中文信息处理丛书

统计自然语言处理

宗成庆　编著

清华大学出版社

中文信息处理丛书

统计自然语言处理

宗成庆　编著

清华大学出版社
北京

内 容 简 介

本书全面介绍了统计自然语言处理的基本概念、理论方法和最新研究进展，内容包括形式语言与自动机及其在自然语言处理中的应用、语言模型、隐马尔可夫模型、语料库技术、汉语自动分词与词性标注、句法分析、词义消歧、统计机器翻译、语音翻译、文本分类、信息检索与问答系统、自动文摘和信息抽取、口语信息处理与人机对话系统等，既有对基础知识和理论模型的介绍，也有对相关问题的研究背景、实现方法和技术现状的详细阐述。

本书可作为高等院校计算机、信息技术等相关专业的高年级本科生或研究生的教材或参考书，也可供从事自然语言处理、数据挖掘和人工智能等研究的相关人员参考。

图书在版编目（CIP）数据

统计自然语言处理/宗成庆编著. —北京：清华大学出版社，2008.5
（中文信息处理丛书/倪光南主编）
ISBN 978-7-302-16598-9

Ⅰ.统…　Ⅱ.宗…　Ⅲ.统计方法－应用－自然语言处理　Ⅳ.TP391

中国版本图书馆 CIP 数据核字(2007)第 189530 号

责任编辑：赵彤伟
责任校对：刘玉霞
责任印制：王秀菊

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京鑫丰华彩印有限公司
装 订 者：北京市密云京文制本装订厂
经　　销：全国新华书店
开　　本：185×260　　**印　张**：32　　**字　数**：732 千字
版　　次：2008 年 5 月第 1 版　　**印　次**：2008 年 5 月第 1 次印刷
印　　数：1～3000
定　　价：66.00 元

序言

自然语言处理技术的产生可以追溯到20世纪50年代，它是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科。近几年来，随着计算机网络技术和通信技术的迅速发展和普及，自然语言处理技术的应用需求急剧增加，人们迫切需要实用的自然语言处理技术来帮助人们打破语言屏障，为人际之间、人机之间的信息交流提供便捷、自然、有效的人性化服务。但是，自然语言处理中的若干科学问题和技术难题尚未得到解决，有待于来自不同领域的学者深入研究和探索。

中文信息处理作为自然语言处理中的一个分支，近几年来备受关注。一方面，随着中国经济的迅速发展和中国国力的不断增强，汉语正在成为一种新的强势语言而被世人瞩目，汉语理解所涉及的科学问题让国际计算语言学界无法回避；而另一方面，汉语使用者所拥有的巨大市场潜力令国际企业界不敢轻视。因此，中文信息处理成为全球自然语言处理研究者们共同关注的问题已经是不争的事实。目前国际上每年举行的颇具影响的几种技术评测，包括机器翻译评测、信息抽取评测和句法分析评测等，无不与汉语密切相关。因此，作为炎黄子孙，我们没有理由不在这一领域的研究中做出应有的贡献。

中文信息处理所面临的困难既有其他任何一种自然语言处理都会遇到的共性问题，如生词识别问题、歧义消解问题等，也有中文处理本身所具有的个性化问题，如汉语自动分词问题、词性定义规范问题等。因此，从某种意义上讲，中文信息处理更具挑战性。值得欣慰的是，中文信息处理在引起国际学术界和企业界关注的同时，得到了中国政府的重视和大力支持，它已经被列入国务院批准的“国家中长期科学技术发展规划纲要”。因此，中文信息处理面临着前所未有的大好机遇。

近几年来，我国的中文信息处理技术得到了快速发展，无论是在基础理论研究方面，还是在技术开发和产业化发展方面，都取得了显著成绩，一大批青年学者投身到这一领域中。为了使这一领域的广大学者，尤其是青年学生，全面了解中文信息处理的技术现状，进一步推动中文信息处理及其相关学科的快速发展，我们组织编写并出版了这套中文信息处理丛书，力求将

这一领域的最新技术和理论方法全面、系统地介绍给广大读者。随着丛书的陆续出版，我们相信这套丛书必将在促进中文信息处理技术的发展和培养后继人才队伍方面发挥它应有的作用。

感谢清华大学出版社给予的支持。

倪光南
中国工程院院士
中国中文信息学会理事长
2007 年 12 月 20 日

序一

自然语言理解和处理是近几年来发展迅速的一门自然语言学、数学(尤其是代数、概率)与计算机科学交叉的学科,如何让计算机正确、有效地理解和处理人类语言,是当今具有巨大挑战性的理论和技术问题。从研究现状来看,自然语言理解和处理的理论体系尚未真正建立,技术方法仍然十分初步,正如作者在绪论中指出的,如何建立语言、知识与客观世界之间的关系,尤其是可计算的关系,如何揭示人类理解及处理自然语言的认知过程等一系列科学问题尚未找到答案。自然语言理解和处理不仅是一门社会需求十分巨大的应用技术,而且也是一门具有非常重要科学意义的自然科学。

由于统计法能使自然语言处理的正确率从比较低的水平有较快增长,引起人们广泛注意,所以近十多年来有比较快的发展。

为了帮助读者把握这一领域的发展脉络,本书对统计自然语言处理的相关理论和实现方法给予了较为全面的系统阐述,尤其对中文信息处理研究的最新进展和成果有较全面的阐述,通过总结和归纳这一领域已有的理论方法,既可以为新理论方法的研究提供依据,又可以为这一领域的入门者提供向导。

虽然近几年来,国外学者已经编写出版了一些关于统计自然语言处理的专著,有的还被国内学者翻译成汉语,但是,关于中文信息处理的最新成果却未能在那些专著中得到充分的介绍。本书比较详细地介绍了近几年来,国内学者在汉语语料库和词汇知识库建设、自动分词(包括分词方法和命名实体识别等)与词性标注、句法分析以及口语信息处理等方面研究的最新成果,全面反映了中文信息处理研究的现状,这是在国外学者编写的专著中难以看到的。而且,本书还详细介绍了这一领域的一些经典论文和近几年来获得国际计算语言学大会(ACL)最佳论文奖的部分论文,这些工作都将促进中文信息处理研究的开展。

本书比较全面地涵盖了自然语言处理的相关理论和应用技术,既有形式语言与自动机、语言模型和隐马尔可夫模型等基础理论介绍,也有汉语自动分词、句法分析和词义消歧等基本方法描述,还有统计机器翻译、语音翻译、信息检索、文本分类和口语信息处理等应用技术的全面阐述,其涉及范

围之广，参考文献之详尽，在本领域的专著中尚属少见。作者在本书的编写过程中付出了辛勤的劳动，他能直接与相关文献的作者讨论，而且在初稿完成后广泛征求有关学者的意见，这种认真负责的治学态度十分可贵。

我们相信，本书的出版将对国内自然语言处理的理论研究和技术开发，以及中文信息处理研究和技术的进一步发展发挥积极的作用。

高庆狮

中国科学院院士

2007 年 12 月 20 日

序二

我在1996年出版的《自然语言的计算机处理》中曾经说过："自然语言处理(natural language processing,NLP)就是利用计算机为工具对人类特有的书面形式和口头形式的语言进行各种类型处理和加工的技术。"[①]这个定义是正确的,它的缺点是比较笼统。我一直不太满意这个定义。

后来,我在1999年出版的《计算机进展》(*Advanced in Computers*)第47卷上,看到了美国计算机科学家马纳瑞斯(Bill Manaris)在《从人-机交互的角度看自然语言处理》一文中给自然语言处理提出的如下定义："自然语言处理可以定义为研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力(linguistic competence)和语言应用(linguistic performance)的模型,建立计算框架来实现这样的语言模型,提出相应的方法来不断地完善这样的语言模型,根据这样的语言模型设计各种实用系统,并探讨这些实用系统的评测技术。"这个定义的英文如下："NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement process incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the result systems."[②]

马纳瑞斯的这个定义更加完善,把自然语言处理的研究过程也清楚地反映出来了。我觉得,这是目前在汗牛充栋的各种文献中可以找到的关于自然语言处理的一个比较好的定义。我原则上认同这个定义。

根据这个定义,自然语言处理要研究"在人与人交际中以及在人与计算机交际中的语言问题",既要研究语言,又要研究计算机,因此,它是一门交叉学科,它涉及语言学、计算机科学、数学、自动化技术等不同的学科。

---

① 冯志伟.自然语言的计算机处理.上海:上海外语教育出版社.1996

② Bill Manaris. Natural language processing: A human-computer interaction perspective. *Advances in Computers*. Volume 47,1999

近年来，由于自然语言处理的发展，不同学科的专家络绎不绝地参加到自然语言处理的队伍中来。这些来自不同学科领域的专家，对于他们自己原来的本行，当然都是精研通达的内行，但是，他们当中的很多人，对于自然语言处理这门交叉学科本身，并没有接受过专门的学习和训练，有必要进行更新知识的再学习，除了学习不同于他们自己本学科的相关学科的知识之外，还有必要学习自然语言处理这门交叉学科本身的知识。

自然语言处理已经有五十多年的发展历史了，在这一漫长的发展过程中，自然语言处理形成了自己特有的理论和方法，成为一门独立的学科，有自己特定的科学内容。关于自然语言处理本身的这些知识，绝不是不学而能的，而是需要经过艰苦的学习之后才可以逐步地掌握。学习自然语言处理这门学科的专门知识，正如学习语言学、计算机科学、数学和自动化技术一样，非下苦功不可。

正是基于这样的理解，中国科学院研究生院专门开设了“自然语言理解”的课程，讲授自然语言处理这门学科特有的专门知识。中国科学院自动化研究所国家模式识别重点实验室研究员宗成庆博士从事自然语言处理研究多年，他从 2004 年开始，每年的春季学期在中国科学院研究生院讲授这门课程。这门课程受到了学生们的欢迎，2005 年被评为中国科学院研究生院的优秀课程。在这门课程的基础之上，宗成庆博士写成了这本《统计自然语言处理》的专著。我国过去曾经出版过一些关于自然语言处理和计算语言学的教材，这些教材中，除了翻译的外版教材之外，大多数只是讲授基于规则的自然语言处理，没有专门讲授基于统计的自然语言处理。这本《统计自然语言处理》弥补了我国缺少自然语言处理教材的缺陷，起了填补空白的作用。这本书纳入“中文信息处理丛书”并由清华大学出版社出版，是我国自然语言处理教材建设的一件值得庆幸的好事。

《统计自然语言处理》一书的整体规划和部分章节是宗成庆博士于 2004 年底在法国格勒诺布尔信息与应用数学研究院（Institut d'Informatique et Mathématique appliquée de Grenoble，IMAG）的自动翻译研究组（Groupe d'Etude de la Transduction Automatique，GETA）完成的。我在 1978 年至 1981 年期间，也曾经在 IMAG 的 GETA 师从著名数学家沃古瓦（B. Vauquois）做过机器翻译的研究，建立了汉-法/英/日/俄/德多语言机器翻译系统，使我对于自然语言处理这个神奇的研究领域产生了越来越浓厚的兴趣，从此就义无反顾地投身于自然语言处理的事业。岁月不饶人，将近三十年光阴匆匆地流逝，当年我还是风华正茂的青年人，而今，已经变成白发苍苍的老人了，我为这个事业坎坷地奋斗了大半生时间，其间甘苦难以言表。三十年来，不论是处于顺境还是逆境，我对于 IMAG 和 GETA 始终怀着难分难解的深厚感情，这种感情当然主要是对于我们共同的自然语言处理事业的感情。宗成庆博士 2004 年底恰巧在 IMAG 的 GETA 写作《统计自然语言处理》一书，说明他和我之间确实有缘分，这样的缘分促使我们这两个年龄相差甚大的人，在自然语言处理这个领域里风雨同舟，休戚与共，一起克服攀登科学高峰的困难，共同分享探索语言奥秘的愉快，成为忘年之交。

宗成庆博士在此书完稿之后，也许是由于他知道我对于 IMAG 和 GETA 的这种特殊感情，马上就给我送来了此书的打印稿，我得以先睹为快。

我带着极大的热情和浓厚的兴趣一口气读完此书。觉得此书覆盖全面，论述清楚，实例丰富，逻辑严密，既有深入的理论分析，又有实际的应用研究。它既是初学者学习统计

自然语言处理的入门初阶，又是这个领域的专家深入钻研统计自然语言处理的导航指南。不禁为之拍手叫绝！

本书在内容的安排方面别具匠心。第 1 至 9 章主要介绍统计自然语言处理的理论，第 10 至 15 章主要介绍统计自然语言处理的应用。

在统计自然语言处理的理论方面，首先介绍有关的基础知识，例如，概率论和信息论的基本概念、形式语言和自动机的基本概念。这些基础知识，对于以语言学为背景的读者是非常有用的，对于理科背景的读者，可以略过这一部分。由于统计自然语言处理是以语料库和词汇知识库为语言资源的，因此，在介绍了有关的基础知识之后，本书讲解了语料库和词汇知识库的基本原理，使读者对语言资源的建造技术获得清楚的认识。语言模型和隐马尔可夫模型是统计自然语言处理的基础理论，在统计自然语言处理中具有重要的地位。因此，本书介绍了语言模型的基本概念，并讨论了各种平滑方法和自适应方法，又介绍了隐马尔可夫模型和参数估计的方法。接着，本书分别论述了在词法分析与词性标注中的统计方法，在句法分析中的统计方法，在词汇语义中的统计方法。

在统计自然语言处理的应用方面，本书对统计自然语言处理的各个应用领域进行了系统的、详细的介绍，分别介绍了统计机器翻译、语音翻译、文本分类、信息检索与问答系统、信息抽取、口语信息处理与人机对话系统等各种应用系统中的统计自然语言处理方法。

从篇幅来看，本书的理论部分与应用部分几乎各占一半，可以说是理论与应用并重。

近年来，统计自然语言处理发展迅速，取得了令人瞩目的成绩。统计自然语言处理的理论逐渐完善，形成了科学的体系，统计自然语言处理的应用硕果累累，产生了很好的社会效益和经济效益，在文字识别、语音合成等领域的技术已经达到了实用化的水平。统计自然语言处理的技术，还进一步应用到网络内容管理、网络信息监控、不良信息的过滤和预警等方面，并且与网络技术、图像识别和理解技术、情感计算(affection computing)技术结合起来，由此而产生了一些新的研究方向，在现代信息科学的发展中，起着越来越重要的作用。

面对统计自然语言处理取得的这些令人鼓舞的辉煌成绩，有些学者的头脑开始发热起来，他们轻视自然语言处理中基于规则的方法，甚至贬低那些从事研究基于规则的自然语言处理的学者。这种局面使我感到困惑。

IBM 公司的杰里内克(Fred Jelinek)是一位使用统计方法研究语音识别与合成的著名学者，他在统计自然语言处理研究中取得的成绩是人所共知的，我也很佩服他的成就，可是，他却看不起使用规则方法研究自然语言处理的人。他于 1988 年 12 月 7 日在自然语言处理评测讨论会上表述了这样的意思：每当一个语言学家离开我们的研究组，语音识别率就提高一步(Anytime a linguist leaves the group the recognition rate goes up)。根据一些参加这个会议的人回忆，杰里内克原话很尖刻，他说："每当我解雇一个语言学家，语音识别系统的性能就会改善一些。"("Every time I fire a linguist the performance of the recognizer improves.") 杰里内克的这些话，把基于规则的自然语言处理研究贬低到了一无是处的程度，把从事基于规则的自然语言处理研究的人，贬低到了一钱不值的程

度，对于基于规则的自然语言处理，采取了嗤之以鼻的态度。[①]

2000 年，在美国约翰·霍普金斯大学（Johns Hopkins University）的暑期机器翻译讨论班（workshop）上，来自南加州大学、罗切斯特大学、约翰·霍普金斯大学、施乐公司、宾夕法尼亚州立大学、斯坦福大学等学校的研究人员，对于基于统计的机器翻译进行了讨论，以德国亚琛大学（Aachen University）年轻的博士研究生奥赫（Franz Josef Och）为主的 13 位科学家写了一个总结报告（final report），报告的题目是“统计机器翻译的句法”（Syntax for Statistical Machine Translation），提出了统计机器翻译的基本框架。奥赫在国际计算语言学 2002 年的会议（ACL-2002）上又发表论文，题目是：“统计机器翻译的分辨训练与最大熵模型”（Discriminative Training and Maximum Entropy Models for Statistical Machine Translation），进一步提出统计机器翻译的系统性方法，获 ACL-2002 大会最佳论文奖。2003 年 7 月，在美国马里兰州巴尔的摩（Baltimore，Maryland）由美国商业部国家标准与技术研究所 NIST/TIDES（National Institute of Standards and Technology）主持的机器翻译评比中，奥赫获得了最好的成绩，他使用统计方法从双语语料库中自动地获取语言知识，建立统计机器翻译的规则，在很短的时间之内就构造了阿拉伯语和汉语到英语的若干个机器翻译系统。伟大的希腊科学家阿基米德（Archimedes）说过：“只要给我一个支点，我就可以移动地球。”（“Give me a place to stand on，and I will move the world.”）而奥赫也模仿着阿基米德说：“只要给我充分的并行语言数据，那么，对于任何的两种语言，我就可以在几小时之内给你构造出一个机器翻译系统。”（“Give me enough parallel data，and you can have translation system for any two languages in a matter of hours.”）奥赫在统计机器翻译方面的成就使我们高兴，使我们看到了未来的机器翻译的曙光，令人鼓舞。[②] 可是，2006 年 6 月奥赫在西班牙巴塞罗那举行的 TC-STAR 机器翻译系统评测研讨会上的特邀报告“机器翻译的挑战”（Challenges in Machine Translation）中却认为：在统计机器翻译中，语料库的规模起着举足轻重的作用，而词法、句法和语义等语言知识对于机器翻译系统的性能几乎没有什么帮助，甚至有些语言知识还会起副作用，帮倒忙。他也开始贬低语言规则在自然语言处理中的正面作用。

杰里内克和奥赫都是在自然语言处理中卓有成就的学者，他们上述的言论值得我们中国的自然语言处理工作者注意，也值得我们深思。

基于统计的自然语言处理的理论基础是哲学中的经验主义，基于规则的自然语言处理的理论基础是哲学中的理性主义。这些问题，说到底，是关于如何处理经验主义和理性主义关系的问题。为了追本溯源，在这里，我愿意回顾一下哲学中经验主义与理性主义，并且考察一下它们对于语言学和自然语言处理的影响，这样，也许能够帮助我们更清楚地认识到这个问题的实质。

自从人类有哲学以来，在认识论中就产生了经验主义（empiricism）和理性主义（rationalism）这样两种不同的倾向。在欧洲哲学史上，当近代哲学家把这两种倾向的冲

---

① Palmer M.，Finin T. Workshop on the evaluation of natural language processing systems. Computational Linguistics，1990，16(3)：175—181

② 冯志伟. 当前自然语言处理发展的四个特点. 暨南大学华文学院学报，2006，第 1 期（总 21 期）

突以及解决这一冲突的不懈努力提到全部哲学的中心地位上来之前，无数的哲学家就已经对此进行了艰苦卓绝的研究，走过了崎岖漫长的探索道路。

人类哲学从它产生的第一天起，就在自身之内包含着一个深刻的矛盾：哲学来自经验，但它又是超越经验的结果；哲学思想的发展是理性思维、范畴和概念的运动，但又只有经验才能推动它。感性与理性的这种矛盾实质上也就是经验主义和理性主义的矛盾，它作为存在和思维的矛盾在认识论方面的一个表现，自开始的时候起，就是人类哲学思想发展的内在动力之一。

这种矛盾在人们的思想中有不同程度、不同形式的表现，但是，经验主义和理性主义作为比较典型的认识论的理论，形成了两个既互相对立、互相斗争，又互相影响、互相渗透的哲学流派而在哲学史上出现，并且在西欧早期资产阶级反封建革命时期前后，成为16世纪末期到18世纪中期重要的历史现象。

在16世纪到18世纪的欧洲，经验主义哲学以培根(Francis Bacon，1561—1626)、霍布斯(Thomas Hobbes，1588—1679)、洛克(John Locke，1632—1704)、休谟(David Hume，1711—1776)为代表，他们都是英国哲学家，因此，经验主义也被称为“英国经验主义”。培根批评“理性派哲学家只是从经验中抓到一些既没有适当审定也没有经过仔细考察和衡量的普遍例证，而把其余的事情都交给了玄想和个人的机智活动”①。他提出“三表法”，制定了经验归纳法，建立了归纳逻辑体系，对于经验自然科学起了理论指导作用。霍布斯认为归纳法不仅包含分析，而且也包含综合，分析得出的普遍原因只有通过综合才能成为研究对象的特殊原因。洛克把理性演绎隶属于经验归纳之下，对演绎法作了经验主义的理解，他认为，一切知识和推论的直接对象是一些个别、特殊的事物，我们获取知识的正确途径只能是从个别、特殊进展到一般，他说，“我们的知识是由特殊方面开始，逐渐才扩展到概括方面的。只是在后来，人们就采取了另一条相反的途径，它要尽力把它的知识形成概括的命题”②。休谟运用实验推理的方法来剖析人性，试图建立一个精神哲学体系，他指出“一切关于事实的推理，似乎都建立在因果关系上面，只要依照这种关系来推理，我们便能超出我们的记忆和感觉的见证以外”③，他认为“原因和结果的发现，是不能通过理性，只能通过经验的”④，经验是我们关于因果关系的一切推论和结论的基础。

现代自然科学的代表人物牛顿(Isaac Newton，1642—1727)建立了经典力学的基本定律，即牛顿三大定律和万有引力定律，使经典力学的科学体系臻于完善。他的哲学思想也带有明显的经验主义倾向。他认为自然哲学只能从经验事实出发去解释世界事物，因而经验归纳法是最好的论证方法。他说：“虽然用归纳法来从实验和观察中进行论证不能算是普遍的结论，但它是事物本性所许可的最好的论证方法，并随着归纳的愈为普遍，这种论证看来也愈有力。”⑤他把经验归纳作为科学研究的一般方法论原理，认为“实验科

---

① 北京大学哲学系外国哲学史教研室编译. 十六—十八世纪西欧各国哲学. 第23页，北京：商务印书馆，1975

② 洛克. 人类理解论. 第598页，北京：商务印书馆，1959

③ 休谟. 人类理解研究. 第27页，北京：商务印书馆，1983

④ 北京大学哲学系外国哲学史教研室编译. 十六—十八世纪西欧各国哲学. 第634页

⑤ 塞耶编. 牛顿自然哲学著作选. 第212页，北京：商务印书馆，1974

学只能从现象出发，并且只能用归纳来从这些现象中推演出一般的命题”[①]。正是由于牛顿遵循经验归纳法，才在物理学上取得了划时代的伟大成就。

法国启蒙运动的代表人物伏尔泰（Voltaire，1694—1778）也有明显的经验主义倾向。他以洛克的经验主义为武器去反对教会至上的权威，否定神的启示和奇迹，否认灵魂不死。他赞美经验主义哲学家洛克：“也许从来没有一个人比洛克头脑更明智，更有条理，在逻辑上更为严谨”[②]。他积极地把英国经验主义推行到法国，推动了法国的启蒙运动。

哲学中的这种经验主义深刻地影响到自然语言处理中基于统计的经验主义方法，它是自然语言处理中经验主义方法的哲学基础。

在自然语言处理中，除了基于统计的经验主义方法之外，还同时存在着基于规则的理性主义方法。自然语言处理中的理性主义来源于哲学中的理性主义。

在欧洲，这种理性主义源远流长，到了16世纪末至18世纪中期更加成熟，出现了笛卡儿（Rene Descartes，1596—1650）、斯宾诺莎（Benetict de Spinoza，1632—1677）、莱布尼兹（Cottfried Wilhelm Leibniz，1646—1716）等杰出的理性主义哲学家。笛卡儿改造了传统的演绎法，制定了理性的演绎法，他认为，任何真理性的认识，都必须首先在人的认识中找到一个最确定、最可靠的支点，才能保证由此推出的知识也是确定可靠的。他提出在认识中应当避免偏见，要把每一个命题都尽可能地分解成细小的部分，直待能够圆满解决为止，要按照次序引导我们的思想，从最简单的对象开始，逐步上升到对复杂事物的认识。斯宾诺莎把几何学方法应用于论理学研究，使用几何学的公理、定义、命题、证明等步骤来进行演绎推理，在他的《论理学》的副标题中明确标示“依几何学方式证明”。莱布尼兹把逻辑学高度地抽象化、形式化、精确化，使逻辑学成为一种用符号进行演算的工具。笛卡儿是法国哲学家，斯宾诺莎是荷兰哲学家，莱布尼兹是德国哲学家，他们崇尚理性，提倡理性的演绎法。他们都居住在欧洲大陆，因此，理性主义也被称为“大陆理性主义”。

在哲学领域中，始终都存在着经验主义和理性主义的矛盾和斗争。这种矛盾和斗争，当然也会反映到自然语言处理中来。

早期的自然语言处理研究带有鲜明的经验主义色彩。

1913年，俄国科学家马尔可夫（A. Markov，1856—1922）使用手工查频的方法，统计了普希金长诗《欧根·奥涅金》中的元音和辅音的出现频度，提出了马尔可夫随机过程理论，建立了马尔可夫模型，他的研究是建立在对于俄语的元音和辅音的统计数据的基础之上的，采用的方法主要是基于统计的经验主义的方法。

1948年，美国科学家香农（Shannon）把离散马尔可夫过程的概率模型应用于描述语言的自动机。他把通过诸如通信信道或声学语音这样的媒介传输语言的行为比喻为“噪声信道”（noisy channel）或者“解码”（decoding）。香农还借用热力学的术语“熵”（entropy）作为测量信道的信息能力或者语言的信息量的一种方法，并且他采用手工方法来统计英语字母的概率，然后使用概率技术首次测定了英语字母的不等概率零阶熵为

---

① 塞耶编. 牛顿自然哲学著作选. 第8页，北京：商务印书馆，1974

② 北京大学哲学系外国哲学史教研室编译. 十八世纪法国哲学. 第59页，北京：商务印书馆，1963

4.03比特。香农的研究工作基本上是基于统计的，也带有明显的经验主义倾向。①

然而，这种基于统计的经验主义的倾向到了乔姆斯基(Noam Chomsky)那里出现了重大的转向。

1956年，乔姆斯基从香农的工作中吸取了有限状态马尔可夫过程的思想，首先把有限状态自动机作为一种工具来刻画语言的语法，并且把有限状态语言定义为由有限状态语法生成的语言，建立了自然语言的有限状态模型。乔姆斯基根据数学中的公理化方法来研究自然语言，采用代数和集合论把形式语言定义为符号的序列，从形式描述的高度，分别建立了有限状态语法、上下文无关语法、上下文有关语法和0型语法的数学模型，并且在这样的基础上来评价有限状态模型的局限性，乔姆斯基断言：有限状态模型不适合用来描述自然语言。这些早期的研究工作产生了“形式语言理论”(formal language theory)这个新的研究领域，为自然语言和形式语言找到了一种统一的数学描述理论，形式语言理论也成为了计算机科学最重要的理论基石。

乔姆斯基在他的著作中明确地采用理性主义的方法，他高举理性主义的大旗，把自己的语言学称之为“笛卡儿语言学”(Descartes linguistics)，充分显示出乔姆斯基的语言学与理性主义之间不可分割的血缘关系。乔姆斯基完全排斥经验主义的统计方法。在1969年的“Quine's Empirical Assumptions”一文中，他说：“然而应当认识到，‘句子的概率’这个概念，在任何已知的对于这个术语的解释中，都是一个完全无用的概念”②。他主张采用公理化、形式化的方法，严格按照一定的规则来描述自然语言的特征，试图使用有限的规则描述无限的语言现象，发现人类普遍的语言机制，建立所谓的“普遍语法”(universal grammar)。转换生成语法在20世纪60年代末到70年代在国际语言学界风靡一时，转换生成语法作为自然语言的形式化描述方法，为计算机处理自然语言提供了有力的武器，有力地推动了自然语言处理的研究和发展。

转换生成语法的研究途径在一定程度上克服了传统语言学的某些弊病，推动了语言学理论和方法论的进步，但它认为统计只能解释语言的表面现象，不能解释语言的内在规则或生成机制，远离了早期自然语言处理的经验主义的途径。这种转换生成语法的研究途径实际上全盘继承了理性主义的哲学思潮。

在自然语言处理中的理性主义方法是一种基于规则的方法(rule-based approach)，或者叫作符号主义的方法(symbolic approach)。这种方法的基本根据是“物理符号系统假设”(physical symbol system hypothesis)。这种假设主张，人类的智能行为可以使用物理符号系统来模拟，物理符号系统包含一些物理符号的模式(pattern)，这些模式可以用来构建各种符号表达式以表示符号的结构。物理符号系统使用对于符号表达式的一系列操作过程来进行各种操作，例如，符号表达式的建造(creation)、删除(deletion)、复制(reproduction)和各种转换(transformation)等。自然语言处理中的很多研究工作基本上

---

① 冯志伟在20世纪70年代末和80年代初，模仿香农的工作，采用手工查频的方法测定出汉字的不等概率零阶熵为9.65比特。他的方法也是一种基于统计的经验主义方法。

② Chomsky N. Quine's Empirical Assumptions. In: Davidson D., Hintikka J, eds. Words and Objections, Dordrecht: Reidel, 1969

是在物理符号系统假设的基础上进行的。

这种基于规则的理性主义方法适合于处理深层次的语言现象和长距离依存关系，它继承了哲学中理性主义的传统，多使用演绎法（deduction）而很少使用归纳法（induction）。

自然语言处理中，在基于规则的方法的基础上发展起来的技术有：有限状态转移网络、有限状态转录机、递归转移网络、扩充转移网络、短语结构语法、自底向上剖析、自顶向下剖析、左角分析法、Earley 算法、CYK 算法、富田算法、复杂特征分析法、合一运算、依存语法、一阶谓词演算、语义网络、框架网络等。

在 20 世纪 50 年代末期到 60 年代中期，自然语言处理中的经验主义也兴盛起来，注重语言事实的传统重新抬头，学者们普遍认为：语言学的研究必须以语言事实作为根据，必须详尽地、大量地占有材料，才有可能在理论上得出比较可靠的结论。

自然语言处理中的经验主义方法是一种基于统计的方法（statistic-based approach），这种方法使用概率或随机的方法来研究语言，建立语言的概率模型。这种方法表现出强大的后劲，特别是在语言知识不完全的一些应用领域中，基于统计的方法表现得很出色。基于统计的方法最早在文字识别领域中取得了很大的成功，后来在语音合成和语音识别中大显身手，接着又扩充到自然语言处理的其他应用领域。

基于统计的方法适合于处理浅层次的语言现象和近距离的依存关系，它继承了哲学中经验主义的传统，多使用归纳法而很少使用演绎法。

这个时期自然语言处理中的经验主义派别，主要是一些来自统计学专业和电子学专业的研究人员。在 20 世纪 50 年代后期，贝叶斯方法（Bayesian method）开始被应用于解决最优字符识别的问题。1959 年，布来德索（Bledsoe）和布罗宁（Browning）建立了用于文本识别的贝叶斯系统，该系统使用了一部大词典，计算词典的单词中所观察的字母系列的似然度，把单词中每一个字母的似然度相乘，就可以求出字母系列的似然度来。1964 年，墨斯特莱（Mosteller）和华莱士（Wallace）用贝叶斯方法成功地解决了文章“联邦主义者”（The Federalist）中原作者的分布问题，显示出经验主义方法的优越性。

20 世纪 50 年代还建立了世界上第一个联机语料库——布朗美国英语语料库（Brown corpus）。这个语料库包含 100 万单词的语料，样本来自不同文体的 500 多篇书面文本，涉及的文体有新闻、中篇小说、写实小说、科技文章等。这些语料是布朗大学（Brown University）在 1963 年至 1964 年收集的。随着语料库的出现，使用统计方法从语料库中自动地获取语言知识，成为了自然语言处理研究的一个重要方面。

20 世纪 60 年代，统计方法在语音识别算法的研制中取得成功，其中特别重要的是隐马尔可夫模型（hidden markov model）和噪声信道与解码模型（noisy channel model and decoding model）。这些模型是分别独立地由两支队伍研制的。一支是杰里内克（Jelinek）、巴勒（Bahl）、梅尔塞（Mercer）和 IBM 公司华生研究中心的研究人员，另一支是卡内基-梅隆大学（Carnegie Mellon University）的拜克（Baker）等。AT&T 的贝尔实验室（Bell laboratories）也是语音识别和语音合成的中心之一。

在自然语言处理中，在基于统计的方法的基础上发展起来的技术有：隐马尔可夫模型、最大熵模型、n 元语法、概率上下文无关语法、噪声信道理论、贝叶斯方法、最小编辑距

离算法、Viterbi 算法、A* 搜索算法、双向搜索算法、加权自动机、支持向量机等。

不过，在 20 世纪 60 年代至 80 年代初期的这一时期，自然语言处理领域的主流方法仍然是基于规则的理性主义方法，经验主义方法并没有受到特别的重视。

这种情况在 20 世纪 80 年代初期发生了变化。在 1983 年至 1993 年的 10 年中，自然语言处理研究者对于过去的研究历史进行了反思，发现过去被忽视的有限状态模型和经验主义方法仍然有其合理的内核。在这 10 年中，自然语言处理的研究又回到了 20 世纪 50 年代末期到 60 年代初期几乎被否定的有限状态模型和经验主义方法上去，之所以出现这样的复苏，其部分原因在于 1959 年乔姆斯基对于斯金纳(Skinner)的“言语行为”(Verbal Behavior)的很有影响的评论在 20 世纪 80 年代和 90 年代之交遭到了学术界在理论上的强烈反对，人们开始注意到基于规则的理性主义方法的缺陷。

这种反思的第一个倾向是重新评价有限状态模型。由于卡普兰(Kaplan)和凯依(Kay)在有限状态音系学和形态学方面的工作，以及丘奇(Church)在句法的有限状态模型方面的工作，显示了有限状态模型仍然有着强大的功能，因此，这种模型又重新得到自然语言处理学界的注意。

这种反思的第二个倾向是所谓的“重新回到经验主义”。这里值得特别注意的是语音和语言处理的概率模型的提出，这样的模型受到 IBM 公司华生研究中心的语音识别概率模型的强烈影响。这些概率模型和其他数据驱动的方法还传播到了词类标注、句法剖析、名词短语附着歧义的判定以及从语音识别到语义学的联接主义方法的研究中去。

从 20 世纪 90 年代开始，自然语言处理进入了一个新的阶段。1993 年 7 月在日本神户召开的第四届机器翻译高层会议(MT Summit IV)上，英国著名学者哈钦斯(J. Hutchins)在他的特邀报告中指出：自 1989 年以来，机器翻译的发展进入了一个新纪元。这个新纪元的重要标志是，在基于规则的技术中引入了语料库方法，其中包括统计方法，基于实例的方法，通过语料加工手段使语料库转化为语言知识库的方法，等等。这种建立在大规模真实文本处理基础上的机器翻译，是机器翻译研究史上的一场革命，它将会把自然语言处理推向一个崭新的阶段。

在过去的四十多年中，从事自然语言处理系统开发的绝大多数学者，基本上都采用基于规则的理性主义方法。这种方法主张，智能的基本单位是符号，认知过程就是在符号的表征下进行符号运算，因此，思维就是符号运算。

著名语言学家弗托(J. A. Fodor)在 *Representations* 一书中说：“只要我们认为心理过程是计算过程(因此是由表征式定义的形式操作)，那么，除了将心灵看作别的之外，还自然会把它看作一种计算机。也就是说，我们会认为，假设的计算过程包含哪些符号操作，心灵也就进行哪些符号操作。因此，我们可以大致上认为，心理操作跟图灵机的操作十分类似。”[①]弗托的这种说法代表了自然语言处理中的基于规则(符号操作)的理性主义观点。

这样的观点受到了学者们的批评。舍尔(J. R. Searle)在他的论文“Minds, Brains and

① Fodor J A. Representations, MIT Press, 1980

Programmes"[①]中，提出了所谓"中文屋子"的质疑。他提出，假设有一个懂得英文但是不懂中文的人被关在一个屋子中，在他面前是一组用英文写的指令，说明英文符号和中文符号之间的对应和操作关系。这个人要回答用中文书写的几个问题，为此，他首先要根据指令规则来操作问题中出现的中文符号，理解问题的含义，然后再使用指令规则把他的答案用中文一个一个地写出来。比如，对于中文书写的问题 Q1 用中文写出答案 A1，对于中文书写的问题 Q2 用中文写出答案 A2，如此等等。这显然是非常困难的，而且几乎是不能实现的事情，而且，这个人即使能够这样做，也不能证明他懂得中文，只能说明他善于根据规则做机械的操作而已。舍尔的批评使基于规则的理性主义的方法受到了普遍的怀疑。

理性主义方法的另一个弱点是在实践方面。自然语言处理的理性主义者把自己的目的局限于某个十分狭窄的专业领域之中，他们采用的主流技术是基于规则的句法分析技术和语义分析技术，尽管这些应用系统在某些受限的"子语言"(sub-language)中也曾经获得一定程度的成功，但是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大的困难。因为从自然语言系统所需要装备的语言知识来看，其数量之浩大和颗粒度之精细，都是以往的任何系统所望尘莫及的。而且，随着系统拥有的知识在数量上和程度上发生的巨大变化，系统在如何获取、表示和管理知识等基本问题上，不得不另辟蹊径。这样，就提出了大规模真实文本的自然语言处理问题。1990 年 8 月在芬兰赫尔辛基举行的第 13 届国际计算语言学会议(即 COLING'1990)为会前讲座确定的主题是"处理大规模真实文本的理论、方法和工具"，这说明，实现大规模真实文本的处理将是自然语言处理在今后一个相当长的时期内的战略目标。为了实现战略目标的转移，需要在理论、方法和工具等方面实行重大的革新。1992 年 6 月在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议(即 TMI-1992)上，宣布会议的主题是"机器翻译中的经验主义和理性主义的方法"。这里的所谓"理性主义"，就是指以生成转换语法为基础的基于规则的方法，所谓"经验主义"，就是指以大规模语料库的分析为基础的基于统计的方法。从中可以看出当前自然语言处理所关注的焦点。当前语料库的建设和语料库语言学的崛起，正是自然语言处理战略目标转移的一个重要标志。随着人们对大规模真实文本处理的日益关注，越来越多的学者认识到，基于语料库的分析方法(即经验主义的方法)至少是对基于规则的分析方法(即理性主义的方法)的一个重要补充。因为从"大规模"和"真实"这两个因素来考察，语料库才是最理想的语言知识资源。

在这样的情况下，人们开始深入地思考，乔姆斯基提出的形式语法规则是否是真正的语言规则？是否能够经受大量的语言事实的检验？这些形式语言规则是否应该和大规模真实文本语料库中的语言事实结合起来考虑，而不是一头钻入理性主义的牛角尖？

乔姆斯基作为一位求实求真、虚怀若谷的语言学大师，最近也开始对理性主义进行了反思，表现出与时俱进的勇气。在最近提出的"最简方案"中，他认为，所有重要的语法原则直接运用于表层，不同语言之间的差异通过词汇来处理，把具体规则减少到最低限度，开始注重对具体词汇的研究。可以看出，乔姆斯基的转换生成语法也开始对词汇重视起

① Searle J R. Minds, Brains and Programmes. Behavioral and Brain Sciences, Vol. 3, 1980