·国防科技大学语言文学博士文库·
主 编 刘载锋 刘 晶

敖 锋 著

Identification of Positive/Negative Implicit Sentiment at the Document Level

肯定/否定篇章隐含情感倾向性分析

河南大学出版社
HENAN UNIVERSITY PRESS

Identification of Positive/Negative
Implicit Sentiment at the Document Level

肯定/否定篇章隐含情感倾向性分析

敖　锋　著

河南大学出版社
·郑州·

图书在版编目(CIP)数据

肯定/否定篇章隐含情感倾向性分析/敖锋著.—郑州:河南大学出版社,2011.4

ISBN 978-7-5649-0436-4

Ⅰ.①肯… Ⅱ.①敖… Ⅲ.①英语-语言学-研究 Ⅳ.①H31

中国版本图书馆 CIP 数据核字(2011)第 072248 号

责任编辑 薛巧玲
责任校对 付理明
封面设计 马 龙

出 版 河南大学出版社
地址:郑州市郑东新区商务外环中华大夏 2401 号 邮编:450046
电话:0371-86059701(营销部) 网址:www.hupress.com
排 版 郑州市今日文教印制有限公司
印 刷 开封市精彩印务有限公司
版 次 2011 年 9 月第 1 版 **印 次** 2011 年 9 月第 1 次印刷
开 本 720mm×1000mm 1/16 **印 张** 12.25
字 数 188 千字 **定 价** 26.00 元

国防科技大学语言文学博士文库

序　言

敖锋是我2006年招收的博士生，也是我门下年龄最小的弟子。他刻苦有加，乐于探索，勇于创新。经过四年的苦读和探索，他的论文《肯定/否定篇章隐含情感倾向性分析》即将付梓。他要我写上几句以为序，我自欣然应之。

敖锋攻读的是“心理语言学”研究方向，但是他对自然语言处理和计算语言学颇有涉猎。鉴于他的知识结构，我建议他将心理语言学与自然语言处理技术结合起来，争取通过不同学科的交叉在理论上有所创新，在实践上有所突破，同时这也是我校语言研究下一步的发展方向。但是，将语言研究与自然语言处理的计算机应用技术相结合需要研究者同时了解这两个领域内的相关理论和实现工具，并且必须找到一个合适的切入点进行深入研究，需要做许多原创性的工作，这必将为论文的撰写带来不小的难度。幸而敖锋饱含开拓创新之激情，知难而进，又深谙科学研究之淡定，静思笃行，最终克服艰难完成了论文，令人欣慰。

本书以篇章隐含情感倾向性分析为研究对象，借助心理语言学的理论和实验，利用构式特征和机器学习算法建立了倾向性分析模型。文章不蔓不支，大含细入，其中不乏真知灼见和创新之处。

首先，综合了语言学和自然语言处理等多个学科的研究方法，研究思路和技术手段对于倾向性分析乃至语言研究都提供了很好的借鉴。

其次，首次将构式语法理论用于隐含倾向性分析，并通过心理语言学实验验证了构式对文本的隐含情感倾向性的作用。

最后，构建了基于构式语法特征的隐含情感倾向性分析模型，设计了从文本中抽取构式语法特征的方法，并运用机器学习技术进行倾向性判断，取得了较好的实验结果。

一言以蔽之，敖锋的论文结构蹙金结绣，内容枝叶扶疏，可谓一篇衔华佩实之作。虽然其中仍有微瑕，但实属迈出了可喜的一步。以上当否，自有此文的读者鉴别，就此打住。诚望敖锋能有更好更多的成果问世，为心理语言学和自然语言处理技术又好又快发展作出贡献。

是为序。

李绍山

2011 年 2 月

前　言

倾向性分析又称观点挖掘或立场判断等，是自然语言处理以及数据挖掘领域中的热点问题之一，其分析结果旨在发现文本作者对该文本涉及话题所持的态度或者观点。本书在倾向性分析中区分出隐含倾向性分析，二者之间的差异主要存在于分析的对象。传统的倾向性分析主要针对的是通过明示的词语来表达态度的文体，即主观文体；而隐含倾向性分析针对的对象是客观文体，即表面上描述客观事实，实际上含有情感倾向的文体。前人的倾向性分析研究主要是针对主观文体展开，有些倾向性研究就建立在主观/客观分析结果之上。这些研究忽略了表面上具有客观描述特征的语言，因而对于隐含倾向性缺乏解释力。

有鉴于此，本书的隐含倾向性分析针对的是没有明示其主观观点或者态度、表面上描述客观事实的文本。这样的文体无法通过传统的统计方法来进行倾向性分析，它们具有的倾向性是篇章作者所持的视角。此外，前人的研究主要使用统计方法，对语言知识的利用十分有限。本研究将结合构式语法（Construction Grammar）理论提取可用的语言特征，利用机器学习算法建立模型，然后使用这些特征对模型进行训练。在特征提取阶段，本研究还设计了心理语言学实验作为实证支持。

在研究步骤上，本研究首先陈述选择构式语法作为本研究特征提取理论基础的理由，然后分别介绍将被本研究用于特征提取的两种构式语法理论。在分析了构式与隐含倾向性之间的关系以后，本研究采用了心理语言学实验来为前面的理论叙述提供实验依据。而后，本研

究将理论叙述的内容应用到隐含倾向性分析实践，以构式作为特征提取的对象，建立巴以冲突语料库作为训练和测试模型的材料，对基于机器学习算法的隐含倾向性模型进行了测试。此外，本研究还将基于构式特征提取的隐含倾向性分析模型应用到影评文本和议会辩论文本，其间涉及了确定文本主题相关中心词汇的方法和基于最小割的图联合方法。

本书主要包括以下六章：

第一章为绪论，该章对倾向性分析进行了概述，并介绍了倾向性分析面临的挑战和本研究将遵循的研究步骤。

第二章主要对此前的倾向性分析研究进行了回顾，介绍了倾向性分析中涉及的概念和方法，着重介绍了分类、特征提取和倾向性分析采用的基本方法。鉴于本研究重在将语言学研究与倾向性分析相结合，因此本研究在此处独辟一节介绍了涉及语言学知识的倾向性分析研究，并指出前人研究中存在的不足，随之提出本研究的研究思路。

第三章是本研究的理论分析部分，是本研究的理论基础。本章首先论述了选择构式语法作为本研究理论基础的主要原因，然后简要介绍了在研究中需要借鉴的具体构式语法理论，并分析了将构式语法用于隐含倾向性分析的途径。最后，在理论分析结束以后，本章还进行了心理语言学实验来验证构式对隐含情感倾向的影响作用。

第四章将理论分析的结果投入隐含倾向性分析实践。具体而言，本章首先定义了用于隐含倾向性分析模型的构式特征。在确定特征提取方案以后，本研究建立了巴以冲突语料库作为隐含倾向性分析模型训练集和测试集的特征数据来源，并随即进入隐含倾向性分析模型的训练和检验。模型分别采用了两种机器学习算法。其目的并非对比这两种算法在隐含倾向性分析中孰优孰劣，而在于更加充分地说明本研究设定的特征提取方案能够在隐含倾向性分析中发挥积极的效果。此外，本研究还制定了一个方法以确定与篇章文本主题相关的中心词汇，并根据确定出来的中心词汇进行构式特征的提取。结果表明，基于中心词汇的构式特征可以帮助隐含倾向性分析模型取得更好的分析效果。

第五章将第四章中建立的模型扩展应用到两个其他领域，即影评

文本和议会辩论文本，以此进一步验证该模型的分析效果和适用性。结果表明，本研究建立的隐含倾向性分析模型针对这两种文本也有不错的分析效果。在针对议会辩论语料库进行分析时，本研究还采用最小割的方法提取了文本间的关系特征。结果表明，本研究定义的构式特征不仅可以涵盖单个篇章文本特征中的有效信息，而且可以发挥某些文本间关系特征在隐含倾向性分析中的作用。

第六章是对本研究的总结，指出了本研究的主要工作、主要贡献以及存在的不足，并为进一步的研究工作提供了建议。

本研究首次将构式语法与隐含倾向性分析相结合，基于构式语法进行特征提取，将这些特征用于训练和测试基于机器学习算法的隐含倾向性分析模型。它在理论、研究方法和实用性方面都具有自身的特点和值得借鉴之处。在理论层面，本研究解释了构式与隐含倾向性之间的关系，并运用心理语言学实验验证了构式对隐含情感倾向的影响作用。此外，本研究以客观文本为分析对象，突出了语言学知识在自动化倾向性分析中的作用。在研究方法层面，本研究在训练和测试隐含倾向性分析模型时采用了多种软件和方法，其经验可供其他相关研究参考借鉴。在实用性方面，本研究所建立的隐含倾向性分析模型也表现出了较高价值。

本研究的不足之处在于：其一，本研究在解释构式与隐含倾向性之间的关系以及在对构式特征进行提取时以构式语法作为理论基础，具体采用的构式语法理论为题元结构理论和词汇语义理论。这两种理论在构式语法理论中具有重要地位，但是它们也不能代表所有的构式语法理论的思想。如果能够在构式特征的提取过程中充分考虑到其他构式语法理论的研究成果，那么构式特征的提取必将更为完善，模型分析的效果也将有所提高。其二，本研究使用的研究工具比较简单，还难以实现将研究中建立的模型投入市场化应用。其三，本研究没有发现可以提高原有模型分析效果的文本间关系特征，仍需进一步研究以确定能够提高隐含倾向性分析模型分析效果的关系特征。

敖　锋

2011 年 2 月

目　录

第一章 绪 论

1.1 概述

随着信息技术的不断发展，信息的获取变得越来越容易，随之而来的问题是如何提高信息处理的效率。利用计算机进行自然语言处理为此提供了便利。在面对大量的语言材料时，人们往往不满足于根据内容将材料分类或者进行简单的关键词检索，很多时候人们还需要了解这些材料所表达的观点和态度，这就需要进行情感倾向性分析(sentiment analysis)。为叙述方便，本书将情感倾向性分析简称为倾向性分析。倾向性分析是涉及自然语言处理、计算语言学以及文本挖掘的一个研究课题。总的来说，倾向性分析的目的是希望发现某文本的作者(或者说话者)对该文本中涉及的话题所持的态度或者观点。这里的态度可以是文本作者做出的判断或者评价，可以是文本作者在文中表达出的观点和态度，也可以是文本作者希望使读者(或者听话者)产生的观点或者态度。

尽管倾向性分析被认为是最近才兴起的一项研究，但实际上以前已有不少研究都体现出了倾向性分析的思想(比如 Wilks & Bien，1984；Carbonell，1979)。2001 年可以说是倾向性分析研究得到突飞猛进的一年。在这一年中，研究者们发表了大量有关倾向性分析的论文。这些论文有力地传播了学界对倾向性分析的认识，点明了该类研究中

存在的难点以及面临的机遇。引发这次腾飞的原因主要有三个:1)自然语言处理和信息挖掘领域中机器学习方法的发展为倾向性分析提供了技术基础;2)随着互联网的发展,网络上出现了可供机器学习算法进行训练的大量数据库,其中包括众多具有意见收集性质的网站;3)倾向性分析在商业和信息情报领域中的成功应用增加了相关方面对倾向性分析研究的投入,这也推动了倾向性分析研究的发展(Pang & Lee, 2008)。

倾向性分析可以发挥作用的领域很广,比如:

- 商业咨询:工厂和企业很在意社会大众对他们生产的产品或者提供的服务所做的评价、所持的态度。他们不能寄希望于通过人工阅读来分析使用者提供的反馈信息以及从各种渠道搜索获取的评论性文本,这样太不经济。此时,利用计算机进行自动倾向性分析是理想的。这种方法可以高效率、低成本地分析出篇章的倾向性信息,工厂和企业可以根据倾向性分析的结果再进行进一步的研究。尽管倾向性分析并不能实现对文本的最终处理,比如自动生成摘要,但是它无疑为生产者节约了两个重要的资源:时间和劳动力。

  Zabin & Jefferies(2008)也指出了倾向性分析在商业中的价值:随着博客、论坛、同事间的网络以及其他各种各样的媒介的出现,消费者在前所未有的广度和深度上对商品和服务的信息进行交流,其中有肯定的态度也有否定的态度。许多大型企业已经意识到,在媒体上传播出来的消费者的意见会极大地影响其他消费者的观点和态度并从而影响商品或者服务的销量。而企业可以根据这些媒介上消费者提供的意见来调整自己的营销策略、品牌定位、产品发展等商业决策。Kim & Hovy(2006)指出,企业需要对有关其品牌信息的传媒进行管控,其中涉及公共关系处理、预防盗版以及获取竞争信息等。但是,传媒分布的广泛性和人们消费行为的改变使传统的管控方式难以应付新形势的需要,而基于信息技术的倾向性分析则正好满足时代的要求。有调查表明,互联网上每天会出现 75000 多个新博客以及 120 多万条新发布的帖子,其中大部分会涉及某种商品和服务。

如此大量的信息使剪报服务等传统管控策略鞭长莫及。

- 媒体分析：随着信息渠道越来越开放，各种媒体往往需要发表出不同的声音才能引起受众的关注，因此各种媒体对某个事件所持的观点就不尽相同。另外，媒体的不同立场也会产生不同的情感倾向。比如媒体监督网站 HonestReporting.com[①] 就指出，英国 BBC 在有关巴以冲突的报道中偏向巴方，而对以色列持否定态度。2008 年中国网络流行语中有一条是“做人不要太 CNN”，其由来是美国 CNN 电视台在对当年在中国西藏发生的打砸抢杀事件进行报道时对事实有所扭曲，网民因此用“CNN”来指代故意歪曲事实真相的行为。可见，倾向性分析对媒体研究乃至政治经济学研究都是很有意义的。
- 网络信息过滤：网络是便捷的信息通道。正是由于它便捷的特点，它不仅传播了正面的积极向上的信息，也为负面的、可能对社会造成危害的信息提供了传播的平台。由于网络信息的海量性和高更新率，人为地进行信息过滤是不现实的。此时，基于网络的倾向性分析系统就大有用武之地，它对社会稳定、经济发展都有十分重要的意义。
- 军事情报处理：军事情报和上述信息之间有一个重大区别，即军事情报的保密性。由于不能公开，军事情报无法利用大量的人力进行分析，加之信息量庞大，军事情报的处理面临着两难的困境。如果能利用计算机进行情报的初加工，区分出情报的倾向性，那必将为军事情报的处理提高效率，并解放出大量的人力资源以进行精度更高的情报处理工作。

可见，倾向性分析研究具有十分重要的理论价值和应用价值，如何提高自动倾向性分析的效果成为自然语言处理中的一个重要课题，这主要体现在篇章层面和篇章以下层面的倾向性分析中。篇章层面倾向性分析主要是要分析出整个篇章内容对其中涉及的话题持何种态度，这方面的典型案例是影评分析和产品评论分析。篇章以下层面的倾向

① 可浏览网址 http://www.honestreporting.com/articles/critiques/Study_Reuters_Headlines.asp。

性分析是指基于小句、句子以及段落的倾向性分析，这方面的典型案例是观点挖掘(opinion mining)，即区分出描述主观内容和客观内容的句子或者段落。本书关注的是篇章层面的倾向性分析，因为该层面分析的结果可以直接用于实践，其现实价值大于篇章以下层面的分析。

从倾向性分析的结果类型来看，本研究主要针对以视角为对象的隐含倾向性分析。根据 Lin 等人(2006)的观点，倾向性分析的对象涉及多种类型，其中包括主/客观分析(如 Wiebe *et al.*，2004；Riloff *et al.*，2003)、观点篇章识别(如 Yu & Hatzivassiloglou，2003)、观点句子识别(如 Yu & Hatzivassiloglou，2003；Riloff *et al.*，2003)以及肯定/否定分析(Pang *et al.*，2002；Morinaga *et al.*，2002；Yu & Hatzivassiloglou，2003；Turney & Littman，2003；Dave *et al.*，2003；Nasukawa & Yi，2003；Popescu & Etzioni，2005；Wilson *et al.*，2005)和视角分析(Lin *et al.*，2006)。主/客观分析指的是区分对象文本的主观性或客观性。肯定/否定分析的最终结果是判定对象文本对其中的主题所持的态度是肯定还是否定。至于视角分析，根据 Lin 等人(2006)的解释，视角和观点(opinion)之间有所区别。观点持有者针对不同的对象往往会形成不同的观点，但是他们看待问题的视角却是一种相对稳定的世界观，这就是本书要研究的重点。根据 Greene(2007)的理解，本书将这一领域称为隐含倾向性分析，在后文有详细阐述。

1.2 倾向性分析面临的挑战

首先我们来看一个选自 Brown 语料库的英文句子："California's newest anti-secrecy law was as dismaying as it was disappointing."该句子的主体是"California's law"，句中"dismaying"和"disappointing"的出现表明了整个句子的否定倾向性。从这个角度来看，倾向性分析似乎难度不大。只要找到文本中具有情感倾向性的关键词，就可以确定整个文本的倾向性。

但实际上，Pang 等人(2002)在早期的倾向性分析研究中就发现，要在文本中找到正确的关键词并非易事。在该研究中，两位研究人员

通过人工分析挑选出影评文本中可以表现出肯定或者否定倾向的关键词。根据这些关键词进行统计得到的倾向性分析准确率可以达到接近60%的水平。相反,使用语料库统计方法得到的同样规模的关键词却可以使倾向性分析的准确率达到70%左右。在这些关键词中,有些词汇似乎并不会表现出态度上的倾向,比如"still",也许正是这些词汇导致了不同关键词选取方法在倾向性分析效果上的差异。这一研究使我们更加深入地了解到采用关键词统计方法来进行篇章情感倾向性分析所面临的挑战。

人工选词的低准确率似乎让人觉得倾向性分析相对于根据内容的文本分类来说难度更大。那么倾向性分析具体面临着哪些因素的挑战呢?从下面的例子中我们就可见一斑:

> "If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut. (如果你阅读这篇文章的理由是因为这款香水是你的最爱,那么我们建议你只在家里使用这款香水,并且别忘了把家里的窗户封好。)"(Viking,2008)

这个例子选自一篇香水评论文章。香水本身是人们作为添加自身魅力的辅助工具,如果一款香水只能在家里使用,使用时还需要将家里的窗口封闭起来,那么这款香水肯定不是那么使人觉得愉悦。但是在该例子中没有出现一个具有明示否定倾向的词汇;相反,其中还有"darling"和"fragrance"这种具有明示肯定倾向的词汇出现。这就形成了词汇倾向性与篇章整体倾向性之间的矛盾。此外,例子由三个小句组成,孤立地分析其中任何一个小句都不能得出正确的篇章倾向性,并且这些小句并没有表现出具有倾向性的特点。换句话说,尽管这些小句看上去是没有倾向性的,但是它们构成一个完整的句子后该句子却表现出否定倾向。这也就是客观文体表现出主观倾向的例子,是本研究的研究对象。

由此可见,倾向性是一种微妙的篇章内容。在对其中的句子或者词汇进行孤立的分析时不一定能得出正确的篇章倾向性结果。与此同时,通过简单的关键词统计方法来进行主观性分析和倾向性分析可能无法获得理想的准确率。即使句子中出现"事实上"这样的字样,也不

能保证该句子描述的就一定是客观事实；而“没有意见”这种字眼的出现也并非一定表明文本中真的没有肯定或者否定的意见和观点被表述出来。因此，倾向性分析需要借助深层次的分析来提高其分析效果，其中也包括对语言学知识的利用。

1.3 倾向性分析研究定义

苏格拉底曾经说过，智慧来源于对术语的界定。这句格言在社会科学的研究中尤为适用。在社会科学的研究中，不同学者或者学派对术语的理解往往是不同的，他们很难针对某一具体的构念达到一致的认识。说到倾向性分析，尽管倾向性分析在很大程度上依赖于自然科学的成果，采用的技术也大多是自然科学研究中的技术，从这个方面来说似乎应该具有自然科学领域中术语界定明确的特点。但是，目前学界还没有对倾向性分析形成明确的定义。

首先，在倾向性分析方面，研究者们的称谓多种多样。Zabin 和 Jefferies(2008)为我们提供了很多例子，比如品牌监测、市场影响分析学、对话挖掘以及在线消费者情报等等。其中最为流行的是情感倾向性分析(sentiment analysis)、主观性分析(subjectivity analysis)和观点挖掘(opinion mining)。这些术语体现了它们所指内容之间的差异，我们将在下文中逐一讨论。

其次，在倾向性本身的定义方面，国外研究者们也没能达成一致意见。相关的叫法很多，比如“attitude”，“opinion”，“affect”，“emotion”，“spin”，“perspective”等。这些术语的定义与它们之间的区别还没有得到业界的明确界定。心理学和社会学的学者(比如 Oskamp，1991；Ekman，1994)试图对情感进行精确定义，也同样遇到了一致性和准确性的问题。就以 Oskamp (1991)为例，他定义了情感的不同层面，希望以此建立不同术语之间不同程度的一致性，如表 1。Oskamp 的定义可谓细致，但他还是将“attitude”和“opinion”未加区分地混用(Greene，2007)。