基于最大熵模型的中文词与句情感分析研究*

董喜双，关毅，李本阳，陈志杰，李生
哈尔滨工业大学，哈尔滨，150001
dongxishuang@gmail.com, guanyi@hit.edu.cn, libenyang012566@163.com, ruoyu_928@126.com, lisheng@hit.edu.cn

摘　要：本文将研究焦点对准喜、怒、哀、惧四类情感分析问题，重点解决中文词、句的情感分析问题。将词的情感分析处理为候选词情感分类问题。首先通过词性过滤获得候选词，进而根据特征模板获取候选词情感特征，然后应用最大熵模型判断候选词情感类别，最后应用中性词典、倾向性词典、复句词表、否定词表过滤候选情感词分类错误得到情感词集合。句的情感分析首先根据情感词典和倾向词典提取词特征，并采用规则提取词序列特征，然后采用最大熵模型对句子进行情感分类。在 COAE2009 评测中词与句情感分析取得较好结果。

关键词：情感分析；情感极性；最大熵；分类；

Sentiment Analysis on Chinese Words and Sentences Based on Maximum Entropy Model

Dong Xi-Shuang, Guan Yi, Li Ben-Yang, Chen Zhi-Jie, Li Sheng
Harbin Institute of Technology, Harbin 150001
dongxishuang@gmail.com, guanyi@hit.edu.cn, libenyang012566@163.com, ruoyu_928@126.com, lisheng@hit.edu.cn

Abstract: This paper presents a method to analyze sentiments on Chinese words and sentences, where the sentiments include happy, angry, sad, and fear. In the case of words, sentiment analysis was processed as the sentiment classification of candidate words. The candidate words were firstly obtained by POS filtering, then Maximum Entropy (ME) model was adopted to judge sentiment categories of the words, which sentiment features were gained with feature templates. Finally, errors in the word classification would be removed through filtering with a neutral lexicon, a sentiment polarity lexicon, a connective word list of complex sentences, and a negative word list. In the case of sentences, word features in sentences were extracted on the basic of the sentiment lexicon and the sentiment polarity lexicon, and word sequence features were extracted by rules while processing sentiment analysis on sentences, then ME model was used to classify the sentences. Good performance of sentiment analysis was gained in COAE 2009.

Keywords: Sentiment Analysis, Sentiment Polarity, Maximum Entropy, Classification

1　引言

情感分析的主要任务为识别文本对某一事物的观点[1]。情感包含两方面信息：情感极性与情感强度。情感极性指情感要素（词、短语、句子以及篇章）表达的情感倾向。情感强度指情感要素表达情感的强弱程度。情感分析包含四方面研究内容：词级情感分析、短语级情感分析、句级情感分析以及篇章级情感分析。词级情感分析包括识别候选情感词、判断候选情感词情感极性与强度以及构建情感字典[2]。短语级情感分析为根据情感词识别

*董喜双，1981 年出生，男，黑龙江省哈尔滨市，博士研究生。本项研究受到国家自然科学基金项目支持，项目批准号：60975077，60736044

情感短语并判定情感极性与强度[3]。句级情感分析为识别句级观点持有人、评价对象以及判断句子的情感倾向[2][4]。篇章级情感分析为识别篇章对某一事物的观点[5-6]。文本情感分析可用来决定获取何种信息并且如何呈现和组织信息。例如信息检索系统可应用情感分析过滤、获取支持某一特定政治倾向的文本[7]。问答系统可根据观点扩展查询，获得更加全面、精准的答案[8]。

本文主要涉及情感分析两方面：词级情感分析和句级情感分析。词级情感分析要求在一定的上下文环境中抽取出能够明确表达作者情感的词，并判断该情感词所属的类别。句级情感分析要求在一定的上下文环境中抽取出能够明确表达作者情感的句子，并判断该情感句所属的类别。其中情感类别包括：喜（happy）、怒（angry）、哀（sad）和惧（fear）。两方面问题难点在于情感类别增至四类使分类更加困难。因而本文将这一困难作为研究重点。

本文结构组织如下：第二部分介绍相关研究工作；第三部分简介最大熵模型；第四部分重点描述词、句级情感分析的方法及优缺点；第五部分分析实验结果；最后给出结论与展望。

2 相关研究

词级情感倾向分析主要任务是判断候选词情感。当前方法主要有两种：（1）基于电子词典的候选词情感分析；（2）基于机器学习的候选词情感分析。

利用电子词典判断候选词的情感相关工作包括：文献[9]利用 WordNet 和 General Inquirer(GI)[10]的同义词集和反义词集获取候选词的情感倾向信息；文献[11]利用 HOWNET 提供的语义相似度方法计算词与基准情感词集的语义相似度值，以此推断该候选词的情感倾向；文献[12]利用《同义词词林》中的同义词词群扩展基准情感词集。这些方法缺点在于对已有的电子词典具有较强的依赖性。

基于机器学习的候选词情感分析方法包括基于无监督学习和基于有监督学习的候选词情感分析。文献[3]计算词与种子情感词的点互信息(Pointwise Mutual Information, PMI)，以此推断该词的情感倾向。文献[13]则在 PMI 方法的基础上结合文本中连接上下文的关联词处理，进一步挖掘文本中的情感词。无监督的机器学习方法依赖于处理语料的领域范围，同样存在着对基准情感词的依赖性问题，而且正确率较低。基于有监督学习方法如：文献[14]利用词语搭配模式发现在主观性文本中的倾向性词语及其搭配关系；文献[15]利用从情感标注语料中抽取的上下文模板，统计词与上下文模板之间的关系，进而判断该词的情感倾向。基于有监督学习方法精度较高，但缺陷是人工标注语料库的缺乏以及语料库标注的不一致性。

句级情感倾向分析主要任务是判别句子的情感倾向性。文献[2]通过获取特定区域（窗口 1：句子内部；窗口 2：句中评论人与评价对象之间；窗口 3：窗口 2 前后两个词；窗口 4：窗口 2 到句尾）内的情感特征，分别利用情感倾向累乘模型、情感强度调和平均模型以及情感强度几何平均模型判断句子情感倾向性。实验表明在窗口 4 区域内识别特征并结合情感倾向累乘模型准确度达到 81%。该方法主要缺陷在于需正确标注评价人和评价对象，同时情感累乘模型无法准确判断否定句情感倾向。文献[7]将观点句分析处理为分类问题，并利用朴素贝叶斯分类模型达到 90%精度。该方法难点在于精准的提取情感特征。文献[16]将情感句分析类比为句子的情感序列化标注问题。该方法不仅从句子本身的情感分析角度出发，还考虑其临近句子对其情感倾向的影响以及整个篇章对其情感倾向影响。该方法明显优点是考虑了不同级别情感分析之间的相互影响，但其缺点在于复杂的编码和解码过程。

本文解决词级情感分析问题时首先构建情感词典，然后借鉴文献[17]思想，采用情感

词分类方法，通过提取候选词周围的不同特征，利用最大熵模型判断候选词的情感极性，并以类别概率作为结果的置信度。句级情感分析在文献[7]的基础上，采用情感分类方法判别句的情感极性。首先介绍本文使用的分类模型。

3 最大熵模型

文献[18]基于信息熵理论建立了最大熵模型。在一定的限制条件下，选择一个系统的最优分布时，如果这些限制条件无法确定唯一的系统分布，那么最好的分布就是在满足所有限定条件下，系统信息熵最大的分布。

给定 ***H*** 代表特征集合，最大熵模型的目标是寻找最优的标记 ***T***（使条件概率 $p(T \mid H)$ 的条件熵最大）。由最大熵的独立性假设，不考虑标记之间影响，以 ***t*** 代表一个特定的状态，***h*** 代表该状态的上下文观测值，条件熵可被定义为：

$$H(p) = -\sum_{t \in T,\ h \in H} \tilde{p}(h)\, p(t \mid h)\, f(t,h) \tag{1}$$

其中，$\tilde{p}(h)$ 为特征 ***h*** 的先验概率，$f(t,h)$ 为特征函数，$p(t \mid h)$ 为状态 ***t*** 的条件概率。由于最大熵模型的解是存在的且唯一，因此可通过运算得到：

$$p(t \mid h) = \frac{\exp(\sum_i \lambda_i f_i(t,h))}{Z(h)} \tag{2}$$

其中，f_i 为特征 ***i*** 的特征函数，$Z(h) = \sum_t \exp(\sum_i \lambda_i f_i(t,h))$ 为归一化因子。λ_i 是特征 ***i*** 的权重，训练过程就是用数值算法求每个 λ_i 值的过程。最大熵模型在自然语言处理领域应用广泛，其中文献[19]最早在自然语言处理领域使用最大熵模型。本文在词、句情感分析中采用最大熵模型作为分类器并取得较好效果，下面详细描述词、句情感分析过程。

4 词句情感分析

4.1 词级情感分析

首先通过三种模型构建情感词典。然而情感词典中情感词数量有限，因此采用最大熵分类模型进一步挖掘情感词。本文在搜狗实验室[20]提供的互联网语料库(SogouC)以及聚友网[21]上的博客、论坛文章上构建情感词典，语料规模有近 15000 篇文章，大小近 70M。应用如下三种模型构建情感词典：

(1)字符情感分值计算方法

Ku 在 NTUSD[22]词典的基础上引入了字符情感分值计算方法。字符情感分值计算方法的基本思想是以构成词的字符的情感倾向表征词的情感倾向。

(2)语义相似度计算方法

词语的语义相似度反映了文本中词语的可替换程度。两个词语，如果在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。通过计算未知情感的候选词与基准情感词集的语义相似度，进而判断该词

的情感倾向。

(3)点互信息计算方法

点互信息是用来计算分布中两个特定样本点之间的互信息，即衡量两个元素之间的关联程度。通过计算未知情感的候选词和基准情感词集的点互信息，进而判断该词的情感倾向。

情感词典构建具体流程如图 1 所示。

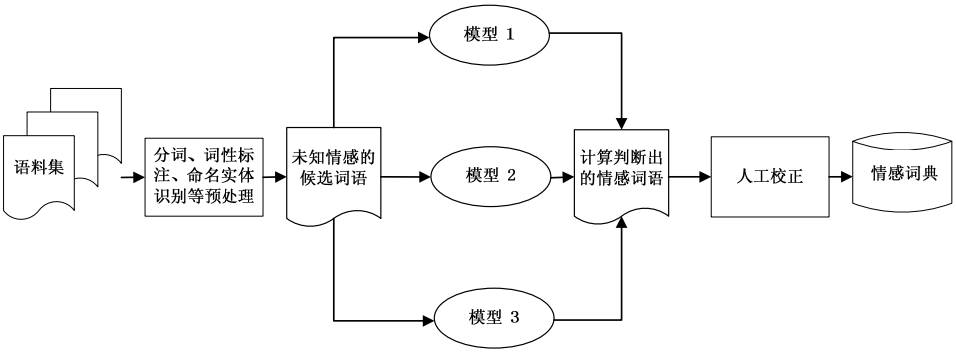

图 1 中文情感词典构建流程图

Fig.1 the Flow Chart of Constructing the Chinese Sentiment Lexicon

图 1 中，模型 1 表示字符情感分值计算方法，模型 2 表示语义相似度的计算方法，模型 3 表示点互信息计算方法。中文情感词典的构建流程如下：

中文情感词典的构建流程如下：

(1) 对语料进行分词、词性标注等预处理，并依据初始情感词典发现文本中未知情感的候选词语。

(2) 将候选词语及其上下文送入到模型中，判别候选词语的情感倾向。

(3) 对各个模型的判别结果进行人工校正，形成情感词典。

由于情感词典是一个有限集合，不能保证最大程度识别语料中的情感词，因此借鉴文献[17]思想通过选择 3 大类特征，采用最大熵模型挖掘情感词。

(1) 词性特征

本文采用 LMR 情感词模板[17]获取词性特征。LMR 设定为三个词，符号 L、M、R 分别表示左一词、待测定词和右一词。所使用的特征为三个词的词性信息。如待预测词“欣然”。

例如：李明/n 欣然/d 接受/v 任务/n 。/w

则获取的词性特征为 n_d_v。

(2) 句型特征

通过观察评测数据发现部分句型影响情感词表达情感，因而选择句型信息明显的问句、感叹句以及表示祝愿的祈使句作为句型特征。

(3) 字特征

将预测词拆成单字作为统计特征。例如“欣然”拆成“欣”与“然”分别作为统计特征。

当获取候选情感词特征后，使用最大熵模型挖掘情感词，流程如图 2 所示。

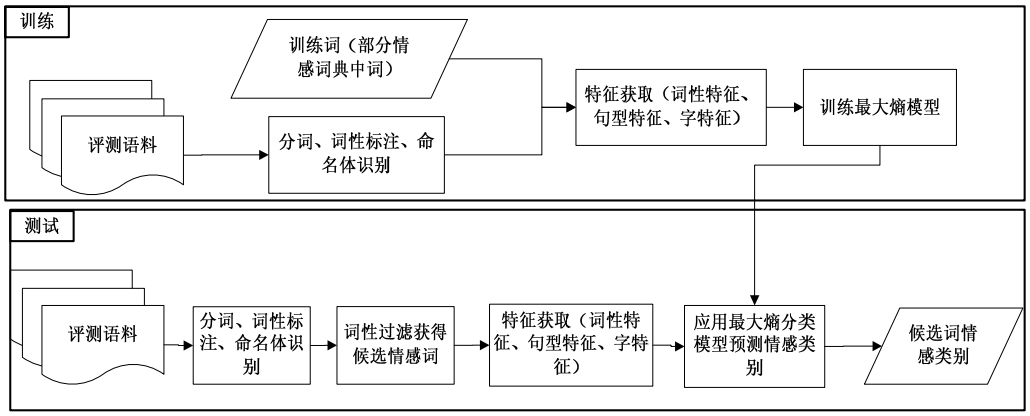

图 2 识别情感词流程图

Fig.2 the Flow Chart of Recognition of Sentiment Words

经上述流程获得候选情感词情感类别，但观察发现候选情感词中存在错误，因而使用中性词典、否定词表、复句词表以及倾向性词典过滤部分错误得到最终情感词集合。

4.2 句级情感分析

本文主要采用分类的方法判断句子情感。分类的方法能够很容易的融入更多的有用的特征，同时也可获得较高的准确率[3]。但分类方法需要情感标注语料作为训练语料，因此首先人工构造四类情感标注语料作为训练语料。

本文对比了最大熵和支持向量机分类模型,实验结果发现两者 F 值差别在 1%-2%之间，但处理情感分类时最大熵模型具有更好的表现，最终采用最大熵模型。具体流程如图 3 所示。

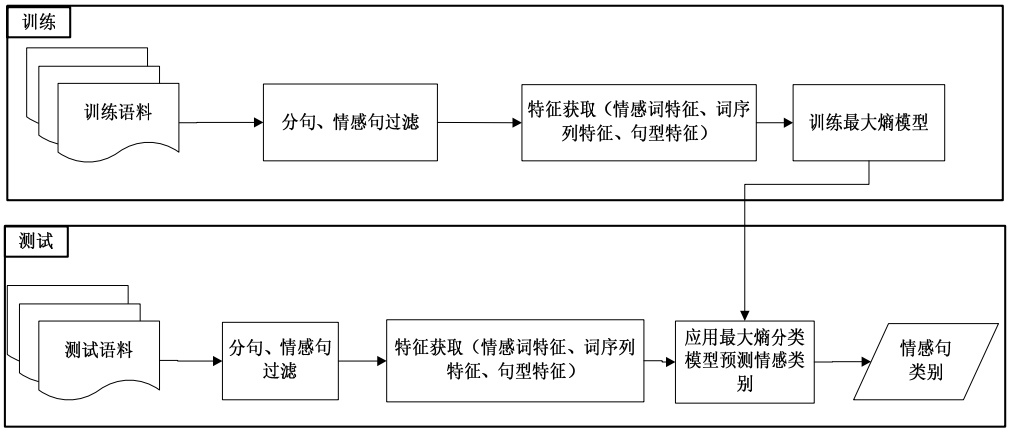

图 3 识别情感句流程图

Fig.3 the Flow Chart of Recognition of Sentiment Sentences

1. 情感句过滤

由于情感句在句子集合中所占比例小而导致语料不平衡，对于分类结果影响较大。另一方面情感句大部分含有情感词，因此使用情感词典过滤不含有情感词的句子，并可有效的减少语料不平衡造成的影响。同时书名号里面是作品名，对于情感分析没有帮助，因此去掉书名号中的内容。

2. 特征获取

选取的特征主要包括三类：

(1)词和词序列

除了识别情感词语倾向词特征外，否定词也会对情感句分析产生影响，而且单纯使用词汇无法体现出词语的顺序关系，因此对于含有情感词、否定词的一定距离内的词序列进行了提取。根据实验选定词间距离为 6 的词序列。

(2)情感词强度累加值

在简单句中，通常某一类比重大的情感词决定该句情感，因此情感词的数量以及情感词强度对句子情感分析有较大影响。因此将不同类别的情感词强度累加值以及强度最大的类别作为特征。强度累加值特征函数如式 3 所示。

$$C(x)=\begin{cases}0 & x=0\\ 1 & 4>x>0\\ 2 & 7>x\geq 4\\ 3 & 10>x\geq 7\\ 4 & x\geq 11\end{cases} \tag{3}$$

其中，$\boldsymbol{x}$ 表示情感词强度累加值。

(3)其他

句型信息有助于判断情感类别，选择问句与感叹句作为句型特征；通常句子越长出现情感词等特征就会越多，所以句长也作为特征。句长的特征函数如式 4 所示。

$$F(t)=\begin{cases}1 & t<24\\ 2 & 100>t\geq 24\\ 3 & 300>t\geq 100\\ 4 & t\geq 300\end{cases} \tag{4}$$

其中，$\boldsymbol{t}$ 表示句子中词的个数。

5 实验与分析

5.1 数据

采用COAE2009 评测提供的 dataset1 数据集，文件格式是简体中文的 txt 文本文件。dataset1 语料集由 4 万篇文本构成，主观文本与客观文本混合，主观文本不少于 20%，包括真实用户评论和新闻报道评论等，涉及财经、娱乐、影视、教育、房地产、电脑、手机等领域，文章长度从几个句子到上百个句子不等。

5.1 实验结果与分析

词级情感分析实验结果如表 1 所示。

表 1 词级情感分析

Tab.1 Sentiment Analysis on Words

	Precision	P@1000	Recall	F1	R-accuracy
HITLTRTask1-angry-1.txt	0.486239	0.106	0.114101	0.18483	0.114101
HITLTRTask1-fear-1.txt	0.560284	0.079	0.116691	0.193154	0.116691
HITLTRTask1-happy-1.txt	0.569892	0.159	0.102913	0.174342	0.102913
HITLTRTask1-sad-1.txt	0.465565	0.169	0.137287	0.212045	0.137287
AVG	0.520495	0.128	0.117748	0.1910928	0.117748
MEDIAN	0.479425	0.14263	0.1254997	0.1775775	0.1254997
MAX	0.619184	0.192	0.1795165	0.2078715	0.166954

其准确度好于平均值并且 F1 值略低于最大值，但其 P@1000 值、召回率、以及 R-accuracy 值要低于平均值，其原因主要在于在挖掘情感词时过滤了部分单字词以及三字词，另一方面，分词错误也会影响最终情感词识别结果。

句级情感分析结果共两组，结果分别如表 2 和 3 所示。

表 2 句级情感分析 1

Tab.2 Sentiment Analysis on Sentences 1

	Precision	P@1000	Recall	F1	R-accuracy
HITLTRTask2-angry-1.txt	0.374618	0.245	0.149299	0.213508	0.149299
HITLTRTask2-fear-1.txt	0.404124	0.392	0.238298	0.299809	0.238298
HITLTRTask2-happy-1.txt	0.212308	0.414	0.201853	0.206948	0.182838
HITLTRTask2-sad-1.txt	0.154557	0.351	0.182243	0.167262	0.136033
AVG	0.28640175	0.3505	0.19292325	0.22188175	0.176617
MEDIAN	0.158407	0.366	0.201223	0.1246522	0.15746
MAX	0.370662	0.6805	0.374071	0.232405	0.189788

表 3 句级情感分析 2

Tab.3 Sentiment Analysis on Sentences 2

	Precision	P@1000	Recall	F1	R-accuracy
HITLTRTask2-angry-2.txt	0.301822	0.265	0.161487	0.210401	0.161487
HITLTRTask2-fear-2.txt	0.422311	0.424	0.257751	0.320121	0.257751
HITLTRTask2-happy-2.txt	0.242439	0.497	0.242321	0.24238	0.210629
HITLTRTask2-sad-2.txt	0.140903	0.34	0.176532	0.156718	0.129283
AVG	0.27686875	0.3815	0.20952275	0.232405	0.1897875
MEDIAN	0.158407	0.366	0.201223	0.1246522	0.15746
MAX	0.370662	0.6805	0.374071	0.232405	0.189788

从表 2 数据可观察到，准确度、P@1000 以及召回率较大的低于最大值，但 F1 值以及 R-accuracy 已接近最大值，这主要是模型在准确率和召回率方面表现比较平衡。另一方面，

惧（fear）类准确度高于最大值近 10%，这主要是由于这类情感特征明显、易提取。而表达喜（happy）和哀（sad）精度较低原因在于部分情感特征不易区分。表 3 中大体反映出与表 2 一致的现象，当其 F1 值已经达到最大值，产生这一结果的原因在使用了一个相对丰富的情感词典。

6 结论与展望

本文采用分类思想处理喜、怒、哀、惧四类情感的词、句情感分析问题。词级情感分析首先构建情感词典，然后借助情感词典中部分情感词并采用分类思想挖据语料中新情感词。句级情感分析通过选取情感特征（情感词、倾向词、否定搭配等），采用最大熵分类器对句进行情感分类并在句级情感分析中取得较好结果。

下一步工作将尝试深入挖掘情感词的各方面特征，同时在处理句级情感分析时将考虑句子间情感影响以及篇章对句子的影响。另一方面，分别以词级和句级情感分析为基础将研究扩展到短语以及篇章情感分析。

感谢

在此感谢孙慧、薛璐影、李超、张书娟、季知祥、阎于文等同学为本文工作提出的宝贵建议和付出的努力。

参 考 文 献

[1] Bo Pang, Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the ACL, 2004, pp.271-278.

[2] Soo-Min Kim, Eduard Hovy. Determining the sentiment of opinions. In Proceedings of COLING, 2004, pp.1367-1373.

[3] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp.417-424.

[4] Y. Mao, G. Lebanon. Isotonic conditional random fields and local sentiment flow. In Proceedings of NIPS, 2006.

[5] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of EMNLP, 2002, pp.79-86.

[6] Zhang, Y., Li, Z., Ren, F., et al. Semi-automatic emotion recognition from textual input based on the constructed emotion thesaurus. In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2005, pp.571-576.

[7] Hong Yu, Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions. In Proceedings of EMNLP-03, 2003, pp.129-136

[8] Cardie Claire, Janyce Wiebe, Theresa Wilson, et al. Combining low-level and summary representations of opinions for multi-perspective question answering. In AAAI Spring Symposium on New Directions in

Question Answering, 2003, pp.20-27.

[9] Vasileios Hatzivassiloglou, Kathleen R. McKeown. Predicting the Semantic Orientation of Adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, 1997, pp.174-181.

[10] General Inquirer. http://wjh.harvard.edu/~inquirer.

[11] 朱嫣岚, 闵锦等. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报, 2006, Vol.20, No.1: 14-20.

[12] 路斌, 万小军, 杨建武, 陈晓鸥. 基于同义词词林的词汇褒贬计算. 第七届中文信息处理国际会议, 2007: 17-23.

[13] 姚天昉, 娄德成. 汉语情感词语义倾向判别的研究 [A]. 中文计算技术与语言问题研究-第七届中文信息处理国际会议论文集 [C], 2007: 221-225.

[14] Janyce Wiebe, Rebecca Brucey, Matthew Bell, et al. A Corpus Study of Evaluative and Speculative Language. In Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, 2001, pp.1-10.

[15] Qi Zhang, Xi-Peng Qiu, Xuan-Jing Hung, Li-De Wu. Learning Semantic Lexicons using Graph Mutual Reinforcement based Bootstrapping. Acta Automatica Sinica, Vol.34 (10), 2008: 1257-1261.

[16] R. McDonald, K. Hannan, T. Neylon, et al. Structured models for fine-to-coarse sentiment analysis. In Proceedings of the 45th Association of Computational Linguistics, 2007, pp.435-439.

[17] 何慧，李思，肖芬，徐蔚然，郭军. PRIS 中文情感倾向性分析技术报告. 第一届中文情感分析测评. 2008: 46-55.

[18] T. Jaynes. Information Theory and Statistical Mechanics. Physics Reviews. 1957(106): 620-630

[19] Adam L. Berget, Stephen A. Della Pietra, Vincent J.Della Pietra. A Maximum Entropy Approach Natural Language Processing. Computational Linguistics, Vol. 22, No. 1. (1996), pp. 39-71.

[20] Sogou Labs. http://www.sogou.com/labs/.

[21] Myspace.cn. http://www.myspace.cn/.

[22] Lun-Wei Ku, Yu-Ting Liang, Hsin-Hsi Chen. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006, pp.100-107.