硕士学位论文

面向产品评价的细粒度情感分析技术研究

FINER GRAINED OPINION ANALYSIS ON PRODUCT REVIEWS

张玥

2012 年 12 月

国内图书分类号：TP391.4　　学校代码：10213
国际图书分类号：621.3　　密级：公开

工学硕士学位论文

面向产品评价的细粒度情感分析技术研究

硕士研究生：张玥
导　　师：徐睿峰副教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：深圳研究生院
答辩日期：2012 年 12 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.4
U.D.C.: 621.3

Dissertation for the Master Degree in Engineering

FINER GRAINED OPINION ANALYSIS ON PRODUCT REVIEWS

Candidate: Zhang Yue

Supervisor: Asso. Prof. Xu Ruifeng

Academic Degree Applied for: Master of Engineering

Specialty: Computer Science and Technology

Affiliation: Shenzhen Graduate School

Date of Defence: December, 2012

Degree-Conferring-Institution: Harbin Institute of Technology

摘　　要

近年来，随着电子商务的迅猛发展，互联网中出现了大量的产品评价文本。人们开始寻求通过自动的方法在这些海量的主观文本中抽取出有价值的信息，于是情感分析研究应运而生。作为情感分析的一个重要的子任务，细粒度的情感分析,如评价发出者和评价对象的识别，由于可以获得用户评价的精细化信息，因此越来越受到研究者的关注。

目前细粒度情感分析研究中，基于模板和规则的方法来抽取细粒度要素的方法是其中的一种主要途径，然而这种方法存在灵活性弱，扩展性差，召回率低等缺点。另一种主要途径是将细粒度情感要素的抽取视为序列标注问题，采用基于随机条件场、隐马尔可夫模型等序列标注方法来抽取特定的情感要素，但是这些方法无法很好地处理评价文本中大量存在的评价元素之间的长距离依赖，这也降低细粒度情感分析性能上提升的空间。

围绕着对产品评价文本细粒度情感分析任务，本课题进行了一系列系统化的工作。第一，提出了面向产品评价文本的细粒度情感标注体系。该体系使用领域本体的形式组织和表示产品的相关概念节点。依据该体系，对1000短篇相机的产品评论文本进行了标注，建立了一套高质量、细颗度情感分析语料(CUHIT Opinmine)。第二、文本提出了一种使用基于依存句法树结构的条件随机场模型对评价对象与评价描述进行结合抽取的方法，该模型改善了线性条件随机场在标注细粒度情感要素时无法适应情感要素长距离语义依赖的问题，使用树边特征表达了细粒度要素中的句法相关性。在CUHIT Opinmine语料库与COAE2011任务三数据集对该模型分别进行了实验和评估。最后，为了进一步提高产品评价的评价对象的识别效果，本文提出了一种基于半监督的学习本体节点新实例的方法来处理产品评价文本出现的词典未登入领域专有词。实验中将该方法的输出结果构建为一套单独特征集提供给细粒度情感分析模型使用。

本课题的贡献如下：一方面，标注的一套产品评价细粒度语料为后续的情感分析提供数据支持；另一方面，提出的使用基于依存句法树结构的条件随机场模型的方法性能更优，验证了使用树边能更好的表示评价文本的语义的相关性的推断；最后，课题提出了加强产品评价细粒度情感分析中对词典未登入领域专有词的识别方法，实验证明该方法能显著提高评价对象识别的召回率。

关键词： 情感分析；领域本体；细粒度分析；条件随机场；依存分析

Abstract

In recent years, with the blooms of Internet, there comes out a lot of subjective texts. Especially for the product evaluation texts, it's encouraged to share comments of product evaluation early by the electricity supplier. As an important subtask of sentiment analysis, fine-grained emotion analysis can analyze the specific emotional elements in subjective texts, which has gained more and more attention by researchers.

In previous researchs of grained emotional analysis, some approachs use the templates to extract the fine-grained elements, however, which has a poor flexibility and low recall for extraction. Many methods are fail to classify the polarity of the pair of the product attribute and the key word of opinion expression.

To meet the need for finer-grain sentiment analysis, we proposed a strategy of the corpus annotation system which organizes the product attributes in the domain ontology style. Then manually annotate 1000 reviews crawled from the digital cameras web sites to be a set of open corpus, which notes the appearance of the pair of product attribute and the key word of opinion expression that provide data support for fine-grained sentiment analyse. In this article, a semi-supervised learning approach is presented to extract the instances of the product attributes. In practical corpus, especially for the corpus of specific domain, the instances of product attribute are always explained in a out dictionary way, for which general segmenter performs poor. This approach can mining the out-dictionary words and provide the feature support to the further domain-specific analyse. We alse propose to apply tree-structured CRFs model to extract the sentiment elements. Instead of the word sequence, the use of dependency features is helpful to describe the direct restrictions which are insensitive to the word distance in the sentence. The experimental results on COAE 2011 dataset and CUHK Opinmine dataset show that the proposed tree-structured CRFs achieves better performance comparing with CRFs without edge features and CRFs with linear features.

Keywords: Sentiment Analyse, Ontology, Fine-grained, Conditional Random Fields, Dependency Parsing

目 录

第 1 章 绪论

1.1 课题来源

课题来源于国家自然科学基金《文本情绪分析中的关键问题研究》及深圳市基础研究计划项目《面向社会媒体的文本情绪分析与预测研究》。

1.2 研究背景与研究目的

电子商务的迅猛发展潜移默化地改变了人们在互联网的习惯。各大电子商务公司，包括国外的亚马逊、eBay以及国内的淘宝、天猫、京东等都鼓励消费者为购买的商品提交评论，久而久之，消费者也更愿意主动分享自己的产品评价。由此，网络中出现了越来越多的来自于用户的对消费产品的评论，领域涉及方方面面。这些包含着大量用户主观性文本引来了越来越多研究者的关注。

来至用户的产品评论携带了人们对产品的主观情感信息，包括对产品主观的喜恶，也包括用户所阐述的正面或负面的客观评价。这些信息能让消费者更容易掌握产品的目标信息，它们已经成为了现在消费者在进行电子商务活动里最终决策的重要依据。另一方面，对于生产商来说，用户评价信息属于消费者的反馈信息，它也为厂商进行产品的改进提供了宝贵的资料。所以对这些信息进行分析和处理所能带来的价值是不言而喻的。

可是对庞大的评价数据来说，基于人工的情感分析的速度是显然不能满足要求的。那么如果能使用机器以自动的方式来处理这些产品评价信息，并且对特定的评价要素进行分析，便可以达到海量情感分析的目的。在这样的需求下，情感分析（Sentiment Analyse）技术开始进入研究者的视野。

情感倾向分析的研究在最近的十几年有了可观的发展，它也被称为意见挖掘（Opinion Mining）、倾向性分析（Opinion Analysis）、情感分类（Sentiment Classification）等。情感分析是自然语言处理的子任务，它通过对评价文本进行分析，对情感的相关信息进行分析抽取。情感分析的主要任务包括判断文本的感情色彩，文本是否为主观句等。文本的大小可以是篇章级、段落级、句子级，也可以是词级。文本的种类可以是新闻文本或产品评价文本。

作为情感分析的一个重要子任务，细粒度的情感分析可以对产品评价文本进行更深入的分析。除了判断文本的倾向性以外，细粒度的情感分析还会关注其它的情感要素分析，如在新闻文本中对评论发起者（Holder）的抽取，特别

是在产品评价文本中，用户的评价常常会与某个产品或产品的具体属性相关联，形成一个”产品-评价词“的搭配对。对产品评价文本中的搭配对、评价描述、意见发起都的分析与抽取也者属于细粒度情感分析的内容。

对评价文本进行细粒度的情感分析能提取更多的有价值的更细粒度的情感信息，基于这些信息可以开发各种应用。首先情感分析可以用于生成产品的口碑，通过分析大量的相同的产品的评价文本，可以统计出广大用户群体对特定产品的总体评价，为生产厂家制定销售策略与营销策略提供数据支持。

另外，来自于其它用户的使用体验也可以为产品的潜在购买者提供购买依据。通过对海量的评价文本进行细粒度的情感分析，可以挖掘出用户对某个产品特定属性的评价，潜在的购买者可以将这些评价结合自己的需求来寻找出最适合的产品。

可见，情感分析具有很高的实际应用的价值。

1.3 国内外研究现状

1.3.1 粗粒度情感分析研究

情感倾向分析的主要内容就是抽取和分析文本中的主观情感和评价[1]。情感倾向分析的早期研究主要集中在词语的倾向性分析，它包括判断目标词语是否为情感词、判断情感词的倾向性（Polarity）。然后，基于篇章级、段落级、句子级的情感倾向分析技术也得到了充分研究，这些研究也叫做粗粒度情感分析研究。粗粒度情感倾向分析可以被描述为一个典型的分类过程。分类可以是二值的，主观或者客观，也可以是多值的，包括正面、负面、中性等[2]。分类主要有两种基本思路：基于情感词典的、上下文语言知识的；基于构建特征并依据特征进行分类的。分类方可以是基于有指导的监督学习，或基于无指导的无监督学习、半监督学习。

基于情感词典与上下文语言知识的方法主要是使用情感词典中标注好的情感词倾向性以及文本中同义词关联、反义词关联等关联规则来对情感词进行分类。例如Turney[3]首先分析文本中评价词的倾向性，通过计算句子中的形容词和副词的极性然后进行加权求和来求得句子最终的倾向性。

在基于特征分类的粗粒度情感分析研究中，不同的特征与分类模型被使用和比较。Pang[4]等人的研究首次涉及这个思路，他们使用贝叶斯、最大熵与支持向量机模型进行比较，选择了n-gram特征与词性特征，并且显示了词是否出现的特征比出现的频率特征能带来更好的性能。Hatzivassiloglou等人[5]的研究表

明形容词、副词是倾向性分类的很好的指示词。Xu[6]的研究表明类似赞扬、批评等隐含了观点倾向的动词也是重要的极性标志。Qiu等人[7]在研究中引入了依存句法分析信息作为特征，实验显示了依存依赖关系、依存路径等语义特征的引入对情感分析有帮助。

无监督的方法一般用于情感词的倾向极性的生成[3, 8]，通过定义词语或短语的相似函数，使用聚类的方法或者参考与种子情感词点式互信息（Point Mutual Information）的方法来获得目标词的极性。无监督的方法希望在大规模的未标注语料中自动生成出情感倾向性信息，这缓解了有指导的分类方法过度依赖情感词典的限制。不过，虽然使用无监督的方法来生成词语以及短语倾向性的研究取得了一定的进展，可是性能还是没有完全满足人们的要求。

另一方面，很多对情感倾向性的研究只限于对观点中心词倾向性的分析上，却忽略了评价搭配对（Pair）的极性判断的重要现象。特别是在产品评价文本中，评论的主观评价往往都会与一个特定的产品属性相关联，而整个观点的极性是由整个搭配以及搭配的上下文共同决定的。如句子“油耗高”与句子“性能高”，两个句子中的观点中心词都是“高”，但两个句子的极性却截然不同。

1.3.2 细粒度情感分析研究

与粗粒度情感分析所处理的层面不同，细粒度的情感分析可以抽取与识别粒度上更细的情感信息。包括情感表述范围、意见的发起者、评价对象与评价词、各情感要素的情感极性等等。细粒度情感倾向分析技术获得了研究者越来越多的关注。

其中观点发起者是指观点表达的行使者。在评价文本中，观点的发起者是显式给出的，或在转述其它人的观点过程中间接出现，甚至更多是缺省的。不过在一些博客以及论坛的评论文本中，观点的发起者通常是作者本人。

评论对象是指主观评论的承受者。一般来说，新闻文本中的评论对象的概念比较宽泛，不好处理，但是在一些产品评价的评论文本中，评论对象就会具体为某一个产品或者产品的具体属性。

评价中心词是指评论中对评论的倾向性关键作用的中心词，如“漂亮”、“优秀”等词，评价中心词大多为形容词与副词[3]，另外动词、名词也是倾向性的重要指示器。在产品评价文本中，评价中心词往往会与某个特定的评价对象相关联。

Kobayashi等人[9]将情感倾向表达形式进行了统一定义。每一个情感表达都

会对应一个五元组，包括意见发起者、评价对象、评价中心词、评价强度、极性。Wu[10]等认为用户的主观评价不适合用简单的极性来描述，提出用图的方法来表示文本的情感信息。图中顶点由观点发起者、评价对象、评价中心词等实体组成，而实体间的关系则形成了图的边。

目前细粒度情感分析方法中，使用模板和规则来抽取细粒度要素的方法被大量使用。Yi等人[11, 12]使用上下文模板在一个给定大小的窗口来描述抽取目标的环境同时抽取评价对象。另一些学者[7]使用了语法级的特征来构建模板。Xu等人[6]的工作采用了两步分析策略。在识别情感句的基础上，利用结构特征和搭配关系，首先识别出情感发出者的核心词语，而后使用句法特征和动词、介词等来从核心词语进行扩展，将获得的最长名词短语作为完整的情感发出者。

然而模板与规则是对数据先验知识的总结，并且模板与规则常常是工人制定的，这会引入一些缺陷。一方面，由于模板的制定都有局限性，模板的方法不能很好的覆盖所有的可能实例，这使得基于模板的方法虽然有着较高的精度，但是召回率都不高。另一方面，制定模板与规则时往往会引入领域相关性，这使得基于模板与规则的方法的扩展性很差。

另一些研究者将细粒度情感要素的抽取当作序列标注问题来研究，Choi等人[13, 14]使用线性的条件随机场来识别主观文本中的观点发起者以及观点描述。然而条件随机场具有马尔科夫性质，即某时刻的条件概率只与当前状态有关。这个性质会使模型无法很好地处理由句法等特征引入的长距离依赖，而句法信息在细粒度的情感分析中有很重要有作用[7]，这降低了基于序列标注方法的细粒度情感分析模型的性能。

1.4 本文的主要研究内容和组织结构

围绕着对产品评价文本细粒度情感分析任务，本课题进行了一系列系统化的工作。第一，提出了面向产品评价文本的细粒度情感标注体系。该体系针对产品评价文本建立其特有的领域本体库，通过领域库的形式维护和组织产品的相关概念。体系除了要求对一般的产品评价细粒度情感要素进行标注以外，还需要标注产品评价中的多评论问题。实际标注中，依据该体系对1000短篇相机的产品评论文本进行了标注，建立了一套高质量、细颗度情感分析语料(CUHIT Opinmine)。

第二、文本提出了一种使用基于依存句法树结构的条件随机场模型对评价对象与评价描述进行结合抽取的方法，该模型引入了文本的深度句法分析，借

助依存句法的树结构来缩短评价文本中相关情感要素之间语义距离，使用树边特征表达了细粒度要素中的句法相关性，改善线性条件随机场在标注细粒度情感要素时无法适应情感要素长距离语义依赖的问题。在CUHIT Opinmine语料库与COAE2011任务三数据集中的电子类数据对该模型分别进行了实验和评估。

最后，为了进一步提高产品评价的评价对象的识别效果，解决产品评论文本中出现的未登入产品属性词的识别为抽取。本文提出了一种基于半监督的学习本体节点新实例的方法来处理产品评价文本出现的词典未登入领域专有词。方法首先通过无监督策略迭代得到初始的分词模型，然后通过人工制定的模板对分词模型进行修订并同时依据修订效果来调整该模板的权重，而权重会影响到最后对属性的抽取。实验中将该方法的输出结果构建为一套单独特征集提供给细粒度情感分析模型使用。

本课题的贡献如下：一方面，标注的产品评价细粒度语料为后续的情感分析提供数据支持；另一方面，通过基于依存句法树结构的条件随机场模型的方法的实验验证了使用树边能更好的表示评价文本中情感要素间的相关性，而线性条件随机场模型的边特征在细粒度情感要素抽取任务里无法描述相关性；最后，课题提出了加强产品评价细粒度情感分析中对词典未登入领域专有词的识别方法，实验证明该方法能显著提高评价对象识别的召回率。

本论文分为五个章节：

第一章为绪论，主要介绍了课题的研究背景和意义，国内外的研究现状以及论文的主要工作以及组织结构。

第二章为情感分析相关技术概述部分，对情感分析技术进行概括的介绍。

第三章为领域本体与基于领域本体的语料建设的工作介绍。

第四章介绍了使用基于依存树结构的条件随机场模型进行细粒度情感要素进行抽取的方法以及实验

第五章介绍了基于领域本体的半监督概念节点学习的分析、方法和实验。

最后总结了本论文的研究工作和存在的不足。

第 2 章 情感分析相关技术概述

情感分析的相关问题的讨论与研究从2001年开始大量出现[15–18]。总结其背后的原因，一方面是由于自然语言处理与信息检索领域的机器学习基础研究长足发展的促进作用，另一方面，得益于互联网的迅猛发展，情感分析相关的语料数据特别是情感分析的应用需求也加快了情感分析发展步骤。

根据分析对象和任务的不同，情感分析可以归纳为两种：情感的倾向性分析（粗粒度情感分析）和细粒度的情感分析。情感分析的分析对象可以是话题、篇章级、句子级以及词级。细粒度的情感分析中除了对分析对象进行倾向性判断以外，还有对细粒度单元进行挖掘抽取。举一个例子，句子“我哥哥说其实iPhone4外观太丑了。”，那么对这个句子中可以提取出来的细粒度信息包括：观点发起者“我哥哥”，评价的对象“iPhone”，评价的属性“外观”，以及评价中心词“丑”等等。对这些情感要素进行进一步分析可以发现，评价中心词“丑”修辞对象是评价对象“iPhone”的属性“外观”，它的情感倾向性是负面的。不同的细粒度情感分析系统对情感要素分析会有不同的侧重。

2.1 粗粒度情感分析研究

粗粒度情感分析也称为情感倾向性分析，它是情感分析的重要的子任务。情感倾向性的目的是判断篇章、句子或者词的总体情感倾向性。情感倾向性分析可以看为一个典型的分类问题[2]：给定一段文本，无论作者评论的对象是什么，也无论观点作者的视角，将这段文本按照情感倾向性的极性进行分类。

对文本的极性两分法有时并不是一种严格的方法。例如新闻文本可以被分为正面新闻与负面新闻，但是新闻同时也可以是正面或负面的事实描述类的。虽然说有刻意的将文本的情感极性细分为太多的类型没有实际意义，但是这一现象也给了我们一些指示：一方面一些具有情感极性的客观描述的句子可以帮助判断主观的情感极性。另一方面，有些情况下，文本是客观的还是主观的界限不明确。最后，特别是在产品评价情感分析方面，客观的极性信息与主观的极性信息同样重要。

一些研究希望判断目标文本是否含有人的主观情感[19]，另一些研究将不同级别的文本分为褒义与贬义两类[3]，或综合了主观判别、极性判别等粗粒度判别的多分类问题[20]。根据情感倾向分类使用的知识来源和采用的分类策略，大多数的粗粒度情感倾向分类技术可以分为基于词典和语言知识的、基于监督学

习和基于非监督学习、半监督学习的三大类方法。随着情感倾向分析技术的不断深入，基于抽取的细粒度情感倾向分析技术吸引了越来越多的注意。识别的细粒度情感表达信息包括情感表达子句、情感/意见的发出者、情感/意见的对象、情感的陈述、情感极性等等。

2.1.1 有监督学习方法

得益于充分研究的机器学习算法和情感标注语料库的建设，基于监督学习的情感倾向分类技术取得了良好的进展。目前主要使用的分类器包括Naïve Bayesian、支持向量机(Support Vector Machine, SVM)、最大熵(Maximum-Entropy, ME)以及条件随机场(Conditional Random Fields, CRFs)等[4, 17, 21]。

朴素贝叶斯经典的分类模型，在粗粒度情感分析中也有相应应用。可以将带有极性特征的POS信息，还有句子中的中心动词，主语，以及修辞词都可以做为分类器的特征。通过以下机制对文档d_j进行分类（d_j表示为由特征组成的向量），将d_j判定为类别c_i,这个c_i是将最大化$P(c_i|d_j)$:

$$P(c_i|d_j) = \frac{P(c_i)P(d_j|c_i)}{P(d_j)} \tag{2-1}$$

$P(d_j)$表示随机选取的文档中特征是d_j的概率，$P(c_i)$表示随机选取的文档以属于类别c_i的概率。为了得到式2-1中的$P(c_i|d_j)$，朴素贝叶斯模型假设d_j中的所有特征都具有条件独立性，即：

$$P(c_i|d_j) = \frac{P(c_i)\prod_{ii=l}^{m} P(f_i i|c_i)}{P(d_j)} \tag{2-2}$$

对朴素贝叶斯进行一个改进，为了防止对训练集的过拟合，多朴素贝叶斯分类器将对特征集F的若干子集训练出多个子分类器，最后将这些分类器的结果进行归并处理。

在给定特征的子集$F_1, F_2, ..., F_m$后，对他们进行分别训练得到相应的分类器$C_1, C_2, ..., Cn$，接下来用训练集对C_1进行训练，然后使用训练后的C_1分类器进行分类，把分类结果中的错误的句子移除，将剩下的句子重新训练分类器$C_2$2，以此类推。Yu和Hativassiloglou[18]在多朴素贝叶斯模型中使用了5个特征子集，从最开始的单词特征，到多元词特征以及POS特征极性词特征等。

Pang等人[22]提出了一种使用采用最小切割（Minimum s-t cut）的方法，首先使用PMI信息以及SVM的方法对语料进行学习，得到一个判断句子是否有倾向性的分类器，然后将判断为倾向句的句子做为节点，以倾向句间的相似度做为图的权重，构建为图后。最小切割法求得了损失函数最小的一种分割，也就是

图里两种类型相差最大的一种分割。

主观文本所转化的特征是分类器能够成功分类的关键，对于同一种分类器，一个更好的特征集会让模型达到更高的性能。借用传统的信息检索系统中用词与词频所组成的向量来表示整个空间中的文档的方法，在粗粒度情感分析过程中也使用相似的方法构建特征。除了最基本的将词是否存在表示为一个二值的维度外，还有使用df-idf来进行表示文档的方法。考虑到主观文本中特定顺序组合对的重要性，也有使用bigram或N-gram来进行特征构建的方法。表示方法有自己的特点，通过Dave的文章[17]以及Hativassiloglou[18]可以看出，每一种方法都有自己的特点，对相应的分类器要选择合适的表示方法。

词性与句法的特征使用了句子语议级的信息。Turney在[3]中表示了形容词是倾向性的很好的指示词。Hatzivassiloglou等人的在主观性识别的工作[8]也显示了形容词与主观性的相关性。虽然形容词是主观性的很好的指示词，但这并不表示其它类型的词不能用于主观点的判断，Pang等人[4]在对电影评论的研究发现，如果只使用形容词性的词语为特征的效果低于使用相同数据的高频词做为特征，研究还指出名词、动词也是主观词很好的指示器。另外Riloff使用了Bootstrapping的方法抽取了有倾向性的名词。

句法信息也可以被用于情感分析研究中，Kudo等人[23]讨论了两个层次的倾向性分类过程，发现依存句法树信息的使用效果超过了使用bag-of-words特征的效果。Yi则在[11]文章中依靠句法信息，对特定的词语组合进行抽取。符合一定特征的模式可以被抽取。连词，包括否连词也提供了倾向性一致性的信息，Hatzivassiloglou在[5]就使用了连词信息对倾向词进行了预测。另一些研究[24, 25]一些固定搭配以及深层次的语法模板对主观性的判断也有帮助。

对否定词的处理也是倾向性分析的生活要研究内容，然而当使用bag-of-words来表示文本的时候，会为否定词的分析带来困难。早期的研究[16]中将否定词是否出现单独作为特征来进行分析。然后并不是所有的否定词都会将倾向性反置，Na等人[26]尝试用更精确的方式来处理否定词，他们使用了基于词性模板的方法来匹配挖掘否定性对主观性进行反义的情况，这使得他们在电子评论文本中的正确率提升了3%。Kennedy等人[27]则使用了更深层次的句法树分析。

2.1.2 无监督学习方法

有监督的学习方法可能在精确度上有一些优势，不过当训练数据相对目标数据非常小时，无监督的学习方法将会有不可替代的优胜。并且，有标注的训练数据相比没有标注的训练数据要小得多。

对种子词进行扩展，可以使用WordNet，同义词典等词典类工具进行扩展。Hatzivassiloglou与McKeown[5]讨论了一种借助评价词在句子中并列关系来进行评价词扩展的方法，方法中，人工找出了一些极性很强的，并且在不同领域极性不会发生变化的词做为种子词，通过对大量语料做句法分析后，用二层的有穷文法从语料中抽取了13,426个并列词对，包括15,048形容词以作为后选极性词。每一个词对都有三个属性，并列的形式（and, or, but, either-or, neither-nor），修辞的种类（定语，表语，动结语），已经被修辞词的数量（单数或复数）。使用这些特征为参数，构建一个评函数来描述词与词之间的相似程度。最后依据这个函数对最有的修行极性词进行一次聚类，以分出不同倾向性的词语。

在句子的层面，对主观句与客观句子的判断里，通过比较一主观句与客观句的相似度是一种简单的实现方法。Yu与Hatzivassiloglou[18]指出，通过定义句子的相似度，基于词、短语的统计，依据WordNet的运算，可以发现主观相互之间的相似度要大于主观句与客观句之间的相似度。那么通过计算句子间的相似度就可以分别出主观句与客观句。相似度的计算分为三个步骤：首先，使用IR的方法，对句子所在的文档进行相关性的计算（向量空间夹角等传统IR模型），以确保所比较的句子是在同一个领域中。再者，将句子与所在文档的其它句子进行相似度计算，得到一个平均相似度。第三步，让句子与两个已经分好的集合进行相似度计算（主观句集，客观句集），以观察相对于哪一个集合的相似度较高。最终，定义一个阈值对句子进行分类。

在判断句子极性的研究中，Turney[3]介绍了一种通过点互信息（PMI：point mutual information）来进行判断的方法，这种方法假设具有正向极性的短语通常会与其它正极的短语共现，同样，一个负面极性的短语通常会出现坏的负面共现。他尝试计算短语与其它相应极性短语的相似度，最后讲正面的相似度与负面的相似度做减法。

自举（Bootstrapping）的方法也被用于倾向性分析中，它的思想是用一些简单原始的分类器来产生活数据的标注信息。Riloff and Wiebe[24]使用了自举方法来学习观点表述的抽取，Pang等人[28]，他们主要考虑如何对观点查询所返回的带有主观的文本结果进行排序。借助信息检索的思想，高频率的出现在少数领域的词更据有表达性，但是，人们的观点表达形式却会不同，主观性的文本更倾向于由那些不是那么罕见的词来构成。

2.2 细粒度情感分析研究

细粒度的情感分析是指从目标文本中获得更具体的情感要素，如评价发

起者、被评价对象、评价观点短语等更细粒度要素的抽取工作。与粗粒度的情感分析相比，它是更深层次的任务，通过挖掘出文本中更加深入全面的情感信息，也让细粒度的情感分析变得更加有实用价值。

在细粒度的表示方面，Kobayashi等人将情感倾向表达表示为一个五元组，包括评价者、评价对象、评价词、强度、极性。而Wu等人提出了情感倾向的基于图的表示方法，图的顶点由评价者、评价对象、评价词、原因、条件等实体组成，边则表示顶点间的关系。对这些情感倾向表达的元素进行抽取的工作，一般称为基于抽取的细粒度情感倾向分析。

基于规则的方法的方法中，Yi[11]使用了不同等级的模板来得到候选评价对象并从中抽取符合要求的对象。Xu[29]使用了启发式的规则方法对评价来源与评价对象进行了抽取，在情感倾向分析之前，加入了一个基于Discourse的预处理模块，对文本中省略的动作发出对象进行还原和补充，提高了情感发出者，特别是文本中省略的发出者的识别正确率。而Bloom[12]在对电影的评价形容词的识别中使用了人工创建的模板。

Hu与Liu[30]认为高频率的名词或名词性短语更可能成为评价对象，他们使用启发试的方法对一些情况下的对象后候选进行修剪，首先，那些由多个词组成但是词却不符合特定顺序的候选；单词组成的后选词被发现包含它的父文本也成为了候选。

Popescu与Etzioni在文章中[31]描述了三种倾向性变化的情况，分别是单词、产品属性与单词搭配、产品属性与单词搭配并且考虑领域与语境。使用文法信息对显示的评价句进行抽取后，依据词语对倾向性的贡献值，建立了一个可迭代的模型。在三种倾向变化的情况中进行迭代，以找出词级，属性评价搭配以及语境级的倾向。

Hu等人[30]以电子产品为研究对象，使用三个步骤对产品信息进行抽取：（1）从评论文本中找到产品的属性。（2）将判定为倾向句的句子进行抽取判别。（3）最后给出结果，结果中将某一个产品的属性与对这个属性进行描述的用户评价联系了起来，并且对极性进行判定与统计。

情感分析的要素还可以包括作者的态度、与对某个话题的观点。例如Grefenstette[32]在研究中通过分类来判断新闻文本的作者的政治观点的方向。

Gamon等人[33]的研究基于半监督的机器学习方法，以句子为对象，将关键词与同句中相关的产品属性进行聚类。并用这一个算法开发了产品评介系统Pulse。系统对每一个特征所相关的观点进行数量统计和情感倾向差异的对比。

Wilson等人[34]研发了OpinionFinder系统，这个系统首先对主观性的语句进行识别，当文本被判别为主观句后，再对其进行分析，抽取出意见主题、事件、情感倾向等细粒度的情感要素。

评价搭配是用户观点中评价对象及其评价修辞的搭配对，在产品评价的分析任务中，对评价搭配对的分析和抽取显得尤其重要。产品评价中，用户情感的往往不是单独表达还是倾向于针对某一个具体的产品对象。如“待机长”中的的对象与评价对“待机”与“长”，极性为褒义。不难发现，评价搭配对的极性有很强的领域性与上下文相关性。相同的评价修辞“长”，在搭配对“延时长”中极性却是为贬义。

一些研究都探索使用基于规则的方法对评价搭配对进行抽取。Kobayashi等人[35]通过对评价对象与评价词之间的观察，总结出了若干简单的模板来进行抽取任务。不过模板只是对词词的共现关系进行描述，而并没有引入深层的语义信息，最后在结果中引入了过多的噪音。

Qiu等人[7]在抽取中结合了依存句法信息，通过总结评价对象与评价词间的依存路径来构建抽取模板，取得了更好的结果。[36]通过对情感句进行语义成分分析(Semantic Role Labeling)来识别特定的谓词词组，进而实现情感发出者和情感表达的同步识别。他们的方法基于语义Parsing，通过识别句子中的谓词和基于谓词的语义框架(Semantic Frame)，找出符合情感表达特征框架的句子。在识别出情感句的同时，利用Frame的信息识别情感发出者。

另一些学者将细粒度情感分析过程看作为序列标注问题，Choi等人[13, 14]使用条件随机场模型(CRFs)尝试对不同的细粒度要素进行了抽取，Shariaty等人[37]提出一种使用线性CRFs模型对评价搭配对进行联合抽取的方法。

Kim等人[19]首先识别了情感表述，然后再通过分析已经识别出来的结果来找到情感发起者。其中，包括语法树在内的一些结构特征被用于找到文本中的情感表述与情感发起者之间的长距离语义关系。他们使用情感发起者与情感表述间的句法路径的模式做为特征，结合另一些局部特征。

2.3 情感标注资源与评测

近年来，国内外的机构和研究者陆续发布了一些语料集以支持情感分析的相关研究。

Wiebe等人[38]发布的MPQA（multiple-perspective QA）语料库，对新闻语料进行了深度标注，以句子为单位对观点持有者、评价对象、主观表达式、以及极性等情感要素进行了手工标注。

Hu与Liu[30]发布了产品领域的评价标注语料。语料评价文本来至于网络中数种电子产品的用户评论。语料的标注同样以句子为单位，标注者对产品评价对象、评价的极性以及评价的强度进行了手工标注。

Barzilay等人构建并发布了餐馆评论语料。语料中标注者要根据评论文本的描述分别在食物风味、用餐环境、服务质量、用餐价格以整体体验五个方面做出评分。该语料以篇章级评论为单位，语料标示出了用户对特定属性进行评论的现象。

随着情感分析逐渐得到重视，国内外许多研究机构组织了一些情感分析相关的评测。

国际文本检索会议（TREC）至2006年以后开始发布情感分析评测的任务，评测任务的主题从博客观点句检索，发展到文档情感分类、主观评论与客观事实的判断、博主身分判断等。

NTCIR（NII test collection for IR system）的首届评测也举办于2006年，它的主要任务是提取和分析新闻文本中的相应的主观性信息，包括需要判断新闻主题与目标句子是否相关，抽取句子中观点发起者、评价对象、评价极性等。评测语料涵盖了中文、英文、日文三种语言。其中NTCIR-8评测中还包括了情感问答任务、跨语言的情感分析等。

在国内，中国倾向性分析（COAE：Chinese Opinion Analysis evaluation）评测始办于2008年。历年的情感分析评测任务紧跟情感分析的热点，同时也会发布相应的情感分析语料，其中2011年评测中共设置了四个任务[39]：一是句子级的粗粒度情感倾向性分析；二是评价中心词的整体抽取；三是产品评价对象-评价词搭配对及其极性的三元组的抽取；四是观点检索评测。同时发布的评测语料涉及电子产品、电影影评、金融新闻三个领域的评论文本。

2.4 领域本体研究

产品评价的领域本体的构建是对产品所具有的各产品属性进行归纳的过程，其所建立的领域本体库呈现了产品属性概念间的相互关系。其中一般包括继承关系、同义关系等。

产品评价文本具有很强的领域相关性，很多的评价对象都由领域的特有词构成。正常识别和处理这些领域特有词是产品评价情感分析的研究中的一个重要任务。在不同领域中，本体的构建层次与概念差别可能会很大。那么如果本体的构建全部采用人工构建的成本将会很高，而且领域迁移的过程也会遇到很多困难。所以，自动或半自动构建的领域本体库构建有着重要的意义。

针对这个问题，一些研究者通过引入领域本体来提升产品评价对象的抽取性能。其中Agichtein等人[40]使用标准的Bootstrapping结构，结合一种新的对模板以及所抽取的关系进行评价的算法，将其应用到大量语料中进行关系的抽取。Etzioni等人[41]提出的KnowItAll系统通过使用一组与领域无关的模板来生成候选答案，依靠PMI信息对候选的答案进行选择。Thelen等人[42]提出了Basilisk算法，使用全部抽取模板所得到的信息进行判断。

在概念节点的相似度计算方面，研究可以分为两种：基于信息容量的方法[43]与基于概念距离的方法[44]。这里的信息容量指的是一个概念节点代表的语义所包含的内容大小。通过比较两个有共同父节点的概念节点的信息信息容量可以比较他们的相似度。而在利用概念距离的方法中，距离的衡量手段有很多种。首先有学者使用现有的本体库的节点信息来计算概念距离，另一些研究者[41]则节点实例在大语料中的文本相似度来计算本体节点的相似度。

另一方面，在产品评价文本中，评论的极性也具有领域相关性，相同的评价文字在不同的领域有时会有截然相反的主观极性。研究都通过引入领域本体来指导评论的观点极性的判断，Zhao[45]在电影评价情感分析的研究中，首先定义了严格的电影领域的本体结构，本体中记录了电影评论中的评价对象的相互关系，并使用基于规则的方法对本体节点进行扩充。当需要对含有多个评价对象的文本进行倾向性判断时，各评价对象的极性和评价对象在本体中的层次关系会被综合考虑。

Wei[46]在建立了领域本体库后，将评价文本的倾向性判断过程当作层分类过程（hierarchical classification）。分类过程会训练出一个权重矩阵，这个矩阵描述了预先定义的词对本体的贡献程度。依据本体库中节点的层次关系，对权重矩阵与文档向量相乘的结果进行修定后得到最后的分类结果。

2.5 本章小结

本章对粗粒度情感分析与细粒度情感分析的国内外相关研究进行了介绍和分析。还对情感分析的语料建设以及领域本体的相关工作进行了介绍。

第 3 章 领域本体与情感语料库的建立

情感分析的研究离不开已标注的情感数据集的支持。而细粒度情感分析由于需要挖掘评价文本中更细粒度的情感要素，细粒度情感分析的研究更依赖语料的支持。

目前的产品评价的语料建设中，Wiebe等人[38]发布的MPQA（multiple-perspective QA），对535篇新闻语料进行了标注工作，对包括观点持有者、评价对象、主观表达式、以及极性等情感要素进行了手工标注。Hu与Liu[30]发布的产品领域的评价标注语料。

由于中文的细粒度情感分析语料资源较少，而且现有的一些语料库又往往存在细粒度标注不够细、不全面的问题。针对这些问题，本章介绍了根据制定的语料标注规范，基于领域本体的形式对产品评价文本进行了标注工作。

3.1 领域本体的构建

在情感分析中，特别是对产品进行细粒度分析的过程中，对产品属性节点的本体建立会带来两方面的好处。首先通过建立各属性的继承关系，规范节点的共性与异性，可以用统计的方法来在未标注的文本中找出新的与相应概念节点相关联的新词，为接下来的细粒度感情分析提供更多的信息，从而提高分析的召回率与粒度。另一方面，在对文本进行细粒度的情感分析后，将挖掘出的评论对象与提前构造好的本体库节点进行关联，再统计节点的倾向性，就可以得到产品全面的细粒度的产品分析。

文本针对从网络文本中关于相机的评论文本进行研究。通过人工对语料的审阅，对相机的概念节点做出了归纳。本体节点的设计遵循以下原则，

1. 除了根节点以外，所有节点都有一个父节点（单继承）。
2. 节点间的继承关系路径不能成环。
3. 叶节点与对应的实例为多对多关系。

每一个节点都会有对应的实例，例如，节点“电池功耗”对应着实例“电量”、“待机时间”、“电池”以及自身文本“电池功耗”等。实例代表了文本中对本体概念节点的具体表述，从刚才的例子也可以看出，一个概念节点一般对应着多个实例。图3-1是针对数码相机的本体构建的一部分。

可以看出，本体库中的节点呈树状结构，树的最大深度为3，叶节点绘制为方开，它们都是可以被用户直接评论的概念节点。中间节点与根节点将它们

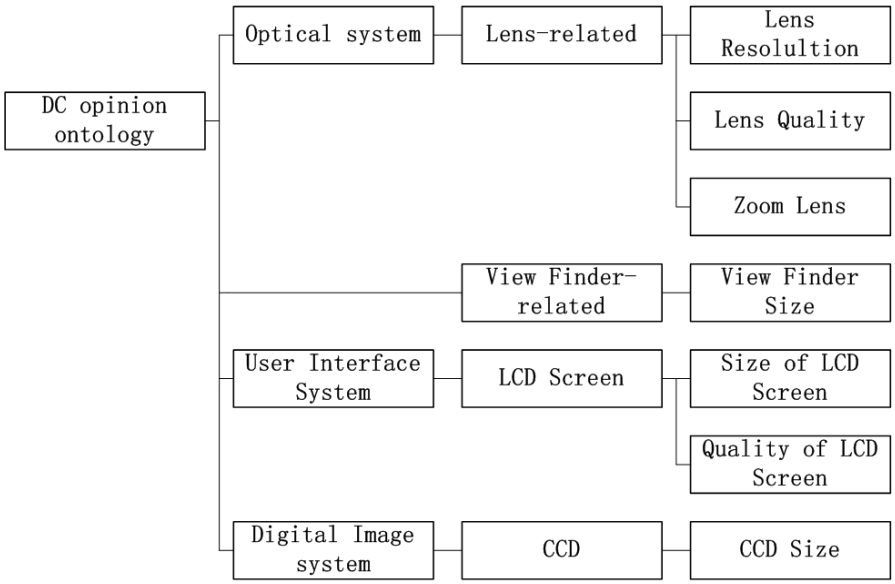


图 3-1 数码相机本体构的一部分

Fig.3-1 Illustration of DC Ontology

的子节点进行归纳，总结出子节点的同共点、相关性。将属性节点分为三层结点。第一层结点对应着功能部分，第二层对应着语义部分，第三层则为文中能直接关联的字符层。如表3-1所示，其中一级节点为12个，二级节点为41，三级节点为60.

3.2 情感语料的标注

为了更好的对细粒度的产品情感分析进行研究，也为了后续的细粒度情感分析提供数据支持，我们组织了研究者对数码相机领域的评论文本进行了人工标注。

3.2.1 语料标注属性节点定义

一些研究者已经从事了中文情感语料的修订工作，其中[47]标注了一份来关于中文新闻文本的情感语料，语料定义了五个关键标签：

1. **观点的范围** 标示了句子中观点内容的范围
2. **观点来源** 观点的发起者
3. **观点的关键词** 观点的中心词
4. **倾向性关键词** 反应倾向性的中心词
5. **倾向性** 观点的倾向性

不过产品评价的特点与新闻语料不同，在产品评价文本中，用户更倾向于将一个有极性的观点赋于或一个产品对象，而产品评价的分析的一个需求与主要作用就是要统计这些产品对象以及相应的评价[17, 30]。具体来说，首先，产品评价的来源一般来说是省略的，在作者没有显示说明引用其他人的评论的情况下，评价来源都是作者自己。第二，产品评价中的观点关键词相对较少，并且重要性不高。第三，产品评价中的评论大多都有评价对象。第四，产品评价听倾向性关键词的倾向性与上下文相关，特别是一些描述数量程度多少的关键词的倾向性必须要根据它所修饰的对象来判断。

基于之前的讨论，语料中每个观点需要标注一共10个主要键的信息，它们是，

1. **观点来源** 观点来源为观点的发起者，如果未显示指明刚默认为作者本人

2. **观点范围** 观点范围标示出了句子中主感倾向性句子的范围。如句子“传统家能的操作，很让人喜爱。”中，观点范围为“很让人喜爱”，之前的则是对事实的描述不以属于观点句范围。

3. **产品对象** 产品对象为句子中评价对象的文本。当评价对象省略时置空。

4. **本体概念节点** 本体概念节点为产品对象相联系的本体概念节点，被省略的产品对象也将与概念节点相关联。

5. **观点范围** 观点范围只的是句子中属于观点的最小独立观点句子的范围。例如句子“机器是黑色的，彩色很好看”，句子前后分句就是本句的最小独立观点句，观点范围就是“彩色很好看”。

6. **观点表述范围** 观点的表述范围标示出了观点评价句中用户对产品对象的评价表述的范围。如句子“菜单结构很清晰”中的观点表述范围为“很清晰”。

7. **观点中心词** 观点中心词为观点表述中的核心词。语料中存在一些评价观点中心词为空的情况，如句子“价格！还是价格！”，结合上下文可知用户在评价中省略了“（价格）太高”。由于这种特征的用法所占比例很小，实际中不考虑。

8. **否定词** 否定词为观点表述中存在的否定词，可为空，如句子“亮度不大”中的“不”。

9. **程度词** 程度词为观点表述中存在的程度词，可为空，如句子“价格非常诱人”中的“非常”。

10. **倾向性** 倾向性为观点的极性，值可为正面（记为“1”）或负面（记为“-1”）。

11. **倾向程度** 倾向程度描述了倾向性的程度，程度一共为两级，普通程度记

为“1”，而更强烈的程度记为“2”。

3.2.2 语料的标注规范设计

标注人员逐条对评论文本进行标注。对目标句会首先进行是否为观点句进行判断。如果句子被判为非观点句，句子不会被添加评论属性，但会被以原始句的形势保存。也就是说，非观点句也会被包括在标注结果中。

对于判断为观点文本的句子，便根据定义好的元素依次进行标注。在标注记录形势中，一个句子中可以存在大于一个的评论（comment），在一个comment中，存在等于或大于一个评论实例，例如在处理下列例句时，

1. 这款相机比较轻巧但却很复杂。
2. 功能很棒!

句一中用户在对相机的评价中对评价对象进行了省略，用一个节点实例“相机”来充代替了子概念节点“便携性”与“操作”。这样的评价将被标注成两个独立的“评论(comment)”，两个观点有着相同的评价对象与概念节点，不同点是评价表述与观点中心词。两个观点表达共享了部分内容，但在语料中被视为用户做了两次评论。句二中，由于评价对象的实例可以对应多个概念节点，用户使用的对象实例可以同时回溯到多个概念节点中。这样的评论被视为一次评论，但评价会与多个概念节点相关联，这样的情况会被标注为一个评论中的多个“评论实例(commentInstance)”。

在标注过程中，首先需要确定评论的范围，当评论中只存在一个评论实例时，那么评论的范围等于评论实例的范围。当评论有多个评论实例时，评论的范围是所有子评论实例的范围的并集。在每一个评论实例中，分别标注预先定义好的属性：观点范围，观点中心词，观点表述短语，否定词，程度词倾向性等。

标注的结果会以XML格式进行储存，例如句子“解像力高，有金属质感”，进行标注为XML格式为，

```
<Sentence SentID="1">
        <Comment CommID="1" is_holder_visible="false"
        is_target_visible="false">
          <CommentContent>解像力高</CommentContent>
          <CommentInstance InstID="1" InstanceContent="解像力高"
          is_attr_visible="true" attr_expression="解像力"
          is_attr_NIL="false" attr="清晰度" is_opinion_visible="true"
```

```
        opinion_expression="高" opinion_keyword="高"
        is_keyword_NIL="false"
        negation="" is_negation_NIL="false" modifier=""
        is_modifier_NIL="false" polarity="1" degree="1"/>
      </Comment>
      <SentenceContent>解像力高，有金属质感</SentenceContent>
      <SentencePolarity>0:1:0:0:0</SentencePolarity>
</Sentence>
```

每个键都会对应一个或多个XML属性节点。其中Comment标签对应一个用户的观点。上述的多重回溯的情况一产生多个Comment。而多评论实例的情况会在同个Comment里产生多个CommentInstance以关联不同的概念节点。为了方便标注人员的标注工作，一个网页版的标注器被设计以辅助标注工作。标注器将各属性节点以直观的形式呈现出来，并且将设计好的本体节点信息固化入选项中，方面标注人员进行选择、使用。附录图A-1为标注器的界面。

3.2.3 标注结果的分析

一共三人参与了标注工作，为了实现语料标注的一致性，三位标注人员首先重复标注了相同的100个评论贴，在讨论了标注结果，对标注标准达成一致后，三人再合作将余下100贴文本进行标注。

语料来源于数码相机爱好者的网站“色影无忌（www.xitek.com）”。评论贴（Review）是用户对产品的基本单位，一个Review中含有若干评价句。人工标注一共处理了1000个Review共计7611个评价句。一共标记了8683个主观评论（Comment），含有9021个评论实例，其中正向实例5774个，负向实例3247，评论实例的数量比主观评论数多出了338，说明多重回溯现象在相机的评价语料中确实存在一定比例。使用了873次否定词以及3304次程度词。

评论与59个概念节点相关联。按照节点被关联次数降序排列可以得到图3-3,图中显示了语料中存在高热度的概念节点，也存在一些频度较低的概念节点。语料中与59个概念节点关联的对象实例数量达到了1867种，这也说明用户在评论中评价对象的文字表述方式多样化。其中TOP-10的概念节点的具体情况如表3-1。TOP-10的概念节点所关联的评论数据已经占到了评论总数的超过50%的份额。

其中节点“其他”频度是355个，节点“其它”是为了统计那些领域本体所能预测的节点外节点，一般来说，“其他”节点所占的比例越大，也就说明人

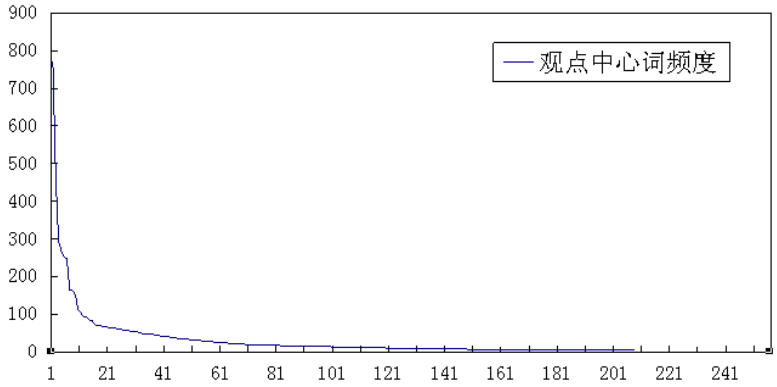


图 3-2 观点中心词的频度

Fig. 3-2 Frequency of opinion key word in annotation

表 3-1 要素按频度排序

Table 3-1 Ranking of elements

(a) 概念节点

节点名	频度
总体	1048
画面	658
功能设计	403
便携性	374
操作	361
其他	355
色彩	347
电池功耗	300
镜头质量	268
对焦速度	251

(b) 观点中心词

节点名	频度
好	770
不错	470
慢	296
方便	269
大	251
高	247
差	165
快	164
小	152
贵	113

工构建的本体库所对目标数据集作出的预测越不精确。在本数据集中，其它节点所占比例为3.9%，还是可以接受的。

观点中心词的种类达到了1573种，提取频度大于4的观点中心词并按照频度降序排列可以得到图3-2，可以发现也存在高热度的中心观点词，TOP-10的中心词就已经关联了50%的评论句，然而，中心词的长尾效应确更为明显。从表3-1可见，TOP-10的观点中心词中，更多的为短的评价词。

另外，TOP-10中的“好”、“差”、“方便”等极性很明显的评价词使用的频度靠前，并且数据表明与他们相关联的概念节点多为概括性节点如“总体”、“功能”等。

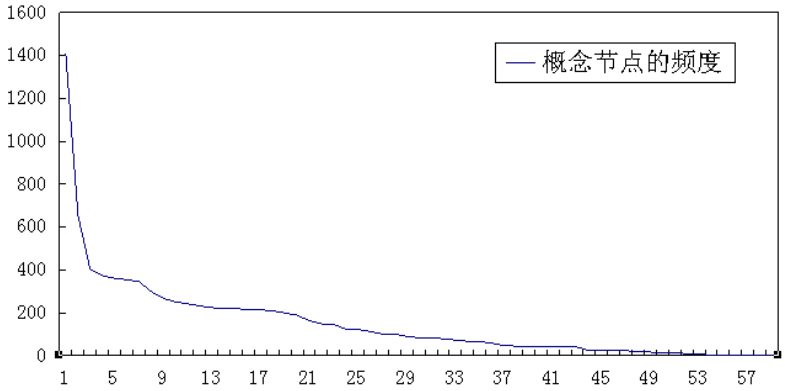


图 3-3 语料中概念节点的频度

Fig.3-3 Frequency of ontology nodes in annotation

而评价词“高”，“大”，“小”的极性不能单独确定，它们的使用属于也就是前面提到的评价搭配，极性需要根据上下文进行判断。这些评价词更多的与一些具体的概念节点相关联。这也暗示了领域本体的信息事实上也含有一定的情感倾向性的先验知识。

在所有257频率高于4的观点中心词中，一共有25个观点词同时存在被标注者标注为正极与负极，即有9.8%的观点词占13.71%的评论的极性判断需要与它搭配的评价对象共同确定，即前面提到的评价搭配现象。标注数据结果说明了对了产品评价文本来说评论搭配的处理的重要性。

3.3 本章小结

本章中提出的一种针对数码相机在线评价文本的细粒度标注策略，并使用这个策略对文本进行了标注工作。标注信息为情感分析特别是细粒度的要素分析提供了资源。

第 4 章 基于依存句法树结构的细粒度情感要素抽取

细粒度情感分析是指从目标文本中获得评价者、被评价对象、评价观点短语，情感极向等更细粒度要素的抽取工作。研究者们提出了基于统计的方法对不同的细粒度要素进行抽取，如评价对象的抽取[48]、评价来源的抽取[13]等。本章提出了一种使用基于依存句法树结构的条件随机场模型对评价对象与评价描述进行结合抽取的方法，该模型改善了线性条件随机场在标注细粒度情感要素时无法适应情感要素长距离语义依赖的问题，使用树边特征表达了细粒度要素中的句法相关性。最后在CUHIT Opinmine语料库与COAE2011任务三数据集对该模型分别进行了实验和评估。

4.1 细粒度情感分析问题定义

具体的评价对象与评价描述的定义可以参看下例。

1. (按键手感) 太硬 而且(声音) 非常大 。
2. 华丽 的(皮草设计) 。

在这些例句中，所有括号中的短语为要抽取的评价对象，所有下划线标注的短语为评价描述。其中评价对象的定义扩展于[48]中的定义：产品的本身、产品的属性、产品的部分以及有联系的概念节点都将作为产品对象。评论短语通常包括了用户对产品评价的核心词。

4.2 条件随机场

在之前的一些研究中[13, 14]，将文本标注的过程运用在细粒度抽取上取得了较好的效果，由于线性的条件随机场模型(Conditional Random Fields, CRFs)在序列文本标注的问题中被广泛适用，一些学者[14, 37, 49]也将细性条件随机场运用在细粒度的抽取工作中并取得了一定效果。但是由于CRFs模型要运算复杂度带来的运算量过高的问题以及数据稀疏等原因[50]，基于CRFs的序列标注会存在无法处理长距离依赖的问题。在细粒度的要素抽取工作中，类似语义关系等的长距离依赖却是感情分析的关键特征。本章将讨论引用树状的CRFs模型来解决长距离依赖问题，通过在本文标注语料库以及第三届中文情感倾向性分析测评电子领域数据集上的实验结果可以看出，基于依存句法树的CRFs模型在细粒度的评价对象，评价描述的挖掘中获得了更好的结果。

线性的条件随机场被Lafferty[50]等人首次用于序列标注问题中。在一个序列标注问题中，输入一串已知的序列$\mathbf{y} = \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$而得到同样大小的序列输出$\mathbf{y} = \mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}$，其中对于输出的每一个$y$，都属于一个有限固定的状态集$S = s_1, s_2, ..., s_k$。在线性的CRFs中，在给定了输入序列后，输出序列的条件概率被定义为，

$$P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp \sum_{t=1}^{T} \sum_{k} \lambda_k f_k(t, \boldsymbol{y}, \boldsymbol{x}) \tag{4-1}$$

其中$Z(.)$为正规化因子，它的作用是确保概率P小于1。$f_k(.)$是二值特征函数，它可以被分为两部分，边特征（edge function）以及点特征（state function）。将$Z(.)$改写为边特征与点特征后，概率等式重写为，

$$P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp \sum_{t=1}^{T} \left(\sum_{k} \lambda_k g_k(t, y_t, \boldsymbol{x}) + \sum_{j} \lambda_j h_j(t, y_{t-1}, y_t, \boldsymbol{x}) \right) \tag{4-2}$$

点特征g只与当前时间点的输出状态y_t相关，而边特征h需要考虑当前时刻以及前一时刻的输出状态y_{t-1}, y_t。权重λ由训练所得。那么，在给定了输入序列后，CRFs模型通过求得一组状态输出$\mathbf{y}$，使得条件概率最大，

$$\boldsymbol{y}^* = \text{argmax}_{\boldsymbol{y}} P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \text{argmax}_{\boldsymbol{y}} \exp\left(\sum \sum \lambda_k f_k \right) \tag{4-3}$$

这组$\mathbf{y}$即为输出。

然而，如果序列中存在长距离依赖，而在式4-2中的边特征却只考虑前一结点的状态，那么长距离的状态依赖向后传递的能力就不强了。在细粒度感情分析中，长距离依赖是一类很重要的特征，图4-1为一个文本的例子。

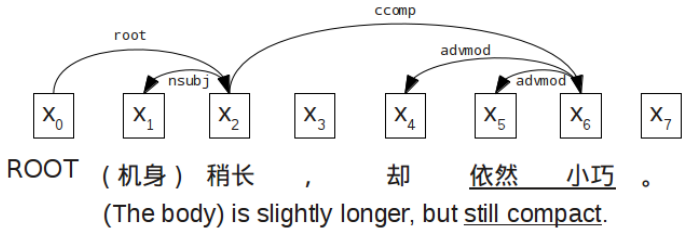


图 4-1 例句的依存分析。其中x_0为虚根节点，弧的标签是依存的类型。

Fig.4-1 x_0 is the virtual root node, the labels of the arcs indicate the types of the dependency.

评价表述的核心词“小巧”与评价对象“机身”线性的距离为5。CRFs模型对节点“小巧”进行标注时，“机身”的标签对它影响不大，而这一影响是与常识不符的，在实验中也显示出线性CRFs性能上的劣势。

4.3 基于依存句法树的条件随机场

为了克服线性条件随机场的缺点，本章提出讨论使用树条件随机场对细粒

度情感要素进行抽取。为了缩短词之间的依赖距离，文本会被预处理为依存树的形式。给定输入以依存句法树结构组织的**x**，输出对应的树结构标签结果**y**，对于每一个y有$y \in S = 'a', 'h', 'v'$，输出**y**的条件概率为，

$$P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{v \in C_1} \sum_k \lambda_k g_k(v, y|_v, \boldsymbol{x}) + \sum_{u,v \in C_2} \sum_j \lambda_j h_j(u, v, y|_{u,v}, \boldsymbol{x})\right) \tag{4-4}$$

C_1为树中的点集合，C_2为树中的边集合。可以看出，与线性CRFs模型不同，树CRFs的边特征考虑当前结点的输出标签以及树结构中的父节点输出标签。由于节点是以依存树的形式进行处理的，所以相关集的语义节点的距离会更接近。

在树结构的条件随机场模型训练过程中，模型会搜索权重段集合$\theta = \{\lambda_1, \lambda_2, \ldots\}$空间来最大化对数形势的似然函数，

$$L = \sum_{t=1}^{T} log\left(p_\theta(y^t|x^t)\right) - \sum_k \frac{2\delta^2}{\lambda_k^2} \tag{4-5}$$

其中函数的第二项是基于权重特征的高斯先验，方差为δ^2，它避免了训练过程中的过拟合。对数似然函数是一个凸函数，树结构CRFs的训练过程可以使用quasi-Newton方法的实现方法：Limited-memory BFGS(L-BFGS)方法来获得最优解。L-BFGS方法与一般的Newton方法不同，在迭代中不需要计算Hessian矩阵，而是通过维护过去几次迭代路径来达到相同的目的，从而找到更优的梯度方向。在求解(Inference)过程中，树CRFs使用Viterbi算法。

4.4 点特征与边特征

树模型中的点特征与线性模型中的点特征基本相同。而边特征却是两种模型的最大差异。边特征是一种对的由同一条边相连的两个输出标签的相关性的描述。从图4-2中可以看出两个模型中边特征的差别。图中变量y_1、y_2与y_6，它们之间有语义相关性，依直觉来说，一个正向的形容词描述一个评价对象的现象更为合理，“机身”被标注为“评价对象”的事件为事件“小巧”被标注为“评价描述”提供很大的可能性，而树CRFs中的边特征是这些相关性的衡量。相比于线性的CRFs模型，边特征却不能很好的表示这样一种相关性。

在CRFs模型的训练过程中，相关性差的二值特征的权重会被逐减小，线性CRFs中的边特征因此存在被忽略的可能性。如果一个线性CRFs模型去除了所有的边特征，即如下式中将所有边特征权重置为0，

$$P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp \sum_{t=1}^{T} \left(\sum_k \lambda_k g_k(t, y_t, \boldsymbol{x})\right) \tag{4-6}$$

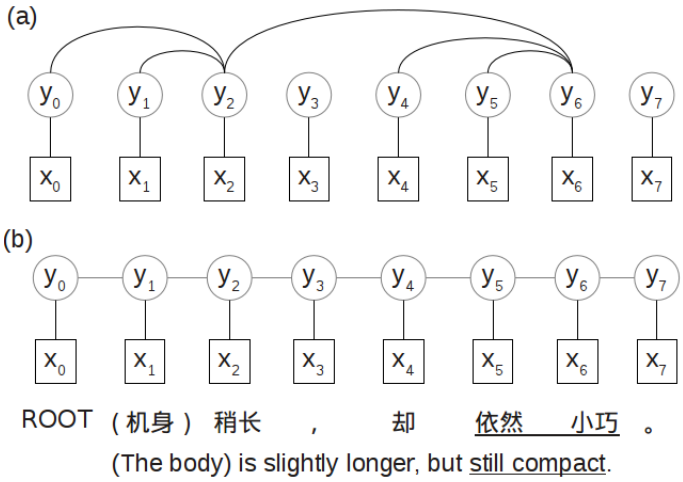


图 4-2 树结构的CRFs与线性CRFs。

Fig.4-2 Tree-structured CRFs and linear-chain CRFs for an opinionated sen- tence.

那么CRFs便退化为了一个逻辑回归模型。它将不再考虑当前时间点以外的其它相关元素的输出标签的信息，这样就丢失了CRFs的一个重要优势，会让细粒度的情感分析丢失重要的语义特征。

基于上面的讨论，为了在实验中比较不同特征对情感分析的影响，以及不同模型的特点，本章中设置了以下四种特征，

- **词法特征（LPF）** 词法特征涵盖了一些以基于词的基本特征，包括特点窗口大小内的POS-tag特征，词本身，例如特征w_t（当前词）、w_{t-1}、$w_{t-1}:w_t$。这些特征都以属于前面提到的点特征。点特征是线性CRFs模型与树CRFs模型共同使用的特征。
- **依存语法特征（DF）** 依存句语法特征包括依存弧序列，如dep_t、$w_{p(t)}{:}dep_t$等，其中$p(t)$是t的父亲节点的下标。DF同样属于点特征。
- **线性边特征（SEF）** 线性边特征在线性CRFs模型中使用，在LPF与DF的基础上，它还考虑序列中所有边所连接的点输出标签。
- **树边特征（TEF）** 树边特征在LPF与DF的基础上，考虑树结构中所有边所连接的点输出标签。

4.5 实验以及讨论

4.5.1 数据集与评测方法

实验使用了两组数据集进行评估。首先是第3章中标注的语料（CUHIT Opinmine），另一组是第三届中文情感倾向性分析测评电子领域数据（COAE）

)[51]。在数据集COAE中没有本体的相关信息，所以在此数据集中进行的实验不会使用本体特征。在预处理过程中，所有的实验都使用了Stanford parser对语料文本进行了依存分析。

在对树结构CRFs模型的抽取实验中，随机选择30%的数据进行训练，剩余的70%的数据进行测试。在测试中，使用两种方法来进行评估，中心匹配（HM）精确匹配（EM）。其中，中心匹配允许抽取的元素与标准答案有差异，当答案的中心词与预测结果的中心词相同时即被判为正确。精确匹配则要求预测与答案的完全匹配。每次评估方法都会用精确度，召回率，F_1值来评估。

4.5.2 Baseline

一些基线方法在实验中用于比较。首先，一个基于字典的方法被实现，它的判定完全取决于训练语料数据所建立的字典，该方法扫描训练语料，记录所有已出现的评价对象与评价短语，分别建立词典，在后续的测试过程中，将每一个词与字典进行查询，直接根据查询结果输出标签。其次，实验实现了一个没有没有边特征的CRFs方法，这样的CRFs会退化为基本的逻辑回归模型，实验中会通过这个方法来比较语义点特征DF的引入对结果的影响。第三个基线方法为一个线性CRFs模型，它与[49]中的方法非常类似。

4.5.3 实验结果以及讨论

表4-1与表4-2文本中标注语料进行的评价对象与评价表述的实验结果。表4-3与表5-2为在COAE数据集上进行实验的结果。表实验表明树结构的CRFs模型获得了比较好的实验结果。基于字典的基线方法获得了HM中较高的精确性，这表明在评价表述中，评价中心词是相对固定的。比较两个无边特征的线性CRFs模型，可以发现DF特征的加入提高了召回率，显示了语义信息的引入提高了细粒度情感分析的性能。比较无边特征CRFs模型与线性CRFs模型，后者在性能上并没有太大提高，反而在少数结果中落后于无边特征的CRFs模型。这也验证了我们之前关于线性边特征表示细粒度情感要素间的尤其是长距离语义相关要素的相关性差的猜测，即线性的边特征带来的影响很小，甚至还有负面的影响。

比较两个不同数据集的结果可以发现，CUHIT Opinmine数据集上的得出的结果普遍比COAE的结果高，暗示了CUHIT Opinmine的数据评论内容更规范。另外，比较CUHIT Opinmine实验结果不难发现，仅靠简单的字典方法，基线方法dictionary-based方法观点表述的F值居然也能达到以60.90%这说明在CUHIT

表 4-1 评价对象抽取的实验结果(CUHIT Opinmine)

Table4-1 Performance on product attributes identification (COAE 2011)

Model		Header Match			Exactly Match		
method	features	precision	recall	F_1	precision	recall	F_1
dictionary-based		88.24	38.61	53.72	56.08	26.46	35.96
non-edge CRFs	LPF	77.92	66.64	71.84	59.33	61.51	60.40
non-edge CRFs	LPF+DF	76.91	73.05	74.93	61.97	63.13	62.55
linear CRFs	LPF+DF+SEF	74.32	74.42	74.37	60.58	62.14	61.35
tree CRFs	LPF+DF+TEF	76.58	73.91	**75.22**	66.90	63.80	**65.32**

表 4-2 观点表达抽取的实验结果(CUHIT Opinmine)

Table4-2 Performance on appraisal expression identification (COAE 2011)

Model		Header Match			Exactly Match		
method	features	precision	recall	F_1	precision	recall	F_1
dictionary-based		83.12	56.91	67.56	70.53	53.59	60.90
non-edge CRFs	LPF	80.20	74.71	77.36	72.59	71.88	72.23
non-edge CRFs	LPF+DF	84.88	76.42	80.42	70.36	73.82	72.05
linear CRFs	LPF+DF+SEF	81.99	82.03	82.01	73.94	73.01	73.47
tree CRFs	LPF+DF+SEF	88.83	79.61	**83.97**	73.85	73.85	**73.85**

Opinmine数据集观点表述的形式是相对固定的。

表 4-3 评价对象抽取的实验结果(COAE 2011)

Table4-3 Performance on product attribute identification (Opinmine)

Model		Header Match			Exactly Match		
method	features	precision	recall	F_1	precision	recall	F_1
dictionary-based		86.56	21.57	34.54	51.49	19.80	28.60
non-edge CRFs	LPF	49.40	27.10	35.00	27.97	23.76	25.69
non-edge CRFs	LPF+DF	57.18	44.42	50.00	35.69	39.09	37.31
linear CRFs	LPF+DF+SEF	56.72	38.79	46.07	47.40	27.79	35.04
tree CRF	LPF+DF+TEF	56.36	46.96	**51.23**	46.72	35.28	**40.20**

4.6 树结构条件随机场工具TCRFs

目前现有的条件随机场的工具中，比较常用的有使用C++语言实现的组件CRF++，日本研究人员开发的同样基于C++语言的FlexCRFs，以及来至马萨诸塞大学的控件mallet的子组件等，但是他们都不符合本研究的要求。前两个

表 4-4 观点表达抽取的实验结果(COAE 2011)

Table4-4 Performance on appraisal expressions identification (Opinmine)

Model		Header Match			Exactly Match		
method	features	precision	recall	F_1	precision	recall	F_1
dictionary-based		81.07	30.50	44.33	24.12	17.06	19.99
non-edge CRFs	LPF	76.79	43.48	55.52	29.83	28.33	29.06
non-edge CRFs	LPF+DF	61.14	64.15	62.61	26.28	38.57	31.26
linear CRFs	LPF+DF+SEF	58.53	68.27	63.02	37.97	30.72	33.96
tree CRFs	LPF+DF+TEF	60.29	70.55	**65.02**	38.02	35.04	**36.47**

工具实现的是线性的CRFs模型，无法满足本研究的要求，而mallet的子组件扩展性不好，为后续进行实验比较带来了困难。

为了解决这个问题，本课题开发了一个可以实现树结构的条件随机场工具，在满足了本次实验的同时，也强调了工具的易用性与可扩展性，希望将其开发为可供其它研究者使用的公共控件。目前此控件的源代码可以通过https://github.com/gagazhn/Segmenter.git下载。

工具提供了二个公用接口。首先，编译接口(Encoder)封装了工具将实例数据文件录入的过程。研究人员在使用工具进行模型的训练或测试时，使用到的数据文件格式可能是多种多样的，当需要使用特定的格式的数据时，研究人员可以根据编译接口实现特定的编译器。

然后，解释接口(Decoder)封装了工具中条件随机场模型进行推演(Inference)过程。树结构随机场模型与线性条件随机场模型的推演过程不同，研究人员可以通过调用不同的实现或者是自己进行实现接口来满足自己研究的需要。

4.7 本章小结

本章使用树CRFs抽取评价对象与评价短语的搭配。研究表明，在评价搭配的抽取中，引入线性的边特征并没有使系统性能提升。与之对应的是，树结构特征却能更好的表达语义信息，缩短了搭配对在CRFs模型中的推演路径。实验表明基于树结构的CRFs模型获得了更好的结果。

第 5 章 基于半监督学习的本体库节点实例扩展

产品评价的文本，特别是类似相机这样的具体领域的评价文本，会有非常多的词典外的领域专有词语出现。在处理这些文本时，无论是最基本的分词，还是其后的细粒度情感分析，都必须要考虑这些重要的字典未登录词。例如相机的评论中，“无敌兔”（Canon 5D II）是一个经常被使用的未登录词，分词程序无法正常对其进行处理，可是“无敌兔”指的是一款经典的数码单反，如果不能正常的识别它，会使细粒度的处理效果大大降低。

为了解决这个问题，本章中提出了一种半监督的概念节点实例的学习方法。它引用人工构建的本体库的各节点，以及少量人工标注的语料，对大量未标注语料进行分析，挖掘出其中的与概念节点相关的词语。

概念节点实例是本体节点在评价文本中可能的表现形式，它是一个短词成为评价对象的必要条件。为了找出这些未登入的实例后选词，方法将分为两部分。首先，模型在无监督的环境下对文本的特征进行为学习，。然后，在已有标注信息的指导下，对前面部分生成的模型进行修证，最后产生本体节点的实例候选。

5.1 基于无监督学习的产品属性新词发现

对目标数据进行无监督的学习可以让模型总结出数据自身的特点，一方面它可以直接产生有意义可直接使用的特征，另一方面，对数据的提前无监督学习可以减少后面的有监督调整的所需要的时间与数据规模。

挖掘新词的传统方法是，先对文本进行分词，然后猜测未能成功匹配的剩余片段就是新词。这似乎陷入了一个怪圈：分词的准确性本身就依赖于词库的完整性，如果词库中根本没有新词，这又降低了分词结果的可靠性。此时，一种大胆的想法是，首先不依赖于任何已有的词库，仅仅根据词的共同特征，将一段大规模语料中可能成词的文本片段全部提取出来，不管它是新词还是旧词。然后，再把所有抽出来的词和已有词库进行比较，用这样的方法来找到新词。

Sproat等人[52]首先将引入互信息来处理分词问题。Peng等人[53]使用了最大熵的方法联合一些基于素质信息的裁剪的策略来进行分词。一些无监督的分词系统基于很简单的原理，比如Cohen等人[54]研究的Voting experts中使用到的分词方法，以及Wang在[55]中的方法对本章中的方法较有启迪。然而已标注的数据对

结果的影响还是很强的，特别是在领域中已经存在有人工构建的本体信息的情况下，对文本进行半监督学习一方面可以发现文本自身所存在数据特点，另一方面又可以根据人工对领域本体节点的总结而生成更高层次的数据表示。

本节中提出了一种半监督的新发现方法，如图5-1，处理目标文本时，采用迭代的方法进行处理，然后根据结果与人工的小量标注进行比较评分，如果迭代没有达到停止条件，就进行新的一次迭代，否则输出结果。

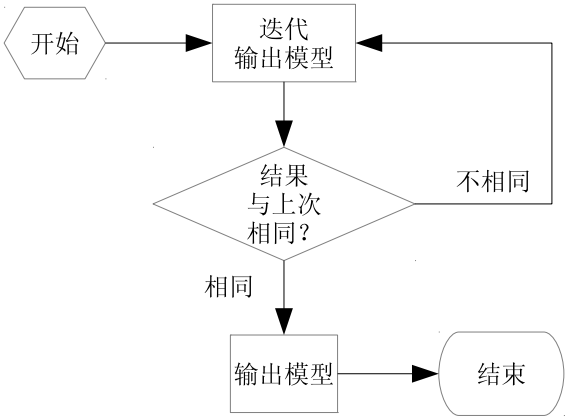


图 5-1 新词发现流程图

Fig.5-1 Flowchart of new word mining

无监督的分词基于二个观点。首先，词的确定是由所组成的字符的结合的确定性与分离的不确定性决定的；然后，通过已有的少量的标注文本可以对当前的分词模型进行修定，而生成更好的模型。

待处理的文本可以看作是一个字序列集合。对于任意一个长度（字数）大于1的字序列，它可以分裂为两个相邻的子序列。而序列是否需要分为两个子序列是由上面提到的子序列结合的确定性与分享的不确定性共同决定的。

另一方面，对于任意一个字序列，它都可以以不同的分裂方案分为子序列。例如如序列“ABC”，

A_BCD

AB_CD

ABC_D

可见对于一个长度为L的序列，如果将“不分裂”也看作为分裂的一种方案，那么它一共有N种分裂方式。无监督分词的过程就是一系列关于字序列的分裂方案是否执行的决策。在给定输入字序列后，依据序列之间的结合的确定性与分开的不确定性，来获得最终的分裂方案。

5.1.1 分词决策过程

在每一次迭代过程中，已知当前的所有参数后，在给定输入文本序列**x**后，使用一个评分函数来联系之前所述的确定性与不确定性，定义为，

$$f(S) = \prod_{t=1}^{T} \mathrm{PV}(s_t) * \prod_{t=1}^{T-1} \mathrm{SPV}(s_t, s_{t+1}) \tag{5-1}$$

其中**S**为分词策略，它的子元素s_t是对输入**x**进行分词后按照顺序进行排序的词的序列。一个字序列会对应一个非常大的所有分词策略的集合：**S** = 该序列所有的s，如图5-2所示，图中的所有节点都是一个分词策略，空格表示分隔符。箭头指的是子序列的再分裂过程，序列分裂后，可以递归地进行再分。但是图中“A_BC”与“AB_C”都再分裂为了“A_B_C”，而分词策略的得分与它是由哪个父状态分裂而来无关。

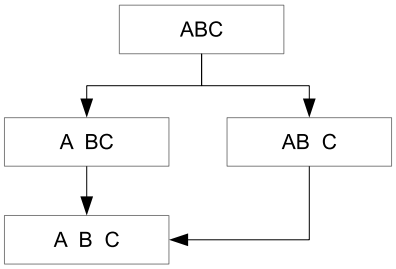


图 5-2 字序列所有的分词策略

Fig.5-2 Enumeration of segmenting strategies

函数PV是对词的结合确定性的评价，SPV则是对两个相邻分词t与$t+1$分离不确定性的评价。

PV值衡量词结合确定性，它的定义如下，

$$\mathrm{PV}(x) = (\frac{\mathrm{F}_x}{\mathrm{MF}_L})^L \tag{5-2}$$

其中F_x为序列x在文本中做为独立的词出现的频率，MF_L是长度为L的所有序列在文本中平均频率。在第一次迭代过程中，由于缺乏独立词的信息，F_x都取1。PV的值反映了词组的结合的强度，PV值与SPV值的结合可以用来评价整个句子或整个文本集合的分词得分。SPV的定义如下，

$$\mathrm{SPV}(x, y) = \frac{\mathrm{LE}(x) * \mathrm{RE}(x)}{\mathrm{MRE}_{L1} * \mathrm{MLE}_{L2}} \tag{5-3}$$

其中LE与RE分别是序列x的右熵与左熵，而$\mathbf{MRE}_L$与$\mathbf{MLE}_L$为所有长度为L的序列所有左熵以及右熵的平均值。

在一次迭代中，分词的输出就是求出让S最大化的分词集Ex即，

$$S_x = \text{argmax} f(S_x) \tag{5-4}$$

可是，文本可能的分词策略的数量是随着文本长快速增长的，所以不能采用枚举的办法来遍历所有分词策略。由于整个文本的评估过程中会重用很多子过程，所以可以使用动态归化的办法来降低复杂度。整个推导类似Viterbi推导，在此过程中，需要预先设定一个参数：分词的最大长度。它对算法带来的影响后面会提到。由于中文词长一般来说不会超过6个，本章中设定分词的最大长度为6。于是在给定一个最大词长后，使用一个前向数组保存子序列的分词评分，如图5-3所示，其中数组的高度由分词的最大长度决定，宽度由字序列的长度决定。数组的建立过程也是推导过程。数组自上而下，从左向右处理节点，将到达此路径点的最大路径记录下来。

考虑图中的节点“CD”，当在计算到达它的最大路径时，遍历它所有的前继节点，将前继节点所储存的值与LPV(前继节点,CD)相乘，最终得到最大的值并记录相应的前继节点，在整个前向数组建立完成后，路径也计算完毕了，此时只需要在所有的终结节点（含有最后一个字的序列节点）的储存值取出最大的，此时相对应的路径便是分词的结果。图中显示了一条分词路径“A-B-CD-E”。那么，整个推导过程可以在$O(N * L^2)$的时间复杂度内完成一次迭代。

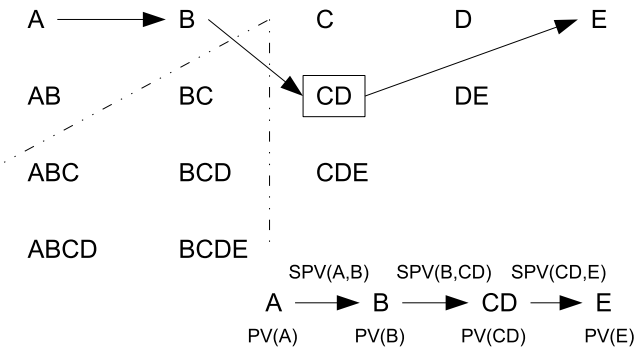


图 5-3 前向数组，以及分词策略“A B CD E”的路径

Fig.5-3 Forward array and path of segment strategy

见算法5.1，在算法某个句子时，算法的主要过程就是在创建前向数组，在考虑数组中每一个单元时，搜索它的所有前继节点以获得到当前点的最大评分，并记录前继节点的序号。当前向数组建满后，找到所有路径的最大评分路径，通过回溯的方法求得分词结果，交将这个结果插入分词结果S中。

Require: Set $X = x$; Foward[][]; Path[][]; F; S
Ensure:
  for all $x \in X$ **do**
    for $i \Leftarrow 1toL$ **do**
      for $j \Leftarrow 1toMAX_LENGTH$ **do**
        $Foward \Leftarrow max_score$
        $Path \Leftarrow max_path$
      end for
    end for
    $p \Leftarrow max \in Path$
    $S inserts(p)$
  end for
  update F

算法 5.1: 权值更新迭代过程

5.1.2 权值更新迭代过程

在一次迭代后，算法得到了一个让分词评分最大的分词策略。此时，更新的过程就是根据瓣的分词结果重新计算分词频率F_x的过程。例如当序列“ABCAB”的上一次分词结果为“AB C AB”时，词频统计为“AB:2,C:1”。当新的一轮迭代分词结果变为“AB ABC”时，那么对应的词频统计将变为“AB:1,ABC:1”。

模型中PV值是以指数的形势计算的，并且指数是文本段的长度。这样就确保了字符与多字符段的PV值在整个分词评分中的重要性是等价的。例如文本ABC，它可以有四种分词策略，分别为ABC、AB_C、A_BC以及A_B_C。此时他们的的PV值的乘积分别为，

$$\frac{F_{ABC}}{MF_3} \times \frac{F_{ABC}}{MF_3} \times \frac{F_{ABC}}{MF_3}$$
$$\frac{F_{AB}}{MF_2} \times \frac{F_{AB}}{MF_2} \times \frac{F_C}{MF_1}$$
$$\frac{F_A}{MF_1} \times \frac{F_{BC}}{MF_2} \times \frac{F_{BC}}{MF_2}$$
$$\frac{F_A}{MF_1} \times \frac{F_B}{MF_1} \times \frac{F_C}{MF_1}$$

不同的分词策略中，每段的PV值都被等价的考虑到了，那么在分词策略的总体

分词的评分中，不会有对细分或粗分词的倾向。

频率信息更新后，即可以进行下一次迭代。而迭代的终止条件是当新的迭代产生的分词策略的分词结果与上一次结题相同。在迭代过程停止后，得到了对目标文本的初始的分词策略。

5.2 基于半监督学习的实例扩展

在学习了文本内在的分词特征后，一下步的是对模型进行有半监督学习。无论是人工总结的模板，还是文本的标注信息，都可以为进一步优化模型的而提供有一定指导学习。在半监督的参数调整过程中，采用迭代的方法，根据模型的评价结果进行调整，直到达到停止条件，过程与无监督的训练过程相似，如图5-4所示。

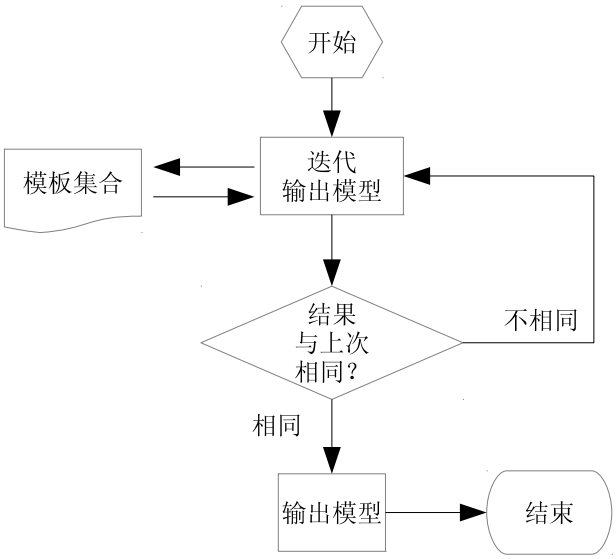


图 5-4 根据评价对象抽取模板进行调整过程

Fig.5-4 Flowchart of updating based on pattern

5.2.1 评价对象抽取模板

抽取模板在算法中有两个作用。首先，就像其它基于模板的属性新词学习中模板的作用一样，它通过匹配文本，一旦匹配成功便将相应的要素抽取出来。另外，本章中模板的还具有指导模型进一步优化分词的作用。制定模板的过程也是一个对文本特征进行总结泛化的过程。

本章中使用上下文模板，定义形式如下,

$$\underline{W_{-2}}\ W_{-1}\ [W_0],$$
$$W_{-1}\ [W_0]\ \underline{W_1},$$
$$[W_0]\ \text{Neg}\ W_2,$$
$$[W_0]\ W_1\ \text{Modi}\ \underline{W_3},$$
$$[W_{-1}]\ W_0\ [W_1]\ \underline{W_2}$$

其中W代表需要匹配的具体词素，小标表示相对位置，中括号标注的词素为抽取的目标本体实例。模型中也包括通配符匹配，如例子中的Neg（否定词）与Modi（程度词）。一些定义中被加下划线的符号的被称之为得分点，得分点在模型调整过程中参与评估模板质量的运算。实验训练中总结使用了19个模板。

5.2.2 基于模板指导的模型调整与本体实例的抽取

前面提到过，模板具有指导模型进行分词的作用。对于某个模板p，当模型输出的分词结果与模板p匹配的事件，本章中称之为模板p的命中。一个好的模板能准确抽取本体实例的同时，也能将其它所指定的得分点准确抽取，即命中的次数高。

基于这样的假设，本章使用迭代的方法，来调整每个模板的权重，达到最佳的分词效果，此时就认为模型对本体实例也有了最佳的抽取效果。

在基于模板指导的调整过程中，分词结果的评分重写为，

$$f'(S) = \prod_{t=1}^{T} \text{PV}(s_t) \times \prod_{t=1}^{T-1} \text{SPV}(s_t, s_{t+1}) \times \prod_{i=1}^{P} \lambda_t h(p_i, S) \tag{5-5}$$

与式5-1不同，分词评分添加了模型调整的的部分。P为模板的总数，函数$h(p,S)$评估模板p的得分，定义为下，

$$h(p,S) = \prod_{o \in O(p,S)} e^{L(o)} \tag{5-6}$$

其中$O(p,S)$为所有被模板p命中后，被成功识别以及模板设置的得分点的已标注词。可以看出，式5-5中的调整部分是通过对模板的引入而加强了模板相关词的PV值。

模型在分词过程中，模型求得让分词评分最大的分词策略，

$$S' = \text{argmax} f'(S'_x) \tag{5-7}$$

与5-4不同，新的求解过程中，模型需要考虑是否需要让模板命中。如图5-5所示，模板

[*] B CD

由于有命中路径“A_B_CD”的潜力，在构建向前数组的节点“CD”时，“CD”节点在遍历比较与前继节点的乘积时，模板会加大前继节点“B”的值，“CD”节点便更可能选择前继节点“B”作为路径前继节点，即模板指导了分词。当分词模型最终采用的分词策略匹配了模板p时，这时p就命中了一次。

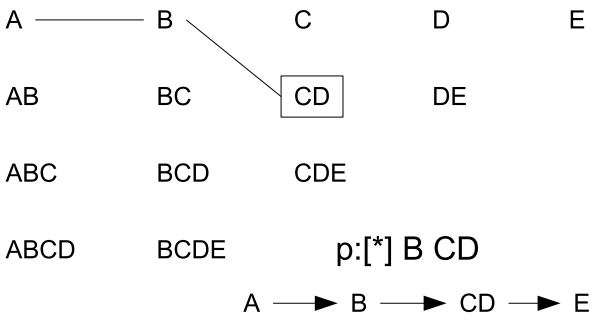


图 5-5 模板对分词的影响

Fig.5-5 The influence of the pattern to the segmentation

5.2.3 抽取模板权重的更新

一次的迭代更新中，一方面分词策略S会被不断更新，表现为模型中独立分词频率F_x的更新。另一方面，需要根据引入模板p_i而带来的影响，更新与其对应的权重λ_i，

$$\lambda_i(t) = \lambda_i(t-1) + \left(\sum_{o \in O} (\frac{F_o(t) - F_o(t-1)}{MF_L})^L \right) \tag{5-8}$$

其中$F_o(t)$第t次迭代时词o的词频。可以看出模板权重更新的策略是，查看那些模板在命中时，对已有的标注语料所标注的词素区能力的强弱来调整模板的权重。如果模板能很好的识别词素，就加大该模板的权重，让分词模型更倾于让该模板命中。反之便要降低权重。

迭代的终止条件是当新的迭代产生的分词策略的分词结果与上一次结果相同。在模型训练完毕后，模型将所有命中模板抽取出来的词作为新实例，然后进行输出。

5.3 实验以及讨论

5.3.1 产品属性新词发现模块分词性能的评估

属性新词发现模块本质是一个无监督的分词系统，但是对它的评估方法完全套用分词系统的评估方法是不合实际的。这是因为模块的最终目的不是输出整个分词策略，而是生成具有特定性质的词段集。属性新词发现模块主要的任务是要找出产品评价文本中评价对象的边界，根据这一任务我们设计了评估属性新词发现模块性能的实验。

实验评估新词发现模块的分词结果与细粒度情感语料中带有的边界信息匹配的结果。举一个例子，情感语料中对主观文本“电池很不耐用”的标注结果为：

```
CommentContent="电池很不耐用"
modifier="很"
negation="不"
opinion_keyword="耐用"
opinion_expression="很不耐用"
attr_expression="电池"
```

通过提取标注信息中的细粒度的情感要素，包括评价修辞（modifier）、否定词（negation）、评价中心词（opinion_keyword）以及产品对象（attr_expression）就可以得到文本中的边界信息：“电池_很_不_耐用”。对属性新词发现模型进行评估就是计算属性新词模型的分词结果与标注结果的边界的匹配结果。

实验中将CUHIT Opinmine的7611行评论文本随机分析为两份：记为D_1与D_2，另外还准备了COAE2011中随机抽取的5000句电子产品评价文本D_3进行无监督训练。在不同的数据集中训练，最后将CUHIT Opinmine标注的文本中有细粒度标注的句子进行边界对比。

前文提到过抽取模板除了具有评价对象抽取的作用以外，还具有进一步指导模块进一步优化分词的作用。为了评估模板的优化作用，本实验中将由$D_1 + D_2 + D_3$训练生成的属性新词模块作为输入，使用抽取模板调整后的结果也进行了对比。

经过迭代后的结果如表5-1如示。

可以发现，无监督的属性新词发现模块与训练数据的规模相关性很大，并且从第三行与第四行结果中可以发现，COAE2011电子产品评论文本的加入也

表 5-1 基于细粒度情感标注语料边界信息的属性新词发现模块评估

Table5-1 Performance of segmenter module

dataset	precision
D_1	51.21
$D_1 + D_2$	72.54
$D_1 + D_3$	68.60
$D_1 + D_2 + D_3$	78.19
$D_1 + D_3$ + adjust	76.33
$D_1 + D_2 + D_3$ + adjust	83.79

提升了分词效果。COAE2011中的文本与CUHIT Opinmine评论文本有一定的领域差异，但是这个差异在实验结果中并不明显，这说明模块具有一定的领域通用性。

通过比较第三行与第五行的结果以及比较第四行与第五行的结果可以发现，抽取模板的引入进一步提高了分词结果的边界匹配结果，验证了前面我们的假设。

5.3.2 概念节点实例扩展的性能评估

无监督的学习过程中，使用所有的语料进行学习，得到的初始的分词模型后，随机选择30%的数据进行训练，在实验中，使用了28个模板指导模型的更新，模型的得分点包括语料标注的产品对象元素，观点中心词元素，否定词元素以及程度词元素。模型产生实例的候选词后，为了评估候选词的可靠性与选择策略的关系，使用选择函数$r(w) = \sum_p c(w,p)\lambda_p$对候选词$w$进行排序，其中$c(w,p)$是候选词被模板$p$抽取的次数，$\lambda_p$为模板$p$的权重，将对在不同选择条件下的候选词进行评估比较，具体会用节点实例预测的精确度、召回率和F_1值来进行评估。

图5-6实验结果。实验测试数据中所有频度大于4的节点实例为288个，依据选择函数对预测词排序，实验结果图中记录了选择TOP-N的实验结果。其中X坐标为TOP-N的N值。当选择前90词时，结果的正确率最高。由于采用选择TOP-N的形式来评估结果，这也造成了召回率从一个很低的值逐步上升。模型输出结果可以为后续的研究提供支持。

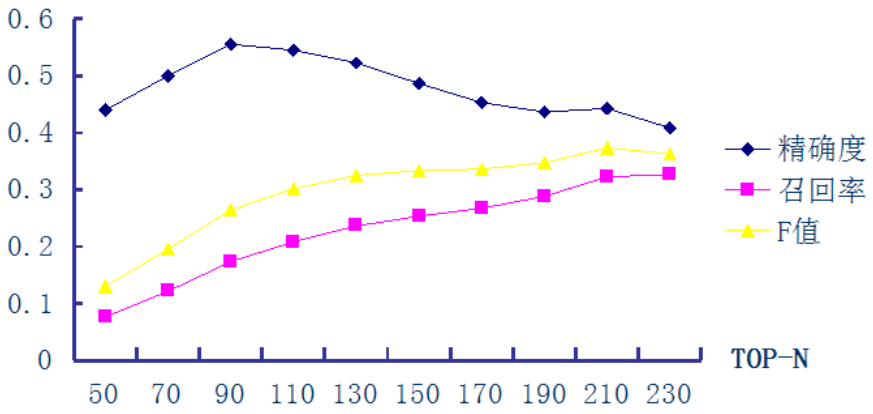


图 5-6 模板对分词的影响

Fig.5-6 The influence of the pattern to the segmentation

5.3.3 领域本体节点属性新词对细粒度情感分析的影响

本章中实验设置与第4章中实验基础上进行了比较设置。产生的新本体节点实例一方面可以做为概念节点的候选实例词，以二值特征的形式做为细粒度产品评价分析中评价对象抽取的特征使用；另一方面，输出的候选词做为分词系统的用户自定义词典，以加强分词的分词的准确性。实验中将200个后选词录入了的自定义词典，词性设置为名词。

实验在数据中进行：CUHIT Opinmine，实验随机选择30%的数据进行训练，剩余的70%的数据进行测试。在测试中，只使用精确匹配（EM）。每次评估方法都会用精确度，召回率，F_1值来评估。

表 5-2 引入本体节点实例候选特征的评价对象抽取实验结果(Opinmine)

Table5-2 Performance on opinion target identification (Opinmine)

Candidate Num	Precision	recall	F_1
0	66.90	63.80	65.32
50	67.22	67.83	67.52
100	67.96	69.78	68.86
150	67.41	70.01	68.69
200	66.45	68.24	67.33

最高F值出现在第三行，使用了前100个候选词。实验显示实例候选词特征的引入明显提高了评价对象的召回率，其中最高的召回率出现在第四行结果中。但是并不是候选词使用越多召回率越高，从第五行数据可以看出，它的召

回率较第四行有回落，这是因为由于过多噪音的引入，后选词特征在训练过程中逐渐被CRFs模型忽略。

5.4 细粒度产品评价展示程序

为了更直观的呈现细粒度产品评价分析的相关结果，我们搭建了一个基于HTML5技术的展示系统。系统以相机型号为基本单元，组织与该型号相关的属性，也就是前面提到了评价对象，以及与之对应的评价表述等细粒度性感信息。

5.4.1 关联评价对象实例与领域本体概念节点

程序的数据来源于前节提出的模型出输，而将模型输出中的评价对象实例与本体的概念节点的关联问题看作为一个标准有监督的多元分类问题。通过使用K近邻（KNN）的方法，将每一个实例与领域本体中的第三级节点进行关联。对于给定的评价对象实例w，我们使用点式互信息（PMI）来描述它与领域文体节点的相似度。即，

$$\mathrm{PMI}(w_1, w_2) = \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{5-9}$$

通过比较待分析实例与标注数据中已知实例的点信息来对其进行分类。对模型输出的所有的节点进行分类后，对所有节点的相关信息进行统计：首先是统计节点被使用的次数，因为实例只与本体中三级概念节点相关联，所以三级节点的计数首先加一。然后该节点的所有父节点次数同时加一。另外，与上一点类似，需要统计三级节点以及父节点的极性进行统计。

5.4.2 数据源与数据的处理方式

程序数据来源于CUHIT Opinmine的处理结果。以线下处理的方式对细粒度处理的结果进行批处理，保存在结果当中。在所有的评价对象实例统计完毕后，提升可视化效果，将更好的结果优显示，所以需要对所有的输出结果进行可靠性排序，可靠性的计算依据第5章中最终计算的分词总评价进行排序，在给定输入文本$\mathbf{x}$与分词评分y后，句子的可靠性，

$$\mathrm{score}(\mathbf{x}) = \mathbf{y}^{\frac{1}{L}} \tag{5-10}$$

其中L为句子的长度。对所有的数据进行线下处理后，就形成了可视化界面的核心数据。

5.4.3 界面显示

在得到所有的数据源后，程序以线下的方式将这些数据显示出来。界面采用目录索引的方式，将用户的点击引入结果显示页面。

界面包括两个主要部分：相机型号列表与相机评价信息列表。相机型号列表如图5-7所示，图中列出了所有的相机型号，这里的型号信息都来至在线的已有信息。界面的上方是根据各品牌分类，用户选择某一品牌时，会列出品牌下的所有型号。

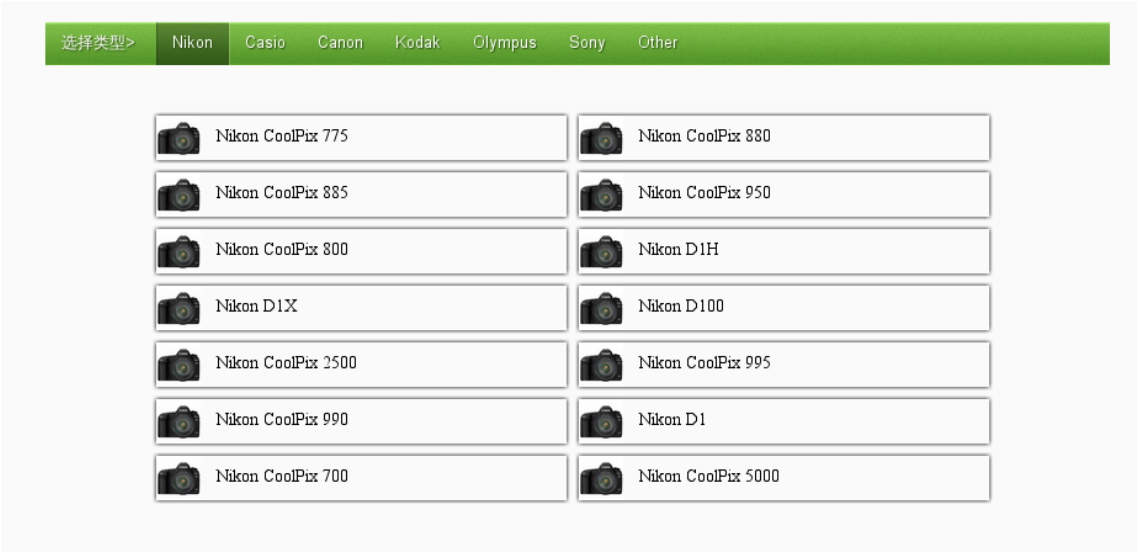


图 5-7 DC的型号列表界面

Fig.5-7 View of the DC list

用户单击某一个相机后，就会转入界面的第二部分，即相机评价信息。如图5-8所示，界面会以时间轴的形式对当前型号的相机本体概念中的第二级节点的所有属性按被评论次数进行排序，及被评论次数多的属性会被显示在时间轴的左侧，所有属性节点都可以通过托拽来移动位，以便居中所需的属性节点。属性节点的下方显示出评论次数，本体库中的没有被评论过的节点将不矛显示。

用户点击具体的属性时，如图5-9，会显示该二级节点属性的相关的所有评论句。在文本中会用不同颜色标记出不同类型的细粒度情感要素。另外，为了提高显示的用户体验，会按前面提到的方法对文本进行排序，将高评分的文本置前。

整个界面使用HTML5的技术进行实现，并使用jQuery等多个兼容HTML5.0的三方JavaScript库进行开发。使用HTML5规范进行构建的网页应

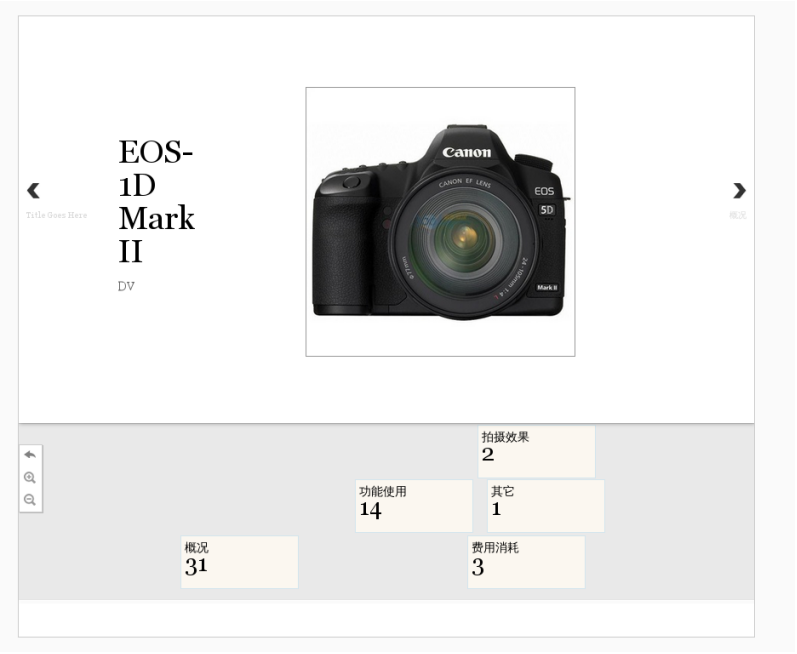


图 5-8 DC的型号属性界面

Fig.5-8 View of the DC list

图 5-9 DC的属性详细页面

Fig.5-9 View of the DC list

用可以支持大多数当前的浏览器，不需要使用Flash就可以实现高品质的动画效果。例如图片元素圆滑的移动、背景的渐入渐出的效果，以及透明效果等。

5.5 本章小结

本章中介绍了一种半监督的本体概念实例的学习方法。它通过一次无监督的分词学习，首先对目标领域的文本进行了自学习，总结出了文本的自身特征。然后基于人工建立的模板以及少量标注信息来对模型参数进行半监督的调

整，来进行对新实例的学习。由于使用了半监督的学习方法，模型对标注数据的数量要求较少，实验表明该方法是有效的。最后，本章的模型输出的结果将作为第4章的特征。

结　论

本文中针对产品评价的细粒度分析需要，首先提出了产品评价的语料标注体系，该体系针对产品评价文本建立其特有的领域本体库，通过领域库的形式维护和组织产品的相关概念。体系除了要求对一般的产品评价细粒度情感要素进行标注以外，还需要标注产品评价中的多评论问题。依据该体系组织研究者对1000短篇相机的产品评论进行了标注，得到了一套开放语料，通过语料可以发现产品评价文中出现单词对应多本体结点的现象、评价中心词在不同的评价搭配对中呈现不同倾向性等现象，为细粒度的产品评价情感分析提供数据支持。

第二、课题还提出了一种基于依存句法树结构的条件随机场模型，该模型为了解决线性条件随机场在标注细粒度情感要素时无法适应情感要素长距离语义依赖的问题，使用树边特征表达了细粒度要素中的语义相关性。实验显示使用树边能更好的表示评价文本的语义的相关性，基于依存句法树结构的条件随机获得了更好的性能。

第三、本文还讨论了一种半监督的学习本体节点新实例的方法。由于在相机等特定领域的产品评价文本中，有大量未词典未登入的领域特有词，一般的分词系统很难对这些词进行区分。方法首先通过无监督策略迭代得到初始的分词模型，然后通过人工制定的模板对分词模型进行修订并同时依据修订效果来调整该模板的权重，而权重会影响到最后对属性的抽取，对预先构建好的本体节点的实例进行扩展，为在后续的研究里提供了特征支持。

本文的贡献主要有三方面，一方面，标注的产品评价细粒度语料为后续的情感分析提供数据支持；另一方面，通过基于依存句法树结构的条件随机场模型的方法的实验验证了使用树边能更好的表示评价文本中情感要素间的相关性，而线性条件随机场模型的边特征在细粒度情感要素抽取任务里无法描述相关性；最后，课题提出了加强产品评价细粒度情感分析中对词典未登入领域专有词的识别方法，实验证明该方法能显著提高评价对象识别的召回率。

参考文献

[1] 姚天昉, 程希文, 徐飞玉, 等. 文本意见挖掘综述[J]. 2008:71–80.

[2] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 2010, 21(8):1834–1848.

[3] P. D. Turney. Thumbs up Or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews[C]//Proceeding ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002:417–424.

[4] L. L. Pang, 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.

[5] V. Hatzivassiloglo, K. R. McKeown. Predicting the Semantic Orientation of Adjectives[C]//Proceeding ACL '98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1998:174–181.

[6] K. F. W. R. F. Xu, Y. Q. Xia. Opinmine - Opinion Analysis System by Cuhk for Ntcir-6 Pilot Task[C]//Proceedings of NTCIR-6, 2007.

[7] B. L. J. B. Qiu, Guang, C. Chen. Expanding Domain Sentiment Lexicon Through Double Propagation[C]//Proceedings of the 21st international jont conference on Artifical intelligence, 2009:1199–1204.

[8] V. Hatzivassiloglou, J. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity[C]//Proceedings of the International Conference on Computational Linguistics (COLING), 2000:299–305.

[9] e. a. N. Kobayashi, K. Inui. Extracting Aspect-evaluation and Aspect-of Relations in Opinion Mining[C]//Proceedings of EMNLP-CONLL, 2007:1065–1074.

[10] Q. Z. Y. Wu. Structural Opinion Mining for Graph-based Sentiment Representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011:1332–1341.

[11] R. B. Jeonghee Yi, Tetsuya Nasukawa, W. Niblack. Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques[C]//Data Mining, 2003. ICDM 2003. Third IEEE International Conference, 2003:427–434.

[12] N. G. Kenneth Bloom, S. Argamon. Extracting Appraisal Expressions[J]. 2007:308–315.

[13] E. R. Yejin Choi, Claire Cardie, S. Patwardhan. Identifying Sources of Opinions with Conditional Random
elds and Extraction Patterns[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005:355–362.

[14] C. C. Eric Breck, Yejin Choi. Identifying Expressions of Opinion in Context[C]//Proceedings of the 20th international joint conference on Artifical intelligence, 2007:2683–2688.

[15] T. W. Claire Cardie, Janyce Wiebe, D. Litman. Combining Low-level and Summary Representations of Opinions for Multi-perspective Question Answering[C]//Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, 2003:20–27.

[16] S. Das, M. Chen. Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards[C]//Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.

[17] S. L. Kushal Dave, D. M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classi cation of Product Reviews[C]//In Proceedings of WWW, 2003:519—528.

[18] H. Yu, V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[C]//Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003:129–136.

[19] S.-M. Kim, E. Hovy. Automatic Detection of Opinion Bearing Words and Sentences[C]//Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2005:61–66.

[20] B. Pang, L. Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales[C]//Proceeding of the Association for Computational Linguistics, 2005:115.

[21] 樊娜, 蔡皖东, 赵煜, 等. 中文文本情感主题句分析与提取研究[J]. 2006, 29(4):1171–1176.

[22] B. Pang, L. Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004.

[23] T. Kudo, Y. Matsumoto. A Boosting Algorithm for Classification of Semistructured Text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004, 4.

[24] E. Riloff, J. Wiebe. Learning Extraction Patterns for Subjective Expressions[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003:105–112.

[25] R. B. M. B. J. M. Wiebe, T. Wilson, M. Martin. Learning Subjective Language[J]. 2004, 30:277—308.

[26] C. K. S. C. J.-C. Na, H. Sui, Y. Zhou. Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews[J]. 2004, 9:49–54.

[27] A. Kennedy, D. Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters[J]. 2006, 22:110–125.

[28] B. Pang, L. Lee. Opinion Mining and Sentiment Analysis[M]. Foundations and Trends in Information Retrieval, 2008.

[29] K.-F. Xu R., Xia Y.Wong. Opinion Annotation in Online Chinese Product Reviews[C]//Proceedings of LREC, 2008.

[30] M. Hu, B. Lu. Mining and Summarizing Customer Reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004:168–177.

[31] A.-M. Popescu, O. Etzioni. Extracting Product Features and Opinions from Reviews[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005:339–346.

[32] J. G. S. G. Grefenstette, Y. Qu, D. A. Evans. Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application[M]. in Proceedings of Recherche d' Information Assiste par Ordinateur (RIAO), 2004:186–194.

[33] S. C.-O. Michael Gamon, Anthony Aue, E. Ringger. Pulse: Mining Customer Opinions from Free Text[C]//Proceedings of the International Symposium on Intelligent Data Analysis (IDA), number 3646 in Lecture Notes in Computer Science, 2005:121–132.

[34] S. S.-K. J. W. J. Wilson T, Hoffmann P. Opinionfinder: A System for Subjectivity Analysis[C]//Proceedings of HLT/EMNLP 2005 Demonstration abstracts, 2005:34–35.

[35] Y. M.-K. T. Nozomi Kobayashi, Kentaro Inui, T. Fukushima. Collecting Evaluative Expressions for Opinion Extraction[J]. 2004:596–605.

[36] A. T.-V. H. S. Bethard, H. Yu, D. Jurafsky. Automatic Extraction of Opinion Propositions and Their Holders[C]//2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text, 2004:2224.

[37] S. Shariaty, S. Moghaddam. Fine-grained Opinion Mining Using Conditional Random Fields[C]//Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference, 2011:109–114.

[38] C. C. Wiebe J, Wilson T. Annotating Expressions of Opinions and Emotions in Language[J]. 2007:316–322.

[39] T. Y.-X. L. Hongbo Xu, Le Sun. Overview of Chinese Opinion Analysis Evaluation 2011[C]//In Proceedings of COAE 2011 workshop, 2011.

[40] E. Agichtein, L. Gravano. Snowball: Extracting Relations from Large Plain-text Collections[C]//Proceedings of the 5th ACM International Conference on Digital Libraries, 2000:85–94.

[41] D. D. Oren Etzioni, Michael Cafarella, S. Kok. Web-scale Information Extraction in Knowitall[C]//Proceeding of the 13th international conference on World Wide Web, 2004:100–110.

[42] M. Thelen, E. Riloff. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 2002:214–221.

[43] R. P. Using Information Content to Evaluate Semantic Similarity[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995:448–453.

[44] C. D. Jiang J. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy[M]. In Proceeding of International Conference on Research on Computational Linguistics, 1997.

[45] L. Zhao, C. Li. Ontology Based Opinion Mining for Movie Reviews[C]//Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management, 2009:204–214.

[46] J. A. G. Wei Wei. 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.

[47] L.-Y. L. Lun-Wei Ku, Tung-Ho Wu, H.-H. Chen. Construction of an Evaluation Corpus for Opinion Extraction[J]. 2005:513–520.

[48] A.-M. Popescu, O. Etzioni. Extracting Product Features and Opinions from Reviews[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005:339–346.

[49] I. G. Niklas Jakob. Extracting Opinion Targets in a Single- and Cross-domain Setting with Conditional Random Fields[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010:1035–1045.

[50] F. C. P. John Lafferty, Andrew McCallum. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[J]. 2001.

[51] T. Y. X. L. Hongbo Xu, Le Sun. Overview of Chinese Opinion Analysis Evaluation 2011[C]//Proceedings of COAE 2011 workshop, 2011.

[52] R. Sproat, C. Shih. A Statistical Method for Finding Word Boundaries in Chinese Text[C]//Computer Processing of Chinese and Oriental Languages, 1990, 4:336––351.

[53] F. Peng, D. Schuurmans. Self-supervised Chinese Word Segmentation[C]//Proceedings of the Fourth International Symposium on Intelligent Data Analysis, 2001:238–247.

[54] N. A. Cohen, Paul, B. Heeringa. Voting Experts: An Unsupervised Algorithm for Segmenting Sequences[J]. 2007, 11:607—625.

[55] S. T. Hanshi Wang, Jian Zhu, X. Fan. A New Unsupervised Approach to Word Segmentation[J]. 2011, 37:421–454.

攻读硕士学位期间发表的论文及其他成果

（一）发表的学术论文

1 **Yue Zhang**, Ruifeng Xu, Jun Xu. Joint Extraction of Product Attribute and Appraisal Expression Using Tree-Structured Conditional Random Fields. To appear in Joural of Computational Information Systems (EI indexed). Vol. 9, 2013.

2 徐睿峰，王亚伟，徐军，**张玥**，郑海清，桂林，叶璐. 基于多知识源融合和多分类器表决的中文观点分析[C]. 第三届中文倾向性分析评测论文集. 北京：中国中文信息学会，2011：77-87.

第 A 章 附录

A.1 细粒度情感分析语料标注程序

图 A-1 语料标注器界面

Fig. A-1 A view of annotator

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《面向产品评价的细粒度情感分析技术研究》，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。

作者签名：  日期： 2013年 1月 11日

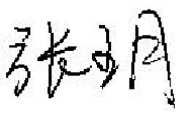
哈尔滨工业大学硕士学位论文使用授权书

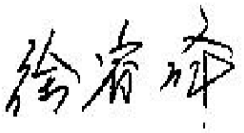
《面向产品评价的细粒度情感分析技术研究》系本人在哈尔滨工业大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归哈尔滨工业大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅，同意学校将论文加入《中国优秀博硕士学位论文全文数据库》和编入《中国知识资源总库》。本人授权哈尔滨工业大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

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： 2013年 1月 11日

导师签名：  日期： 2013年 1月 11日

致　谢

本论文是在徐睿峰老师的悉心指导下完成的，论文只是两年半研究生学业的总结，然而徐老师所告诫的严谨处事的态度却可以受用一生。衷心感谢徐老师两年多以来对我的学习甚至生活的帮助！徐老师不厌其烦的教导我如何使用科学的方法做研究，让我真正领悟了学而不思则惘，思而不学则怠的含义，避免了我在学习研究中走过多的弯路。徐老师是也是我们的益友。无论是初到深圳的安置，佳节每逢倍思亲时的照顾，研究学习之余的户外游玩，以及最近的对我哈尔滨寻找工作之旅的关心与帮助，徐老师在生活上的照顾给一个异乡学子添置了太多温暖。也感谢实验室其他老师们，特别是王晓龙老师为实验室创造了良好的研究环境与学习气氛，以及陈清才与刘滨老师对我的论文提出的宝贵意见。

感谢徐军博士给我的无私帮助和积极支持。在DOTA的研究过程中桂林同学的创造力让我受益匪浅，叶璐同学的吐槽也给矛了我反思与欢乐，感谢智能计算实验室所有的兄弟姐妹们，陪伴我度过了这快乐充实的研究生阶段。感谢球友们的热血支持，无论射门还是投篮都少不了你们的激情四射。也感谢全体室友，有了你们的胡来与义气，让我不再孤单。

最后，还要感谢我的亲人们，他们对我要求甚少，但给予我的都是关怀、支持和理解。特别感谢我的女友秦晓娟同学，在论文撰写期间保证了我正常的营养摄入以及论文文字校对工作，适度的无理取闹进一步激发了我论文写作的灵感。感觉父母对我的论文撰写所用电脑的赞助，特此为你们提供内容为“永远爱你们”的广告条位！