基于最大熵模型和最小割模型的中文词与句褒贬极性分析*

董喜双，邹启波，关毅，高翔，闫铭
哈尔滨工业大学，哈尔滨，150001
dongxishuang@gmail.com，zouqibo2009@163.com，guanyi@hit.edu.cn，hngaoxiang@gmail.com，
mingitouch@gmail.com

摘　要：本文运用最大熵模型和最小割模型预测中文词和句子的褒贬极性。词级情感分析首先构建领域情感词典，然后根据领域情感词典提取候选词，并使用最大熵模型预测候选词的极性，最后采用最小割模型优化极性结果。句级情感分析首先根据领域情感词典识别观点句，将观点句切分成短句并基于规则提取特征，应用最大熵模型预测短句的极性，最后根据短句的极性预测长句的极性。

关键词：情感极性，情感分析，最大熵，最小割；

Positive and Negative Polarity Analysis on Chinese Words and Sentences Based on Maximum Entropy Model and Min-Cut Model

Dong Xi-Shuang, Zou Qi-Bo, Guan Yi, Gao Xiang, Yan Ming
Harbin Institute of Technology， Harbin，150001
dongxishuang@gmail.com，zouqibo2009@163.com，guanyi@hit.edu.cn，hngaoxiang@gmail.com，
mingitouch@gmail.com

Abstract: In this paper, Maximum Entropy Model and Min-Cut Model were adopted to predict the positive and negative polarities of Chinese words and sentences. First, we built a domain sentiment lexicon. Then, the candidate sentiment words were recognized by the lexicon, and the polarity was predicted by Maximum Entropy model. Finally, we used Min-Cut model to optimize the polarity results. On the side of sentiment analysis on sentences, opinion sentences were obtained by the lexicon. These opinion sentences were split up into short sentences, and sentiment features were extracted by rules. Then the polarity of short sentences was predicted by Maximum Entropy model. Finally, polarities of opinion sentences were predicted according to polarities of short sentences.

Keywords: Sentiment Polarity, Sentiment Analysis, Maximum Entropy, Minimum Cut;

1　引言

情感分析的基本任务指对给定文本的极性进行分类[1]。按照处理的粒度不同，情感倾向性分析主要包括四个方面的研究内容：词级情感倾向性分析、短语级情感倾向性分析、句子级情感倾向性分析和篇章级情感倾向性分析。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[1]。

本文主要研究包括词和句子级情感倾向性分析。词级主要完成的任务是根据上下文抽

*

取出观点词，并判断观点词的褒贬极性，本文采用最大熵模型预测观点词的情感极性，然后通过最小割模型优化极性结果。句子级主要完成的任务是从文本中识别出观点句，再判断观点句的褒贬极性。本文将观点句切分成短句并用最大熵模型预测短句情感极性，然后用短句预测长句的情感极性。

本文的组织结构如下：第二部分介绍相关研究；第三部分介绍相关的模型；第四部分具体介绍词和句子的情感分析方法；第五部分分析实验结果，最后是结论和展望。

2 相关研究

词汇的情感倾向性分析研究主要包括两类：语义方法和统计方法[17]；语义的方法主要是通过一个现有的知识库，然后通过计算候选词和知识库里面的基准词的语义距离，进而得到候选词的情感倾向。在英文方面，Kamps, Marx, Mokken 和 Rijke 于 2002 年提出了基于 WordNet 的同义结构图计算候选词和知识库基准词的语义距离，进而判断候选词的情感倾向性[6]；在中文方面，朱嫣岚、闵锦、周雅倩于 2006 年提出了基于 HowNet 的词汇语义计算方法[7]；路斌、万小军、杨建武于 2007 提出了基于《同义词词林》来计算词汇的褒贬性[8]。统计方法主要是基于有监督和无监督的机器学习[5]，利用文本中词汇间的共现关系来计算词汇的倾向性；Peter D. Turney 和 Michael L. Littman[9]于 2003 年提出了基于搜索引擎的“NEAR”操作计算候选词和种子词之间的相关性，从而得到候选词的情感倾向性。Yu 和 Hatzivassiloglou[10]于 2003 年提出了一个基于种子词典的方法, 首先选择部分情感词作为种子词构建情感词典，通过计算候选词和种子词的共现概率来判断候选词的倾向性。

句级情感分析的研究方法主要是统计方法。Soo-Min Kim 和 Eduard Hovy[2]等人于 2004 年提出了情感倾向累乘模型、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的方法。Hong Yu 和 Vasileios Hatzivassiloglou[10]于 2003 年提出了通过朴素贝叶斯模型来预测句子情感极性。McDonald[14]于 2007 年将句子情感极性分析转化成为情感极性序列化标注，该方法考虑到篇章中上下文对当前句子的影响，但是编码解码过程比较复杂。

本文对词的情感分析研究借鉴了文献[12]的思想，首先采用最大熵模型预测词的极性，然后采用最小割模型优化极性结果。句子情感分析借鉴文献[5]中关于最大熵模型在句子分类中的研究，本文将句子划分为短句并基于规则提取特征，然后采用最大熵模型预测短句极性，最后用短句极性预测长句极性。

3 相关理论和模型

本文用最大熵模型用来预测词和句子的极性，最小割模型优化词的极性结果。

3.1 最小割原理

文献[11]给出了流网络的定义。流网络 $G=(V,E)$ 的割 (S,T) 将 V 划分为 S 和 $T=V-S$ 两个不同的集合，满足条件 $s\in S,t\in T$，穿过割 (S,T) 的净流量定义为 $f(S,T)$，割 (S,T) 的容量定义为 $c(S,T)$ [11]，所以一个流网络的最小割就是这个流网络所有割中具有最小容量的割。

流网络 $G=(V,E)$ 的最小割的容量 $c(S,T)$ 等于其最大流 $|f|$ 值，文献[11]给出了证明，本文就不再赘述。1965 年 Ford 和 Fulkerson 给出了计算最大流的方法 Ford-Fulkerson 算法，文献[11]给出了详细介绍。其基本思想就是不断地寻找增广路[11]，当找不到增广路

时，表示流网络中没有流量可以扩展，此时流量达到了最大值。

3.2 最大熵模型

文献[15]基于信息熵理论建立了最大熵模型。在一定的限制条件下，如果这些限制无法确定唯一的系统分布，那么信息熵最大的分布是最优系统分布。最大熵模型被广泛用于自然语言处理中，文献[16]最早采用最大熵模型来处理自然语言问题。

4 词句情感分析

4.1 词级情感分析

词级情感分析首先构建领域情感词典，然后采用领域情感词典抽取出观点词并用最大熵模型预测观点词极性，最后构建词的加权无向图并用最小割模型优化极性结果。

4.1.1 领域情感词典构建

本研究以谭松波构建的数码情感语料[13]为实验基础，采用中科院 ICTCLAS 分词系统分词、专有名词识别并使用哈工大网络智能研究室情感词典过滤出情感词，得到原始领域情感词典，然后人工标注原始领域情感词典词的褒贬极性得到种子词并使用 HowNet 扩充种子词，扩充包括同义词、对义词、反义词；我们构建了一个训练词集[5]，将训练词集使用 HowNet 扩展同义词生成同义词特征并训练最大熵模型，再使用最大熵模型预测扩展词的极性并人工剔除非领域性词，得到领域情感词典。其流程图如图 1 所示。

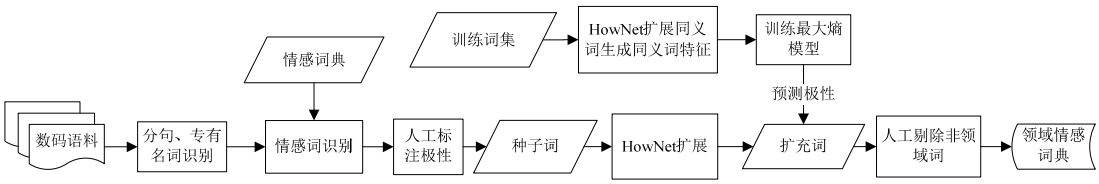

图 1 领域情感词典构建流程图

Fig.1 the Flow Chart of Building Domain Sentiment Lexicon

4.1.2 观点词抽取

首先用中科院分词系统 ICTCLAS 对测试语料分词，然后采用基于领域情感词典的方法、词性判断提取出观点词，我们假设“ ”、”“、【】、（）、{}、[ ]这些符号中间的内容不会影响句子的情感，本文将这些符号中间的内容去掉。由于通用的分词系统不能很好地把特定领域一些专用的词汇准确分词，为此我们新增加了一个专有名词词典[18]，把一些公司名字、节目名称、艺人姓名等专有名词从文件里面识别出来，提高了分词的准确度。我们认为大多数观点词都是形容词，为了简单起见，我们只对形容词进行判断，对于每个形容词我们查找它是否在该领域的观点词词典里，如果不在则忽略，如果在，则抽取出该词以及该词附近的文本做答案中的第五部分。

4.1.3 观点词情感分类

观点词的情感分类首先通过HowNet扩展训练语料生成同义词特征并训练最大熵模型；对测试语料分词并识别出语料中的专有名词、人名等以提高分词正确率，然后采用领域情感词典提取观点词并用最大熵模型预测观点词情感极性，最后建立基于词汇的加权无向图并用最小割模型优化极性结果，具体过程如图 2 所示。

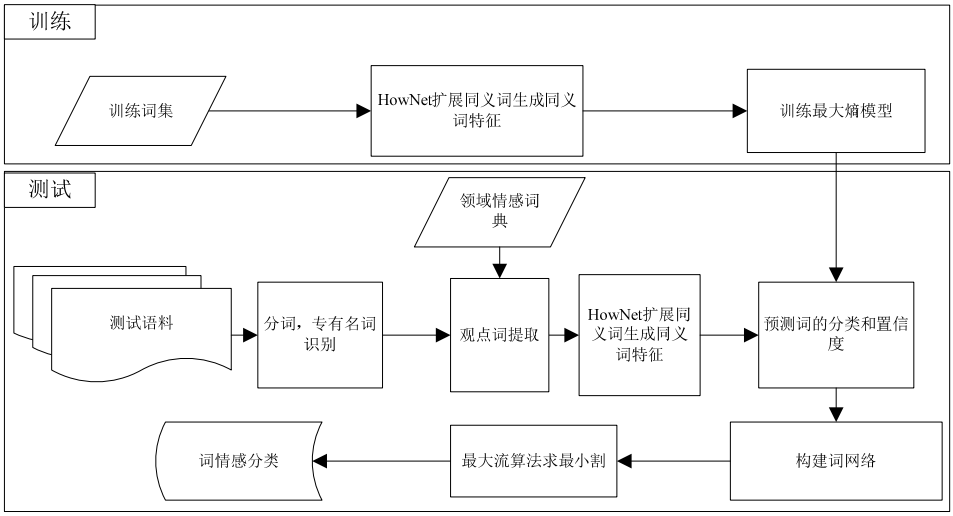

图 2 词情感分类流程图

Fig.2 the Flow Chart of Classification of Sentiment word

下面具体介绍最小割模型对结果的优化过程。

4.1.4 构建词的加权无向图

构建基于词汇语义关系的加权无向图，其中词语对应于图中的节点，边上的权重 c_i 是这两个节点(词)的同义词关系权重，权重 d_i 表示反义词关系权重，可以通过知识库得到，例如采用 WordNet 或者 HowNet 来计算词语语义关系来计算；引入两个节点：源点(s)和汇点(t)；其他节点到 s 和 t 的边上的权重 (b_i)和(a_j) 是这个节点到 s 和 t 的权重，本文通过最大熵模型预测词的褒贬极性，并将预测概率值作为该词节点到源点和汇点的权重构建加权无向图，如图 3 所示。

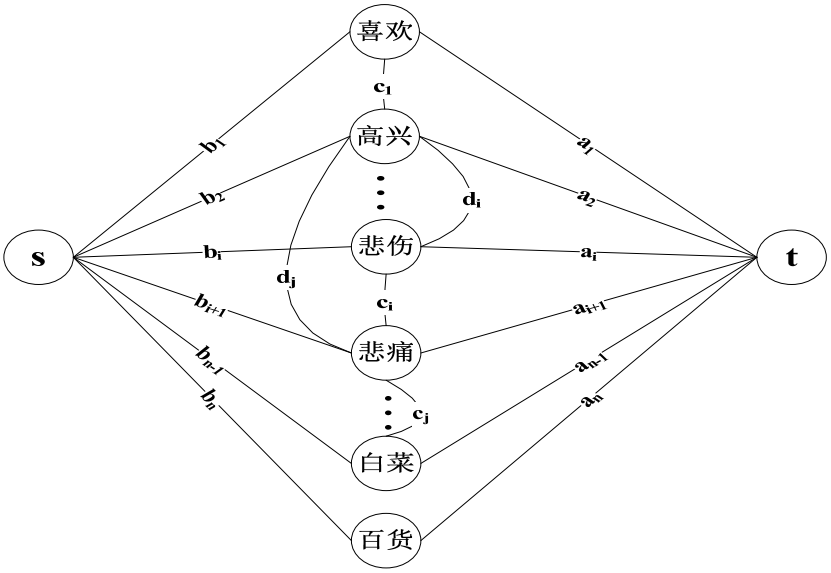

图 3 词汇加权无向图

Fig.3 A Weighted Undirected Graph of Words

其中，b_i和a_j分别代表词节点到源点和汇点的权重，本文采用最大熵模型预测词的褒贬极性，并用预测概率值作为这个权重。d_i和c_i分别代表反义词关系权重和同义词关系权重，采用 HowNet 计算这个权重。

4.1.5 基于最小割模型的词褒贬极性分析

图论中最小割的二值划分是基于相似的元素总应该会被分到一个相同集合的假设[12]；词对应于无向图的顶点，一个二值划分等价于把这个词的加权无向图划分为两个子集 S 和 T，$s \in S, t \in T$，即从这个无向图中移除了一些边，使得这个连通的词加权无向图被分成两个不连通的联通子图。那么我们希望相似的元素被分到一个相同的集合，所以最好的分割就是把相似的元素被分割到相同集合中去，也就是移去的边的权重和最小，换句话来说就是求解一个最小割问题，被移去的边的权重表示如下：

$$W(S,T) = \sum_{u \in S, v \in T} w(u,v) \qquad (1)$$

其中 $w(u,v)$ 是词语 u 和 v 之间的权重，那么问题其实就是求 $\arg\min W(S,T)$。由于词的褒贬极性判断问题是一个二值划分问题，所以可以用最小割理论来处理，其次 $W(S,T)$ 达到最小，意味着我们的词的加权无向图从源点 s 到汇点 t 的概率达到最大，也就是破坏了最小量的语义关系和最大程度地保留了分类结果，所以这个分类是一个有效的分类。通过最小割分割图 3 得到图 4，如图 4 所示。

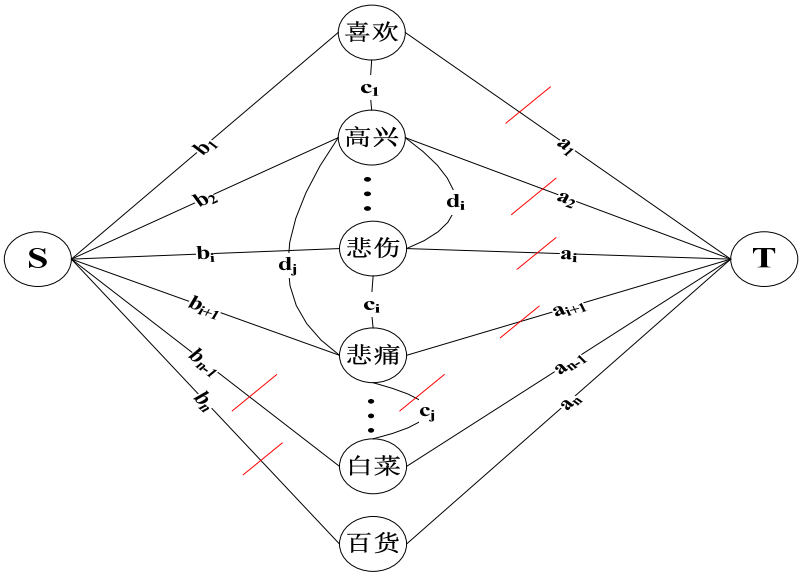

图 4 最小割图

Fig.4 A min-cut graph of networks of words

4.2 句级情感分析

本文主要采用最大熵分类的方法判断情感句类别，因为分类的方法能够融入比较多的特征，同时也可以获得比较高的准确率[3]；本文采用的语料包括 2000 篇褒义评价文章和 2000 篇贬义评价文章[13]。句级情感分析首先采用词典筛选观点句，然后将观点句切分成短句并用最大熵模型预测短句极性，最后通过短句预测长句极性。

4.2.1 观点句识别

抽样分析后发现，在整个测试语料中，观点句和非观点句的比例严重不平衡，所以将会对分类造成严重的影响；我们认为观点句里面基本上都包含了观点词，所以我们通过领域情感词典来筛掉非观点句，可以大大减少测试语料的不平衡给分类带来的影响。我们假设“ ”、”“、【】、（）、{}、[ ]这些符号中间的内容不会影响句子的情感，本文将这些符号中间的内容去掉。

4.2.2 长句切分

实验表明长句分类的效果不是特别好，实验对比了长句分类和短句分类，发现短句分类效果更好，所以本文采用短句分类来处理。一个长句包含的情感相对短句较复杂，特征也比较多且难以准确提取，分类效果比较差；短句包含的结构简单、语义容易判断、特征容易抽取。本文主要采用的分割符有逗号“，”、“：”、“；”、“。”；比如句子“这款笔记本外观造型比较时髦，但是显示器的显示效果总是感觉有点儿差，还有就是内存挺大的；所以总而言之来说还是比较实惠的”，按照我们的方法，这个句子会被分成四个子短句，分别是：

A 这款笔记本外观造型比较时髦

B 但是显示器的显示效果总是感觉有点儿差

C 还有就是内存挺大的

D 所以总而言之来说还是比较实惠的

4.2.3 特征获取

文献[5]给出了 3 个特征获取的方法，这些方法在 COAE2009 的数据上实验取得了比较好的效果。提取的特征主要包含词和词序列、情感词强度累加、句子的句型。词和词序列主要包括识别情感词，否定词和情感词周围的词序列；情感词强度累加指通过对识别的情感词的情感强度累加；句型特征包括问句、感叹句和长句。

4.2.4 长句情感倾向性计算

由于长句被分成了短句，长句情感极性由短剧情感极性决定，因此本文主要采用下面的决策思想：

（1） 如果一个长句的所有短句都是褒义，那么这个长句的倾向性就是褒义。

（2） 如果一个长句的所有短句都是贬义，那么这个长句的倾向性就是贬义。

（3） 如果一个长句里面既有褒义又有贬义，那么这个句子的倾向性是褒贬混合。

长句置信度计算采用下列方法，如果一个长句包含 n 个短句，每一个短句的置信度是 x_1、x_2、$x_3 \cdots x_n$，那么长句置信度 $p = \dfrac{(x_1 + x_2 + x_3 + \cdots + x_n)}{n}$。句级情感褒贬分析的全部具体过程如图 5 所示。

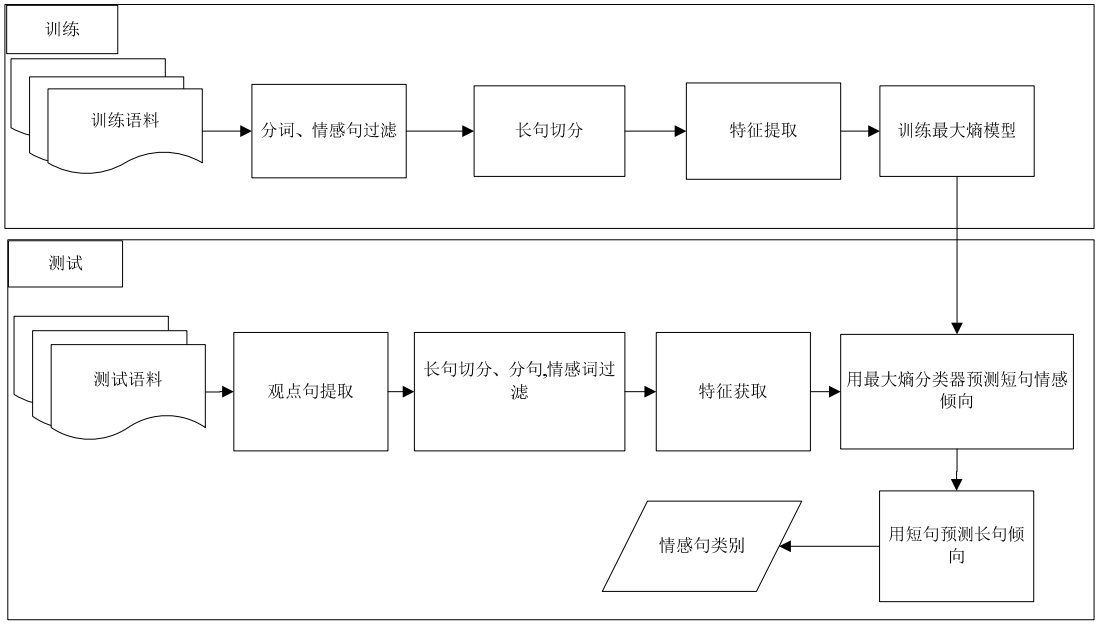

图 5 情感句褒贬极性分析流程图

Fig.5 the Flow Chart of Positive and Negative Polarities Analysis on Sentiment Sentences

5 实验与分析

词级情感分析实验结果如表 1 所示。

表 1 词级褒贬极性分析

Tab.1 Positive and Negative Polarities Analysis on Words

	Precision	P@1000	Recall	F1	Raccuracy
HITWI_D	0.3927	0.577	0.0956	**0.1538**	0.0956
HITWI_E	0.4528	0.596	0.1009	**0.1651**	0.1009
HITWI_F	0.6125	0.641	0.0936	**0.1624**	0.0936
Median	0.3430044	0.57126	0.0947444	**0.1475933**	0.0947444
Best	0.6125	0.674	0.1194	**0.1833**	0.1194

从表 1 的实验结果可以看出，电子数码领域和娱乐领域的准确度、召回率、Raccuracy 都比平均值高且接近最大值，F 值接近最大值，这主要是因为这两个领域的情感倾向容易判断，观点容易抽取。而在财经领域召回率和 Raccuracy 都比平均值低，原因是这个领域的观点词情感比较模糊以及领域情感词典不是很丰富所致。

句级情感分析实验结果如表 2 所示。

表 2 句级褒贬极性分析

Tab.2 Positive and Negative Polarities Analysis on Sentences

	Precision	P@1000	Recall	F1	Raccuracy
HITWI_D	0.49268	0.608	0.391297	**0.436174**	0.391297
HITWI_E	0.303493	0.278	0.2262	**0.259207**	0.2262
HITWI_F	0.204598	0.089	0.172147	**0.186975**	0.172147
Median	0.240815	0.290183	0.3979462	**0.27632415**	0.2554445
Best	0.729751	0.8	0.798097	**0.693304**	0.660324

从表 2 的实验结果可以看出，电子领域的准确度、P@1000、F 和 Raccuracy 值都大于平均值，主要是电子这个领域的特征比较明显容易提取，而召唤率小于平均值，这主要是领域情感词典不够丰富所致。娱乐和财经这两个领域的召回率和 Raccuracy 都小于平均值，原因是领域词典内容不够丰富、情感特这难以准确提取、分词也不够准确致使召回率不够高。

6 结果与展望

本文采用了最大熵模型和最小割模型来处理词和句子的褒贬极性分析问题。词级情感分析通过最大熵分类模型预测候选情感词的褒贬极性，然后构建基于词的加权无向图并采用最小割模型优化极性结果，该方法取得比较好的实验结果。句级情感分析将观点句切分成短句并提取短句的特征并使用最大熵模型来预测短句的褒贬极性，最后用短句情感极性预测长句的情感褒贬极性。下一步工作将丰富领域情感词典以及如何准确的提取句子的特征。

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] 董喜双，关毅，李本阳，陈志杰，李生. 基于最大熵模型的中文词与句情感分析研究. 第二届中文情感倾向性分析会议，2009：1-8

[6] J. Kamps, M. Marx, R. J. Mokken and M. D. Rijke. Using WordNet to measure semantic orientation ofadjectives. In: Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, Lisbon, 2004, 1115-1118.

[7] 朱嫣岚,闵锦, 周雅倩, 黄萱菁, 吴立德. 基于HowNet 的词汇语义倾向计算. 中文信息学报，2006,20(1): 14-20.

[8] 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.

[9] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems. 2003. 21 (4): 315 – 346.

[10] H. Yu and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences [A], In: M. Collins and M. Steedman(eds). Sapporo. Japan: 2003.129-136

[11] Thomas H. Cormen Charles E. lerserson Ronald L. Rivest Clifford Stein. Introduction of Algorithms. Second Edition. China Machine Press.2006.396-419

[12] Fangzhong Su; Katja Markert. Subjectivity Recognition on Word Senses via Semi-supervised Mincuts. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 1–9, Boulder, Colorado, June 2009.

[13] 情感语料：http://www.searchforum.org.cn/tansongbo/corpus-senti.htm

[14] 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.

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

[16] 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.

[17] 何婷婷，闻彬，宋乐. 词语情感倾向性识别及观点抽取研究. 第一届中文情感倾向性分析会议，2008：1-8

[18] 专有名词词表：http://list.video.baidu.com/manhotlist/