文章编号：2096-1472(2017)-01-49-06

微博多领域情感分析系统研究与实现

张 超，汪龙庆
（东华大学计算机科学与技术学院,上海 201620）

摘 要：微博每天产生的大数据成为市场调研、产品宣传、舆情监控等实际应用高度关注的目标。目前微博要素级的情感分析缺乏领域针对性，人工处理工作量大。本文提出了基于领域自动分拣的情感要素分析模型，通过实验获取有价值的博文特征表示，训练评价对象抽取模型和情感倾向性判别模型。本文实现的MSAS(Microblog Sentiment Analysis System)系统能够自动地完成微博数据预处理、情感要素分析和统计分析功能，为相关的应用提供有价值的分析工具。

关键词：情感要素分析；评价对象抽取；MSAS系统
中图分类号：TP399 **文献标识码**：A

Research and Implementation of the Multi-Domain Sentiment Analysis on Micro-Blog

ZHANG Chao,WANG Longqing
(*College of Computer Science and Technology,Donghua University,Shanghai* 201620,*China*)

Abstract: Big data generated on Micro-blog has caught the attention of various aspects of practical application,such as market survey,product promotion,public opinion monitoring,etc.The current sentiment analysis method is severely lacking in pertinence to specific domains,and the pre-processing work is labor intensive.The paper proposes a feature-level sentiment analysis model based on automated domain data filtering,and conducts experiments to achieve the valuable feature presentation of micro-blog articles and train the opinion target extraction model and the sentimental classification model. The MSAS(Micro-blog Sentiment Analysis System)constructed in this study can automatically implement the pre-processing of micro-blog raw data,the analysis of sentimental factors and the statistics and analysis function,offering valuable analysis tools for relevant application.

Keywords:sentiment factor analysis;opinion target extraction;MSAS

1 引言(Introduction)

微博(Microblog)是当今人们分享与传播信息的重要平台，每天产生的数据仅文本信息就达到上百GB，通过对博文大数据展开情感倾向性分析，可以实现微博营销、品牌宣传、客户关系管理、舆情监控等有价值的应用[1]。微博上每个博主随时随地发布自己的所见所思，内容无限制，其涉及评论对象十分广泛且多变，因此针对微博的情感分析系统需要实现要素级的意见挖掘，即从博主发表的意见中准确地抽取出所针对的评论对象以及对应的情感倾向性。

目前研究主要针对专业网站的评论数据展开分析[2-4]，取得了较好的结果。不同于专业网站的用户评论，微博数据具有数据量大、话题分散、垃圾信息多等特点，很难直接将要素级情感分析的方法用于大规模的日常微博数据进行分析应用，研究[5]更多针对博文的情感极性的判别，而不考虑评价对象的识别；也有研究[6]使用经过人工预处理的数据集设计微博评论目标的分析模型和算法。

本文面向微博日常发布的文本大数据，研究如何快速地从中获取用户关注目标及情感倾向性。舆情监控、微博营销等应用都需要获取不同领域的分析结果，本文基于此提出基于领域数据进行模型训练，应用于多领域数据判别的微博情感分析模式，设计实现了一套完整的系统MSAS(Microblog Sentiment Analysis System)，可以自动从海量博文中过滤出兴趣领域的文本，应用跨领域的分析模型获得大众感兴趣的评价对象，进行相关的情感倾向性统计分析。系统利用专业网站产生的领域相关评论数据作为样本，自动产生领域高频词集，用于领域微博过滤和观点句识别，大量减少了人工预处理工作。不仅可直接用于面向领域的微博用户情感监测，也可通过集成为微博营销等应用提供有价值的数据来源。

2 基于机器学习的情感要素分析(Analysis of emotional factors based on machine learning)

情感要素分析包括对文本中情感对象的抽取以及相应的

情感倾向性判别。本文首先提取博文的有效特征，基于CRFs抽取情感要素，针对获得的情感评价对象再建立特征集，训练分类器进行情感倾向性判别。多个领域之间有许多共通的特性，选择适当的特征，能够帮助我们更好的实现多领域之间的应用，下面介绍MSAS系统所提取的博文特征，以及对应的情感对象抽取和极性判别的建模方法。

2.1 特征提取

(1)Window特征

一段文本中有时会存在多个“评论对象”，情感倾向性也不尽相同，如果每次选取的都是“评论对象”所在的整条语句，那么同一语句中不同的“评论对象”生成的特征集将基本相同，显然无法区分出不同的情感极性。因此需要通过Window的设置，将用于情感倾向性判别的特征限定在一定的词序范围内。

例如情感要素为词，当Window设置为4的时候，所处的语境就变成了。如果前面或后面的词长度不足则仅考虑到句子句首或末尾。如果没有情感目标，则整句都当作情感目标处理，即不再考虑window特征。

(2)Emoticons特征

微博中人们常用符号来表示自己的情感和对一个问题的态度，不同的符号通常对应不同的情感倾向性，表征这些表情符号能够在一定程度上帮助系统预测情感目标的极性。表1是部分常见的表情符号与情感的映射关系。

表1 微博表情符号

Tab.1 Micro-blog expression symbol

表情符号	表情含义
n_n	SMILEYEMOT
T[_]T	CRYEMOT
:-O+	SHOCKEMOT
:-X+	MUTEEMOT
ò_ó	ANGRYEMOT
o3o	KISSEMOT
U_U	SADEMOT

MSAST系统对分词的结果进行识别，如果当前词是表情符号，则将表情符号转化为该表情符号对应的表情含义，如果不是表情符号，则转化为“NULL”，表情含义作为Emoticons特征对应的值，该特征能够很好的提高情感极性判别的准确性。

(3)Word2vec特征

word2vec[7]将文本中的词转换成向量表示，以反映文本语法规则和语义特性。通过向量空间上的相似度，表示文本语义上的相似度。为了获得更好的领域迁移效果，本文将包含不同领域的微博数据一起作为word2vec的输入进行向量化，然后将得到的向量采用K-means进行聚类，最终这些博文中的词被分成100个类别，得到了词和类别的映射关系，词映射得到的类别作为一个特征进行建模。

(4)其他特征及汇总

表2列出了在MSAS系统中用于情感分析的所有博文特征，以及在文中的标识。

表2 特征标识符及其含义

Tab.2 Feature and its implication

标识符	特征	特征含义
F1	Ngram	N元词组
F2	RST	依赖关系
F3	Window	窗口
F4	Emoticons	表情符号
F5	Pos-tag	词性标注
F6	SentenceLength	语句长度
F7	NER	命名实体
F8	Word2Vec	Word2Vec聚类
F9	Clark	Clark聚簇

2.2 基于CRFSuite的情感要素抽取模型

条件随机场(CRFs)模型是由Lafferty[8]在2001年提出的一种典型的判别式模型。它在观测序列的基础上对目标序列进行建模，重点解决序列化标注的问题。条件随机场模型既具有判别式模型的优点，又像产生式模型那样考虑了上下文标记间的转移概率，以序列化形式进行全局参数优化和解码，解决了其他判别式模型(如最大熵马尔科夫模型)难以避免的标记偏置问题。

本文采用CRFSuite[9]来建立情感抽取模型。该实现不关心标签和属性的命名方式，也不关心其含义，只是将它们视为单纯的字符串。CRFSuite学习特征和特征值之间的关联权重(要素权重)。MSAS系统给出每条博文分词结果的特征值集合，其中，这里的是上文列表2给出的特征，是人工标注得到的当前分词是否是情感要素，此时的还不能直接拿来进CRFsuite训练，还得将特征和特征值进行关联，最终我们的输入是，输出即为情感要素抽取模型。

例2：博文“这种耳机是有线的，可能是目前earpods的改进版。”它的特征集合如表3所示。该微博中的评价对象是“耳机”，将它的词语、词性、命名实体、分类、根节点、句法等特征和对应的标签关联后，形成了“N word[-2]=是word[-1]=目前word[0]=earpods……”训练样本序列，其中“N”代表我们模型的分类果，“word[-2]”表示特征名称，“是”代表该特征对应的值，输入CRFsuite中进行训练生成相应的评价对象抽取模型。

表3 特征抽取示例

Tab.3 Feature extraction example

词语	词性	命名实体	分类	根节点	句法	情感目标
耳机	名词	否	76	否	主语	是

2.3 基于SVM的情感极性判别模型

SVM是Cortes和Vapnik[10]1995年首先提出的，它在解决小样本非线性及高维模式识别中表现出许多特有的优势，其分类的查全率和查准率几乎超过了现有的所有方法，具有很好的泛化能力及其他机器学习方法不可比拟的优势。

情感极性判别其实是一个多分类的问题，MSAST系统通过SVM来构建情感极性的分类器。SVM的输入是一系列具有相同维度的向量，对与MTSAT系统而言，得到整个微博语料的特征集合，并将所有情感要素对应特征集合的值进行向量化即可，，其中是表2中的特征对应的值，是人工标注得到的该评价对象的情感极性，SVM训练输出是情感极性判别模型。

3 MSAST训练系统(MSAST system)

微博的数据限制用户发布的文不得超过140个中文字符，通常包含多个句子，且每个句子涉及多个和可能表达不同的情感可能不同，评论相对自由，与基于商品评论有较大区别。本文经过研究和大量数据实验发现，使用NLPCC提供的已标注的三个领域数据集(共658条数据)进行抽取模型的训练，应用于微博数据情感分析时，由于数据量较小，且语言风格与微博数据不同，导致情感要素抽取和情感倾向性分析效果都不佳，在不同领域之间应用时，效果特别差。因此本文考虑直接采用微博数据进行模型的训练。

MSAST系统(如图1所示)主要包括数据预处理模块、人工标注模块、模型训练模块这三个模块。大致分为几个步骤：

(1)对微博数据中的垃圾数据进行过滤，并建立索引，通过筛选和分类得到相关领域的微博原数据。

(2)对筛选得到原数据进行人工标注，标注每条微博的评价对象及其对应的情感倾向。

(3)用标注好的微博数据来训练抽取模型和极性判别模型。

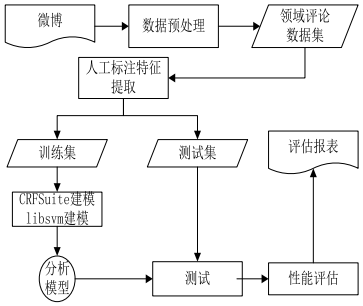

图1 MSAST系统流程图

Fig.1 MSAST system flow chart

3.1 数据预处理模块

由于微博的数据量很大，内容涉及面广，垃圾数据相对较多。如果从微博中自动过滤出与用户兴趣领域相关的评论数据是系统重点解决的问题。MSAST利用专业领域网站评论针对性强的特点，从这些网站中获取数据作为带有领域标记的数据。利用词的IF-IDF特征筛选出领域高频词并自动训练出相应的领域分类器。

(1)去除转发消息和极短的句子，由于转发的内容是别人的博文，不代表博主本身的意见。另一方面大部分是转发为广告，如“关注常州网微信，转发此微博即有机会获得五星级酒店双人自助餐券、viva西班牙餐厅50元现金券！”因此对情感分析的意义不大。

(2)lucene是一个全文检索引擎的架构[11]，提供了完整的查询引擎和索引引擎，部分文本分析引擎。系统对清洗后的微博数据使用ansj分词工具进行中文分词在建立lucene索引，索引的域包括微博ID、用户ID、微博内容，其中微博ID和用户ID不进行分词。

(3)本文选取了三个领域酒店、手机、电脑微博数据作为标注训练集。为了获取领域高频相关词，自动爬取了携程的酒店评论数据和京东的手机和电脑评论数据，基于词的TF-IDF特征筛选兴趣领域的种子词集。

将酒店、手机、电脑、其他类数据(任选6万条微博数据)作为输入，本文使用fudannlp提供的文本分类工具，通过选取经过互信息刷选的BI-Gram和Tri-Gram作为特征，领域名称作为分类结果获得领域分类模型。

(4)利用兴趣领域的种子词对步骤2中得到的微博数据集的lucene索引进行筛选，得到三个领域每个领域各1.5万条数据，再将这些数据用步骤3建立好的分类模型进行检验，如果得到的分类结果和之前获取来源对应的分类不一致，则舍弃该数据，筛选得到的数据作为下一个模块的输入。

3.2 人工标注模块

人工标注模块就是通过人工的方式来判别一条微博数据的情感目标和该情感目标的极性。该模块能够方便的选择情感目标和极性，并将人工标注的结果写入文件。

人工标注模块的实现如图2所示，将领域微博集作为输入，标注系统自动断句，然后对每一句进行中文分词，用户使用图2所示界面，选择评价目标和对应的极性。如果没有评价目标，记为“NULL”，如果评价目标由多个词组成，标记目标开始词、中间词和结束词。MSAST训练系统仅包含3种极性正(positive)、负(negtive)、中立(netural)。

图2 标注训练系统

Fig.2 Label training system

3.3 模型训练模块

利用人工标注模块人工标注出来的微博数据作为训练集，评价目标抽取我们用的是CRFSuite训练模型，主要参考了NLANGP[12]系统构建特征集，而极性判别阶段是SVM训练模型。系统实现的重点是进行特征的提取和构造特征集合fSet，如算法1所示。

算法1:特征选取算法

输入:courps//已经进行过分词、词性标注等处理的微博语料

fQueue//特征处理队列

window//窗口大小

输出:fSet//特征集合

BEGIN

01 FOR each opinion∈courps.getOpinionsDO

02 courp=courps.getCourp(opinion.id)

03 IF (opinion.hasTarget) THEN//观点中含有情感目标

04start=opinion.from//起始位置

05 end=opinion.to//结束位置

06IF (window>0 && end>0) THEN

07from=start window

08IF (from<0) THEN

09from=0

10 END IF

11 to=end+window

12 IF (to>courp.length-1) THEN

13 to=courp.length-1

14 END IF

15 courp=courp.sublist(courp, start, end)//获取在窗口中的语料

16END IF

17 WHILE(!fQueue.isEmpty())

18 fProcess=fQueue.deQueue()//获取特征处器

19 features=fProcess(courp)

20 fSet.add(features)//加入特征集合

21 END WHILE

22 END FOR

23 RETURNfSet

END

上述算法中01—02步通过情感要素ID选取对应的语料；03—16步首先判断是否存在评价对象，然后判定评价对象窗口起始和结束位置，提取在窗口中的语料相关特征，如果没有评价对象则窗口为整个句子；17—23步遍历特征处理器队列，对语料进行处理，将得到的特征依次加入到特征集合中。

3.4 MSAST系统性能分析

(1)实验数据

本实验首先爬取了携程的酒店评论数据和京东的手机和电脑评论数据各2万条，并任意选取6万条微博数据作为其他领域数据，用FudanNLP[13]训练得到领域过滤器。

本实验获取了新浪微博某一天的评论数据约30.1GB，去除转发消息和极短的句子后剩下15.7GB的数据，对这15.7GB的数据建立索引，使用手机、电脑、酒店的种子词集进行筛选，然后使用领域分类器分类后，自动获得手机数据9754条，电脑数据10730条，酒店数据6020条。本实验对三个领域任意选取2000条进行人工标注，最终得到的训练数据集分布如表4所示。

表4 训练数据集分布

Tab.4 Training data

领域	positive	negtive	netural	total
手机	689	551	760	2000
电脑	624	423	953	2000
酒店	462	365	1173	2000

(2)评价指标

评价微博情感分析系统的指标主要有准确率、召回率和F值。准确率计算公式为：

$$Precision = M/N \quad (1)$$

其中，M指的是极性判断正确的微博数据条目数，N则表示总共的微博数据集的条目数。准确率越高则表示系统进行极性判别时越准确。

召回率计算公式为：

$$Recall = M/R \quad (2)$$

其中，R表示所有带有极性的情感目标的数量，表示了系统能够发现情感极性的能力。

然而在大多数情况下，情感分析系统的准确率越高，系统的召回率却很低，类似的召回率越高，准确率也必定会有所降低，于是为了综合考虑准确率和召回率，引入F值公式：

$$F = (2 * Precisio * Recall)/(Precision + Recall) \quad (3)$$

(3)结果分析

本实验用已经标注好的手机领域的数据集作为训练集，使用交叉验证法对情感要素抽取模型进行了评估，得到了如表5所示的数据表，本实验将仅使用Tri-gram特征作为实验的基准线，得到的58.8%的准确率，增加词性特征(Pos)后准确率提高到了62.1%，不断增加word2vec、Dependency、Lexicons、PoS特征后，准确率提升到了69.7%。

表5 不同特征组合对情感要素抽取准确率的影响

Tab.5 The influence of different feature combinations on the extraction accuracy of affective factors

特征	Precision
Baseline (3gram)	58.8
3gram+Pos	62.1
3gram+Lexicons	64.9
3gram+word2vec+Dependency+Lexicons+PoS	69.7

为了说明不同特征对跨领域情感极性分析效果的影响，本实验使用已经标注好的手机领域的数据集作为训练集，通过在系统中不断增加新特征，使用MSAST系统来训练不同的情感极性判别模型，分别用如表4所示的手机、电脑、酒店三个领域已经标注的数据集进行测试，得到表6的实验结果。

表6特征选取对跨领域情感分析的影响

Tab.6 The influence of feature selection on cross domain affective analysis

特征	训练集：手机 测试集：手机			训练集：手机 测试集：电脑			训练集：手机 测试集：酒店		
	Precision	Recall	F-value	Precision	Recall	F-value	Precision	Recall	F-value
ngram	0.517	0.552	0.533	0.502	0.411	0.452	0.489	0.406	0.444
+sentence Length	0.524	0.559	0.526	0.513	0.457	0.483	0.487	0.431	0.457
+pos-tag	0.552	0.579	0.565	0.516	0.501	0.508	0.491	0.447	0.468
+word2vec	0.551	0.582	0.566	0.525	0.529	0.526	0.493	0.449	0.470
+Emoticons	0.582	0.594	0.588	0.531	0.537	0.534	0.515	0.502	0.508
+RSV	0.612	0.603	0.607	0.540	0.551	0.545	0.519	0.517	0.518
+Window	0.659	0.624	0.641	0.537	0.572	0.554	0.521	0.518	0.524

通过表6可以看出，当增加Emoticons、RSV、Window特征后，准确率、召回率和F值都有了明显的提升。增加新特征对当前领域的情感极性判别影响比较明显，因为选取的特征越多，特征的差异也会随之增加。但的F值并不是很高，因为微博数据本身比较自由和散乱的，导致召回率很难提升。

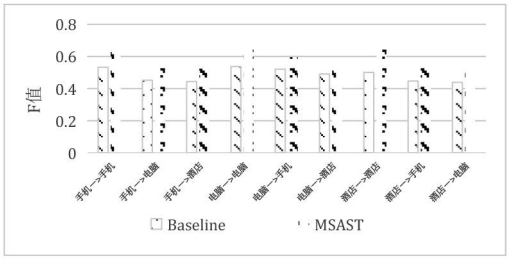

图3 跨领域间情感分析对比实验

Fig.3 Comparative analysis of cross domain affective analysis

图3中给出的是3个领域之间，相互作为训练集和测试集得到的F值，图3中可以看到，由于手机和电脑领域的相关度比较大，两两之间的测试结果要比差异性较大的领域酒店效果好，也说明了相似领域之间的特征分布比较近似。

4 MSASA分析应用系统(MSASA system)

MSASA应用系统是应用MSAST训练系统得到的情感要素抽取模型和情感极性判别模型，自动从微博中筛选出用户感兴趣的领域微博数据，进行情感要素分析，通过进一步的挖掘，即可了解这些领域中人们感兴趣的话题及其分布，了解人们对这些话题的态度。MSASA应用系统主要包括数据预处理模块，情感要素抽取模块，情感极性判别模块，统计分析模块。图4给出了MSASA系统的详细的框架设计。

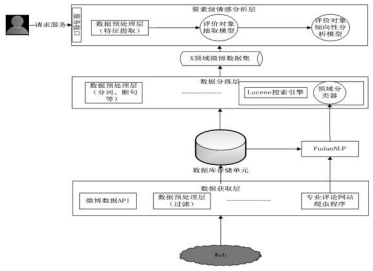

图4 MSASA系统框架

Fig.4 MSASA system flow chart

MSASA系统的实验大致分为几个步骤：

(1)对已经建好索引的15.7GB的数据，通过电脑领域的种子词集选取了20万的初始微博数据集。获得高度相关的5万条该领域的微博评论集。这里选择采用了手机领域训练出来的情感目标抽取模型和情感极性判别模型。

(2)经过情感要素抽取，这5万条数据共抽取到2069个情感目标，经过统计得到513个不同的评价对象，去除其中出现频率小于5的评价对象，得到了33个常见评价对象，剩余的被归为其他，其关注程度如图5所示。可以看出屏幕、版本、电池、密码、摄像头被提及的比较多，当然也发现有一些与电脑领域无关的情感对象，比如“眼睛”和“洗澡”。

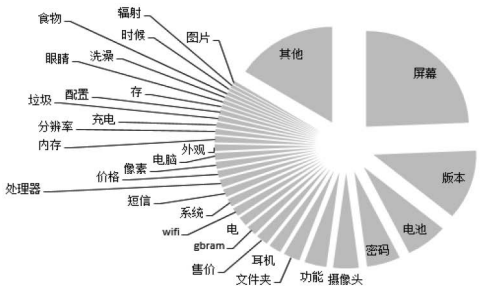

图5 电脑微博数据热门评价目标分布

Fig.5 Micro-blog computer field popular evaluation target distribution

(3)经过情感极性判别和统计，得到了这513个评价对象的情感极性分布(如图5所示，篇幅所限，仅列出部分数据)。图6清晰地反应了用户对于电脑一些评价对象的态度，可以帮助决策人掌握用户的需求和心理。

图6给出了评价人数较多的前17个评价对象的情感极性统计，尽管大家的观点存在差异，但在大数据分析的基础上，可以看到如“耳机”反对的人明显比支持的多，说明需要提升耳机的质量，如“屏幕”支持的人就比较多，说明目前大家对市场上的电脑屏幕满意度还不错。

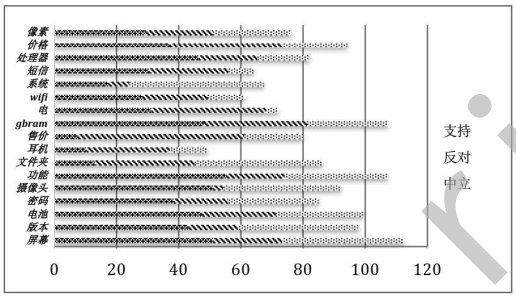

图6 电脑微博数据情感极性统计(部分)

Fig.6 Computer micro-blog field sentiment polarity statistics (part)

5 结论(Conclusion)

本文提出了多领域微博要素级情感分析系统解决方案，并用实验证明了该方案的可行性。通过对系统性能的评估，分析了博文特征对情感要素分析的影响，通过具体的应用实验给出了完整的情感分析应用实例，展示了系统的应用价值。未来将考虑使用领域语义模型进一步归并评价目标，降低领域之间的迁移损失，使分析结果更便于第三方系统应用。

参考文献(References)

[1] Hanhua Chen,Hai Jin,Xiaolong Cui.微博系统中一种混合关注对象推荐方法[J].Science China Information Sciences,2017,60(1):012102.

[2] Agarwal B,et al.Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing:A Novel Approach[J].Cognitive Computation,2015,7(4):487-499.

[3] Cambria E.Affective Computing and Sentiment Analysis[J].IEEE Intelligent Systems,2016,31(2):102-107.

[4] Liu B.Sentiment Analysis:Mining Opinions,Sentiments,and Emotions[J].Computational Linguistics,2015(3):1-4.

[5] 孙建旺,吕学强,张雷瀚.基于词典与机器学习的中文微博情感分析研究[J].计算机应用与软件,2014,31(7):177-181.

[6] 丁晟春,李霄.基于CRFs和领域本体的中文微博评价对象抽取研究.第六届中文倾向性分析评测(COAE2014)报告论文集,2014:131-136.

[7] Goldberg Y,Levy O.Word2vec Explained:Deriving Mikolov et al.Negative-Sampling Word-Embedding Method[J].EprintArxiv,2014.

[8] Lafferty J D,Mccallum A,Pereira F C N.Conditional Random Fields:Probabilistic Models for Segmenting And Labeling Sequence Data[C].2001:282-289.

[9] Tang B,et al.A Comparison of Conditional Random fields and Structured Support Vector Machines for Chemical Entity Recognition in Biomedical Literature[J].Journal of Cheminformatics,2015,7.

[10] Cortes C,Vapnik V.Support-Vector Networks[J].Machine Learning,1995,20(3):273-297.

[11] 郎小伟,王申康.基于Lucene的全文检索系统研究与开发[J].计算机工程,2006,32(4):94-96.

[12] Toh Z,Su J.NLANGP:Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction[C].International Workshop on Semantic Evaluation,2015.

[13] Qiu X,Zhang Q,Huang X.FudanNLP:A Toolkit for Chinese Natural Language Processing[C].Meeting of the Association for Computational Linguistics:System Demonstrations,2013:49-54.

作者简介：

张 超(1990-)，男，硕士，助教.研究领域：自然语言处理，机器学习.

汪龙庆(1988-)，男，硕士，助教.研究领域：自然语言处理，大数据.