硕士学位论文

面向产品领域的细粒度情感分析技术

FINE-GRAINED SENTIMENT ANALYSIS ORIENTED IN PRODUCT DOMAIN

王山雨

哈尔滨工业大学

2011 年 6 月

国内图书分类号：TP391.2　　学校代码：10213
国际图书分类号：681.37　　密级：公开

工学硕士学位论文

面向产品领域的细粒度情感分析技术

硕士研究生：王山雨
导　　师：郑德权 副教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：计算机科学与技术学院
答辩日期：2011 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.2
U.D.C: 681.37

Dissertation for the Master Degree in Engineering

FINE-GRAINED SENTIMENT ANALYSIS ORIENTED IN PRODUCT DOMAIN

Candidate: Wang Shanyu

Supervisor: Associate Prof. Zheng Dequan

Academic Degree Applied for: Master of Engineering

Specialty: Computer Science and Technology

Affiliation: School of Computer Science and Technology

Date of Defence: June, 2011

Degree-Conferring-Institution: Harbin Institute of Technology

摘　要

随着网络信息化的发展，网络上涌现出大量的用户评论信息，如何对这些文本情感信息进行分析整理并抽取出用户关心的问题引起了研究工作者的广泛兴趣。文本情感分析作为自然语言处理的一个应用方向逐渐成为新的热点。文本情感分析又称为意见挖掘，是指对具有感情色彩的主观性文本进行分析、处理、归纳和推理的过程。

文本情感分析涉及了多项具有挑战性的研究任务，根据研究内容的不同，文本情感分析可以分为情感分类、情感信息抽取和情感信息检索和归纳等。其中研究情感信息抽取是指抽取情感文本中有价值的情感信息，情感信息抽取具有实际应用价值。围绕文本情感分析，文本的工作主要包括以下几个内容：

（1）研究文本情感资源的构建技术。为解决目前情感资源缺乏的问题，文本借鉴已有的研究成果并使用了三种方法进行情感词词典扩展，扩大了情感资源的规模、提高情感资源的可信度，并对扩展后的情感词词典进行了校验，最终构建了一个可信度比较高的情感词词典。

（2）研究产品评论语料中的产品属性（评价对象）抽取方法。产品属性抽取是一个很有价值的研究任务，为了提高评论语料中产品属性抽取的准确率，本文采用条件随机场模型和最大熵模型在产品属性抽取任务中进行比较分析。此外，还对词性、浅层句法等相关特征选取进行了细致的介绍，分析加入这些句法特征对产品属性特征抽取的影响。

（3）探索文本情感分析中产品属性跨领域移植的方法。针对目前文本情感语料缺乏的现状，提出了一种基于主动学习的评价对象跨领域移植方法。本文将模型从电子产品领域移植到汽车领域，实验结果证明本文提出的基于主动学习的方法在情感领域移植中起到了很好的作用。

（4）设计并实现了一个细粒度情感分析系统，主要的功能包括评价对象抽取、极性词识别、评价对象极性判断等，综合了本文在文本情感分析中的研究成果。

关键词：文本情感分析；情感词典；产品属性抽取；领域移植；条件随机场

Abstract

With the development of Internet information in the whole world, a large number of user reviews increase with exponential shape. Thus, how to analyze this sentimental information and extract user-needed information arouses the interest of Natural Language Processing (NLP) researchers. Sentiment Analysis as a branch of NLP with potential applications is gradually becoming a new hot research issue. Sentiment Analysis, also known as opinion mining, is to find relevant sources, extract related information with opinion, process this text and set summarization for further tasks.

Sentiment Analysis involves a number of emotional and challenging research tasks. According to the different research field, Sentiment Analysis involves sentiment classification, sentiment extraction, opinion search and opinion summary. Sentiment extraction aims to extract text that is more important and valuable for users. Sentiment extraction has more practical value. Around the sentiment analysis, the thesis includes the following contents:

(1) Research on the construction of sentiment resources. To solve the lack of sentiment resources, the thesis presents three methods to expand sentiment lexicon. The experimental results show that the so achieved sentiment lexicon is more effective.

(2) Research on the extraction of product attributes. Product attribute extraction is a very valuable task for sentiment analysis. In order to improve the precision of product attributes extraction, we apply Conditional Random Field (CRF) model and Maximum Entropy (ME) model with the combination of several important linguistic features. We also analyze the role of each features such as Part of Speech (PoS), shallow parsing to improve our result.

(3) Research on the sentimental domain adaptation. To deal with the lack of sentimental domain corpus, we propose an active learning method to label the data in CRF modeling. The experimental results show that the domain adaptation of active learning achieves the effective performance from electronic product domain to car domain.

(4) Design and implementation of a sentiment analysis system. This system

includes opinion objects extraction, opinion polarity recognition and sentiment text classification and integrates our proposed method in sentiment analysis research.

Keywords: sentiment analysis, sentiment lexicon, product feature extraction, domain adaptation, Conditional Random Field

目　录

第1章 绪 论

1.1 本文研究的背景和意义

获取他人的主观意见已经成为我们决策过程中一个重要的环节。在互联网普及之前，人们为了获取对某个事物或者某个事件的评论信息，要么通过询问朋友亲人，要么借助咨询公司进行调查；公司为了分析商品的出售情况和客户的购买意向，也常常需要通过组织人力对市场进行调研。这样的行为不仅仅需要花费大量的人力物力，而且得到的结果往往带有局部性，不能真实完整的获取当事人所需要的信息。随着web2.0时代的到来，网络上涌现了大量的反应人们真实感情和态度的信息。网络表达方式的变革，改变了单单依靠口口相传的交流模式，互联网使人们的情感信息以文本的形式存储和传播。文本信息不仅易于查找和永久性的存储，而且可以获取大量网友的评论信息，不仅仅包括朋友，更多的是对此有丰富经验陌生人。这样的虚拟的交际模式，更易于获取他人的真实意见和客户的商品需求意向。越来越多的人们乐于接受这种交流方式，并开始在互联网上分享自己的观点、态度、感受和情感等。随着这类信息的爆炸式增长，仅仅靠人工进行信息处理已经成为一个非常棘手的问题。人们迫切需要一种技术来简化自己的工作量，尤其在信息和观点的获取上。在这样的背景下，如何借助计算机对文本信息进行组织处理、挖掘特定信息成为当前研究学者关注的一个热门课题。文本情感分析（sentiment analysis）技术作为自然语言处理的一个研究方向逐渐被研究学者们重视。

文本信息总体上可以分为两大类，一类是事实描述，另一类是情感表达。事实描述是对事物、事件及其属性的客观表达；而情感是表达人们对客观事物的观点、偏见、感觉及其态度等[1,2]。文本情感分析又称作意见挖掘（opinion mining），是指对具有感情色彩的主观性文本进行分析、处理、归纳和推理的过程[3]文本情感分析的研究工作可以追溯到1997年Hatzivassiloglou[3]对带有感情色彩的形容词进行了分析和分类，如“漂亮”、“美丽”是带有褒义色彩的形容词，应该把这类词语归为一类（positive）；而“丑陋”、“邪恶”这类词语带有明显的贬义色彩，应当归为另一类（negative）。在2006年，国际文

本检索会议TREC（Text REtrieval Conference）首先对文本情感分析方向进行了关注，要求检索博客（blog）具有观点性的文档，以后每年都有情感分析相关的任务出现。在2008年的时候，《新华网》、《环球时报》等一些大众媒体纷纷转载了英国《新科学家》杂志的一则报道，英国Corpora软件公司开发了一款名为“感情色彩(Sentiment)”的软件。这种软件的可以对报纸上的文字进行分析处理，从而识别该文章对英国政党的态度。国内的科研工作者也对文本情感分析表现出了非常高的热情，自2008年每年举办一次中文倾向性分析的评测来分享研究机构各自的研究方法和理论，国内形成了很好的研究氛围。

与传统文本信息检索不同，文本情感分析侧重挖掘文本中的观点持有者的主观倾向，而非文本内容本身。文本情感分析是对传统文本信息检索和信息抽取等研究的进一步扩展和传承。文本情感分析的研究内容非常广泛，涉及人工智能、机器学习、信息检索和数据挖掘等多学科的专业知识，而且相当多的研究任务是具有挑战性的。研究文本情感分析就有非常广泛的应用前景：（1）产品的评价与推荐，商家通过挖掘对自己产品的评论信息，可以改进产品、增强市场竞争力；消费者通过挖掘比对产品，寻找物美价廉的物品等等；（2）社会舆论监督，通过对挖掘文本中的情感信息，分析大众对某一事件的看法和态度，了解基层对政府工作的满意度。这样的研究工作不仅提高了管理者的工作效率，而且还能有效的解决基层最关心的问题；（3）信息过滤系统，自由的网络中充斥的色情暴力的内容不利于社会的安全稳定，将文本情感分析用于自动过滤恶意的、有害的、不健康的垃圾信息，有效的维护了互联网的信息安全和社会的和谐稳定；（4）网络社区信息共享，网民可以分享对某首歌或某部电影的看法，传统的信息分享只能在全局做一个综合评分，文本情感分析可以挖掘用户对电影某个细节的看法和褒贬，如对男主角的着装，对背景环境的构造等等。

文本情感分析是一个新兴的研究课题，具有很大的研究价值，受到国内外许多研究机构的重视。由于文本情感分析涉及多项非常具有挑战性的任务，本文只针对文本情感倾向性展开研究工作。文本情感倾向性分析是以文本观点或态度作为研究对象，不包含作者的情绪分析（如喜怒哀乐）。情感倾向性研究任务主要包括情感信息抽取、情感分类以及情感信息检索等。极性词词典，即包含词语及其对应的感情色彩和强度的计算机用词典，是文本情感分析任务的前提。针对目前中文的情感倾向性分析的资源很少，尤其是极性词典等一些关键性资源不充裕，导致文本情感倾向性分析的效果不理想，构建

一个置信度比较高的词典是本领域迫切需要解决的问题。在产品评论领域中，抽取产品属性挖掘潜在的评论信息、改进商品的质量及不足成为提高厂家竞争力的重要手段，然而现有的产品属性抽取的准确率和召回率都比较低，提高产品属性挖掘的效果是我们需要研究的问题；根据目前的研究结果分析，使用有监督的算法抽取产品属性具有很好的准确率，然而产品属性多数为领域性较强的名词性短语。如何在不同的语料中使用相同的训练语料是文本情感跨领域移植研究的问题。解决文本情感跨领域移植问题可以使文本情感倾向性分析的实用性更强，文本情感跨领域移植在解决情感语料稀少问题上有非常重要的意义。

1.2 相关研究综述

根据研究层次和粒度的不同，文本情感倾向性分析可以分为词语情感倾向性分析、句子情感倾向性分析、篇章情感倾向性分析、海量信息的整体倾向性预测[4]。不同的研究层次所涉及的内容有所差异，词语情感倾向性分析分析并获得词语或短语的极性及其强度，是其他研究层次的基础。句子情感倾向性分析对象主要是在特定上下文中出现的语句。其任务是对句子中的各种主观性信息进行分析和提取，包括对句子情感倾向的判断以及从中提取出与情感倾向性论述相关联的各个要素。要素主要有情感倾向性论述的持有者、评价对象、倾向极性、强度，甚至是论述本身的重要性等。篇章情感倾向性分析是将一篇文章作为一个处理对象，主要任务是对篇章进行主客观分类和褒贬分类。篇章情感分类是文本情感倾向性分析中研究最早的一个领域，涉及的研究者数量也是相当多的。海量信息预测是对某个信息源或者某个话题做整体性的概括情感分析。

根究研究任务的不同，文本情感分析主要包括文本情感分类、情感信息抽取和情感信息检索和归纳等[1]。文本情感分类还可以细分为主客观分类及主观褒贬分类，主客观分类是指将带有主观性的文本从客观文本中分离出来，这是我们获取主观性文本一个很重要的手段；主观褒贬分类是指将带有主观性的文本进一步分类积极和消极两类，即所说的文本情感褒贬分类。情感信息抽取是指在文本信息中抽取有价值的、用户关注的一些特征，如评价信息的持有者、文本信息中评价的对象、评价词语及其与评价对象的关系等等。情感信息检索是指一个以应用为主的研究课题，旨在检索同一主题的相关情感文本，它是传统信息检索的一个延伸应用。

目前研究文本倾向性分析主要是基于统计自然语言处理的理论，多数情况下使用机器学习的方法来分析处理。

1.2.1 文本情感词典建设

判断一篇文章极性的褒贬首先要判断的是文章中词语的极性，词语文本情感倾向性分析是其他研究工作的基础。如何构建一个高质量的情感词典资源是开展文本情感分析研究工作的首要任务。词语情感倾向性分析处理的对象主要是形容词、副词和动词等潜在具有情感倾向性的词语。词语情感倾向分析包括对词语极性、强度和上下文模式的分析[4]，分析后的结果可以写入到语义词典中。目前，词语情感倾向性分析中极性词词典的构建方法大体可以分为两类：

（1）基于已有极性词词典或本体库的方法

此方法通过已有的极性词词典来扩展极性词典。Hovy[5]先通过WordNet扩充词典，然后使用情感分类器来标注情感句子中候选词的极性。Jijkoun[6]使用bootstrapping方法将通用极性词典生成针对特定话题（媒体分析）的词典，以便提高情感分类中的准确率。在中文方面，主要使用HowNet本体库进行扩充，如朱嫣岚[7]使用HowNet语义相似度计算来扩展极性词典。从总体来说，基于词典的方法对种子词的质量要求比较高，对已有词典的依赖性比较强。

（2）基于机器学习的方法

无监督机器学习的方法和基于词典的方法类似，首先人工选择一些词作为种子词，然后判断候选词和种子词的紧密程度。对于英文，Turney[8,9] 使用点互信息（Point Wise Mutual Information, PMI）方法来衡量候选词和种子词的紧密程度。Hassan[10]使用无监督的马尔可夫随机游走（Random Walk）模型来计算单词的极性，此模型的优点是处理速度快并且不需要很多语料，但是在准确率等评价指标上表现的不是很明显。

有监督的方法主要是通过语义或者共现原理将种子词进行扩展。如Hatzivassiloglou[3]使用情感词的共现原理，利用连接词的约束条件来扩充英文词典。孙[27]使用最大熵交叉验证来抽取上下文相关的情感词，使用时验证上下文环境以确定情感词的极性，这种方法在一定程度上提高判断情感词在句子中的极性。为了使词典在领域内的效果更好，Du[11]使用改进的信息瓶颈方法构建了领域词典。另外，标注大量的情感语料可以提高词语情感分析的

效果。

1.2.2 文本情感信息抽取

文本情感抽取是抽取文本信息中有价值的情感信息，它是文本情感倾向性分析中一项很有价值的研究工作。文本情感信息抽取的主要任务是提取出与情感倾向性论述相关联的各个要素，包括情感倾向性论述的持有者、评价对象、倾向极性、强度，甚至是论述本身的重要性等。

观点持有者（opinion holder）是指信息的持有者，可以是表达观点的某一人或者某个组织。信息持有者也被称作观点源[28]，如果文本信息为一篇产品评论或者博客，那么观点持有者一般是该文章的作者。在一篇文章中，观点持有者是一个非常重要的要素，它往往表明了当事人或者组织的对某个事件的态度。Choi[29]使用条件随机场（Conditional Random Field, CRF）来识别文章中观点、情感的持有者，识别的准确率达到了81.2%。KIM[5]首先将名词或者名词性短语作为候选观点持有者，然后使用最大熵（Max Entropy, ME）通过分类的方式识别观点持有者。

评价对象是指文本信息（一般以句子为单元）中被评论的产品、事件或者某个话题，甚至是一个人或一个组织。根据评价对象在句子中与中心评论词的距离关系，评价对象可以分为直接评价对象和间接评价对象，直接评价对象是句子中的直接被评论的对象，间接评价对象和直接评价对象具有从属关系，如在句子中“佳能A系列的做工实在太一般”，直接评价对象是“做工”，间接评价对象是“佳能A系列”。现在的研究工作大多集中于商业产品评论，一般首先将评价对象限定于句子中的名词或者名词性短语。目前抽取评价对象的方法大致可以分为两类：一类是基于规则/模版的匹配方法，Liu[31]使用了被称作标签序列规则（label sequential rule, LSR）来匹配语料中的产品特征，此类识别方法的优点是针对性强并且可以高效的识别结构类似文章中的评价对象，缺点是需要人工编写相应的模版，通用型不强；另一类通过大规模语料统计的方法，使用机器学习模型等来识别句子中的评价对象，Xu[30]使用CRF模型识别电子产品领域语料中的评价对象，通过优化特征选择来提高识别的准确率。

评价词语又称极性词、情感词，特指带有情感倾向性的词语。评价词是句子中对评价对象的褒贬评价的短语，如上述句子中的“一般”是评价对象“做工”的评价词。在评价搭配关系抽取上，Liu[12]首先常用的评价对象和

评价词，然后利用评价对象识别对应的评价词、评价词识别对应的评价对象，相互迭代抽取评价对象和情感词。Wei[13]使用基于情感本体树的层次学习模型来识别产品属性及其极性，来克服知识的层次化结构对实验的影响。在中文方面，王波[14]在小规模的标注语料下，使用半监督自学习的方法识别产品的特征。章剑锋[15]使用最大熵模型进行主观关系<评价对象，评价词>抽取。

1.2.3 文本情感跨领域移植

情感分类的准确率会受到领域的影响，主要基于以下原因：1）同一个短语在不同的领域中表达的情感倾向是不同的，如在影评[1]中说“你应该去读书（go read the book）”，应该贬义的，但在书评中表达的意思却是相反的。2）不同的领域使用不同的表达方式，这是导致领域移植的最主要的困难。针对目前中文情感倾向性分析领域性比较强而且语料资源缺乏，使用情感跨领域移植来解决领域内语料稀少的是一个很好的办法[17-19]。另一种领域移植是跨语言移植（Cross-lingual adaptation），鉴于英文中有大量的资源，利用移植的方法将其应用到另一种语言也是非常有价值的一项研究，目前一般使用双语词典或者平行语料来消除不同语言直接的差异[35]，也可以应用机器翻译模型来处理语料[34-36]。

在文本情感分类领域移植方向，Blitzer[16,17]使用SCL-MI (Structural correspondence learning Mutual Information）方法来减少领域依赖关系,并定义了一种衡量领域移植效果的方法。Andreevskaia[18]使用词典(GI)与语料相结合的办法解决情感领域移植的问题，按照在训练集上的准确率计算两者的权重并在句子和篇章上进行测试。Yang[32]为了解决情感分类训练语料稀少的问题，首先抽取在多个领域中起到重要作用的特征，将这类特征视为具有领域无关性，利用这些特征在另外一个领域中进行情感分类。中文方面，Tan[19]首先提出了一种抽取共同特征的方法，然后使用改进的贝叶斯算法获取新领域的知识。

在文本情感评价对象移植方向，张[33]将汽车领域评论语料作为源领域（source domain）集，电子产品领域评论语料作为目标领域（target domain），使用MAP条件随机场领域自适应算法进行跨领域移植，移植效果有了34%的相对提高。然而此种方法没有充分利用未标注目标语料，仅仅使用两个领域共同的特征，失去了大量的目标领域的先验知识。

1.2.4 语料库资源

在自然语言处理领域中，根据研究方法的不同可以分为基于规则的自然语言处理和基于统计的自然语言处理。基于规则的自然语言处理方法以乔姆斯基（Noam Chomsky）语言体系为主；基于统计的自然语言处理主要以语料库为学习的样本，是文本分析的基础。

在统计自然语言处理中语料库起着非常重要的角色，使用一个高质量的语料库可以使研究工作更加出色。文本情感分析的目的是使计算机更加智能的分析文本情感，这就需要大量的、高质量的语料库。在文本情感分析领域，研究工作者们已经人工标注了部分行业的情感语料。情感分析英文语料主要有影评数据（cornell movie-review datasets）[1]、用户评论数据（customer review data）[2]，前者可以用来在做句子级和篇章级情感分析，后者由亚马逊网站的评论数据组成。在中文方面，文本情感分析主要使用COAE（中文倾向性分析评测）语料。COAE评测是有中国信息学会信息检索专委会[3]主办，在中文情感分析领域起到了重要的作用，并为研究工作提供了丰富的语料。COAE语料是有中科院计算所和洛阳外国语学院共同整理和标注完成的，主要涉及影视娱乐、财经、教育、房产、电脑、手机等领域。

COAE语料库只标注了评价对象及其极性，并没有标注对应的评价词，为了更好的进行研究工作，本实验室将COAE语料中电子产品领域部分进行人工极性词标注。为了保证标注的可信性及防止人工标注的偏差，每篇文档均由三人标注并对其校正。电子产品领域共包括312篇文章，其中相机评论语料136篇，笔记本评论语料53篇，手机评论语料123篇，总共评价对象个数为6000个，评价词个数4747个。文本使用的语料主要是本实验室标注后的COAE语料。

1.3 本文的研究内容

本文的研究内容主要包括四个方面的内容：（1）研究文本情感资源的建设技术，借鉴已有的研究成果并提出一种词典扩展的方法来扩大情感资源的

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[2] http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip

[3] http://www.cipsc.org.cn

规模并提高情感资源的可信度；（2）探索评论语料中的产品属性抽取的研究方法，提高评论语料中的产品属性的准确率以提高产品属性抽取技术的实用性；（3）探索文本情感分析中产品属性跨领域移植的方法，针对目前文本情感语料缺乏的现状，提出一种高效的产品属性跨领域移植方法。（4）在情感分析的研究工作的基础上，本文设计一个情感分析系统。系统的主要功能包括评价对象抽取、极性词识别、评价对象极性判断等，以便直观的展现研究成果。围绕这些工作，文本的章节安排如下：

第1章为绪论部分，主要介绍了文本情感分析的研究背景，介绍了文本情感分析的发展起因，从互联网发展的角度阐述了网络发展对人们生活交流的影响，也进一步影响了研究工作者工作内容，并定义文本情感分析的研究内容和研究意义。相关研究综述介绍了国内外工作者研究现状和工作进展，对文本情感分析研究工作进行总体概述。

第2章为文本情感资源建设，主要分析了现在文本情感分析研究方法及其优缺点，进而详细探讨了三种经典方法在本文的应用。为了提高资源建设的精度，融合上述方法后并进行了词典校验和剥离，提高了词典资源的可信度。

第3章为评论语料中的产品属性抽取，引入相关机器学习模型并介绍了机器模型的原理及其使用方法，进而阐述不同特征对实验结果的影响。通过添加相关特征进而提高抽取的指标和实用性。最后对实验结果进行了分析，表明充分利用机器学习模型的优势和有针对性特征选择可以提高产品属性抽取性能。

第4章为文本情感分析跨领域移植，章节开始介绍了半监督学习和主动学习等涉及到的机器学习方法，阐述了方法原理并比较学习方法之间的差异。然后提出了基于条件随机场的产品属性抽取技术。本文提出了使用主动学习方法策略的跨领域移植的方法，实验证明了此方法可以提高文本情感领域移植的效果。

第5章为文本情感分析系统设计与实现，文本设计了一个文本情感分析系统，整合了现有工作并展现研究成果，开发一个文本倾向性分析的应用系统，提高了文本倾向性分析的实用性。针对电子产品领域和汽车领域评论，可以判别句子极性、识别句子中的评价对象和评价词。此系统可以很好的展示文本情感分析的应用模式并为研究提供一个基础性的平台，方便以后研究成果的集成。

第2章 文本情感资源建设

2.1 引言

在表达对某事物或事件的观点时，人们的情感大致可以分为两类，一类为积极向上、赞扬的看法，另一类表达厌恶、批评的看法。进而可以把带有情感的词语分为两类，一类为正极性词语，另一类为负极性词语；没有表达情感的词语成为中性词（neutral）。词语表达情感的强度称为词语的极性强度。极性词词典是由上述词语及其极性强度构成的计算机可用词典。

中文情感倾向性分析对资源的依赖比较大，而极性词词典是用于研究使用的主要资源。极性词词典不仅仅要确定一个单词的极性，还要确定这个词在上下文中的极性。词语倾向性分析处理的对象主要是形容词、副词和动词等潜在具有情感倾向性的词语。词语情感倾向分析包括对词语极性、强度和上下文模式的分析。其分析结果可以写入到语义词典中。词语情感倾向性分析的一个主要任务就是构建极性词词典。

目前使用的情感词典主要是人工编撰，英文词典有General Inquirer, sentiWordNet, OpinionFinder's Subjectivity Lexicon[3,38]等，新版的General Inquirer[39]合并了Harvard-IV-4和Lasswell词典，根据词语的性质和功能分为182个类别，其中带有正极性的词语有1915个，带有负极性的词语有2291个；中文情感词典有学生褒贬义词典[41]，知网，NTU sentiment Dictionary[40]等。比较常用的为知网（本体库），包含了219个中文程度词语、836个中文正面情感词语和1254个中文负面情感词语。虽然耗费了大量的人力物力进行词典构建工作，文本情感分析中极性词的需求远远大于词典中极性词数量，使用计算机来扩展极性词词典是研究工作者一直在努力的方向。

中文情感资源的缺乏影响了文本情感的应用前景，本章提出了情感词典融合构建方法，使用词语相似度计算、本体库扩展及网络资源将已有的词典资源进行扩展，使情感词典的适用范围更广、效果更好。

2.2 基于词语相似度情感词典扩展

在语言学上，语言粒度从小到大可以分为语素、词语、短语、句子、段落及其篇章，其中文本的语义信息蕴含于各个层次的语言粒及语言粒的各种语法关系中[42]。计算语言学上常常利用小的语言粒度来计算大粒度语言单元，这是一种基于解析思想的常用方法。语素是最小的语法单位，它是最小的语音、语义结合体。目前文本情感倾向性研究的信息主要是文本信息，语素在文本信息中主要起到构成的作用，研究工作者常常把词语作为最小的研究单元，利用词语的情感倾向性来计算分析句子、文本的情感倾向性。

文本相似度是表示两个或多个文本之间匹配程度的一个度量参数。由此可以得出，两个文本之间相似度越高，两个文本的相似程度也就越高，反之就越低。按照文本粒度不同，中文文本的相似度主要包括以下几种关系：词与词、词与句、词与段、句与句、句与段以及段与段之间的关系等。体现在词语级别上，文本相似度主要用来衡量文本中词语的可替换程度，相似度越高说明词语之间的同义程度越高，进而词语的情感倾向性也相同。基于文本相似度理论，通过计算词语之间的相似度，如果相似度大于某个阈值，可以认为候选词的极性与基准词的极性相同。

常用的词语相似度计算方法主要分为两种：基于语料库统计的方法、基于语义词典的方法等。基于语料库统计的方法是近几年来研究比较多的一种词语相似度计算方法，Turney[8]通过上下文的信息来确定词语的情感倾向并使用了PMI-IR（Point Wise Mutual Information and Information Retrieval）方法来衡量候选词与基准词的关系，从而确定候选词的极性。之后又做了进一步的研究工作，将使用PMI-IR计算词语相似度的方法和LSA（Latent Semantic Analysis）的方法进行对比，实验结果表明使用PMI的方法优于LSA的方法

2.2.1 点互信息方法概述

互信息（Mutual Information, MI）是信息论中的一个概念。在信息论中，信息是物质、能量、信息及其属性的标示，信息是确定性的增加。信息论中的互信息是衡量两个随机变量相互依赖、共有信息的程度。在自然语言处理领域中，互信息广泛的应用于语义消歧、词语搭配识别以及聚类等相关课题的研究。

对于随机变量X和Y，两者之间的互信息用I(X;Y)表示，计算公式为：

$$H(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \qquad (2\text{-}1)$$

其中H(X)为随机变量X的熵，H(X)为随机变量X的熵，定义为

$H(X)=-\sum_{x\in X}P(x)\log_2 p(x)$，H(Y|X)为已知随机变量X的情况下Y的条件熵，定义为 $H(Y\mid X)=\sum_{x\in X}P(x)H(Y\mid X=x)$。

1961年，Fano[43]提出了点互信息的概念，用来计算分布中两个特定样本点之间的互信息，公式为：

$$I(x,y)=\log\frac{p(x,y)}{p(x)p(y)} \tag{2-2}$$

PMI可以度量两个词之间的统计依赖性。对于两个词$word_1$和$word_2$，将两个词在语料库中的使用分布视为两个随机变量。用PMI度量两个词$word_1$和$word_2$的统计依赖性，进而得到两个词的近似程度。两个词$word_1$和$word_2$ 之间的点互信息PMI($word_1$，$word_2$)定义为：

$$PMI(word_1,word_2)=\log_2\frac{p(word_1,word_2)}{p(word_1)p(word_2)} \tag{2-3}$$

式中P($word_1$，$word_2$)为$word_1$和$word_2$同时出现的概率，P($word_1$)为$word_1$单独出现的概率，P($word_2$)为$word_2$单独出现的概率。在实际应用中，为了简化计算，概率可以通过两个词在语料库中共现次数进行估计，公式为：

$$PMI(word_1,word_2)\approx\log_2\frac{N\times count(word_1,word_2)}{count(word_1)\times count(word_2)} \tag{2-4}$$

式中N为语料库中句子的总数，count($word_1$,$word_2$)为语料库中同时出现的次数，count($word_1$)和count($word_2$)分别为$word_1$和$word_2$在语料库出现的次数。

为了计算情感词的极性，需要将候选情感词cword与强度为i的所有基准词BSW_i进行极性相似度计算。对于一个候选情感词cword，将cword和每个基准词bsw进行PMI词语倾向性计算，得到当前候选词cword的总的互信息，计算公式为：

$$sentiOrient(cword,BSW_i)=\sum_{bsw\in BSW_i}PMI(cword,bsw) \tag{2-5}$$

在情感词典中，通常使用一个整数来表示情感词的极性强度，根据经验定义整数的范围为{-3，+3}。其中正数表示正极性，负数表示负极性，0表示情感词为中性词，整数的绝对值越大表明该情感词极性越强。由候选词cword和基准词的互信息可以得出cword的极性强度为：

$$sentiPolarity = \arg \max_{i \in \{-3,-2,-1,0,1,2,3\}} sentiOrient(cword, BSW_i) \quad (2\text{-}6)$$

式中i为情感词的强度，范围为{-3，3}。文献[42]分别用40对、10对、5对正负情感基准词进行了相关实验，结果证明基准词对数越多，识别新情感词的效果越好。

2.2.2 相似度扩展算法

基于PMI的词典相似度算法是一个高效的情感词词典扩展算法，本文在已有的电子产品领域评论语料中实现了基于PMI的词典扩展方法，并作为词典构建融合算法的一部分。根据已有的研究证明，使用基准词数越多识别新情感词的效果越好。所以，本文在实验室现有情感词典的基础上进行了情感词典扩展，基准词的数目远远超过了已有的情感词典的规模。

对于一个候选情感词cword，将cword和强度为i的所有情感词典极性词LEX_i中的每个词进行PMI词语倾向性计算，得到当前候选词cword的总的互信息，计算公式为：

$$sentiOrient(cword, LEX_i) = \sum_{lex \in LEX_i} PMI(cword, lex) \quad (2\text{-}7)$$

从而获得候选词的极性为：

$$sentiPolarity = \arg \max_{i \in \{-3,-2,-1,0,1,2,3\}} sentiOrient(cword, LEX_i) \quad (2\text{-}8)$$

本文基于PMI词语相似度扩展的基本思想是：将语料库中的文本评论信息以句子作为处理单元，针对句子中的词语建立倒排索引以便获取每一个词的在语料库中出现的频数，然后计算索引中候选词和情感词的PMI，每一个候选词选择PMI信息最大的极性强度作为自身的极性强度，并舍弃候选词为中性的词语。基于PMI词语相似度扩展的具体算法如图2-1所示。

在本实验中，使用了基础情感词典作为扩展情感词的依据。基础情感词典是本实验人工标注的极性情感词典，共计有3235个情感词，其中包括1754个褒义词，1481个贬义词，每个情感词都用整数{-3，3}标注极性强度。实验使用的语料库为COAE2008（2008年中文情感倾向性分析评测会议）电子产品领域评论语料，抽取符合条件的句子共302,103句，按词进行倒排索引后得到候选词46,832词，去除中性词后得到情感词38,668词。大部分情感的极性可以正确的识别，但其中掺杂一部分中性词，下一步工作对其做一步的实验

处理，以便使词典达到实用的水平。

PMI_EXPAND(corpus,lexicon)

输入：

corpus：电子产品领域评论语料库;

lexicon：情感词词典;

输出：

new_senti_word & polarity：新情感词及其极性;

算法：

step 1：对于 corpus 的每一个文本文件，按照中文语义切分为句子 sentences;

step 2：将 sentences 进行分词，去除停用词，对整个 corpus 按单词 word 进行倒排索引，索引值为 sentenceID;

step 3：分离倒排索引中情感词 $senti\_word_i$ 和候选词 unkown_word，并在 lexicon 中查询每个情感词的极性，计算情感词的频数 $count(senti\_word_i)$;

step 4：对于每一个 unkonw_word:

计算候选词的频数 count(unkown_word)，unkown_word 和每个 $senti\_word_i$ 共现频数 co_count;

按照公式(2-7)计算每个候选词的 uw_pmi_j ($-3 \leqslant j \leqslant 3$,j 为整数);

$polarity=argmax(uw\_pmi_j)$;

step 5：过滤 Ploarity=0 的 unkown_word，剩余的候选词作为新情感词 new_senti_word 返回。

图2-1 PMI_EXPAND算法实现

2.3 使用本体库扩展情感资源

基于统计的词语相似度计算被广泛的应用在信息检索、文本分类等众多领域。在语料库充分的前提下，基于统计的词语相似度可以达到很好的效果，然而面对目前语料库不充裕的情况下，统计词语相似度面临着自身的瓶颈。本节将基于语义相似度进行情感词扩展，以弥补语料库的匮乏对情感词扩展造成的影响。

正如上节所说，词语相似度计算大体可以分为两类，一类是基于统计的

词语相似度计算，另一类是基于语义词典的相似度计算。第2.2节使用了第一类方法基于统计的PMI词语相似度计算进行词典扩展。基于语义词典的相似度计算主要使用的是本体知识（Ontology）如WordNet、HowNet中的同义词或者词语的树状体系结构，通过计算两个词的信息熵或者语义距离来获得两个词的相似度。

在进行语义相似度计算时需要考虑语义关系和语义距离，语义关系也称为语义结构，是指句子的语义组合关系。词语语义关系包括继承关系、整体部分关系和同义词关系三个关系，不同关系的概念之间其相似度是不相同的，同义词关系的词相似度要大于继承关系和整体部分关系。语义距离一般指两个概念在Ontology语义树上的距离，是连接两个概念之间的最短长度。语言学的学者认为，两个词的语义相似度越大表明这两个词的相似度越大，反之亦然。根据这一理论，本节以HowNet中文本体库为基础进行情感词扩展，以便得到高质量的情感词。

2.3.1 本体库概述

在自然语言处理中，目前常常使用辅助资源如词典、知识库和统计相结合的方法进行句法分析、语块识别等。本体库广泛应用于自然语言处理中多项任务中，本体库是对共享概念的正规、明确的表述[44]。本体库的概念来自于哲学，1991年，Neches[45]将本体库的概念引入人工智能领域，主要用来在语义和知识层面上描述信息系统的概念。目前，英文中应用比较多的本体库为WordNet，它按照单词的语义将单词分组，同一组的单词成为一个同义词集合（Synset）。中文本体库主要有HowNet[4]，“是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库”[46]。它使用树状结构将中文词语、字联系在一块。在知网中，词语使用概念或者称为义项来描述，概念可以清楚的描述词语的各个属性，它是用一种“知识表示语言”来描述的，每个概念又通过“义原”来描述，并定义义原是“最基本的、不易于再分割的意义的最小单位”。更确切的说，义原可以来描述概念之间的关系，进一步可以描述词与词之间的关系。刘群[47]使用HowNet进行词语相似度计算，并应用在机器翻译领域中。根据义原在HowNet中的作用，他将HowNet中的义原分为三类，

[4] http://www.keenage.com

分别为基本义原、语法义原、关系义原。两个概念的相似度是由第一独立义原、其他独立义原、关系义原和关系符号这四个特征组成，第一独立义原是基本义原的第一个义原描述，其他独立义原是基本义原中除第一独立义原以外的其他义原描述。

2.3.2 本体库扩展方法

使用HowNet相似度计算可以获得许多高质量的情感词，本文使用电子产品领域评论语料，实现了基于情感词的HowNet词语相似度计算。计算两个词语 W_1 和 W_2，假设在知网中的义项（概念）分别为 $C_{11},C_{12},...,C_{1m}$ 和 $C_{21},C_{22},\ldots,C_{2n}$，规定 W_1 和 W_2 的相似度是各个义项相似度中最大值，即：

$$Sim(W_1,W_2)=\max_{i=1...m,\,j=1...n} Sim(C_{1i},C_{2i}) \tag{2-9}$$

这样，将词语的相似度问题转化为概念的相似度问题，文献[47]认为两个概念的相似度由第一独立义原、其他义原、关系义原和符号义原组成。用公式表示为：

$$Sim(C_1,C_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \tag{2-10}$$

Sim(S1,S2) 为两个义原之间的相似度，$\beta_1+\beta_2+\beta_3+\beta_4=1$ 且 $\beta_1\geq\beta_2\geq\beta_3\geq\beta_4$，此参数反映了每种义原在计算相似度时的贡献值。在知网中，所有义原都根据上下位的关系构成了树状结构，这样可以使用语义树的距离来衡量两个义原的相似程度，公式表示为：

$$Sim(S_1,S_2)=\frac{\alpha}{Dis(S_1,S_2)+\alpha} \tag{2-11}$$

其中，$Dis(S_1,S_2)$ 表示 S_1 和 S_2 在语义树的路径长度，α 为可调节的参数。

通过上述公式可以求得两个词之间的相似度，在情感分析中为了计算候选词的极性，我们定义K对褒贬基准情感词，候选词uk的极性为正极性词与负极性词之间的差值，公式表示为：

$$Polarity(uk)=\sum_{w\in positveLex}Sim(uk,w)-\sum_{w\in negtiveLex}Sim(uk,w) \tag{2-12}$$

在式（2-12）中，当极性P大于阈值T(T>0)时，将uk标为正极性，极性P小于阈值-T时，将uk标为负极性。在实验中，人工挑选了10组可信度高的情感词，褒义词为健康、安全、天下第一、美丽、超级、保险、卫生、天使、

英雄、精选；贬义词为不合作、黑客、疯狂、错误、事故、非法、失败、背后、麻烦、不良。根据经验设定阈值T为0.3，经过语义相似度计算后，得到相似度高的6000词，褒贬义词各3000词。在文献[7]的实验中将阈值设为0，然而得到的效果并不是很理想，经过实验证明将阈值设为0.3后发现情感词的质量可以大幅度的提高。

2.4 使用网络资源扩展情感词典

互联网的出现已经改变了我们的生活方式，网络上出现了大量的资源信息，这些资源信息往往是人们手工编辑的，具有较高的可信度。为了充分利用这些资源以及完善情感词词典，本文使用网络同义词来扩展情感词词典。笔者发现一些知名网站提供同义词资源，比如金山词霸[5]、百度词典[6]等。网站上的同义词都是经过人工核对的，这对情感词扩展非常有帮助。扩展的思路如下：读取基本情感词词典，对于每一个情感词，通过网络查询每个词的同义词和反义词，将同义词标记为情感词的极性、反义词标记为情感词的相反的极性。若候选词被标注多个极性时，使用投票加权法确定候选词的极性。网络资源扩展情感词算法如图2-2所示。

最终获取网络资源扩展情感词14,516个，包括7,364个褒义词和7,152个贬义词。网络扩展的情感词有较多极性冲突的现象，这表明有些词的极性是随着上下文的语境而变化，导致与多个极性相反的情感词同为同义词。

2.5 多方法融合的词典构建

上几节分别使用了基于统计的词语相似度扩展情感词、使用本体库扩展情感词及使用网络资源扩展情感词，虽然在数量上远远的超过了基本情感词词典，但是得到的情感词包含很多噪声，质量有待于改进。为了提高情感词的质量，得到一个实用价值高的词典，本文将三种方法进行融合，将新的情感词词典应用于第三章的产品属性抽取实验，获得了很好的效果。

[5] 金山词霸 http://hanyu.iciba.com

[6] 百度词典 http://dict.baidu.com

INTER_EXPAND(lexicon,Sites)

输入：

lexicon：情感词词典;

Sites：网站词典的链接地址

输出：

new_senti_word & polarity：新情感词及其极性;

算法：

step 1：对 lexicon 进行预处理，获取褒义词和贬义词的极性;

step 2：for word ←lexicon

step 3：for site ←Sites

step 4：获取 word 的同义词和近义词，将结果放入结果集 ResultSet;

step 5：for unkown ← ResultSet

if unkown 存在多个极性;then

对 unkown 的所有极性加权，获得 unkown 极性 ploarity

step 6：合并 ResultSet 和 lexicon 词典

图2-2 INTER_EXPAND算法实现

使用基于统计的词语相似度方法扩展情感词词典得到了38,668个情感词，由于语料库的不规整性，情感词中存在大量的中性词，而通过本体库扩展的6000个情感词也有部分词的极性不一致。网络资源虽然使用近义词和反义词扩展了14,516词，但是由于中文词语使用范围较多，有些情感词具有多种极性，即需要在上下文环境中判断情感词的极性。为了使情感词词典达到高实用性，本节将三种方法进行了相互校验。算法的基本思想是：使用三种方法扩展后对于极性相同的情感词认为是可信的，对于两种方法相同的情感词进行人工校验。通过程序处理和人工校验后，得到静态情感词14,810个，与上下文相关的动态情感词170个。

为了验证情感词典的效果，本文将情感词典应用于产品属性抽取过程。产品属性抽取实验使用条件随机场模型，除了使用词性、语块信息等特征外，情感词也作为模型的一个重要特征。本实验使用的训练集和测试集均为COAE2008电子产品领域评测语料，条件随机场使用CRF++工具软件。精确

评价是指产品属性识别严格匹配时视为正确匹配，覆盖评价是产品属性识别包含于标注答案时视为正确匹配，实验结果如表2-1所示。

表2-1 情感词典应用对比实验

特征选择	精确评价			覆盖评价		
	准确率	召回率	F1值	准确率	召回率	F1值
旧情感词典	58.5	45.0	50.9	74.9	57.7	65.1
新情感词典	59.0	45.6	**51.4**	76.2	58.4	**66.1**

由表2-1可以看出，使用新情感词后精确评价F1值和覆盖评价F1值都有明显的上升，这证明新情感词词典质量要优于原情感词词典质量。然而我们也发现结果上升幅度并不明显，我们分析认为是由于训练语料局限于电子产品领域，语料中情感词的范围有限从而并不能发挥新词典的真正作用。

2.6 本章小结

本章主要进行了情感词词典资源扩展工作，分别使用了基于统计的词语相似度扩展、本体库扩展、网络资源扩展多种方法对已有的情感词词典进行了扩展。为了保证词典的高可信度，扩展时舍弃了大量的不一致的词语。虽然情感词在总量要少于使用扩展方法得到的词数，但是一个高质量的词典对我们来说更为重要。另外，在词典扩展时简单介绍了本体库的原理和应用，本章最后介绍了一下情感分析常用的语料库以及我们在语料库建设上的工作。

第3章 基于条件随机场的产品属性抽取技术

3.1 引言

随着信息时代的到来，网络上存在大量的产品评论信息。通过分析大量的产品评论语料，我们发现每个人对产品的评论并不是单单对某产品的评论，而是通过分析产品的某些属性来评价该产品。不同的人关注的产品属性也是不一样的。对于客户来说，希望在众多评论中不但找到对某件商品的整体看法，而且还需了解对产品某个细节的评价，如手机电池的耐用性；对于商家来说，他们也想了解顾客对自己产品的各方面的细节。面对这样的任务仅仅依靠人力来处理信息是远远不够的，人们迫切需要使用计算机来自动分析这些情感评论语料，这些需求都是以往商品评分系统无法胜任的。文本情感分析可以提取评论中的评论对象，针对某个细节进行褒贬分析，即文本情感信息抽取。

文本情感信息抽取的主要任务是提取出与情感倾向性论述相关联的各个要素，包括情感倾向性论述的评价对象、持有者、倾向极性、强度，也可以是论述本身的重要性等[4]。其中，评价对象是指句子中被评论的产品、事件或者某个话题，甚至是一个人或一个组织。在产品评论语料中，人们更关注的是评价对象，即产品属性。产品属性是产品本身所固有的性质，是产品在不同领域差异性的集合。比如“佳能A系列的做工实在太一般”这句话中“做工”是产品“佳能A系列”的属性。

目前属性抽取的方法大致可以分为两类：一类是基于规则/模版的匹配方法，如Liu[31]使用了被称作标签序列规则（label sequential rule, LSR）来匹配语料中的产品特征，此类识别方法的优点是针对性强，可以高效的识别结构，即评价对象，缺点是需要人工编写相应的模版，通用型不强；另一类方法是大规模语料统计的方法，使用机器学习模型等来识别句子中的评价对象，Xu[30]使用CRF模型识别电子产品领域语料中的评价对象，通过优化特征选择来提高识别的准确率。本文的研究方向是采用基于统计学习的方法抽取评价对象，并使模型的训练过程不依赖于特定的领域知识，通过引入多种有效的语言学特征来提高系统的性能。其中，基于统计的产品属性是本文的研究的

方向。

3.2 相关机器学习模型

机器学习是使用计算机来模拟或者实现人类的学习行为，以获取新的知识或技能。在统计自然语言处理中，机器学习模型起着重要的作用，常用的机器学习模型有最大熵模型、隐马尔科夫模型（Hidden Markov Model，HMM）、条件随机场、支持向量机（Support Vector Machines，SVM）等等。

3.2.1 最大熵模型

熵的概念来源于信息论，信息论是香农在20世纪40年代建立的理论体系。熵是用来表示信息不确定性的均值。信息的不确定性越高，其熵值就越大。熵的计算公式为：

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \tag{3-1}$$

其中，p(x)为随机离散变量X的概率密度函数，x属于某个符号或者字符的离散集合X，此公式表明随机变量X的熵越大，则不确定性越大，能估计X值的概率也就越小。

最大熵的主要思想是在掌握未知分布知识的前提下，应该选择符合这些知识但是熵值最大的概率分布。最大熵模型正是一种基于最大熵原理的统计预测模型。从随机变量的角度来说，最大熵实质是在已知知识的前提下，选择未知分布符合已知知识中不确定性最高的那个分布作为预测的结果，即尽可能的选择均匀分布作为预测结果。最大熵保证了除限定条件以外，对未知的分布不做任何人为的假设。

自然语言处理中的许多问题可以被看作统计分类问题，即估计在上下文b的发生条件下，被标记为类别a的概率。对于文本分类问题，已知训练语料为（X，Y），X为输入文本信息，Y为输出类别，则p(y|x)表示当输入为x时，标记为y的概率。 $\tilde{p}(x, y)$为训练样本中x的经验概率。为了表示语料库中包含的各种知识，我们使用二值特征函数来表示：

$$f(x, y) = \begin{cases} 1, & if\ y = 0 \\ 0, & other \end{cases} \tag{3-2}$$

则特征函数f对于样本的期望为：

$$\tilde{p}(f)=\sum_{a\in\{x,y\},b\in\{0,1\}}p(a,b)f(a,b) \qquad (3\text{-}3)$$

从而，我们可以将语料库中的样本统计表示为特征f的期望，f对于p(y|x)的期望可以定义为：

$$p(f)\equiv\sum_{x,y}p(x,y)f(x,y)\approx\sum_{x,y}\tilde{p}(x)p(y\mid x)f(x,y) \qquad (3\text{-}4)$$

在训练过程中，当我们发现有价值的统计数据时，调整模型以适应该数据，以示此数据的重要性。为了达到这一目标，我们通过约束 $p(f)=\tilde{p}(f)$ 来实现，用公式表示为：

$$\sum_{x,y}\tilde{p}(x)p(y\mid x)f(x,y)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \qquad (3\text{-}5)$$

最大熵的原理是在满足公式（3-5）的模型p中，选择一个分部最均衡的模型。为了衡量条件概率分布p(y|x)的均衡性，我们引入条件熵的概念，H(Y|X)为已知随机变量X的情况下Y的条件熵，定义为 $H(Y\mid X)=\sum_{x\in X}p(x)H(Y\mid X=x)$。在最大熵中，可以使用条件熵H(p)来衡量p（y|x）的均衡性，即：$H(p)\equiv-\sum_{x,y}\tilde{p}(x)p(y\mid x)\log p(y\mid x)$，H(p)的取值范围为[-∞,0]，如果H(p)的值为0说明模型是确定的并不存在不确定。根据这一理论，我们得出：在所有的条件概率分布p中，选择是条件熵H(p)最大的模型p*，即得 $p^{*}=\arg\max_{p\in C}H(p)$。

求解最大熵模型，可以看作一个约束条件下求极值的问题，这类问题我们通常使用拉格朗日乘数法来确定。为了解决此类问题，Darroch提出了称为GIS算法（Generalized Iterative Scaling Algorithm）的优化方法，D.Pietra等改进了原有的最大熵模型求解算法，通过降低求解算法的约束条件，提出了IIS（Improved Iterative Scaling Algorithm）。IIS增加了算法的适用性，目前该方法是求解最大熵参数中常用的算法，鉴于网络上有大量的实现算法，本文不做详细的证明和解释。

最大熵模型是在满足约束条件下的模型中选择信息熵最大的一个模型，在出现的概率上也是占优势的。另外，最大熵可以很灵活的选择特征，而不需要特别关注如何使用这些特征，因为 $p(f)=\tilde{p}(f)$ 这样的约束形式可以不用关心特征是否重叠，一个实例可以出现任意多个特征。这一特性为特征的选

取和模型构建提供了很大的便利。

3.2.2 条件随机场模型

条件随机场是一种用于序列标注的机器学习模型，它是基于概率结构的模型。隐马尔科夫模型和最大熵隐马尔可夫模型（Maximum Entropy Markov Models，MEMM）的基础上发展而来的。条件随机场被广泛用于序列标注、数据分割、组块分析等自然语言处理任务中；中文分词、命名实体识别、歧义消解等任务也常常使用CRF模型。鉴于条件随机场发展于隐马尔可夫模型，先简单介绍一下HMM。

马尔可夫（Markov）模型是一种广泛应用于语音识别、词性标注、语音转换、概率文法等自然语言处理任务的一种统计模型。对于一个随机变量序列称为马尔可夫过程（链）需要满足两个条件：历史有限性假设和时间不变性假设，即

$$P(X_{t+1}=s_k \mid X_1,...,X_t)=P(X_{t+1}=s_k \mid X_t) \quad (3\text{-}6)$$

$$\forall i \in \{1,2,3,...,T\}, \forall x,y \in S, P(X_i=y \mid X_{i-1}=x)=p(y \mid x) \quad (3\text{-}7)$$

式(3-6)描述了历史有限性，X_{t+1}只与X_t有关，即一阶马尔可夫链。式(3-7)表示历史不变性，在任何时间i，p(y|X)不随时间改变而改变。

马尔可夫模型可以表示为一个三元组(S,Π,A)，其中S是状态的集合，Π是初始状态的概率，A是状态间的转移概率。隐马尔可夫模型是在马尔可夫模型的基础了加了一层隐藏的状态。隐马尔可夫模型被广泛应用主要有两个原因，一个是隐马尔可夫模型有丰富的数学理论结构，另一个原因是在某些重要的应用中恰当的出色的运用表现。隐马尔可夫模型可以表示为一个五元组(S,K,Π,A,B)，S表示状态的集合，模型的各个状态之间是相互连接的，任何状态都能到达其他的状态；K为模型中每个状态对应的不同的观察符号，即输出字符集合；Π为模型的初始状态的概率，A为状态转移的概率，表示为 $A=\{a_{ij}\}$，是指从某一状态i到状态j的概率；B为在状态转移的过程中输出字符的概率，$B=\{b_{jk}\}$是指在状态j是观察到字符为k的概率。隐马尔可夫模型主要用于解决以下三个问题：

1）在给定HMM模型μ=(S,K,Π,A,B)的基础上，如何高效的计算某一字符的输出概率p(O|μ)，即HMM评估问题；

2）给定了一个输入字符序列O和模型μ，如何确定产生序列O概率最大的状态序列X，即HMM解码问题；

3）给定一个输出字符序列O，如何调整模型的参数使模型μ产生序列O的概率最大，即训练隐马尔可夫模型。

三个问题对应的解决方法分别为向前算法、维特比（Viterbi）算法及向前向后算法（Forward-backward algorithm也称Baum-Welch算法）。然而隐马尔可夫模型有一个致命的缺陷假设：输出独立性假设，即输出序列只与当前状态有关，输出序列字符之间无关。这样的假设导致HMM不能考虑上下文的特征，限制了特征的选择。为了解决HMM的假设问题，Andrew[48]提出了最大熵马尔可夫模型（MEMM），将最大熵模型和马尔可夫模型进行了融合，MEMM模型解决了HMM的输出独立性假设，可以任意的选择特征。但是MEMM在每个节点都要进行归一化操作，导致模型只能找到局部最优值，同时也出现了标记偏见问题（label bias problem）。

在2001年Lafferty[49]提出了条件随机场模型，CRF模型不仅优于隐马尔可夫模型，而且解决了最大熵马尔可夫模型中存在的问题。条件随机场是一种判别式无向图模型，模型的思想来源于最大熵模型。CRF是一种用来标记和切分序列化数据的统计模型，它在给定欲标记的观察序列下，计算整个标记序列的概率。为了求解CRF无向图所对应序列的联合概率，需要寻找一种方法来分解无向图，我们按图中最大连通子图也称为最大团C（clique）来分解无向图，并在上面定义一种函数，这样保证了分解后没有边的两个节点在不同的函数中，并将此函数称为势函数（potential functions）。势函数是严格的正实数值的函数。为了保证一组正实数函数的乘积满足概率公理，我们引入归一化因子Z：

$$Z=\sum_{v_i}\prod_{c\in C}\Phi_{v_c}(v_c) \tag{3-8}$$

其中C为最大团的集合，$\Phi_{v_c}(v_c)$就是势函数，势函数本身并不具有概率意义。利用Hammerslery-Clifford定理，可以得到无向图标记序列联合概率为：

$$P(v_1,v_2,\cdots,v_N)=\frac{1}{Z}\prod_{c\in C}\Phi_{v_c}(v_c) \tag{3-9}$$

对于标记序列(X,Y)，无向图的标记序列的联合概率为：

$$p(Y\mid X)=\frac{1}{Z(x)}\exp(\sum_k l_k f_k(y_{i-1},y_i,X,i)) \tag{3-10}$$

其中，z(x)为归一化因子，以使P(Y|X)的所有可能的序列和满足概率公理。

$$z(x)=\sum_{y\in Y}\exp(\sum_{k}l_k f_k(y_{i-1},y_i,X,i)) \quad (3\text{-}11)$$

定义每个势函数的形式如下：

$$\Phi_{y_c}(y_c)=\exp(\sum_{k}\lambda_k f_k(c,y\mid c,x)) \quad (3\text{-}12)$$

通过对联合概率的分解得到转移特征和状态特征的权重参数，然后通过最大似然估计原理和L-BFGS算法进行参数训练，使得P(Y|X)的对数似然度最大，最后使用动态规划的算法得到最优序列Y*:

$$Y^{*}=\arg\max_{Y}(p_l(Y\mid X)) \quad (3\text{-}13)$$

条件随机场在自然语言中有非常重要的应用，根据研究成果表明[49]，条件随机场在序列标注问题上优于其他算法。

3.3 特征选取

在机器学习理论中，特征选择对性能的提升起着关键性的作用。在文本情感分析中需要将文本信息转换为特征向量或者其他机器学习模型所需的形式，如何选择对文本情感分析有重要意义的特征是研究学者一直在探索的问题。目前常用的特征有词特征、词频特征、词性、句法信息、否定特征、主题相关特征等等。在英文中我们还可以使用首字母大写、词根等英语语言专有的一些特征；在中文处理中，常常以词作为一个单位而不像英文以字为单位，我们可以使用中文分词等可以作为一个很重要的特征加入模型中。

3.3.1 词性特征

词性是在自然语言处理中经常使用的语法特征，文本情感分析也不例外。词性指作为划分词类根据的词的特点。词性表明了一个词在句子中的作用。

按照现代汉语词的划分共有12类词性，其中实词包括名词、动词、形容词、数词、量词和代词共6种；虚词包括副词、介词、连词、助词、拟声词和叹词。在文本情感分析中，极性词往往为形容词性或副词性短语，评价对象常常是名词或者名词性短语。正确识别词性对文本情感倾向性分析具有很大的帮助。例如句子“操作系统使用起来非常繁琐，比Linux差远了”，进行词

性标注后得到“操作系统/nl 使用/v 起来/vf 非常/d 繁琐/a ，/wd 比/p Linux/x 差/a 远/a 了/ule”，添加词性后，产品属性“操作系统”的词性为nl（名词），可以作为一个名词来看待，评价词“繁琐”的词性为a（形容词），这样对产品属性抽取提供了很大的便利，也提高了产品属性抽取的效果。另外，根据研究表明除名词和形容词外，动词对整个句子的情感表达也起到了一定的作用[51]。

3.3.2 浅层句法特征

浅层句法分析（shallow parsing），也称为组块分析或者部分句法分析（partial parsing），用来识别句子中某些句法结构相对简单、功能和意义比较重要的成分，如主语、谓语、宾语等。浅层句法分析的结果并不是构建一棵完整的句法分析树，它的目的是简化自然语言处理分析的复杂度，以便提高分析的性能。起初句法分析只是针对基本的名词性短语和介词短语，后来慢慢扩展到对所有类型的短语都进行识别分析[50]。例如，对句子：

佳能A系列的经典毋庸置疑，让用户可以花较少的钱买到功能强大的手动相机，适合摄影启蒙和入门学习。

进行浅层句法分析后得到以下结果序列：

BNP[佳能/nz A/x 系列/n] 的/udeq 经典/n 毋庸置疑/vl ，/wd 让/v 用户/n 可以/v 花/n BNP[BADJP[较/d 少/a] 的/ude1 钱/n] 买/v 到/v BNP[功能/n] 强大/a 的/ude1 手动/b 相机/n ，/wd 适合/v BVP[摄影/vn 启蒙/vn] 和/cc BVP[入门/vn 学习/vn] 。

在上述的例子中，“佳能A系列”是一个产品属性名，经过分词后该短语被分为“佳能/A/系列”。当产品属性为一个短语时，这样的分词结果导致程序无法正确的识别该短语，然而这样的短语恰恰是我们需要的产品属性，影响了我们抽取产品属性的精度。

在产品评论中，产品属性往往是名词或名词性短语，我们可以利用浅层句法分析技术识别出这些名词或者名词性短语，这样不仅有利于产品属性的识别，而且还能对句子结构进行简单的分析处理，极大地方便了句子情感极性的识别。

3.3.3 其他特征

（1）词特征及字特征

在英文中，由于每个单词即为一个词，可以标注单词的词性而无需再分词；在中文中词是由字来组成，为了在计算机上进行深层次的分析，需要对一句话进行分词。在中文自然语言处理任务中分词比英文要复杂，处理的方式也有所不同。例如句子“操作系统使用起来非常繁琐，比Linux差远了”，进行分词后得到“操作系统/使用/起来/非常/繁琐/，/比/Linux/差/远/了”，在使用的机器学习模型“操作系统”当作一个单元来看待，免去了因识别词而导致的错误。在我们的产品属性抽取实验中，我们发现使用词特征比使用字特征效果有很大的提升。

（2）上下文信息

上下文信息对信息处理起到了非常重要的作用，在产品属性抽取中，产品属性的识别下上文特征起到了重要的作用。产品属性附近一般存在修饰性的形容词或者评价词。在文本实验中，大多数实验把词看作一个单位，每个词可以对前后4个词进行联合分析，即窗口为[-4,+4]。

3.4 算法描述

目前，抽取句子中的产品属性有很多方法，但是总体效果不是很理想。基于以上研究工作，文本将条件随机场模型引入产品属性抽取任务，并充分利用了条件随机场模型在序列标注的优势。我们将产品属性的识别看作序列标注问题，每一个句子作为一个处理独立单元。在每个句子中，以句子分词后的词为单位，按是否为产品属性进行标记：B-t标记是产品属性的开始，I-t标记是产品属性的中间部分，O标记表明该词不是产品属性。

除了选择性能好的模型外，特征也是提高产品属性抽取效果的重要一部分。经过大量实验分析得出，同时使用词、词性、浅层句法特征和上下文信息对产品属性识别有很大的帮助。训练集特征如表3-1所示。

产品属性抽取算法的主要思想是将训练集以句子为单位进行切分，对于每个句子进行中文词性标注工作，并利用浅层句法分析计算技术识别句子中的名词性短语等组块，然后将识别的组块和词性标注都作为CRF的特征来使用。将训练集表示为这些特征后在CRF模型中进行训练。本文产品属性抽取的亮点是加入了浅层句法特征，这一特征的加入使得我们的实验效果有了很大的提升。产品属性抽取系统的过程如图3-1所示。

表3-1 产品属性抽取特征示例

词	词性	浅层句法	产品属性标记
32	m	BNP	B-t
和弦	n	BNP	I-t
铃声	n	O	I-t
非常	d	O	O
动听	a	O	O
。	wj	O	O

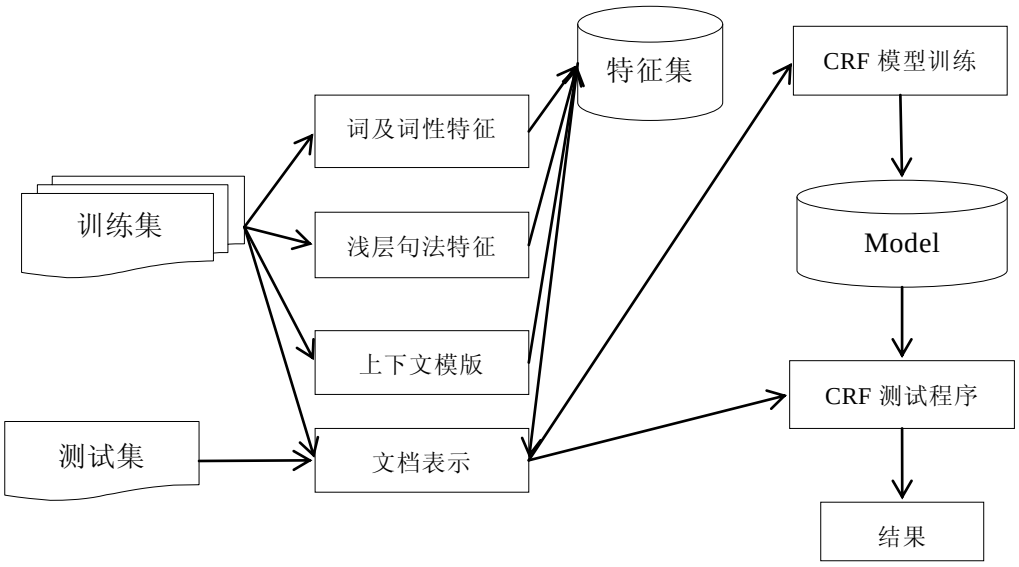

图3-1 产品属性抽取过程

3.5 实验结果与分析

3.5.1 实验设置

在产品属性抽取实验中采用的机器学习模型为最大熵模型和条件随机场模型，使用的特征有词及词性特征、字特征、浅层句法特征和上下文特征。

语料：使用本实验标注后的COAE中电子产品领域语料，共包括312篇文

章，其中相机评论语料136篇，笔记本评论语料53篇，手机评论语料123篇，总共评价对象个数为6000个，评价词个数4747个。

评价指标：本文采用准确率、召回率及F1值作为评价指标，

其中，准确率是程序识别正确的数目与程序识别数目的比值。召回率是程序识别正确的数目与语料中正确答案的比值。F1值是一个综合衡量指标，它综合考虑了准确率和召回率，使用两者的调和平均数。为了更准确的衡量实验的性能，本文中的准确率、召回率及F1值有精确评价、模糊评价之分，精确评价是指产品属性识别严格匹配时视为正确匹配，覆盖评价是产品属性识别包含于标注答案时视为正确匹配。

本实验采用的CRF++0.53工具包，通过实验分析采用表3-2所示的上下文模版可以取得较好的实验效果。

表3-2 条件随机场模型上下文窗口模版

词/字特征	词性特征	浅层句法信息特征
U03:%x[-4,0]	U20:%x[-4,1]	U220:%x[-4,2]
U04:%x[-3,0]	U21:%x[-3,1]	U221:%x[-3,2]
U05:%x[-2,0]	U22:%x[-2,1]	U222:%x[-2,2]
U06:%x[-1,0]	U23:%x[-1,1]	U223:%x[-1,2]
U07:%x[0,0]	U24:%x[0,1]	U224:%x[0,2]
U08:%x[1,0]	U25:%x[1,1]	U225:%x[1,2]
U09:%x[2,0]	U26:%x[2,1]	U226:%x[2,2]
U10:%x[3,0]	U27:%x[3,1]	U227:%x[3,2]
U11:%x[4,0]	U28:%x[4,1]	U228:%x[4,2]
U14:%x[-2,0]/%x[-1,0]	U31:%x[-4,1]/%x[-3,1]	U231:%x[-4,2]/%x[-3,2]
U15:%x[-1,0]/%x[0,0]	U32:%x[-3,1]/%x[-2,1]	U232:%x[-3,2]/%x[-2,2]
U16:%x[0,0]/%x[1,0]	U33:%x[-2,1]/%x[-1,1]	U233:%x[-2,2]/%x[-1,2]
U17:%x[1,0]/%x[2,0]	U34:%x[-1,1]/%x[0,1]	U234:%x[-1,2]/%x[0,2]
	U35:%x[0,1]/%x[1,1]	U235:%x[0,2]/%x[1,2]
	U36:%x[1,1]/%x[2,1]	U236:%x[1,2]/%x[2,2]
	U37:%x[2,1]/%x[3,1]	U237:%x[2,2]/%x[3,2]
	U38:%x[3,1]/%x[4,1]	U238:%x[3,2]/%x[4,2]

表3-2 条件随机场模型上下文窗口模版（续）

特征之间关系上下文信息
U39:%x[-3,1]/%x[-2,1]/%x[-1,1]
U40:%x[-2,1]/%x[-1,1]/%x[0,1]
U41:%x[-1,1]/%x[0,1]/%x[1,1]
U42:%x[0,1]/%x[1,1]/%x[2,1]
U43:%x[1,1]/%x[2,1]/%x[3,1]
U44:%x[-4,1]/%x[-3,1]/%x[-2,1]/%x[-1,1]
U45:%x[-3,1]/%x[-2,1]/%x[-1,1]/%x[0,1]
U46:%x[-2,1]/%x[-1,1]/%x[0,1]/%x[1,1]
U47:%x[-1,1]/%x[0,1]/%x[1,1]/%x[2,1]
U48:%x[0,1]/%x[1,1]/%x[2,1]/%x[3,1]

本文使用最大熵模型作对比实验，使用的特征和条件随机场模型是相同的，ME模型的上下文窗口为1，即当前词仅与上一个词和下一个词有关系。特征如表3-2所示。

表3-3 最大熵模型上下文窗口模版

MaxEnt特征	说明
T	评价对象
TPOS	评价对象词性
Tchunk	chunk所标注的块名
Tleft	评价对象左边的词
Tright	评价对象右边的词
POSTleft	评价对象左边词的词性
POSTright	评价对象右边词的词性
Tchunkleft	chunk所标注的块名_左侧
Tchunkright	chunk所标注的块名_右侧

3.5.2 实验结果分析

3.5.2.1 字、词比较实验

为了比较字特征和词特征的在产品属性抽取中的影响，本实验数据质采用手机领域语料包括123篇文档，2377个句子，2396个产品属性。使用的机器学习模型为条件随机场，CRF_word特征为词、词性及上下文信息，CRF_letter的特征为字、字类型、字位置特征。训练集与测试集的比例为5:1。结果如表3-4所示。

表3-4 phone数据集字、词比较实验

特征选择	精确评价			覆盖评价		
	准确率	召回率	F1值	准确率	召回率	F1值
CRFs_letter	28.97	32.88	30.80	54.37	**61.71**	57.81
CRFs_word	**53.17**	**39.64**	**45.42**	**71.30**	53.15	**60.90**

由实验结果可以看出，使用词作为特征来识别产品属性起到了很好的效果，无论精确评价还是覆盖评价，准确率和召回率都有很好的提升，其中精确F1值高出使用字特征比例22%。本实验说明产品属性识别是以词为单位的，而不是以字为单位。为了进一步提高实验结果，下一步我们将对其他特征进行分析实验。

3.5.2.2 CRF和MaxEnt多特征比较实验

为了继续验证其他特征，我们将进行更多的实验。实验数据为COAE2008电子产品领域语料，为了测试条件随机场模型在产品属性抽取的性能，本实验采用CRF模型和MaxEnt模型进行比较，并在原来的基础上添加浅层句法特征（chunk）。为保证实验的合理可信，我们进行六折交叉验证，即任意5/6的数据作为训练集，剩下1/6的数据作为测试集，循环测试。结果如表3-5所示。

表3-5 CRF和MaxEnt多特征比较实验

特征选择	精确评价			覆盖评价		
	准确率	召回率	F1值	准确率	召回率	F1值
CRF_word	54.84	41.74	47.40	74.81	56.94	64.66
MaxEnt_word	36.67	37.23	36.94	59.60	60.52	60.04
CRF_word_chunk	**57.95**	**45.80**	**51.16**	**78.61**	**62.13**	**69.41**
MaxEnt_word_chunk	38.49	38.13	38.30	63.15	62.54	62.82

由实验结果可以看出，由CRF_word与CRF_word_chunk可以看出，添加浅层句法特征后，准确率和召回率都有明显的提升，这说明数据集中的产品属性更多时候为名词或者名词性短语，而进行中文分词时并不能识别名词性短语。实验证明添加浅层句法特征可以很好的提高产品属性抽取效果。

通过对比CRF_word、MaxEnt_word和CRF_word_chunk、MaxEnt_word_chunk，我们可以看到条件随机场模型的性能要好于最大熵模型，这说明条件随机场模型更能胜任序列标注任务。

3.5.2.3 与COAE结果对比实验

本文对phone、camera、notebook三个领域数据集分别进行实验，使用CRF模型，特征为词、词性、浅层句法特征，并与COAE2008的最好成绩进行比较。实验结果如表3-6所示。

表3-6 与COAE结果对比实验

数据集	精确评价			覆盖评价		
	准确率	召回率	F1值	准确率	召回率	F1值
camera	50.97	35.79	42.06	72.95	51.22	60.19
phone	56.79	36.31	44.30	77.61	49.62	60.54
notebook	58.43	42.00	48.87	76.92	55.29	64.33
平均值	55.40	38.03	**45.08**	75.83	52.04	**61.69**
COAE2008_max	**56.41**	**41.72**	39.76	72.06	**57.88**	51.69

本实验方法无论在精确评价还是覆盖评价的F1值都要好于COAE结果，说明本实验方法在抽取产品属性这一任务上具有明显的优势。然而，我们的实验结果在绝对值上远远小于实用的标准，一方面，我们使用的语料直接取自网络真实数据，数据噪声很大；另一方面，我们使用的分词工具和浅层句法分析工具并不完善，导致实验有累计误差。我们还要不断的改进方法使产品属性抽取结果更好。

3.6 本章小结

本章开始介绍了产品属性抽取的相关技术，引入了最大熵、条件随机场

模型并对介绍模型的理论。本章对条件随机场模型和最大熵模型在产品属性抽取任务中进行了实验分析比较。此外，还对词性、浅层句法等相关特征选取进行了细致的介绍，分析加入这些句法特征对产品属性特征抽取的影响。在实验的基础上提出基于条件随机场模型的产品属性抽取技术，对产品属性抽取实验进行了细致的分析并指出了实验中的不足。

第4章 文本情感分析跨领域移植技术

4.1 引言

在统计自然语言中需要使用大量的语料，在条件随机场模型中也需要一定规模的语料进行训练。研究工作表明，使用有监督学习方法的精确率要高于无监督学习方法，然而要想进行有监督学习需要标注大量的相关语料，并且当训练好的模型移植到新的领域后，需要重新标注大规模训练语料，这是非常费时费力的。文本情感分析跨领域移植研究的就是如何在尽量少的标注训练语料的条件下，将在原有领域上的模型移植到新的领域上，并且新的模型在目标领域上的性能高于原模型的性能。

为了进行跨领域移植，大多数做法是在训练语料中加入小部分标注的目标语料，比如文献[33]将汽车领域评论语料作为源领域（source domain）集，电子产品领域评论语料作为目标领域（target domain），训练语料使用汽车领域语料和小部分人工标注的电子产品领域语料，并使用MAP条件随机场领域自适应算法进行跨领域移植，移植效果有了34%的相对提高。本文提出基于主动学习的产品属性领域移植的算法，产品属性抽取性能要好于文献[33]的算法。

4.2 主动学习方法

在机器学习的算法中，基本上都是在无论是样本分类还是序列标注，都是类别归属问题。即给定一个或多个数据集，判断测试集数据的类别。按照是否使用标注数据，可以将机器学习方法分为有监督学习(Supervised learning)、无监督学习(Unsupervised learning)以及半监督学习(Semi-supervised learning)。有监督学习是指通过已知的输入数据及其输出数据的对应关系生成一个映射函数，测试时输入数据通过此函数计算对应输出。比如文本分类算法大多是有监督学习的方法；而无监督的学习方法是在输入数据集上直接进行建模分析，不需要训练数据的指导，比如文本聚类算法属于无监督的方法。半监督学习方法是介于有监督和无监督学习方法之间的一类方法，它综

合利用了有标注的数据集和无标注的数据集，通过两者进行映射一个函数。

在一般情况下，使用有监督的学习的效果往往要好于使用无监督的学习方法，然而在实际情况中可以获得某个领域大量的未标注数据，针对海量的数据标注某个领域的语料是不仅低效而且不现实的。半监督学习的目的是如何在训练数据缺失的条件下，研究如何修改方法提高系统性能。常见的半监督学习算法有EM算法、直推向量机和协同训练等等。

主动学习[52]是一种有监督的学习方法，通过机器学习方法迭代的形式从候选样本中按照某种策略选择样本进行训练，即使用人机交互的查询方式以便在新输入数据上获得输出数据。主动学习也是在大量未标注语料的基础上，通过能够为数据标注的“神谕”（oracle，通常是领域专家）进行交互学习。主动学习可以有效地降低样本的复杂度，近年来该研究已经取得了较大的发展。

主动学习的主要思想是，存在一个已经标注的数据语料K（K可以为空），针对未标注的数据语料U（U中有大量的未标注数据），挖掘K中的文本信息，通过某种策略在集合U中找出一个子集C，将集合C提交给oracle进行数据标注后，进行下一次循环迭代。主动学习的学习过程可以分为两部分：学习部分和选择部分，学习部分提供一个分类器，使用在标注的数据集上进行有监督学习；选择部分是使用学习后的模型，用于未标注数据集，并按一定的方法选择子集交给oracle标注。主动学习与半监督学习不同，主动学习需要进行人工交互的学习；而半监督学习仅仅通过自我学习更新模型。主动学习的目的是尽可能的减少语料标注以取得与标注所有语料相同的结果。

主动学习一个很重要的问题是如何选择样本。根据获取未标注数据样本方式不同，主动学习算法可以分为基于流（stream-based）和基于池（pool-based）的学习策略，基于流的方法是对依次到来的每一个查询样例，判断其是否需要提交给oracle进行标注；基于池的方法是针对新来的样例进行缓存，当缓存到一定数目时，在缓存中按照标准选择需要oracle标注的样例。基于池的学习策略是当前应用最多的、理论最充分的一种策略。

基于池的主动学习方法根据选择样本标注的方法不同，可以分为基于不确定度的方法（Uncertainty Sampling）、投票选择法（Query-By-Committee）、模型变化期望（Expected Model Change）、误差减少期望（Expected Error Reduction）、方差减少原则（Variance Reduction）及重量密度（Density-Weighted Method）等多种方法[54]。其中最常用的是基于不确定度的方法，主动学习器选择最不确定的那些样本提交给oracle进行标注。这种方法对概率模型来说是

非常便捷的选择策略，比如二元分类问题，不确定度选择方法只需要找到那些值为0.5的中间数据提交oracle标注[55]。投票选择法是指使用多个模型在标注语料中训练，对所有模型结果值进行加权统计确定样本是否进行oracle标注。模型变化期望方法会预计每个样例对当前模型的改变程度，并对那些最能影响模型变化的样例进行oracle标注。误差减少期望通过衡量样本对泛化误差减少的幅度，来决定该样本是否进行oracle标注。方差减少原则和误差减少期望类似，只是衡量误差的方法不同。每一种方法都有适用情况，选择合适的主动学习策略对实验结果有很大的帮助。

4.3 基于主动学习的文本情感移植

为了提高产品属性的领域移植效果，文本提出了基于主动学习的条件随机场模型适应算法，算法的主要思想是给定源领域标注数据集L和目标领域未标注数据集U，将数据集L作为训练集并训练一个产品属性标注器M，利用标注器M对U数据集进行标注，使用不确定度方法选择结果中的句子集S进行人工标注，将数据集L和数据集S合并为训练集L’，如此循环迭代一直到达到某个阈值。算法描述如如图4-1所示。

为了验证选择不确定度方法对提高产品属性抽取性能的影响，在实验中我们选择了不确定最高的样本和确定性最高的样本分别标注，分析样本选择方法对产品属性抽取的影响。在条件随机场模型中通过计算每个句子的置信度作为该句的不确定度。句子标注置信度的格式如表4-1所示。

对于每一个句子，通过CRF结果计算一个置信度，然后按照置信度选择句子中最不确定/确定的句子进行oracle标注，并将标注后的句子从未标注数据集中移除，加入到训练集。

4.4 实验结果与分析

4.4.1 实验设置

由第3章的实验结果分析，条件随机场模型在序列标注问题上有很大的优势，在文本情感跨领域移植实验中产品属性抽取仍使用条件随机场模型。CRF模型使用的特征有词及其词性特征、字特征、浅层句法特征和上下文特征。

本实验语料为COAE2008语料，其中电子产品领域语料312篇文章，包括

相机评论语料136篇，笔记本评论语料53篇，手机评论语料123篇，总共评价对象个数为6000个，评价词个数4747个。汽车领域共161篇文章，句子数4347个，包含3033个评价对象。电子产品评论语料在网络上是比较丰富且容易获得的，可以作为源领域语料，汽车评论语料相对规模较小，作为目标语料。我们将汽车评论语料按7:3比例分为未标注语料和测试语料，以便观察领域移植实验的效果。

DOAMIN_ADAPTATION(L, U, T)

输入：

L：源领域标注数据集 source labeled data;

U：目标领域未标注数据集 target unlabeled data;

T：目标领域测试集 targe test data

输出：

M：产品属性序列标注模型;

算法：

step 1：将数据集 L 和数据集 U 中数据处理为 CRF 所需的格式;

step 2：Iter←1,

step 3：在数据集 L 上使用 CRF 模型训练出模型 M;

step 4：使用模型 M 在测试集 T 上进行测试，得到结果 T_precision

step 5：while T_precision < threshold

step 6：使用模型 M 在数据集 U 上输出结果 U_result;

step 7：使用不确定度方法选择 U_result 中数据子集 S;

step 8：将数据集 S 进行 oracle 标注答案;

step 9：L←L+S;

step10 U← U-S;

step 11：Iter←Iter+1;

step 12：跳转到 step 3;

step13：返回标注模型 M;

图4-1 基于主动学习的领域移植算法

本实验条件随机场模型采用CRF++0.53工具包，使用的上下文特征与第3章的上下文特征相同。根据经验值，每次迭代选取50个句子进行oracle标注，并分析加入目标领域标注样本对跨领域移植的性能影响。

表4-1 句子置信度标注实例

词	词性	浅层句法	产品属性标记
#0.747176			
要	v	O	O
把	qv	O	O
赚钱	v	O	B
环节	n	BNP	O
多元化	vn	BNP	O
。	wj	O	O

4.4.2 实验结果

4.4.2.1 基于不确定性的主动学习策略

为了测试加入目标语料数目对实验结果的影响，利用训练出的模型对未标注语料进行序列标注，然后选择不确定性最高的1%样本进行oracle标注。实验结果如图4-2和图4-3所示。从图中可以看到，在未加入目标领域数据时，即使用电子产品领域数据集训练，直接对汽车领域数据集进行测试，精确评价F值为30%，模糊评价F值为40.1%。第一次迭代加入1%的目标样本后，无论是精确评价还是模糊评价，各个指标都有较大幅度的提升，其中精确评价F值达到36.3%，模糊评价F值达到48.2%。我们发现迭代次数超过11次后（图中未画出），各个指标趋于平缓。从总体上来说，各个指标随着迭代次数的增加而增加。然而，随着目标领域数据的不断加入，F值提升的幅度趋于缓和。

实验的详细结果如表4-2所示，使用电子产品领域语料直接对汽车领域测试称作直接测试。加入10%的目标语料时精确召回率提高了55%，精确评价F1值提高了35.5%。当加入5%的目标语料时，精确评价F1值提高了31%，覆盖评价F1值提高了32.4%。当加入目标语料超过10%时精确评价和模糊评价指标均趋于平缓。目标语料为5%时，各个指标处于次高峰期，这时曲线的斜率趋向0。权衡标注工作和实验效果，选择标注5%的目标语料是性价比最高点。

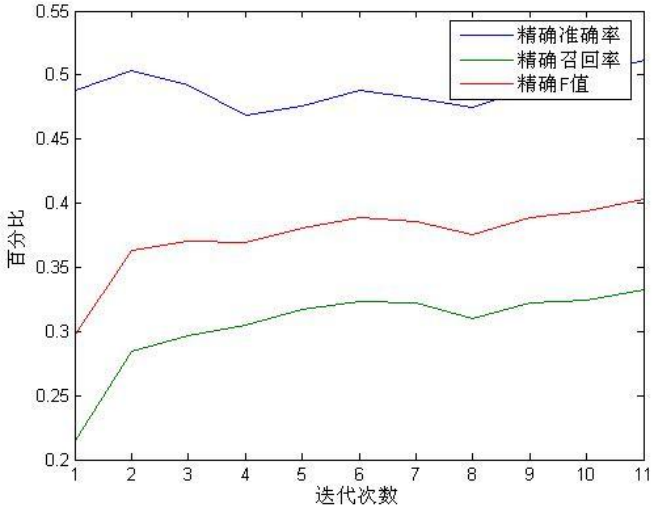

图 4-2 不确定性样本选择精确评价指标趋势

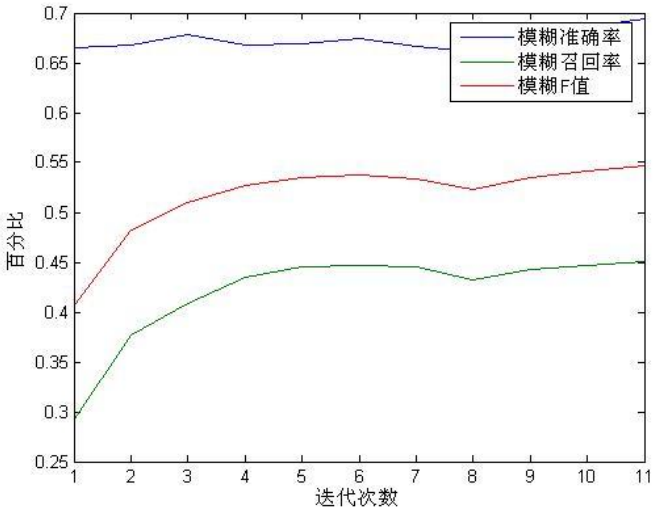

图 4-3 不确定性样本选择模糊评价指标趋势

4.4.2.2 基于确定性的主动学习策略

为了验证选择不确定性主动学习提高的效果，本实验选择确定性高的特征进行oracle标注，实验效果如图4-4和图4-5所示。

表4-2 基于主动学习的评价对象领域移植结果

训练数据	精确评价			覆盖评价		
	准确率	召回率	F1值	准确率	召回率	F1值
直接测试	48.75	21.37	29.71	66.57	29.18	40.58
1%目标语料	50.32	28.45	36.35	66.74	37.73	48.21
5%目标语料	48.80	**32.36**	38.91	67.40	44.69	53.74
10%目标语料	**51.13**	33.21	**40.27**	**69.36**	**45.05**	**54.63**

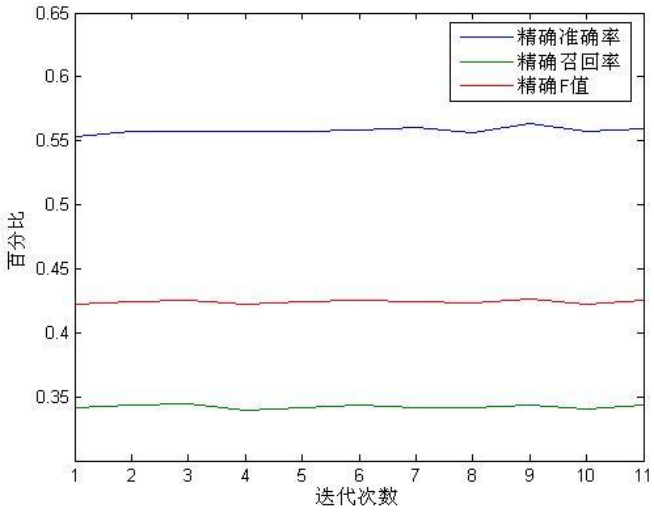

图4-4 确定性样本选择精确评价指标趋势

由实验结果可以看出无论是精确评价还是模糊评价，各个指标的增长趋于平缓，这说明选择不确定性样本进行标注对效果提升有很大的帮助。我们分析原因可能是确定性的句子都是比较短的句子，含有的信息量比较少；不确定性的句子大多数是长句子，含有的信息量比较大。从概率上讲，分类器对确定性的样本置信度高，不需要过多的额外信息即可正确标注；而分类器对不确定性的样本不能以高置信度进行标注。另外语料库规模和标注的不规整性也可能导致评价对象识别率不高，从整体来说，主动学习中选择不确定性样本的方法可以很好的提高产品属性抽取的移植效果。

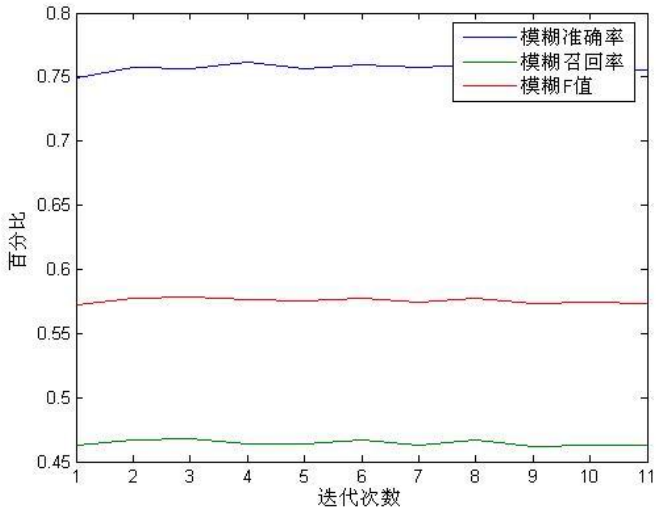

图4-5 确定性样本选择精确评价指标趋势

4.5 本章小结

本章研究了文本情感中评价对象跨领域移植的相关工作，首先介绍了主动学习方法，并比较了半监督学习和主动学习之间的区别。针对目前文本情感语料缺乏的现状，文本将主动学习方法引入评价对象跨领域移植任务。文本在产品评论语料进行实验并从电子产品领域移植到汽车领域，实验结果证明了本方法在情感领域移植具有很好的效果。最后，对可能影响实验结果的因素进行分析。

第5章 文本情感分析系统设计与实现

5.1 引言

随着文本情感分析研究的深入和不断应用，越来越多系统加入了情感分析模块，比如twitter[7]加入了多种情感分析工具，用来监控社交舆论和分析用户的情绪；语义搜索引擎Evri[8]推出了情感网络API，可以用于市场调研、体育娱乐、品牌管理、产品评测等多个应用功能。

文本情感分析的主要功能是对文本信息进行极性分析并抽取出用户所需的信息。根据分析粒度的不同，文本情感倾向性分析可以分为词语情感倾向性分析、句子情感倾向性分析、篇章情感倾向性分析、海量信息的整体倾向性预测。词语情感分析系统主要是对词语的极性进行判断、标注等一体化工作，如du[11]系统实现了对网络中的产品情感词进行挖掘抽取工作。句子情感分析主要是对句子的极性进行判断,抽取句子中用户所关心的情感属性信息。本文实现的系统主要面向句子的情感分析，实现了对产品领域中句子极性的判断，抽取产品属性及其对应的评价极性词。篇章情感分析主要是对一篇文档在总体上进行极性判断，篇章情感分析主要用于情感信息统计，如twitter中的Tweetfeel[9]。海量情感分析是对网络中的信息进行趋势分析，如Evri搜索引擎。

在情感分析的研究工作的基础上，本文设计了一个情感分析系统，主要的功能包括评价对象抽取、极性词识别、评价对象极性判断等。

5.2 评价对象和评价词一体化识别

识别产品属性等评价对象不仅仅是找到句子中的产品属性，还要识别句子中所对应的极性词并判别极性。识别情感单元的极性词是文本情感分类的一个重要环节。根据文本的研究工作，文本提出了评价对象、评价词一体化

[7] http://www.twitter.com

[8] http://www.evri.com

[9] http://www.tweetfeel.com

识别方法。识别后的结果可以直接提供给文本情感分析系统。

5.2.1 一体化识别方法

在抽取产品属性实验中，基于条件随机场模型的情感抽取方法取得了很好的效果。CRF模型识别产品属性模型可以看作序列标注问题，因此可以再CRF模型中同时识别评价对象和评价词。为了提高识别的效果，CRF模型中除了加入词、词性、浅层句法和上下文特征以外，还加入了情感词信息特征，以便提高极性词识别的效果。情感词信息用二元特征来表示当前词是否为情感词词典中的词语。对比产品属性标记，在训练语料中引入了两个标记B-s和I-s，B-s标记分词后的词语是情感词的开始部分，I-s标记分词后的词语是情感词的中间和结尾部分。标记后的训练数据如表5-1所示。

表5-1 一体化识别标注范例

词	词性	浅层句法	情感词信息	产品属性标记
如此	rzv	O	0	O
奢华	a	O	1	B-s
的	ude1	O	0	O
配置	v	O	0	B-t
展现	v	O	0	O
出	vf	BVP	0	O
强劲	a	BVP	1	B-s
的	ude1	BNP	0	O
性能	n	BNP	0	B-t
，	wd	BNP	0	O
有点	d	BNP	0	O
台式	b	O	0	O
机	ng	BNP	0	O
的	ude1	BNP	0	O
感觉	n	BNP	0	O
了	y	BNP	0	O

其中第4列特征使用0、1进行标注，1表示当前词为极性词，0表示当前词

非极性词。

5.2.2 实验结果与分析

本实验使用本实验标注后的COAE中电子产品领域语料。本实验使用的评价指标仍为精确评价和模糊评价的准确率、召回率和F1值，并分别对评价对象和极性词极性进行评估。实验结果如表5-3所示。

表5-2 评价对象评价词一体化识别结果

评价目标	精确评价			覆盖评价		
	准确率	召回率	F1值	准确率	召回率	F1值
极性词	80.26	63.99	71.21	84.65	67.48	75.10
评价对象	**58.94**	45.69	**51.49**	76.59	59.36	66.89
CRF_word_chunk对比	57.95	**45.80**	51.16	**78.61**	**62.13**	**69.41**

表5-2中，CRF_word_chunk是由第三章使用的模型训练得到的结果，我们可以看出一体化识别后评价对象的精度并不会显著降低，这证明了评价对象评价词一体化识别的可行性。另外，极性词识别效果也达到了预期的水平。我们将本节实验用于情感系统数据处理。

5.3 情感分析系统架构

文本情感分析系统包括三个子模块：词语情感分析、句子情感分析、评价对象跨领域移植。词语情感分析实现了词语极性查询和情感词典操作功能；句子情感分析实现了句子极性的判断，抽取产品属性及其对应的评价极性词；评价对象跨领域移植是对第4章研究成果的实现，实现了从电子产品领域到汽车领域移植并展示了移植后产品属性识别效果。系统架构图如5-1所示。

浅层句法分析器将数据集进行语块识别，CRF模型使用情感词词典lexicon、浅层句法进行模型训练CRF_model，另外还需对评价对象跨领域移植进行模型训练。MaxEnt通过训练集训练模型MaxEnt_model。

模型CRF_model和MaxEnt_model用来评价对象、评价词一体化识别，模型adaption_model进行领域移植。

词语情感分析主要调用了情感词词典lexicon，句子情感分析使用了浅层

句法分析器、CRF模型、MaxEnt模型。领域移植模型训练需要oracle参与标注工作。

5.4 系统设计与实现

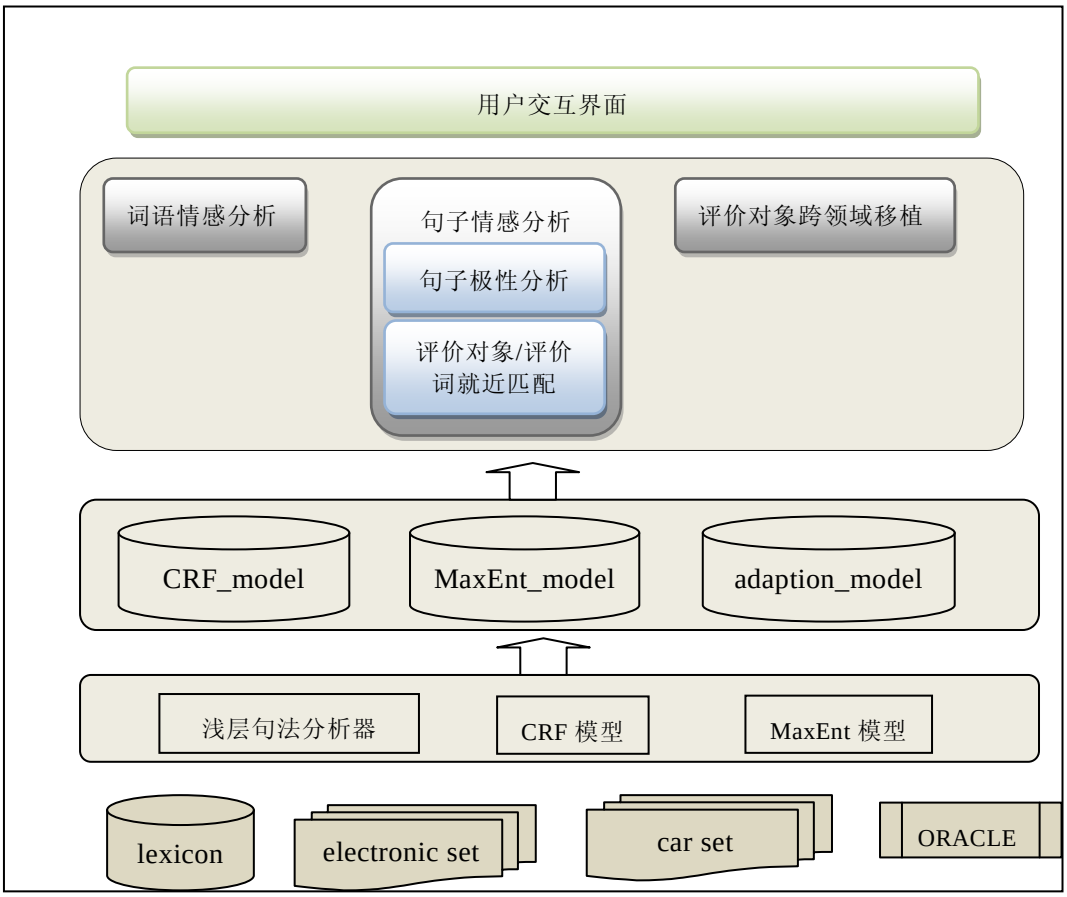

图5-1 情感分析系统架构图

5.4.1 算法设计

情感系统中句子情感分析是最复杂的一个模块，模块中的评价对象、评价词识别使用CRF模型标注的结果，如果CRF模型只识别了其中一个，我们利用邻近匹配的名词短语/形容词来作为评价对象/者评价词；如果两者都没有识别出，我们则认为此句子为客观句。

为了判断句子的极性，首先我们找到评价词，如果评价词前方视野内存

在否定词则翻转极性，句子的极性即为评价词极性的加权统计值。算法如图5-2所示。

Sentence_polarity(data)
输入： data：该句子在一体化识别后的标注结果; 输出： polarity：句子极性; 算法： step 1：polarity ← 0; sentiword_count ← 0; step 2：对于 data 中每个评价对象 target: step 3：若无评价词 setiword: step 4：sentiword←就近的形容词/形容词短语; step 5：对于每个极性词 sentiword step 6：tmp_polarity ← sentiword 在情感词词典极性 step 7：如果 sentiword 前 n 字内存在否定词: step 8：tmp_polarity ← - tmp_polarity; step 9：sentiword_count ← sentiwrod_count + 1; step 10：polarity ← polarity + tmp_polarity; step11：polarity ← polarity / sentiword_count; step 12：return polarity;

图5-2 句子极性分析算法

5.4.2 详细设计

本系统使用java语言实现界面功能，开发平台为eclipse3.4，操作系统为WindowsXP。浅层句法分析部分使用实验室现有功能模块，采用C++语言实现。

句子级情感分析系统中，评价对象、评价词抽取使用了CRF模型和MaxEnt模型两种机器学习模型。

图5-3 电子产品领域细粒度情感分析界面

图5-3展示了句子级情感分析的结果，可以识别大部分领域专有产品属性，总体识别效果比较好。对句子极性判断用汉字“褒义”、“贬义”来表示，标注句中的紫色字体为评价对象，红色字体为极性词。当前演示例句使用CRF模型和MaxEnt标注的结果相同。

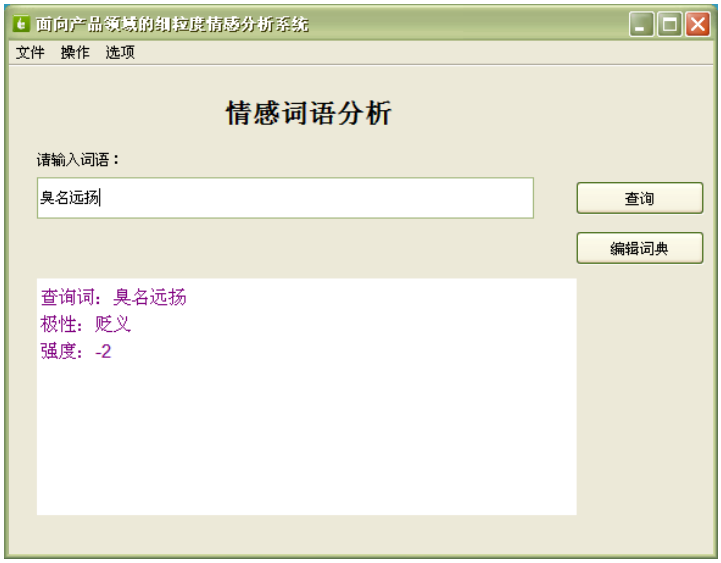

图5-4 情感词语查询分析界面

图5-4展示了情感词查询分析，系统调用词典里的查询词。另外，还可以对情感词词典进行编辑工作。

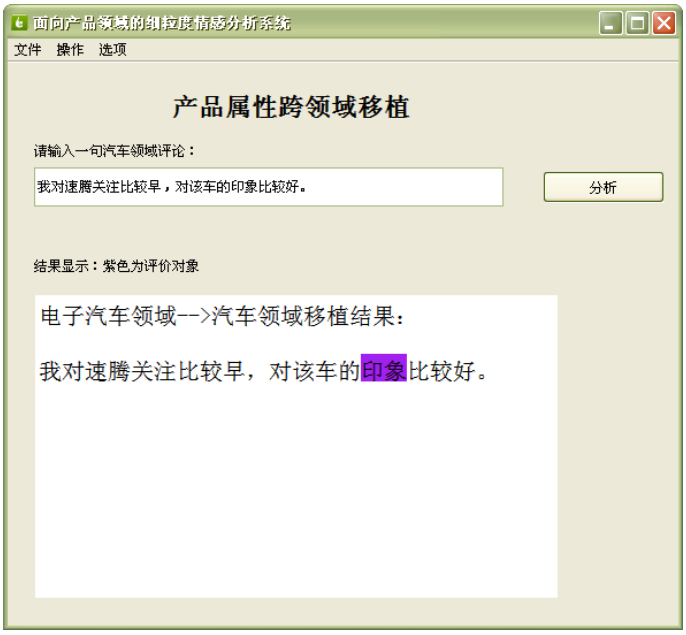

图5-5 产品属性跨领域移植效果图

图5-5展示了产品属性跨领域移植结果，语料使用电子产品领域数据集移植到汽车领域。根据第4章实验结果和多个演示示例检查分析，情感语料领域移植效果总体比较好。

5.5 本章小结

文本首先对目前情感分析应用情况进行了分析，然后提出了评价对象、评价词一体化识别算法并进行了相关实验分析。根据一体化识别的结果，本文设计了文本情感分析系统，对系统的架构和算法进行了说明。系统功能包括情感词查询、评价对象及评价词识别，句子极性计算，还展示了评价对象的领域移植效果。

结　论

文本情感分析是自然语言处理中的一个热门研究方向，它的目的是对文本中主观性的文本进行分析、整理及综合，以便得到人们期望的信息。从研究内容上来看，文本情感分析主要包括文本情感分类、情感信息抽取和情感信息检索和归纳等多个研究内容。

本文的研究内容主要包括以下几个方面的内容：

（1）研究文本情感资源的建设方法，借鉴现有的研究方法引入一个高效的方法，扩大情感资源的规模并提高情感资源的可信度，文本使用了三种方法进行词典扩展并对结果进行了校验，得到了一个可信度比较高的词典。

（2）探索评论语料中的产品属性（评价对象）抽取的研究方法，提高评论语料中的产品属性的准确率，目的是提高产品属性抽取的可应用性。通过选择机器学习模型和选取特征，提高了产品属性抽取的性能。

（3）探索文本情感分析中产品属性跨领域移植的方法，作为一个很新颖的课题，运用一种可适用性的方法来提高跨领域移植的精度，文本提出了基于主动学习策略的评价对象移植方法，从电子产品领域语料移植到汽车评论领域起到了很好的效果。

（4）在情感分析的研究工作的基础上，本文设计了一个情感分析系统，主要的功能包括评价对象抽取、极性词识别、评价对象极性判断等，可以方便直观的展现研究成果。

文本在文本情感倾向性分析研究工作中虽然有一些创新的应用，仍然还有一些不足和改进的地方，以后在以下几个方面加强研究：

（1）文本情感词的极性可以根据上下文的语境改变，如何计算动态情感词的极性是下一步词典构建重点考虑的问题。

（2）产品属性常常和极性词同时出现，如何抽取两者的搭配关系，识别产品属性对应的极性词对人们判别产品属性的情感极性有很大的帮助。搭配抽取是情感信息抽取一个非常重要的研究任务。

（3）文本情感分析研究对领域资源依赖比较大，面对目前标注语料比较稀缺的现状，如何充分利用已标注的语料是情感移植需要考虑的问题。文本情感跨领域移植的效果还不能达到实用要求。如何充分利用源领域和目标领域的关系，充分的挖掘两者的关系信息仍是文本情感移植努力的方向。

参考文献

[1] Bing Liu. Sentiment Analysis and Subjectivity[C]//N. Indurkhya and F. J. Damerau. Handbook of Natural Language Processing (Second Edition). USA：Chapman and Hall/CRC, 2010：627-666.

[2] B. Pang and L. Lee. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2（1-2）：23-60.

[3] Hatzivassiloglou and McKeown, Predicting the Semantic Orientation of Adjectives[C]//association for Computational Linguistics. proceedings of 35th Annual Meeting of the Association for Computational Linguistics, 1997：174-181.

[4] 黄萱菁，赵军. 中文文本情感倾向性分析[J]. 中国计算机学会通讯. 2008，4(2).

[5] Kim, S. and Hovy, E. Determining the sentiment of opinions[C]// Proceedings of the 20th international conference on Computational Linguistics (COLING), 2004.

[6] Valentin Jijkoun, Maarten de Rijke, Wouter Weerkamp. Generating Focused Topic-Specific Sentiment Lexicons[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 585-594.

[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[J]. 中文信息学报，2006，20(1)：14-20.

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

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

[10] Ahmed Hassan, Dragomir R. Radev. Identifying Text Polarity Using Random Walks[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010：395-403.

[11]Weifu Du, Songbo Tan, Xueqi Cheng, Xiaochun Yun. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon[C]. WSDM2010， 2010: 111-120

[12]G. Qiu, B. Liu, J. Bu and C. Chen. Expanding Domain Sentiment Lexicon through Double Propagation[C]//International Joint Conference on Artificial Intelligence (IJCAI-09), 2009：1199-1204.

[13]Wei Wei, Jon Atle Gulla. Sentiment Learning on Product Reviews via Sentiment Ontology Tree[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 404-413.

[14]王波，王厚峰. 基于自学习策略的产品特征自动识别[C]//内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集. 2007：509-514.

[15]章剑锋，张奇，黄萱菁，吴立德. 中文评论挖掘中的主观性关系抽取[C]//第三届全国信息检索与内容安全学术会议， 2007：675-681.

[16]John Blitzer, Ryan McDonald, and Fernando Pereira. Domain Adaptation with Structural Correspondence Learning[C]//Empirical Methods in Natural Language Processing (EMNLP). 2006：120-128.

[17]J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification[C]//Proceedings of the Association for Computational Linguistics (ACL), 2007：440-447.

[18]A Andreevskaia, S Bergler. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging[C]// Proceedings of ACL-08: HLT, 2008：290-298.

[19]Songbo Tan, Xueqi Cheng, Yuefen Wang, Hongbo Xu. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis[C]// European Conference on Information Retrieval (ECIR2009). 2009: 337-349.

[20]B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2002：79-86.

[21]B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web[C]//Proceedings of WWW, 2005.

[22]姚天昉. 一个用于汉语汽车评论的意见挖掘系统[C]//中文信息处理前沿进展-中国中文信息学会二十五周年学术会议论文集. 北京:清华大学出版社, 2006：260-281.

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

[24]A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: A case study[C]// Proceedings of Recent Advances in Natural Language Processing (RANLP), 2005.

[25]G. Ganapathibhotla, B. Liu. Mining Opinions in Comparative Sentences [C]//Proceedings of the International Conference on Computational Linguistics, COLING. 2008：241-248.

[26]L Zhang, SH Lim, B Liu, EO'Brien-Strain. Extracting and Ranking Product Features in Opinion Documents[C]//Proceedings of the 23rd International Conference on Computational Linguistics. 2010 ： 1462-1470.

[27]孙慧，关毅，董喜双. 中文情感词倾向消歧[C]//第六届中文信息检索学术会议. 黑龙江. 2010： 660-666

[28]J. Wiebe, T. Wilson and C. Cardie. Annotating expressions of opinions and emotions in language[J]. Language Resources and Evaluation, 2005, 1(2)：165-210.

[29]Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns[C]// Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). 2005.

[30]Xu Bing, Zhao Tiejun, Zheng DeQuan, Wang Shanyu. Product Features Mining Based on Conditional Random Fields Model[C]//International Conference of Machine Learning and Cybernetics, 2010：3353-3357.

[31]B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web[C]//Proceedings of WWW, 2005： 342-351.

[32]Hui Yang, Luo Si, Jamie Callan. Knowledge Transfer and Opinion

Detection in the TREC2006 Blog Track[C]//In proceedings of Text REtrieval Conference (TREC). 2006.

[33]张奇. 细颗粒度情感倾向分析若干关键问题研究[D]. 上海：复旦大学，2008.

[34]S.-M. Kim and E. Hovy. Identifying and analyzing judgment opinions[C]//Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL), 2006：200-207.

[35]R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections[C]//Proceedings of the Association for Computational Linguistics (ACL). 2007：976-983.

[36]M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs[C]//Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2008.

[37]Massimiliano Ciaramita, Olivier Chapelle. Adaptive Parameters for Entity Recognition with Perceptron HMMs[C]//Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL, 2010：1-7.

[38]Theresa Wilson, Janyce Wiebe, Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis[C]//Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language (HLT-EMNLP), 2005：347-354.

[39]Inquirer. http://www.wjh.harvard.edu/~inquirer. 2010.

[40]L.-W. Ku, Y.-S. Lo, H.-H. Chen. Test collection selection and gold standard generation for a multiply-annotated opinion corpus[C]// Proceedings of the ACL Demo and Poster Sessions, 2007：89-92.

[41]张伟，刘缙. 学生褒贬义词典[M]. 中国大百科全书出版社. 2004.

[42]王素格，李德玉，魏英杰. 基于同义词的词汇情感倾向判别方法[C]. 中文信息学报, 2009，23(5): 68-74.

[43]Robert Fano. Transmission of Information[M]. MIT Press and John Wiley and Sons, 1961.

[44]郑德权. 本体论和统计语言模型相结合的跨语言信息检索研究[D]. 哈尔滨：哈尔滨工业大学. 2006.

[45]R. Neches, R. E. Fikes, T. R. Gruber, et al. Enabling Technology for Knowledge Sharing[C]. AI Magazine, 1991, 12(3) :36-56.

[46]董振东，董强.“知网”. http://www.keenage.com. 1999.

[47]刘群, 李素建. 基于《知网》的词汇语义相似度的计算[C]. 第三届汉语词汇语义学研讨会, 2002.

[48]Andrew McCallum, Dayne Freitag, Fernando Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation[C]// Proceedings of the 17th International Conf. on Machine Learning, 2000：591-598.

[49]John Lafferty, Andrew McCallum, Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proceedings of the Eighteenth International Conference on Machine Learning, 2001：282-289.

[50]S.Buchholz, J.Veenstra, W.Daelemans. Cascaded Grammatical Relation Assignment[C]//Proceedings of EMNLP-99, 1999：239-246.

[51]Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL, 2003：25-32.

[52]1 A. P. Engelbrecht. Incremental Learning Using Sensitivity Analysis[C]. International Joint Conference on Neural Networks. 1999，2: 1350−1355.

[53]龙军, 殷建平, 祝 恩 赵文涛. 主动学习研究综述[J]. 计算机研究与发展. 2008，45(zl)：300-304.

[54]Burr Settles. Active Learning Literature Survey[M]. Computer Sciences Technical Report 1648 University of Wisconsin-Madison. 2010.

[55]D. Lewis and W. Gale. A sequential algorithm for training text classifiers[C]//Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, ACM/Springer, 1994:3-12.

攻读硕士学位期间发表的论文及其它成果

（一）发表的学术论文

[1] 徐冰,**王山雨**. 句子级文本倾向性分析评测报告. 第二届中文倾向性分析评测(COAE 2009), 2009.

[2] Xu Bing, Zhao Tiejun, Zheng DeQuan, **Wang Shanyu**. Product Features Mining Based on Conditional Random Fields Model. International Conference of Machine Learning and Cybernetics 2010.7:3353 – 3357.(EI 收录号：20104613374735)

[3] 徐冰，赵铁军，**王山雨**，郑德权. 基于浅层句法特征的评价对象抽取研究.自动化学报.2011（已录用）

（二）参与的科研项目情况

[1] 参与哈尔滨工业大学重点实验室开放基金项目“文本情感倾向性分析关键问题研究”.课题编号：HIT.KLOF.2009019

哈尔滨工业大学学位论文原创性声明及使用授权说明

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向产品领域的细粒度情感分析技术》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：王山雨　　日期：2011 年 06 月 27 日

学位论文使用授权说明

本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，即：

（1）已获学位的研究生必须按学校规定提交学位论文；（2）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（3）为教学和科研目的，学校可以将学位论文作为资料在图书馆及校园网上提供目录检索与阅览服务；（4）根据相关要求，向国家图书馆报送学位论文。

保密论文在解密后遵守此规定。

本人保证遵守上述规定。

作者签名：王山雨　　日期：2011 年 06 月 27 日

导师签名：郑德权　　日期：2011 年 06 月 27 日

致谢

时光如梭，两年硕士学习生涯即将结束。在哈尔滨工业大学的两年求学生活让我受益匪浅。在论文完成之际，由衷的感谢所有帮助过和支持我的老师、同学和家人。

感谢我的导师郑德权副教授，感谢两年来对我学习、科研和生活无微不至的关心和指导。郑老师渊博的知识、严谨的治学态度和丰富的科研经验深深的感染了我。当在学术上遇到困难时，郑老师总能帮我找到前进的方向。郑老师和蔼可亲的态度让我感觉到了家的温暖，让我非常感动。

感谢徐冰老师在学术上对我的指导，不断指导我做科研和思考问题的方法，在很多科研问题上进行了具体指导。徐老师学术和教学上敏锐的洞察力让我深深的佩服。

感谢赵铁军教授对我指导和教育，让我有机会在一个科研氛围浓厚的大家庭中学习和成长，您平易近人的态度让我由衷的感动。感谢李生教授，您严谨的学风、认真的工作态度是我一直学习的榜样。感谢杨沐昀老师的帮助和指导，辛勤负责的工作态度让我佩服不已。感谢李晗静老师对我的教导，让我明白了许多做学问的方法。感谢朱聪慧老师、赵华老师对我的指导和帮助，对我学习生活提供了许多便利。

感谢师兄刘水、李世奇、林建方、张春越、师姐刘莎莎等等对学习生活的帮助，让我感受到了你们的温暖。

感谢我的同学胡亚楠、刘海波、郑宏、郑博文、朱晓宁、李大任、于墨、王超、李震、王垚尧，在两年的时间里共同学习生活，留下了美好回忆。

感谢师弟孙振龙、吴建伟、胡鹏龙、何小春等对我的帮助，从你们身上学到了很多。

感谢父母和家人对我支持和帮助，是你们默默无闻的让我全身投入学习生活中。感谢爷爷奶奶的关心支持，给了我无尽的动力。

感谢所有曾经帮助过我的人，感谢机器智能与翻译实验室这个大家庭，给了我无尽的快乐和支持。