DUTIR中文微博情感分析评测报告 *

杨亮，王昊，李雪妮，任巨伟，魏现辉，何文译，林鸿飞
大连理工大学计算机科学与技术学院，大连，116024
E-mail: yangliang@mail.dlut.edu.cn

摘　要：中文情感分析的发展离不开相关评测的推动作用。中文微博情感分析包括了观点句识别，观点句的倾向性判断，观点句情感要素抽取三个子任务。本文介绍了大连理工大学信息检索研究室在微博情感分析评测中的所用的各种方法及资源。评测结果表明了，大连理工大学信息检索实验室的大连理工大学文本倾向性分析知识库对于中文情感分析工作具所起到的辅助作用，但是在机器学习方法的利用，以及在召回率方面都还需要进一步的提升。

关键词：微博情感分析；NLP&CC；文本倾向性分析知识库

DUTIR at COAE2011

Yang Liang, Wang Hao, Li Xueni, Ren Juwei, Wei Xianhui, He Wenyi, Lin Hongfei
Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024
E-mail: yangliang@mail.dlut.edu.cn

Abstract: Chinese Sentiment Analysis Evaluation is indispensable to the development of Chinese sentiment analysis. There are three subtasks in NLP&CC, including opinion sentence detection, opinion sentence identification, and comment object extraction. In this paper, the methods and resources of DUTIR at NLP&CC are detailed introduced. The evaluation result proved the effectiveness of the sentiment ontology constructed by DUTIR. On the other hand, the method innovation still needs to be improved in machine learning and recall.

Keywords: Micro-blog sentiment analysis; NLP&CC; sentiment ontology

1　引言

大连理工大学信息检索研究室（DUTIR）参加了中文微博情感分析&词汇语义关系抽取评测任务中的中文微博情感分析评测部分，其中包括三个子任务。本次评测的对象是面向中文微博的情感分析核心技术，包括观点句识别、情感倾向性分析和情感要素抽取。本文介绍了DUTIR在三个子任务中所用方法及评测结果的分析。

2　观点句识别

该任务采用规则过滤与机器学习相结合的方法完成。基本流程如图 1 所示。

首先我们对针对标注样例进行规则总结，将不作为观点句处理的情况直接过滤掉，主要规则如下：

（1）只有标点符号的语句；

（2）只有网址的语句；

*基金项目：国家自然科学基金项目（60673039，60973068）；国家社科基金（08BTQ025)；教育部留学回国人员科研启动基金和高等学校博士学科点专项科研基金资助课题（20090041110002，20110041110034）。
作者简介：王昊，李雪妮，任巨伟，魏现辉，何文译，研究方向为情感计算；林鸿飞，男，博士，教授，博士生导师，研究方向为搜索引擎、文本挖掘和自然语言处理，hflin@dlut.edu.cn。

（3）只含有 tag 的语句；

（4）含有特殊标点的语句，特殊标点如“【”，“】”，这样的句子有可能是观点句，但是概率较低，因此直接过滤；

（5）只含有表情的句子；

（6）含有“必须”，“希望”，“但愿”等词语的句子。

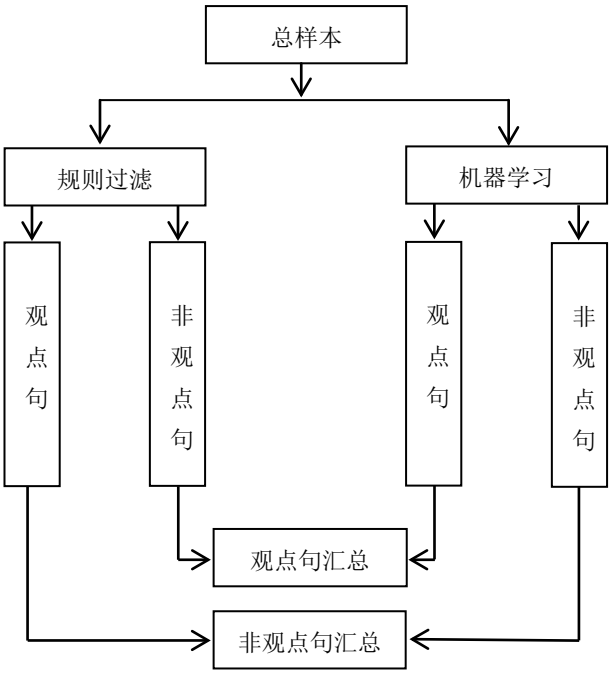

图 1 观点句识别流程图

Fig.1 Opinion Sentence detection flow chart

对于规则过滤后剩余的样本，我们通过机器学习的方法实现观点句的识别。具体方法如下:

首先在规则过滤之前随机选取 10%的样本进行标注作为训练集。标注完成后，统计标注为观点句和非观点句的数量 NY、NN。计算样本比例系数 R=NY/NN。由于标注样本是随机选取，所以可以近似认为总样本中，观点句与非观点句的比例大致也为 R。

接来下将样本表示成 VSM 形式，词的权重用 TF*IDF 计算。对用训练集，用 Lib-SVM 工具进行模型训练，对测试集（未标注的样本）进行分类。Lib-SVM 是一个已经被广泛认可的分类工具。在此,我们使用线性核，并将其所有参数都设为缺省值。

通过标注的训练集可以发现，非观点句的数量大于观点句的数量。因此，作为一个不平衡的数据集，分类时，不可避免的有一部分观点句被错误的分到非观点句一类中。为此，我们需要将此部分判断错误的观点句重新修正。具体方法为如下：将分类结果按照被分为观点句的置信度降序排序。假设测试样本为 n 条，被分为观点句的为 m 条。由于训练集和测试集数据分布大致一致，我们近似认为测试集中观点句有 n*R 条。因此，可以近似认为

有 n*R-m 条观点句被错分为非观点句。因此，我们需要将排好序的样本中第 m+1 条到 n*R 条重新修正为观点句。

通过分析语料可以发现，陈述句形式的观点句比较规整，表达的观点比较明显，所以采用机器学习的方法效果相对较好。而一些非陈述形式的观点句，比如感叹句、反问句，表达的观点常常比较隐含。为了减小此部分给整体带来的噪音，我们在提交的第二组结果中，将陈述句和非陈述句分别分开再分类。实验结果（如表 1 所示）表明，该组结果略逊于全部样本一起做分类的第一组结果。究其原因，我们认为是非陈述句的表达方式多样，观点表达不明显，而且非陈述句的训练集较小，所以导致第二组结果相对略差。

表 1 任务一结果

Tab.1 The results of task 1

结果编号	微平均			宏平均		
	准确率（P）	召回率（R）	F 值	准确率（P）	召回率（R）	F 值
DUTIR-1	0.825	0.603	0.697	0.828	0.589	0.679
DUTIR-2	0.822	0.592	0.688	0.824	0.581	0.674

从提交的 52 组结果来看，我们的召回率偏低（29,30/52），比最高值低大约 0.3，而准确率较高（3,4/52），比最好结果低 0.01，从而使得整体 F 值（23,27/52）比最好结果低大约 0.1。这说明我们返回的观点句个数太少，原因有两个：1）直接匹配观点句的规则过于死板，2）标注语料时，对于观点句的判断规则过紧，导致观点句的数量比实际要少。

3 观点句的情感倾向性判断

该任务是在任务一的基础上完成的。我们采用机器学习的方法进行倾向性判断。通过分析标注的训练集我们发现，NEG 的比例远远大于 POS 和 OTHER。对于这样不平衡的语料，分类结果中显然会将部分 POS 和 OTHER 的句子错分为 NEG。类似于任务一，我们根据分类结果的置信度调整测试集中各极性样本的比例，使之与训练集中的各部分比例一致。

表 2 任务二结果

Tab.2 The results of task 2

结果编号	微平均			宏平均		
	准确率（P）	召回率（R）	F 值	准确率（P）	召回率（R）	F 值
DUTIR-1	0.841	0.507	0.633	0.849	0.497	0.62
DUTIR-2	0.833	0.493	0.619	0.843	0.487	0.611

该任务的提交的两组结果是在任务一两组结果的基础上完成的。从提交的 47 组结果来看，我们召回率偏低（21,23/47），比最好队伍低大约 0.3，准确率（12,13/47）比最好结果低约 0.1，整体 F 值（18,21/47）比最好值低约 0.2。造成我们 F 值不理想的主要原因是召回率欠佳。这是由于任务一召回率欠佳，造成部分观点句未识别出来造成的。

4 观点句情感要素抽取

本任务要求找出微博中每条观点句作者的评价对象，即情感要素。同时判断针对情感

要素的观点极性。评测数据集包含每条微博中的各个句子，参赛队伍需要先进行观点句识别再进行情感要素抽取。

4.1 抽取流程

本任务中，结合大连理工大学 IR 实验室的情感本体库[1]，我们使用条件随机域（Confidence Random Field，简称 CRF）与语义规则相结合的方法进行抽取。抽取流程如图 2 所示。

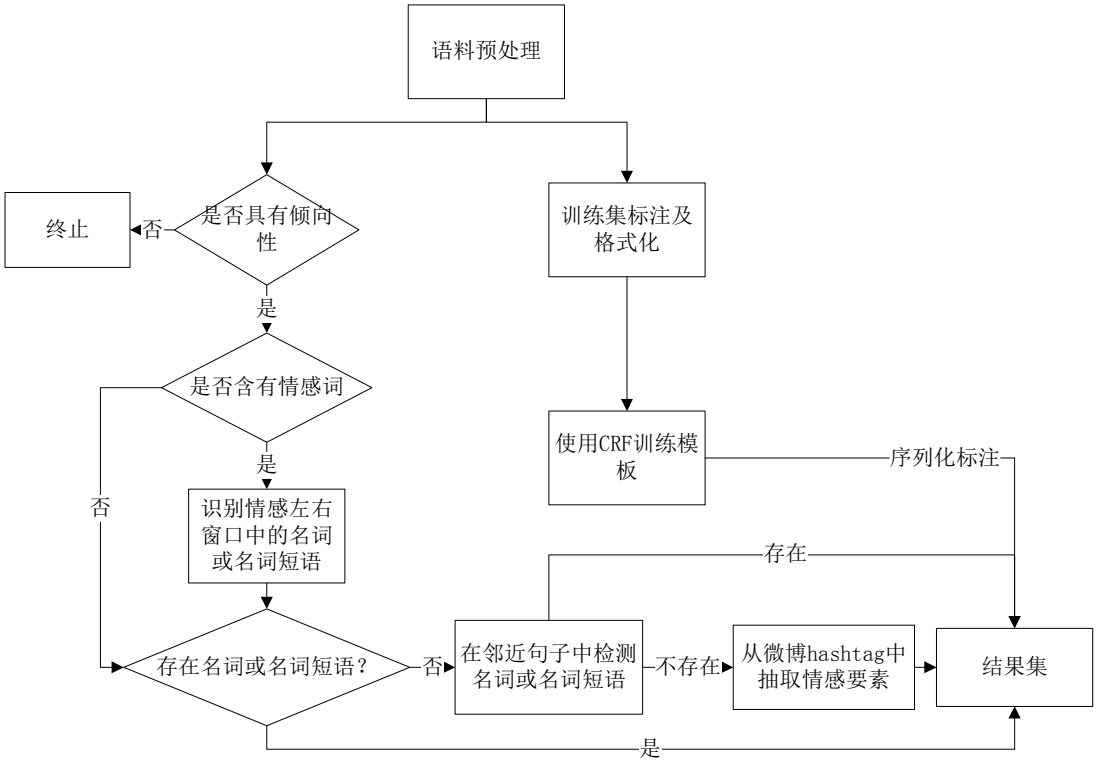

图 2 观点句识别流程图

Fig.2 Opinion Sentence detection flow chart

步骤 1：语料预处理。以句子为单位，使用中科院计算所的汉语分词系统 ICTCLAS50 进行分词和词性标注。

步骤 2：训练集选取。从语料每个文件中随机抽取总句子数的 10%作为训练集，对于每个句子，人工标注其观点倾向性及所包含的情感要素。

步骤 3：格式化训练集和测试集，使用 CRF，选取词性和词的情感类别特征加入 CRF 中，构造模板文件，训练模型并进行序列化标注，对情感要素做初步筛选。

步骤 4：使用规则方法做进一步抽取。在任务 1 的基础上，对于具有情感倾向性的句子，若其中含有情感词，则以情感为中心，选取一定大小窗口内的名词或名词短语作为候选情感要素；对于窗口中不含名词或名词短语的句子，则转步骤 5。

步骤 5：在同一微博中的上一句或下句中寻找距离情感词最近的名词或名词短语作为

候选情感要素。

步骤 6：对于任务 1 中含有情感倾向性的句子，若其中不存在名词性短语，则从微博的 hashtag 中抽取情感要素。

4.2 CRF 抽取情感要素

情感要素的抽取可以转换为序列标注的问题，而对于序列标注问题，我们利用条件随机域（Confidence Random Field，简称 CRF）实现。CRF 是由 Lafferty 等人在 2001 年新提出的一个分割和标注序列的框架[5]。给定序列 X，标注为序列 Y 的概率公式如下：

$$P_\theta(y \mid x) = \exp(\sum_{e \in E,k} \lambda_k f_k(e, y \mid e, x) + \sum_{v \in V,k} u_k g_k(v, y \mid v, x))$$

这里x是观察值序列，y是标记序列，$y|s$表示子图s中和y相连的顶点f_k和g_k都是预先设定的特征。

本文使用的 CRF 工具为 CRF++-0.54，分别使用了两种方法训练模板，方法一是单独选取词性作为特征，以基准词前后 4 个词项作为相关词项设计模板，训练模型；方法二是在词性特征的基础上加入词的情感类别特征，同样以基准词前后 4 个词项作为相关词项设计模板，训练模型。两组结果表明：只添加词性特征的模型召回率和准确率更高，而后者则由于情感类别过多以及词在不同语境下的情感变化差异带来了不必要噪音，在降低了准确率和召回率的同时也降低了算法的效率。所以我们选择前者训练的模板对语料进行序列化标注，抽取情感要素。

4.3 结果分析

从表 1 表 2 可以看到，本文的方法在准确率上获得了不错的效果，但是因为召回率太低影响了本文方法的 F 值。

本文的方法在训练集上测试时，准确率也不高，产生这个问题的原因主要是我们在选择模型的时候对原始数据处理不够。本文主要采取了词特征作为 CRF 的基本特征。在语料集比较小的时候，有限的训练集制约了 CRF 的表现。虽然 CRF 在识别短语上有很好的效果，但在未登录词的标注上，CRF 的作用很有限。在以后的评测中本文会考虑采用把词聚类或者分类的方法对特征进行进一步改进。

表 3 任务三结果 a（严格评价指标）

Tab.3 The results of task 3(a) (Strict)

结果编号	微平均			宏平均		
	准确率（P）	召回率（R）	F 值	准确率（P）	召回率（R）	F 值
DUTIR-1	0.485	0.066	0.116	0.474	0.066	0.113
DUTIR-2	0.425	0.077	0.131	0.417	0.076	0.126
Best	0.303	0.275	0.288	0.306	0.263	0.278

表 4 任务三结果 b（宽松评价指标）

Tab.4 The results of task 3(b) (Loose)

结果编号	微平均			宏平均		
	准确率（P）	召回率（R）	F 值	准确率（P）	召回率（R）	F 值
DUTIR-1	0.636	0.086	0.152	0.643	0.086	0.149
DUTIR-2	0.569	0.099	0.169	0.572	0.098	0.165
Best	0.388	0.354	0.371	0.393	0.342	0.359

同时在组委会给出最终答案的时候，我们对本文的标注和答案进行了比对，结果是我们的训练集在答案上的召回率比较低（因为放出的答案不是全部的，准确率无法测算），这也是制约本文结果的一个原因。

参 考 文 献

[1] 徐琳宏, 林鸿飞, 潘宇等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185

[2] 陈建美，林鸿飞，杨志豪. 基于语法的情感词汇自动获取[J].智能系统学报, 2009, 4(2):100-106

[3] 潘凤鸣，王宇轩，常富洋等. DUTIR COAE2009 评测报告[].

[4] 宋锐,洪莉,林鸿飞. 基于 ChunkCRF 的观点持有者识别及其在观点摘要中的应用[J].小型微型计算机系统，2009，7:1462-1466

[5] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//18th International Conf on Machine Learning. San Francisco, USA: Morgan Kaufmann,2001: 282-289.