

微博情感倾向性分析

娄鑫坡, 柴玉梅, 咎红英, 韩英杰

(郑州大学信息工程学院, 郑州 450001)

摘要: 本文主要介绍郑州大学自然语言处理实验室(ZZUNLP)在参加NLP&CC2012中文微博情感分析评测中完成的系统。其中评测共分3个任务,包括观点句识别、情感倾向性分析和情感要素抽取。对于三个任务,其中观点句识别和情感倾向性分析采用了基于表情符号的规则方法和基于情感词典的规则方法,情感要素抽取采用了基于依存句法分析的评价对象抽取方法。

关键词: 情感分析; 情感词; 句法分析; 微博

Micro-blog sentiment analysis

Abstract: This paper mainly introduces a system, which we completed when we take part in the evaluation of the NLP&CC 2012 Chinese microblog sentiment analysis. Evaluation is contains three tasks, perspectives sentence recognition、sentiment orientation analysis and sentiment factors of extraction. For three tasks, perspectives sentence recognition and sentiment orientation analysis were performed using a method based on the sentiment rules and sentiment dictionary based rules, emotional factors of extraction based Dependency Parser evaluation object extraction method.

Key Words: Sentiment Analysis; Polar Word; Parsing; MicroBlog

1 简介

微博,即微博客(MicroBlog)的简称,是一个基于用户关系的信息分享、传播以及获取平台,用户可以通过WEB、WAP以及各种客户端组建个人社区,以140字左右的文字更新信息,并实现即时分享。微博自问世以来,迅速吸引了大众的眼光,蓬勃发展。国内以新浪微博为例[1],截至2012年6月30日,注册用户达到3.68亿,平均每日活跃用户达到3650万,每天需要发送近5000万条微博内容,其中也吸引了大量的名人和政客。规模庞大的微博文本也给自然语言处理带来了新的机遇和挑战。在这海量的文本信息中,有很大一部分是带有情感的文本信息。这些文本信息是非常宝贵的意见资源。针对产品的评论,不管是对商家还是对买家都是非常有价值的;对于热点事件的评论,对于政府了解网民对待特定事件的观点也是十分重要的。

微博作为近几年发展起来的一种应用,由于微博信息数量大,更新快,吸引了一大批学者对其进行研究。针对微博的自然语言处理也成为了当前的一个研究热点和前沿课题,情感分析又是当前自然语言研究的热点。情感分析又称意见挖掘(opinion mining),它的目的是通过分析带有主观情感或者褒贬倾向的主观性文本,挖掘其中的观点或评价信息,以其更加直观的方式或者形式呈现给大家。目前已经有了很多针对英文的微博情感分析系统,例如TweetFeel, Twendz,和Tweeter Sentiment。在这些系统中,只要用户给予一个查询,系统就会自动的给出查询的内容是正向情感、负向情感或中性情感。在情感分析方面,主要是采用了两中技术,一种是采用情感词典和规则相结合的方法。另一种是采用基于机器学习的方法,这种方法主要是使用情感词、主题相关特征等作为分类特征,标注训练集和测试集,使用机

器学习的分类方法进行情感倾向性分析，常用的分类器包括朴素贝叶斯 (Naïve Bayes) [1,2]、支持向量机 (support vector machine) [2,3]、最大熵 (Maximum Entropy) [4] 等分类器来进行情感分类。

在本次评测任务的描述中，观点句的定义不包括表达自我情感、意愿或心情的句子，只限定于对特定事物或对象的评价（例如“我真心喜欢 iphone 的屏幕效果。”），不包括内心自我情感、意愿或心情。根据评测任务定义，本文使用人工创建的情感词典和抽取的一些规则实现观点句识别和情感倾向性分析，利用哈尔滨工业大学信息检索和社会计算实验室的句法分析[5]对观点句进行句法分析，根据句法分析结果和褒贬词位置自动的识别出评价对象。本文主要介绍了我们实验室参加评测的系统所使用的方法，并对实验过程以详细的介绍。

2 微薄情感分析模型

2.1 观点句识别

针对于任务1，在没有带标注训练语料的前提下，比较可行的方法是考虑使用情感词典来判断一个句子是否为观点句，本文实验使用的情感词典是人工通过浏览大量微博语料抽取的。情感词是情感极性分析中重要的依据，在传统的情感分析中，把情感词分为正、负情感词，并构建正、负情感词词典。例如，“善良”是带有褒义色彩的词语，而“丑陋”是带有贬义色彩的词语等。

在新浪微博上，微博平台提供了一些默认的表情符号，表情符号在抓下来的文本中的表现形式为“/”加文本，如“/哈哈”。一条信息中可以包含多个文本符号，本文针对微博平台的表情符号进行正、负向表情符号的分类。我们认为只有含有情感词或含有带有情感的表情符号的句子才是表达情感明显的句子，所以含有情感词或者带有情感的表情符号句子为主观句。

2.2 观点倾向性识别

对于微博的情感倾向性的判断具体是正面、反面或者其他，通常都是看修饰的情感词的极性。如果微博中只有一个情感词，那么这条微博的倾向性由这个情感词决定；如果这条微博中含有多个存在的极性冲突的情感词，本系统所使用的方法是分别计算褒贬情感词的个数，考虑两者的总和，然后确定这条微博的极性。

否定副词往往是情感片段极性逆转，所以在本系统中构建了含有 64 个否定词的否定因子表，处理否定词带来的句子极性逆转的问题。其中对于两个棘手问题的处理：

1) 双重否定问题：本系统考虑只有在同一个句子片段（以符号标点为界）中出现两个否定词，并且两个否定词之间没有情感词，才能构成双重否定。

2) 否定词的位置问题：在本系统中，情感词的位置在否定因子表中已经指定，例如“不必/d An/n”其中 An 代表情感词的位置，这里是把情感词放在否定词的后面。

$$\text{情感极性} = \begin{cases} \text{正向情感} & (\text{如果正向情感词数} + \text{正向表情符号数} > \text{负向情感词数} + \text{负向表情符号数}) \\ \text{负向情感} & (\text{如果正向情感词数} + \text{正向表情符号数} < \text{负向情感词数} + \text{负向表情符号数}) \\ \text{中性情感} & (\text{如果正向情感词数} + \text{正向表情符号数} = \text{负向情感词数} + \text{负向表情符号数}) \end{cases}$$

基于情感词词典和表情符号的微博倾向性识别算法如下：

输入：微博集合 $T = \{t_1, t_2, t_3, \dots, t_n\}$, 情感词词典 $Dict$, 表情符号集 $Emotions$;

1) 对于每条微博 t_i 按公式 $Score(t_i)$ 打分

$$Score(t_i) = \sum_{w_{ij} \in t_i} f(w_{ij}, Dict) + \sum_{e_{ij} \in t_i} \delta(e_{ij}, Emotions)$$

其中 w_{ij} 是微博 t_i 中的词, e_{ij} 是微博 t_i 中的表情符号;

$$f(w, Dict) = \begin{cases} 1 & w \text{ 在 } Dict \text{ 词典中的极性为褒义} \\ -1 & w \text{ 在 } Dict \text{ 词典中的极性为贬义} \end{cases}$$

$$\delta(e, Emotions) = \begin{cases} 1 & e \text{ 在表情符号集中为正向} \\ -1 & e \text{ 在表情符号集中为负向} \end{cases}$$

2) 如果这条微博的打分大于 0, 那么这条微博的情感倾向性为正向, 如果这条微博的打分等于 0, 那么这条微博的情感倾向性为负向, 若小于 0, 则为贬义。

2.3 情感要素抽取

本任务要求找出微博中每条观点句中作者的评价对象, 即情感对象。同时判断针对情感对象的观点极性。针对任务, 我们采用了基于句法分析的情感对象抽取。利用 2.1 所描述的微博文本主观性特征进行分类后, 我们对句子进行句法分析及其利用依存关系抽取其中的情感对象。本文利用哈尔滨工业大学信息检索和社会计算实验室[5]开发的汉语句法分析器对句子进行句法分析, 该分析器采用的句法分析形式为依存关系形式, 对输入的句子进行处理。在本系统中主要是查找情感词, 根据情感词往前找情感对象, 一般情感词都出现在情感对象的后面。句法分析首先会对句子进行分词和词性标注, 并在句子的每个词及词性前面加上序号。然后输出句子中所有词与词之间的依存关系, 依存关系中, 每个关系以一个依存对表示, 依存对中第一个词是核心词, 支配第二个词, 如“Ipad3 就要上市了哦”的依存关系对 “[2] 上市_[1]ipad3(SBV)” 这个依存对表示“ipad3”和“上市”存在依存关系 SBV (主谓关系)。我们抽取 SBV 结构中的主语作为情感要素, 然后判断主语前面是否有修饰词, 即与主语构成 ATT(定中关系)关系的词, 如果有则一块抽取。如图 1

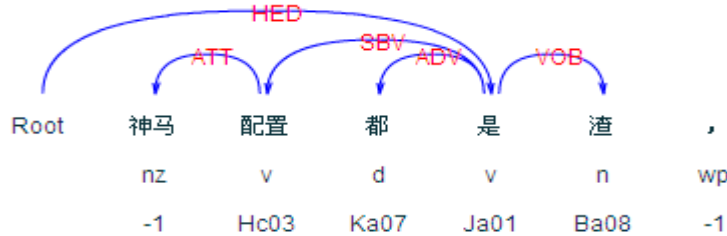


图 1

首先通过情感词“渣”找情面的 SBV 结构, “[2]是_[1]配置(SBV)” 构成 SBV 依存关系, 而依存关系中的主语为配置, 而配置前面有修饰词“神马”, 与配置构成依存关系, 其依存

关系为“[2]配置_[1]神马 (ATT)”，则抽取“神马配置”作为该条微博的情感对象。

情感要素抽取过程如下：

- Step1:** 对每条微博用哈尔滨工业大学的句法分析处理。
- Step2:** 对于句子中每个词对应情感词词典 Dict 判断是否为情感词。如果是情感词转（3），否则继续查找，如果到句子末尾转（6）
- Step3:** 如果该词为情感词，则查找该情感词前面是否含有 SBV 依存关系，是则判断该依存关系中主语是否为第一人称，如果主语不是第一人称则转（4），否则一直往前查找 SBV 结构，如果到句首依然没有则转（5）。
- Step4:** 抽取该主语并且查找主语前是否有词与该主语形成 ATT 依存关系，有则一块抽取并一直往前找，直到没有存在 ATT 依存关系为止。
- Step5:** 抽取主题作为情感对象。转（2）
- Step6:** 转下一条微博处理。

4 实验及分析

NLP&CC2012 的评测语料包含大约 30, 000 篇微博，给定了 20 个主题，其中包括产品名，事件名等。情感词词典是通过人工浏览大量的微博语料抽取的，其中褒义词 577 个，贬义词 3548 个。

4.1 任务 1

任务描述：针对每条微博的中各个句子，本任务要求判断该句是观点句还是非观点句。观点句的定义不包括表达自我情感、意愿或心情的句子，只限于对特定事物或对象的评价，不包括内心自我情感、意愿或心情。

表一 任务 1 的评测结果

评测结果	微平均			宏平均		
	正确率	召回率	F 值	正确率	召回率	F 值
ZZU_opinion	0.765	0.647	0.701	0.760	0.640	0.680
平均结果	0.727	0.615	0.647	0.727	0.607	0.634
最佳结果	0.671	0.944	0.784	0.674	0.942	0.783

4.2 任务 2

任务描述：本任务要求判断微博中每天观点句的情感倾向。评测数据集包括每条微博中的各个句子，需要在任务 1 观点句识别的基础上再进行观点句的倾向性分许。观点句的情感倾向性可以分为正面（POS），负面（NEG）和其他（OTHER）。

表二 任务 2 的评测结果

评测结果	微平均			宏平均		
	正确率	召回率	F 值	正确率	召回率	F 值
ZZU_polarity	0.902	0.584	0.709	0.899	0.578	0.690
平均结果	0.745	0.455	0.552	0.744	0.450	0.542
最佳结果	0.853	0.743	0.794	0.854	0.745	0.794

4.3 任务 3

任务描述：本任务要求找出微博中每条观点句作者的评价对象，即情感对象。同时判断针对情感对象的观点极性。

表三 任务 3 的评测结果

评测结果	微平均			宏平均		
	正确率	召回率	F 值	正确率	召回率	F 值
ZZU_object	0.260	0.164	0.201	0.400	0.172	0.205
平均结果	0.172	0.098	0.112	0.173	0.098	0.112
最佳结果	0.303	0.273	0.288	0.306	0.263	0.278

从表 1，表 2 的对比结果可以看出，两项任务的正确率都高于评测的最佳结果，但由于召回率偏低使得 F 值低于评测最佳结果。分析其原因：在观点句识别和情感倾向性分析上，我们采用了基于词典的方法，而词典本身具有一定的局限性。另外，微博中存在一些不包含情感词的情感句，也影响了实验结果。

表 3 是情感对象抽取的对比结果，本系统的准确率和召回率都高于评测的平均结果，说明采用依存句法分析的方法进行对象的抽取还是比较可行的。在之后的学习研究中，进一步进行深入的分析，使系统的性能更好。

5 结束语

本文主要介绍了郑州大学自然语言实验室在参加 NLP&CC2012 评测任务所做的工作，评测任务包括观点句识别、情感倾向性分析和情感要素抽取等。对系统采用的方法和关键技术进行了详细的阐述，尽管都是采用了比较简易实现的方法。从评测结果来看，简单的方法性能也是比较有效的，但也有不足的地方，总体来说，尚待改进的地方还很多。网络用语词典的构建，利用社交网络关系或信息与信息之间的关系来提升结果。

参考文献

- [1] E. Riloff, J. Wiebe and T. Wilson, "Learning Subjective Nouns Using Extraction Pattern Bootstrapping," In CoNLL03, PP.25-32,2003
- [2] W.H.Lin, T. Wilson, J. Wiebe and A. Hauptmann, "Which Side are You on? Identifying Perspectives at the Document and Sentence Level," In: CoNLL06, PP.109-116, 2006
- [3] E. Riloff, S. Patwardhan and J. Wiebe, "Feature subsumption for opinion analysis," In EMNLP06, 2006
- [4] B. Pang and L. Lee, "A sentiment Education: Sentiment Analysis Using Subjectivity Summarization based on Minimum Cuts." In ACL04, pp.271-278, Spain, 2004
- [5] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations. 2010.08, pp13-16, Beijing, China.