

硕士学位论文

情感标签抽取相关技术研究

RESEARCH ON SENTIMENT LABEL EXTRACTION

刘鸿宇

哈尔滨工业大学

2010 年 6 月

国内图书分类号: TP391.2
国际图书分类号: 681.37

学校代码: 10213
密级: 公开

工学硕士学位论文

情感标签抽取相关技术研究

硕士研究生: 刘鸿宇
导师: 秦兵教授
申请学位级别: 工学硕士
学科、专业: 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: 20010年6月24日
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

RESEARCH ON SENTIMENT

LABEL EXTRACTION

Candidate:	Liu Hongyu
Supervisor:	Prof. Qin Bing
Academic Degree Applied for:	Master of Engineering
Specialty:	Computer Science and Technology
Affiliation:	School of Computer Science and Technology
Date of Defence:	June 24, 2010
Degree-Confering-Institution:	Harbin Institute of Technology

摘要

随着 Web2.0 技术的蓬勃发展,互联网上产生了海量的用户评论信息,如何从这些评论中挖掘出有价值的信息,越来越受到研究者的关注。从产品评论中挖掘意见信息,一方面可以帮助用户在购买产品时作出决策,另一方面可以帮助商家即时了解用户对自己产品的意见。情感标签由评价对象和评价词组成,包含了用户评论的详细信息,能够有效地体现出用户评论的核心内容。为此,本文针对情感标签抽取中的三项任务:评价词集构建、评价对象识别以及情感标签抽取进行了深入研究。

在评价词集构建中,为了构建准确且全面的评价词集合,本文采用了融合语义知识库和大规模语料库的方法来获取候选评价词集合,进而通过候选评价词在语料库中的上下文为其设置置信度,根据置信度来度量候选评价词正确的可能性,最终选取置信度较高的评价词构成评价词集合。使用本方法构建的评价词集合参加了第一届中文倾向性分析评测中的任务一,取得了较好成绩。

在评价对象识别中,本文首先基于短语结构获取候选评价对象集合,进而针对评价对象具有领域相关性的特点引入了基于网络挖掘的 PMI(Pointwise mutual information)值过滤方法,针对评价对象中的名词冗余现象引入了名词剪枝算法,最终对评论句进行分类,以评价对象集为基础找出评论文本中用户进行评论的评价对象。本方法实现的系统参加了第一届中文倾向性分析评测中的任务三,取得了较好成绩。

情感标签抽取方面,本文提出了基于短语句法路径的情感标签抽取方法,本方法通过句法关系挖掘评价词与评价对象之间的修饰关系,解决了就近查找方法存在的经验性过强问题,同时,本文采取全自动的句法路径库获取方法,解决了传统人工制定规则方法存在的规则覆盖不全的问题,最后,本文在句法路径匹配的过程中引入了编辑距离进行松弛,从而有效的提高了系统召回率。

最后,针对传统情感标签抽取方法无法解决的隐式标签抽取问题,本文尝试使用主题模型对文本进行情感标签标注,提出了基于互信息和基于概率相似度的两种具体标注方案,实验结果表明主题模型在抽取隐式标签时能够起到一定的作用,本文最终对主题模型应用于情感标签标注存在的问题进行了详细的分析。

关键词: 情感分析;情感标签;评价对象;评价词;句法路径

Abstract

With the rapid development of web2.0 technology, the Internet produced lots of online reviews, more and more researchers focused on extracting useful information from these reviews. Mining opinion information from reviews can not only help consumers to make a decision when buying products, but also help manufacturers to know the users' proposals in time. Sentiment label is a collection of a polarity word and its related target, contains detailed information on user reviews, can effectively reflect the core content of user reviews. In this paper, we deal with three key issues in sentiment labels: polarity word set construction, target extraction and sentiment label extraction.

For polarity word set construction, the goal of this paper is to construct an accurate and comprehensive polarity word set. We first integrate the semantic knowledge-base and large-scale corpus to get the candidate polarity word set, and then get the context of the candidate polarity word in the corpus, using context to set confidence to the candidate polarity word, which reflects the probability that polarity word is correct. Finally, we choose the polarity word having high confidence to form the polarity word set. We use this polarity word set to participate the task one of the first COAE and get a good result.

For target extraction, we first get the candidate target set using phrase structure, and then several target filtering algorithms are proposed. Firstly, the targets are domain-dependent, so we use the web-mining PMI to filter the candidate target. Secondly, the noun targets contained in phrase targets are usually redundant, so we use the noun pruning algorithm to filter them. After target set constructed, we classify the sentences in the reviews, and then find the appraised targets in the reviews based on target set. The system based on this method participates the task three of the first COAE and get a good result.

For sentiment label extraction, this paper proposes a novel method that uses syntactic paths to automatically recognize the sentiment labels. By using the syntactic path to extract the relationship between polarity word and target, we solved the too strong empirical problem of the nearest approach. By using an automatic method to construct the syntactic path set, we solved the problem that rules

constructed by human are always incomplete. Finally, we use the edit distance when match the syntactic path, which improves the system recall effectively.

Finally, to solve problem that traditional methods can't extract implicit sentiment labels, this paper tries to use topic model to annotate the sentiment labels for text, and proposes two methods based on PMI and probability distribution similarity. The results show that topic model can play a role in implicit sentiment tag extraction, we analyze the existed problems that topic model used in sentiment label extraction at last.

Keywords: Sentiment Analysis, Sentiment Labels, Target, Polarity Word, Syntactic Path

目 录

摘要	I
Abstract	II
第 1 章 绪论	1
1.1 课题背景与意义	1
1.2 情感标签抽取研究现状	2
1.2.1 评价词识别	3
1.2.2 评价对象识别	5
1.2.3 情感标签抽取	6
1.2.4 情感倾向性分析系统	7
1.3 相关评测会议	8
1.4 本文的主要研究内容	8
第 2 章 基本情感单元识别	10
2.1 评价词集构建	10
2.1.1 候选评价词集构建	11
2.1.2 评价词置信度设置	11
2.1.3 实验结果及分析	13
2.2 评价对象识别	14
2.2.1 候选评价对象获取	14
2.2.2 候选评价对象筛选	15
2.2.3 评价对象抽取	17
2.2.4 实验结果及分析	18
2.3 本章小结	20
第 3 章 基于句法路径的情感标签抽取	22
3.1 句法路径简介	22
3.2 标准句法路径库构建	23
3.2.1 句法路径生成	23
3.2.2 句法路径泛化	23
3.3 基于句法路径的情感标签抽取	25
3.3.1 基本情感标签抽取算法	26
3.3.2 编辑距离在句法路径匹配中的作用	27

3.4 实验结果及分析.....	29
3.4.1 实验数据及评价方法.....	29
3.4.2 对比实验设置.....	29
3.4.3 实验结果分析.....	30
3.5 本章小结.....	34
第 4 章 基于主题模型的情感标签标注.....	35
4.1 主题模型理论基础.....	35
4.1.1 LDA 数学描述.....	36
4.1.2 Gibbs Sampling 算法.....	38
4.2 基于主题模型的情感标签标注方法.....	39
4.2.1 语料及情感标签库获取.....	40
4.2.2 基于互信息的情感标签标注.....	41
4.2.3 基于概率分布相似度的情感标签标注.....	42
4.3 实验结果及分析.....	43
4.4 存在的问题及展望.....	44
4.5 本章小结.....	45
结论.....	46
参考文献.....	47
攻读学位期间发表的学术论文.....	51
致谢.....	53

第1章 绪论

1.1 课题背景与意义

Web2.0 时代的到来，为互联网提供了一个全新的发展模式，互联网逐渐倾向于“以用户为中心”的开放式架构理念。传统的互联网上用户只能被动的接受信息，网页由网站管理员、网站编辑等撰写好放到网上，用户只是单纯的去读网页。而 Web2.0 则为用户提供了一个主动创造互联网信息的途径，用户可以参与到建设互联网中来，可以自由的编辑、撰写网页，尤其是博客和论坛的产生，可以让用户自由、集中的发表自己的看法和意见，因此，互联网上产生了大量用户发表的评论信息，包括产品、时事等方面。这些评论信息表达了用户对于某一事物的评价，通常体现出一定的情感倾向性，如“批评”、“赞扬”、“喜”、“怒”、“哀”、“乐”等。这些评价信息具有十分重要的作用，对于产品方面，潜在的用户可以通过阅读产品的相关评价来决定自己是否购买某一产品；对于时事方面，可以通过分析这些评论信息进而获取大众的舆论导向。随着网民数量的急速增长，这类评价信息的膨胀速度十分惊人，仅通过人工的方法无法应对网上海量信息的收集和处理，使用计算机对评论信息进行快速的获取、整理和分析的需求十分迫切，为此，情感倾向性分析技术随之产生，并在近几年来成为广大学者们研究的热点。

情感倾向性分析，就是对带有情感色彩的文本进行分析、处理、归纳和推理的过程。根据处理文本粒度的不同，情感倾向性分析可分为词语级、短语级、句子级、篇章级以及多篇章级几个研究层次。情感分析技术产生初期，学者主要集中于研究词语级的情感色彩分析，如“漂亮”是含有褒义色彩的词语，而“丑陋”是含有贬义色彩的词语。随着互联网的发展，产生了越来越多的评论文本，仅对词语的情感倾向性分析不能满足互联网技术的需求，为此，学者们就逐渐从简单的词语情感分析逐渐过渡到短语级、句子级等更加复杂的文本单元进行研究。

目前情感倾向性分析领域通常将评论文本分为两类，即新闻评论和产品评论，对于新闻评论的情感倾向性分析，主要是为了了解大众对于某一事件的看法，如“我觉得奥运会的成功举办极大地提高了中国的国际声望。”这句话，表达了观点持有者“我”对于“奥运会成功举办”的立场。对于产品评论的倾

向性分析，主要是为了了解用户对于某一产品的看法，如“这款相机的外观很时尚。”这句话，表达了作者对评价对象“相机的外观”的评价“时尚”是褒义的。产品评论的倾向性分析具有重大的应用价值，一方面，商业网站需要了解用户对自己产品的看法，清楚的认识到自己产品的优缺点，从而更加迅速的响应消费者的需求，另一方面，消费者在购买某一产品前需要了解此产品的详细信息，尤其是消费者比较关心的重要属性，为此，产品评论的倾向性分析受到很多消费者和商业网站的青睐，目前学者的主要研究方向也集中在产品评论上。

情感信息抽取的目的在于抽取情感文本中有价值的信息，它是情感倾向性分析的基础任务。就目前学者们的的主要研究内容来看，这些有价值的信息主要包括评价对象（如“外型”、“镜头”）、评价词语（如“美丽”、“漂亮”）等。然而，随着研究工作的进行，研究者们逐渐发现组合搭配在情感分析中有更加直接的帮助，这种搭配主要体现为评价对象与评价词语之间的搭配，如“外型—时尚”、“价格—便宜”等，我们将这种搭配称为情感标签。情感标签包含了用户对于商品的详细评论信息，通过对情感标签进行汇总、整理，能够很方便的让用户以及商家了解产品的使用情况，因此，抽取评论文本中的情感标签在情感倾向性分析的任务中具有十分重要的意义。

1.2 情感标签抽取研究现状

情感标签由评价对象和评价词组成，依据评价对象和评价词是否均在评论文本中显式出现，可以将情感标签分为显式标签和隐式标签两类。例如“The camera has a good zoom.”，此句话包含一个情感标签“good—zoom”，此情感标签的评价词“good”以及评价对象“zoom”均在句子中出现，则“good—zoom”为一个显式标签。考虑“I never thought that battery can take so many pictures.”这句话，其隐式的表达了“long—battery life”这一含义，但此标签的评价词“long”并未在句子中显式出现，因此，“long—battery life”为一个隐式标签。目前已有的研究工作大多集中于显式标签的研究，隐式标签方面的研究较少。

对于情感标签抽取，目前的方法通常是基于已有的评价词集或评价对象集合，在文本中首先找出评价词或评价对象，进而以找出的评价词或评价对象为基础，使用规则、模板等信息挖掘情感标签。由此可见，评价对象和评价词的识别是情感标签抽取的基础任务，也是情感标签抽取中重要的研究方向，本文

将评价词与评价对象统称为基本情感单元。本文主要对评价词识别、评价对象识别以及情感标签抽取三个方面进行研究，对于评价词识别，已有的方法主要分为基于规则以及语料库两个方面，对于评价对象识别，已有的方法主要是基于规则以及指示词两个方面，下面，本文对评价词识别、评价对象识别以及情感标签抽取三个方面的研究现状进行详细介绍。

1.2.1 评价词识别

评价词语又可称为极性词、情感词，是指带有情感倾向性的词语，如“美丽”、“漂亮”等。评价词语在倾向性分析中有着十分重要的地位，任何情感评价单元（情感标签、句子、文本）的情感倾向性均由评价词语来决定。为此，评价词语的识别及其极性判断是情感倾向性分析的基础任务，也是被研究最多的一项任务。可以说，正是评价词语识别任务拉开了情感倾向性分析领域的序幕。评价词集构建通常由两部分组成，一是识别评价词，另外就是判断识别出来的评价词的极性。前人的研究工作往往同时对这两个任务同时进行解决，目前的评价词集构建主要包括基于语料库和基于词典这两种方法，下面对其进行详细介绍。

基于语料库的评价词语抽取和倾向性判别方法主要是利用大规模语料库的统计特性，根据统计特性制定相应规则来挖掘语料库中的评价词语并判断极性。早期的学者研究发现，由连词连接的形容词之间往往存在一定的关联性。如“and”连接的两个形容词之间往往具有相同的极性，而由“but”连接的形容词之间往往具有相反的极性。基于此规则，Hatzivassiloglou和McKeown^[1]从华尔街日报(Wall Street Journal)这个大规模语料库中挖掘出大量的形容词性的评价词语，实验表明此规则是比较有效的。Wiebe^[2]等人使用了一种基于相似度分布的词聚类方法来抽取评价词语，此种方法在大语料库上完成了形容词性的评价词语的获取。

以上两种方法均默认仅有形容词可以作为评价词，但在实际情况中，名词、动词等词性的词也可以作为评价词使用。为了解决这个问题，Riloff^[3]等人手工制定了一些模板，进而采用迭代的方式不断从语料库中抽取词性为名词的评价词语。之后，Turney和Littman^[5]提出了使用点互信息的方法来判别某个词语是否是评价词语，对于被抽取出的评价词，再进一步判定其极性。点互信息可以给出两个词语在大规模语料库上统计之后得到的含义相近的概率，其核心思想是利用共现信息，此方法首先需要定义两个种子词集合，分别为褒义和贬

义,进而计算词语与每一种子集合的点互信息,根据此值来判定词语是否为极性词,进而判定极性,此方法对种子集的依赖性极强,不同的种子集可能得到差异很大的结果。此外,Kaji^[6]等人利用网页文本HTML结构中的一些规律来构建日文的评价词词典,但此种方法对HTML结构的依赖性很强,且HTML结构的规律需要经过大量观察得到,不便于移植。基于语料库的评价词抽取方法简单易行,依据语料库的规律可以较准确的抽取极性词并进行极性判断,但它的缺点在于语料库的规律需要经过大量观察得到,评价词语在语料库中的分布现象较难归纳。

基于词典的评价词语抽取及判别方法主要是使用词典中的词语之间的词义关系。这里的词典是广义的词典,通常是指带有语义关系的语义知识库,如WordNet和HowNet。最初的学者人工采集一些种子评价词语^[7-9],在语义知识库中使用同义词以及词义进行扩展,这种方法实现简单,十分依赖于种子评价词语的质量,而且语义知识库中词语的多义性现象较为普遍,会给这种方法带来一定的噪声。有部分学者使用词典中词语的注释信息来识别评价词语及其倾向性^[10-13],这种方法能从一定程度上解决词语多义性带来的噪声。J. Kamps等人^[14]沿用了Turney等人的点互信息方法,将WordNet看做一个大语料库,选取种子词与形容词之间计算关联度,进而识别评价词语及其情感倾向性,这种方法的主要改进在于WordNet中词义之间的关系定义明确,相较于Turney使用的网络语料噪声更小,计算得到的结果也就更加准确。F. Su等人^[15]使用多个已有的语言知识库来对词语的极性进行判别,这种方法的缺点在于当某些语种的语言知识库不足时无法移植,为了解决这个问题,R. Mihalcea等人^[16]将词典资源丰富的语种的评价词典翻译到资源较少的语种中,但是实验结果表明,翻译之后的评价词语极性有时会发生改变,带来噪声。综上可知,基于词典的方法优点在于评价词语扩充方便,且极性的计算比较准确,难点在于大量的词含有多种语义,如何选取合适的语义来定义词与词之间的联系是基于词典的方法需要解决的问题。

无论是基于语料库的方法,还是基于词典的方法,其核心思想都是通过词与词之间的联系来判定评价词的情感倾向性。如果将评价词看做顶点,评价词之间的联系看为边,则可以将评价词之间的关系转化为图的形式。基于此,有部分学者采用基于图的方法来识别评价词的极性,H. Takamura等人^[17]利用两个词语之间的注释信息来表征词语之间的联系,继而使用Spin模型对图进行迭代概率计算,进而得到词语的极性。D. Rao等人^[18]尝试使用多种图模型,如:最小切分模型、随机最小切分模型、标签迭代模型等实现评价词语的极性判

定。相对于基于语料库和基于词典的方法，基于图的方法是一种比较新颖的方案，图通过给边赋权值的方式可以很方便的把词语之间的各种联系引入图中，进而进行计算，实验表明基于图的方式在评价词的倾向性判定上达到了比较好的效果。如何更加合理的利用边来表征词语间的联系以及选取更加合适的图算法都是此类方法主要研究的内容。

评价词仅考虑一个词语的倾向性，当评价词在具体句子中出现时，可能通过与其他词语组合共同表现出情感倾向性，我们将这种组合称为评价短语，如“not good”、“very bad”等。基于此，Moilanen等人^[19]和Choi等人^[20]提出了“复合语义单元”的概念。这里的“复合语义单元”就是指评价短语，表示的是一组有相互作用共同表达倾向性的词语，例如情感句“[I did [not] have any [doubt] about it.]⁺”中，“doubt”本身是贬义词，但经过“not”的修饰，使得此句的情感倾向性发生变化，最终表达了褒义的含义，在此句中，“not doubt”即可看作一个组合评价单元。目前的研究多使用人工总结或半自动生成的模板来识别评价短语。

1.2.2 评价对象识别

按照评论文本划分，评价对象也可以分为新闻领域的评价对象与产品领域的评价对象，新闻领域的评价对象通常是某个事件或某个人，而产品领域的评价对象可以被看成产品的某种属性（如“镜头”、“屏幕分辨率”），目前学者所作的研究工作大多集中在产品领域的评价对象抽取，下面对目前的工作进行详细介绍。

基于规则的方法是当前评价对象抽取中的主流方法，这类方法的优点在于领域可移植性强，适用于类别纷繁复杂的产品领域。为了增强规则的通用性，规则的制定通常需要经过一系列语言分析和处理的过程，包括分词、词性标注、句法分析以及命名实体识别等，从而在词性、短语结构等层级上制定相应的规则。J. Yi等人^[21]使用三条限制等级逐渐递进的词性规则从候选评价对象中抽取出真正的评价对象。B. Liu等人^[22]使用关联规则挖掘以及规则过滤的方法构建评价对象集，他们首先使用语料库构建共现项集，然后挖掘出共现项集中的频繁子项集作为候选评价对象，进而使用两条规则分别对候选评价对象中的短语评价对象和词评价对象进行过滤，得到频繁评价对象集合。为了挖掘非频繁的评价对象，他们以评价词为中心，对于不含有频繁评价对象的句子，在评价词附近查找评价词所修饰的评价对象，从而对频繁评价对象集合进行补充，

得到最终的评价对象集。这两种方法的优点在于领域可移植性强，存在的问题在于在构建评价对象集时没有充分考虑到各领域的特点，忽略了评价对象的领域相关性。

评价对象本身为产品属性，可以看作是产品的一个组成部分，如对数码相机领域而言，“相机滑盖”就是数码相机的一个组成部分。A.-M. Popescu等人^[23]从这个角度来实现评价对象的抽取技术，首先找出领域的指示词（如“整体-部分”关系的指示词为“camera has”或“of camera”），继而考察候选评价对象与领域指示词之间的关联度来获取真正的评价对象，这种方法取得了比较好的实验效果，指标上超过了基于规则的方法，但缺点在于领域可移植性差，领域的指示词只能通过人工构建，难度较大。

王波等人^[24]尝试使用半监督的机器学习的方法来解决评价对象抽取的问题，他们首先在语料中找出候选评价对象，然后为候选评价对象定义了 14 维特征，进而将评价对象识别的问题转化为二元分类问题，通过小规模标注语料以及大量的未标注语料，迭代的在语料中搜索评价对象，这种方法与基于规则/模板的方法有本质不同，它将评价对象抽取的研究重点转移到特征选择问题上。

近年来，随着主题模型^[25,26]的逐渐兴起，一些学者开始考虑将主题模型应用到情感分析领域。评价对象本身是对产品属性的一种描述，恰好符合主题模型中主题的概念，因此，主题模型理论上可以应用于评价对象的识别。I. Titov 等人^[27]采用多粒度的主题模型挖掘产品领域评论文本中的评价对象，并将含义相似的评价对象进行聚类，这种方法理论上能够提高评价对象抽取的召回率，但目前的实验并没有充分验证其有效性。

1.2.3 情感标签抽取

Hu等人^[28]首先使用评价对象抽取技术获取情感句中的评价对象，进而以评价对象为中心，选取距离评价对象最近的形容词作为修饰其的评价词语，Kim等人^[29]在Hu的方法上进行了一定的改进，他们仅在评价对象周围k个词语内选择修饰评价对象的评价词。这两种方法在某些条件下适用，但经验性太强，对于稍微复杂一些的句子就无能为力了，例如“the pictures taken by this amazing camera is very good.”，修饰评价对象“pictures”的评价词为“good”而使用就近的方式会使得其找出“amazing”作为其评价词。

评价词与评价对象之间具有一定的修饰关系，这种修饰关系通常可由一些

特殊的句型来体现。为此，Kobayashi等人^[30]采用模板的方式来表征这种修饰关系，他们为评价词语和评价对象之间的关系定义了 8 个模板，如“<Attribute> of <Subject> is <Value>”，具体的表现形式为一个评价单元三元组，<evaluated subject, focused attribute, value>，其中“focused attribute”表示评价对象，“value”表示修饰评价对象的评价词，“evaluated subject”表示他们之间的修饰关系。这种方法的缺点在于模板过于简单，而且模板均需手工构建，难以覆盖全面，模板匹配的过程中会产生大量的候选评价对象和评价词，需要人工筛选来完成情感标签的获取。

模板仅从句子表面给出了评价词语和评价对象之间的修饰关系，为了更加深入的挖掘评价对象和评价词之间的修饰关系，一部分学者尝试将句法分析技术应用于此项任务中。Bloom等人^[31]利用Stanford Parser手工构建了 31 条句法规则来描述评价词语与评价对象之间的关系，如 $target \xrightarrow{nsbj} x \xleftarrow{dobj} y \xleftarrow{amod} polarityword$ ，此外，Popescu等人^[32]利用MINIPAR Parser手工构建了 10 条依存句法抽取模板来获取情感标签。国内的姚天昉等人^[33]分析总结过“上行路径”和“下行路径”的匹配规则，进而利用“SBV”，“VOB”以及“ADV”三中结构总结过SBV极性传递的一些规则。通过句法分析技术可以更加深入的挖掘评价对象和评价词之间的修饰关系，避免了单纯停留在词表面上的规则，但由于匹配规则或模板的制定有过多的人工参与，覆盖率较低。

1.2.4 情感倾向性分析系统

目前，国内外有很多针对于各个领域开发的情感倾向性分析系统，用以帮助用户浏览网上的海量评论信息。例如，Liu等人^[34]研发的OpinionObserver系统可以处理网上在线顾客的产品评论，并给出评论的文摘表现形式，使得用户可以很方便的对产品各评价对象的质量进行比较。Wilson等人^[35]研发的OpinionFinder系统可以自动识别主观性句子以及抽取句子中的情感信息。国内上海交通大学^[33]开发了一个用户汉语汽车论坛的情感倾向性分析系统，挖掘并概括人们对各种汽车品牌的评论和意见。北京问天信息技术有限公司研发的“爱搜车”众评在线情感倾向性分析系统¹按照汽车的各种评价对象将评论信息进行呈现，方便用户的查询决策，以及多个车型之间的对比。

¹ <http://zp.isoche.com/>

1.3 相关评测会议

情感倾向性分析相关的评测主要有 TREC、NTCIR 以及中文倾向性分析评测 (COAE)，其中，TREC 主要针对英文文本中观点信息的检索，NTCIR 主要针对日、韩、英、中文文本的情感分类以及观点持有者的获取，包括句子主客观判别、句子倾向性判别、观点持有者识别、观点目标识别以及句子相关性判别 5 项任务。与情感标签抽取相关的评测会议目前仅有中文倾向性分析评测，第一届 COAE 主要包括以下六项任务：

- 1、评价词识别：要求自动识别出评测集中包含的评价词，输出结果要求按照置信度排序。
- 2、评价词极性判别：要求对任务 1 识别出来的评价词进行褒贬极性判别，即褒义或贬义。
- 3、评价对象抽取：针对给定的评论语料，找出其中作者评价的对象，同时对于作者对所评价对象的倾向性做出判别。
- 4、文本主客观判别：对评测集中的文本是否包含倾向性观点进行判别，包含倾向性观点的为倾向性文本，不包含倾向性观点的为客观性文本。
- 5、文本褒贬极性判别：对任务 4 判别出的倾向性文本的观点倾向性进行褒贬极性判断。
- 6、观点检索：针对给定主题，要求找出包含关于该主题的倾向性观点的文章。

第一届 COAE 最早提出了针对评价词与评价对象的评测方案，为情感标签抽取的相关技术提供了一个统一的评测平台，为情感标签抽取技术的发展做出了巨大的贡献。

1.4 本文的主要研究内容

本文首先对情感标签抽取的基础任务：评价词识别和评价对象识别进行了研究，使用基于这两项技术开发的系统参加了第一届 COAE 评测，并对评测结果进行了详细的分析。在基本情感单元识别技术的基础上，本文进一步提出了一种以评价词为中心、基于句法路径的情感标签抽取方法，实验结果表明此方法在情感标签抽取问题上达到了比较好的效果。最后，为了解决隐式情感标签抽取的问题，本文尝试了两种基于主题模型的情感标签标注方法，并对存在的问题进行了分析。

本论文的内容安排如下：

第一部分为绪论，首先介绍了课题的背景与研究意义，介绍了情感标签抽取相关技术当前主要的技术路线以及优缺点，并介绍了相关评测会议情况。

第二部分为情感标签抽取的基础任务研究，包括评价词识别与评价对象抽取两部分，使用基于此部分技术的系统参加了第一届 COAE 评测，并对评测结果进行了详细的分析。

第三部分提出了一种以评价词为中心、基于句法路径的情感标签抽取方法，在评价词集已知的条件下识别句子中的情感标签，通过与之前学者的方法进行比较，验证了句法路径在情感标签抽取研究中的作用。

第四部分为基于主题模型的标签抽取方法研究，为了解决隐式标签抽取的问题，提出了两种基于主题模型的情感标签标注方法，对实验结果进行了分析，并对存在的问题进行了总结。

最后一部分是结论。

第2章 基本情感单元识别

情感标签包括评价词和评价对象这两个基本情感单元，识别这两个基本情感单元是情感标签抽取的基础任务。本章针对这两项任务，首先提出了融合语义知识库与大规模语料库的评价词集构建方法，然后提出了基于评价对象库的评价对象识别方法，下面对这两部分内容进行详细介绍。

2.1 评价词集构建

评价词相较于评价对象来说结构相对简单，通常为一个单一的词，如何构建准确并且全面的评价词集合是情感分析相关研究中较为重要的一项任务。为了解决这一问题，本文提出了一种结合语义知识库与大规模语料库的评价词集构建方法。首先，本文构建一个种子评价词集合，进而使用 HowNet 对种子评价词集进行扩充，同时，在大规模语料库中进行评价词挖掘，融合两部分结果从而获取一个较大规模的候选评价词词典，此时评价词词典的噪声较大，为此，本文引入置信度来对候选评价词体现情感倾向性的可能性进行度量。获取候选评价词词典后，本文在大规模语料库中搜索评价词，基于句法关系信息获取其上下文，进而通过上下文信息为评价词设置置信度，选取置信度较高的评价词作为最终结果，系统框架如图 2-1 所示。

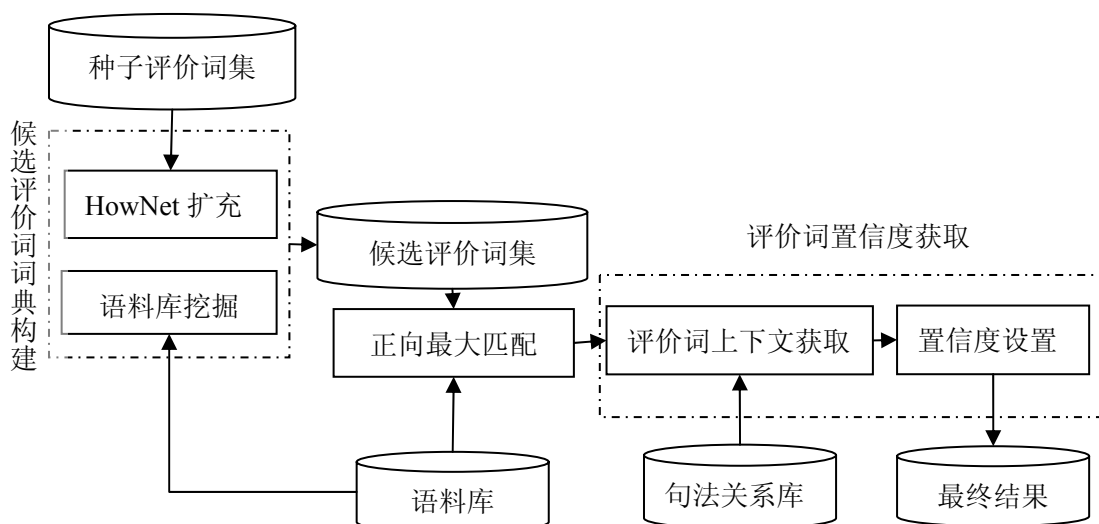


图 2-1 评价词识别系统框架

2.1.1 候选评价词集构建

本文方法首先需构建一个较为全面的候选评价词词典，本节首先基于语义知识库扩展的方式来构建这个候选评价词词典，具体步骤如下：

step1: 手工构建一个种子评价词词典；

step2: 以 HowNet 为基础，获取此种子评价词词典在 HowNet 中的义原集 M ；

step3: 用 HowNet 对种子评价词的义原集 M 进行同义词扩充，形成评价词词典 D_1 ；

step4: 将 D_1 与 HowNet 中原有的评价词词典 D_2 合并形成最终的评价词词典 D 。

为了获取覆盖面更广的评价词，本文对测试集中的词语进行了挖掘，定义了某个词语 w 成为评价词的可能性如下：

$$PRO = \frac{w \text{ 扩展的词语个数}}{w \text{ 的义原个数}} \quad (2-1)$$

1)

其中，公式的第一项“ w 扩展的词语个数”指的是词语 w 通过 HowNet 进行同义词扩展后，扩展出来的同义词在评价词词典中出现的个数；第二项“ w 的义原个数”指的是词语 w 在 HowNet 中的词义的个数。通过设定阈值，我们获取了这部分评价词，将这两部分评价词集合融合，即可得到候选评价词集合。

2.1.2 评价词置信度设置

候选评价词集构建完毕后，首先使用正向最大匹配算法在测试集中搜索评价词，进而利用句法关系为评价词找出最为准确的上下文，最终根据上下文为评价词设置置信度，下面对具体方法进行详细介绍。

2.1.2.1 评价词上下文获取

本文使用正向最大匹配 (Forward Maximum Matching method, FMM) 算法在语料库中搜索评价词。正向最大匹配实际上是一种简单的中文分词方法，其基本思想是：假设词典中最长的关键字的长度为 i ，取当前待处理文本的前 i 个字作为匹配字段 w ，在词典中查找，若词典中有 w ，则匹配成功， w 即做为一个词被切分出来；如果匹配失败，则去掉 w 的最后一个字或字符，继续去字典中查找；如果 w 的长度为 1，仍然没有匹配到，则也将其切分。切分出 w 后，继续对 w 之后的字词进行上面步骤的切分，直到切分出所有的词为止。

若评价词本身具有一定的情感倾向性，则其必然在语料库中的某个句子中体现出情感倾向性，本文认为这个句子即为评价词在文本集合中的上下文。由于评价词可能在文本中的多个句子中出现，我们要找出其最可能体现情感倾向性的那个句子作为上下文，进而对置信度进行设置。

本文使用句法路径来表征一个评价词在某个情感句中体现情感倾向性的可能性。句法路径是获取评价词是否修饰句子中其他成分的有效方式，例如“目前 学校 体育 设施 不足 较为 普遍。”这句话，此句话的评价词为“不足”，其评价的对象为“设施”，通过短语句法分析，我们可以获取它们之间的句法路径“NN↑NP↑IP↓VP↓VA”。本文的方法基于这样一个假设：若评价词在句子中表现出一定的情感倾向性，则其必然与句子中的某个目标词产生修饰关系，这种修饰关系越强，则评价词在此句中体现出情感倾向性的可能性也就越大。例如“他想总冠军戒指已经想到快疯了。”、“这款相机的外观时尚。”这两句话，第一句话中的“快”并不直接修饰某个目标词，因此其不具有情感倾向性，而第二句话中的“时尚”是在修饰目标词“外观”，从而具有一定的情感倾向性。本文使用句法路径的优先级来表征评价词与目标词之间的修饰关系强度，通过获取评价词在所有句子中与目标词的句法路径，即可找出包含评价词且评价词与目标词之间句法路径优先级最高的句子，此句即为评价词在语料库中的上下文。

为了构建标准句法路径库并给出优先级，本文从 sina 网上下载了主题为“体育”、“财经”、“娱乐”等领域的多篇文章，对其进行分句、分词、词性标注等预处理，并找出含有评价词的句子。对含有评价词的句子进行人工标注，给出评价词在这些句子中所修饰的目标词，经过句法分析后即可获取评价词与目标词之间的句法关系，对句法路径进行频率统计，并以此为优先级将句法路径排序，优先级越高，说明评价词修饰目标词的可能性越大，评价词体现情感倾向性的可能性也就越大。

标准句法路径获取后，即可获取评价词在文本集合中的上下文，具体步骤如下：

step1：预处理，对语料库进行分句（短句，以“，”、“。”、“？”、“！”等分隔），分词，词性标注以及短语句法分析。

step2：遍历评价词词典，针对每一个评价词，获取含有该评价词的所有句子。

step3：针对每一个含有评价词的句子，获取该评价词和句中所有的候选目标词（名词，动词）之间的句法关系，选取优先级最高的句法路径所在的句子作为该评价词的上下文。

2.1.2.2 评价词置信度设置

通过以上步骤，可以获取大部分评价词在文本集合中的上下文，但是我们发现，有部分评价词在分词时会出现错误，导致无法根据句法路径获取其上下文，如“死心塌地”分词为“死心 塌 地”，“寡淡”分词为“寡 淡”，为此，我们分别为含有上下文以及分词错误的评价词制定相应的置信度设置规则。

评价词可以分为上下文无关与上下文相关两种，上下文无关是指在修饰任何目标词时，其始终表达相同的情感倾向性，上下文相关是指在修饰不同目标词时，其表达的情感倾向性可能不同。本文的置信度设置方式基于如下原则：形容词更可能在文本集中体现情感倾向性；评价词的结构越复杂，其越可能体现情感倾向性；上下文无关的评价词更可能在文本集中体现情感倾向性。根据此三条原则，本文的评价词置信度设置规则定义如下：

- 置信度为 1000：上下文无关的评价词，且评价词在上下文中是形容词
- 置信度为 1000：分词错误的上下文无关的评价词，且词语的字个数大于等于 3
- 置信度为 900：上下文相关的评价词，且评价词在上下文中是形容词
- 置信度为 800：上下文无关的评价词，且评价词在上下文中是除形容词外的其他词性
- 置信度为 500：上下文相关的评价词，且评价词在上下文中是除形容词外的其他词性
- 置信度为 50：分词错误的其他评价词

2.1.3 实验结果及分析

依照前面介绍的方法，本节实现了评价词识别系统，并参加了第一届 COAE 评测的任务一。任务一提供了一个近 4 万篇文本构成的语料库，包括真实用户评论和新闻报道评论等，涉及财经、娱乐、影视、教育等多个领域，任务要求以此语料库为基础，抽取出 6000 个评价词，给出置信度并将评价词按照置信度从高到底的顺序排序，此次评测包括以下两种评测方式：

(1) pooling 方式：每个提交结果取前 100 和 1000 条记录组成评测池，经人工评判后作为答案，对各自提交的前 100 和 1000 条记录进行评判打分，评测指标为 $P@100$ ， $P@1000$ 。

(2) 自动评判方式：采用事先人工标注的词典对各个提交结果的所有记录进行评测打分，评测指标为提取正确的评价词总数。

系统评测结果如表 2-1 所示，其中 HIT-IR 表示的是本文系统结果，Best 表示的是所有评测系统中的最好结果，Median 表示的是所有评测系统的平均结果。

表 2-1 评价词识别的评测结果对比

系统名称	P@100	P@1000	Right_by_Lexicon
HIT-IR	0.97	0.974	3038
Best	1	0.984	3097
Median	0.925	0.9335	2602

由结果可以看出，无论从哪个指标上来看，本文的方法均很接近评测最好值，并远远高于各评测系统的平均结果，这充分说明本文使用方法的有效性，另外，我们在 P@100 和 P@1000 上都取得了很好的成绩，这说明本文的置信度设置方式是比较合理的，即越靠前的评价词识别的越准确。但同时我们也可看出，系统最终仅有 3000 个左右的评价词被正确找出，我们认为主要有以下两点：一是我们采用的方法基于规则，经验性较强，具有一定的缺陷；二是本文的候选评价词词表很大，由于评价词的识别具有一定主观性，因此，其中的某些评价词可能并不被官方认可。总体来说，本文的方法还有很多需要改进的地方，这也表明了评价词识别未来还有很广阔的研究空间。

2.2 评价对象识别

本文使用基于评价对象库的方式对评价对象进行识别，对于给定输入，首先对其进行分词、词性标注以及句法分析等处理，然后提取语料中特定句法结构的短语作为候选评价对象；继而对初步获取的候选评价对象使用频率过滤、PMI(Pointwise mutual information)值过滤和名词剪枝等算法进行筛选得到最终的评价对象库，评价对象库获取后，在测试集中基于正向最大匹配算法抽取有意义的评价对象，系统框架如图 2-2 所示。

2.2.1 候选评价对象获取

考察情感句“这款电脑的续航能力很强，但价格太高。”，其句法结构图 2-3 所示。我们发现“续航能力”和“价格”是该情感句的评价对象且其句法成分分别为 NP（名词短语）和 NN（名词）。此外，通过对大量的情感句中的评价对象进行观察，我们发现其句法成分基本均为 NP 或 NN。基于此，本文将语料中的所有 NN 和 NP 作为候选评价对象，得到初始评价对象集。对于图 2-3 中的样例，按照我们的句法结构限制，此句的评价对象为“续航能力”、“续

航”、“能力”、“价格”，可以看出，通过使用句法信息，我们能够避免抽取出“能力价格”这种错误评价对象，但同时也会产生一定的噪声（续航、能力），为此，针对不同的噪声产生原因，我们引入了三种过滤技术，并对三种过滤技术的作用进行了实验对比，下面对其进行详细介绍。

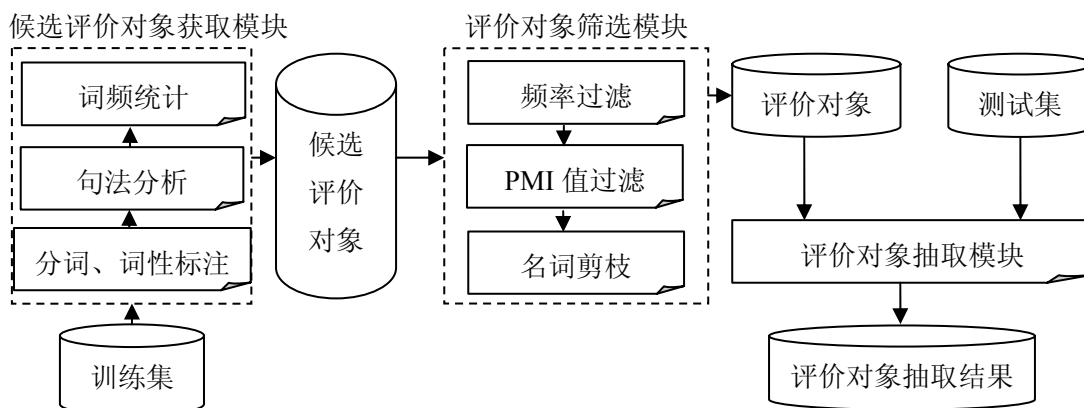


图 2-2 评价对象识别系统框架图

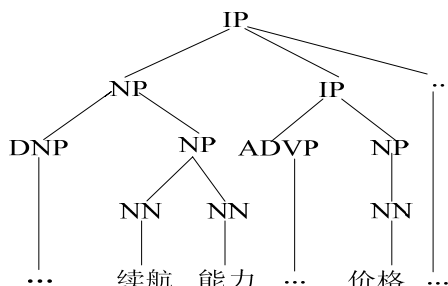


图 2-3 句法分析示例

2.2.2 候选评价对象筛选

由句法分析结果得到的候选评价对象集存在一定的噪声，为此，我们加入了相应的过滤技术。评价对象筛选流程如下：首先，部分候选评价对象在训练集中出现的次数很少，这类评价对象很有可能是分词、句法分析等预处理带来的噪声，为此，首先使用词频信息进行过滤；其次，评价对象具有很强的领域相关性，候选评价对象可能在训练集中出现的次数很多，但并不应该作为此领域的评价对象，针对这种现象，使用基于网络挖掘的 PMI 算法进行二次过滤；最后，候选评价对象之间存在一定的包含关系，为此，使用名词剪枝技术解决单个词的冗余现象。下面分别

对这三项技术进行介绍。

2.2.2.1 词频过滤

词频过滤即只选取语料中出现次数比较多的 NN 或 NP 作为候选评价对象。词频信息过滤的加入主要基于以下两点考虑：1、评价对象更倾向于在评论中多次出现，一些不相关的 NN 或 NP 应该在商品中很少出现，如“有限公司”、“多图”等。2、词频信息过滤可能会过滤掉一些评价对象，但这对系统的结果影响不会很大，因为那些出现次数较少的评价对象并不被大多数人所关心，属于次要属性。

2.2.2.2 PMI 值过滤

PMI 值提供了一种量化词与词之间关系的方法，在一定的文本集合中，词 a 和 b 的 PMI 值定义如下：

$$PMI_{a-b} = \frac{N_{ab}}{N_a * N_b} \quad (2-2)$$

其中， N_{ab} 表示既包含 a 又包含 b 的文本数， N_a 表示含有 a 的文本数， N_b 表示含有 b 的文本数。从公式中可以看出，PMI值的计算使用了统计的思想，同时基于这样一个假设：两个单词共现的次数越多，则它们之间的联系也就越大。PMI值计算的难点在于大规模文本集合的获取，理论上讲，文本数越多，则统计效果越明显，PMI值的计算也应该越准确。

本文使用PMI值来进一步挖掘评价对象的领域相关性。针对每一领域，选取其最具代表性的词语 w_p （也就是在此领域语料中出现次数最多的词语）作为领域代表词，继而对于每一候选评价对象 w ，计算其与相关领域代表词 w_p 的PMI值，PMI值越大，则说明候选评价对象 w 与此领域更相关，更可能成为此领域的一个评价对象。为了得到足够大的语料，本文充分利用了网络资源，选取百度的搜索结果作为大语料库。具体方法如下：为了计算候选评价对象 w 是否是某一领域的相关词（假设该领域的代表词语是 w_p ），本文在www.baidu.com中分别搜索 w 、 w_p 、和“ $w w_p$ ”，返回的网页数量为 N_a 、 N_b 、 N_{ab} ，则 $PMI_{a-b} = N_{ab} / (N_a * N_b)$ 。如 w 为“电池寿命”，其所在领域为手机领域，手机领域的代表词 w_p 为“手机”，则通过计算“电池寿命”和“手机”的PMI值来确定“电池寿命”是否应该是手机领域的评价对象。计算PMI值后，通过设定阈值对候选评价对象进行过滤。本文最后取定阈值为 5.0，实验表明这种方法取得了比较好的效果。

2.2.2.3 名词剪枝

此技术主要应用于冗余名词的过滤。为了说明什么是冗余，本文首先定义

s-support: 对于名词 t , 设包含 t 的句子数为 s , 在 s 个句子中, t 单独作为评价对象出现的句子数为 k (这 k 个句子中不含有包含 t 的候选评价对象), 则 $s\text{-support} = k/s$ 。对于 $s\text{-support}$ 值小于 0.5 的名词评价对象, 本文认为它是冗余的, 过滤掉。

考虑这样一个例子: 系统识别出两个评价对象“(NN 能力)”和“(NP (NN 续航) NN (能力))”, 由于“(NN 能力)”包含在“(NP (NN 续航) (NN 能力))”中, 其出现频率不会低于“(NP (NN 续航) (NN 能力))”的频率, 因此, 使用词频和 PMI 值是不能将其过滤的。考虑“(NN 能力)”的 $s\text{-support}$ 值, 包含“(NN 能力)”的句子数为 50, 其中“(NN 能力)”单独出现的次数仅为 10 次, $s\text{-support} = 10/50 = 0.2$, 通过设定 $s\text{-support}$ 阈值, 我们可以将这类评价对象过滤掉 (本系统设定 $s\text{-support}$ 值为 0.5)。

2.2.3 评价对象抽取

评价对象在文本中出现时存在两种情况, 一是对评价对象进行客观描述, 如“这款相机的镜头是佳能的。”, 此句的“镜头”为评价对象, 但实际上作者并未对其作任何评价; 二是对评价对象进行带有情感倾向性的评论, 如“这款相机的镜头很不错。”, 作者在此句中就表达了对评价对象“镜头”的情感倾向性。可以看出, 仅有第二种情况下评价对象的出现才有意义, 为此, 本文的评价对象抽取是针对第二种情况下的评价对象抽取。此时, 抽取过程包括两个步骤, 首先使用正向最大匹配算法识别出评论文本中的所有评价对象, 包括上面描述的两种情况, 然后对第一种情况的评价对象进行过滤, 从而得到最终评价对象抽取结果, 下面对方法进行详细介绍。

2.2.3.1 评价对象识别

本文使用正向最大匹配算法识别文本中的评价对象。评价对象抽取的目标是尽可能找出信息完全的短语作为评价对象, 如“这款相机的屏幕分辨率很高。”这句话, 我们的目标是找出“屏幕分辨率”这个评价对象, 而不是“屏幕”或者“分辨率”, 正向最大匹配算法恰好能解决这个问题, 即总能找出评价对象表中最长的短语作为抽取结果。

2.2.3.2 无意义评价对象过滤

本文使用的评价对象抽取方法基于这样一个假设: 若评价对象所在的句子具有情感倾向性, 则评价对象本身也具有一定的情感倾向性。因此, 评价对象是否具有情感倾向性的问题就转化到评价对象所在的句子是否具有情感倾向

性。通过观察，本文将评论文本中的句子分为以下四类：

类别一：句子带有明显的倾向性，即句子中带有一种倾向性（褒义或贬义）的上下文无关评价词明显多于另一种。例如：“这款数码相机的质量不错外形也很漂亮”，这句话的情感词为“不错”和“漂亮”，均为褒义，则此句明显含有褒义的倾向性。

类别二：句子不带有明显的倾向性，但句子中含有评价词，且褒义和贬义评价词的个数相同。例如：“这款相机的质量不好但外形很漂亮”，这句话中含有“不好”和“漂亮”两个情感词，极性分别为贬义和褒义，但不能说此句的倾向性是褒义还是贬义。

类别三：句子不带有明显的倾向性，且句子中没有评价词，但其上下文的句子带有明显倾向性。例如：“这款相机的质量非常好，照出的照片也一样”，“照出的照片也一样”这句话本身不含有评价词，但此句的前一个句子“这款相机的质量非常好”带有明显的倾向性。

类别四：句子不带有明显的倾向性，句子中没有评价词，且其上下文的句子也不带有明显倾向性。例如：“我购买相机主要看中质量。”、“锐志和皇冠有很多类似的地方，如外型、内饰等”。

作者撰写评论文本的主要目的在于对评价对象进行评价，因此，只要句子体现了一定的情感倾向性，且句子中包含评价对象，则此时的评价对象就是具有意义的评价对象。对于类别一中的句子，其句子本身具有情感倾向性，则其含有的评价对象也应具有情感倾向性，应被抽取出来，对于类别二来说，句子本身的情感倾向性不能判定，但是句子中出现了多个评价词，同样，作者使用评价词的目的在于对评价对象进行评价，因此，类别二句子中的评价对象也应抽取出来。对于类别三中的句子，虽然不能通过是否含有评价词判断其倾向性，但其上下文句子均含有明显的情感倾向性，而通常情感倾向性具有一定的传递性，为此，本文认为类别三中的句子也带有一定的情感倾向性，应当抽取出其含有的评价对象。相应的，类别四的句子通常为客观描述，过滤掉其所含有的评价对象。

2.2.4 实验结果及分析

本系统参加了第一届中文倾向性分析（COAE）评测中的任务三中评价对象抽取这一评测任务，此任务要求从评论文本中抽取出用户表达情感倾向性的评价对象。

此评测任务的语料主要涉及汽车、手机、数码相机以及笔记本四个领域，共有 473 篇文档（3000 个句子，10000 个具有倾向性评价的评价对象）。评测方法有精确评价（Strict）和覆盖评价（Lenient）两种，具体如下：

- 精确评价：抽取的评价对象与答案完全匹配才算正确，如答案为“屏幕分辨率”，如果提交结果为“屏幕”或者“分辨率”都不算正确；
- 覆盖评价：抽取的评价对象与答案有重叠就算匹配的正确，其中重叠部分大小可以调整，用多个参数进行评价。例如上面的例子，提交结果为“屏幕”或者“分辨率”都算正确。

表 2-2 给出了系统在这两种情况下评价对象抽取的性能指标。由结果可以看出，本文使用的评价对象抽取方法在评测中取得了较好的成绩，无论在 Strict 还是在 Lenient 评测方式下，本文的方法均远远高于平均指标，同时，在 Lenient 的评测方式下取得了最好成绩。

表 2-2 评价对象提取的评测结果对比

RunID	DataSet	属性抽取的结果(Strict)			属性抽取的结果(Lenient)		
		Precision	Recall	F-measure	Precision	Recall	F-measure
HIT_IR	Car	0.3126	0.4127	0.3557	0.4513	0.5958	0.5136
	Camera	0.3219	0.4006	0.3569	0.4556	0.567	0.5052
	Phone	0.3472	0.381	0.3633	0.4976	0.546	0.5206
	NoteBook	0.3379	0.4566	0.3883	0.4693	0.6342	0.5394
	All	0.3275	0.4058	0.3625	0.467	0.5788	0.5169
所有评测结果平值:		0.287749	0.22697	0.233079	0.442581	0.347527	0.357563
所有评测结果最值:		0.5641	0.4172	0.3976	0.7206	0.5788	0.5169

为了验证本文提出三种过滤方法在过滤候选评价对象噪声中的作用，本文依次在笔记本领域的语料上依次加入这三种过滤方式进行实验，评测方式为严格匹配。实验结果如表 2-3 所示。

表 2-3 三种过滤技术对系统指标的影响

	属性抽取结果(Strict)		
	Precision	Recall	F-measure
未对过滤候选评价对象集	0.0853	0.2780	0.1305
仅使用频率过滤方法	0.0904	0.3581	0.1444
频率过滤+名词剪枝	0.1170	0.4044	0.1815
使用三种过滤方法	0.3379	0.4566	0.3883

由表 2-3 可以看出，三种方法都有一定的过滤效果，尤其是 PMI 值过滤，使系统的 F 值提升了近 20%，这也与本文开始的设想一致，这也充分说明了评价对象具有一定的领域相关性。可以看到，名词剪枝方法使得系统 F 值提升了 4%，这说明候选评价对象确实存在着一定的包含信息，去除这种信息对构建

准确的评价对象库很有帮助，这种思想在评价对象抽取时使用正向最大匹配算法也有一定体现。

本文的评价对象抽取依赖于评论句的类型识别，因此，评论句类型识别的性能指标很大程度上影响了评价对象抽取的性能，为了验证本文对于评论句类别识别的精度，本文分别从四个领域各随机抽取 50 个句子，人工对这 200 个句子进行标注，表 2-4 给出了这四部分语料的情感句分类精度。

表 2-4 情感句分类精度

	数码相机	手机	笔记本	汽车
精确率	0.78	0.76	0.84	0.84

可以看出，情感句类别精度在各个类别大体相当，均为 80%左右。情感句型识别是以评价词表为基础，而评价词表本身会存在噪声和不完全两类问题，即评价词表中的词可能是错误的，且部分评价词可能并未包含评价词表中，这样就会造成情感句分类不准的情况，如“按键设计凹凸有质。”这个句子，本应属于第一类情感句，但由于评价词“凹凸有质”并不在评价词表中，导致此句子会被错误分到第四类句子中。

由结果可以看出，本文的评价对象提取技术在 Strict 评价和 Lenient 评价中均远远高于平均水平，且在 Lenient 评价中取得了最好的成绩，这也进一步肯定了本文方法的有效性。但尽管如此，我们发现该系统还是存在的问题。通过观察可以看出，系统的准确率偏低，尤其是 strict 评价方法下。由于本文的评价对象提取使用的是无指导方法，为了得到尽可能多的评价对象，我们放宽了过滤条件（如频度、PMI 值），使得获取评价对象中存在较大噪声，因而造成了系统召回率较高而准确率较低的局面。

中文的评价对象表示非常丰富，存在大量的长度较长的评价对象，这种评价对象的抽取十分困难，如“空调系统面板的布局”、“P. A. T. S. 发动机电子控制防盗模块”等。本次的评测结果普遍偏低，这也充分说明了中文评价对象抽取具有相当大的难度，中文领域的评价对象抽取还有相当大的研究空间。

2.3 本章小结

本章对情感标签抽取任务中的基本情感单元识别工作进行了研究，包括评价词集构建以及评价对象识别两部分。评价词集构建部分，本章主要研究如何构建准确并且全面的评价词集合，为此，本章首先基于 HowNet 以及大规模语料库构建了一个候选评价词词典，进而通过评价词在语料库中的上下文以及置

信度设置规则识别文集中的评价词。对于评价对象识别部分，本章主要研究在特定文本集合中识别有意义的评价对象，即用户进行评论的评价对象，为此，本章首先基于短语结构以及三种过滤方式获取了评价对象集，进而通过分析评价对象所在句子得到有意义的评价对象。基于本章技术的评价词集构建系统以及评价对象识别系统参加了第一届中文倾向性分析评测，取得了较好的成绩，说明了本章方法的有效性，最后，本章对两种方法存在的问题进行了分析和总结。

第3章 基于句法路径的情感标签抽取

第 2 章介绍了评价词与评价对象识别的相关技术，在此基础上，本章着重研究如何从情感句中抽取正确的情感标签，提出了一种以评价词为中心、基于句法路径的情感标签抽取方法，与之前学者的两种抽取方法进行了对比实验，并对三种方法的实验结果进行了详细的分析，下面对其进行详细介绍。

3.1 句法路径简介

本章的句法路径是针对短语句法结构而言的，对于一棵短语结构句法树，任意两个节点之间存在一条有向路径，本文将这条有向路径称为句法路径。本章通过挖掘评价词语和评价对象之间的句法结构关系从而识别情感标签。首先看下面两个句子，图 3-1 分别显示了它们的短语句法路径。

Sen1: The camera's image is perfect.

Sen2: The camera has very perfect image.

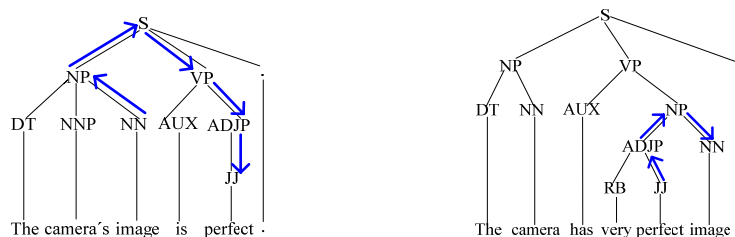


图 3-1 句法路径示例

观察两个句法树可以发现，Sen1 中评价对象“image”和评价词语“perfect”之间的句法路径描述了一种特殊的“主系表关系”，此句法路径可以记为“NN↑NP↑S↓VP↓ADJP↓JJ”。Sen2 中评价对象“image”和评价词语“perfect”之间的句法路径则描述了一种“修饰语-中心语”关系，此句法路径可表示为“JJ↑ADJP↑NP↓NN”。可以看出，句法路径能够从一定程度上表示出评价对象及其对应评价词语之间的修饰关系。为了得到准确且覆盖较全的句法路径规则，本文采用自动构建句法路径库的方式。由于句法路径精确匹配会导致系统召回率较低，本文在句法路径匹配阶段引入编辑距离进行松弛，以达到更好的匹配效果。下面对本章使用的技术进行详细介绍。

3.2 标准句法路径库构建

本节首先基于评价词语在语料库中自动发现句法路径，构建初步的句法路径库，此时获取的句法路径较多而且复杂，为此，本节引入了句法路径泛化这一步骤，从而得到最终的标准句法路径库。

3.2.1 句法路径生成

生成句法路径的前提是要确定情感句中的评价词和评价对象。本文选取了 Hownet 中的英文评价词典作为评价词集合，上一章我们提到过，评价对象可以是名词或名词短语，本章主要研究句法路径在情感标签抽取中的作用，因此，仅选取名词作为候选评价对象，对于短语评价对象，只要构建好评价对象集，本方法可以很方便的扩展到上面。除去名词外，英文中的指代现象较多，如“it”、“they”等，为此，本文设定情感句中词性为名词(NOUN)或代词(PRON)的词语作为候选评价对象。英文中的名词有多种表现形式，选取的具体词性描述如下：

- NOUN: NN, NNS, NNP, NNPS
- PRON: PRP

基于此，对于一个情感句来说，本文首先基于评价词集识别出其中的评价词，再找出其包含的所有候选评价对象，然后两两组合得到所有的句法路径。本文使用出现频率来衡量句法路径正确度，为了体现统计规律，本文在数码相机领域构造了一个含有 20000 个情感句的无人工标注的大规模语料库。对于一个情感句来说，如果其含有 n 个评价词语， m 个名词或者代词，则会产生 $n*m$ 条句法路径，基于我们的语料规模，此时得到的句法路径库是十分庞大的。

3.2.2 句法路径泛化

经过第一步获取的句法路径纷繁复杂，且数量庞大。观察后发现，其中大部分句法路径之间仅由于一些细小的差别而成为两种句法路径，而这些细小的差别对评价词与评价对象之间的修饰关系不产生任何影响，如“JJ↑ADJP↑NP↓NN”与路径“JJ↑ADJP↑NP↓NNS”，无论最右端的元素是“NN”还是“NNS”，均与本句话中的“JJ”具有修饰关系。为了解决这类问题，本文对初始的句法路径做了如下两步泛化：

- 泛化 1：若某些句法成分标签在句法路径中的作用相同，则用一个规范化

的标签对它们进行统一表示,如图 3-2 所示,两个情感句的句法结构完全相同,但由于第 2 个句子中比较级“better”和名词复数“images”的使用,导致两个句子中评价词与评价对象之间的句法路径产生差别。不难发现,对于描述评价词和评价对象之间的关系来说,句法成分标签“JJ”和“JJR”的作用完全相同,“NN”和“NNS”也不存在任何差别,因此,可以将“JJR”用统一的标签“JJ”来代替,将“NNS”用统一的标签“NN”代替。经过句法成分标签统一后,右边句子评价词与评价对象之间的句法路径“JJR \uparrow ADJP \uparrow NP \downarrow NNS”即可被泛化为“JJ \uparrow ADJP \uparrow NP \downarrow NN”,本文定义了 5 类统一标签,具体如表 3-1 所示。

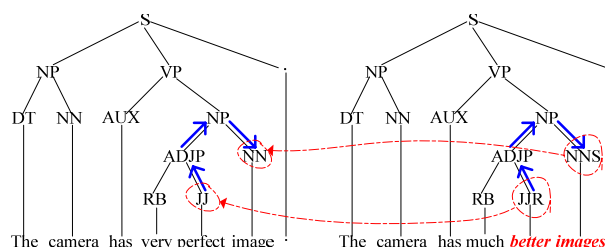


图 3-2 句法路径的第一种泛化示例

- **泛化 2:** 在描述评价对象与评价词的关系时,句法路径中一连串相同的句法成分标签通常与一个单一的句法成分标签具有相同的作用。如图 3-3 所示,右侧句子中评价词与评价对象之间的句法路径“NN \uparrow NP \uparrow S \downarrow VP \downarrow VP \downarrow ADJP \downarrow JJ”中的成分“VP \downarrow VP \downarrow ”与左侧的句法路径“NN \uparrow NP \uparrow S \downarrow VP \downarrow ADJP \downarrow JJ”中的“VP \downarrow ”作用完全相同,这种情况下,完全可以用“VP \downarrow ”替换“VP \downarrow VP \downarrow ”,对描述评价词与评价对象的修饰关系不产生任何影响。

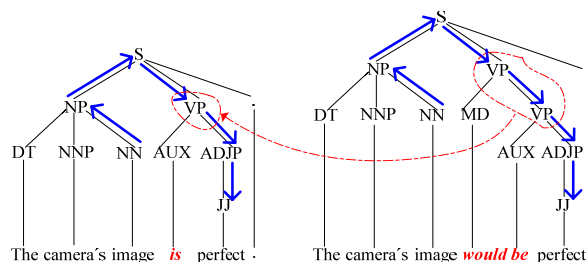


图 3-3 句法路径的第二种泛化示例

本文使用频率作为选取最终句法路径的标准,对于经过泛化后的句法路径,首先统计出各句法路径出现的总次数,然后设定阈值来选取出现次数较多

的句法路径构成最终的句法路径库。通过频率选取句法路径的方法基于这样一个假设：评价词与评价对象之间的修饰关系是具有一定规律性的，这种正确的修饰关系应该是有限的，并且在语料库中出现较频繁，对于出现不频繁的句法路径，通常表达的是错误的修饰关系，应该过滤掉。

表 3-1 相似句法成分标签的泛化

泛化标签	相似句法成分标签
JJ	JJR, JJS
NN	NNS, NNP, NNPS, CD
RB	RBR, RBS
VB	VBD, VBG, VBN, VBP, VBZ, VV
S	SBAR, SBARQ, SINU, SQ

在这个假设的基础上，部分研究者通过人工总结的方式来识别情感标签，他们工作的有效性也充分验证了此假设的正确性。基于这条假设，在本文构建的大规模语料库中，能够充分体现出具有正确修饰关系的句法路径的统计特性，可以真实的反映出句法路径是否正确。

在本文的句法路径中，评价对象与评价词具有较严格的位置关系限制，对于任何一条句法路径来说，评价词与评价对象的相对位置是确定的。表 3-2 列出了一些最为频繁的句法路径示例，其中第一列表示了句法路径内容，第二列给出了该句法路径在语料库中的统计频率，本文根据此频率为句法路径设定匹配优先级，频率越高，则其正确的可能性越大，优先级也就越高。当为评价词寻找其修饰的评价对象时，优先级高的句法路径将被优先选择，最后一列“位置”具体描述了评价对象与评价词之间的相对位置，“Front”代表了评价词位于评价对象的前面，即句法路径的最左端；而“Back”代表了评价词位于评价对象的后面，即句法路径的最右端。

表 3-2 高频句法路径示例

句法路径	出现频率(次数)	位置
JJ↑NP↓NN	26,223	Front
NN↑NP↑S↓VP↓ADJP↓JJ	17,528	Back
NN↑NP↑S↓VP↓NP↓JJ	5,413	Back
JJ↑NP↑PP↓NP↓NN	5,142	Front
.....

3.3 基于句法路径的情感标签抽取

本章基于句法路径库在评论文本中抽取情感标签，首先使用精确匹配句法路径库的方法，此时存在召回率偏低的情况，因此，在匹配句法路径库时加入

了编辑距离计算进行松弛，下面进行详细介绍。

3.3.1 基本情感标签抽取算法

情感标签包括评价词与评价对象两部分，本文首先通过评价词典（HowNet 的评价词典）在情感句中进行正向最大匹配，找出情感句中的评价词。进而基于句法路径使用精确匹配的方式寻找评价词对应的评价对象，匹配流程如下：首先，对情感句进行短语句法分析，获取其短语结构句法树，进而，获取情感句中评价词与所有候选评价对象之间的短语句法路径（候选评价对象为名词或代词），然后对获取的所有短语句法路径进行泛化，将泛化后的句法路径与标准句法路径库中的句法路径进行精确匹配，匹配包括句法路径内容以及评价词和评价对象的相对位置，最终，选取匹配成功的优先级最高的句法路径作为结果，抽取此句法路径的评价对象与评价词构成情感标签，对于无法成功匹配任何标准句法路径库中的路径的情况，则认为评价词在此句中不具有修饰的评价对象，过滤掉。对于评价词含有多个相同最优句法路径的情况，选取出现在句子最前端的评价对象与其组成情感标签，这种情况通常由于评价对象为多个名词组成的名词短语造成，因此，抽取其中的任何一个名词作为结果均可。

句法路径精确匹配的方法可以准确的识别出情感句中的情感标签，但是存在以下两点不足：

- 句法路径库使用出现频率作为阈值获取标准句法路径，这就导致部分出现次数较少的句法路径会被过滤掉，导致句法关系涵盖不全。因此，句法路径精确匹配的算法会使得部分潜在的情感标签被漏掉。
- 语料本身存在一定噪声，并不是所有的句子都严格按照语法表述，同时，句法分析器的分析结果并不完全准确，也会导致一定的错误，而精确匹配算法对语料噪声以及错误的句法分析结果十分敏感，会直接导致这些句子中的情感标签无法找出。

图 3-4 给出了一个包含以上两点不足的实例。右侧虚线括起的部分包含一个情感标签“poor—quality”，评价词与评价对象之间的句法路径为“JJ↑ADJP↑NX↓NN”，此句法路径虽然可以体现出评价词与评价对象间的修饰关系，但是由于其出现频率较低，并不包含在标准句法路径库中，导致情感标签丢失。对于左侧虚线括起的部分，“poor—color”是其包含的一个情感标签，但由于句法分析出现错误，导致识别出“JJ↑NP↓NX↓NN”这一错误结

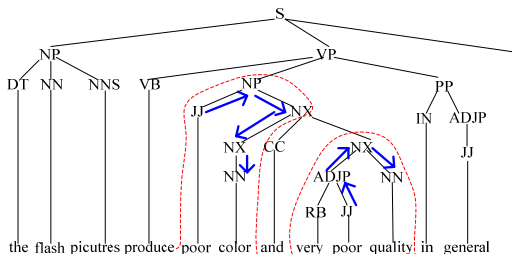


图 3-4 包含两点不足的句法分析结果

通过观察，我们发现句法路径库中的路径“JJ↑NP↓NN”和路径“JJ↑ADJP↑NP↓NN”可以分别近似的描述图 3-4 中的两个情感标签所在的句法路径“JJ↑NP↓NX↓NN”和“JJ↑ADJP↑NX↓NN”。针对此类情况，本文在句法路径匹配中加入了编辑距离的计算，进而通过计算结果判定两个句法路径是否表示了相同的修饰关系。下面对具体方法进行详细介绍。

编辑距离又称Levenshtein距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。例如，将字符串“kitten”转换成“sitting”共需要三步操作：(1)sitten (k→s)；(2)sittin (e→i)；(3)sitting (→g)。若每步操作定义的cost均为 1，则这两个字符串之间的编辑距离为 3。

编辑距离的算法目前已经相当成熟，定义好编辑操作的代价之后，可用动态规划算法对两个句法路径的编辑距离进行求解，对于分别含有 m 和 n 个基本单元的句法路径来说，计算它们之间编辑距离的时间复杂度为 $O(mn)$ 。其算法核心思想如下：设含有 m 个基本单元的句法路径为 a ，含有 n 个基本单元的句法路径为 b ，使得 a 与 b 变换成相同句法路径的方式有三种：1、将 b 的前 $n-1$ 个基本单元所构成的句法路径变换成 a ，然后删除 b 的最后一个单元；2、将 a 的前 $m-1$ 个基本单元所构成的句法路径变换成 b ，然后删除 a 的最后一个单元；3、将 a 的前 $m-1$ 个基本单元与 b 的前 $n-1$ 个基本单元变换成相同句法路径，然后将

a和b的最后一个基本单元变换成相同单元。根据这三种情况，我们可以得到求解编辑距离的递推式，格式如下：

$d[i][j] = \min\{d[i-1][j]+1, d[i][j-1]+1, d[i-1][j-1] + \text{cost}(a[i],b[j])\}$ ，算法的伪代码描述图 3-5 所示：

```
//初始化
For i: 0...m
  d[i, 0] = i
For j: 0...n
  d[0, j] = j
//用动态规划方法计算Levenshtein距离
For i: 1...m
  For j: 1...n
  {
    //d[i,j]的Levenshtein距离
    d[i, j]= minimum( d[i-1, j] + 1,  d[i, j-1] + 1,  d[i-1, j-1] + cost)
  }
//返回d[m, n]
return d[m, n]
```

图 3-5 编辑距离算法伪代码描述

3.3.2.2 编辑距离应用于句法路径匹配

为了将句法路径之间的相似度转换到编辑距离算法上，首先需要定义句法路径中的基本转换单元以及相应的转换代价。句法路径中包括句法成分标签以及指示方向两个要素，本文将这两个要素合并作为一个基本的句法路径转换单元，如图 3-6 中的“JJ↑”、“NX↓”等，相应的，针对这个转换单元，定义与字符串编辑距离相同的添加、删除以及替换三种操作，并且这三种操作的转换代价均为 1。

基于如上定义，我们可以通过计算两个句法路径的编辑距离来考察二者的相似程度。按照此算法，图 3-6 左侧图的句法路径“JJ↑NP↓NX↓NN”与“JJ↑NP↓NN”之间的编辑距离为 1，表现为图中用虚线指示的一个删除操作；相似的，右侧图中的句法路径“JJ↑ADJP↑NX↓NN”和“JJ↑ADJP↑NP↓NN”之间的编辑距离也为 1，表现为虚线指示的一个替换操作。

如图 3-6 所示，如果 $path$ 与 $path'$ 之间的编辑距离小于阈值 th_{ed} ，则本文认为 $path$ 与 $path'$ 之间表现了相似的评价对象与评价词之间的修饰关系，相应的，我们将 $path$ 修改为 $path'$ ，同时依据 $path'$ 的句法路径优先级抽取原句中的情感标签。

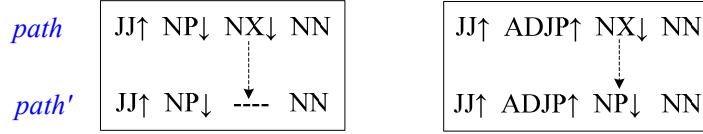


图 3-6 句法路径之间的编辑距离计算示例

3.4 实验结果及分析

3.4.1 实验数据及评价方法

情感标签抽取方面的研究目前还处于初级阶段，国内外并没有标准的情感标签抽取相关的测试语料，为此，本文从 www.epinions.com 的数码相机和 MP3 播放器两个典型的电子产品领域选取语料，每个领域各找出 600 句情感句，经过人工标注后获取标准评测集合，标注后数码相机领域获取情感标签 894 个，MP3 播放器领域获取情感标签 703 个。本文使用 P、R、F 值的方式进行评价，具体计算公式定义如下：

$$P = \frac{\text{系统识别出的正确情感标签个数}}{\text{系统识别出的情感标签总个数}} \quad (3-1)$$

$$R = \frac{\text{系统识别出的正确情感标签个数}}{\text{标准评测集中情感标签总个数}} \quad (3-2)$$

其中， P 为准确率， R 为召回率，调和平均值 F 定义如下：

$$F - score = \frac{2PR}{P + R} \quad (3-3)$$

3.4.2 对比实验设置

为了验证本文方法的有效性，本文选取了目前已有的两种情感标签抽取方法作为对比实验，具体如下：

- 1、**最近邻方法**：很多情感分析技术中，学者们都近似的认为评价词语修饰的是距离其最近的评价对象。基于此，本文使用最近邻方法作为第一个对比

实验。首先，我们基于评价词典识别出情感句中的所有评价词，进而，为每个评价词选取距离其最近的名词或代词作为其对应的评价对象，得到的二元对即为情感标签抽取结果。

- 2、**Bloom 等人的方法**：Bloom 等人首次提出了情感标签的概念，并且手工构建了 31 条依存句法规则，工作较其他研究者更为细致，可以看作是传统的基于规则/模板方法的代表，为此，本文重现了他们的方法作为对比实验。同样的，首先需要基于评价词典找出情感句中的所有评价词，进而，使用他们定义的规则为评价词查找评价对象，最终将获取的二元对作为情感标签抽取结果。

3.4.3 实验结果分析

本文选取 Charniak 短语句法分析器对情感句进行句法分析，经过实验观察，发现当出现频率最高的前 70 条句法路径用于匹配时，系统性能达到最优。为此，本文首先选取前 70 条句法路径，使用精确匹配的算法，与设置的两个对比实验结果进行比较，进而，分别对编辑距离以及句法路径的数量对结果的影响进行了详细分析，下面进行具体介绍。

3.4.3.1 对基于句法路径精确匹配的情感标签抽取方法的评价

表 3-4 首先给出了本文基于句法路径精确匹配算法与两个对比实验（最近邻方法、Bloom 等人的方法）的结果，可以看出，在两个领域的标准测试集上，本文的方法在各个实验指标上均超过了这两个对比实验。

表 3-4 情感标签抽取中各种方法的对比实验结果


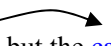

实验方法	测试领域	Camera			MP3 player		
		<i>P</i> (%)	<i>R</i> (%)	<i>F</i> (%)	<i>P</i> (%)	<i>R</i> (%)	<i>F</i> (%)
最近邻方法		65.80	65.44	65.62	60.65	66.00	63.22
Bloom 等人的方法		85.16	66.11	74.43	84.88	63.87	72.89
本文的方法		86.19	78.19	81.99	85.53	76.53	80.78

从表 3-4 中可以看出，对于两个领域来说，最近邻方法的准确率均最低。这是因为最近邻方法定义的规则过于简单，没有考虑情感句中评价词与评价对象之间的修饰关系，经验性过强。同时，其默认评价词只要在情感句中出现，就一定会为其寻找一个评价对象组成情感标签，这是导致其准确率偏低的主要原因。表 3-5 给出了一些典型的情感句实例，首先看其中的前两个句子，其中箭头的起始点为评价词，终止点为使用最近邻方法得到的评价对象，箭头表示评价词与评价对象之间的修饰关系。可以看出，由于最近邻方法仅找距离评价

词最近的评价对象，导致了第一句话中的评价词“friendly”被错误的与“user”匹配形成情感标签，同时，第二句中的评价词“good”也会被错误的与“camera”组成情感标签。同最近邻方法相比，本文提出的基于句法路径的方法可以很好的解决这两个问题，通过句法路径对修饰关系进行分析，可以非常准确的找出第一句话中的“friendly-interface”以及第二句话中的“bad-work”这两个情感标签，同时，又可以很好的将第二句话中错误的情感标签“good-camera”过滤掉。

可以看出，对于两个领域来说，Bloom 等人的方法以及本文的方法的结果均要好于最近邻方法，这充分说明了深入挖掘评价词与评价对象之间关系要比单纯的考虑位置更有价值，同时也说明句法关系在挖掘情感标签中的有效性。

表 3-5 情感句实例对比

编号	情感句	标注结果	句法路径
1	The interface of this camera is very  user friendly.	friendly-interface	NN↑NP↑S↓VP↓ADJP↓JJ
2	I was  good but the camera did a  bad work.	bad-work	JJ↑NP↓NN
3	Very satisfied with this camera's picture.	satisfied-picture	JJ↑ADJP↓PP↓NP↓NN
4	The only drawback is the speed between shots.	drawback-speed	NN↑NP↑S↓VP↓NP↓NN

可以看出，虽然 Bloom 等人的方法准确率较高，但召回率却比较低，造成这个局面的原因主要是 Bloom 等人的方法句法规则均由人工整理制定，在制定的过程中必然会出现定义不全的情况。而本文的方法能够从未标注语料中自动挖掘句法路径，能够保证句法路径的准确并且全面，这点可以从系统指标看出，本文的方法在准确率以及召回率上都达到了最好的效果，充分说明了本文方法在挖掘评价对象与评价词之间修饰关系上的有效性。观察表 3-5 中的 3、4 两个句子，由于手工制定的句法关系规则并不全面，导致 Bloom 等人的方法无法找出这两个句子中的情感标签，而本文自动构建的标准句法路径库包含“JJ↑ADJP↓PP↓NP↓NN”、“NN↑NP↑S↓VP↓NP↓NN”这两条句法路径，从而能够准确的找出这两个句子中的情感标签“satisfied-picture”以及“drawback-speed”。

综上所述，通过对表 3-4、表 3-5 中实验结果以及实例的深入分析，充分说明了本文所找出的高频句法路径在大部分测试实例中起到很好的效果，也说明了本文自动句法路径获取的方法是有效的。同时，本文所获取的句法路径是使用数码相机领域的语料获取的，但通过观察 MP3 播放器领域的结果可以发现，虽然 MP3 播放器领域的系统指标较数码相机领域稍低一些，但这主要是

因为 MP3 播放器领域的评论语料撰写格式更加随意，噪声明显多于数码相机领域。可以看出，使用数码相机领域语料构建的标准句法路径库在其它领域仍然使用，这充分说明了标准句法路径具有领域适应性，领域移植十分方便。

通过对实验结果的观察，我们发现基于句法路径精确匹配的情感标签抽取方法召回率偏低，在两个领域分别为 78.19%和 76.53%，为了进一步提高系统召回率，减少精确匹配带来的局限性，本文在句法路径匹配的过程中加入了编辑距离算法来提升系统召回率，下面对编辑距离对系统指标的影响进行详细分析。

3.4.3.2 对基于编辑距离的句法路径匹配算法的评价

基于句法路径精确匹配的方法存在召回率偏低的问题，因此，本文引入编辑距离来对句法路径进行模糊匹配，更进一步的使句法路径在情感标签抽取中发挥作用。在编辑距离实际应用于模糊匹配时，不同的编辑距离阈值设置会造成系统最终指标的不同，图 3-7 给出了两个领域系统指标随编辑距离阈值的变化情况。

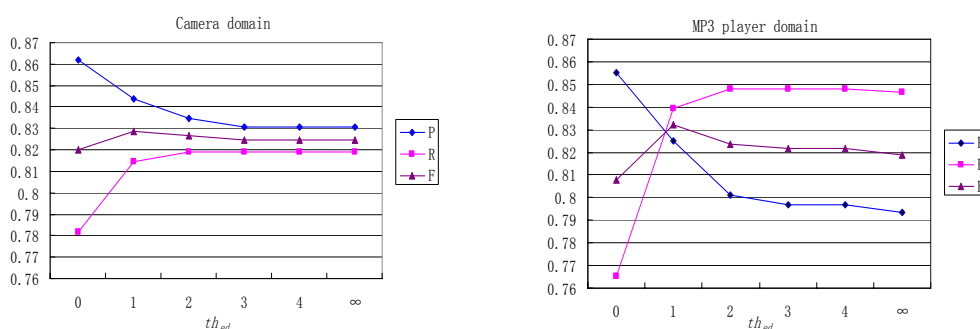


图 3-7 编辑距离阈值 th_{ed} 对系统性能指标的影响

观察图 3-7 中两个领域的曲线,我们可以看出,随着编辑距离阈值的不断提升,准确率在持续下降,而召回率则不断上升。当编辑距离阈值设为 1 时,情感标签抽取系统的性能达到最优, F -score 分别达到了 82.87%和 83.22%。同时,召回率得到了显著的提升,较精确匹配的召回率分别提高了约 3%和 7%。

当编辑距离阈值设为 0 时,系统即为基于句法路径精确匹配的情感标签抽取系统,相对于阈值为 1 的情况下,虽然它的准确率相对高一些,但是过于严格的匹配会导致召回率下降很多。由图 3-7 可以看出,当编辑距离阈值提升到一定程度后,系统指标趋于平稳,这是因为系统此时已经转变为为情感句中的每个评价词语选出一个评价对象,即任何两个句法路径之间经过阈值步的修改

均可互相转换。仅管如此，通过观察图 3-7 中的数据曲线，我们可以发现无论编辑距离取何值，基于编辑距离的句法路径匹配算法性能都要优于基于句法路径精确匹配的方法，这也进一步证明了编辑距离在句法路径匹配过程中的作用。

3.4.3.3 句法路径的数量分布对情感标签抽取性能的影响

如前面所述，句法路径的完备性对情感标签识别系统的性能有很大影响。但实际上，并不是句法路径的数量越多，系统的性能就越好。图 3-8 给出了数码相机领域系统指标随句法路径数量增多而发生的变化，句法路径的选取依照出现频率排序，如“10”代表取最频繁的前 10 条句法路径进行实验。

从图 3-8 可以看出，当句法路径库规模从 1 变到 10 时，系统指标得到了迅速提升，这充分说明高频句法路径在情感标签抽取中的有效性，也间接证明了本文标准句法路径库构建方法的合理性。当句法路径数量从 10 变化到 70 时，系统指标增长速度越来越慢，并在取前 70 条句法路径时系统性能达到最优。虽然句法路径数量增加了很多，但系统 F-score 大概仅增加了约 4%，这种现象可以用句法路径在语料中出现的次数来解释。情感标签中的评价对象与评价词之间的修饰关系规律性很强，且人们日常表达时频繁使用的句法关系数量有限，自然可以用较少的句法路径来识别大部分的情感标签。其余的 60 条次频繁句法路径则用来解决评价对象与评价词二者句法关系不太常见的情感标签的识别，进一步改善性能。当句法路径的数量超过 70 时，引入的噪声开始增加，导致系统性能开始缓慢下降。

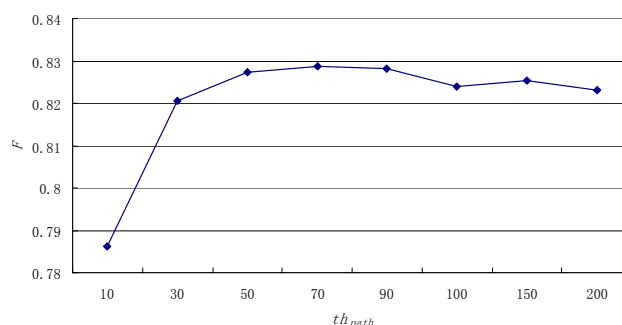


图 3-8 数码相机领域句法路径库规模 th_{path} 对系统性能的影响

3.5 本章小结

本章提出了一种基于句法路径的情感标签抽取方法。本方法能够自动从大规模未标注语料库中获取标准句法路径，从而建立评价对象与评价词之间的句法路径关系集合。使用此标准句法路径库在情感句中匹配即可抽取出情感标签，实验表明此方法取得了比较好的效果，更进一步，本章在句法路径匹配时引入了编辑距离计算，改进了情感标签抽取系统的性能。最后，本文对句法路径数量对系统指标的影响进行了总结、分析，再次验证了本文自动构建句法路径算法的正确性。

第4章 基于主题模型的情感标签标注

第 3 章提出了一种基于句法路径的情感标签抽取方法，实验表明此方法在抽取情感标签时达到了比较好的效果，但对于评价对象或评价词在文本中未显式出现的隐式标签来说，此方法无法对其识别。针对此问题，本章尝试将主题模型应用于情感标签标注中，下面对具体方法进行详细介绍。

4.1 主题模型理论基础

主题模型本质上是一种文档生成模型，其核心思想在于认为文档是若干主题的混合分布，而每个主题又是关于单词的一个概率分布。主题示例如图 4-1 所示。

Topic 1		Topic 2		Topic 3		Topic 4	
word	prob.	word	prob.	word	prob.	word	prob.
drugs	.069	red	.020	mind	.081	doctor	.074
medicine	.027	blue	.099	thought	.066	patient	.061
effects	.026	green	.096	remember	.064	hospital	.049
body	.023	yellow	.073	memory	.037	care	.046
pain	.016	white	.048	thinking	.030	medical	.042
person	.016	color	.048	professor	.028	nurse	.031
marijuana	.014	bright	.030	forget	.012	health	.025
label	.012	cross	.029	moment	.020	medicine	.017
alcohol	.012	orange	.027	thing	.016	dental	.015
dangerous	.011	brown	.027	wonder	.014	physician	.012

图 4-1 主题样例

主题模型定义了这样一个文档生成过程：为了生成新的文档，主题模型首先为要为该文档生成一个关于主题的概率分布，对于新文档要生成的每一个词，基于已生成主题的概率分布随机得到某个主题，接着再通过该主题的单词分布随机得到一个单词。图 4-2 给出了一个简单的基于主题模型生成文档过程的示意图，主题 1 和 2 分别与 money 和 river 相关，图中给出了这两个主题包含的词以及相应的生成概率。根据要生成文档的主题概率分布，即可为文档选择主题并进而选择文档的生成词。图 4-2 中主题到文档的箭头表示文档生成主题的概率，可以看出，文档 1 仅包含主题 1，文档 3 仅包含主题 2，因此，这

两个文档中的词均来自于其所在的单一主题，而文档 2 包含主题 1 和主题 2，所以文档 2 中的词由这两个主题混合生成，单词右上角的标号表示的是生成当前词的主题标号。

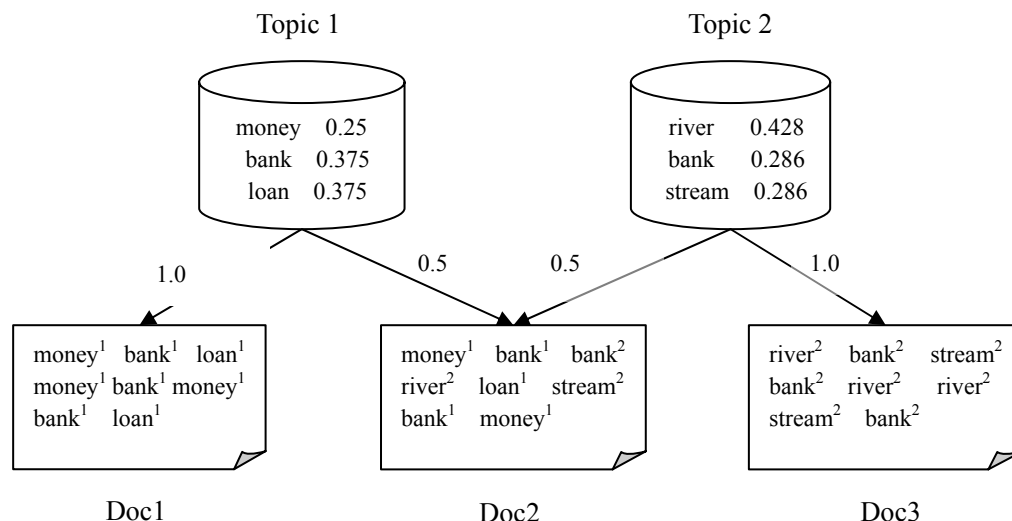


图 4-2 文档生成示例

LDA (Latent Dirichlet Allocation) 是目前应用最广泛的一种主题模型，相比于其他模型来说，LDA 的文本生成假设更加全面。下面对 LDA 进行详细描述。

4.1.1 LDA 数学描述

对于给定的文本，设文本在主题 z 上的概率分布为 $p(z)$ ，设特定主题 z 上单词 w 的概率分布为 $p(w|z)$ 。对于特定文本，我们用 $p(z_i = j)$ 表示在生成文本的第 i 个单词时选择第 j 个主题的概率， $p(w_i | z_i = j)$ 表示在主题确定为 j 的情况下，生成单词 w_i 的概率。进而，一个文本生成某个词的概率可表示为公式 4-1：

$$P(w_i) = \sum_{j=1}^T P(w_i | z_i = j) P(z_i = j) \quad (4-1)$$

其中， T 表示主题数目，公式分为两部分，左侧表示单词在主题 j 上的概率分布，右侧表示主题在文档 d 上的概率分布，这也是主题模型中最重要的两

个分布，为了便于后续说明，令 $\phi^{(j)}$ 表示单词在主题 j 上的概率分布， $\theta^{(d)}$ 表示主题在文档 d 上的概率分布。为了使模型更加合理，LDA 模型在计算分布 ϕ 和 θ 时加入了共轭先验。LDA 模型采用 Dirichlet 作为共轭先验，对于 T 维多项式分布 $p = (p_1, \dots, p_T)$ 来说，其对应的概率密度定义如公式 4-2 所示：

$$Dir(\alpha_1, \dots, \alpha_T) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^T p_j^{\alpha_j - 1} \quad (4-2)$$

超参数 α_j 可看作文本中主题 j 的出现次数的先验知识，即在观察到文本中任何一个单词之前就已经假定主题 j 已经在文本中出现了 α_j 次。LDA 采用平衡的 Dirichlet 分布作为先验，即令超参数 $\alpha_1 = \alpha_2 = \dots = \alpha_T = \alpha$ ，这样可以使模型更加简化。图 4-3 揭示了超参数 α 是如何对主题分布产生影响的，其中，三角形代表了一种可以表示所有可能的主题概率分布的坐标系，三角形的每个顶点代表一个主题，对于三角形中的任意一个点 $p = (p_1, p_2, p_3)$ ，有 $p_1 + p_2 + p_3 = 1$ ，其中，颜色较深的地方表示其概率更高。由图 4-3 中左右两个图形对比可知，若超参数 α 越大，主题分布受先验知识的影响也就越大，则多项式分布越向三角形的中心收敛，三个主题的分布也就越均匀。实际应用中， α 通常不会取的很大，这样使得概率分布更加趋近于角落，从而达到区分主题的目的。

类似的，LDA 模型在分布 ϕ 上也加入一个对称 Dirichlet 先验 β ，结构如图 4-4 所示，其中， D 表示文本， w 表示单词， z 表示主题， N_d 表示文本 d 含有词数，加入超参数 α 和 β ，我们可以通过 $Dir(\alpha)$ 和 $Dir(\beta)$ 来控制主题分布 θ 和主题上的单词分布 ϕ ，进而通过 θ 获取主题 z ，再通过 z 与 ϕ 得到单词 w 。

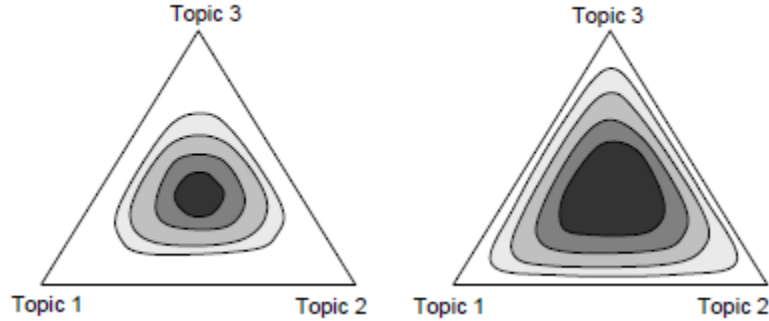
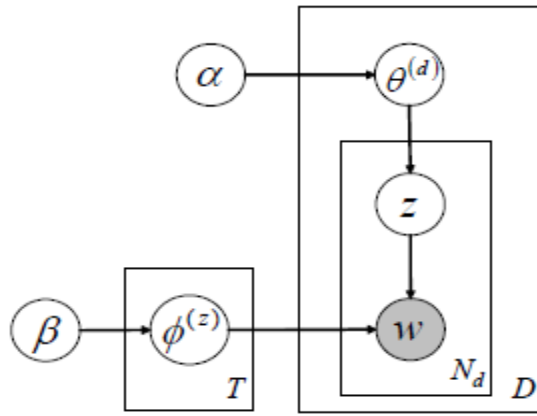

 图 4-3 Dirichlet 先验对主题概率分布的影响，左侧 $\alpha=4$ ，右侧 $\alpha=2$


图 4-4 LDA 的图形化表示

4.1.2 Gibbs Sampling 算法

LDA模型的关键在于获取文本上主题的概率分布 θ 以及主题上单词的概率分布 ϕ 。Minka T等人最先EM算法来构建LDA模型^[37]，但其往往无法找到最优解。D. M. Blei等人提出了变分法来构建LDA模型^[25]，但这种方法构建的模型与真实情况有一定偏差。Griffiths TL等人提出了Gibbs抽样算法^[38]，此方法是一种马尔科夫链蒙特卡洛方法，它的公式相对简单，容易理解和实现，抽取主题非常高效。因此，目前Gibbs抽样算法是最流行的LDA模型构建算法。

Gibbs抽样算法并不直接对 θ 和 ϕ 进行计算，而是根据文本集中可见的单词序列，为文本中的每个词符赋予一个主题，进而估计出 θ 和 ϕ 。为此，文本集中的每个词符 i 均对应一个主题变量 z_i ，在算法迭代的过程中， z_i 被不断赋予某个主题 t 。我们用 w_i 和 d_i 来表示词 i 在文本集合中的词汇索引与文档索引。该算

法依次考察文本集中的每个词符，假设文本集中其他词符的主题赋值已经确定，进而估计出当前词符的主题赋值。词符在主题上的概率分布公式为 $P(z_i = j | z_{-i}, w_i, d_i, \bullet)$ ，其中 $z_i=j$ 表示词符 i 的主题赋值为 j ， z_{-i} 表示除当前词符外文本集合中的此词的主题赋值， \bullet 表示其他所有的已知或者可见的信息，如其他所有词的词汇索引 w_{-i} 和文档索引 d_{-i} ，以及超参数 α 和 β 等。其具体计算公式 4-3 所示^[38]：

$$P(z_i = j | z_{-i}, w_i, d_i, \bullet) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^W C_{w j}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^T C_{d_i t}^{DT} + T\alpha} \quad (4-3)$$

其中， C^{WT} 、 C^{DT} 分别是 $W \times T$ 、 $D \times T$ 维的整数矩阵； $C_{w j}^{WT}$ 表示除当前词符 i 外单词 w 被赋予主题 j 的次数， $C_{d j}^{DT}$ 表示除当前词符 i 外文本 d 中的词符被赋予为主题 j 的次数。Gibbs 抽样过程直接给出了每个词符的主题赋值。通过对主题赋值的统计，我们可以得到 ϕ 和 θ 的近似值 ϕ' 和 θ' ，计算公式如 4-4 所示：

$$\phi_i^{(j)} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^W C_{kj}^{WT} + W\beta} \quad \theta_i^{(j)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^T C_{dk}^{DT} + T\alpha} \quad (4-4)$$

4.2 基于主题模型的情感标签标注方法

主题模型包含 ϕ 和 θ 两个关键分布，分别表征了文本与主题、词与主题之间的联系。主题模型相当于将主题作为文本和词的中间媒介，通过将文本、词映射到主题这个中间媒介上，进而获取文本、词之间的联系。情感标签标注的本质就是为文本选取合适的标签，通过建立情感标签和文本之间的联系，就可以找出适合文本的标签，从而实现情感标签标注，系统框架如图 4-5 所示。

从系统结构可以看出，为了对文本进行情感标签标注，我们首先需要获取一个情感标签集合，进而建立起情感标签与主题之间的联系，再以主题为中间媒介为文本选取合适的情感标签进行标注，可以看出，解决此问题的关键在于如何将情感标签与主题之间建立联系，为此，本节提出了两种建立联系的方式，下面对其详细介绍。

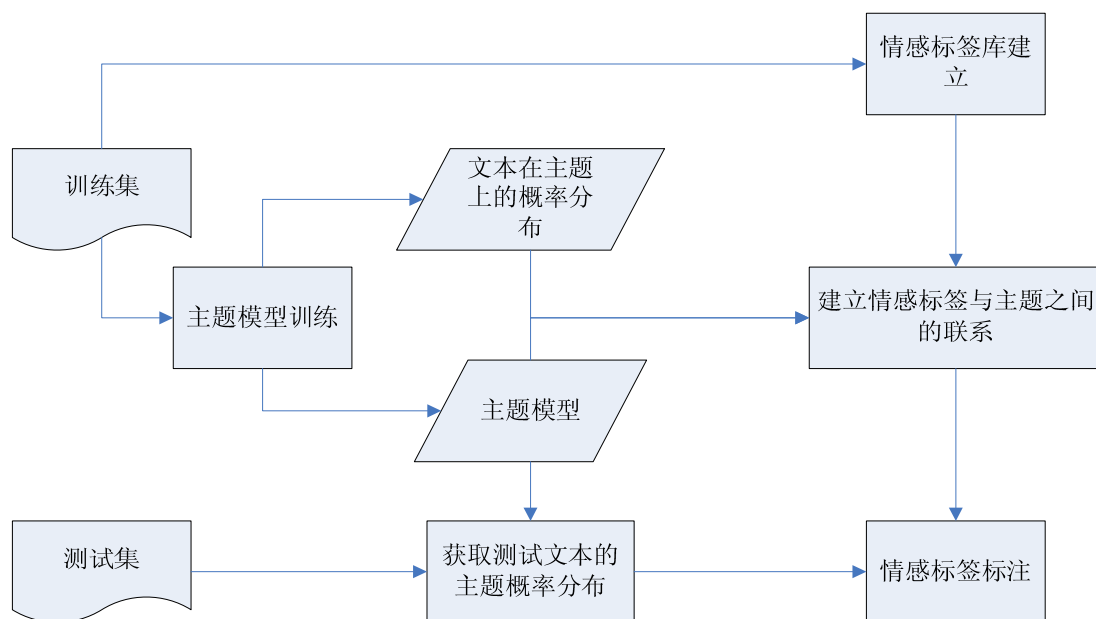


图 4-5 基于主题模型的情感标签标注系统结构图

4.2.1 语料及情感标签库获取

为了获取主题模型所需的训练集和测试集，本文从 www.epinions.com 网站上下载了数码相机领域的评论语料 1000 余篇。此网站的评论均由用户撰写，在撰写评论的同时，网站要求用户为自己所撰写的评论标注情感标签，其具体形式如图 4-6 所示：



图 4-6 评论文本样例

图 4-6 左侧的部分表示商品，右侧带有“Pros”与“Cons”标记的两行为用户为此评论标注的情感标签，“Pros”表示褒义，“Cons”表示贬义，此两行下面的文本为用户撰写的评论。

通过对分析评论格式，我们可以将评论中的情感标签抽取出来构成情感标签库。此时的情感标签库存在以下两个问题：

- 存在一定量的内容性标签：用户在撰写网络评论时，用词随意，写出的标签不一定都能作为情感标签，如“10x zoom”、“2.5" LCD”等，这些标签体现的是产品的固有属性，并不是情感标签。
- 标签本身是一个句子：用户有时不愿意花时间去想能够概括评论内容的情感标签，就把评论中的某一句话提取出来作为情感标签标注，如“Compact with 8.1 pixel and many fun features including manual aperture and speeds”等，这种标签对系统没有实际意义，不应该算作情感标签。

针对以上两个问题，本文首先使用评价词表对情感标签库进行了过滤，如果标签不含有评价词，则认为其不具有情感倾向性，直接过滤掉，使用的评价词表与第 3 章相同；对于第二类问题，本文将含有超过 6 个词的情感标签全部过滤掉，从而形成最终的情感标签库。

4.2.2 基于互信息的情感标签标注

主题模型是一个生成模型，对于给定文档 d 和特定词 w 来说，可通过公式 4-5 给出 d 生成 w 的概率：

$$P(w|d) = \sum_{j=1}^T P(w|z_i = j)P(z_i = j) \quad (4-5)$$

其中， z 表示主题， T 为主题个数。如果不考虑情感标签内部结构，仅将其看作一个词集合的话，则文档 d 生成情感标签 q 的概率就等于 d 生成 q 所包含的所有词的概率连乘，公式如 4-6 所示：

$$P(q|d) = \prod_{w_k \in q} P(w_k|d) = \prod_{w_k \in q} \sum_{j=1}^T P(w_k|z = j)P(z = j|d) \quad (4-6)$$

通过设定概率阈值，我们即可从标签库中选取标签对 d 进行标注，从而得到 d 的情感标签结果。但可以发现，使用概率连乘的方法存在这样一个问题，即情感标签包含的词越多，它由公式 4-6 得出的生成概率也就越小，即此公式偏好于为文档标注包含词数更少的标签，例如“good zoom”与“long battery life”这两个标签，即使文档表达了“long battery life”的含义，但由于此标签包含三个词，连乘之后概率较小，导致其不能被正确标出。为了解决这个问题，本节使用互信息来表征情感标签与文档之间的关系，互信息相当于在概率连乘的基础上加入了语言模型进行修正，从而很好的解决了连乘造成的长标签概率值过低的问题，互信息公式计算公式 4-7 所示：

$$I(q, d) = \log \frac{P(q, d)}{P(q) * P(d)} = \log \frac{P(q|d)}{P(q)} \quad (4-7)$$

其中， $P(q|d)$ 表示的是文档 d 生成标签 q 的概率，可由公式 4-6 计算得到， $P(q)$ 则为情感标签出现的概率，在具体计算时使用 q 在训练集中出现的概率对其进行近似。通过设定 $I(q, d)$ 的阈值，即可为文本标注情感标签。

4.2.3 基于概率分布相似度的情感标签标注

上节我们通过互信息的方式来为评论文本进行情感标签标注，这种方法忽略了训练集中情感标签与文本之间的联系这一重要信息。考虑这样一个例子：标签库含有“fast start-up”、“long battery life”、“long start-up”三个标签，训练集中有文本 d ，其包含两个标签“fast start-up”、“long battery life”，训练得到主题模型后，我们仍然使用这个文本 d 进行测试，则上节提出的方法很有可能为其标注“long start-up”这个标签。

为了解决上节方法的不足，本节通过训练集直接获取情感标签在主题上的概率分布，进而通过计算文本与情感标签在主题上的概率分布相似度来为文档选取标签。本方法基于这样一个假设：若两个文本表达的含义相近，则它们对应的情感标签应该相同。方法流程如下：

- 1、对于训练集中的文本，通过训练之后可以得到其在主题上的计数向量 X_d ，此向量的每一维表示训练结束后文本中的词被赋予某一主题的次数。
- 2、对于情感标签，其在主题上的计数向量 X_q 等于包含它的所有文本在主题上的计数向量的平均值，计算公式 4-8 所示，其中， $count(d)$ 表示包含此情感标签的文本个数。
- 3、对于测试文本，主题模型可以给出其在主题上的计数向量 X'_d 。
- 4、通过计数向量 X_q 与 X'_d 可以得到情感标签与测试文本在主题上的概率分布，概率计算公式 4-9 所示，进而可以通过概率分布计算得到测试文本与情感标签之间的相似度，从而为测试文本进行情感标签标注。

$$X_q = \sum_{d \in C, q \in d} X_d / count(d) \quad (4-8)$$

$$P(z = j | d) = \frac{X_j^{d'} + \alpha}{\sum_{k=1}^T X_k^{d'} + T\alpha} \quad P(z = j | q) = \frac{X_j^q + \alpha}{\sum_{k=1}^T X_k^q + T\alpha} \quad (4-9)$$

本节选择 Kullback-Leibler(KL)距离来计算文本与情感标签之间的概率分布

相似度, KL 距离又称相对熵, 能够表示两个随机分布之间的差异程度, 当两个随机分布的差异增大时, 它们的相对熵也增大, 其计算公式如 4-10 所示:

$$D(p, q) = \sum_{j=1}^T p_j \log_2 \frac{p_j}{q_j} \quad KL(p, q) = \frac{1}{2} [D(p, q) + D(q, p)] \quad (4-10)$$

通过设定相似度阈值, 即可为文档标注标签。

4.3 实验结果及分析

本文选取了 50 篇数码相机领域的评论人工标注情感标签, 评测采用标准的 P、R、F 方式, 指标定义如下:

$$P = \frac{\text{系统识别出的正确情感标签个数}}{\text{系统识别出的情感标签总个数}} \quad (4-11)$$

$$R = \frac{\text{系统识别出的正确情感标签个数}}{\text{标准评测集中情感标签总个数}} \quad (4-12)$$

其中, P 为准确率, R 为召回率, 调和平均值 F 定义如下:

$$F - score = \frac{2PR}{P + R} \quad (4-13)$$

两种方法的结果如表 4-1 所示, 基于概率分布相似度方法得到的结果要明显高于基于互信息方法的结果, 这说明训练集中情感标签与评论文本之间的对应关系是很有价值的信息, 同时也验证了本文对于相似评论文本之间情感标签应该相似这一假设。

表 4-1 实验结果表

	P(%)	R(%)	F(%)
互信息方法	20.01	24.25	21.93
概率相似度方法	37.05	47.02	41.44

为了验证主题模型在抽取隐式情感标签中的作用, 本文对隐式标签的数量进行了统计, 统计结果如表 4-2 所示。

表 4-2 隐式标签结果统计表

	隐式标签数(正确)	总标签数(正确)	隐式标签百分比(%)
标准答案	43	268	16.04
互信息方法	13	63	20.63
概率相似度方法	19	124	15.33

由统计结果可见, 标准答案中隐式标签约占 16%, 这表明隐式标签在情

感标签中占有一定的数量，具有一定的研究价值。同时，两种标签抽取方法抽取出的正确标签中均含有一定量的隐式标签，且与标准答案中隐式标签的百分比大体相当，这说明基于主题模型的情感标签标注方法并不对显式标签和隐式标签进行区分，能够抽取出一定数量的隐式标签，证明了主题模型能够在隐式标签抽取这项任务中发挥一定作用。

4.4 存在的问题及展望

对比两组实验，可以看出指标均不理想，虽然基于概率分布相似度的标注方法效果较好，但 F 值仍仅有 41%。通过对结果进行分析，本文认为造成指标较低的原因主要有以下几点：

- 1、主题模型获取文本上主题的分布时考虑的是文本中的所有词，而情感标签在文本中具有一定的局部性，即使是隐式标签，也通常在一句话中即可体现出来。当标注时引入文本中所有的词时，就会带来极大的噪声，例如，文本在第一段表达了“good”的主题，在最后一段表达了“zoom”的主题，则主题模型很有可能为文本标注“good—zoom”这一情感标签。
- 2、主题模型中的主题无法体现出情感倾向性。主题模型根据词出现的上下文以及文本集合中的主题分布情况为词划分主题，划分的过程中无法引入情感倾向性的信息，因此，即使是含义完全相反的两个标签，也有可能具有十分相似的概率分布，这就导致同一篇文档可能被同时标注两个含义完全相反的标签，这显然是不合理的。
- 3、用户在提供情感标签时，通常会有所遗漏，造成文本的情感标签标注不完全，这就导致基于概率分布相似度的方法无法充分获取情感标签与文本之间的关系，进而不能达到最佳效果。

情感标签抽取的研究工作目前仍处于起步阶段，目前对情感标签抽取问题的研究工作仍然很少，主题模型的优势在于不需要基于情感标签抽取的底层研究内容，避免了级联错误。实验结果表明，基于主题模型的情感标签标注方法能够克服传统方法无法处理隐式标签的问题，但当前系统得到的指标并不理想，仍有很多问题需要解决。目前主题模型已经被应用于很多问题的解决，也产生了越来越多的主题模型的变型，相信随着主题模型的发展，未来会在情感标签抽取方面发挥重要的作用。

4.5 本章小结

为了解决当前情感标签抽取方法无法处理隐式标签的问题，本章尝试将主题模型应用于情感标签抽取这项任务中，提出了基于互信息和基于概率分布相似度的两种具体标注方案，通过对两种方法的比较，验证了本文关于相似文本之间标签也应相似的假设，进而对隐式标签进行统计，证明了隐式标签具有一定的研究价值，验证了主题模型在抽取隐式标签任务中能够发挥一定的作用，并对目前存在的问题进行了分析和总结。

隐式标签抽取任务涉及到语义层面，是一项十分困难的任务，本章仅进行了探索性的实验，因此结果并不十分理想，这也说明了隐式标签抽取这项任务也需要未来更多、更加深入的研究去解决。

结论

情感标签包含了用户对于商品的详细评论信息，通过对情感标签进行汇总、整理，能够很方便的让用户以及商家了解产品的使用情况，因此，抽取评论文本中的情感标签在情感倾向性分析的任务中具有十分重要的意义。基于此，本文针对情感标签抽取的相关问题，主要进行了评价词集构建、评价对象抽取以及情感标签抽取三方面的研究。

评价词集构建方面，本文提出了融合语义知识库与大规模语料库的评价词集构建方法，首先以语义知识库以及大规模语料库为基础，构建一个候选评价词集合，进而在语料库中定位最能体现评价词情感信息的上下文，从而根据上下文为评价词制定相应置信度设置规则，为候选评价词排序，从而得到最终的评价词表。实验结果表明，本文的评价词集构建方法取得了比较好的效果。

评价对象抽取方面，本文首先基于短语结构获取候选评价对象集，进而提出了三种过滤方法对候选评价对象进行过滤，最终通过对评论文本中的句子进行分类来寻找用户评论的评价对象，实验结果表明了本文方法对于评价对象抽取的有效性，同时验证了本文对评价对象领域相关性较强这一假设。

情感标签抽取方面，本文提出了基于短语结构句法路径的情感标签抽取方法。首先，以评价词为中心在未标注的情感句语料库中挖掘句法路径，将得到的句法路径泛化、统计后得到标准句法路径库，进而基于标准句法路径库在测试集中挖掘情感标签，实验表明，本文方法在情感标签抽取问题上取得了比较好的效果，同时，也验证了本文标准句法路径库构建方法的合理性。为了解决挖掘情感标签过程中句法路径精确匹配带来的问题，本文引入编辑距离来计算句法路径之间的相似度，从而有效的提高了系统召回率。

为了解决传统情感标签抽取方法无法抽取隐式标签的问题，本文尝试将主题模型应用于情感标签抽取问题中，并提出了相似文本间情感标签也应相似的假设，实验验证了此假设的合理性，同时，通过对结果进行统计，表明了主题模型在抽取隐式标签问题中能够发挥一定作用，最后，本文对主题模型应用于情感标签抽取存在的问题进行了详细分析。

情感标签抽取技术还处于初级阶段，因此还有很广阔的研究空间，如：如何自动识别复杂评价对象，如何构造适合于情感标签抽取问题的主题模型变型等等，都将作为我们的下一步工作。

参考文献

- 1 Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995(12):1137~1143
- 2 V. Hatzivassiloglou, K. R. McKeown. Predicting the Semantic Orientation of adjectives. Proceedings of EACL-1997. 1997:174-181
- 3 J. Wiebe. Learning Subjective Adjectives from Corpora. Proceedings of AAAI. 2000. 2000:108-113
- 4 E. Riloff, J. Wiebe. Learning Extraction Patterns for Subjective Expressions. Proceedings of EMNLP-2003. 2003:105-112
- 5 P. Turney, M. L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information Systems (TOIS). 2003, 21(4):315-346
- 6 N. Kaji, M. Kitsuregawa. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007:1075-1083
- 7 S.-M. Kim, E. Hovy. Automatic Detection of Opinion Bearing Words and Sentences. Proceedings of IJCNLP-2005. 2005:61-66
- 8 S.-M. Kim, E. Hovy. Identifying and Analyzing Judgment Opinions. Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). 2006:200-207
- 9 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 年第 1 期: 14-20
- 10 A. Andreevskaia, S. Bergler. Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL). 2006:209-216
- 11 F. Su, K. Markert. Subjectivity Recognition on Word Senses via Semi-supervised Mincuts. Proceedings of NAACL-2009. 2009:1-9
- 12 A. Esuli, F. Sebastiani. Determining the Semantic Orientation of Terms

- Through Gloss Analysis. Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM). 2005:617–624
- 13 A. Esuli, F. Sebastiani. Determining Term Subjectivity and Term Orientation for Opinion Mining. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL). 2006:193-200
- 14 J. Kamps, M. Marx, R. J. Mokken, et al. Using WordNet to Measure Semantic Orientation of Adjectives. LREC. 2004: 1115-1118
- 15 F. Su, K. Markert. Fromwords to Senses: A Case Study of Subjectivity Recognition. Proceedings of Coling-2008. 2008:825–832
- 16 R. Mihalcea, C. Banea, J. Wiebe. Learning Multilingual Subjective Language via Cross-lingual Projections. Proceedings of the Association for Computational Linguistics (ACL). 2007:976–983
- 17 H. Takamura, T. Inui, M. Okumura. Extracting Semantic Orientation of Words Using Spin Model. Proceedings of the Association for Computational Linguistics (ACL). 2005:133–140
- 18 D. Rao, D. Ravichandran. Semi-supervised Polarity Lexicon Induction. Proceedings of EACL-2009. 2009:675–682
- 19 K. Moilanen, S. Pulman. Sentiment Composition. Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007). Borovets, Bulgaria, 2007:378–382.
- 20 Y. Choi, C. Cardie. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. Proceedings of EMNLP-2008. 2008:793–801
- 21 J. Yi, T. Nasukawa, R. Bunescu, et al. Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques. Proceedings of the IEEE International Conference on Data Mining (ICDM). 2003:423-434
- 22 M. Hu, B. Liu. Mining Opinion Features in Customer Reviews. Proceedings of AAAI-2004. 2004:755–760
- 23 A.-M. Popescu, O. Etzioni. Extracting Product Features and Opinions from Reviews. hltemnlp2005. 2005:339–346
- 24 王波, 王厚峰. 基于自学习策略的产品特征自动识别. 全国第九届计算语言学学术会议. 2007

- 25 D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003:993–1022
- 26 M. Steyvers, T. Griffiths. Probabilistic Topic Models. In T. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Erlbaum, 2007:427-447
- 27 I. Titov, R. McDonald. Modeling Online Reviews with Multi-grain Topic Models. *Proceedings of WWW-2008*. 2008:111–120
- 28 Hu MQ, Liu B. Mining and summarizing customer reviews. In: *Proceedings of KDD-2004*. New York: ACM, 2004. 168–177.
- 29 Kim S.M, Hovy E. Determining the sentiment of opinions. In: *Proceedings of Coling-2004*. Morristown, NJ, USA: Association for Computational Linguistics, 2004. 1367–1373.
- 30 Kobayashi N, Inui K, Matsumoto Y, Tateishi K, Fukushima T. Collecting evaluative expressions for opinion extraction. In: *Proceedings of the International Joint Conference on Natural Language Processing*. New York: Springer, 2004. 584–589
- 31 Bloom K, Garg N, Argamon S. Extracting appraisal expressions. In: *HLT-NAACL 2007*. Association for Computational Linguistics, 2007. 308–315.
- 32 Popescu A.M, Etzioni O. Extracting product features and opinions from reviews. In: *hltemnlp2005*. Morristown, NJ, USA: Association for Computational Linguistics, 2005. 339–346.
- 33 姚天昉, 聂青阳, 李建超等. 一个用于汉语汽车评论的意见挖掘系统. *中国中文信息学会成立二十五周年学术年会论文集*. 2006, 260–281.
- 34 B. Liu, M. Hu, J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of WWW-2005*. 2005:342–351
- 35 T. Wilson, P. Hoffmann, S. Somasundaran. Opinionfinder : A System for Subjectivity Analysis. *Proceedings of HLT/EMNLP2005 Demonstration Abstracts*. 2005:34–35
- 36 Girolami M, Kaban A. On an equivalence between plsi and lda[c], *Proc of the 26th ACM SIGIR*. 2003:433-434
- 37 Minka T, Lafferty J. Expectation-propagation for the generative aspect model[C], *Proc of UAI 2002*.
- 38 Griffiths TL, Steyvers M. Finding scientific topics[J]. *The National Academy*

- of Sciences. 2004. 101:5228-5235
- 39 W. Buntine, J. Lofstrom, J. Perkio, S. Perttu, V. Poroshin, T. Silander, H. Tirri, A. Tuominen, V. Tuulos. A scalable topic-based open source search engine. In Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence, pages 228–234, 2004.
- 40 Limin Yao, David Mimno, Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In Proceedings of KDD, pages 937–946. 2009.
- 41 Hanna M. Wallach. Topic modeling: beyond bag of words. In International Conference on Machine Learning. 2006:977–984
- 42 Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 569–577. 2008.
- 43 林洋港. 概率主题模型在文本分类中的应用研究[M]. 硕士毕业论文. 中国科学技术大学. 2009:19–27

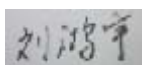
攻读学位期间发表的学术论文

- 1 刘鸿宇, 赵妍妍, 秦兵, 刘挺. 评价对象抽取及其倾向性分析. 中文信息学报. 2010 年第 1 期:84-88

哈尔滨工业大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《情感标签抽取相关技术研究》，是本人在导师指导下，在哈尔滨工业大学攻读硕士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。

作者签字：



日期：2010 年 6 月 24 日

哈尔滨工业大学硕士学位论文使用授权书

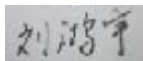
《情感标签抽取相关技术研究》系本人在哈尔滨工业大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归哈尔滨工业大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅，同意学校将论文加入《中国优秀博硕士学位论文全文数据库》和编入《中国知识资源总库》。本人授权哈尔滨工业大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

本学位论文属于（请在以下相应方框内打“√”）：

保密☐，在 年解密后适用本授权书

不保密☒

作者签名：



日期：2010 年 6 月 24 日

导师签名：



日期：2010 年 6 月 24 日

致谢

转眼间已经在 CIR 度过了近三年的时光。在这三年中，我体会到了如沐春风般的集体的温暖，老师孜孜不倦的教诲，参与到科研项目中的酸甜苦辣，最重要的是，在这里我懂得了如何保持良好的心态。在即将离开实验室之际，我谨在此向所有在工作、生活中帮助过我的老师、同学、朋友表示深深的谢意。

首先，感谢我的导师秦兵老师。她不仅在学术上对我悉心指导，更在日常生活中关心我。秦老师严谨的治学态度，敏锐的洞察力，丰富的研究经验是我首先要学习的，但更重要的是秦老师对待生活积极、乐观的心态深深的影响了我。感谢秦老师对我的关心，对我学习和工作上的帮助，对我成长的指导和影响！

感谢刘挺老师，是刘老师给了我进入 CIR 这个温暖集体的机会，让我学到这么多宝贵的知识和经验。刘老师高瞻远瞩的目光，灵活开放式的思维，深厚的文化底蕴，一直都是我所追求的一种目标和境界。

感谢我的 mentor 赵妍妍师姐对我的照顾，当我遇到学习和工作上的困难时，总是耐心的帮助我解决问题，教会了我很多书本上看不到的知识。

感谢 TM 组全体成员，我一个战壕里的战友，你们是我学习和生活中的催化剂，让我的研究生生活变得更加充实，更加精彩。

感谢 08 级的所有硕士生们，我们并肩作战，互相鼓励和帮助，结下了深厚的友谊。

感谢实验室所有的老师、师兄师姐、师弟师妹们，是大家使我感到了 CIR 大家庭的温暖，是大家激励我向着更高的目标努力。

最后，感谢所有关心、支持和帮助我的亲人和朋友们。我将把母校的培养和亲友们真挚的感情永藏心底。