基于依存分析和褒贬指向的微博情感对象抽取方法
（NLP&CC 2012 南京大学参评系统介绍）

高磊[1,†] 李斌[1,2] 戴新宇[1] 黄书剑[1] 陈家骏[1]

1.南京大学 计算机软件新技术国家重点实验室 南京 210046；

2.南京师范大学 语言信息科技研究中心 南京 210097

† 通讯作者，E-mail: gaol@nlp.nju.edu.cn

摘要 本文介绍了南京大学计算机科学与技术系自然语言处理实验室(NJU-NLP)参加第一届 CCF 自然语言处理与中文计算会议(NLP&CC2012)组织的面向微博的情感分析评测的情况。我们参加了面向微博的情感分析任务中的观点句识别、情感倾向性判断和情感要素抽取全部三个子任务。在观点句识别子任务和情感倾向性判断子任务中，我们采用分词后使用情感词典识别情感词的方法识别观点句并判断情感倾向。在情感要素抽取子任务中我们在句法分析的基础上使用了情感词的褒贬指向规则来确定情感对象。本文详细介绍了参评系统以及参加评测的结果。

关键词 微博；情感分析；褒贬指向

Extraction of Sentiment Targets of Chinese Micro-blog Based on Dependency Parsing and Derogatory and Commendatory Direction Tendency(DCDT)

Lei Gao[1,†], Bin Li[1,2], Xinyu Dai[1], Shujian Huang[1], and Jiajun Chen[1]

1.State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing 210046;

2.Research Center of Language and Informatics, Nanjing Normal University, Nanjing 210097

†Corresponding Author, E-mail: gaol@nlp.nju.edu.cn

Abstract This paper describes our (NJU-NLP) participation for the evaluation of the 1st CCF Conference on Natural Language Processing & Chinese Computing. We submit results for all three subtasks of the sentiment analysis of Chinese micro-blog task, including opinion sentences recognition, sentiment tendency analysis, and sentiment targets extraction. In opinion sentences recognition and sentiment tendency analysis, we use

tokenization and emotional directories to recognize the opinion sentences and judge the sentiment tendency of each opinion sentence. In sentiment targets extraction, we use the rules of Derogatory and Commendatory Direction Tendency(DCDT) of sentiment words to acquire the sentiment targets. This paper makes a detailed description of our system, as well as the evaluation results.

Keywords micro-blog; sentiment analysis; derogatory and commendatory direction tendency

1. 引言

第一届 CCF 自然语言处理与中文计算会议(NLP&CC2012)评测项目共有两个任务: 面向微博的情感分析任务和中文词汇语义关系抽取任务。其中，面向微博的情感分析任务又分为三个子任务：观点句识别、情感倾向性判断和情感要素抽取。南京大学计算机科学与技术系自然语言处理实验室(NJU-NLP)作为参评单位之一参加了面向微博的情感分析任务的所有三个子任务。本文主要介绍我们的参评系统、相关技术，以及在评测任务上的性能表现。

2. 参评系统描述

微博是一种新兴的信息分享平台，许多微博文本的内容表达了对某个事件的个人看法，带有明显的情感倾向。对微博的情感分析逐渐成为研究热点[1,3,7]。我们参考前人的研究，在此次评测的面向微博的情感分析任务中，提出了一种基于依存分析和褒贬指向的获取情感对象的方法。

在观点句识别和情感倾向性判断子任务中，本文参考 Kim 等人的方法[4]，根据情感词典等资源识别出文本中的情感词，进行加权求和，计算出句子的情感倾向。考虑到微博文本一般比较简短[13]，含有的情感词相对较少，因此我们仅通过比较句子中情感词的数量和一些启发式的规则来确定观点句及其极性。

在情感要素抽取子任务中，本文采用了一种以情感词的褒贬指向识别情感要素的方法。褒贬指向是指，由褒贬词语的褒贬义所决定的，评价者对褒贬对象的态度在语义角色上呈现的指向性[11]。基于已有的研究，我们以依存树为基础，根据一些褒贬指向的规则，确定情感对象。对于作为动词的情感词，以依存树中谓词的语义角色，自动获取其施事方作为情感对象；对于作为名词和形容词的情感词，根据褒贬指向规则，在依存树中选择相应的句法成分作为情感对象。

根据三个子任务，我们的系统分为三个处理模块：

(1) 参照 Che Wanxiang 等人的方法和工具对微博句子进行分词[9]，使用情感词典识别出情感词，从而识别出观点句。

(2) 对观点句进行依存分析[5]，得到微博句子的依存树，再根据修饰情感词的词中是否含有否定词进行情感倾向性判断[12]。

(3) 根据依存树中情感词的语义角色信息[10]，并参考“词语褒贬指向总表”[11]，识别出情感对象。

2.1 分词与依存句法分析

与普通文本相比，微博文本具有如下特性：文本长度短，非正式（即口语化），半结构化，包含大量的省略和指代等[13]。所以在分析之前，需要对微博语料做相应的预处理，以消除某些内容对分词和依存句法分析的不良影响。

由于目前尚没有微博文本专用的分词和依存分析工具，我们采用哈工大社会计算与信息检索研究中心发布的语言技术平台(LTP)中文语言处理系统[9]，对微博语料进行分词和依存分析一体化分析。

2.2情感词典

情感词典 HowNet2007 和 NTUSD 收录的褒贬词比较丰富，包含了相当数量的简体和繁体褒贬词，我们认为可以用于评测任务。表 1 给出了情感词典相关信息。

表 1 情感词典信息

Table 1 Sentiment Dictionaries

词典名称	语言	词条数(褒/贬)	url
HowNet2007	简体	3103/3288	http://www.keenage.com/
NTUSD	繁体/简体	2812/8276	http://nlg18.csie.ntu.edu.tw:8080/opinion/

2.3词语褒贬指向规则

考虑到微博句子的文本长度短和非正式化的特性[13]，我们采用一些结构相对简单的词语褒贬指向规则[11]。表 2 给出了本文所使用的褒贬指向规则。

比如“球队凯旋归来”，就是一价动词做谓语中心语的结构，情感词“凯旋”指向“球队”。再如“他很有风度”，为“施事+有+名词”的结构，情感词“风度”指向“他”。

其中对于一价动词做谓语中心词的情况，由于依存树中对于每个动词都会给出语义角色信息，我们直接根据其指向的施事方来确定情感对象。其他的词则按照表 2 中的规则在依存树中寻找情感对象。

表 2 情感词褒贬指向规则

Table 2 DCDT rules

词语类别	句法功能或格式	态度持有者	褒贬对象
一价动词	做谓语中心语	说话人	施事
一价名词	施事+有+名词	说话人	施事
一价形容词	做谓语中心语	说话人	施事
一价形容词	做定语	说话人	修饰的对象

3. 实验

3.1预处理

在进行依存分析前对微博语料进行预处理，是大多数微博分析系统的必需步骤[1,2,6]。本文采取的预处理主要包括过滤一些结构或功能符号(如“@”)，错误词修正，移除停用词等。

我们通过观察本次评测的测试语料，人工设定了一些需要预处理的内容，主要包括：

(1) 主题词。例如“#90 后当教授#当什么教授 怎么玩是吗”中的“ #90 后当教授#”。

(2) 特殊符号。例如“呃，无语了……..-_-”中的“-_-”。

(3) 表情符号。例如“这酒店浪费可耻/敲打这大姐尊守纪律就没事了”中的“/敲打”。

(4) url。例如“建议每一位朋友都可以看看《三个白痴》这部影片，http://url.cn/4bZYQt”中的“http://url.cn/4bZYQt”。

(5) 文本结构字符。例如“希望大家给转播一下\'希望别人不要再受骗上当”中的“\&apos”。

3.2分词和依存句法分析

我们使用工具 LTP 对预处理后的微博语料进行分词和依存分析一体化分析，能够得到每个句子的依存树。但是依存树的结构中缺乏一些表示句子情感属性的元素，其已有元素还不能满足我

们进行后续分析的需要。因此我们在依存树的结构中添加了几个表示观点句、句子极性，以及情感词数组的元素，如表 3。

表 3 依存树结构中添加的元素

Table 3 Elements Added into Dependency Tree

元素名称	功能	类型	默认值
opinionated	表示句子是否为观点句	布尔值	“N”
polarity	表示句子的极性	布尔值	“N”
sent_pos	表示句子中含有的所有正面情感词	数组	“NULL”
sent_neg	表示句子中含有的所有负面情感词	数组	“NULL”

3.3观点句识别和情感倾向性判断

在得到各句子的依存树之后，我们使用情感词典进行观点句识别和情感倾向性判断，主要算法为：

(1) 在各情感词典中查找句子中的每个词。若在正面情感词典中找到，则将词的名称、位置存入 sent_neg 项；如果在负面情感词典中找到，则将词的名称、位置存入 sent_pos 项。

(2) 观察句子中的 sent_neg 和 sent_pos 项。若都为空，则认为句子不是观点句，否则为观点句。

(3) 对于确定为观点句的句子，判断情感倾向的伪代码见图 1。先观察句子中是否含有否定词(根据句法结构中的项如“sweep=“不_没””)，然后再比较句子中的 sent_neg 和 sent_pos 项，以确定句子的极性。若没有否定词且 sent_neg 项含有的词数大于等于 sent_pos 含有的词数，则认为此句为负面情感句，否则为正面情感句；若含有否定词且 sent_pos 项不为空，则认为此句为负面情感句，否则为正面情感句。这里的规则是启发式的，根据人的日常表达习惯确定，其准确性需要进一步的研究。

```
IF (不存在否定词)
    IF (sent_neg 中元素数 >= sent_pos 中元素数) THEN 情感倾向为正面;
    ELSE 情感倾向为负面;
ELSE
    IF(sent_pos 中元素数 >= 1) THEN 情感倾向为负面;
    ELSE 情感倾向为正面;
```

图 1 情感倾向性判断伪代码

Fig. 1 pseudo code of judging the sentiment tendency

3.4情感要素抽取

我们利用依存树中谓词的语义角色信息以及非谓词的褒贬指向规则，对各观点句进行情感要素抽取，主要分为以下两部分：

(1) 根据依存树中的语义角色信息，选取情感对象，如图 2。这里的情感词“完蛋”为施事方，根据其指向的 id，指向的情感对象为“老匹夫”。

(2) 由于只有依存树中的谓词才有相应的语义角色信息，对于非谓词的情感词，根据“词语褒贬指向总表”的部分规则，确定其情感对象，如图 3。这里的情感词“缺德”为形容词，根据规则“一价形容词做谓语中心语，对施事进行褒贬”，选择情感词修饰的名词短语作为情感对象。这里选用情感词前面最近的一个主谓关系词(“SBV”)及其修饰成分作为情感对象，即“这孩子”。

```
<sent id="0" cont="老匹夫不完蛋，世界不会太平。">
    <word id="0" cont="老" pos="a" ne="O" parent="1" relate="ATT" wsd="Eb15" wsdexp="嫩_老" />
    <word id="1" cont="匹夫" pos="n" ne="O" parent="3" relate="SBV" wsd="Aa01" wsdexp="人_人民_众人" />
    <word id="2" cont="不" pos="d" ne="O" parent="3" relate="ADV" wsd="Ka18" wsdexp="不_没" />
    <word id="3" cont="完蛋" pos="v" ne="O" parent="-1" relate="HED" wsd="Ib03" wsdexp="活_死">
        <arg id="0" type="施事" beg="0" end="1" />
```

图 2 依存树示例 1

Fig. 2 instance no.1 of dependency tree

```
<sent id="0" cont="这孩子太缺德了。">
    <word id="0" cont="这" pos="r" ne="O" parent="1" relate="ATT" wsd="Ed61" wsdexp="这个_那个_某个_各个_其他_何" />
    <word id="1" cont="孩子" pos="n" ne="O" parent="3" relate="SBV" wsd="Ab04" wsdexp="婴儿_儿童" />
    <word id="2" cont="太" pos="d" ne="O" parent="3" relate="ADV" wsd="Ka02" wsdexp="最_至多_至少" />
    <word id="3" cont="缺德" pos="a" ne="O" parent="-1" relate="HED" wsd="Ee02" wsdexp="厚道_刻薄_缺德" />
    <word id="4" cont="了" pos="u" ne="O" parent="3" relate="MT" wsd="Kd05" wsdexp="吗_的话_罢了_了_哉" />
    <word id="5" cont="。" pos="wp" ne="O" parent="-2" relate="PUN" wsd="-1" wsdexp="" />
```

图 3 依存树示例 2

Fig. 3 instance no.2 of dependency tree

4. 结果与分析

我们根据评测结果中的微平均值(以整个数据集为一个评价单元)，分别观察三个子任务的得分情况(正确率 P，召回率 R，F 值 F)，以及相应的排名，对我们的方法进行分析。

4.1观点句识别

本子任务共有 53 支队伍提交了结果，我们的得分为 P＝0.695，R=0.473，F＝0.563，排名分别是 42 名、40 名和 42 名。与其它参赛组的结果相比，正确率和召回率都偏低。虽然与其他组仍有差距，但是近 70%的正确率说明，使用分词加情感词典的方法识别观点句是可行的。召回率偏低则说明，这种方法对分词的正确性和情感词典的完备性要求较高。

对于这一子任务，我们的方法在以下方面仍有提高的空间：

(1) 所选用的分词工具是面向一般文本的，对于微博文本，可能由于领域特殊性[13]而达不到在一般文本上的分词效果，造成某些情感词没有被正确分词。

(2) 所选用的情感词典也不是专门面向微博文本的，对于微博文本来说可能很不完备，有些微博特有的情感词不会被识别出来。例如：在句子“最好是把扣扣马化腾给弄下来，太坑爹了”中，情感词“坑爹”就没有被识别出来。

4.2情感倾向性判断

本子任务共有 53 支队伍提交了结果，我们的得分为 P＝0.803，R=0.379，F＝0.515，排名分别是 23 名、34 名和 32 名，与前一子任务相比有所上升。这里召回率进一步下降，但应当看到在有限的(约 47%)已识别出的观点句中，有约 80%的句子的情感倾向性被正确识别出来，说明我们的用于判断观点句情感倾向性的方法是可行的。

这里可能改进的地方为，所使用的启发式的规则还是比较粗糙的，可能对一些结构稍复杂的句子(如双重否定句)无法做出正确判断，对这些规则进行优化可能会得到更好的效果。

4.3情感要素抽取

本子任务参考严格评价指标，共有 22 支队伍提交了结果，我们的得分为 P＝0.182，R=0.100，F＝0.129，排名分别是 8 名、9 名和 7 名。在这一子任务中，多数参赛组的得分与上个任务相比下降的很多。相比之下，在前两个子任务的结果并不突出的基础上，我们的得分比较靠前，这就说明我们的用于情感要素抽取的方法具有相对较好的性能。

本子任务的召回率和正确率都较低，我们认为有以下原因：

(1) 情感词作为谓词并不占多数，依靠依存树中谓词的语义角色信息只能识别一小部分情感对

象。

(2) 由于前面分词和依存句法分析的结果可能不恰当，即使识别出情感词，所得到的情感对象成分也是不恰当的，例如句子“这是个物欲横行的社会”，其虽然识别出情感词“横行”，句法分析却将“横行”分析为动词“v”，最后选择出来的情感受体为“个物欲”。

(3) 在利用词语褒贬指向规则时，选用情感词前面最近的一个主谓关系词(“SBV”)及其修饰成分作为情感对象这种方式，可能不适用于所有句子。

综上，我们认为，前期对观点句选择的不够充分，利用词语褒贬指向规则的方式不够多样，以及依存树中谓词的语义角色信息所占比例较小，都造成了结果的召回率较低；而部分不恰当的分词和依存句法分析结果则造成了正确率较低。此外，由于目前缺乏有情感对象标注的大量的微博语料，我们难以对微博情感词和情感对象的性质进行全面的观察比较，也限制了我们对自己方法性能的评价和改进。而且，如果有更多的标注数据，则可以尝试采用一些机器学习方法[8]，可能会有更好的效果。

5. 总结

本文系统地描述了南京大学计算机科学与技术系自然语言处理实验室(NJU-NLP)参加第一届 CCF 自然语言处理与中文计算会议(NLP&CC2012)组织的面向微博的情感分析评测的情况。我们使用了一种基于依存分析和褒贬指向的微博情感对象抽取方法，结果表明：依存句法分析的结果，有助于分析情感态度；褒贬指向是有用的，但限于所采用的分词和句法分析工具的领域适用性以及褒贬词典规模，自动分析的效果并不理想。

在今后的工作中，我们将进一步加强面向微博领域的词法和句法分析技术研究，标注更大规模词表的褒贬指向，将两者更好的结合起来，从而提高微博的情感分析精度。

致谢

本文承国家自然科学基金（课题编号：61170181）、国家社科基金（课题编号：10CYY021）、中国博士后基金（课题编号：2012M510178）的资助。

参考文献

[1] Go, A., Bhayani, R., Huang, L. Twitter Sentiment Classification using Distant Supervision. Technical report, Stanford Digital Library Technologies Project, 2009.

[2] Pak, A., Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of LREC, 2010:1320–1326.

[3] Davidov, D., Tsur, O., and Rappoport, A. : Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Coling, 2010

[4] Soo-Min, K., Hovy E. Automatic Detection of Opinion Bearing Words and Sentences. Proc. of International Joint Conference on Natural Language Processing. Jeju Island, Korea: [s. n.],2005

[5] Ting, L., Ma, J., Li, S. Building a dependency treebank for improving Chinese parser. Journal of Chinese Language and Computing, 2006, 16(4): 207–224

[6] Barbosa, L., Feng, J. Robust Sentiment Detection on Twitter from Biased and Noisy Data. Coling, 2010

[7] Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J. Twitter polarity classification with label propagation over lexical links and the follower graph. Proceedings of the EMNLP First workshop on Unsupervised

Learning in NLP, 2011: 53–63

[8] Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment Classification Using Machine Learning Techniques. Proc. Of Conference on Empirical Methods in Natural Language Processing [S. l.]: ACM Press, 2002

[9] Wanxiang, C., Zhenghua, L,. Ting L. : LTP: A Chinese Language Technology Platform. In Proceedings of the Coling , 2010

[10] 晋耀红, 苗传江. 一个基于语境框架的文本特征提取算法. 计算机研究与发展, 2004, 4

[11] 李斌, 陈小荷. 汉语褒贬词语的褒贬指向问题. 语言文字应用, 2009, 3

[12] 姚天昉, 娄德成. 汉语情感词语义倾向判别的研究. ICCC2007:第七届中文信息处理国际会议论文集.北京: 电子工业出版社, 2007: 221-225.

[13] 张剑峰, 夏云庆, 姚建民. 微博文本处理研究综述. 中文信息学报, 2012, 26(4)