厦门大学 NLP&CC 2012 微博情感分析评测报告

罗凌 吴昌 陈毅东* 史晓东 曹茂元

厦门大学信息科学与技术学院智能科学与技术系 厦门 361005

* 通讯作者，E-mail: ydchen@xmu.edu.cn

摘要 本文描述了厦门大学智能科学与技术系参加 NLP&CC 2012 中文微博情感分析评测的系统。系统参加了本评测中任务一观点句识别和任务二情感倾向性分析，文章分别对系统在这两个任务中所使用的技术做了介绍，并详细描述了参与评测中的数据情况和结果。

关键字 观点句识别；情感倾向性分析；机器学习；规则

一、 引言

本文将对厦门大学智能科学与技术系参加 NLP&CC 2012 中文微博情感分析评测的系统进行描述。本次评测由中国计算机协会主办，共包含 3 个子任务，分别是：（1）观点句识别（2）情感倾向性分析（3）情感要素抽取。我们参加了任务一和任务二的评测。

文章的第二节将介绍我们系统参加任务一观点句识别中所使用的技术及相关实验结果；第三节将介绍我们系统参加任务二情感倾向性分析中所使用的技术及相关实验结果；第四节是总结。

二、 任务一：观点句识别

本节介绍此次评测中我们参加任务一的系统情况，首先概述现有的主要观点句识别模型，然后详细介绍我们所采用的规则与机器学习相结合的观点句识别方法，最后是实验和讨论。

2.1 概述

观点句（也称为主观句），即在表达的过程中带有某种情感和观点的句子，这种观点可以是作者本人的、引用于他人的、或是某群体、组织发表的。国外对观点句的研究起步较早，较有代表性的工作有：Wiebe（2001）[1]选择某些词类（代词、形容词、序数词、情态动词和副词）、标点和句子位置作为特征，实现对观点句识别。Riloff（2003）[2]等人利用 boot-strapping 算法学习得到主观性名词，单独使用主观性名词为特征，采用朴素贝叶斯分类器对主观句识别。Wiebe 和 Riloff（2003）[3]他们依靠先前研究中确定的主观特征，分别建立了主观分类器和客观分类器，自动从未标注的文本中获得大量主观句和客观句，再从这些句子中得到更多主观性词语搭配，再用准确性很高的词语搭配更新原始的主观特征。Yu 和 Hatzivassiloglou（2003）[4]利用相似性方法、朴素贝叶斯分类和多重朴素贝叶斯分类等三种统计方法进行观点句识别研究。

国内较早开始该工作的是姚天昉，彭思威（2007）[5]使用了机器学习的方法进行分类识别。叶强等（2007）[6]提出了一种根据连续双词词类组合模式(2-POS)自动判断句子主观性程度的方法。王根，赵军（2007）[7]提出了一种基于多重冗余标记的 CRFs。蒙新泛，王厚峰（2009）[8]通过对比试验，分析了上下文信息对于主客观分类的影响。经过早期的研究，近几年系统融合和多分类器结合的研究热潮，张博（2011）[9]进行了模块串行的方法。宋乐等人[10]在 2009 年的第二届 COAE 评测中文观点句抽取的任务中使用了一种类似最小图个的方法。在 2011 年第三届 COAE 评测中，徐瑞峰等人[11]提出一种基于图的句子排序算法 SentenceRank。

NLP&CC 2012 中文微博情感分析评测任务旨在从 20 个领域的数据集中自动识别观点句，本评测定义的观点句只限定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。我们采用了规则与机器学习相结合的方法进行中文观点句的识别。下面对其进行介绍。

2.2 规则与机器学习相结合的观点句识别方法

2.2.1 方法概述

此次评测中，我们的系统采用了一个多系统融合的观点句分类框架，在此框架下，该系统结合了规则方法和机器学习方法。其训练和分类流程如图 1 所示：

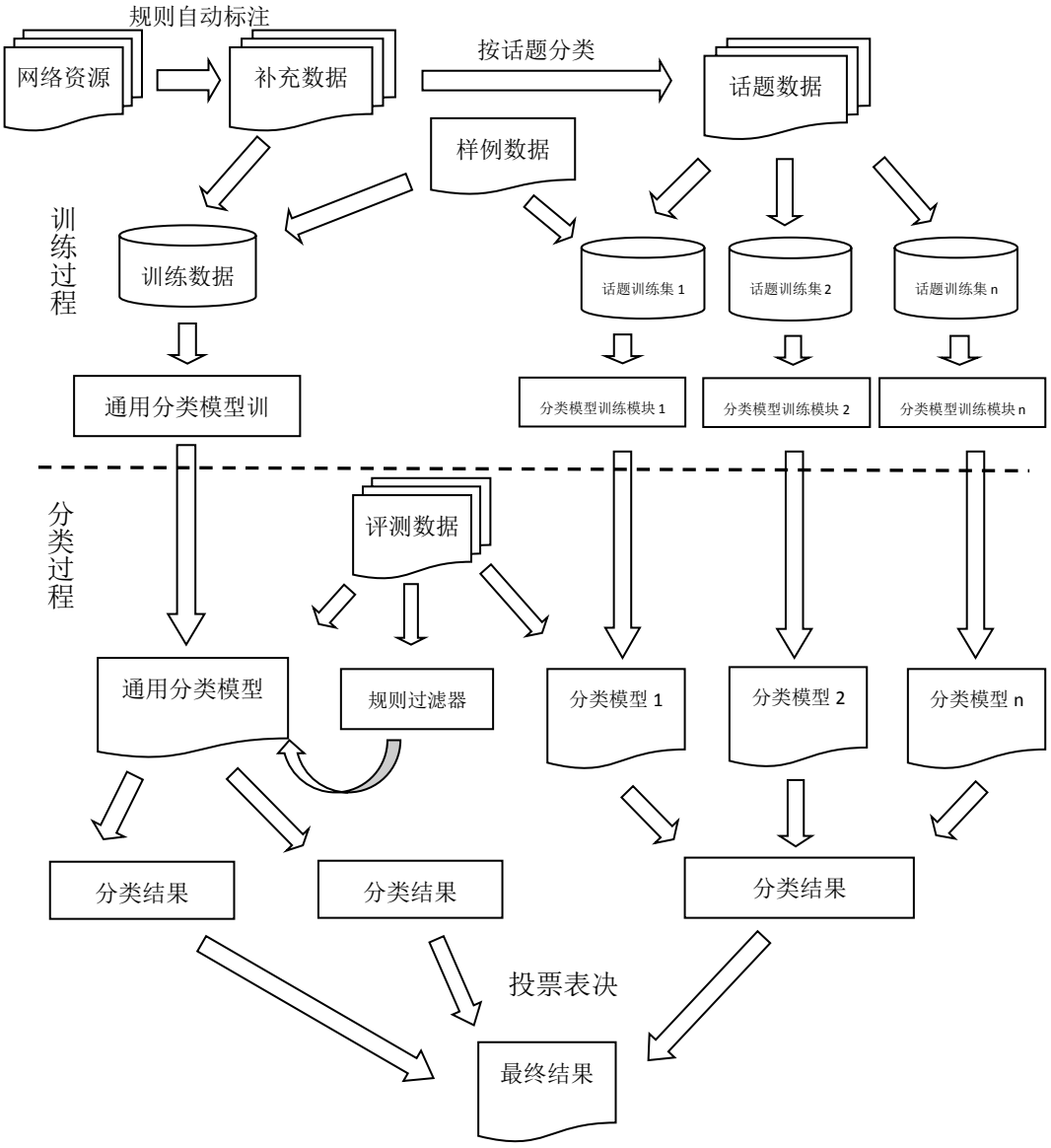

图 1　规则与机器学习相结合的观点句识别方法流程图

可以看到，系统的主体采用了机器学习的方法，但在两个地方应用了规则方法：一处是利用规则系统对从网络中自动挖掘的微博数据进行分类以扩充训练集；另一处则是将规则系统作为机器学习分类器的预处理模块。此外，考虑到不同话题的数据在分类特征方面可能存在差异，除了通用的分类模型外，本系统还引入了分话题模型，这些分话题模型仅由相关话题的微博数据训练所得。

2.2.2 基于规则的观点句识别

如前所述，在本系统中，规则方法在训练过程和分类过程中均得了应用。但是，针对不同数据规模，两个规则系统所采用的规则粒度也有差异。

训练过程中使用的规则：

观点句规则：

- …+（代词|人名|地名|专有名词）+….+是+名词+….
- …+（代词|人名|地名|专有名词）+….+副词+形容词+….

非观点句规则：

- 仅包含表情符号。
- 仅包含网址。
- 不满足观点句规则且以动词开头的短句。

以上的规则我们用于在训练过程中对从网络中挖掘的微博数据进行自动标注，凡是满足上面规则的句子我们将其抽取出来进行标注，作为训练语料的补充。

分类过程中使用的规则：

- 非法词、乱码过滤器：非法词、乱码超过 6 个判定为非观点句。
- 链接过滤器：含有链接，无实际信息的判定为非观点句。
- 单句过滤器：在单句中不含网络新闻，且不是反问句式判定为非观点句。
- 单一特征过滤器：只有 hashtag，表情符合，标点符号的句子判定为非观点句。
- 主观愿望过滤器：只包含愿望词判定为非观点句。
- 词典过滤器：包含情感词、网络新词、成语、否定词等信息判定为观点句。

在分类过程中，我们通过上面 6 个过滤器，将评测数据进行了过滤标注预处理，其中不能判定的句子再通过基于机器学习方法进行识别。

2.2.3 基于机器学习方法的观点句识别

我们在进行分类训练时采用了通用和分话题两种训练方法。通用模型是通过将所有话题的训练数据全部拿给分类器训练出一个通用模型；话题模型是通过该话题相关的训练数据给分类器分别训练出 20 个话题模型。这两种方法我们都采用了在姚天防[5]，张博[9]使用的特征基础上，加入了主题相关的人名特征作为分类特征：

1. 情感词，我们整合了知网和清华的情感词典，总共约 16000 个词。
2. 指示性动词，我们使用了张博论文[9]中的指示性动词表和根据数据集自己添加的一些动词，总共约 100 个词。
3. 人称代词。
4. 叹词和语气词。
5. 副词。
6. 主题相关的人名。
7. 标点符号。
8. N-POS，N-POS 是指语句中 N 个连续词性的顺序组合，系统中我们采用了 1-pos 和 2-pos。

我们使用了以下 5 种分类器进行了结果分类：（1）标准概率朴素贝叶斯分类算法（Naïve Bayes）（2）支持向量机分类算法（SVM）（3）用于支持向量分类的连续最小优化算法（SMO）（4）随机森林算法（Random Forest）（5）分类与回归树算法（Classification Via Regression）。最终的分类结果是将所有模型的分类结果进行一个投票表决所产生的。

2.3 实验结果及讨论

2.3.1 训练语料

由于本次评测未供训练语料，只有少量的样例数据，所以我们从网络上爬取了与评测数据相关主题的微博，并使用了一些规则自动标注了部分数据，进行训练语料的补充。

1. 样例数据：评测主办方提供的样例数据。共包含已标注毁容案话题约 240 句和 Ipad 话题约 220 句。
2. 补充数据：从腾讯微博上爬取的与评测数据相关主题的微博。共包含菲军舰恶意撞击、疯狂的大葱等 20 个话题，每个话题约 2000 句。接着使用基于规则的方法对其进行了部分自动标注，标注后每个话题约 600 句。

这些标注语料将用于分类器训练。本文使用的是 weka[1]平台进行训练和分类

2.3.2 实验结果

本系统在 NLP&CC 2012 中文微博情感分析评测的观点句识别任务中取得的评测结果如表 1 所示。微平均以整个数据集为一个评价单元，计算整体的评价指标；宏平均以每个话题为一个评价单元，计算参评系统在该话题中的评价指标，最后计算所有话题上各指标的平均值。

表 1　系统任务一评测结果

微平均			宏平均		
正确率	召回率	F 值	正确率	召回率	F 值
0.74	0.646	0.690	0.744	0.639	0.680

上表是将多分类器表决后的结果，在主办方发布评测答案后，我们对每个分类器分别进行了评测，下面用标号来表示各个分类器：（1）标准概率朴素贝叶斯分类算法（NB）（2）支持向量机分类算法（SVM）（3）用于支持向量分类的连续最小优化算法（SMO）（4）随机森林算法（RF）（5）分类与回归树算法（CVR）（6）使用规则过滤器后再使用 SVM 分类算法（ALL-Rule-SVM）。通用分类器我们在简写前加 ALL-，不加 ALL 的表示话题分类器，实验结果如表 2，以下的实验结果我们都是采用微平均计算：

表 2　通用模型和话题模型对比结果

标号	正确率	召回率	F 值
ALL-NB	0.742	0.376	0.499
NB	0.744	0.432	0.547
ALL-SVM	0.735	0.675	0.704
SVM	0.735	0.682	0.708
ALL-SMO	0.737	0.609	0.667
SMO	0.747	0.623	0.679
ALL-RF	0.727	0.657	0.690
RF	0.728	0.684	0.705
ALL-CVR	0.720	0.657	0.687
CVR	0.725	0.720	0.722
ALL-Rule-SVM	0.733	0.683	0.707

从表 2 结果可以看出在所有分类算法中朴素贝叶斯的分类结果较为糟糕，对系统融合的影响比较大，此外，由于训练数据规模小，我们在进行系统时是采用简单的投票，并没有采用权重，因而也是这次结果不理想的原因。从表 2 对比同样分类器中通用模型和话题模型的分类结果可知，分话题进行训练得到的分类结果都比通用模型的分类结果要好，说明领域间存在着差异。

在进行分类器训练时，由于提供的训练语料过少，我们通过上面提到的规则自动标注了从网络中挖掘的微博数据，并将这部分数据作为扩充语料进行对分类器的训练。为了证明我们加入这些自动语料对分类器训练是有帮助的，我们做了下面的实验。我们用起初标注好的样例集中约 200 句毁容案话题数据作为训练，和加上了自动标注的扩充数据约 600 句作为训练，对同样的约 200 句 Ipad 话题数据测试，得到了各个分类器的对比结果，见表 3：

表 3　加入扩充数据后对比结果

标号	正确率	召回率	F 值
NB	0.645	0.396	0.491

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

NB+Extra	0.578	0.515	0.545
SVM	0.560	0.782	0.653
SVM+Extra	0.575	0.762	0.655
SMO	0.578	0.515	0.545
SMO+Extra	0.583	0.733	0.649
CVR	0.538	0.624	0.578
CVR+Extra	0.570	0.802	0.667
RF	0.560	0.644	0.599
RF+Extra	0.549	0.782	0.645

从表 3 的结果我们可以看出加入了自动标注的扩充数据进行训练后，基本每个分类器都有或多或少的提升，这表明我们加入的这部分自动标注数据是对分类有所帮助的。

三、 任务二：情感倾向性判断

3.1 概述

近年来，文本情感分析吸引了国内外众多研究者的目光。从研究方法来看，大致可以分为两类方法：一、基于词汇的文本情感分析；二、基于语料的文本情感分析。

基于词汇的文本情感分析是指根据情感词汇、短语的情感度中计算出文本的情感倾向性，其研究中心是已标注好的情感规则词典。2002 年，Turney[12]等人的语义倾向方法通过语料中含有形容词、动词和短语的情感倾向平均值来预测文本的情感倾向性。Kim[13]等人建立三个模型，连接情感承载词的独立情感，对句子的进行情感分类。

基于语料的文本情感分析是指利用已标注好的语料，通过机器学习等方法建立分类器进行情感分类，其研究中心是语料。2002 年，Pang[14]等人利用语言信息，分别使用最大熵分别使用最大嫡 （Maximum Entropy， ME）、朴素贝叶斯 （Naïve Bayes， NB）和支持向量机（Support Vector Machine， SVM）三种分类方法对英文电影评论进行情感分类研究。2007 年 Tan[15]等人利用图书、旅馆和笔记本电脑三种中文评论语料，分别使用互信息（MI），信息增益（IG），CHI 和文档频率（DF）三种特征值，对比了 Centroid Classifier、KNN、NB、Winnow、SVM 五种分类器效果，得出 IG 和 SVM 分类器效果比其它组合的精度更高的结论。

在本次评测任务中，任务二要求判断任务一中识别的观点句的情感倾向。观点句的情感倾向分为正面（POS），负面（NEG）和其他（OTHER）。我们结合基于词汇、语料的方法，参考了谢丽星[16]的工作，采用分两个步骤的多策略方法进行情感倾向性判断。第一步采用规则方法进行判断，第二步利用 SVM 的方法进行判断。

3.2 外部资源

在规则方法中，词汇是研究的中心，在我们的工作中，使用了几个词典，分别是：情感词典、网络新词词典、成语词典、叠词词典、表情词典、愿望词典和否定词典。其中情感词典和表情词典已经标注极性。词典详细情况如下表：

表 4：词典表

词典名称	**说明**
情感词典	词典采用清华大学的情感词典,其中正向情感词 5567 个,负向词 4468 个
网络新词词典	词典来源于搜狐输入法词库,共 21401 个
成语词典	词典采用搜狐输入法词库,共 48967 个

叠词词典	词典采用搜狐输入法词库,共 525 个
表情词典	词典采用 QQ 表情词典，共 106 个,并对极性进行人工标注
愿望词典	自定义词典,共 8 个
否定词典	自定义词典,共 23 个

3.3 方法简介

我们参考了谢丽星[16]的工作，采用分两个步骤进行情感倾向性判断。第一步采用规则方法进行判断，第二步利用 SVM 的方法进行判断。

在第一步规则方法中，使用了三个规则方法，利用词典信息进行情感计算，通过正、负向情感数值的差来判别主观句的情感倾向。通过三个规则判断后，对标注为 OTHER 的句子，再利用 SVM 方法进行第二步分析。

规则如下表：

表 5 任务 2 规则表

序号	*规则方法*	**说明**
1	情感计算方法	在没有否定词的情况下，通过计算正向情感词个数与负向情感词之差来判定，大于 0 时为正向，小于 0 时为负向，等于 0 时为 OTHER。
2	基于标点符号否定处理	句子存在否定词时，如果否定词与情感词同在两个标点之间，则认为否定词否定了该情感词，原情感词情感反置，再通过序号 1 的规则情感计算方法进行计算情感。
3	表情符号情感计算方法	句子中存在表情符号时，利用人工标注极性的 qq 表情，通过计算正向表情个数与负向表情之差来判定，大于 0 时为正向，小于 0 时为负向，等于 0 时为 OTHER。

通过第一步规则的方法后，大概有 60%的微博句子标注了极性，剩下 40%无法通过规则的方法确定地标注极性。

在第二步方法中，我们再通过对“IPAD”、“毁容案”两个样例数据进行训练，得到的模型用于评测集数据的情感倾向性判断。在选取特征方面，我们主要选用了字典信息、句式信息和话题信息。特征选取如下：

1. 正负情感词计数
2. 是否仅有正向情感词或负向情感词?
3. 网络新词个数
4. 叠词个数
5. 成语个数
6. 否定词个数
7. 是否有问号？
8. 话题编号

3.4 实验结果及分析

经过两步的分析后，我们通过对任务 1 中识别的观点句进行倾向性判断，实验结果如下表：

表 6：任务 2 实验结果表

微平均			宏平均		
正确率	召回率	F 值	正确率	召回率	F 值
0.725	0.495	0.588	0.725	0.490	0.583

跟其它队伍比，我们的实验结果处于中等，跟最好的队伍比，我们还有很大的差距，并没有达到预期目标。究其原因主要有两点：一是过度依赖词典信息，规则方法过于严格。这样造成召回率不够高，影响整体性能比较严重。二是由于语料不够充分，训练的模型对情感判断不够灵敏。由于微博句子都比较短，因此，基于小样本的语料的方法我们还需进一步努力。

四、 总结

本文描述了厦门大学智能科学与技术系参加 NLP&CC 2012 中文微博情感分析评测的系统，在任务一观点句识别中使用了规则的方法对从网络上挖掘的数据进行了自动标注，通过实验证明，我们加入的这部分自动标注数据是对训练分类模型有帮助的，这解决了在机器学习训练过程中语料不足的问题。

本次评测分了 20 个话题，我们针对话题进行了分话题模型的训练，5 种分类算法结果都表明分话题模型比通用模型分类的结果要理想，这说明了话题间的分类特征是存在差异的。

该系统在参加此次评测的所有系统中总体处于中上水平，从实验看出，在任务一中朴素贝叶斯的分类结果较为糟糕，对系统融合的影响比较大，此外，我们在进行系统时是采用简单的投票，并没有采用权重。在任务二中，由于过度依赖词典信息，规则方法过于严格。这样造成召回率不够高，影响整体性能比较严重。还有由于语料不够充分，训练的模型对情感判断不够灵敏。这些都是这次评测结果不理想的原因

本次评测的与以往评测较为不同的是，处理的数据是来自于微博，而不是规范化的文本。微博的最大特点是简短，不规范，里面不仅包含了大量的网络术语，表情，还有很多错别字，病句，这对我们进行分词，提取特征都有很大的影响。如今，由于网络的迅速发展，微博等形式的网络数据大量出现，我们能否从中挖掘到有效信息，如何处理这类形式的数据，都需要我们更深入的研究。

参考文献

[1] Janyce Wiebe, Rebecca Bruce,Matthew Bell, et al. A corpus study of evaluative and speculative language. acm. 2001.

[2] Ellen Riloff, Janyce Wiebe, Theresa Wilson. Learning Subjective Nouns using Extraction Pattern Bootstrapping. CoNLL-03. 2003: 25-32

[3] Ellen Riloff, Janyce Wiebe. Learning Extraction Patterns for Subjective Expressions. EMNLP-03. 2003: 105-112

[4] Hong Yu, Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. EMNLP. 2003.

[5] 姚天昉，彭思威. 汉语主客观文本分类方法的研究. 第三届全国信息检索与内容安全学术会议论文集, 2007: 117-123

[6] 叶强，张紫琼，罗振雄. 面向互联网评沦情感分析的中文主观性自动判别方法研究. 信息系统学报, 1(1), 2007: 79-91

[7] 王根，赵军. 基于多重冗余标记CRFs的句子情感分析研究. 中文信息学报, 21(5), 2007: 51-55

[8] 蒙新泛，王厚峰. 主客观识别中的上下文因素的研究. 中国计算机语言学研究前沿进展(2007-2009), 2009: 594-599

[9] 张博. 基于SVM的中文观点句抽取[D]. 北京: 北京邮电大学计算机学院, 2011.

[10] 徐睿峰等. 基于多知识源融合和多分类器表决的中文观点分析. 第三届中文倾向性分析评测会议(COAE). 济南. 2011.

[11] 宋乐等. 中文情感词句识别及文本观点抽取研究. 第二届中文倾向性分析评测会议(COAE). 上海. 2009.

[12] Turney, Peter. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]. In Proceedings of 40th Meeting of the Association for Computational Linguistics, Philadelphia, PA. 2002:417-424.

[13] Kim, Soo-Min and Eduard Hovy. Determining the sentiment of opinions[C]. In Proceedings of COLING 2004, Geneva. 2004:1367-1373.

[14] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques[C]. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA. 2002:79–86.

[15] Songbo Tan and Jin Zhang. An empirical study of sentiment analysis for chinese documents [J]. Expert Systems with Applications: An International Journal. 2008-5:(34)4:2622-2629.

[16] 谢丽星. 基于SVM的中文微博情感分析的研究[D]. 北京: 清华大学, 2011.