设计研究与应用

基于 NLP 的短文本观点抽取和极性词分析

晏丞骁
（毕索大学，加拿大希尔布鲁克 J1M 1Z7）

摘 要：随着现代信息技术的发展，各种网络平台快速普及，大众已经习惯于通过如微博、微信等网络媒体表达他们的观点和意见，且用户发表的多为短文本，其中包含大量有价值的信息。因此，网络短文本成为自然语言处理（Natural Language Processing，NLP）领域的研究热门。本文以短文本为研究对象，对短文本进行语义极性分析，利用计算机自动分析包含观点信息的句子，抽取主题词、特征词，利用主谓极性传递算法提取句子中的观点。

关键词：短文本；观点抽取；极性词；自然语言处理
中图分类号：TP391 **文献标识码：**A **DOI：**10.3969/j.issn.1003-6970.2022.03.038
本文著录格式：晏丞骁.基于NLP的短文本观点抽取和极性词分析[J].软件,2022,43(03):121-123

NLP-based Short Text Opinion Extraction and Polarity Word Analysis

YAN Chengxiao
（Bishop's University, Sherbrooke Canada J1M 1Z7）

【Abstract】：With the development of modern information technology, various network platforms are rapidly popular. People has become accustomed to expressing their views and opinions in short text by network media such as Weibo, WeChat, which contains a lot of valuable information. Therefore, short text is a popular research field in Natural Language Processing. In this paper, the short text is treat as a research object, and the phrase of the short text is used to automatically analyze the sentence containing the synthesis of the viewpoint information, the topic words, feature words, and use the primary polarity transmission algorithm to extract the opinion in the sentence.

【Key words】：short text;opinion extraction;polar word;natural language processing

0 引言

随着互联网的不断进步，个人观点通过网络迅速传播，短文本主要通过网络论坛、微博、微信等社交平台传播，很多短文本具有重要的商业和社会研究价值，因此，对这些信息进行分析具有非常重要的意义。短文本观点抽取和极性词分析是现阶段自然语言处理领域的研究热点。本次研究主要分为两部分：一是，短文本中词的极性分析；二是，短文本中观点的提炼。极性词分析综合运用人工和自动方法，即先人工标示极性比较强的词，建立极性词集合，然后运用相关算法自动分析新词、原词之间的极性传递关系，根据极性传递关系给新词打分，判断词的极性[1]。相关语义极性分析研究很多，比如检测单个词的语义极性，分离主客观句子，计算整篇文章或段落的极性。也出现了一些成熟的算法。然而，观点抽取研究才刚刚起步，目前没有成熟稳定的算法。

1 极性词分析

极性词（Polar Word）是指句子中具有明显情感倾向的词语[2]，例如，“新款卡罗拉 2020 车型是卡罗拉车系优秀的设计水准的又一次提高”。其中，“优秀”“提高”就是带有情感倾向的极性词。基于极性词可以评价句子乃至文本的极性。Hong Yu 等人选择多个极性明显的形容词创建集合。通过计算集合中部分词的共生度，据此确定新词的极性。如果集合中已含有“优秀”，且极性为“+”，则由于“提高”“优秀”同时出现，故将“提高”的极性也定为“+”。但是，此类算法有两个缺陷：一是，计算复杂度高，每个新词极性的计算需要时间，资源消耗大；二是，忽视了集合中的词语的强度如何，不能为计算新词的强度提供有效依据，为此，本文提出基于中国知网情感词典的解决方案。

作者简介：晏丞骁（1995—），男，四川成都人，硕士研究生，研究方向：自然语言处理。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.1 词典构建

CNKI 情感词典共收录 6564 个词条，单个词条包含极性、强度两大属性。其中极性属性分为三种：正极性、负极性和中性 [3]。强度属性分为 5 级：1.0、0.5、0、-0.5、-1.0。没有进一步细分强度等级的原因是手动极性标记考验个人经验，有时候人很难区分词的极性强弱，多人独立对词进行极性、强度判断，取平均值，确保 CNKI 情感词典中包含大部分常用极性词，降低时间复杂度，如（优秀，+，1.0）、（提高，+，0.5）、（执行，0，0.0）、（差，-，1.0）、（弱，-，-0.5）。细分 2565 条关于汽车的在线评论短文本中的词语，人工筛选极性词，共 1516 个，四人独立筛选，取相同的词，忽略不同的词，得到 642 个正面词、405 个负面词，这些极性词的强度与 CNKI 词典中类似，均标注极性、强度，得到极性词典。

1.2 基于混合方法词语极性判断

无论使用哪种方法都存在不足。首先，只有约 25% 的正确句子作为相关词出现，因此第一种方法，召回率较低。其次，约 5.5% 的句子中词之间有转折词存在，说明即便极性词在同一个句子中出现，其极性也会不同。因此，本文结合两种算法得到混合算法 [4]。算法如下：

（1）对于每个潜在的极性词，首先搜索 CNKI 情感词典和极性词典，确定词的极性、强度。

（2）对于查找不到的潜在极性词，可以扩大搜索范围，找到前后极性词中间的相关词。

（3）如果一个潜在极性词与前后已知的极性词之间有关联词，可以先将关联词的类型确定下来，然后依据以下规则将这个词的极性、强度计算出来：

1）如果是并列词，则已知极性词与潜在极性词的极性、强度相同。

2）如果是递进词，则已知极性词与潜在极性词的极性相同，强度不同，后者是前者的两倍。

3）如果是转折词，则潜在极性词的极性与已知极性词的极性相反，强度与已知极性词的强度相同。

2 基于极性词的短文本观点提取

2.1 指代消解

观点是由（Topic, Holder, Claim, Sentiment）构成的四元组合，其中，观点拥有者确信存在有关特定主题的声明，通常这种确信包括观点拥有者的主观情感因素。例如，“我的卡罗拉已经购买了 3 个月，我认为它在油耗方面表现很好。”句子中有两个声明：一是，我的卡罗拉购买了 3 个月；二是，我认为卡罗拉油耗表现很好。两个声明的主题词都是“卡罗拉”，但前一个声明只是客观描述，不存在极性词，后一个声明中“很好”是极性词，属于主观陈述。观点抽取的关键是基于 NLP 明确四元组：即主题、特征词抽取、观点分析、极性分析、观点总结。

当分析一个句子时，需要确保句子的结构完整，词没有歧义。首先要解决的是指代消解问题。指代消解主要分为两部分。首先，对指代消解的必要性进行判断，其次，结合语境等约束条件建立先行词集合，最终确定先行词，具体过滤规则如下 [5]：

（1）查看分词列表时，如果不属于实体类、功能类或已消解的词，则剔除。

（2）如果词与所指对象的单复数信息有冲突发生，则过滤该词。

（3）如果词与宾语的语法作用不相容，则将该词过滤掉。比如，句子动词对一个子句进行引导成为宾语，如果人称代词是句子的主语，那么被引导的句子中的名词不能作为代词的先行词。

（4）处理句子中的先行词时，筛除“其他”“长期以来”等干扰词；

（5）划分句子的层次，找出先行词；

（6）对各个潜在的先行词进行权重计算，权重值 = 角色权值 + 单复数权值 – 距离权值。

（7）明确句子中最有可能的先行词，进行下一步处理。

2.2 特征词和主题词抽取

在实践中，用户通常会关注事物的某些特征。如果有多个特征词出现在一个句子中，那么必须将这些特征词的从属关系明确下来 [6]。本体论能够将这一问题解决掉，可以在层次分析体系的定义中应用本体论，用树状结构将该体系表示出来，通过遍历树，将父子关系充分利用起来，能够将特征词之间的从属关系寻找出来。本文用树状层次结构将汽车行业的主体表示出来，为了为用户短文本中系统表示最终图像提供便利，组件和技术统称为特征，汽车本体结构如图 1 所示。

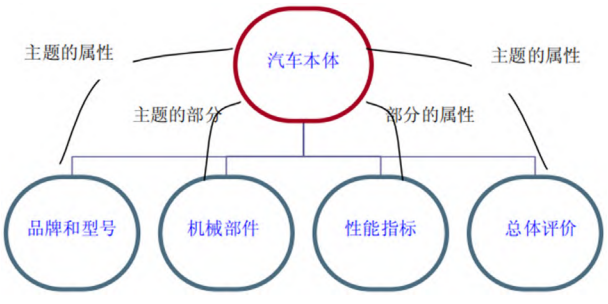

图 1 汽车本体结构及特征

Fig.1 Structure and characteristics of automobile body

2.3 句子的观点抽取

如果句子包含观点，那么只知道观点的极性并不够。用户想知道意见领袖是谁，意见讨论的话题 / 特点是什么，而传统的统计方法不能可靠地解决这个问题，需要

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

从语法上解析，可以看出句子存在主谓结构，可以提供主谓变化关系等信息。在主谓结构中，主语要么是主题或特征，要么是意见领袖。而谓语可能形容词或动词。

句子主谓结构中的极性传递：

（1）将所有含有主谓结构的关系对寻找出来，执行（2）。

（2）如果谓语的上下文极性不等于 0 如果谓语是形容词，则主语的上下文极性等于谓语的上下文极性。结束。否则表示谓语为动词，则执行（3）。

（3）如果谓语的上下文极性不等于 0，则主语的上下文极性等于谓语的上下文极性 .，否则表示谓语动词无极性，执行（4）。

（4）将含有该动词的动宾结构关系对寻找出来，如果宾语为形容词，且主语的上下文极性等于形容词 1 的上下文极性。如果宾语是名词，则找到含有名词的关系对，其中，形容词标记为形容词 2，主语的上下文极性等于形容词 2 的上下文极性。

（5）将含有谓语的副词关系对寻找出来，标记其中的形容词为形容词 3，将谓语动词赋予其上下文极性。

（6）主语的上下文极性等于谓语的上下文极性。

（7）算法结束。

虽然对上述算法进行了应用，可以将句子的主语从主语与谓语关系中寻找出来，并向主语上转移主语的极性，但是讨论的主题并不总是主语。因此，需要改进算法。如果谓语的上下文极性不等于 0，则主语的上下文极性等于谓语的上下文极性。否则，表示谓语动词没有极性，继续（4）。如果谓词的上下文极性不等于 0，在主语是特征词的情况下，主语的上下文极性等于谓语的上下文极性，否则，对含有谓语动词的动宾关系进行继续查找，如果特征词为该关系对含有的名词，则名词的上下文极性等于谓语的上下文极性 . 否则表示谓语动词没有极性，执行（4）。除此意外，由于长句中包括子句，网络短文本存在很多断句不规范的问题，常常句子太长，主谓分析的方法覆盖范围有限。

例：卡罗拉就感觉有点老气，大片暗色的内部装饰大气得很但又感觉刻板。依存关系对如下：

{(01)[4] 有点 (副词) 〜 [5] 老气 }
{(02)[2] 就 (副词) 〜 [3] 感觉 }
{(03)[8] 暗色 (介词) 〜 [9] 的 }
{(04)[9] 的 (助词) 〜 [10] 内部装饰 }
{(05)[14] 又 (副词) 〜 [15] 感觉 }
{(06)[13] 但 (副词) 〜 [15] 感觉 }
{(07)[1] 卡罗拉 (主谓) 〜 [3] 感觉 }
{(08)[16] 刻板 (动宾) 〜 [5] 感觉 }
{(09)[15] 感觉 (动宾) 〜 [12] 得很 }
{(10)[10] 内部装饰 (主谓) 〜 [11] 大气 }
{(11)[5] 老气 (动宾) 〜 [3] 感觉 }
{(12)[7] 大片 (主谓) 〜 [11] 大气 }
{(13)[11] 大气 (副词) 〜 [12] 得很 }
{(14)[12] 得很 (副词) 〜 [3] 感觉 }
{(15)[3] 感觉 (核心词) 〜 [18]<EOS>}

通过句子主谓结构分析得到 (07)、(10)、(11) 三个主谓关系对，其中“大片”不属于特征词，同时找不到含有“大气”的动宾关系对，因此“大气”的极性不会传递。此时，根据汽车本体结构判断句子中的主语是否为主题词或特征词。如果是主题词或特征词，则做好标记，用到的极性词也打上标签 (Marked)，用主谓结构分析法分析后，查找没有打标签的极性词，并剔除，最后得到结果如表 1 所示。

表 1 观点抽取结果

Tab.1 Opinion extraction results

主谓结构分析	主语	观点	极性
(04) [3] 喜欢 ~[1] 我 (SBV)	[1] 卡罗拉	< 不 >+“喜欢”	+1.0×（-1）=-1.0
(07) [3] 感觉 ~[1] 卡罗拉 (SBV)	[1] 卡罗拉	< 比较 >+“老气”	0.5×2=-1.0
(10) [1] 大气 ~[10] 内部装饰 (SBV)	[10] 内部装饰	< ? >+“大气” < ? >+“刻板”	(+1.0×1) + (-0.5×1) =+0.5

3 结语

综上所述，本文提出句子中的词的基于依存关系来确定极性词的上下文极性，基于极性词和句子主谓结构提取句子中词的关系对，以提取主题词和特征词。然后与手动标记的结果进行比较。但由于网络短文本断句不规范，本文的工作只是分析，这方面的研究不成熟，参考方法不多。本文中介绍的方法只是一种经验建模方法，句子不规范的情况下，算法的有效性会大大降低。因此，对句子观点的抽取和极性词分析算法有待进一步改进。

参考文献

[1] 宋静静.中文短文本情感倾向性分析研究[D].重庆:重庆理工大学,2013.

[2] 钟丁媛,高峥洲,金皓辰,等.基于NLP文字处理的评论有用性探究[J].科技风,2020(31):150-153.

[3] 翟晓晓.基于NLP的产品中文评论特征词识别与语义倾向分析[D].天津:南开大学,2012.

[4] 杨传龙,王金龙.基于NLP的企业供应关系自动抽取研究[J].计算机科学与应用,2018,8(12):1823-1832.

[5] 化柏林.基于NLP的知识抽取系统架构研究[J].现代图书情报技术,2007(10):38-41.

[6] 陈永俊,夏艳锋,高宇航,等.基于NLP技术的警情文本数据分析应用[J].警察技术,2021(2):39-42.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net