# Description: 东大校徽1111

学校代码： 10286

分 类 号： TP311

密 级： 公开

U D C : 004.4

学 号： 131508

**基于依存句法分析的中文评价对象抽取**

**和情感倾向性分析**

研究生姓名**：贾治中**

导师姓名**：高志强 教授**

|  |  |
| --- | --- |
| 申请学位类别 工学硕士 | 学位授予单位 东 南 大 学 |
| 一级学科名称 计算机科学与技术 | 论文答辩日期 2016 年 6 月 5 日 |
| 二级学科名称 计算机软件与理论 | 学位授予日期 2016 年 月 日 |
| 答辩委员会主席 | 评 阅 人 匿名评阅人 |
|  | 匿名评阅人 |

2016 年 5 月30日

**CHINESE OPINION TARGET EXTRACTION AND ORIENTATION ANALYSIS BASED ON SYNTACTIC DEPENDENCIES**

A Thesis Submitted to

Southeast University

For the Degree of Master of Engineering

BY

JIA Zhizhong

Supervised by

Professor Gao Zhiqiang

School of Computer Science and Engineering

Southeast University

June 2016

**东 南 大 学 学 位 论 文 独 创 性 声 明**

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名： 日期：

**东 南 大 学 学 位 论 文 使 用 授 权 声 明**

东南大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学研究生院办理。

论文作者签名： 导师签名： 日期：

# 摘要

随着互联网的发展，包含有观点和评论的文本大量涌现。人们一方面在浏览别人发表的评论，一方面也在不停地分享自己对于某些人或物的观点和情感。群体的观点发掘能更好地为决策者提供更好的决策依据，所以情感分析无论对于经济发展、政治决策还是社会个体的行为都有着极其重要的指引作用。目前中文评价对象抽取和情感倾向性分析的效果还很不理想，本文在前人研究的基础上，进行了一些有益的探索

本文主要利用语句中各个成分之间的依存关系，提取出中文短文本的评价对象和评价词，并分析出评价对象的情感倾向性。论文的主要工作如下：

1. 在现有词典的基础上，构建用于情感分析的情感词典，主要包括：正面情绪词典、反面情绪词典、正面评价词典、反面评价词典、观点引述词典、虚拟语气词典、转折词典、名词性情感词典。
2. 由于传统的基于依存句法分析的评价对象抽取和倾向性分析效果不理想，本文在依存句法分析的基础上，利用语义角色标注，并添加了一系列的语义规则进行情感分析。实验结果表明在NLPCC2013的微博评测语料上，添加语义规则的基于依存句法分析的方法，能够显著提高识别效果。
3. 结合词义和词语相似度计算算法，提出一种评价对象搜索方法，用于改善在只抽取出代词或句法关系中无评价对象的情况下，搜索上下文中真正的评价对象的精确度。

本文实验表明：通过完善情感词库和添加语义知识，能够较为显著地提高评价对象抽取和情感倾向性分析效果。

**关键词：**依存句法分析，评价对象抽取，倾向性分析，语义规则

Abstract

With the development of the Internet, a large number of text with opinions and comments appear on the Internet. People browse in other people's comments on the one hand, on the other hand also share their views on some people or things continually. Mining the group view can provide better decision-making basis for policymakers, so sentiment analysis plays an important guidance role for economic development, political decisions and behaviors of the individual or social behavior. However, in the domain of the Chinese, the performance of opinion target extraction and orientation analysis is not very ideal. Based on the research results in this area, some useful exploration is carried out in this thesis.

Syntactic dependencies of components of the sentence are used to extract opinion targets and opinions from the Chinese short text, then sentiment orientation is analyzed in this thesis. Our major work is as follows:

1. On the basis of the existing dictionaries, sentiment dictionaries are built for sentiment analysis such as Positive Emotion Dictionary, Negative Emotion Dictionary, Positive Opinion Dictionary, Negative Opinion Dictionary, Opinion Statement Dictionary, Subjunctive Mood Dictionary, Adversative Dictionary and Noun Sentiment Dictionary.
2. As a result of non-ideal performance of traditional opinion target extraction and orientation analysis based on syntactic dependencies, semantic role annotation and some semantic rules are introduced into sentiment analysis. The experimental results show that in the NLPCC 2013 micro-blog evaluation corpus, the method with semantic rules improve the performance significantly.
3. With combination of lexical meaning and word similarity computation algorithm, a opinion target search method is proposed, which is used to improve the precision of real opinion target search when the system only extract pronouns or opinion target is not found in syntactic dependencies.

The experimental results show that through improving the sentiment word library and adding the semantic knowledge, the performance of opinion target extraction and orientation analysis can be improved significantly.

**Key Words:** Syntactic Dependency, Opinion Target Extraction, Orientation Analysis, Semantic Rule

# 目录

[摘要 I](#_Toc358238750)

[Abstract II](#_Toc358238751)

[目录 III](#_Toc358238752)

[缩略词表 V](#_Toc358238753)

[第一章 绪论 1](#_Toc358238754)

[1.1 研究背景 1](#_Toc358238755)

[1.2 命名实体识别的基本概念 1](#_Toc358238756)

[1.3 研究现状 2](#_Toc358238757)

[1.4 本文的研究目的和主要内容 3](#_Toc358238758)

[第二章 相关研究 4](#_Toc358238759)

[2.1 中文分词 4](#_Toc358238760)

[2.2 中文命名实体识别的难点 5](#_Toc358238761)

[2.2.1. 中文人名的特点 5](#_Toc358238762)

[2.2.2. 中文地名的特点 6](#_Toc358238763)

[2.2.3. 中文组织名的特点 6](#_Toc358238764)

[2.3 命名实体识别的方法 6](#_Toc358238765)

[2.3.1. 基于规则的方法 7](#_Toc358238766)

[2.3.2. 基于统计的方法 7](#_Toc358238767)

[2.4 本章小结 8](#_Toc358238768)

[第三章 条件随机场模型 9](#_Toc358238769)

[3.1 CRFs模型 9](#_Toc358238770)

[3.2 线性链CRFs模型 10](#_Toc358238771)

[3.3 CRFs特征函数的选择 11](#_Toc358238772)

[3.4 CRFs模型的参数估计 11](#_Toc358238773)

[3.5 迭代缩放算法 12](#_Toc358238774)

[3.5.1. 基本原理和过程 12](#_Toc358238775)

[3.5.2. GIS算法 13](#_Toc358238776)

[3.5.3. IIS算法 14](#_Toc358238777)

[3.6 概率计算 14](#_Toc358238778)

[3.7 本章小结 15](#_Toc358238779)

[第四章 基于条件随机场的中文命名实体识别系统 16](#_Toc358238780)

[4.1 系统框架设计 16](#_Toc358238781)

[4.2 系统模块 17](#_Toc358238782)

[4.2.1. 词典生成模块 17](#_Toc358238783)

[4.2.2. 预处理模块 18](#_Toc358238784)

[4.2.3. 实验模块 18](#_Toc358238785)

[4.2.4. 实际应用模块 21](#_Toc358238786)

[4.3 本章小结 21](#_Toc358238787)

[第五章 实验与分析 22](#_Toc358238788)

[5.1 实验条件 22](#_Toc358238789)

[5.2 性能指标 22](#_Toc358238790)

[5.3 特征阈值筛选实验 22](#_Toc358238791)

[5.3.1. 实验原理 22](#_Toc358238792)

[5.3.2. 实验内容与结论 23](#_Toc358238793)

[5.4 中文命名实体识别性能研究实验 25](#_Toc358238794)

[5.4.1. 识别性能与训练语料规模的关系 26](#_Toc358238795)

[5.4.2. 识别性能与语料标注粒度的关系 29](#_Toc358238796)

[5.4.3. 识别性能与模型类型的关系 30](#_Toc358238797)

[5.5 本章小结 32](#_Toc358238798)

[第六章 总结和展望 33](#_Toc358238799)

[6.1 工作总结 33](#_Toc358238800)

[6.2 未来展望 33](#_Toc358238801)

[致谢 35](#_Toc358238802)

[参考文献 36](#_Toc358238803)

[附录 39](#_Toc358238804)

# 缩略词表

|  |  |  |
| --- | --- | --- |
| **缩略词** | **英文全称** | **中文全称** |
| CRF/CRFs | Conditional Random Fields | 条件随机场 |
| NER | Named Entity Recognition | 命名实体识别 |
| MUC-6 | The Sixth Message Understanding Conference | 第六届信息理解会议 |
| NE | Named Entity | 命名实体 |
| MEMs | Maximum Entropy Models | 最大熵模型 |
| MEMMs | Maximum Entropy Markov Models | 最大熵马尔可夫模型 |
| HMMs | Hidden Markov Models | 隐马尔可夫模型 |
| MRFs | Markov Random Fields | 马尔可夫随机场 |
| MLE | Maximum Likelihood Estimation | 极大似然估计 |
| MAP | Maximum a Posteriori | 最大后验估计 |
| BE | Bayes Estimation | 贝叶斯估计 |
| GIS | Generalized Iterative Scaling | 广义迭代收缩（算法） |
| IIS | Improved Iterative Scaling | 改进迭代收缩（算法） |
| MI | Mutual Information | 互信息 |
| CE | Conditional Entropy | 条件熵 |
| PMI | Point-wise Mutual Information | 点式互信息 |

# 绪论

本章主要阐述情感分析的研究背景、情感分析的基本概念、中文评价对象抽取和倾向性分析的难点，最后介绍本文的研究目的和主要研究内容。

## 研究背景

观点或情感对于人类的一切活动和行为起着决定性的作用。我们的信仰、对现实的认知以及我们所作出的任何决定，从某种程度上，都取决于自身以及他人如何看待和评价所在的世界。正因为如此，每当我们需要做决定的时候，都会习惯性地先参考别人的观点。这对于个体不仅如此，对于大多数群体和组织而言更是如此。在现实生活中，个体消费者在购买某产品前往往想知道该产品使用者的评价；在选举中，选民也希望在投票前知道其他人对于候选者的观点；商业组织必须清楚地了解顾客或公众对于他们的产品或服务的看法，才能更好地做出商业计划和营销决策。

随着互联网的飞速发展，人们一方面在浏览别人发表的评论，一方面也在不停地分享自己对于某些人或物的观点和情感。尤其是微博、论坛、购物平台、新闻网站等公众媒体的快速发展，评论性信息迅速膨胀，仅靠人工的方法难以处理海量信息的收集、分类和分析。于此同时，仅仅获取个体的评价信息是没有意义的，群体的观点发掘才能更好地为决策者提供更好的决策依据。所以情感分析无论对于经济发展、政治决策还是社会个体的行为都有着极其重要的指引作用。因此，迫切需要计算机帮助用户快速获取和整理这些相关评价信息，情感分析（Sentiment Analysis）技术应运而生。观点以及与之相关联的概念包括情感、评价、态度和情绪等，都是情感分析和观点挖掘的主要研究对象。自从2000年开始，情感分析已经逐步成为自然语言处理领域中最为活跃的研究领域之一[1]。在数据挖掘、网络挖掘、文本挖掘等领域，它同样被广泛研究。由于情感分析对于商业和社会的重要性，它已经从计算机科学领域扩展到管理科学和社会科学领域。

## 情感分析的基本概念

情感分析，又叫做观点挖掘，主要研究人们对于实体（如产品、服务、组织、个体、议题、事件、话题和它们的属性）的观点、情感、评价、评估、态度和情绪。根据任务的侧重，情感分析还有其他不同的名称，例如观点抽取、情感抽取、主观性分析、影响分析、情绪分析、回顾分析等。

按照处理文本的粒度不同，情感分析主要分为文档级、句子级、实体和特征级。按照处理文本的类别不同，可分为基于新闻评论的情感分析和基于产品评论的情感分析。

目前情感分析有六个比较关键的任务，分别是实体抽取和分类，方面（特征）抽取和分类，观点持有者抽取和分类，时间抽取和标准化，实体方面情感分类，观点元组生成。基于这个框架的情感分析常常被称为基于方面或者基于特征的情感分析[2][3]。

基于方面的情感分析又可分为两个子任务，分别是评价对象抽取和情感倾向性分析。文档级和句子级的情感分析在现实使用中往往不足以满足需求，因为它们无法识别出每个实体或属性及其对应的情感倾向性。即使我们假设每个文档评估一个单独的实体，一个含有关于某个实体正明评价的文档并不意味着作者对该实体的所有方面或属性都持有正面评价。对于更加完整的情感分析，我们需要抽取出所有的评价对象并判断对于该对象的评价是正明还是负面的。

评价对象抽取（Opinion Target Extraction）是文本情感分析中的重要研究任务之一，评价对象是指一段文本中所讨论或者描述的主题、实体或者实体的某些方面和属性等，具体表现为评论文本中修饰词语或评论词语所修饰的对象，如新闻评论中的某个事件/话题或者商品评论中的某个产品的组件（如“手机电池”）等。本文研究主要针对中文短文本中显式的评价对象进行抽取，而非隐藏在文本背后的评价对象。

倾向性分析（Orientation Analysis），又可以称作情感分类，指的是抽取出文本中描述评价对象的情感词、极性词，对其做情感倾向性分析（正面评价或负面评价），这里的情感词特指的是具有情感倾向性的词语。对于有些具体场景，比如购物网站上的商品评论，其评价对象可能就已经给定了，此时我们就不必再做评价对象的抽取工作，只需要判断该评价对象的情感倾向性。

## 中文评价对象抽取和倾向性分析的难点

由于中文词语语义的多样性，在不同语境下可能会有不同的情感表述效果，加上网络用语的随意性，每个人表述情感的方式个性化，中文评价对象抽取和倾向性分析在实际应用中的效果很不理想，尤其是微博。目前，中文评价对象抽取和倾向性分析的难点主要有以下几点：

1. 同一句中，评价对象与评价词距离较远，同时，二者之间存在其他干扰性的评 价对象和评价词，导致二者无法被正确地关联和抽取出来。
2. 情感词存在，但是句中无明显的评价对象或只有相应的代词存在时，即评价对 象缺失，如何根据上下文正确找到情感词关联的评价对象。
3. 句子含有情感词，但是并不包含情感倾向性，可能只是表达某种心理活动或非 评价性的想法、愿望等。如，我希望小明能取得好成绩！
4. 句子中无明显的情感词存在，但依旧表达情感倾向性。如，他的店铺很有人气！
5. 转折词、否定词、情态动词等表达方式及其位置给倾向性分析带来不确定性。
6. 一些与情感词关联的名词，干扰了真正评价对象的抽取，例如，查韦斯病逝， 又一个伟大的英雄离开了。该句中可能抽取出来的评价对象和评价词是[英雄， 伟大]。
7. 情感分析基础工作的错误带来的干扰，例如中文分词、句法分析、语义分析过 程中的标注错误可能给后续的分析带来影响，尤其是后两者的效果还不理想。
8. 英文情感词典已经有非常完备的SentiWordNet，而中文情感词典资源却质量不 高，不够细 致，前期的基础不牢固，要得到准确的效果就不容易。
9. 大多数语料库都是基于商品评论的，如电脑、酒店等，标注细致的新闻评论、 微博等语料库太少，而这些正是中文评论语境最为丰富的领域。
10. 新词不断产生，基本常识、上下文知识和各式各样的实体词，以及中文句式类 型的复杂性都对中文评价对象的抽取带来很大的影响。

## 本文的研究目的和主要内容

使用依存句法分析能够很好地揭示出评论语句中各成分之间的修饰关系或依赖关系，通常观点词都是和评价对象相互关联的，而多数情况下观点词可以根据词典或分词的标注结果而事先得知，所以利用依存关系分析可以通过观点词抽取出一些不频繁出现的评价对象。同时，通过依存句法关系分析，可以获得观点词的上下文特征（如转折关系、递进关系等），进而有利于分析出观点词的情感倾向。所以，基于句法依存分析的方法能够很好地对评价对象和观点词进行抽取和分析。然而，目前对于中文文本进行评价对象抽取和倾向性分析的精度还不是很好，因此，本文希望通过丰富情感词典、添加语义知识、改进评价对象搜索方法来改善目前基于依存句法分析的中文评价对象抽取和倾向性分析的性能。

由于中文本身具有的复杂性与特殊性，决定了中文评价对象的抽取和倾向性分析所面临的困难更大。本文将对此展开研究，主要内容如下：

1. 构建用于情感分析的情感词典，主要包括：正面情绪词典、反面情绪词典、正 面评价词典、反面评价词典、观点引述词典、虚拟语气词典、转折词典、名词 性情感词典。
2. 针对实验所使用微博评测语料的语言成分的多样性，进行了大量的预处理工 作，包括去除无关字符、链接等，同时对非核心句子成分进行过滤，进而缩短 评价对象和评价词的关联距离。
3. 由于传统的基于依存句法分析的评价对象抽取和倾向性分析效果不理想，本文 在依存句法分析的基础上，利用语义角色标注，并添加了一系列的语义规则进 行情感分析。
4. 结合词义和词语相似度计算算法，提出一种评价对象搜索方法，用于改善在只 抽取出代词或句法关系中无评价对象的情况下，搜索上下文中真正的评价对象 的精确度。

## 论文结构

本文的后续章节组织如下：

第二章介绍中文评价对象抽取和倾向性分析的研究现状，并加以比较。

第三章构建用于情感分析的情感词典，介绍各个词典的用途。

第四章详细介绍基于依存句法分析的中文评价对象抽取和倾向性分析的实验设计与分析。首先介绍本文使用的基于图的依存句法分析模型，然后对本文使用的评价对象和情感词抽取规则及相应的语义规则进行详细介绍。最后设计并实现中文评价对象抽取和倾向性分析系统，并在NLPCC 2013微博情感要素抽取评测语料上进行实验，给出实验结果和分析。

第五章介绍评价对象搜索方法，并进行相应的实验和分析，验证该方法的有效性。

第六章对全文的工作进行总结和展望。

# 相关研究

尽管语言学和自认语言处理有着很长的历史，但在2000年以前，关于人们观点和情感的研究非常少。随着互联网的飞速发展，这个领域的研究开始变得活跃起来。出现这种情况有如下几个原因：第一，情感分析有着广阔的运用领域，无论是商业还是政治决策。第二，情感分析领域有很多具有挑战性的问题并且以前从未被研究过。第三，这是人类历史上第一次出现这样的情况，通过网络上的公众媒体，我们能够拥有如此海量的情感数据。正因为此，情感分析得到了飞速的发展。

## 相关评测和情感资源

情感分析最先引起了国际文本检索会议TREC1(Text Retrieval Evaluation Conference)的关注，该会议是由美国国家标准技术研究所（NIST）和美国国防部（DOD）联合主办的，自从1992年开始就作为TIPSTER Text Program下的附属计划，每年召开一次。该会议主要关注如下几个方向：问答系统、特定领域搜索、传统Web检索等。其目标是：1）鼓励基于大型语料库上的信息检索研究；2）通过创建一个开放的平台来促进工业界、学术界和政府间的交流与合作；3）通过验证优秀的检索方法和技术对于解决现实问题的效果，加速技术从研究领域转向商业产品；4）提供更为先进有效的评估技术供工业界和学术界使用。情感分析相关的任务从2006年开始加入到会议评测中，并且微博检索任务TREC Blog Track[4-5]是TREC最先关注的情感分析任务，主要针对英文文本中的情感词或观点信息进行检索。

NTCIR（NACSIS Test Collection for IR）2计划是由日本学术振兴会（JSPS）和日本国家科学咨询系统中心（NACSIS）自1997年开始联合主办的，每年举行一次，并拥有中、英、韩、日四种语言的标准语料库，主旨是为了加强信息索引、问答系统、文本摘要、信息抽取等信息获取技术的研究。其目标与TREC很相近，都是为了促进相应技术的发展，提供各个领域的交流平台，研究更好的评估方法以及构建和提供用于信息获取领域实验的大规模数据集。NICIR的情感分析评测同样出现在2006年，主要针对文本的情感分类以及观点持有者的抽取等[6-7]。

随着中文文本的情感分析得到了越来越多的学者和研究机构的关注，为了推动中文情感分析技术的发展，国内第一个情感分析方面的评测COAE（Chinese Opinion Analysis Evaluation）[8]于2008年举办第一届。COAE目的在于推动中文情感分析理论和技术的研究和应用，同时建立相关的分析语料库。COAE 共设置6个任务，可分为3个方面：一是中文评价词语的识别和分析，属于词语级的情感分析评测；二是中文文本倾向性相关要素的抽取，主要是抽取句子中的评价对象，以及对于其观点的倾向性判别，属于语句级的情感分析评测；三是中文文本主客观性及倾向性的判别，属于篇章级的情感分析评测。

除了COAE提供了产品类的评价语料库，中国科学院计算技术研究所的谭松波博士提供的较大规模的中文酒店评论语料，约有10000篇，并标注了褒贬类别，可以为中文的篇章级的情感分类提供一定的帮助。

1. http://trec.nist.gov
2. http://ntcir.nii.ac.jp

另外自然语言处理与中文计算会议NLPCC（The conference on Natural Language Processing and Chinese Computing ）自2012年开始也设置了一些热门主题的研究任务，包

括情感分析、信息关联和组织、问答系统等。其中对情感分析提供了中文微博评测语料和三个任务，分别是微博情感分类、情感句识别和分类、情感要素抽取。

中文的评价词词典资源有NTU评价词词典（繁体中文）和HowNet评价词词典。NTU 评价词词典由台湾大学收集，含有2812个褒义词与8276个贬义词。HowNet评价词词典包含9193个中文评价词语/短语，9142个英文评价词语或短语，并被分为褒贬两类。而且该词典提供了评价短语，为情感分析提供了更丰富的情感资源。

## 评价对象抽取的研究现状

情感分析问题的一些特征可以用来帮助评价对象的抽取[1]。其中一个关键特征就是一个观点往往都会有一个对应的评价对象，这个对象往往是句中的某一属性或话题。所以识别出句子中的每一个观点表达式和它的评价对象非常重要。同时我们也需要注意到，一些观点表达式可能扮演两个角色，一个是表示正面或负面情感的情感词，另一个是隐式的评价对象。例如，“这本书很便宜”。该句的评价词时“便宜”，对应的隐式评价对象是“价格”。处理该类型的文本时，需要构建一个评价词和隐式评价对象的对应词典。本文主要研究句中出现的显式评价对象的抽取工作。

目前针对显式评价对象的抽取主要有四种方法，分别是基于寻找频繁出现的名词和名词词组的抽取方法，利用观点和评价对象的关系进行抽取的方法，使用有监督学习进行抽取的方法以及使用话题模型进行抽取的方法。

现在既有的研究主要是针对网络评论，其评论格式主要有下面两种：

1. 赞成意见、反对意见和详细评论：评论者首先会分别描述简要的优缺点，然后 再写一个完整的评论内容。
2. 自由格式：评论者可以自由发表观点。

格式（1）主要来自于商品、酒店评论等，相对于格式（2）要容易处理。Liu等[9]在2005年提出一种基于序列学习的方法抽取评价对象，主要利用赞成和反对意见的文本特征，如它们通常比价简短，句块由标点符号、连词等隔开，每个句块最多含有一个评价对象等，这些特征都较为有效地提高了抽取精度。所以目前针对商品评论的情感分析效果要优于微博评论。本文主要研究自由格式文本的评价对象抽取。

### **基于寻找频繁出现的名词和名词词组的抽取方法**

这种方法在给定的领域中从大规模的评论数据中寻找那些名词或名词性短语的显式评价对象表达式。Hu等[10]在2004年使用了一种数据挖掘算法，通过词性标注器识别出名词和名词性短语。计算这些潜在评价对象出现的频率，频繁出现达到阈值的对象才被保留下来。这种方法的依据是人们评论某个实体的不同属性时，所用词汇往往会有一定的聚集性。所以，那些频繁用到的名词往往是真正重要的评价对象。同样，那些非频繁的名词更可能是非评价对象或不重要的评价对象。尽管这种方法很简单，但是非常有效。有一些商业公司已经使用这种方法，效果也有所提高。

Hu的算法在Popescu等[11] 2005年的研究中得到了改进。他们的算法计算每个被发现的名词或名词短语与鉴别词（已经和某些实体类别关联的词）的点互信息（Pointwise Mutual Information, PMI）的评分来试图移除那些可能不是评价对象的名词或名词短语。以英文为例，对于汽车类的属性识别特征（或称为识别器）可能有：“of car”，“car has”，“car comes with”等，这些可以通过网络搜索来找到文本中的评价属性。PMI计算方式可以表示成下式：

（1）

其中，a表示使用频率搜索的方法识别出的潜在评价属性，d表示属性识别特征。网络搜索用于查找单独的评价属性和识别特征出现的次数和它们共现的次数。如果潜在评价对象的PMI值过低，则它不大可能是一个产品的某个组件或属性，因为a和d没有频繁共现。

Blair-Goldensohn等[12]在2008年通过只考虑含有情感词或某种句法模式发现情感倾向的语句，进一步改进了基于名词和名词性短语频率的抽取方法。Long等[13]通过基于频率和信息距离的方法抽取出评价对象。他们的方法首先使用基于频率的方法找到核心评价对象词，然后使用信息距离[14]来寻找其他与评价对象相关联的词语。

### **利用观点和评价对象的关系进行抽取的方法**

既然观点总是有对象的，所以他们总是相关联的。通常观点词都是很容易发现的，观点和评价对象之间的关系可以用来抽取评价对象。这种方法被Hu等[10]用来抽取非频繁出现的评价对象，他们没有使用任何分析器，当句子中不含有频繁出现的评价对象但含有一些情感词，那么离每个情感词最近的名词或名词短语被抽取出来。邻近评估函数的计算依赖于情感词和其所修饰的名词或名词短语的依存关系。Blair-Goldensohn等[12]基于情感模板的方法也是与此类似。这种基于关系的方法对于从含有观点的句子中发现重要的评价对象或主题都非常有效，因为无任何观点或情感词修饰的对象往往都不那么重要。

Zhuang等[15]在2006年使用句法依存关系分析器来抽取出依存关系，用于评价对象的抽取。Somasundaran[16]，Kobayashi[17]等人也相继使用类似的方法。Qiu等[18]在2011年进一步利用依存关系的思想融入双向传播的方法，同时抽取评价对象和情感词。Wu[19]等使用一个短语依存关系分析器从候选评价对象中抽取名词短语和动词短语，然后系统过滤掉那些非评价对象。目前利用依存关系进行抽取的方法已经被很多研究者运用到不同的研究任务中去[16]。

### **使用有监督学习进行抽取的方法**

评价对象抽取可以被看做一般信息抽取问题的特例。在过去的信息抽取领域，很多基于有监督学习的算法被提出。占主导地位的方法是基于序列标注的方法，因为这些都是有监督的学习方法，所以需要大量的标注数据进行训练。目前的序列标注学习方法有隐马尔科夫模型（Hidden Markov Models, HMM）[21]和条件随机场（Conditional Random Fields, CRF）[22]。Jin等[23]在2009年使用了词汇化隐马尔科夫模型来学习用于抽取评价对象和观点表达式的模板。Jakob等[24]在2010年利用词、词性标注、句法依存关系、词距离等特征，使用条件随机场在不同领域的评价语料上训练一个领域独立的评价对象抽取模型。Li等[25]融合了两种CRF的变种（Skip-CRF和Tree-CRF）来抽取评价对象和观点。不同于普通的链式CRF只能利用词序列进行学习，变种CRF可以进一步利用文本的结构特征。使用条件随机场方法的研究工作还有很多。

Kobayash等[26]在2007年提出的方法首先使用依存关系分析树寻找到候选的评价对象和观点词对，然后利用树结构分类算法学习和划分之前找到的评价对象和观点词对。Yu等[27]在2011年使用一个单类SVM[28]来标注某些是评价对象的正例。Ghani等[29]使用传统的有监督学习和半监督学习进行评价对象抽取。Kovelamudi等[30]利用维基百科上相关的信息进行有监督的学习和抽取。

### **使用话题模型进行抽取的方法**

近年来基于统计的话题模型开始成为从大量文本文档中抽取主题的一个主流方法。话题模型是一种无监督的学习方法，它假设每个文档是由多个话题混合组成，每个话题又是一个在词之上的概率分布。一个话题模型也就是一个文档的概率生成模型，其输出是一系列词的集合，每个集合构成了一个话题，同时也是文档集合中所有词的概率分布。目前有两种主要的基本模型：pLSA(Probabilistic Latent Semantic Analysis)[31] 和 LDA(Latent Dirichlet allocation)[32] 。

话题模型是一种基于贝叶斯网络（Bayesian Networks, BN）的图模型。话题模型尽管是用来对文本集合进行建模和话题抽取，但同时也可以对许多其他种类的信息进行建模。例如，在情感分析的上下文，基于观点词与评价对象存在对应关系，可以设计一个联合模型同时对观点词和话题进行建模。对于使用话题模型抽取评价对象仍然有一些不同之处，话题可以同时覆盖评价对象和评价词，对于情感分析而言，二者则需要分开。这种分开处理可以通过扩展基本模型（如LDA）来对评价对象和评价词联合建模。

Mei等[33]在2007年提出了一种联合模型用于情感分析。他们构建了一个评价对象和情感的混合模型（基于话题（评价对象）模型、正面情感模型和负面情感模型）。他们的方法主要是基于pLSA，其他大多数研究所使用的话题模型实施基于LDA的。

Titov等[34]在2008年提出像LDA的全局话题模型或许并不适合用于评价对象抽取。原因是LDA是利用文档间词共现和话题分布的不同来识别话题和话题间的词概率分布。然而，特定产品的评论所涉及的评价实体是相同的，这意味着全局话题模型将失效。Titov提出了多颗粒的话题模型，全局模型找到评价实体的同时，局部的模型找到实体的属性或某些方面（评价对象）。在这里，每个找到的评价对象都是一个一元语言模型。不同的词若表述相同或关联的事物则会自动聚合在同一个对象之下。然而，这种技术并没有分开评价对象和评价词。

Brody等[35]在2010年提出的方法首先利用话题模型识别评价对象，然后识别和评价对象相关的观点词（只考虑形容词）。Li等[36]在2010年提出两种联合模型（Sentiment-LDA和Depen-LDA），用来抽取带有正面和负面情感的评价对象。模型不是单独地寻找评价对象，而是将评价对象和观点词一并抽取。Mukherjee等[37]在2012年提出一种半监督的联合模型，模型允许用户提供评价对象的种子用来推理生成符合用户需求的评价对象分布。

尽管话题模型是一种可以扩展处理多种信息的概率推理模型，但它有一些缺点限制了它在情感分析应用中的实际使用。一个主要问题就是为了实现较理想的效果需要大量的数据和相当多的参数调优。更为糟糕的是，大多数话题模型使用了吉布斯采样（Markov chain Monte Carlo sampling, MCMC的一个特例），它会在每次运行时产生不同的结果。话题模型寻找频繁出现的评价对象并不难，难的是寻找那些局部的或全局非频繁的评价对象。那些局部频繁的评价对象往往在实际应用是最为重要的，因为它们往往是用户感兴趣的实体的某个方面或属性。

总之，话题模型是一个灵活而强大的建模工具，但对于现实情感分析而言，仅仅使用现有的话题模型很难得到理想的结果。一个有希望的研究方向是更多地融合自然语言和领域知识到模型中，这方面的工作已经有所进展[38-41]。

## 情感倾向性分析的研究现状

# 基于条件随机场的中文命名实体识别系统

## 系统框架设计

# 实验与分析

## 实验条件

实验主机配置：主频2.13GHz，内存2GB，操作系统Windows 7。

编程语言：C++。

使用工具：Visual Studio 2010，CRF++ 0.58，Ultra Edit。

实验平台：CNER Experimental Platform。

实验语料：北大富士通人民日报1998年1月份的语料库 (共计2,305,896字)，平均每100个字包含命名实体1.192个(不记数词、时间词)。

## 性能指标

对于中文命名实体识别，本文主要使用三个性能指标进行评估，分别为：准确率p、召回率r和F1值。事实上，很多情况下，准确率与召回率是相互制约、相互影响的，因为召回率的提高有时候会同时引起准确率的下降，反之亦然。为避免片面使用准确率或召回率所带来的片面性，研究人员引进了F1值。

命名实体识别的准确率p、召回率r和F1值分别定义如下：

 （5-1）

其中： N1为测试文件中实际的命名实体的总数；N2为系统识别出来的命名实体的总数；N3为系统正确识别出来的命名实体数目。是准确率与召回率之间的权衡因子，在本次研究中取l。

准确率能够反映识别结果的可靠成都，召回率能够反映实体提取的效率，而F1值是对实体识别的综合评价。

一个命名实体被正确识别出来当且仅当：

* 1. 在命名实体的头部有一个边界分割标注。
  2. 在命名实体的尾部有一个边界分割标注。
  3. 在命名实体的内部没有边界分割标注。

## 特征阈值筛选实验

## 本章小结

# 总结和展望

## 工作总结

## 未来展望

# 致谢

首先，衷心地感谢高志强老师，无论是在技术实现、问题解决还是论文的修改上，都给我细心的指点和极大的帮助。他一丝不苟的治学态度、严谨的学术思维都是值得我敬佩和学习的！

其次，非常感谢实验室的师兄师姐在实验进行和论文修改中提出的宝贵建议。特别是鲁廷明、杨沛和吴柯师兄在此方面的悉心帮助，感谢鲁师兄在项目繁忙时给了我极大的空间去认真完成毕业设计，并帮我审阅论文，感谢他对我的诸多指点。这些无论是在以后的学习还是工作，都将对我有着很大的影响！

同时，对评审论文的各位专家和教授表示衷心的感谢！他们给我提出的宝贵意见和真知灼见，在我日后的工作和学习上一定会受益匪浅！

最后，我要感谢我的家人和朋友，是他们在精神上了给我极大的支持和鼓励！

再一次对他们表达我深深的谢意和敬意！

# 参考文献

1. Bing Liu. *Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers*, May 2012.
2. Hu, Minqing and Bing Liu. *Mining and summarizing customer reviews.* in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*. 2004.
3. Hu, Nan, Paul A Pavlou, and Jennifer Zhang. *Can online reviews reveal a product's true quality: empirical findings and analytical modeling of Online word-of-mouth communication. in Proceedings of Electronic Commerce (EC-2006)*. 2006.
4. Ellen M, Voorhees. *Overview of TREC 2010[C] // Proceedings of the Nineteenth Text Retrieval Conference Proceedings (TREC 2010)*, NIST, 2010.
5. Iadh Ounis, Craig Macdonald, Ian Soboroff. *Overview of the TREC 2010 Blog Track[C] // Proceedings of the Nineteenth Text Retrieval Conference Proceedings (TREC 2010)*, NIST, 2010.
6. Ku L W, Lo Y S, Chen H H. *Using Polarity Scores of Words for Sentence-level opinion extraction[C] // Proceedings of the NTCIR-6 Workshop Meeting, 2007:316-322*, 2007*.*
7. *Hideo Joho, Tetsuya Sakai. Overview of NTCIR-10[C] // Proceedings of the 10th NTCIR Conference, Tokyo*, Japan, 2013.
8. 赵军, 许洪波, 黄萱菁等．中文倾向性分析评测技术报告[R]．北京:中文信息学会,2008．
9. Liu, Bing, Mingqing Hu, and Junsheng Cheng. *Opinion observer: Analyzing and comparing opinions on the web.* in *Proceedings of International Conference on World Wide Web(WWW-2005)*. 2005.
10. Hu, Minqing and Bing Liu. *Mining and summarizing customer reviews*. in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*. 2004.
11. Popescu, Ana-Maria and Oren Etzioni. *Extracting product features and opinions from reviews.* in *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2005).* 2005.
12. Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar. *Building a sentiment summarizer for local service reviews.* in *Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era.* 2008.
13. Long, Chong, Jie Zhang, and Xiaoyan Zhu. *A review selection approach for accurate feature rating estimation.* in *Proceedings of Coling 2010:Poster Volume*. 2010.
14. Cilibrasi, Rudi L. and Paul M. B. Vitanyi. *The google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 2007. 19(3): p. 370-383.*
15. Zhuang, Li, Feng Jing, and Xiaoyan Zhu. *Movie review mining and summarization.* in *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006).* 2006.
16. Somasundaran, S., J. Ruppenhofer, and J. Wiebe. *Discourse level opinion relations: An annotation study.* in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue.* 2008.
17. Kobayashi, Nozomi, Kentaro Inui, and Yuji Matsumoto. *Extracting aspect evaluation and aspect-of relations in opinion mining. in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007.
18. Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. *Opinion Word Expansion and Target Extraction through Double Propagation. Computational Linguistics, Vol. 37, No. 1: 9.27,* 2011.
19. Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. *Phrase dependency parsing for opinion mining.* in *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2009).*2009.
20. Kessler, Jason S. and Nicolas Nicolov. *Targeting sentiment expressions through supervised ranking of linguistic configurations.* in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM-2009).* 2009.
21. Rabiner, Lawrence R. *A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989. 77(2): p. 257-286*.
22. Lafferty, John, Andrew McCallum, and Fernando Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* in *Proceedings of International Conference on Machine Learning (ICML-2001)*. 2001.
23. Jin, Wei and Hung Hay Ho. *A novel lexicalized HMM-based learning framework for web opinion mining.* in *Proceedings of International Conference on Machine Learning (ICML-2009).* 2009.
24. Jakob, Niklas and Iryna Gurevych. *Extracting Opinion Targets in a Singleand Cross-Domain Setting with Conditional Random Fields.* in *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).* 2010.
25. Li, Binyang, Lanjun Zhou, Shi Feng, and Kam-Fai Wong. *A Unified Graph Model for Sentence-Based Opinion Retrieval.* in *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010).* 2010.
26. Kobayashi, Nozomi, Kentaro Inui, and Yuji Matsumoto. *Extracting aspectevaluation and aspect-of relations in opinion mining.* in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* 2007.
27. Yu, Jianxing, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. *Aspect ranking: identifying important product aspects from online consumer reviews.* in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.* 2011.
28. Manevitz, Larry M. and Malik Yousef. *One-class SVMs for document classification. The Journal of Machine Learning Research, 2002. 2: p. 139-154*.
29. Ghani, Rayid, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. *Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter, 2006. 8(1): p. 41-48.*
30. Kovelamudi, Sudheer, Sethu Ramalingam, Arpit Sood, and Vasudeva Varma. *Domain Independent Model for Product Attribute Extraction from User Reviews using Wikipedia.* in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2010).* 2011.
31. Hofmann, Thomas. *Probabilistic latent semantic indexing.* in *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI-1999)*. 1999.
32. Blei, David M. and Jon D. McAuliffe. *Supervised topic models.* in *Proceedings of NIPS.* 2007.
33. Mei, Qiaozhu, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. *Topic sentiment mixture: modeling facets and opinions in weblogs.* in *Proceedings of International Conference on World Wide Web (WWW-2007)*. 2007.
34. Titov, Ivan and Ryan McDonald. *Modeling online reviews with multi-grain topic models.* in *Proceedings of International Conference on World Wide Web (WWW-2008)*. 2008.
35. Brody, Samuel and Noemie Elhadad. *An Unsupervised Aspect-Sentiment Model for Online Reviews.* in *Proceedings of The 2010 Annual Conference of the North American Chapter of the ACL.* 2010.
36. Li, Fangtao, Minlie Huang, Yi Yang, and Xiaoyan Zhu. *Learning to Identify Review Spam.* in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2011)*. 2011.
37. Mukherjee, Arjun, Bing Liu, and Natalie Glance. *Spotting Fake Reviewer Groups in Consumer Reviews.* in *Proceedings of International World Web Conference (WWW-2012)*. 2012.
38. Andrzejewski, David and Xiaojin Zhu. *Latent Dirichlet Allocation with topic-in-set knowledge.* in *Proceedings of NAACL HLT.* 2009.
39. Andrzejewski, David, Xiaojin Zhu, and Mark Craven. *Incorporating domain knowledge into topic modeling via Dirichlet forest priors.* In *Proceedings of ICML.* 2009.
40. Mukherjee, Arjun, Bing Liu, Junhui Wang, Natalie Glance, and Nitin Jindal. *Detecting Group Review Spam.* in *Proceedings of International Conference on World Wide Web (WWW-2011, poster paper).* 2011.
41. Zhai, Zhongwu, Bing Liu, Hua Xu, and Peifa Jia. *Constrained LDA for Grouping Product Features in Opinion Mining. in Proceedings of PAKDD-2011.* 2011.