单位代码 10006

学 号

分类号

密 级 秘 密

****

毕业设计(论文)

面向产品评论的情感要素抽取及情感倾向性分析

|  |  |
| --- | --- |
| 院（系）名称 | 计算机学院 |
| 专业名称 | 计算机科学与技术 |
| 学生姓名 |  |
| 指导教师 |  |

2016年4月

面向产品评论的情感要素抽取及情感倾向性分析 北京航空航天大学

北京航空航天大学

**本科生毕业设计（论文）任务书**

Ⅰ、毕业设计（论文）题目：

Ⅱ、毕业设计（论文）使用的原始资料（数据）及设计技术要求：

Ⅲ、毕业设计（论文）工作内容：

Ⅳ、主要参考资料：

学院（系） 专业类 班

学生

毕业设计（论文）时间： 年 月 日至 年 月 日

答辩时间： 年 月 日

成 绩：

指导教师：

兼职教师或答疑教师（并指出所负责部分）：

系（教研室） 主任（签字）：

注：任务书应该附在已完成的毕业设计（论文）的首页。

**本人声明**

我声明，本论文及其研究工作是由本人在导师指导下独立完成的，在完成论文时所利用的一切资料均已在参考文献中列出。

作者：

签字：

时间：2011年 6 月

面向产品评论的情感要素抽取及情感倾向性分析

学 生： 李子靓

指导老师：巢文涵

摘要

随着网购的迅速普及，网络用户积极在各电商平台上留下相关的产品评论。这些产品评论对于消费者和商家而言都是重要的参考信息，但评论数量海量，评论文本过长，评论质量低等因素大大降低了阅览者的信息获取效益。本文针对该问题，旨在从产品评论中抽取情感要素，即评价对象和评价词，并分析情感要素的情感倾向，为用户呈现简洁明了的产品信息。

本文的主要工作分为四个方面：

1. 本文使用条件随机场模型抽取评价对象，即产品的特性，共使用了6个学习特征，并使用两种名词剪枝技术过滤被误判的评价对象，能识别词语级和短语级的名词评价对象。
2. 本文使用句法分析器得到依存句法关系，再结合针对特定句式的分析规则抽取评价词。评价词是描述产品特性的词语，本文限定其为形容词。
3. 本文基于HowNet，筛选、扩展、整理出合适的评价词典，分析情感要素的情感倾向。给评价对象定义正向和负向的情感色彩，给评价词定义增强和削弱情感的作用，以此判断上下文相关的评价词的情感倾向。
4. 实现一个面向产品评论的情感要素抽取和情感倾向性分析的可视化系统。

**关键词：**产品评论，情感要素，情感倾向，条件随机场

Sentiment Element Extraction and Sentiment Polarity Analysis from Product Reviews

Author: Ziliang Li

Tutor: Wenhan Chao

Abstract

As online-shopping is swiftly spreading across the public, internet users are producing product reviews on e-commerce platforms. These product reviews provides vital information for both consumers and merchants. Some factors, however, like huge amount, too much content and poor quality, lower the efficiency of acquiring worthy information. In attempt to solve the problem and offer users clear and precise product information, methods mentioned in the article extract sentiment elements from product reviews, namely opinion targets and opinion words, and analyze sentiment polarities of the sentiment elements.

The work of this paper is mainly about four aspects:

1. Conditional Random Field model (CRFs) is applied to extract product traits, which is called opinion targets here. Six learning features are used. And two nouns-pruning techniques is used to eradicate wrongly-selected opinion targets. The method is able to recognize word-level and phrase-level noun opinion targets.
2. Relation between opinion target and opinion word is extracted by the method that uses syntactic dependency, produced by syntactic analyzer, and specifically-designed rules for specific sentence patterns. Opinion words are those describe product traits, confined to objectives here.
3. A sentiment lexicon is made based on HowNet. It is used to determine the sentiment polarity of sentiment element. In order to decide sentiment polarities of context-related opinion words, opinion targets are labeled as positive or negative, and opinion words are labeled as enhancing or weakening sentiment strength.
4. A product-comment-oriented visual system has been established that can extract sentiment elements and analyze sentiment polarities.

**Key words:** product reviews, sentiment element, sentiment polarity, CRFs

**目录**

[1 绪论 1](#_Toc451823332)

[1.1 研究背景及意义 1](#_Toc451823333)

[1.2 研究目标及要点 3](#_Toc451823334)

[1.2.1 研究目标 3](#_Toc451823335)

[1.2.2 研究要点 3](#_Toc451823336)

[1.3 论文结构 4](#_Toc451823337)

[1.4 研究现状 4](#_Toc451823338)

[1.4.1 情感分析 4](#_Toc451823339)

[1.4.2 情感分析的应用 4](#_Toc451823340)

[1.4.3 情感分析的评测 5](#_Toc451823341)

[2 相关技术 7](#_Toc451823342)

[2.1 评价对象的抽取 7](#_Toc451823343)

[2.2 评价词的抽取 7](#_Toc451823344)

[2.3 情感倾向性分析 8](#_Toc451823345)

[3 预处理 9](#_Toc451823346)

[3.1 分词 9](#_Toc451823347)

[3.2 词性标注 11](#_Toc451823348)

[3.3 去停用词 12](#_Toc451823349)

[3.4 句法分析 12](#_Toc451823350)

[4 基于条件随机场的评价对象抽取 13](#_Toc451823351)

[4.1 条件随机场的背景介绍 13](#_Toc451823352)

[4.2 条件随机场的数学定义 14](#_Toc451823353)

[4.3 条件随机场的关键问题 15](#_Toc451823354)

[4.3.1 特征函数的选择 15](#_Toc451823355)

[4.3.2 参数估计 15](#_Toc451823356)

[4.3.3 模型推断 16](#_Toc451823357)

[4.4 评价对象的识别 17](#_Toc451823358)

[4.4.1 标记方法 17](#_Toc451823359)

[4.4.2 学习特征 17](#_Toc451823360)

[4.5 条件随机场实现工具CRF++ 18](#_Toc451823361)

[4.5.1 CRF++文件格式 18](#_Toc451823362)

[4.5.2 CRF++使用方法 20](#_Toc451823363)

[4.6 训练集 21](#_Toc451823364)

[4.7 名词剪枝 21](#_Toc451823365)

[4.7.1 词频剪枝 22](#_Toc451823366)

[4.7.2 PMI剪枝 22](#_Toc451823367)

[5 基于依存句法关系的评价词抽取 23](#_Toc451823368)

[5.1 依存句法关系 23](#_Toc451823369)

[5.2 评价词抽取 24](#_Toc451823370)

[5.2.1 正常抽取 24](#_Toc451823371)

[5.2.2 特殊抽取一 24](#_Toc451823372)

[5.2.3 特殊抽取二 25](#_Toc451823373)

[5.2.1 特殊抽取三 26](#_Toc451823374)

[6 基于情感词典的情感倾向性分析 27](#_Toc451823375)

[6.1 HowNet 27](#_Toc451823376)

[6.2 情感倾向的判断 27](#_Toc451823377)

[6.2.1 上下文无关评价词 28](#_Toc451823378)

[6.2.1 上下文相关评价词 28](#_Toc451823379)

[7 实验 29](#_Toc451823380)

[7.1 实验数据集 29](#_Toc451823381)

[7.2 评判指标 29](#_Toc451823382)

[7.3 实验结果及分析 30](#_Toc451823383)

[8 可视化系统设计与实现 31](#_Toc451823384)

[8.1 系统设计 31](#_Toc451823385)

[8.2 开发环境 31](#_Toc451823386)

[8.3 系统实现 31](#_Toc451823387)

[结论 33](#_Toc451823388)

[致谢 35](#_Toc451823389)

[参考文献 36](#_Toc451823390)

[附录A 38](#_Toc451823391)

[附录B 39](#_Toc451823392)

# 绪论

## 研究背景及意义

随着Web2.0时代的到来，互联网为网络用户提供了方便的交流渠道和丰富的信息来源，导致用户产生的信息在网络上爆炸性地增长，人们越来越习惯于在网上表达自己对各种事件、物体、人物的看法。形式包括长篇幅的博客，和短篇幅的微博、商品评论、空间留言、朋友圈状态等。发表的内容也各色各样，有关于自身的经历，转发的文章，推送的广告或网页，对时事的看法，或产品使用感受等等。英文用户主要使用的社交平台有Facebook、Twitter等，电商平台有Amazon、eBay等。而中文用户主要使用的社交平台有博客、微博、微信、QQ空间等，电商平台有淘宝、京东、亚马逊等。 这些由网络用户产生的文本包含丰富的情感因素，反映了个人喜好。而当下有很多研究和应用都需以这样带有强烈个人情感倾向的文本为研究对象。以舆情系统为例，需要采集和分析大量的用户文本，经过文本处理和统计之后能得到多方面的信息，可用于分析热门话题的产生到消亡的过程，人们对于某些事件的舆论导向，消费者对于某些产品和企业的评价等。研究领域也越来越重视对这些文本的分析算法，“微博”、“博客”等字眼频繁出现在论文标题和主体中。这些应用和研究对于政府、企业、消费者等各类人群都有非常重要的意义，对于总结过往经验和预测未来走势都有指导作用。比如O’Connor等人[[1](#_ENREF_1)]研究了Twitter文本中的信息和民意调查结果之间的相似性，希望把主流社交类文本中的信息作为民意调查的补充和代替。Bollen等人[[2](#_ENREF_2)]利用Twitter中的情绪信息预测股票市场的行为。

作为情感分析文本的一种，产品评论也因为电商行业的迅猛发展如雨后春笋般大量产生。产品评论能很好地反映消费者的态度和观点，这对于商家和消费者而言都是极佳的参考。商家能从产品评论中得到消费者最直接的反馈，第一时间知晓发布的产品在消费者心中的形象；帮助商家总结经验，及时修整这一代产品的缺陷，构想下一代产品的设计方案，调整生产和销售策略。消费者在购物之前，更倾向于查阅产品评论，用他人提供的信息来评估商品是否满足自己的预期。虽然产品评论有巨大的参考价值，但是几个现实问题让人工提取信息的工作变得非常费时费力。第一是产品评论的数量非常庞大。据CNNIC中国互联网络发展状况统计报告称，到2014年6月份为止，我国网购用户数量约为3.32亿，由消费者产生的产品评论的数量更是浩如烟海。面对这样量级的数据，人工处理苍白无力。第二是评论内容各异，甚至相悖，如果不加以归类统计会让人非常困惑，然而人工统计工作量较大。第三是评论质量参差不齐，有些评论内容虽然长但有用信息很少，人工审阅这样的评论收益肯定很低。所以人工直接阅读评论原文只适用于小规模的产品评论，若想更全面地掌握产品信息，需要利用计算机来实现。这也是那么多研究者都致力于文本处理技术研究的原因。本文所做的研究就是为了解决上述的现实困扰，提高商家和消费者从产品评论中获取有用信息的效率。

## 研究现状

### 情感分析

研究带有情感色彩的文本的根本是从中抽取出有价值的信息，这个抽取的过程被称为情感分析，或者观点挖掘。虽然被称作此，但抽取出的信息绝不止用户情感或观点，其他信息如事件属性、物体特征、时间发展等也是可被抽取的对象。

在大量进行情感分析研究之前，前人所做的研究[[3](#_ENREF_3)]主要是围绕分析带有情感倾向的词语展开的。比如“美妙”是带有正向情感即褒义的词语，而“刺耳”是带有负向情感即贬义的词语。随着大量带有主观情感的文本出现在互联网上，研究者们的研究课题不再拘泥于简单的情感词研究，逐步发展到复杂性更高的情感句甚至篇章级别的研究。

根据分析的粒度不同，情感分析可分为粗粒度的情感分析和细粒度的情感分析。粗粒度的情感分析一般指多篇章级、篇章级的情感分析，即分析的最小单位为一篇文章。篇章级的情感分析一般有关键词提取、文档摘要或多文档摘要、情感倾向性分析。细粒度的情感分析一般指句子级、短语级、词语级的情感分析。这类分析的应用有情感要素抽取、情感倾向性分析等等。本文所研究的内容属于细粒度情感分析，达到短语级或词语级的粒度，即从句子级别的产品评论中抽取出短语级或词语级的表述。

根据文本类型可以把情感分析主要分为三类：面向微博的情感分析、面向新闻的情感分析和面向产品评论的情感分析。微博指的是出现在社交平台上的篇幅较短的博文，比如Twitter、Facebook、新浪微博、微信等。这类文本的特点是篇幅短，带有强烈的个人情感色彩，差异性大，实时性强，用语不甚规范，内容多为时下热门话题；因此常被用来研究热点话题、事件。由于近几年网络自媒体的高速发展，这类文本的数量急剧增长，也成为了情感分析的主要对象。新闻与微博相比有很大的不同，新闻一般由官方撰写，代表媒体机构和部分权力机关的观点和态度，所以遣词造句较为讲究、规范，态度也趋于客观、冷静，虽然也隐含情感色彩，但没有微博类文本具有的情感色彩那般强烈。产品评论的描述对象为产品，反映了产品的质量水平，能用来评估大众对产品的期许；语言一般较为简练，不是特别规范，包含大量产品特性的描述，带有较为鲜明的情感色彩，主要来源为电商平台，对于消费者和商家有重要的参考价值。

其中，产品评论是本文研究的对象。产品评论的内容主要关于产品的特性，比如手机领域经常被评价的特性是屏幕、音量、续航能力等。用户在阅读产品评论时最关心的一般就是产品特性的描述，这样的描述被称为情感要素。情感要素一般指评价对象、评价词、观点持有者等内容，在产品评论中观点持有者一般为消费者，所以不需特别考虑；因此产品评论的情感要素一般指的是评价对象和评价词。评价对象是指评价的物体的某种属性；评价词是描述评价对象的词语或短语。情感要素的情感倾向性分析指的是判断情感要素表达的情感是正向、负向还是中立的。以手机领域为例，情感要素的示例有“屏幕大”和“音量小”；前者的情感是正向的，而后者的情感是负向的。

### 情感分析的应用

情感分析是一种面向应用的研究课题，然而用户并不能直接使用情感信息提取和分类的结果。情感信息归纳和情感信息检索是情感分析技术服务用户的主要形式。情感信息归纳指的是从大量包含情感信息且与主题相关的文档中分析、归纳出情感分析结果，作为用户的参考，帮助用户节省查阅文档的精力和时间。而情感信息检索主要用于为用户检索出包含情感信息且与检索词相关的文档。

情感信息归纳的主要内容是生成文档摘要，包括生成新闻评论、产品信息等的摘要。情感信息检索的主要内容有三个方面：首先在传统的信息检索模型的基础上加入情感相关的知识，再对满足主题词的文档做主客观情感识别，根据主观情感词对文档做筛选和排序。

### 情感分析的评测

随着网络上越来越多带有情感色彩的文本出现，情感分析受到更多研究学者和研究机构的注意。近些年以来，许多国内外研究机构组织了大大小小的公开评测，以便交流和发展情感分析的技术。

国际文本检索大会（Text Retrieval Evaluation Conference，简称TREC）是关注情感分析的第一批机构之一，自2006年起，每年都会公开发布情感分析的评测任务。TREC最先关注的是博客检索任务（Blog Track），即要求在博客上检索满足特定话题的带有观点的文档，不能是纯客观的叙述，必须带有主观性信息[[23](#_ENREF_23)]。除此之外，还要对检索得到的文档进行情感分类的子任务，即把文档分类为褒义、贬义或者混合。评测发展到TREC2009，有更多的情感分析元素被加入到微博检索任务。比如，判断博主的性别，博主是专家还是普通人；判断博文是客观事实还是主观看法；判断博文是属于个人发表的还是公司发表的。这些有意思的评测任务吸引了许多相关研究者的参与。

同样在2006年举办的评测有NII Test Collection for IR Systems（简称NTCIR）发布的多语言观点分析任务（Multilingual Opinion Analysis Task, 简称MOAT）。这个测评每年都会如期举办，有中日英三种语言的语料库。NTCIR评测的要求是从新闻媒体报道中抽取带有主观色彩的信息。具体而言，是判断句子是否和文章的主题相关，并从句中得到评价对象、评价词的情感极性和观点持有者等信息。NTCIR的观点分析的目标是进行多信息源、多粒度、多语种、深层次的主观信息抽取[[20](#_ENREF_20)]。到NTCIR-8位置，MOAT任务加入了许多新的元素。比如情感问答，即对于某个给定的情感问题，需要从相关的文档中找出正确的回答；比如跨语言情感分析，即对于某个英文查询，从四种不同语言的文档库中找出相关的文档。

而国内对于汉语的情感分析研究也取得了一定的成绩。COAE（Chinese Opinion Analysis Evaluation）在2008年创立，是国内首个情感分析领域的评测。它推动了中文情感分析的理论研究和实际应用，也为建立中文情感分析基础数据集做出了贡献。到目前为止COAE已经连续举行了7次。另外，中文信息技术专业委员会创办的NLPCC自然语言处理与中文计算会议也成功举办了三届中文评测。自然语言处理和中文计算领域是NLPCC主要关注的方面。国内的情感分析评测任务基本上和学术前沿研究进程保持一致，大体上可分为三个方面：第一个方面是识别和分析中文评价词，倾向于词语级的情感倾向性评测；第二个方面是抽取中文情感要素，重点是抽取评价对象和评价搭配；第三个方面是判断中文文本情感倾向性，偏向于篇章级和句子级的评测。

情感分析领域的公开评测既为研究学者提供了一个学习、交流的机遇，也为该领域的发展搭建了很好的平台。

## 本文的研究内容

本文的主要研究内容有：

1. 基于条件随机场的评价对象抽取

在本文中，产品评论的评价对象为名词或名词短语。本文使用条件随机场模型抽取评价对象。

1. 基于依存句法关系的评价词抽取

依存句法关系描述的是词与词之间的关系。本文利用某些联系评价对象和评价词的依存句法关系，再结合第一部分找到的评价对象，就能抽取出评价词。

1. 基于评价词典的评价词抽取

本文使用的评价词典以第三方的公开词典HowNet为基础，经过人工和自动化的方式整理而成。通过在词典中查找评价词的方式得到评价词的情感倾向，进而判断情感要素的情感倾向。

1. 面向产品评论的情感要素抽取和情感倾向性分析系统的设计与实现

本文利用Java Swing实现系统，能对用户输入的产品评论进行情感要素抽取和情感倾向性分析。

## 论文结构

本文分为以下章节：

第一章是“绪论”，介绍了本课题的研究背景和意义，情感分析领域的研究现状，本文的主要研究内容。

第二章是“相关技术”，介绍了评价对象抽取、评价词抽取和情感倾向性分析的各种解决方法和各自的特点。

第三章是“基于条件随机场的评价对象抽取”，介绍了条件随机场的背景以及如何使用条件随机场抽取评价对象。

第四章是“基于依存句法关系的评价词抽取”，介绍了依存句法关系以及如何使用依存句法关系抽取评价词。最后用实验对本文提出的情感要素抽取算法进行了验证。

第五章是“基于评价词典的情感倾向性分析”，介绍了如何制作评价词典以及如何利用评价词典分析情感要素的情感倾向。最后用实验对本文提出的情感倾向性分析算法进行了验证。

第七章是“可视化系统设计与实现”，介绍了可视化系统的设计方案、实现方法和最终效果。

# 相关技术

本章将介绍情感分析中常用到相关技术，主要分为预处理、评价对象抽取、评价词抽取和情感倾向分析这几个方面。

## 预处理

在计算机领域中，预处理指的是把原始数据数据转换为后续程序的输入数据的过程。文本预处理又被称作文本标准化，它的作用是把原始文本转换为便于后续分析操作的文本，它在文本分析过程中是非常重要的基础。

文本预处理的操作从形式上一般包括删除、纠正、拆分、合并、补全、标注等。不同的分析任务对预处理的步骤会提出不同的要求，比如大多数任务要求的文本预处理步骤有分词、词性标注、去停用词等等，而有些更复杂的任务可能要求句法分析。对于不同语种的文本也有不同的预处理操作。比如英文文本可能需要做词干提取，也被称为词形还原。因为英文单词经常带有前缀和后缀来表示时态的变化，就像“looks，looking，looked”等词有同样的词干“look”。一般而言词干提取是有利的，但是由于这是一个较为复杂的处理过程，所以也会有特例，比如“department”和“depart”表示的含义是完全不同的。再例如英文文本的预处理还有大写字母转换为小写字母，但是这同样也面临一些窘境，比如“US”表示美国，而“us”表示我们，类似这样的情况是比较棘手的。中文文本的预处理也有自己的难处，比如分词的准确性有时难以保证，一旦出现偏差会直接影响后续的所有处理和分析。所以预处理是一个复杂的，必需的，需要谨慎运用的技术。

目前有不少开源的自然语言处理工具都能完成预处理任务。比较常用的有Stanford系列、OpenNLP、FudanNLP、NLTK等自然语言处理工具包。针对不同的分析场景，会选用不同的工具包，比如Stanford系列的工具包常用于英文文本的处理，而FudanNLP常用于中文文本的处理。为了适应越来越复杂和多样的文本分析需求，许多工具包提供了灵活的自定义功能，让使用者根据实际需要调整处理的效果。这些工具的出现大大减少了相关领域开发人员的工作量，本文的开发也用到了其中的处理工具。下面逐一说明本文所运用的文本预处理技术。

### 分词

分词指的是把连续的字符序列切分为若干个有合理意义的词语单元的过程。在英文和其他使用拉丁语字母表的语言中，空格是分隔单词极佳的标志。然而，在有些语言中并没有这种发挥同等做用的符号，比如中文，那么分词就是一个非常重要但是困难的任务。分词的本质是决定一个词的边界，即它从哪里开始到哪里结束。

下面是几个常见中文分词工具在分词方面的评测结果对比，来自Github的Java开源项目cws\_evaluation。

表 2.1 常见中文分词器对比

|  |  |  |  |
| --- | --- | --- | --- |
| 分词器 | 分词速度（字符/毫秒） | 行数完美率 | 字数完美率 |
| word 最大Ngram分值算法 | 370.9714 | 66.55% | 60.94% |
| word最少词数算法 | 330.1586 | 65.67% | 60.12% |
| HanLP标准分词 | 935.7724 | 58.31% | 50.43% |
| Stanford Chines Treebank segmentation | 13.7233 | 58.29% | 51.36% |
| Ansj基本分词 | 1295.5205 | 55.36% | 48.18% |
| Smartcn | 611.1504 | 55.29% | 48.03% |
| HanLP 极速词典分词 | 6015.3677 | 54.25% | 46.12% |
| FudanNLP | 123.4570 | 51.48% | 43.22% |
| Jieba SEARCH | 993.435 | 50.84% | 41.54% |
| Jcseg复杂模式 | 561.55975 | 47.96% | 38.84% |
| MMSeg4j ComplexSeg | 1699.5801 | 38.81% | 29.60% |
| IKAnalyzer智能切分 | 319.2809 | 37.55% | 27.97% |

上面的表格是按照行数完美率进行排行的，表现最好的是word分词器，HanLP和Stanford分词器也表现不错；而从分词速度来看，HanLP标准分词明显优于另外两个分词器。而且HanLP分词器支持用户自定义词库，能满足不同领域专有词的补充或分析任务特殊的要求，所以本文采用了HanLP分词器。

HanLP提供了很多不同算法实现的分词器，本文使用的是标准分词器。下表列出了常被使用的分词器和各自的特点。

表 2.2 Hanlp常用分词器

|  |  |
| --- | --- |
| 分词器 | 特点 |
| 标准分词 | 基于隐马尔科夫模型实现的维特比算法，有中外人名识别功能 |
| NLP分词 | 会执行全部命名实体识别，故速度略慢，而且会出现误识别 |
| 索引分词 | 主要面向搜索引擎，能够全切分长词，还能得到单词在文本中的偏移量 |
| 繁体分词 | 能对繁体中文进行分词 |
| 极速词典分词 | 速度非常快，但准确率一般 |

### 词性标注

词性标注（Part-of-Speech tagging 或POS tagging)，又称词类标注或者简称标注，是指为分词结果中的每个单词标注一个正确的词性，即确定每个词是名词、动词、形容词或其他词性的过程。决定一个词的词性基于词的定义和上下文。比如词语和邻近词语或在短语、句子或语段中的关联词语的关系。

词性标注不是简单地罗列出所有的词和对应的词性，因为很多词在不同的语境下有不同的词性，甚至有些词性很复杂或者难以明确。词性标注工具比如 “这本书的封面设计很别致”和“小张设计了一套管理系统”中“设计”的词性分别为名词和动词，有时词性标注工具不能正确地区分。

现在主流的词性标注算法有Viterbi算法，Brill标注器、约束文法等。一些机器学习方法也被应用到词性标注中，比如支持向量机、最大熵分类器、感知器算法、最近邻算法等，大多数的方法能达到95%以上的准确率。虽然词性标注和分词是两个不同的概念，但是这两个过程一般来说是一起完成的。

### 去停用词

停用词是在处理自然语言数据之前或之后被过滤掉的词语。一般而言去停用词在正式文本分析之前完成。停用词常指的是那些在某个语言中经常出现的，对分析没有帮助甚至起到阻碍作用的词语，这些词收集在一起形成了停用词表。但是没有一个统一的停用词表，根据不同的分析目的，不同的词会被认为是停用词，因此停用词表也有所不同。一般来说，去停用词能减少文本中的干扰项，更便于后续的分析。

## 条件随机场

条件随机场（conditional random fields，简称 CRF，或CRFs），是2001年由Lafferty等人在隐马尔科夫模型和最大熵模型的基础上创造的一种无向图学习模型，是一种用于标注和切分有序数据的条件概率模型。它能完成结构化的预测，因此经常被用于模式识别和机器学习领域。CRFs有很强的推理能力，能学习复杂、非独立、有重叠的特征，能较好地考虑上下文信息。而且CRFs能任意加入其它外部特征，这意味着它能够获取非常丰富的信息，这些特点使得它经常被用于自然语言文本、生物序列的标注任务。

序列标注是机器学习领域的一种模式识别任务，它的具体任务是把某个类别标签赋值给观察值序列。可以把序列标注问题看作分类问题，把序列中的每个单位分到一个标签类别。利用上下文信息达到的全局最优状态能够提高分类的准确率，所以CRFs模型能很好的完成序列标注问题。评价对象抽取也能看作序列标注问题，即给每个词标注其是否是评价对象，所以采用CRFs模型是很好的选择。

## 评价对象的抽取

评价对象是指文本中讨论的主题，一般为名词和名词短语。目前评价对象抽取的研究主要针对产品评论，在产品评论中评价对象一般是产品的特性。

Hu和Liu[[4](#_ENREF_4), [5](#_ENREF_5)]是进行面向网络评价的评价对象抽取研究的早期人员，他们使用基于Aprior算法的联合规则[[6](#_ENREF_6)]提取频繁出现的名词或名词短语作为候选评价对象；然后利用紧密度剪枝和冗余剪枝技术选出真正的评价对象。紧密度剪枝用于过滤没有以固定词序出现的名词短语；冗余剪枝用于过滤经常以与其他词组成短语的形式而非单个词的形式出现的单个名词。有的研究者使用模板来抽取评价对象[[7](#_ENREF_7)]，但这种方法可扩展性差，需要耗费大量人力。Kim和Hovy[[8](#_ENREF_8)]采用语义角色标注的方法进行抽取。Brody和Elhadad[[9](#_ENREF_9)]使用基于LDA（Latent Dirichlet allocation）的非监督方法抽取评价对象。

常见的监督策略是使用HMM（Hidden Markov Model，隐马尔科夫模型）或者CRF（Conditional Random Field，条件随机场）。Jin和Ho[[10](#_ENREF_10)]应用了HMM模型抽取评价对象，在训练模型时使用词特征和词性。词特征是语料进行分词后的词汇本身。CRF模型把评价对象提取的问题转换为序列标注的问题，Jakob和Gurevych[[11](#_ENREF_11)]使用下面这些特征来训练CRF模型：词本身，词性，句法依存关系，词距和观点句。Toh和Wang[[12](#_ENREF_12)]也使用CRF提取评价对象，他们的算法在2014年的SemiEval测评中的多个测试集上均取得了前两名的成绩。他们使用的特征是命名实体识别中常用的特征、当前词特征、当前词的词性、中心词、中心词的极性、当前词的依赖关系和命名清单。

## 评价词的抽取

评价词是那些修饰评价对象的有情感色彩的词，一般为形容词或动词。评价词是分析情感倾向性的关键。[[4](#_ENREF_4), [5](#_ENREF_5)]使用句法分析抽取评价词，在已经确定评价对象的基础上，找到句法路径离评价对象最短的形容词作为评价词。这种方法的抽取准确性非常依赖于句法分析的准确性，而网络文本的表述不一定规范，这会导致句法分析不准确。为了解决上面的问题，Liu等人[[13](#_ENREF_13)]把句法关系和词对齐[[14](#_ENREF_14), [15](#_ENREF_15)]结合起来。仅使用句法关系中准确率较高的几种关系，再利用词对齐策略找到评价对象和评价词之间的对应关系。Turney和Littman[[16](#_ENREF_16)]定义了种子情感词集合，并计算点互信息以判断评价词。

Liu[[17](#_ENREF_17)]利用图论，提出Graph Co-ranking策略来提取评价对象和评价词。他们认为传统的思路只考虑了评价对象和评价词之间的关系，即观点关系；在文中他们另外考虑了评价对象之间和评价词之间的关系，这种关系被定义为语义关系。他们把这些关系用图的方式表示出来，再用随机游走算法计算候选词的可信度分值，当分值大于阈值时，候选词就被认为是真正的评价对象或者评价词。

## 情感倾向性分析

根据粒度的不同，常见的情感倾向性分析有篇章级、句子级和短语级。面向产品评论的情感分析一般为句子级和短语级。一种解决方案是采用基于词典的方式，英文文本常借助WordNet，而中文文本常使用HowNet。常见的策略如Hu等人[[4](#_ENREF_4), [5](#_ENREF_5)]使用WordNet中的同义词集和反义词集对已知倾向性的种子词进行扩展，得到大量形容词的情感极性；再利用评价词的倾向性确定情感要素的情感倾向性。也有不完全依赖公开词典的方法，比如Hatzivassiloglou和 McKeown[[3](#_ENREF_3)]利用对数线性回归模型和形容词之间的关系来预测相关联的形容词之间是相同或者相反情感倾向的。再利用聚类算法判断形容词是正向的还是负向的情感。另一种解决方案是利用机器学习的方式。Pang等人[[18](#_ENREF_18)]最早尝试使用机器学习的策略来判断文档的情感倾向，他使用了朴素贝叶斯、最大熵和支持向量机这三种机器学习方法，效果明显好于人工制作的系统，但不如传统的基于话题进行分类的方法。Lu等人[[19](#_ENREF_19)]则利用分类器对评价对象的上下文做分类，通过上下文的情感倾向确定评价对象的情感倾向。还有的研究者主要集中研究特征和分类器的选取以更好的判断情感倾向[[20](#_ENREF_20), [21](#_ENREF_21)]。当然也有一些另辟蹊径的方式，比如Lu等人[[22](#_ENREF_22)]把情感倾向性分析转换为最优化问题。

# 基于条件随机场的评价对象抽取

情感分析一直以来在工业领域和学术领域都受到十分的关注，在实际应用中，情感分析常作为重要的模块被嵌入到其他系统中。篇章级和句子级的情感分析已经有相当程度的研究了，其性能已经能让其在大多数应用场景中立足。还有一些任务要求进行更精确、粒度更小的分析，可能是对词语、短语或者某个成分进行分析。比如在推荐系统中，应该只推送用户容易感兴趣的商品。在问答系统中，会有相对具体的提问，“鱼的哪些成分对人体有益？”。在情感信息归纳中，要给出给定实体具体的褒义和贬义观点。在这些分析任务中，找出情感文本的观点和评价对象是关键。

评价对象在句子级别的情感分析中，通常指被评论的对象本身，以及被评论对象所具有的某种属性或其组成部分。例如在电影评论中，评价对象可以是影片本身，以及画面、音乐、演员等构成影片的一些要素及其属性。对于微博文本而言，评价对象可以是博主评论的事件、人物等。在产品评论文本中，评价对象一般指的是产品的参数、配置等属性。比如有产品评论“小米手机的屏幕比较大，续航能力不错，但是外观不够好看”。这个句子的评价实体是“小米手机”，评价倾向是褒义的，评价对象是小米手机的若干属性，包括“屏幕”、“续航能力”和“外观”。评价对象抽取子课题的目的就是在产品评论中找出所评论的产品属性。

本文使用条件随机场模型抽取评价对象，下面介绍具体抽取方法。

## 条件随机场的数学定义

设是一个无向图，是以G中节点为索引的随机变量构成的集合。在给定X的条件下，如果每个随机变量服从马尔科夫属性即。u~v表示u和v是相邻的边，则（X,Y）就构成了一个条件随机场。

CRFs是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率，即求解条件分布。而不是在给定当前状态条件下，定义下一个状态的隐马尔科夫模型分布，即求联合分布。



图 3.1 P(Y|O)图示

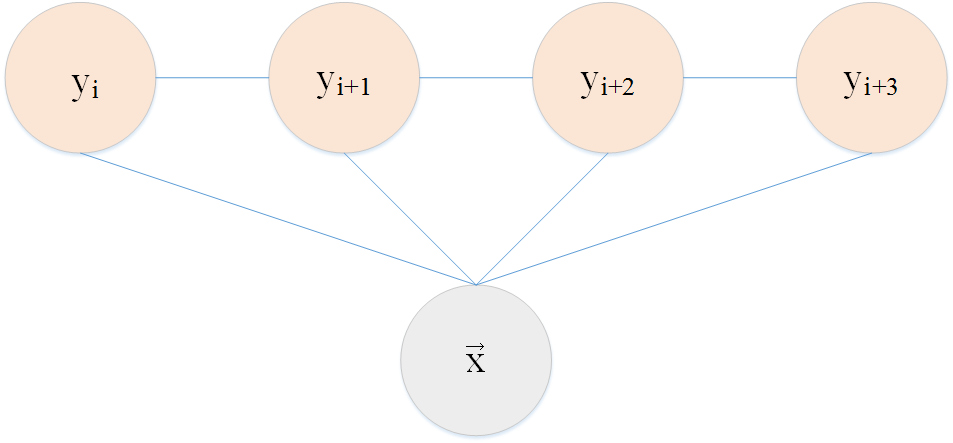


图 3.2 P(Y,O)图示

线性链条件随机场模型（Linear-chain CRFs）是常用的CRFs模型。令表示观察序列，是有限状态的集合，根据随机场的基本理论：

其中是对于观察序列的标记位置i-1与i之间的转移特征函数；是观察序列的i位置的状态特征函数。将两个特征函数统一为，则有：

## 条件随机场的关键问题

### 特征函数的选择

特征函数的选取直接关系模型的性能。CRFs模型中特征函数的形式定义是，它是转移特征函数和状态特征函数的统一形式的表示方法。一般情况下特征函数是二值函数，即取值为0或者1。在定义特征函数的时候，应该首先构造观察序列的实数值特征集合来描述训练数据的经验分布特征。例如：

每个特征函数被表示为观察序列的实数值特征集合中的一个元素，如果当前状态和前一个状态有某个特定的值，那么所有特征函数都是实数值。

### 参数估计

构建CRFs模型的主要任务是从训练数据中估算特征的权重。假设给定训练集，采用极大似然估计法估算参数。条件概率的对数似然函数为：

其中，是训练样本的经验概率，是随机变量在训练样本中的经验分布。它们的计算公式如下：

分别对对数似然函数中的求导，得到：

令上面的式子为0，求解。上面的方法直接使用对数最大似然估计，可能会产生过度学习的问题，所以通常会引入惩罚函数来解决。

使用惩罚项之后，对数似然函数变为：

对上述公式中每个求导数，再令结果为0，求解。但是极大似然估计不一定会得到近似解，所以需要利用迭代技术来选择参数，是的对数似然函数最大。Lafferty提出了两个迭代缩放算法：GIS（Generalized Iterative Scaling）算法，IIS（Improved Iterative Scaling）算法。

### 模型推断

CRFs做的推断可以表述为，对于一个给定观察序列，求一个标记序列使得该观察序列出现概率最大。

对于一个练市条件随机场，在图的模型中添加一个开始状态和一个结束状态。定义一组矩阵，其中是的随机变量矩阵。中的元素如下定义：

由于是从开始节点到结束节点的某条路径的概率，所以，

其中，是归一化因子，是所有路径的概率之和，

## 评价对象的识别

### 标记方法

本文把评价对象抽取问题转换为一个序列标注问题。给定一句话，这个句子就组成了一个序列，为每个词打上一个标记，就会得到标记序列；实际上这也是一个分类问题，即把每个词分类到某个标记。

为了标注某个词是否是评价对象，本文采用BIO标记法，即标签集合是。B代表当前词是评价对象的开头，I代表当前词是评价对象的内部成分，O表示当前词不属于评价对象。如下例所示：



图 3.3 BIO标记示例

### 学习特征

本文使用条件随机场模型抽取评价对象。本文对每个词语提取如下特征用于学习和标记：

1. 当前词对应的字符串；
2. 前一个词对应的字符串；
3. 后一个词对应的字符串；
4. 当前词的词性；
5. 当前词的中心词。中心词是现代汉语语法知识中的一个术语，它也被称为定位词、中心语。中心词出现在偏正短语中，是被修饰语修饰的中心成分。在包含多层状语或定语的偏正短语中，每层状语、定语所修饰的都是中心词。比如：



图 3.4 中心词示例图

这句话要表达的核心含义是“设计新颖”，所以“新颖”是修饰语，“设计”是形容词“新颖”的中心词。“屏幕”是限定“按键”的，所以中心词是“按键”。“按键”是修饰“设计”的，所以中心词是“设计”。而“设计”不修饰任何词，所以没有中心词。从这个例子可以看出来，中心词能较好地体现词与词之间的修饰关系，把上下文信息加入到了条件随机场的学习过程中，对评价对象的抽取有重要作用。

本文抽取的评价对象

1. 中心词的词性；

## 条件随机场实现工具CRF++

常见的条件随机场实现工具有CRF++和CRFSuite，本文采用CRF++完成条件随机场的训练和标注。CRF++是一个简便、开源、支持用户定制的条件随机场实现工具，能完成分词、序列标注任务。在自然语言处理领域的应用有命名实体识别、信息抽取等等。如下图所示是CRF++的工具包。

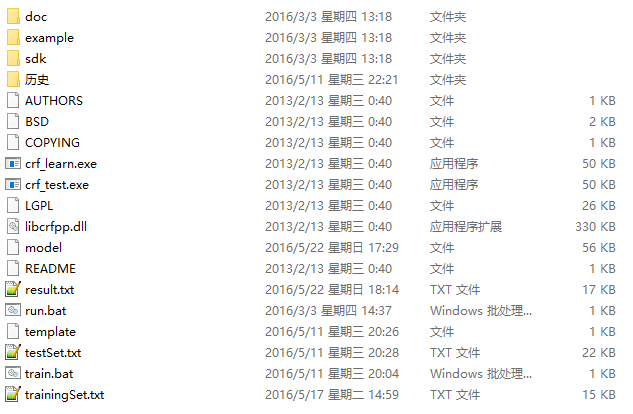


图 3.5 工具包文件

### CRF++文件格式

CRF++使用过程中需要人工配置的文件有三个：训练集、测试集、模板。训练集的内容如下图：

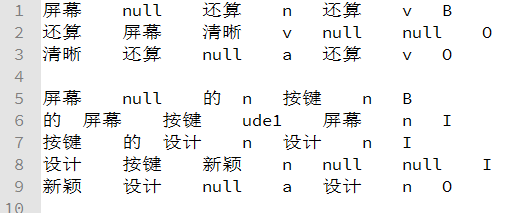


图 3.6 训练集示例图

每个训练样例的原始文本，即一条产品评论，会根据分词后的结果在第一列纵向排列。每行的若干元素是以第一列为当前词提取的特征。最后一列是人工标注的结果。以“屏幕还算清晰”的“屏幕”为例，前一个词没有所以为“null”，后一个词是“还算”，词性是名词所以第四列是“n”，中心词是“还算”，中心词的词性是动词所以是“v”，标注结果是评价对象的开头，所以标注为“B”。

测试集和训练集的格式基本一样，不同在于最后一列是可选的，即可以加上人工标注结果，也可以不加人工标注结果。模板的内容如下图所示：

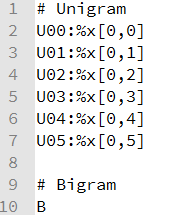


图 3.7 模板内容图

CRF++会根据模板在训练集中寻找当前词的训练特征。%x[a, b]是模板的基本格式，a表示相对于当前词的纵向偏移，b表示列的编号（从0开始计算）。假设当前词是图 3.6中的“屏幕还算清晰”的“还算”，则%x[1,3]提取的特征是“还算”所在行的后一行，即“清晰”行的列号为3的特征，也就是“a”。

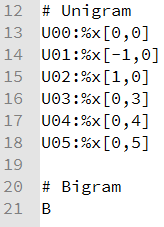


图 3.8 等价模板内容图

使用本文中的训练集，图 3.7和图 3.8所示的模板实际是等价的，因为训练集中的行已经包含了前一个词对应的字符串和后一个词对应的字符串这两个特征。

### CRF++使用方法

CRF++的使用很简单，分为训练模型和使用模型两部分。

训练模型需要模板（template）和训练集，在命令行进入到CRF++所在的目录下，输入命令：crf\_learn template trainingSet.txt model，就能生成训练好的模型。“template”是模板的文件名，“trainingSet.txt”是训练集的文件名，“model”是生成的模型文件。下图是模型训练成功时的效果。

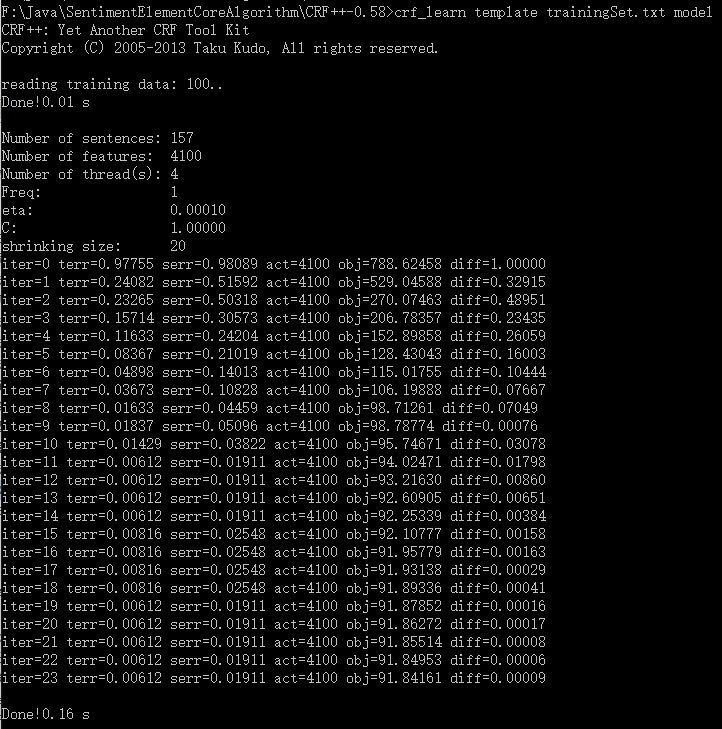


图 3.9 成功训练模型效果图

使用模型需要模型（model）和测试集，命令行指令是：crf\_test -m model testSet.txt >> result.txt。 “testSet.txt”是测试集的文件名，“model”是生成的模型文件，“result.txt”是写入结果的文件。下图是序列标注的结果示例。

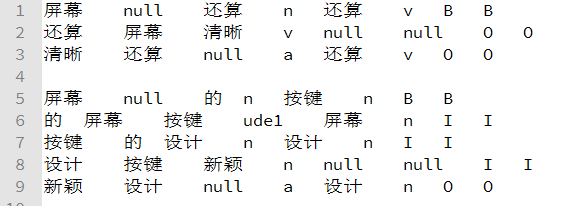


图 3.10 标注结果图

其中，最后一列是程序标注的结果，倒数第二列是人工标注的正确结果。

## 训练集

训练集包含120个正样本，80个负样本。通过观察实际的产品评论语料，发现评论产品属性的句式主要是下面这些，本文采用的正样本覆盖了这些句式。

表 3.1 评论句式及示例

|  |  |
| --- | --- |
| 句式 | 示例 |
| 评价对象 + 程度副词 + 评价词 | 屏幕很大 |
| 评价对象 + 否定词 + 评价词 | 电池不耐用 |
| 评价词 + 评价对象 | 大屏幕 |
| 评价对象 + 有/没有 + 评价词 | 外观没有特色 |
| 评价对象 + 是/不是 + 评价词 | 屏幕是高清的 |
| 评价对象A + 连词 + 评价对象B + 评价词 | 屏幕和电池都很好 |
| 评价对象A + 评价词 + 连词 + 评价对象B + 评价词 | 屏幕很大而且电池很耐用 |

## 名词剪枝

利用条件随机场得到的评价对象可能不一定属于本领域的，比如在手机领域的评论中有这样一条评论“这道菜的味道有点咸”，那么前文所述的算法会把“这道菜的味道”当做评价对象，但是它不是属于手机领域的，所以不属于特定领域的评价对象应该被去掉。为了解决这个问题，本文采用两种名词剪枝技术，词频剪枝和PMI剪枝。

### 词频剪枝

词频剪枝就是过滤掉在语料库中出现频率较低的评价对象。因为评价对象是名词，所以先统计语料库中所有名词出现的频率，并且选出其中频率最多的前N个名词。如果抽取出的评价对象不包含这N个名词中的任何一个，则去掉这个评价对象。

本文对5万条手机评论进行统计，选择了前1000个名词，并且手动筛选了这1000个词。在这些词中有噪声，即有些词并不是手机的属性，也就是不会是评价对象，那么这些词也被人工剔除了，最后只剩下400个左右的名词。排在1000以后的名词的频率已经很低了，第1001个词的词频是57，而且频率越低的名词越可能是噪声。所以选出的400个名词能覆盖绝大部分的手机属性，即使遗漏了一些低频率的手机属性，由于频率很低，本文认为这样的属性不是消费者主要关注的产品属性，所以遗漏了也能接受。

### PMI剪枝

PMI（Pointwise Mutual Information）值描述的是词与词之间的关系，在给定的语料库中，词a和词b的PMI值可用如下公式求得：

其中，表示的是同时包含a和b的文本数量，表示包含a的文本数量，表示包含b的文本数量。PMI值基于的假设是，两个词共现的次数越多，那么这两个词之间的联系越大。为了用PMI值描述某个词的领域相关性，需要计算这个词和领域特征词（集）的PMI值，领域特征词（集）指的是能代表某个领域的词（集）。本文使用特征词集，包含32个特征词。计算评价对象中的名词和该词集的PMI值，由于PMI值很小，所以乘上105，去掉值小于2.2的评价对象。

# 基于依存句法关系的评价词抽取

评价词是描述评价对象的词语。在产品评论中，评价词能描述产品的属性。比如“屏幕很大”中“大”是评价词，它描述了“屏幕”的大小。本文中研究的评价词为形容词或者是“有”、“没有”、“是”、“不是”这些词的直接宾语。本文利用依存句法关系抽取评价词。

## 依存句法关系

依存语法也被称作从属关系语法，是用词和词之间的依存关系来描述文本的语言结构的一种语法框架。现代依存语法理论是由发过的语言学家Lucien Tesnière创立的。在传统的语言学中，主语在一个句子中的地位是最主导的，而其他成分被统称为谓语。而依存语法认为，动词在句子中起主导作用，其他成分直接或间接地和其关联。在依存语法的理论中，“依存”指的是句子中词和词之间的从属关系。根据支配和从属关系，成分被认定为支配者或被支配者，后者也被称为从属者。依存句法结构可以用依存句法树表示。支配者为父节点，而被支配的成分为该父节点的子节点。比如“这个手机的屏幕很大”的依存句法树如下图：



图 4.1 依存句法树

本文关注的依存句法关系是评价对象和评价词之间的关系，比如这个例句中“屏幕”和“大”之间的依存关系是nsubj。

## 评价词抽取

### 正常抽取

本文在第一部分得到了评价对象，然后利用某些特定的依存关系，找到和评价对象有关联的词语，这个词语就被认为是评价词。本文使用Stanford Parser做句法分析。下表列出了本文用到的依存关系：

表 4.1 依存关系表

|  |  |  |
| --- | --- | --- |
| 依存关系 | 说明 | 示例（支配词，从属词） |
| amod | 形容词修饰 | （屏幕，大） |
| rcmod | 关系从句修饰 | （问题，遇到） |
| nsubj | 名词性主语 | （漂亮，外观） |
| assmod | 关联修饰 | （屏幕，手机） |
| ccomp | 从句补语 | （喜欢，屏幕） |
| top | 主题 | （有，屏幕） |
| dep | 依赖关系 | （漂亮，颜色） |

利用这些关系找到的词是候选评价词，从这些词中选出词性为形容词的词作为最后的评价词。

### 特殊抽取一

有几个词要特殊处理，这些词是“有”、“没有”、“是”和“不是”。其中“有”和“没有”是一类的，而“是”和“不是”是一类。

1. 比如“外观有特点”，这句话的依存关系有top(有, 外观)，root(ROOT, 有)，dobj(有, 特点)。准确的评价词应该是“有特点”，所以当出现“有”和“没有”的时候，要另外考虑dobj直接宾语关系。
2. 比如“屏幕是高清的”，这句话的依存关系有top(是, 屏幕), root(ROOT, 是), attr(是, 高清), cpm(高清, 的)。准确的评价词应该是“是高清的”，所以当出现“是”和“不是”的时候，要另外考虑attr属性关系。

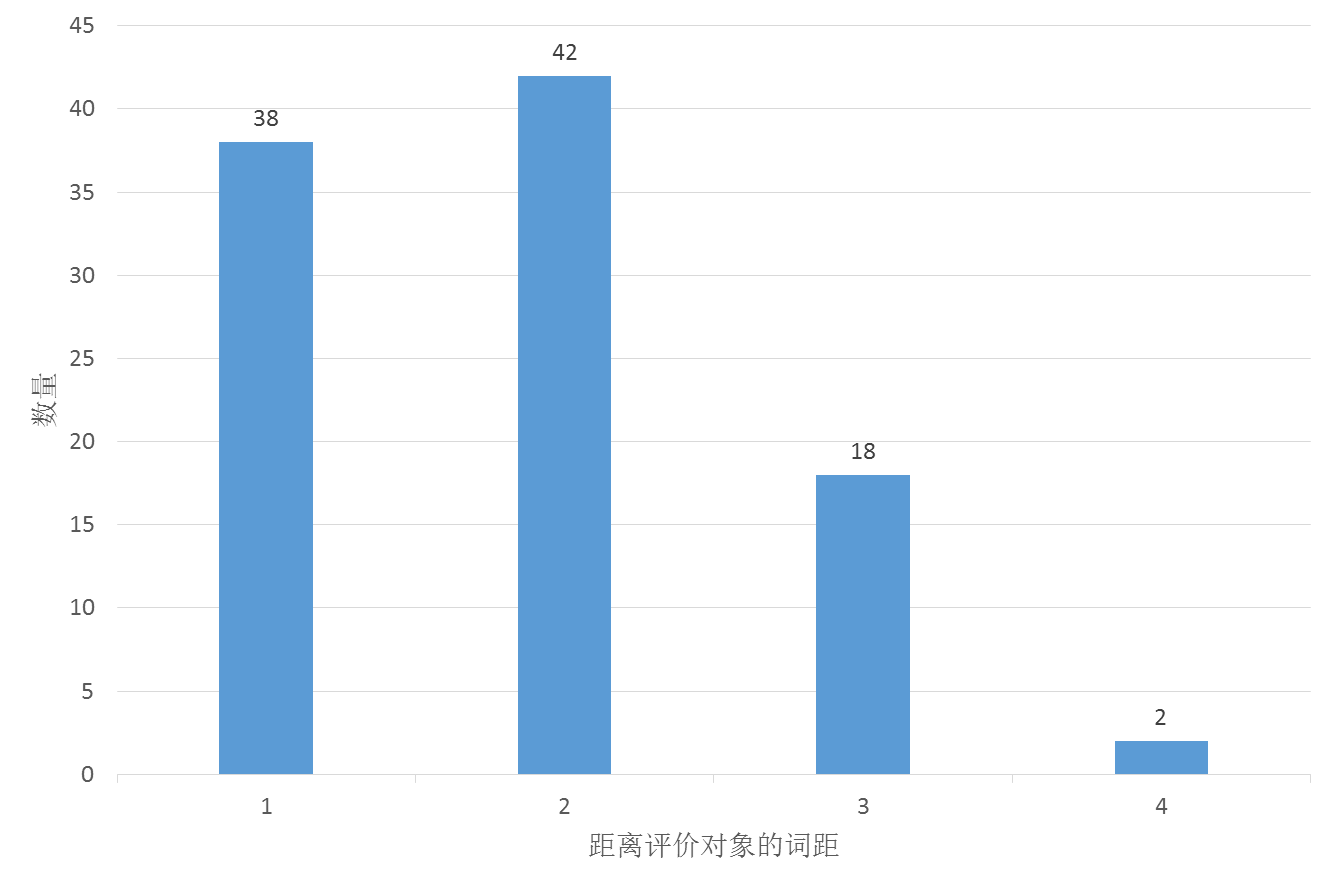
### 特殊抽取二

但是有的句子比较复杂，句法分析器可能无法正确分析出依存关系。比如“我觉得这屏幕似乎有点小”这句话的依存关系有nsubj(觉得, 我)，root(ROOT, 觉得)， advmod(小, 似乎)，advmod(小, 有点)，ccomp(觉得, 小)。并没有“屏幕”和“小”之间的依存关系，为了解决这个问题，本文采用的策略是：如果关于评价对象的依存关系中没有表 4.1所列出的依存关系，就选择距离评价对象3以内的形容词作为评价词。这个句子的各成分到评价对象“屏幕”的词距如下图所示：



图 4.2 词距图

通过这种策略，就能找到评价词“小”。之所以采用这种方法，是因为统计了100条同时有评价对象和形容词的评论，有如下结果：

纵轴表示的是评论数量，横轴表示的是形容词距离评价对象的词距。可以看出来形容词距离评价词的距离主要集中在小于等于3的范围内。在98个词距在3以内的评论中，几乎所有的形容词都是评价词。所以距离评价对象3以内的形容词有很大概率是评价词。

### 特殊抽取三

有些评论中出现了用连词连接两个评价对象或者两个子句的情况。大多数情况下很难正确对这样的句子做句法分析。比如“外形很漂亮而且颜色好看”对应的依存关系有nsubj(漂亮, 外形)，advmod(漂亮, 很)，root(ROOT, 漂亮)，conj(颜色, 漂亮)，cc(漂亮, 而且)，dep(漂亮, 颜色)，dobj(颜色, 好看)。利用这个依存关系抽取出来的情感要素是（外形，漂亮）和（颜色，漂亮）。但是“颜色”的评价词是“好看”，所以出现了评价词抽取错误的问题。为了解决这个问题，本文按照如下规则进行抽取：

1. 判断句子中评价对象的数量和是否有连接词，如果有两个评价对象而且出现了连接词，则进行步骤2；
2. 判断连接词连接的是两个评价对象还是两个子句，如果是两个评价对象，则进行步骤3；如果是两个子句，则进行步骤4；
3. 句式可以表示为（评价对象A，连词，评价对象B，其他成分），把该句子拆分为两个子句（评价对象A，其他成分）和（评价对象B，其他成分），进行步骤5；
4. 句式可以表示为（评价对象A，其他成分A，连词，评价对象B，其他成分B），把该句子拆分为两个子句（评价对象A，其他成分A）和（评价对象B，其他成分B），进行步骤5；
5. 对两个子句分别提取评价词。

## 实验

在上一章和本章先后介绍了评价对象抽取和评价词抽取的算法，即得到了抽取情感要素的算法。本节将测试情感要素抽取的效果。

### 实验数据集

本文实验使用了从数据堂获取的手机评论语料中的一部分，两个测试用的是同一批数据。如下表所示：

表 6.1 数据集

|  |  |  |  |
| --- | --- | --- | --- |
| 领域 | 训练 | 测试 | 合计 |
| 手机 | 157 | 300 | 457 |

实验数据集的标注结果用XML文件记录，格式样例如下：

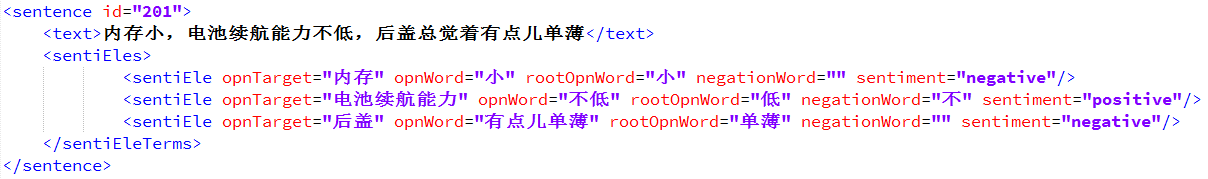


图 6.1 数据集样例

其中opnTarget是评价对象，opnWord是带有修饰词的评价词，rootOpnWord是没有任何修饰词的评价词，isNegation是是否有否定修饰词。

### 评判指标

本文选用的指标是准确率P、召回率R和F1值，它们的计算方式如下：

因为有些程度修饰词会被当作停用词剔除掉，导致opnWord不完全一样，但是抽取结果和标注结果在语义上是一致的。而本文关注的是语义的准确性，即抽取得到的情感要素能正确表达产品的属性。所以本文认为情感要素抽取正确的标准是：opnWord、rootOpnWord、negationWord的抽取结果和标注结果完全一致。比如人工标注结果是“后盖不怎么单薄”，因为“怎么”是停用词，所以抽取结果是“后盖不单薄”；rootOpnWord是“单薄”，negationWord是“不”，这两个参数的抽取结果和标注结果相同，所以这两个表述的语义基本一致，因此抽取结果是可接受的。

而情感倾向的正确标准是sentiment结果完全一样。

### 实验结果及分析

在手机领域的数据集上测试的结果如下：

表 6.2 完整算法的实验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 测试名称 | 标记数量 | 抽取结果数量 | 正确标记数 | 正确率 | 召回率 | F1值 |
| 情感要素抽取测试 | 820 | 669 | 551 | 0.8236 | 0.8158 | 0.8196 |
| 情感要素抽取测试（不用名词剪枝） | 820 | 669 | 551 | 0.8236 | 0.8158 | 0.8196 |

从实验结果可以看到，本文采用的方法能有效地完成情感要素抽取和情感倾向判断任务。可以看到本文算法能使F1值达到0.8以上，使用名词剪枝技术能改善抽取效果。

# 基于评价词典的情感倾向性分析

情感分析分为篇章级、句子级、短语级、词语级的分析，本文属于短语或词语级的情感分析。本文的情感分析对象是情感要素，即从产品评论中抽取出来的评价对象和评价词组成的二元组。情感要素的情感倾向分为正向、负向和中立。比如情感要素（屏幕，很大）的情感倾向是正向，（续航能力，差）的情感倾向是负向，（外观，一般）的情感倾向是中立。本文使用评价词典来判断情感要素的情感倾向。具体操作是在评价词典中查找评价词的对应的情感倾向，以此决定情感要素的情感倾向。本文使用的评价词典是在HowNet的基础上整理而成的。

## HowNet

HowNet的中文名称是知网，是一个以英语和汉语词语的概念为描述对象，以描述概念之间、概念的属性之间的关系为内容的常识知识库。本文使用了HowNet的情感词典，HowNet提供的情感词典如下图所示，本文使用了其中标红的两个词典。



图 5.1 HowNet情感词典

## 情感倾向的判断

本文在HowNet提供的评价词典的基础上进行筛选，最后形成的词典如下表所示：

表 5.1 整理后的评价词典

|  |  |
| --- | --- |
| 词典 | 词典大小 |
| 上下文无关正向评价词典 | 1288 |
| 上下文无关负向评价词典 | 888 |
| 上下文无关中立评价词典 | 15 |
| 上下文相关评价词典 | 26 |

### 上下文无关评价词

对于上下文无关的评价词，如果没有否定修饰词，则情感要素的情感倾向和评价词的情感倾向一致，如（外形，炫酷）是正向的，因为“炫酷”是正向评价词；如果有否定修饰词，则情感要素的情感倾向和评价词的情感倾向相反，如（样子，不时尚）是负向的，因为“时尚”是正向的，“不”是否定修饰词。

### 上下文相关评价词

而对于上下文相关的评价词，需要结合评价对象来决定情感要素的情感倾向。比如（屏幕，大）是正向的，而（噪声，大）是负向的。为了决定上下文相关的情感要素的情感倾向，本文把评价词分为增强类评价词和削弱类评价词。如“大”和“高”等词是增强类评价词，而“小”和“低”是削弱类评价词。把评价对象也分为正向类评价对象和负向类评价对象。如“屏幕”和“音量”是正向评价对象，而“价格”和“噪声”是负向评价对象。再根据如下规则判断情感要素的情感倾向。

表 5.2 上下文相关情感要素的情感判断规则

|  |  |
| --- | --- |
| 评价对象和评价词类别组合 | 情感要素的情感倾向 |
| 正向 + 增强 | 正向 |
| 正向 + 削弱 | 负向 |
| 负向 + 增强 | 负向 |
| 负向 + 削弱 | 正向 |

## 实验

本节将测试情感要素的情感倾向性分析的效果。由于在上一章已经介绍了数据集、评判标准，本节不再赘述。下表是实验结果：

表 6.2实验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 测试名称 | 标记数量 | 抽取结果数量 | 正确标记数 | 正确率 | 召回率 | F1值 |
| 情感倾向性测试 | 820 | 634 | 617 | 0.9731 | 0.7524 | 0.8486 |
| 情感倾向性测试（不用名词剪枝） | 820 | 634 | 617 | 0.9731 | 0.7524 | 0.8486 |

从实验结果可以看出来，情感倾向性分析的正确率较高，达到0.9以上，而召回率相对较低。这个结果和本文采用的方法有很大关系，查词典的方式能够很好地保证正确得到评价词的情感倾向，但是由于采用的是穷举法，词典的覆盖率不一定很高，有些词没有出现在词典中就会导致召回率的降低。

# 可视化系统设计与实现

本文实现了一个面向产品评论的情感要素抽取及情感倾向性分析演示系统。它能用于挖掘用户对产品的真实观点。

## 系统设计

根据对消费者产生的产品评论的信息挖掘的需要，本文设计了一个针对产品评论的情感要素抽取和情感倾向性分析演示系统。系统以用户输入的产品评论作为分析对象，实现了文本预处理、情感要素抽取、情感倾向性分析之后，将分析结果显示在用户界面上。系统的处理流程如下图所示。



图 7.1 系统流程图

## 开发环境

系统的开发和运行环境如下表所示：

|  |  |
| --- | --- |
| 参数 | 配置 |
| 操作系统 | Windows 10 |
| CPU | Intel(R) Core(TM) i5-3210 CPU @ 2.50GHz |
| 操作系统位数 | 64 |
| 内存 | 8GB |
| 编程语言 | Java（1.7.0\_80） |

## 系统实现

### 类介绍

如下图描述的是系统主要的类和类之间的关系。



图 7.2 系统类图

其中UserInterface类负责和用户交互，它直接调用AlgoController类进行分析；AlgoController类负责管理和调用各个分析模块，调用NLPTools类完成预处理和句法分析，调用的包括CRF类来抽取评价对象，调用Pruning类来做名词剪枝，调用OpinionWord类来抽取评价词，调用JudgeSentimentElement类来分析情感倾向。

NLPTools类在本程序中起着非常重要的作用，它的主要功能是进行文本预处理，其处理结果会直接影响后面的所有分析。下面介绍这个类实现预处理的思路。

1. 分词和词性标注

Hanlp分词器的原始分词能完成分词和词性标注，但是分词后的结果包含大量粗粒度的词语，即可以继续分词的词语。比如“很好”的粒度不是最细的，仍然可以分解为“很”和“好”。本文的分析任务要求细粒度的分词效果，因为这会影响情感倾向性分析子课题的开展。如果采用粗粒度分词，诸如“很好”、“很漂亮”、“很差劲”之类由程度修饰副词“很”和形容词组合而成的粗粒度词语会被抽取为评价词，为了判断其情感倾向，要求评价词典中包含这些词；同时评价词典中也会包含“好”、“漂亮”、“差劲”这样没有程度副词修饰的细粒度评价词，从一定程度上来讲，评价词典是冗余的。所以把分词器调整为细粒度分词效果是首要任务。调整方式为整理分词词库，把其中包含程度修饰副词和否定修饰词的复合词语删除掉。除了删除包含这些修饰词的复合词，还要根据语料所属的领域增加领域专有名词和网络流行用语。如下表所示。

表 2.3 删除的修饰词

|  |  |
| --- | --- |
| 程度修饰副词 | 很，最，太，超，超级，越，越来越，极，极其，十分，非常，有点，格外、分外、更、更加、越发、有点、稍、稍微、几乎、略微、过于、尤其 |
| 否定修饰词 | 不、没有、没 |

表 2.4 分词词库新增的部分词语

|  |  |
| --- | --- |
| 手机领域专有名词 | 续航能力、屏显、触控、流畅度、外放…… |
| 网络流行用语 | 给力、坑爹、水…… |

1. 去停用词

在后续操作中要做句法分析以得到句法依存关系，如果句子比较复杂，句法分析器不能正确分析，所以需要利用去停用词来简化句子。比如句法分析器不能正确的分析“这个手机屏幕看起来不错”，若在去停用词时去掉“这个”和“看起来”两个词，句法分析器就能正确分析“手机屏幕不错”。但有时候删掉某些停用词会导致句义的改变。比如“下”是一个停用词，于是“阳光下屏幕不反光”经过去停用词操作后变成了“阳光屏幕不反光”。这句话的含义明显是不正确的，也会导致抽取评价对象时把 “阳光屏幕”被当作评价对象，显然这不是正确的评价对象。所以本课题在筛选停用词时的原则是简化复杂句子的结构，但不改变句义。部分从原始停用词表中被去掉的停用词如下表所示：

表 2.5 从原始停用词表中删去的停用词

|  |  |
| --- | --- |
| 从原停用词表中删除的停用词 | 不是、与、和、且、于、使、去、及、就是、是、有、没有、要、的、上、下、大、内、只…… |

### 可视化实现

本系统采用Java Swing实现用户界面。实现效果如下图所示：

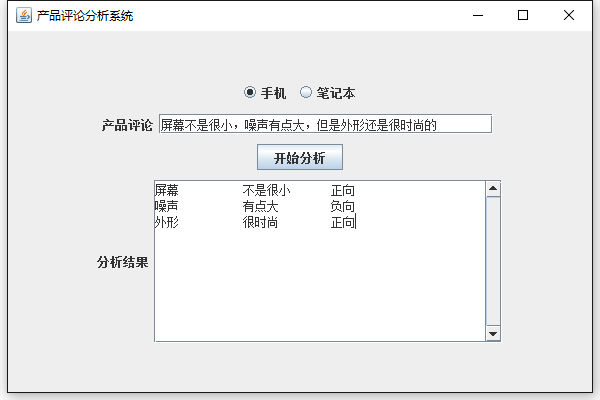


图 7.3 系统界面

界面包含的元素有用于输入产品评论的输入框、分析开始按钮、分析结果展示框。用户在输入框内输入产品评论之后，点击分析开始按钮，系统会在分析结果展示框显示分析后的结果。

# 结论

近些年来，随着互联网的快速发展，网络自媒体的规模得以迅速扩大。导致越来越多的网络用户在网络环境中发表个人的评论、博文、状态等信息。这些包含丰富情感色彩的文本具有巨大的价值，使得更多的学者关注情感分析领域。这几年情感分析的相关研究取得了显著的进步，相关技术也逐渐出现在各种各样贴近人们生活的应用之中。而产品评论作为众多文本种类中的一种，被大量的学者当作研究对象，从中挖掘各种有用的信息。因此本文也以产品评论为研究对象，希望从中挖掘对于消费者和商家有价值的信息，免去人工阅读评论的工作量。

本文的目的是从产品评论中抽取情感要素并分析其情感倾向。情感要素指的是评价对象和评价词。产品评论中的评价对象指的是产品的某些属性，比如手机评论中的评价对象有屏幕、音量等等。而评价词是描述评价对象的词，在本文中评价词被限定为形容词。本文使用条件随机场模型抽取评价对象，使用依存句法关系抽取评价词，使用基于HowNet的评价词典判断情感要素的情感倾向。

本文所做工作的主要意义有以下几个方面：

* + - 1. 本文把评价对象抽取问题转换为序列标注问题，通过观察和总结产品评论中经常出现的句式，使用少量的特征和针对性的训练集训练条件随机场模型，达到较好的抽取效果。
      2. 本文利用依存句法关系，和针对常见评论句式制定的分析规则，比较简便地完成评价词抽取的任务。
      3. 本文经过整理HowNet的词典得到了比较实用的评价词典，并且给评价对象定义正负情感色彩，给评价词定义增强和削弱情感的作用，以此解决判断上下文相关的评价词的情感倾向的问题。

最后，本文对所提出的分析算法进行了实验验证，并实现了一个面向产品评论的情感要素抽取和情感倾向性分析的可视化系统。

本文提出的方法虽然比较简便有效，但有几个问题需要进一步探讨和解决：

1. 本文使用的方法能比较好地分析句式较为简单的文本，对于复杂的文本表现有欠缺。在条件随机场的训练中可以增加更多的特征使其能从更复杂的句式中抽取评价对象；评价词的种类不仅仅有形容词，名词、动词都有可能成为评价词，为了提高抽取程序的适应性，使程序能识别更多种类的评价词是一个突破口。
2. 本文有些工作需要人工参与，导致程序的领域可移植性较差。如果对于专一领域开发应用，本文的方法可以发挥作用，但想要扩展到更多领域，选择无需人工参与的方法显然更合适。在以后的研究中，要着重考虑使用自动化的方式代替本文中人工参与的工作。

# 致谢

# **参考文献**

[1] O'CONNOR B T, BALASUBRAMANYAN R, ROUTLEDGE B R, et al. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series; proceedings of the ICWSM, F, 2010 [C].

[2] BOLLEN J, MAO H, ZENG X-J. Twitter mood predicts the stock market [J]. CoRR, 2011, abs/1010.3003(

[3] HATZIVASSILOGLOU V, MCKEOWN K. Predicting the Semantic Orientation of Adjectives; proceedings of the ACL, F, 1997 [C].

[4] HU M, LIU B. Mining and summarizing customer reviews; proceedings of the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, F, 2004 [C]. ACM.

[5] HU M, LIU B. Mining opinion features in customer reviews, F, 2004 [C].

[6] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules; proceedings of the Proc 20th int conf very large data bases, VLDB, F, 1994 [C].

[7] YI J, NASUKAWA T, BUNESCU R, et al. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques; proceedings of the Data Mining, 2003 ICDM 2003 Third IEEE International Conference on, F, 2003 [C]. IEEE.

[8] KIM S-M, HOVY E. Extracting opinions, opinion holders, and topics expressed in online news media text; proceedings of the Proceedings of the Workshop on Sentiment and Subjectivity in Text, F, 2006 [C]. Association for Computational Linguistics.

[9] BRODY S, ELHADAD N. An unsupervised aspect-sentiment model for online reviews; proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, F, 2010 [C]. Association for Computational Linguistics.

[10] JIN W, HO H H, SRIHARI R K. A novel lexicalized HMM-based learning framework for web opinion mining; proceedings of the Proceedings of the 26th Annual International Conference on Machine Learning, F, 2009 [C]. Citeseer.

[11] JAKOB N, GUREVYCH I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields; proceedings of the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, F, 2010 [C]. Association for Computational Linguistics.

[12] ZHIQIANG T, WENTING W. DLIREC: Aspect Term Extraction and Term Polarity Classification System [J]. 2014,

[13] LIU K, XU L, ZHAO J. Syntactic Patterns versus Word Alignment: Extracting Opinion Targets from Online Reviews; proceedings of the ACL (1), F, 2013 [C].

[14] LIU K, XU L, ZHAO J. Opinion target extraction using word-based translation model; proceedings of the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, F, 2012 [C]. Association for Computational Linguistics.

[15] GAO Q, BACH N, VOGEL S. A semi-supervised word alignment algorithm with partial manual alignments; proceedings of the Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, F, 2010 [C]. Association for Computational Linguistics.

[16] TURNEY P D, LITTMAN M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4): 315-46.

[17] LIU K, XU L, ZHAO J. Extracting opinion targets and opinion words from online reviews with graph co-ranking; proceedings of the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, F, 2014 [C].

[18] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment Classification using Machine Learning Techniques [J]. CoRR, 2002, cs.CL/0205070(

[19] HAMDAN H, BELLOT P, BECHET F. Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis [J]. 2015,

[20] ZHAO J, LIU K, WANG G. Adding redundant features for CRFs-based sentence sentiment classification; proceedings of the Proceedings of the conference on empirical methods in natural language processing, F, 2008 [C]. Association for Computational Linguistics.

[21] KIM S-M, HOVY E. Automatic identification of pro and con reasons in online reviews; proceedings of the Proceedings of the COLING/ACL on Main conference poster sessions, F, 2006 [C]. Association for Computational Linguistics.

[22] LU Y, CASTELLANOS M, DAYAL U, et al. Automatic construction of a context-aware sentiment lexicon: an optimization approach; proceedings of the Proceedings of the 20th international conference on World wide web, F, 2011 [C]. ACM.

[23] CHOI Y, CARDIE C. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis; proceedings of the EMNLP, F, 2008 [C].