LTLAB 中文微博情感分析评测报告

周霄[1]　周振宇[1]　李芳[1,†]

1. 上海交通大学电子信息与电气工程学院计算机科学与工程系，上海 200240
† 通信作者，E-mail: fli@sjtu.edu.cn

摘要　本文介绍了上海交通大学中德语言技术联合实验室（LTLAB）参加 2012 年中文微博情感分析评测的方法实现。在本届评测设立的 3 个评测任务中，LTLAB 分别参加了任务 1（观点句识别）和任务 3（情感要素抽取）：对于任务 1，参评系统使用了基于分类器的方案，特征抽取时考虑到了文本中的词性和句法特征；对于任务 3，参评系统结合了基于模板的抽取和基于分类器的抽取，考虑到了微博特有的话题信息以及词的统计信息。评测结果表明，本文提出的方法在实践中是行之有效的。

关键词　微博; 情感分析; 观点句识别; 情感要素抽取

中图分类号　X123

LTLAB at Chinese Microblog Sentiment Analysis Track

ZHOU Xiao[1]，ZHOU Zhenyu[1]，　LI Fang[1,†]

1. Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240

†Corresponding author, E-mail: fli@sjtu.edu.cn

Abstract　This paper introduces the system that used by UDS-SJTU Joint Research Lab for Language Technology to participate in the first Chinese Microblog Sentiment Analysis Track。LTLAB attended 2 out of all 3 tasks, which are task 1(identifying opinionated sentences) and task 3(extracting sentiment element)。In task 1, system takes advantages of POS tag and syntactic information, and uses a method based on classification to solve the problem. In task 3, classification based method and pattern based method are combined to provide final result; Topic information and statistical information of words are also considered. Evaluation result reveals the soundness of proposed system.

Key words　Microblog; Sentiment analysis; Identifying opinionated sentence; Extracting sentiment element

1　引言

在本届中文微博情感分析评测中，本实验室参加了任务 1 和任务 3，分别对应于句子级别的情感分析和情感要素级别的情感分析。本届评测的要求包含了一些比较特殊的地方：任务 1 中，评测要求识别的观点句不能是表达自我情感、意愿和心情的句子，例如“我感到很高兴”这样的句子是情感句，但不属于本次评测定义的观点句；任务 3 要求抽取的评价对象不限于本句，而是可以在整条微博中寻找，甚至可以包括话题符号中的文本。我们针对这些特殊的要求设计了参评系统，并取得了比较好的结果。下面，本文将在第 2 节中对 LTLAB 参加本次评测的方法进行介绍，在第 3 节介绍实验结果和分析，最后在第 4 节中进行总结。

2　系统描述

本次参评系统的整体架构如图 1 所示。输入关于一个话题的若干微博，系统首先对其进行预处理，这一过程包括分词、句法分析和评价词抽取 3 个步骤；根据预处理的结果，使用一个分类器进行观点句判别，输出即为任务 1 所求；使用上步结果，继续进行情感要素抽取，对于每个观点句输出一组<评价对象，评价极性>的二元组。

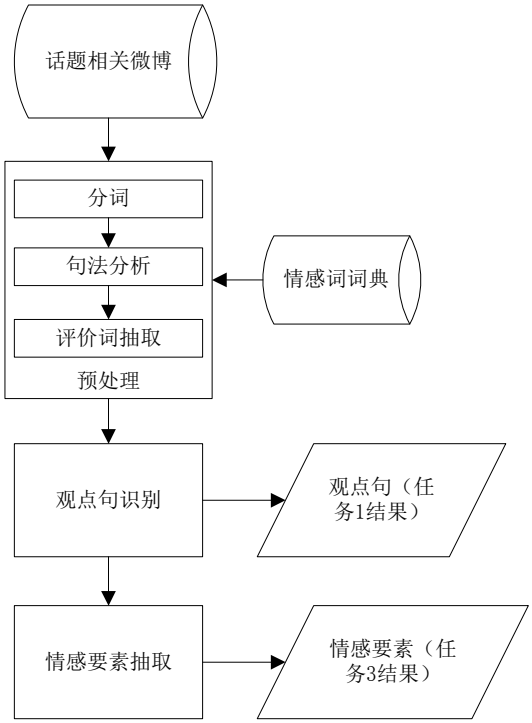

图 1　系统流程图
Fig.1　Framework of the system

2.1　预处理

考虑到微博文本的特殊性，预处理部分并不直接对整句进行操作，而是将句子分成了四种不同的成分：1）URL，是以“http://”开头的网址信息；2）用户名及转发，表现为“@”后跟微博用户名，或者“|| @”后跟用户名；3）话题，即由一对“#”包裹的字符；4）正常文本信息。这四种成分中，只有最后一种成分会被分词和进行句法分析。对句子进行划分一方面改善了后续处理的质量，另外一方面也由于缩短了每次处理的文本的长度从而加快了句法分析。

系统分词使用了中科院的分词系统ICTCLAS[1]，句法分析使用了Stanford Parser[2]。为了改善分词的质量，评测小组成员从网络上和评测给的样例数据集里抽出了部分常用网络用语加入了ICTCLAS的用户词典中。句法分析部分，优先调用Stanford Parser提供的训练好的“xinhuanFactored”模型，如果失败的话（可能由于句子过长或未找到一致解），再调用“xinhuaPCFG”模型进行句法分析。

由于参评系统的观点句识别模块和情感要素抽取模块都使用到了句子中的评价词信息，因此评价词抽取这一步骤被放到了预处理中进行。系统基于评价词词典对句子中的词或词组进行匹配，输出微博中的评价词及其评价极性。具体细节如下：

1) 评价词词典由 3 部分构成：Hownet情感词集合[3]，NTUSD情感词集合[4]，以及来自网络和评测数据样例的评价词。
2) 对于评价词词典中的单字词，人工对其进行了筛选，并对于每个词限定了词性。例如“土”，当其词性为形容词的时候就判定为评价词，当词性为名词的时候就不判定为评价词。
3) 除了词的完全匹配之外，评价词词典中还被添加了少量的正则表达式模板。例如，“丢.{1,4}的脸”可以匹配“丢我的脸”。
4) 对于抽取出的评价词，首先设定初始评价极性为词典中的极性；然后向前一个范围内寻找是否含有表否定的词语，如“不”、“没有”等，每找到一个词就对当前评价极性进行一次反转。

[1] ICTCLAS: http://www.ictclas.org/
[2] Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml
[3] Hownet: http://www.keenage.com/
[4] NTUSD: http://nlg18.csie.ntu.edu.tw:8080/opinion/#

2.2 观点句识别

观点句识别任务要求给定一条微博，判断其中各句是否为观点句。针对这一任务，参评系统采用了一个基于文本分类的方案。

2.2.1 分类器的选择

参评系统选用了VFI（Voting Feature Interval）[1]分类器。VFI分类器的原理较简单，正如其名字表示的那样，它将每个特征划分成若干个区间。在训练的时候，对每个特征的每个区间上的各种类标识（class label）进行计数。在进行预测过程的时候，各个特征对结果进行投票，最后各个类的得分正比于该类得到投票数。

微博中观点句数量显著小于非观点句数量，因此这是一个不平衡分类问题。由于VFI分类器不显式的考虑先验概率，因此对训练数据的不平衡并不敏感，比较适合当前任务。评测小组人工对样例数据集中的“安徽萧县车祸”相关微博进行了人工标注，作为实验语料。结果显示，相比于更加流行的使用SVM、Naïve Bayesian分类器的做法[2]，VFI分类器在保证准确率没有降低很多的情况下极大的提升了召回率，也使得后续任务表现更好。

表 1　不同分类器的表现

Table 1　Results of different classifier

分类器 \ 指标	准确率	召回率	F 值
SVM	0.779	0.726	0.7516
NB	0.708	0.723	0.7154
Tree	0.719	0.689	0.7037
VFI	0.748	0.776	0.7617

针对微博中的每个句子，系统从以下几个方面设计了特征：

1) **基于词性的特征** 选取了一些在观点句中经常出现的词性或词性组合作为特征，包括：连词个数、非“我”代词个数、副词后跟形容词的词组个数、形容词后跟名词的词组个数、“不”后跟形容词的词组个数。表 2 中列出了一些例句。

表 2　词性相关特征

Table 2　POS related features

例句	符合的词性特征
虽然 16 岁还未成年人，但是对社会影响极坏，社会危险性极大，应依法从重判处死刑	“虽然”和“但是”都是连词。连词个数为 2
法律制裁不了这人渣，大家一起搞死他	“他”是非“我”代词。非“我”代词个数为 2
太可恶了，这种人不应该留，太自以为是，把人家毁容了，以后的生活该怎么办?	“太可恶”和“太自以为是”都是副词后跟形容词。副词后跟形容词的词组个数为 2
这么美好的女孩子现在变成这样子了。	“美好的女孩子”是形容词后跟名词。形容词后跟名词的词组个数为 2
男孩的父母心肠不好，应该换位思考如果是他儿子遭毁容了他们怎么想呢	“不好”是“不”之后跟形容词。“不”后跟形容词的词组个数为 2

2) **基于标点符号的特征** 问号与感叹号常伴随主观表达出现，因此也将句中感叹号和问号的数量作为一个特征。

3) **基于评价词的特征** 观点句的一个主要特点就是其中包含评价词，所以这里将句中评价词的个数作为一个特征。

4) **基于主观意愿词的特征** 如在第一节中介绍的那样，本次评测要求观点句不能是纯粹表达主观意愿的句子。为了满足这一要求，评测小组人工收集了一些纯粹表达主观意愿的词，并将句子中这种词的个数也作为一个特征。

表 3 中列举出了一些包含评价词和表达主观意愿的词的句子。

表 3　评价词和主观意愿词特征

Table 3　Opinionated word feature and subjective word feature

例句	评价词或主观意愿词
陶汝坤 让我们记住这个肮脏的名字吧，，，真是个混蛋啊	“肮脏”、“混蛋”是评价词。评价词个数为 2
iPad3 不知道性能好不好，如果好的话可以考虑买一个，永远怀念乔布斯	“好”、“怀念”是评价词。评价词个数为 2
祝愿苹果能出更优秀的产品，以此刺激一下中国睡狮	“祝愿”是主观意愿词。主观意愿词个数为 1
希望周岩能在社会的帮助下，早日康复。	“希望”是主观意愿词。主观意愿词个数为 1

2.2.3　对分类器的改进

参评系统利用了Weka[1]工具包对VFI分类器的实现：对于每一个分类实例，分类器会给出一个 0 到 1 之间的实数值作为其得分，默认情况下得分超过 0.5，就判定候选句为观点句。

在实际使用的过程中，参评系统进行了两个改进：

1) 实验中发现，有些观点句的得分非常接近但低于 0.5。为了确定阈值，评测小组人工标注了评测 20 个事件的前 100 条语料，采用不同的阈值对结果进行评估，实验结果如图 2 所示。我们最终确定了 0.488 为最终判别观点句的阈值。

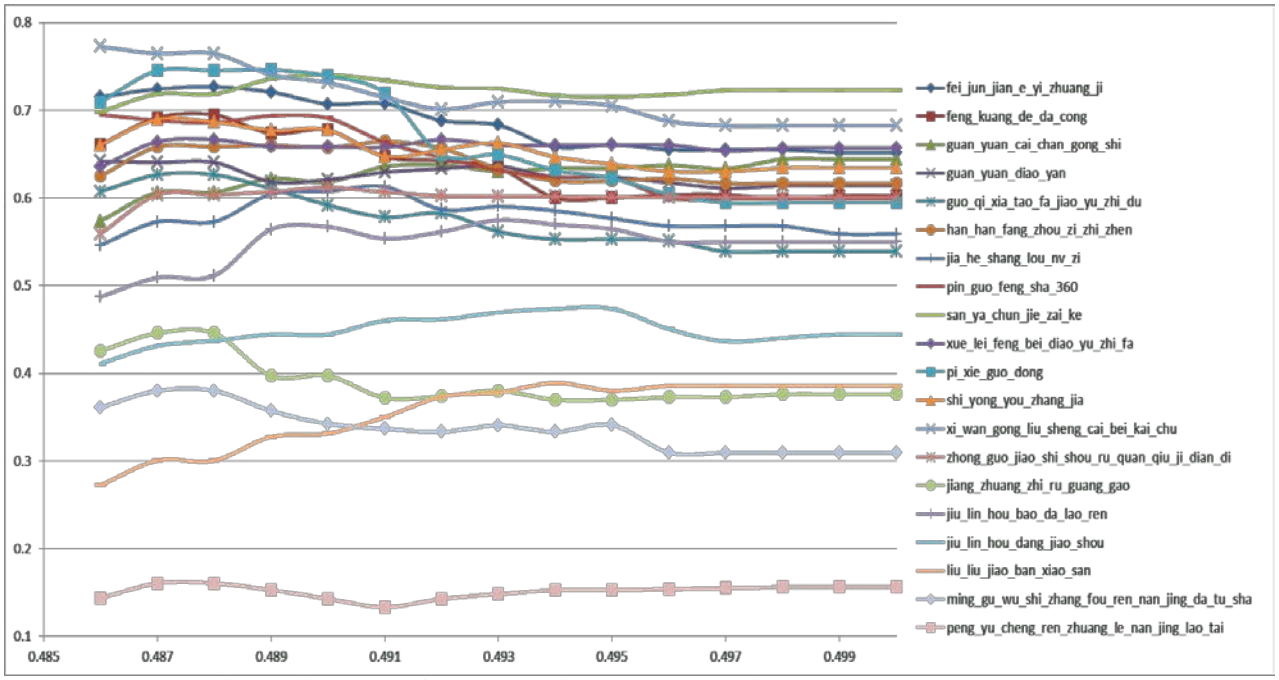

图 2 VFI 模型在 20 个测试语料集的前 100 人工标注结果上的阈值-F 曲线

图 2　采用不同阈值的测试结果

Fig.2　Results of different thresholds

Weka 在原始 VFI 分类器的基础上加入了各个特征权重的概念，然而实验发现该策略发挥作用不大，各个特征之间的权重相差很小，因此系统采取了人工设定权值的方法。有些特征，如评价词的在不同区间的比例差别较为明显，这往往在求总得分时起主导作用。经过实验发现，对评价词适当降低权重效果更好。在最终模型中评价词的权重为 0.8，其他特征的权重为 1。

2.3　情感要素抽取

情感要素抽取任务要求抽取出观点句子中的评价对象，并判断对该评价对象的评价极性。本次评测中对评价对象的位置不做限制，也就是说一个句子的评价对象可能出现在本微博的其它句子中。针对这一特点，参评系统首先使用一个基于分类器的方法抽取评价对象，然后再辅以一个基于模板的方法进行第二次评价对象抽取以提高查全率，最后对于抽取出的评价对象进行评价极性判别。

2.3.1　基于分类器的评价对象抽取

基于分类器的评价对象抽取方案采取的流程如图 3 所示。对于一条微博，首先抽取其中的候选评价对

[1] Weka: http://www.cs.waikato.ac.nz/ml/weka/

象，然后对于微博中的每个观点句，分别判断每个候选评价对象是否是其正确评价对象。

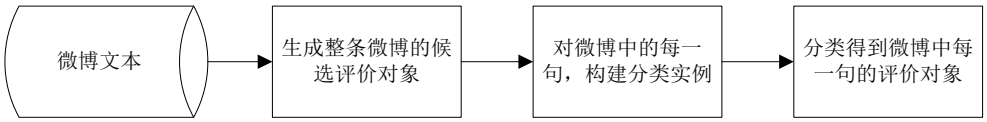

图 3　基于分类器的评价对象抽取
Fig.3　Target extraction based on classifier

候选评价对象生成

在相关工作中[3-7]，评价对象常常被限定为名词或名词词组。另外，通过观察数据，也可以发现话题符号中的词经常被显式或隐式的评论。因此，我们将微博中以下两种文字抽取为候选评价对象：长度小于阈值的、句法成分为NN、NR、NP、IP、FW之一的词组；是当前话题的关键词之一。评测中，我们采取不同的话题关键词抽取策略提交了两轮结果。

第一种策略将话题关键词中的名词作为关键词。例如，“假和尚搂女子”事件中的关键词就是“和尚”和“女子”。

第二种策略将话题关键词扩展到词组、子句的层面：话题符号中的字符串整个是一个关键词；话题中包含的名词词组也是关键词。例如，“假和尚搂女子”事件的话题符号是“#假和尚搂女子#”，那么 “假和尚搂女子”就是一个关键词， “假和尚”和“女子”也是关键词。

从评测的结果来看，两种策略都取得了不错的结果。第二种策略由于抽取到的候选评价对象更加具体，因此在评测中表现更好。

对微博中的每一句，构建分类实例

候选评价对象生成之后，就要判断它们是哪个观点句的正确评价对象。由于评价对象不一定要在当前句中，因此对于一个观点句，需要对其所在微博的所有候选评价对象构建分类实例，以判断该候选评价对象是否是当前句的正确评价对象。

针对每个实例，抽取了以下 5 种特征：

1) 句法成分：即句法分析结果中词组的成分。
2) 候选评价对象离当前句子的距离：根据大纲解释，寻找候选评价对象的时候总是先在当前句中找，如果没有的话再向前面的句子搜索，仍然失败的话才会寻找后面的句子。因此需要一个特征描述候选评价对象与当前句的远近。这个特征定义如下：如果候选评价对象在当前句中，那么距离为0；如果在当前句的前面，那么距离为与当前句间隔的字符数，且符号为负；如果在当前句之后，那么距离为与当前句间隔的字符数，且符号为正。
3) 关键词：该候选评价对象是否是话题关键词。
4) 话题符号：该候选评价对象是否由一对“#”包裹。
5) 候选评价对象频度：观察发现，一个评价对象常在同一话题的多条微博中出现，类似的特点常被用来辅助评价对象的抽取[6,7]。候选评价对象频度这一特征描述了候选评价对象在当前话题下的所有微博中出现的频繁程度，其值等于候选评价对象中所有词出现次数的平均数。

分类

本步仍然使用 Weka 工具包中的 VFI 分类器。对于每一个分类实例，分类器给出一个 0 到 1 之间的实数值作为其得分，得分超过 0.5 就判定候选评价对象是该句的一个正确评价对象。

在实验中，可能会出现对于一个观点句，没有一个候选评价对象得分超过 0.5 的情况，即没有找到正确的评价对象。出现这种情况时取该句所有候选评价对象中得分最高者为正确评价对象。

2.3.2　基于模板的评价对象抽取

为了提高查全率，系统还使用了一个基于模板的方法辅助抽取。评测小组借鉴了史兴等人在COAE 2011 论文集中使用[8]的方法，对微博中所有<依存关系，评价对象句法成分，评价词句法成分>都满足下表的三元组进行抽取。

表 4　基于依存句法的模板

Table 4　Pattern based on dependency parsing

成分	允许的类型
依存关系	rcmod, nsubj, amod, assmod, nn, vmod
评价对象	FW, NR, NT, NN
评价词	VA, JJ, VV

例如“和尚也寂寞”这一句，对其进行句法分析之后得到两组依存关系，分别是：

(1) 和尚（NN）---nsubj-→ 寂寞（VA）；

(2) 也（AD）---advmod-→ 寂寞（VA）；

对比上表可知，第一个依存关系的三元组<nsubj, NN, VA>中各项均出现在了上表中，因此可以抽取出评对象“和尚”；第二个依存关系中，“也”的句法成分为“AD”，不符合要求，因此不能抽取出评价对象。

2.3.3　评价对象的极性判别

对于以上两步抽取出的评价对象，使用如下方式判断其评价极性：

1) 如果评价对象在观点句中，那么搜索一个窗口范围内离评价对象最近的评价词，如果找到的话，使用该词的极性作为评价对象的极性。
2) 如果第一步失败，或者评价对象不在观点句中，那么寻找观点句中的评价词。如果有的话，分别计数评价极性为正和评价极性为负的数目。如果数目相等，判定评价对象极性为 OTHER；如果极性为正的评价词较多，判定评价对象极性为 POS；如果极性为负的评价词较多，判定评价对象极性为 NEG。
3) 如果观点句中也没有评价词，那么将范围扩大到整条微博，计数微博中的正负评价词，判定方法同上。
4) 以上三步都失败的话，直接判定评价对象极性为负。

3　实验结果和分析

评测小组人工标注了评测样例数据集中的“安徽萧县车祸”、“朝鲜发射卫星”和“宝马打死奔驰”三个事件的观点句和情感要素，以此作为训练数据。在这次评测中，本评测小组在参加的两个任务中都取得了比较不错的成绩，也证实了提出的方法的可行性。

表 5　任务 1 评测结果

Table 5　Evaluation result of task 1

指标 / 计算方法	准确率	召回率	F 值
微平均	0.671	0.944	0.784
宏平均	0.674	0.942	0.783

表5展示了任务1的评测结果。与其余参赛的队伍相比，这一结果中的准确率处于中游位置，但由于召回率很高，导致最后的F值排名靠前。可以看出这个结果符合第2节中做出的分析，利用VFI分类器的特点，参评系统的算法在保证一定精度的前提下大幅提升了召回率。

如2.3.1小节所述，针对任务3，评测小组采用不同的话题关键字生成策略提交了两轮。表6a和表6b展示了评测结果，其中编号为33的结果对应前文中的第一种策略，即直接采用话题符号中包含的名词作话题关键字；编号32的结果对应前文中的第二种策略，即使用话题符号中包含的名词词组和话题本身作为话题关键字。

任务3的两轮结果属于各个队伍中的中上游。从结果可以看出，采用更加复杂关键字策略的32轮结果均好于33轮结果，但当使用宽松标准时二者差距不大。这一结果与评测小组之前进行的实验结果相似：在大部分事件中，即使采取简单的话题关键字生成策略仍然能得到比较好的效果；但在少数事件中，大部分微博讨论的内容都是话题本身而非话题中的某一个人或物，此时采取33中关键字生成策略将导致采用严格标准评判时得分大幅度降低。例如“奖状植入广告”这一事件，采取33中关键字生成策略得到话题关键词为“奖状”和“广告”，而采取32中关键字生成策略将得到话题关键词为“奖状”、“广告”和“奖状植入广告”。由于微博中大部分人都直接对这一事件本身表达观点，因此导致结果33使用的系统很难找到完全正

确的评价对象。评测小组自己进行的实验显示，在“奖状植入广告”这个话题上，采取两种策略得到的严格标准下的F值差距超过了0.10。

表 6a　使用严格标准时任务 3 的评测结果

Table 6a　Evaluation result of task 3 with strict standard applied

结果编号	微平均			宏平均		
	准确率	召回率	F 值	准确率	召回率	F 值
32	0.160	0.160	0.160	0.172	0.169	0.170
33	0.112	0.112	0.112	0.121	0.118	0.119

表 6b　使用宽松标准时任务 3 的评测结果

Table 6b　Evaluation result of task 3 with lenient standard applied

结果编号	微平均			宏平均		
	准确率	召回率	F 值	准确率	召回率	F 值
32	0.290	0.220	0.250	0.302	0.229	0.257
33	0.287	0.189	0.228	0.300	0.196	0.234

4　总结

本文总结了上海交通大学中德语言技术联合实验室参加第一届中文微博情感评测所使用的方法和最后的评测结果。在参加的两项任务中，评测小组主要采取了将评测任务转化为分类任务，然后再采用标准分类技术进行解决的策略。其中任务 3 还使用了一个基于模板的方法作为分类方法的补充。评测的结果证明了本文提出方法的可行性。实验过程中，我们发现构建适合网络语言的词典和情感资源将显著改善结果。通过分析任务特点来选用适合的技术而不是盲目使用通用的方法也非常重要。

本次评测也暴露出了评测系统的一些不足。如何在保证高召回率的前提下提升系统的准确率，以及如何提升情感要素抽取方法的有效性是下一步研究的重点。另外，在本次评测中使用到了很多的语义资源，如网络新词表、评价词表、单字评价词表、主观意愿词表等。由于时间所迫，评测小组采用了人工收集的方式，这样做即低效又限制了各种语义资源的规模。能否自动化或半自动化的进行语义资源的收集是微博情感分析系统从实验系统走向实用系统的一个关键点，也是我们以后将要努力的方向。

参考文献

[1] Gülşen Demiröz, H. Altay Güvenir. Classification by Voting Feature Intervals. Machine Learning: ECML-97, 1997:85-92

[2] Bo Pang, Lillian Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval Volume 2, Issue 1-2 (January 2008), 2008: 1-135

[3] Tengfei Ma and Xiaojun Wan. Opinion target extraction in Chinese news comments. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, 2010, pp: 782-790

[4] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. Computational Linguistics Volume 37, Issue 1 (March 2011), 2011: 9-27.

[5] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, and Chun Jin. Red Opal: product-feature scoring from reviews. In Proceedings of the 8th ACM conference on Electronic commerce (EC '07), 2007:182-191.

[6] Minqing Hu, Bing Liu. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, 2004: 168-177

[7] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, 2005:339-346.

[8] 许洪波,孙乐,姚天昉. 第三届中文倾向性分析评测(COAE2011). http://www.ir-china.org.cn/coae2011/第三届中文倾向性分析评测论文集.pdf.