# 宗成庆《统计自然语言处理》[1]一书序言

冯志伟

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。”[2]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.”[3]

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，它涉及到语言学、计算机科学、数学、自动化技术等不同的学科。

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个交叉学科本身，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个交叉学科本身的知识。

自然语言处理已经有五十多年发展的历史了，在这五十多年的发展过程中，自然语言处理形成了自己特有的理论和方法，成为了一门独立的学科，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的。学习自然语言处理这个学科的专门知识，正如学习语言学、计算机科学、数学和自动化技术一样，非下苦功学习不可。

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个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 年春天开始，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

---

[1] 宗成庆，统计自然语言处理，清华大学出版社，2008 年。

[2] 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996 年。

[3] Bill Manaris, Natural language processing: A human-computer interaction perspective, *Advances in Computers*, Volume 47，1999)

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。《统计自然语言处理》一书，弥补了我国自然语言处理教材的这个缺陷，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，这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）在这里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，从此我就义无返顾地投身于自然语言处理的事业。岁月不饶人，将近三十年的光阴匆匆地流逝而去，当年我还是风华正茂的青年人，而今，我已经变成白发苍苍的垂垂老人了，我为这个事业坎坷地奋斗了大半生时间，其间的甘苦有谁知道呢？三十年来，不论是处于顺境还是逆境之中，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 和 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，成为了忘年之交的好朋友。

宗成庆博士完稿之后，也许他知道我对于 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，又是这个领域的专门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。1 至 9 章主要介绍统计自然语言处理的理论，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。语言模型和隐马尔柯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，又介绍了隐马尔柯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部门进行系统的、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、图象识别和理解技术、情感计算（affective 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的发言中曾经说过：“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，当时杰里内克讲的话更为尖刻，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。[4]

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（ACL2002）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 ACL2002 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，他使我们看到了未来的机器翻译的曙光，令人鼓舞。[5]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，他却认为：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

---

[4] M. Palmer and T. Finin, workshop on the evaluation of natural language processing systems, Computational Linguistics, 16(3), 175-181, 1990.

[5] 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 年，第 1 期（总 21 期）。

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，当近代哲学家们把这两种倾向的冲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，无数的哲学家们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，在人们的思想中都有不同程度、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，并且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，则是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，他说，“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”[6]。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，人心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”[7]。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”[8]，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”[9]，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”[10]。他把经验归纳作为科学研究的一般方法论原理，认为，“实验科学只能从现象出发，并

---

[6] 《十六——十八世纪西欧各国哲学》，第 23 页。
[7] 洛克，《人类理解论》，商务印书馆，第 598 页。
[8] 休谟，《人类理解研究》，商务印书馆，第 27 页。
[9] 《十六——十八世纪西欧各国哲学》，第 634 页。
[10] 塞耶编，《牛顿自然哲学著作选》，商务印书馆，第 212 页。

且只能用归纳来从这些现象中推演出一般的命题”[11]。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”[12]。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，出现了笛卡尔（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。笛卡尔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。笛卡尔是法国哲学家，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，俄国科学家马尔柯夫（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，提出了马尔柯夫随机过程理论，建立了马尔柯夫模型，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）把离散马尔柯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 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。[13]

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柯夫过程的思想，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

---

[11] 塞耶编，《牛顿自然哲学著作选》，商务印书馆，第 8 页。

[12] 《十八世纪法国哲学》，商务印书馆，第 59 页，1963 年。

[13] 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。

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“笛卡尔语言学”（Descartes linguistics），充分地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”[14]。他主张采用公理化、形式化的方法，严格地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 年代时期在国际语言学界风靡一时，转换生成语法对于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），或者叫做符号主义的方法（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的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

---

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

验主义的传统，多使用归纳法（induction）而很少使用演绎法（deduction）。

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在《联邦主义者》（The Federalist）文章中的原作者的分布问题，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 1963－64 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柯夫模型（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 的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：隐马尔柯夫模型、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 年代初期的这一个时期，在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 80 年代初期发生了变化。在 1983－1993 年的十年中，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。在这十年中，自然语言处理的研究又回到了 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 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）在他的特约报告中指出，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（MIT Press, 1980）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。"[15] 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 Programmes》[16]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的。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'90）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-92）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

---

[15] J. A. Fodor, Representations, MIT Press, 1980.

[16] J. R. Searle, Minds, Brains and Programmes, 载 Behavioral and Brain Sciences, Vol. 3, 1980.

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，最近他也开始对于理性主义进行了反思，表现了与时俱进的勇气。在最近他提出的“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，把具体的规则减少到最低限度，开始注重对具体的词汇的研究。可以看出，乔姆斯基的转换生成语法也开始对词汇重视起来，逐渐地改变了原来的理性主义的立场，开始与经验主义妥协，或者悄悄地向经验主义复归。

在 20 世纪 90 年代的最后五年（1994-1999），自然语言处理的研究发生了很大的变化，出现了空前繁荣的局面。概率和数据驱动的方法几乎成为了自然语言处理的标准方法。句法剖析、词类标注、参照消解和话语处理的算法全都开始引入概率，并且采用从语音识别和信息检索中借过来的评测方法，统计方法已经渗透到了机器翻译、文本分类、信息检索、问答系统、信息抽取、语言知识挖掘等自然语言处理的应用系统中去，基于统计的经验主义方法逐渐成为自然语言处理研究的主流。

可以看出，在自然语言处理发展的过程中，始终充满了基于规则的理性主义方法和基于统计的经验主义方法之间的矛盾，这种矛盾时起时伏，此起彼伏。自然语言处理也就在这样的矛盾中逐渐成熟起来。

总结自然语言处理发展的曲折历史可以看出，基于规则的理性主义方法和基于统计的经验主义方法各有千秋，因此，我们应当用科学的态度来分析它们的优点和缺点。

我们认为，基于规则的理性主义方法的优点是：

- 基于规则的理性主义方法中的规则主要是语言学规则，这些规则的形式描述能力和形式生成能力都很强，在自然语言处理中有很好的应用价值。
- 基于规则的理性主义方法可以有效地处理句法分析中的长距离依存关系（long-distance dependencies）等困难问题，如句子中长距离的主语和谓语动词之间的一致关系（subject-verb agreement）问题，wh 移位（wh-movement）问题。
- 基于规则的理性主义方法通常都是明白易懂的，表达得很清晰，描述得很明确，很多语言事实都可以使用语言模型的结构和组成成分直接地、明显地表示出来。
- 基于规则的理性主义方法在本质上是没有方向性的，使用这样的方法研制出来的语言模型，既可以应用于分析，也可以应用于生成，这样，同样的一个语言模型就可以双向使用。
- 基于规则的理性主义方法可以在语言知识的各个平面上使用，可以在语言的不同维度上得到多维的应用。这种方法不仅可以在语音和形态的研究中使用，而且，在句法、语义、语用、篇章的分析中也大显身手。
- 基于规则的理性主义方法与计算机科学中提出的一些高效算法是兼容的，例如，计算机算法分析中使用 Earley 算法（1970 年提出）和 Marcus 算法（1978 年提出）都可以作为基于规则的理性主义方法在自然语言处理中得到有效的使用。

基于规则的理性主义方法的缺点是：

- 基于规则的理性主义方法研制的语言模型一般都比较脆弱，鲁棒性很差，一些与语言模型稍微偏离的非本质性的错误，往往会使得整个的语言模型无法正常地工作，甚至导致严重的后果。不过，近来已经研制出一些鲁棒的、灵活的剖析技术，这些技术能够使基于规则的剖析系统在剖析失败中得到恢复。
- 使用基于规则的理性主义方法来研制自然语言处理系统的时候，往往需要语言学家、语音学家和各种专家的配合工作，进行知识密集的研究，研究工作的强度很大；基于规则的语言模型不能通过机器学习的方法自动地获得，也无法使用计算机自动

地进行泛化。

- 使用基于规则的理性主义方法设计的自然语言处理系统的针对性都比较强，很难进行进一步的升级。例如，斯罗肯（Slocum）在 1981 年曾经指出，LIFER 自然语言知识处理系统在经过两年的研发之后，已经变得非常之复杂和庞大，以至于这个系统原来的设计人很难再对它进行一点点的改动。对于这个系统的稍微改动将会引起整个连续的“水波效应”（ripple effect），以至于“牵一发而动全身”，而这样的副作用是无法避免和消除的。
- 基于规则的理性主义方法在实际的使用场合其表现往往不如基于统计的经验主义方法那样好。因为基于统计的经验主义方法可以根据实际训练数据的情况不断地优化，而基于规则的理性主义方法很难根据实际的数据进行调整。基于规则的方法很难模拟语言中局部的约束关系，例如，单词的优先关系对于词类标注是非常有用的，但是基于规则的理性主义方法很难模拟这种优先关系。

不过，尽管基于规则的理性主义方法有这样的或那样的不足，这种方法终究是自然语言处理中研究得最为深入的技术，它仍然是非常有价值和非常强有力的技术，我们决不能忽视这种方法。事实证明，基于规则的理性主义方法的算法具有普适性，不会由于语种的不同而失去效应，这些算法不仅适用于英语、法语、德语等西方语言，也适用于汉语、日语、韩国语等东方语言。在一些领域针对性很强的应用中，在一些需要丰富的语言学知识支持的系统中，特别是在需要处理长距离依存关系的自然语言处理系统中，基于规则的理性主义方法是必不可少的。

我们认为，基于统计的经验主义方法的优点是：

- 使用基于统计的经验主义方法来训练语言数据，从训练的语言数据中自动地或半自动地获取语言的统计知识，可以有效地建立语言的统计模型。这种方法在文字和语音的自动处理中效果良好，在句法自动分析和词义排歧中也初露锋芒。
- 基于统计的经验主义方法的效果在很大的程度上依赖于训练语言数据的规模，训练的语言数据越多，基于统计的经验主义方法的效果就越好。在统计机器翻译中，语料库的规模，特别是用来训练语言模型的目标语言语料库的规模，对于系统性能的提高，起着举足轻重的作用。因此，可以通过扩大语料库规模的办法来不断提高自然语言处理系统的性能。
- 基于统计的经验主义方法很容易与基于规则的理性主义方法结合起来，从而处理语言中形形色色的约束条件问题，使自然语言处理系统的效果不断地得到改善。
- 基于统计的经验主义方法很适合用来模拟那些有细微差别的、不精确的、模糊的概念（如“很少、很多、若干”等），而这些概念，在传统语言学中需要使用模糊逻辑（fuzzy logic）才能处理。

基于统计的经验主义方法的缺点是：

- 使用基于统计的经验主义方法研制的自然语言处理系统，其运行时间是与统计模式中所包含的符号类别的多少成比例线性地增长的，不论在训练模型的分类中还是在测试模型的分类中，情况都是如此。因此，如果统计模式中的符号类别数量增加，系统的运行效率会明显地降低。
- 在当前语料库技术的条件下，要使用基于统计的经验主义方法为某个特殊的应用领域获取训练数据，还是一件费时费力的工作，而且很难避免出错。基于统计的经验主义方法的效果与语料库的规模、代表性、正确性以及加工深度都有密切的关系，可以说，用来训练数据的语料库的质量在很大的程度上决定了基于统计的经验主义方法的效果。
- 基于统计的经验主义方法很容易出现数据稀疏的问题，随着训练语料库规模的增

大，数据稀疏的问题会越来越严重，这个问题需要使用各种平滑（smoothing）技术来解决。

自然语言中既有深层次的现象，也有浅层次的现象，既有远距离的依存关系，也有近距离的依存关系，自然语言处理中既要使用演绎法，也要使用归纳法。因此，我们主张把理性主义和经验主义结合起来，把基于规则的方法和基于统计的方法结合起来。我们认为，强调一种方法，反对另一种方法，都是片面的，都无助于自然语言处理的发展。

英国经验主义哲学家培根既反对理性主义，也反对狭隘的经验主义，他指出，由于经验能力和理性能力这两方面的“离异”和“不和”，给科学知识的发展造成了严重的障碍，为了克服这样的弊病，他提出了经验能力和理性能力联姻的重要原则。他说，“我以为我已经在经验能力和理性能力之间永远建立了一个真正合法的婚姻，二者的不和睦与不幸的离异，曾经使人类家庭的一切事务陷于混乱”[17]。他生动而深刻地说道：“历来处理科学的人，不是实验家，就是教条者。实验家像蚂蚁，只会采集和使用；推论家像蜘蛛，只凭自己的材料来织成丝网。而蜜蜂却是采取中道的，它在庭园里和田野里从花朵中采集材料，而用自己的能力加以变化和消化。哲学的真正任务就正是这样，它既非完全或主要依靠心的能力，也非只把从自然历史和机械实验收来的材料原封不动，囫囵吞枣地累置于记忆当中，而是把它们变化过和消化过放置在理解力之中。这样看来，要把这两种机能、即实验的和理性的这两种机能，更紧密地和更精纯地结合起来（这是迄今还未收到的），我们就可以有很多的希望”[18]。

培根的主张是值得我们深思的。在自然语言处理的研究中，我们不能采取像蜘蛛那样的理性主义方法，单纯依靠规则，也不能采取像蚂蚁那样的经验主义方法，单纯依靠统计，我们应当像蜜蜂那样，把理性主义和经验主义两种机能更紧密地、更精纯地结合起来，推动自然语言处理的发展。

本书讲述的是统计自然语言处理的经验主义方法，这些方法只是自然语言处理的一个方面。我们在阅读本书的同时，不要忘记在自然语言处理中还存在着另外一个方面，这就是基于规则的理性主义方法，我们也应当学习这些基于规则的理性主义方法，并且把这两种方法结合起来，彼此取长补短，使之相得益彰。这样，我们对于自然语言处理这个学科，就可以获得全面而完整的认识。

尽管本书的题目是《统计自然语言处理》，但是，本书作者并不偏袒基于统计的经验主义方法而排斥基于规则的理性主义方法，他对于经验主义和理性主义之间关系的认识是非常清楚的，他说：“尽管目前统计机器翻译研究进展迅速，却并没有一个确切的结论告诉人们究竟哪一种模型和方法可以绝对地取代其它任何模型和方法，或者证明哪一种模型可以被彻底淘汰。而从近期的研究成果来看，多种模型和特征的结合，尤其是句法结构信息的利用，已经成为改进和提高统计翻译系统性能的有效途径，这实际上从另一个角度印证了多种方法结合的必要性和有效性。”他强烈主张：在机器翻译问题彻底解决以前，永远没有过时的理论和方法，也决不应该有哪一种方法可以“藐视天下，惟我独尊”。对于宗成庆博士的这种真知灼见，我举双手赞成。

2006 年 8 月 25 日于北京

[17] 《十六——十八世纪西欧各国哲学》，第 8 页。
[18] 培根，《新工具》，商务印书馆，第 75 页。