文章编号: 1000-5641(2014)04-0062-07

基于词典与语料结合的中文微博主观句抽取方法

朱海欢, 余青松

(华东师范大学 计算中心, 上海 200062)

摘要: 提出一种基于词典与语料结合的中文微博主观句抽取方法, 通过判断句子中是否包含情感表达文本来判断句子是否为主观句. 首先, 从现有的情感词典中挑选出情感倾向较为固定的情感词构建了一个高可信情感词典, 用于抽取句子中的情感表达文本, 保证情感表达文本抽取的准确率; 然后提出 N-POSW 模型, 并基于 2-POS W模型通过语料学习的方法较为准确地抽取句子中的剩余情感表达文本, 保证了情感表达文本抽取的召回率. 实验结果表明, 相比于传统的基于大规模情感词典的方法, 本文方法主观句抽取的 F 值提高了 7%.

关键词: 情感词典; 高可信情感词典; N-POSW 模型; 主观句

中图分类号: TP39 **文献标识码**: A **DOI:** 10.3969/j.issn.1000-5641.2014.04.008

Study on the extraction of Chinese microblog subjective sentences based on lexicon and corpus

ZHU Hai-huan, YU Qing-song

(*Computer Center, East China Normal University, Shanghai* 200062, *China*)

Abstract: In this paper, we propose a new method for the extraction of Chinese microblog subjective sentence, which is based on a combination of lexicon and corpus. By determining whether the sentence contains emotional expressions, it can be classified as a subjective or objective sentence. Firstly, a highly credible sentiment lexicon was built based on the words whose emotional orientation is fixed from the existing sentiment dictionary. Based on the highly credible sentiment lexicon, sentiment expressions can be extracted with assurance of accuracy. Finally, a N-POSW model was proposed for the corpus-based learning method. Through the 2-POSW model, the remained sentiment expressions in the sentence can be extracted, thus guaranteeing the overall recall rate. Experimental results show that the F Value in this paper increases 7% compared with the traditional method, which is based on the large-scale sentiment lexicon.

Key words: sentiment lexicon; highly credible lexicon; N-POSW model; subjective sentence

收稿日期: 2013-07

第一作者: 朱海欢, 男, 硕士研究生, 研究方向为自然语言处理. E-mail: zhhh1988@gmail.com.

通信作者: 余青松, 男, 高级工程师, 硕士生导师, 研究方向为Web应用技术.

E-mail: qsyu@cc.ecnu.edu.cn.

0 引　　言

文本情感分析是自然语言处理领域一个重要的研究方向, 广泛应用于商品推荐、商品调研、舆情分析、事件预测、有害信息过滤等领域, 具有巨大的社会经济价值. 文本主客观分类是文本情感分析的重要组成部分, 也是文本情感分析首要解决的问题.

目前, 英文语句的主客观分类研究较为成熟. Kim[1]通过抽取句子中的情感信息来完成句子的主客观分类. 然而, 包含情感信息的句子不一定都是主观句. Wiebe[2−5]对句子中主观表达式抽取做了深入的研究, 通过抽取句子中主观表达式来提高主客观分类的效率. Pang[6]基于图理论完成了句子的主客观分类. Long[7]分析了抽取内容特征、情感词特征、面向主题特征对于 Tweets 主客观分类效果的影响, 发现抽取面向主题特征时主客观分类效果最好.

中文方面, 叶强[8]根据句子中的连续双词词类组合模型(2-POS)对互联网评论进行主客观分类. 该方法根据CHI统计方法为每个2-POS模型赋予权值, 并根据句子中的2-POS模型计算句子的主观性总分, 若主观性总分大于阈值表示句子为主观句. 张博[9]结合了句法结构模板、依存关系模板、SVM 分类器模板对中文观点句进行抽取. 杨武[10]以特征词和主客观线索做语义特征, 以 2-POS 模型做语法特征, 采用朴素贝叶斯分类器对中文微博句子进行主客观分类, 发现同时考虑语义特征和语法特征的分类效果比只考虑一类特征的效果要好.

微博是一种短文本. 短文本具有长度短、内容集中等特点. 这种短文本特性决定了微博中的句子长度不会太长, 表达的主客观倾向较直接. 因此, 本文提出的中文微博主观句抽取方法通过分析句子中文本的主客观倾向, 抽取句子中的主观表达文本, 来判断句子的主客观倾向符合微博的短文本特征.

本文主要研究中文微博的主观句抽取方法, 首先根据一个高可信的情感词典确保主观句抽取的准确率, 而后通过语料学习提高主观句抽取的召回率. 相比较于已有的方法, 本文所提方法的创新性在于: 对已有的 N-POS 模型进行改进, 提出了中文微博句子的 N-POSW 模型, 并基于 2-POSW 模型, 通过半监督学习方法抽取句子中的主观表达文本.

本文结构组织如下: 第 1 节介绍微博主观句抽取的思路; 第 2 节介绍传统的基于大规模情感词典的中文微博主观句抽取方法, 分析了该方法存在的不足之处; 第 3 节阐述了本文提出的基于词典和语料结合的中文微博主观句抽取方法; 第 4 节是实验与分析; 第 5 节是总结.

1 微博主观句抽取思路

判断一个中文微博句子是否为主观句的关键在于抽取句子中的情感表达文本. 本文所提到的情感表达文本是指句子中能够表达作者情感倾向的文本. 句子中, 情感表达文本可以是一个情感词语、一个情感短语、一个标点符号, 也可以是其他字符. 我们判断一个句子是否为主观句的思路为: 若句子中包含这样的情感表达文本, 则该句子为主观句. 因此, 准确、全面地抽取句子中的情感表达文本是微博主观句抽取要解决的核心问题.

微博句子样例 1: #ipad#我很喜欢.

微博句子样例 2: #ipad#买的很值！！！

微博句子样例 3: #ipad#终于上市了,/微笑.

在中文微博句子中, 情感表达文本主要包括句子中的词序列、标点、表情. 例如微博句子样例 1, 句子中的词序列“喜欢”是能够表达情感倾向的文本. 例如微博句子样例 2, 句子中标点符号“！！！”也能够体现强烈的情感. 例如微博句子样例 3, 句子中表情“/微笑”也能体

现句子的情感倾向. 然而, 中文微博句子的表达形式比较不规范. 微博句子中除了出现词序列、标点、表情外, 还会出现其他对微博句子主观句抽取无关的文本. 这些无关文本主要包括字母、数字、标签、链接等. 这些无关文本的介入不仅增加了微博主观句抽取的复杂性, 而且还会对微博主观句抽取带来一定的噪音. 针对这种现象, 在抽取微博主观句前, 需要对微博句子进行预处理, 预处理部分包括去除微博句子中的所有字母、数字、标签、链接. 图 1 描述了中文微博主观句抽取的主要思路.

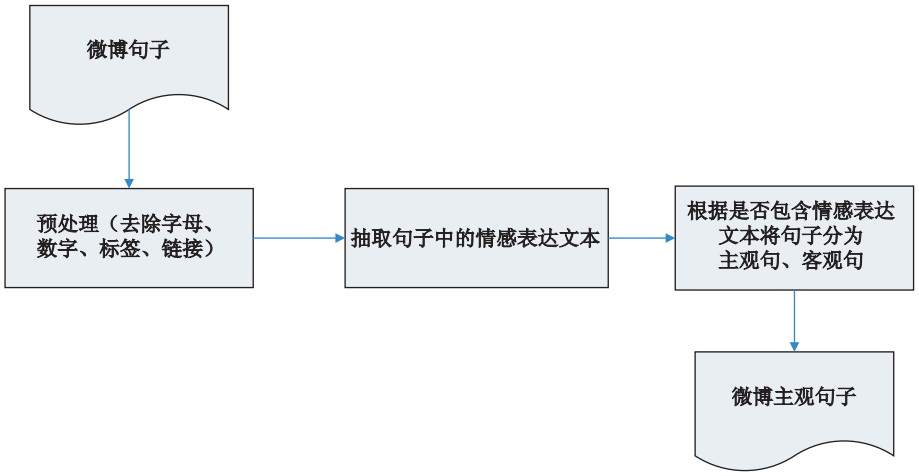

图 1 中文微博主观句抽取的主要思路

Fig. 1 Main thoughts of extracting Chinese microblog subjective sentences

2 传统的基于大规模情感词典的微博主观句抽取方法

传统的基于大规模情感词典的微博主观句抽取方法, 通过判断句子中是否出现情感词典里的情感词, 来判断该情感词是否为句子的情感表达文本. 因此, 需要构建一个庞大的情感词典. 将 HowNet 情感词典[11]、NTUSD情感词典[12]进行融合和修正, 构建了一个包含 11 212 个情感词的情感词典, 记为 SentiDic. 具体描述如表 1 所示.

表 1 基于情感词典的微博主观句抽取算法

Tab. 1 A algorithm for extraction of subjective sentences based lexicon

Input:	中文微博句子 Sentence
Output:	Sentence 是主观句或 Sentence 是客观句
1:	if Sentence 包含 SentiDic 里的某个词 then
2:	Sentence 是主观句
3:	else Sentence 是客观句
4:	end if

传统的基于大规模情感词典的主观句抽取方法很大程度上依赖于所构建情感词典的准确性和全面性. 如果情感词典规模太小, 则主观句抽取的召回率会较低, 如果情感词典中的某些情感词在句子中并没有表达情感, 那么也会影响主观句抽取的准确率. 在中文句子中, 普遍存在一种现象: 一个词语在某个句子中是情感词, 在另一个句子中却不是情感词, 尤其是一些动词和名词. 比如“喜欢”这个词语, 在微博句子: “我超喜欢#ipad#.” 中, 词语“喜欢”是句子的情感词; 而在微博句子: “#官二代求爱不成将少女毁容#周岩, 他喜欢你吗？”中, 词语“喜欢”并不是情

感词.

3　基于词典与语料结合的微博主观句抽取方法

3.1　建立高可信情感词典

虽然某些词语在不同句子中会有不同的情感倾向(正如第 2 节中的分析), 但我们也不能否认有很多词语的情感倾向在多数情况下是固定的, 比如“该死”、“狠毒”、“善良”、“聪明”等. 针对这种现象, 我们一方面肯定情感词典对中文微博主观句抽取的作用, 另一方面也要认识到其中的不足. 因此, 我们从情感词典SentiDic中人工挑选出情感倾向较为固定的情感词构成一个高可信情感词典 R_SentiDic, 用于微博句子的情感表达文本抽取. R_SentiDic 包含情感词 2 143 个. 表 2 列举了 10 个 R_SentiDic 中的情感词.

表 2　R_SentiDic中10个情感词

Tab. 2　Ten words in the R_SentiDic

情感词				
不和谐	妄自尊大	凶狠	愚蠢	歹毒
不懈怠	博学多才	沉稳	诚实	刚正

3.2　N-POSW模型

在阐述 N-POSW 模型之前, 首先介绍 N-POS 模型[8]. N-POS 模型是将句子中的词按照词性进行分类, 用句子中连续的 N 个词的顺序组合作为一个项, 对句子进行表示. 当 N=2 时, 此时的模型为 2-POS 模型. 例如:

语句: 我心情很好.

对其分词及词性标注后的结果为: “我(代词)心情(名词)很(副词)好(形容词). (标点符号)”. 对于这个语句的 2-POS 模型是“代词-名词, 名词-副词, 副词-形容词”, 其中, “代词-名词”是一个 2-POS 项.

文献 [8] 提出了 N-POS 模型. 该模型主要考虑到词性对词语主观倾向的影响. 然而, 仅考虑词性对词语主观倾向的影响具有一定的误差. 本文基于 N-POS 模型, 在考虑词性对词语主观倾向影响的同时也考虑词语本身语义对词语主观倾向的影响, 提出了句子的 N-POSW 模型.

定义 1　S为一个语句.

定义 2　w_i 为 S 中序号为 i 的词. S=$w_1w_2w_{3...}w_{i...}w_n$, n 为 S 词序列的长度.

定义 3　SqPOS=(c_1, c_2,...,c_n) 为 S 的词性序列, c_i 为 w_i 的词性.

定义 4　N-Words=($w_1w_2...w_N$, $w_2w_3...w_{N+1}$, ..., $w_{n-N-1}w_{n-N+1}...w_n$), N-Words 为 S 中连续 N 个词序列组合成的特征项序列.

定义 5　N-POSW=(SqPOS,N-Words)为 S 的 N-POSW 模型. 该模型由两部分组成, 一个是 S 中各词序列对应的词性序列, 一个是 S 中 N 个连续词序组合成的特征项序列.

例如, 对于语句“我心情很好.”. 对其分词及词性标注后的结果为: “我(代词)心情(名词)很(副词)好(形容词). (标点符号)”. 对于这个语句, SqPOS=(代词, 名词, 副词, 形容词), 2-Words=(我心情, 心情很, 很好).

3.3　基于 2-POSW 模型的主观句抽取

上节提出了句子的 N-POSW 模型, 当 N 取 2 时, 该模型即为 2-POSW 模型. 本节主要阐述基于 2-POSW 模型, 通过语料学习的方式抽取句子中的情感表达文本, 从而判断该句子是否为主观句.

定义 6 $\mathrm{T(text)}=\begin{cases}1 & \mathrm{subj(text)}>\alpha\times\mathrm{obj(text)},\ \mathrm{subj(text)}>\beta,\\ 0 & \text{其他.}\end{cases}$ 其中, subj(text)为文本 text 在训练语料中主观句中出现的次数, obj(text) 为文本 text 在训练语料中客观句中出现的次数, α, β 为两个整数阈值.

定义 7 $Q(w_iw_{i+1})=\begin{cases}1 & \exists w\in\{w_i,w_{i+1}\},c\in(a,v,n),\\ 0 & \text{其他.}\end{cases}$ 其中, w_iw_{i+1} 为语句 S 的两个连续词序列, c 为词 w 对应的词性, a, v, n 分别表示形容词, 动词、名词.

定义 8 $P(S)=\sum_{i=1}^{n-1}Q(w_iw_{i+1})\times T(w_iw_{i+1})$ 为语句 S 的主观值. 其中, w_i 为语句 S 序号为 i 的词, n 为 S 词序列长度. $P(S)$ 的值表示语句 S 中情感表达文本的数量. S 中的情感表达文本为两个连续词组成的词序列. 若该词序列在训练语料主观句中出现的次数大于在训练语料客观句中出现次数的 α 倍, 且在训练语料主观句中出现的次数大于 β, 同时该词序列中包含一个形容词或者动词或者名词, 则该词序列为 S 的情感表达文本.

表 3 描述了基于 2-POSW 模型的主观句抽取. 分词及词性标注工具采用的是中科院提供的分词系统 ICTCLAS[13].

表 3 基于 2-POSW 模型的主观句抽取算法

Tab. 3 A algovithm for extraction of subjective sentence base on 2-POSW

Input:	中文微博句子 S
Output:	S 是主观句或 S 是客观句
1:	对 S 分词及词性标注
2:	构建 S 的 2-POSW 模型
3:	计算 $P(S)$
4:	if $P(S)>0$ then
5:	S 是主观句
6:	else S 是客观句
7:	end if

3.4 词典和语料结合抽取微博主观句

基于词典和语料结合的主观句抽取算法综合了基于词典的的主观句抽取算法和基于 2-POSW 模型的主观句抽取算法. 这里采用的词典为高可信的情感词典 R_SentiDic. 首先, 判断微博句子中是否包含 R_SentiDic 中的情感词; 若包含, 则可判断该句子为主观句. 由于 R_SentiDic 中的情感词是可信的, 即该情感词在大多数情况下能表达主观的情感倾向. 因此, 此时抽取的主观句中能保证较高的准确率. 但由于 R_SentiDic 中的情感词数量有限, 因此不能保证主观句抽取的召回率. 为了克服这个缺点, 对剩余的句子采用基于 2-POSW 模型的主观句抽取算法, 抽取剩余的主观句, 从而保证主观句抽取的召回率. 具体描述如表 4 所示.

基于 2-POSW 模型的主观句抽取算法需要一个主客观标注的句子语料作为训练语料. 虽然网络中, 未标注的句子语料较易获取且比较丰富, 但主客观标注的句子语料比较缺乏. 如果通过对未标注语料进行人工标注从而获得主客观标注语料, 需要通入大量的精力. 因此本文根据第 2 章中提出的传统的基于大规模情感词典的主观句抽取算法对大量未标注的句子进行主客观标注, 从而得到所需的训练语料, 这里采用的词典为 SentiDic.

表 4　基于词典和语料结合的微博主观句抽取算法

Tab. 4　A algovithm for extraction of Chinese microblog subjective senterces based on lexicon and corpus

```
Input: 中文微博句子 S
Output: S是主观句或 S 是客观句
if 用基于词典的主观句抽取算法判断 S 为主观句 then
  S 为主观句
else
  if 用基于 2-POSW 模型的主观句抽取算法判断 S 为主观句 then
    S 是主观句
  else S 是客观句
  end if
end if
```

4　实验与分析

4.1　评测语料与训练语料

NLP&CC中文微博情感分析评测[14]提供了“官二代求爱不成将少女毁容”话题的微博评论数据. 该评测已对每条微博评论进行句子级的切分, 共 7 040 个句子. 评测数据记录在 xml 文件, 每条微博的记录形式如下列微博评测样例所示. NLP&CC 中文微博情感分析评测规定: 主观句只限定在对其它对象的评价, 不包括内心自我情感表达. 根据该评测对主观句的鉴定标准, 从“官二代求爱不成将少女毁容”话题微博评论数据中抽取 1 000 个主观句和 1 000 个客观句构成评测数据, 根据本文第 3.4 节中介绍的方法对剩下句子进行机器标注构成训练语料.

微博评测样例:

<weibo id=″89″ >

<sentence id=″1″ >#官二代求爱不成将少女毁容#看了陶汝坤一些资料, 就是感到他太横了, 缺少教养是毋庸置疑的, 问题是谁给他的这么大的底气？</sentence>

<sentence id=″2″ >养不教父之过, 教不严师之惰, 学校的老师呢？</sentence>

<sentence id=″3″ >家里的父母呢？</sentence>

<sentence id=″4″ >怎么教出来这样一个横行无忌, 出口就带脏话的孩子, 常言道, 三岁知老, 今后谁家的女儿嫁给他, 真是瞎了眼了. </sentence>

<hashtag id=″1″ >官二代求爱不成将少女毁容</hashtag>

</weibo>

4.2　标价标准

根据准确率、召回率、F 值来评价算法在主观句抽取中的效果. 计算公式分别是

$$\text{准确率} = \frac{\text{算法识别出的正确主观句数量}}{\text{算法识别出的主观句总数}},$$

$$\text{召回率} = \frac{\text{算法识别出的正确主观句数量}}{\text{评测语料中主观句总数}},$$

$$F\text{值} = \frac{2 \times \text{准确率} \times \text{召回率}}{\text{准确率}+\text{召回率}}.$$

4.3 实验结果与分析

传统的基于大规模情感词典的微博主观句抽取算法实验结果入表 5 所示. 本文提出的基于词典与语料结合的微博主观句抽取算法实验结果如表 6 所示. 对比两个表格数据, 以传统的基于大规模情感词典的微博主观句抽取算法为基线, 本文的方法 F 值提高了 7%.

表 5 传统的基于大规模情感词典的微博主观句抽取算法实验结果

Tab. 5 The experiment result base on the traditional large-scale lexicon algorithm

准确率/%	召回率/%	F 值/%	运行时间/s
57.5	79.2	66.6	3.2

表 6 基于词典和语料结合的微博主观句抽取算法实验结果

Tab. 6 The experiment result base on the lexicon and corpus algorithm

α、β 取值	准确率/%	召回率/%	F 值/%	运行时间/s
α=3, β=6	64.8	80.6	71.9	61.3
α=3, β=9	69.6	78.1	73.6	60.5
α=4, β=8	69.7	72.9	71.3	59.1
α=4, β=12	74.0	67.1	70.4	61.1秒

实验结果表明, 传统的基于大规模情感词典的微博主观句抽取算法的准确率较低. 这说明很多情感词语在某些话题的微博句子中并没有表达明显的主观语义. 然而, 传统的基于大规模情感词典的微博主观句抽取算法只需要遍历句子中是否出现情感词语就能判断该句子是否为主观句. 因此, 算法的执行效率较高. 对于本文提出的基于词典与语料结合的微博主观句抽取算法, 主观句抽取的准确率、召回率随着 α、β 取值的不同而不同. 相比较与传统的基于大规模情感词典的微博主观句抽取算法, 本文提出的算法能够在提高准确率的同时提高召回率. 然而, 本文所提算法达到的主观句抽取效果很大程度上依赖于训练语料的规模及训练语料中主客观标记的正确率. 同时, 相比较于传统的基于大规模情感词典的微博主观句抽取算法, 本文所提算法的运行效率有所下降.

5 总 结

本文根据句子中是否包含情感表达文本来判断该句子是否为主观句. 本文的主要贡献是提出了句子的 N-POSW 模型, 并根据句子的 2-POSW 模型, 通过语料学习的方式抽取句子中的情感表达文本, 从而提高微博主观句抽取的召回率. 本文提出的 N-POSW 模型和文献 [8] 中提到的 N-POS 模式具有本质上的区别. N-POS 模式是将 N 个词的词性序列作为句子的特征项; 而本文提出的 N-POSW 模型中, 是将 N 个词作为句子的特征项, 同时考虑词性对特征项情感倾向的影响. 实验结果表明, 本文方法的主观句抽取 F 值可以达到 73.6%, 说明本文方法的可行性.

[参 考 文 献]

[1] KIM S M, HOVY E. Automatic detection of opinion bearing words and sentences[C]//Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). Berlin: Springer, 2005: 61-66.

[2] WIEBE J, WILSON T, BELL M. Identifying collocations for recognizing opinions[C]//Proceedings of the ACL'01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation. Toulouse, FR: ACL, 2001: 24-31.

[3] WIEBE J, WILSON T. Learning to disambiguate potentially subjective expressions[C]//Proceedings of the 6th conference on Natural language learning-Volume 20. Stroudsburg, PA: Association for Computational Linguistics, 2002: 1-7.

(下转第 87 页)

[25] RAJAGOPALAN P, IROH J O. Characterization of polyaniline-polypyrrole composite coatings on low carbon steel a XPS and infrared spectroscopy study[J]. Applied Surface Science, 2003, 218:58-69.

[26] STREET G B, CLARKE T C, GEISS R H, et al. Characterization of polypyrrole[J]. Journal de Physique, 1983, 44:c3, 559-606.

[27] LIU Y-C, HWANG B-J, JIAN W-J, SANTHANAM R. In situ cyclic voltammetry-surface-enhanced Raman spectroscopy studies on the doping undoping of polypyrrole film[J]. Thin Solid Films, 2000, 374:85-91.

（责任编辑　李　艺）

（上接第 68 页）

[4] WILSON T, WIEBE J, HWA R. Just how mad are you? Finding strong and weak opinion clauses[C]//Proceedings of the National Conference on Artificial Intelligence. Menlo Park, CA: MIT Press, 1999, 2004: 761-769.

[5] WILSON T, WIEBE J, HEA R. Recognizing strong and weak opinion clauses[J]. Computational Intelligence. 2006, 22(2): 73-99.

[6] PANG B, LEE L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]//Proceedings of the 42nd annual meeting on Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2004: 271-278.

[7] LONG J, MO Y. Target-dependent Twitter Sentiment Classification [C]//Proceeding of the 49th Annual meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2011: 151-160.

[8] 叶强，张紫琼，罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报，2007, 1(1): 7-91.

[9] 张博. 基于 SVM 的中文观点句抽取[D]. 北京邮电大学，2011.

[10] 杨武，宋静静，唐继强. 中文微博情感分析中主客观句分类方法[J]. 重庆理工大学学报：自然科学，2013, 27(1): 51-56.

[11] 董振东，董强. 知网简介[DB/OL]. [2013-7-20]. http://www.keenage.com.

[12] 台湾大学. NTUSD-简体中文情感极性词典[DB/OL]. [2013-7-20]. http://www.datatang.com/data/11837.

[13] ICTCLAS. ICTCLAS 汉语分词系统[DB/OL]. [2014-06-10]. http://www.ictclas.org.

[14] 中文信息技术专业委员会. 中文微博情感分析评测[EB/OL]. [2013-7-20]. http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html.

（责任编辑　李　艺）