学习的中文情感词极性判别[J]. 计算机工程与应用, 2011, 47(34): 164-167.
- [29] 曹树金, 张学莲, 陈忆金. 网络舆情意见挖掘中极性词典构建和极性识别方法研究[J]. 图书情报杂志, 2012(1): 60-65.
- [30] 王铁套, 王国营, 陈越, 等. 基于语义模式与词汇情感倾向的舆情态势研究[J]. 计算机工程与设计, 2012, 33(1): 74-77.
- [31] WANG Yang, WANG Haofen, ZHU Haiping, et al. Exploit semantic information for category annotation recommendation in Wikipedia[C]//Natural language processing and information systems. Berlin: Springer, 2007: 48-60.
- [32] WANG R C, COHEN W W. Automatic set instance extraction using the web[C]//Proceedings of ACL-IJCNLP 2009. Suntec, Singapore: ACL, 2009: 441-449.
- [33] 阳爱民, 林江豪, 周咏梅. 中文文本情感词典构建方法[J]. 计算机科学与探索, 2013, 7(11): 1033-1039.