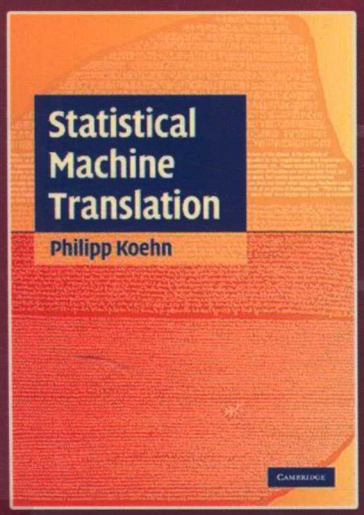
国外计算机科学教材系列

CAMBRIDGE

统计机器翻译

Statistical Machine Translation



[德] Philipp Koehn 著

宗成庆 张霄军 译

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

国外计算

统计机器翻译

Statistical Machine Translation

[德] Philipp Koehn 著

宗成庆 张霄军 译

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书是介绍统计机器翻译理论和方法的教材。全书分三部分(共 11 章),分别讨论基础知识、核心方法和前沿研究。全书首先简要介绍语言学和概率论基础知识,然后全面讨论各种经典统计机器翻译模型和系统实现方法,最后深入探讨统计翻译领域的最新进展和研究热点。对核心方法的论述按照统计机器翻译模型发展的过程逐步展开:基于词的模型、基于短语的模型和基于句法树的模型。从技术实现的角度,本书还介绍了统计翻译模型的参数训练方法、语言模型和参数平滑方法、解码算法和译文自动评测方法及系统整合方法等。

本书是统计机器翻译和自然语言处理课程的理想教材,适合研究生和本科生教学使用,也是所有对机器翻译技术和系统有兴趣的研究者、开发者和使用者的指南性读物。同时,本书还可作为人工智能、语言学等相关专业的辅助读物。

版权合同登记号　图字:01-2011-8104

图书在版编目(CIP)数据

统计机器翻译/(德)科恩(Koehn,P.)著;宗成庆,张霄军译. —北京:电子工业出版社,2012.9
(国外计算机科学教材系列)
书名原文:Statistical Machine Translation
ISBN 978-7-121-17592-3

I. 统…　II. ①科…　②宗…　③张…　III. ①机器翻译-翻译机-研究　IV. ①H085　②TP319.2

中国版本图书馆 CIP 数据核字(2012)第 157998 号

策划编辑:马　岚
责任编辑:刘娴庆
印　　刷:
装　　订:三河市鑫金马印装有限公司
出版发行:电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本:787×1092　1/16　印张:20.25　字数:525 千字　　彩页:2
印　　次:2012 年 9 月第 1 次印刷
定　　价:55.00 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888。

质量投诉请发邮件至 zlts@phei.com.cn,盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010)88258888。

作者写给中国读者的寄语

With the increase in global exchange, translating between languages is becoming more and more important. Especially the strengthening ties between China and the West, both economic and cultural, pose a significant challenge.

Statistical machine translation is hoping to make a contribution by making it easier to bridge the language barrier. In fact, there is a significant effort for the development of automatic translation technology between Chinese and English, carried out by research groups both in China and the West.

It is therefore my great pleasure to see a Chinese translation of my textbook to further these efforts. China is already at the forefront on statistical machine translation research, and hopefully this book can make a contribution to further push stimulate this important work ahead.

Philipp Koehn
Edinburgh, September 6,2011

随着全球化交流的增强，语言之间的翻译正变得越来越重要。尤其是随着中国与西方国家之间经济和文化联系的不断加强，对翻译的挑战也越来越大。

统计机器翻译就是希望为架设跨越语言障碍的桥梁变得更加容易而做出贡献。实际上，汉语和西方国家的众多研究组为研发汉语和英语之间的自动翻译技术已经做出了巨大的努力。

因此，能够看到我的教科书被翻译成中文，继续为机器翻译研究发挥作用，我感到非常高兴！中国已经处于统计机器翻译研究的前沿，我希望这本书能够为进一步推动这项重要工作不断前进做出贡献。

菲利普·科恩
2011 年 9 月 6 日于爱丁堡

译者介绍

宗成庆 1998年3月毕业于中国科学院计算技术研究所，获博士学位。1998年5月至2000年4月在中国科学院自动化研究所从事博士后研究，博士后出站后留自动化所工作至今，现为模式识别国家重点实验室研究员、博士生导师。曾于1999年和2001年两次在日本国际电气通信基础技术研究所(ATR)做客座研究员，2004年在法国Grenoble信息与应用数学研究院机器翻译研究组(GETA-CLIPS, IMAG)做短期高访。

主要研究方向为自然语言处理基础、机器翻译、文本分类和自动文摘等相关技术。作为项目负责人承担国家自然科学基金项目、国家"863"项目、国家支撑计划项目和国际合作研究项目等10余项，在国内外重要学术期刊和会议上发表论文100余篇，其中在*Computational Linguistics*、*Information Sciences*、*IEEE TASLP*、*ACM TALIP*、*Machine Translation*及ACL、COLING、EMNLP等本领域权威期刊和会议上发表论文20多篇，出版学术专著1部，获8项国家发明专利。目前担任国际计算语言学联合会(ACL)汉语特别兴趣组(SIGHAN)候任主席(Chair Elect)和亚洲自然语言处理联合会(AFNLP)执行理事，并担任国际学术期刊*IEEE Intelligent Systems*副主编(Associate Editor)、*ACM TALIP*副主编、*IJCPOL*副主编、*Machine Translation*编委、*JCST*编委、《自动化学报》编委，以及中国中文信息学会常务理事、中国人工智能学会理事和中国计算机学会中文信息技术专委会副主任等职务。2008年获中国科学院研究生院集中教学突出贡献奖。2009年获亚太地区语言、信息与计算国际会议(PACLIC)最佳论文奖，2010年获中国科学院"朱李月华优秀教师"奖。

张霄军 2008年6月毕业于南京师范大学，获博士学位。现为陕西师范大学外国语学院副教授，硕士生导师。2010年至2011年在英国曼彻斯特大学访学，研究方向为现代翻译技术。目前承担国家社科基金项目1项，参与国家自然科学基金项目1项及国家社科基金项目1项。在国际学术期刊*Computational Linguistics*、*Information Retrieval*、*Language Learning & Technology*和*Applied Linguistics*等发表学术论文4篇，在《当代语言学》和《计算机应用研究》等国内期刊发表学术论文50余篇。出版学术专著《语义组合与机器翻译》(科学出版社，2010)，主(参)编教材多部。

序

机器翻译(Machine Translation, MT)是采用电子计算机进行自然语言之间的翻译的一门新兴的实验性学科。这门学科兴起于20世纪50年代初，20世纪60年代中期曾一度低落，20世纪60年代后期又重新兴旺起来，现在仍在不断发展中。

机器翻译是计算语言学的一个应用领域，它的研究建立在语言学、数学和计算技术这三门学科的基础之上。语言学家提供适合于机器进行加工的词典和语法规则，数学家把语言学家提供的材料形式化和代码化，计算技术专家给机器翻译提供软件手段和硬件设备。缺少上述任何一方面，机器翻译就不能实现。机器翻译效果的好坏，也完全取决于上述三方面的共同努力。

机器翻译大致可分为基于规则的机器翻译(Rule-Based Machine Translation, RBMT)和基于语料库的机器翻译(Corpus-Based Machine Translation, CBMT)两种。

基于规则的机器翻译过程一般可分为分析、转换、生成三个阶段，具体地说，这三个阶段如下所示。

- 原文分析。分析原文的形态和句法结构；
- 原文译文转换。把原文词转换为译文词，并进行原文和译文之间的结构转换；
- 译文生成。生成译文的句法和形态，输出译文。

我们把分析阶段记为A(Analysis)，把转换阶段记为T(Transformation)，把生成阶段记为G(Generation)。

在基于规则的机器翻译系统中，有的把转换与分析结合起来，即从双语言转换的角度来进行原语分析，这样的系统称为相关分析、独立生成系统，可表示为AT→G；有的把转换与生成结合起来，独立地进行原文分析，从双语言转换的角度来生成译文，这样的系统称为独立分析、相关生成系统，可表示为A→TG；有的把分析、转换、生成分开来，原文分析时不考虑译文，译文生成时不考虑原文，原译文的差异通过转换来解决，这样的系统称为独立分析、独立生成系统，可表示为A→T→G。

选择哪一种系统要根据具体情况来决定。一般来说，一对一机器翻译(把一种语言译为另一种语言)或多对一机器翻译(把多种语言翻译为一种语言)宜于采用相关分析、独立生成系统；一对多机器翻译(把一种语言翻译为多种语言)宜于采用独立分析、相关生成系统；多对多机器翻译(把多种语言翻译为另外的多种语言)宜于采用独立分析、独立生成系统。为了深入地研究原文和译文各自的语法特点，目前的倾向是把原文分析与译文生成分开，越来越多的机器翻译研制采用了独立分析、独立生成系统，源语言和目标语言之间的差别通过转换(transfer)的方法来解决。

自从1954年美国乔治敦大学的第一次机器翻译试验以来，基于规则的机器翻译研究有了很大的进展，人们一般把基于规则的机器翻译的发展过程分为三代。

第一代机器翻译：以词汇转换为主的机器翻译。翻译时主要是把原文的词转换为译文的词，这种机器翻译的译文质量极低。

第二代机器翻译:以句法为主的机器翻译。翻译时除了进行词汇转换之外，还着重研究句法结构的分析和句法结构的生成。这种机器翻译的译文质量有显著提高。第二代机器翻译采用了数理语言学的一些理论，有的系统采用了形式语法(Formal Grammar)和自动机理论(Automata Theory)，有的系统采用了语言的集合论模型(Set-Theory Model)、有的系统采用了生成转换语法(Generative Transformational Grammar)，有的系统采用了依存语法(Dependency Grammar)，有的系统采用了蒙德鸠文法(Montague Grammar)。他们所用的理论虽有不同，但其共同点是以句法分析为主。

第三代机器翻译:以语义为主的机器翻译。翻译时，先对原文进行语义分析，得出原文的语义内容，然后再把这种语义内容用译文的文本表示出来。

目前，国内外大多数基于规则的机器翻译系统都是以句法为主的机器翻译，由于语义的形式表示十分困难，以语义为主的机器翻译只进行了一些很初步的探索，在这方面的许多研究与自然语言理解有着密切的关系。

在过去的50多年中，从事计算语言学系统开发的绝大多数学者，都把自己的目的局限于某个十分狭窄的专业领域中，他们采用的主流技术是基于规则的句法-语义分析，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远远不及的。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动处理问题。

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，将是计算语言学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①。

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(TMI-92)的主题是“机器翻译中的经验主义和理性主义的方法”。所谓“理性主义”，就是指以生成语言学为基础的方法，所谓“经验主义”，就是指以大规模语料库的分析为基础的方法。从中可以看出当前计算语言学关注的焦点。当前语料库的建设和语料库语言学的崛起，正是计算语言学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。但是，要想使语料库名副其实地成为自然语言的知识库，就有必要首先对语料库中的语料进行自动标注，使之由“生语料”变成“熟语料”，以便于人们从中提取丰富的语言知识。这样的机器翻译就是基于语料库的机器翻译。

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

① 冯志伟，论语言学研究中的战略转移，《现代外语》，第34卷，第1期，p1-11，2011年。

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，它将把机器翻译推向一个崭新的阶段。

基于语料库的机器翻译可以分为基于实例的机器翻译(Example-Based Machine Translation，EBMT)和统计机器翻译(Statistical Machine Translation，SMT)两种。这两种机器翻译都使用语料库作为翻译知识的来源，所以可以统称为基于语料库的机器翻译。在统计机器翻译中，知识的表示是统计数据，而不是语料库本身，翻译知识的获取是在翻译之前完成的，翻译的过程中不再使用语料库。在基于实例的机器翻译中，双语语料库本身就是翻译知识的一种表现形式(不一定是唯一的)，翻译知识的获取在翻译之前没有全部完成，在翻译的过程中还要查询并利用语料库。

本书论述的是统计机器翻译。

统计机器翻译的思想并不是在20世纪90年代才产生的，在机器翻译产生的初期，就有学者提出了采用统计方法进行机器翻译的思想了。

1949年，美国洛克菲勒基金会副总裁Weaver发表了一份以Translation为题的备忘录①，提出了机器翻译问题。在这份备忘录中，他除了提出各种语言都有许多共同的特征这一论点之外，还有如下两点值得我们注意。

第一，他认为翻译类似于解读密码的过程。他说："当我阅读一篇用汉语写的文章时，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，当我在阅读时，我是在进行解码。"②

在这段话中。Weaver首先提出了用解读密码的方法进行机器翻译的想法，这种想法成为后来噪声信道理论的滥觞。备忘录中还记载了一个有趣的故事，布朗大学数学系的R. E. Gilmam曾经解读了一篇长约一百个词的土耳其文密码，而他既不懂土耳其文，也不知道这篇密码是用土耳其文写的。Weaver认为，Gilmam的成功足以证明解读密码的技巧和能力不受语言的影响，因而可以用解读密码的办法来进行机器翻译。

第二，他认为原文与译文"说的是同样的事情"。因此，当把语言A翻译为语言B时，就意味着，从语言A出发，经过某一"通用语言"(Universal Language)或"中间语言"(Interlingua)，转换为语言B，这种"通用语言"或"中间语言"可以假定是全人类共同的。

这是文献中关于统计机器翻译思想的最早论述，但由于当时尚缺乏高性能的计算机和联机语料，采用基于统计的机器翻译在技术上还不成熟。Weaver的这种方法是难以付诸实现的。

现在，这种局面已经大大改变了，计算机在速度和容量上都有了大幅度的提高，也有了大量的联机语料可供统计使用，因此，在20世纪90年代，基于统计的机器翻译又兴盛起来。

在Weaver思想的基础上，IBM公司的Peter Brown等人提出了统计机器翻译的数学模型。

基于统计的机器翻译把机器翻译问题看成一个噪声信道(noisy channel)问题。可以使用噪声信道的观点来看机器翻译：一种语言S由于经过了一个噪音信道而发生了扭曲变形，在信道的另一端呈现为另一种语言T，翻译问题实际上就是如何根据观察到的语言T，恢复最

---

① W. Weaver, Warren Weaver's memorandum in 1949: Translation, Milestones in Machine Translation. In Locke, W. N. and Booth, A. D. (eds.) *Machine translation of languages: fourteen essays*, Cambridge, Mass. Technology Press of the Massachusetts Institute of Technology, 1955.

② 此处"用汉语写的文章"在有的引文中为"用俄语写的文章"，经查对原文，"用俄语"应更正为"用汉语"。原文参见W. Weaver, Warren Weaver's memorandum in 1949: Translation, Milestones in Machine Translation。

为可能的语言 S。语言 S 是信道意义上的输入，在翻译意义上就是目标语言，语言 T 是信道意义上的输出，在翻译意义上就是源语言。

根据这种观点，一种语言中的任何一个句子都有可能是另一种语言中的某几个句子的译文，只是这些句子的可能性各不相同，机器翻译就是要找出其中可能性最大的句子，也就是将对所有可能的目标语言 S 计算出的概率最大的一个作为源语言 T 的译文。由于 S 的数量巨大，可以采用栈式搜索（stack search）的方法。栈式搜索的主要数据结构是表结构（list structure），表结构中存放着当前最有希望的对应于 T 的 S，算法不断循环，每次循环扩充一些最有希望的结果，直到表中包含一个得分明显高于其他结果的 S 时结束。栈式搜索不能保证得到最优的结果，它会导致错误的翻译，因而只是一种次优化算法。

可见，统计机器翻译系统的任务就是在所有可能的目标语言（翻译意义上的目标语言，也就是噪声信道模型意义上的源语言）的句子中，寻找概率最大的那个句子作为翻译结果。其概率值可以使用贝叶斯公式得到（下面公式中的 T 是在翻译意义上的目标语言，S 是在翻译意义上的源语言）：

$$P(T|S) = \frac{P(T)P(S|T)}{P(S)}$$

由于等式右边的分母 $P(S)$ 与 T 无关，因此求 $P(T|S)$ 的最大值相当于寻找一个 $\hat{T}$，使等式右边分子的两项乘积 $P(T)P(S|T)$ 为最大，也就是说：

$$\hat{T} = \arg\max_T P(T)P(S|T)$$

其中，$P(T)$ 是目标语言的语言模型，$P(S|T)$ 是给定 T 的情况下 S 的翻译模型。根据语言模型和翻译模型，求解在给定源语言句子 S 的情况下最接近真实的目标语言句子 $\hat{T}$ 的过程，相当于噪声信道模型中解码的过程。

统计机器翻译翻译系统要解决三个问题：

1. 估计语言模型概率 $P(T)$；
2. 估计翻译概率 $P(S|T)$；
3. 设计有效快速的搜索算法来求解 $\hat{T}$，使得 $P(T)P(T|S)$ 最大。

著名学者刘复提出，翻译应当遵从“信”（faithfulness）、“达”（expressiveness）、“雅”（elegance）三个标准。

估计翻译概率 $P(S|T)$，就是要求目标语言与源语言保持一致，目标语言应完整地传达出源语言的思想，相当于“信”。

估计语言模型概率 $P(T)$，就是要求目标语言通顺可读，相当于“达”。

当然，我们不能要求机器翻译达到“雅”。

英国近代翻译理论家 A. Tytler（泰特勒，1747—1814）的 *Essay on the Principle of Translation*（论翻译的原则）提出了翻译的三原则，即：

1. 译文应完整地传达出原作的思想（A translation should give a complete transcript of the ideas of the original work）；
2. 译文的风格与笔调和原作性质相同（The style and manner of writing should be of the same character as that of the original）；
3. 译文应与原作同样流畅（A translation should have all the ease of the original composition）。

这三个原则，与我国学者严复提出的“信”“达”“雅”有异曲同工之妙，严氏“信”“达”“雅”曾经被美国翻译理论专家 Nide(奈达)推崇为“翻译三原则”(for Chinese translators Yan Fu's triple principle of translation)。可见，“信”“达”“雅”应当成为判定翻译质量的重要依据。

统计机器翻译的基本公式已经反映了翻译三原则中的“信”和“达”两个原则，正在逐渐向翻译三原则靠拢。这是值得我们高兴的。

鲁迅把严复的“信”“达”“雅”三个标准简化为“顺”和“信”两个标准。根据常识，好的机器翻译的译文应当是流畅的，同时又应当是忠实于源语言的，也就是说，既要“顺”，又要“信”。鲁迅的“顺”这个标准反映了语言模型的要求，“信”这个标准反映了翻译模型的要求。

在统计机器翻译中联合使用语言模型和翻译模型，既考虑了译文的“顺”，又考虑了译文的“信”，其效果应该比单独使用翻译模型更好，如果仅仅考虑翻译模型，由于只考虑了“信”而忽视了“顺”，就常常会导致一些不通顺的译文。

年轻的德国亚琛大学博士研究生 Franz Josef Och 在国际计算语言学 2002 年的会议(ACL2002)上发表论文，题目是“统计机器翻译的分辨训练与最大熵模型”[①]，提出了统计机器翻译的系统性方法，获 ACL2002 大会最佳论文奖。

2003 年夏天，在约翰·霍普金斯大学的暑假机器翻译讨论班(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、宾西法尼亚州大学、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 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”[②]，这个报告提出了把基于规则的方法和基于统计方法结合起来的有效途径。

2003 年 7 月，在美国马里兰州巴尔的摩市由美国商业部国家标准与技术研究所 NIST/TIDES 主持的评比中，Och 获最好成绩，他使用统计方法，在很短的时间内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说过：“Give me a place to stand on, and I will move the world.”(只要给我一个支点，我就可以移动地球)，而现在 Och 也模仿着 Archimedes 说：“Give me enough parallel data, and you can have translation system for any two languages in a matter of hours.”(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)。这反映了新一代的机器翻译研究者朝气蓬勃的探索精神和继往开来的豪情壮志。看来，Och 似乎已经找到了机器翻译的有效方法，至少按照他的路子走下去，也许有可能开创出机器翻译研究的一片新天地，使我们在探索真理的曲折道路上看到耀眼的曙光。过去研制一个机器翻译系统往往需要几年的时间，而现在采用 Och 的方法构造机器翻译系统只要几个小时就可以了，研制机器翻译系统的速度已经大大地提高了。

目前，统计机器翻译已经成为机器翻译研究的主流。

越来越多的互联网和软件公司都推出了基于统计的在线机器翻译系统。例如，谷歌的多语言在线机器翻译系统 Google Translator 可翻译的语言有 58 种，翻译方向有 $58 \times 57 = 3306$ 个，也就是说，这个系统可以进行 3306 个语言对的翻译工作，这样的工作显然是由人来翻译难以

---

① J. Och and H. Ney, Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, *In Proceedings of ACL-02*, pp. 295-302, 2002.

② J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, D. Radev, Final Report of John Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation, 2003.

胜任的。如果用户不知道文本的语言是哪种语言，Google Translator 系统还可以帮助用户进行检测，根据文本中字母的同现概率来判定该文本究竟属于哪种语言，从而进行机器翻译，这大大地方便了讲不同语言的人们在互联网上的沟通。可以看出，统计机器翻译目前已经取得了令人瞩目的可喜成绩，值得我们高度关注。

本书全面介绍了统计机器翻译的基础知识和核心方法，并探讨了统计机器翻译中的一些高级问题，是一本学习统计机器翻译的好书。

全书分三部分。第一部分包括第 1 章至第 3 章，讲述统计机器翻译的基础知识；第二部分包括第 4 章至第 8 章，讲述统计机器翻译的核心方法；第三部分包括第 9 章至第 11 章，讲述统计机器翻译的前沿研究。各章内容分述如下。

第一部分讲述统计机器翻译的基础知识，介绍了机器翻译的历史、基本语言学知识、概率论知识等。

第 1 章简要叙述了机器翻译发展的历史，特别说明了统计机器翻译的背景及最新的发展情况，介绍了统计机器翻译的应用状况，提供了一份丰富的语言资源清单。本章着重指出，机器翻译技术的应用价值将随着翻译质量的提升而提升，机器翻译并不要求译文的完美，粗略的译文也可以传递信息，因此也是有实用价值的。

第 2 章介绍了词类、形态学、句法、语义、篇章和语料库等基础知识。简要描述齐夫定律、短语结构语法、依存语法、词汇功能语法、组合范畴语法等形式化模型，并特别指出语体和领域的差别会影响统计机器翻译的效果。

第 3 章介绍了概率论的基本概念，如均值、方差、二项分布、正态分布、联合概率、条件概率和熵等。这些概念对于本书后面要讲述的统计机器翻译，都是非常重要的。

第二部分讲述统计机器翻译的核心方法，包括基于词的翻译模型、基于短语的翻译模型、解码、语言模型和评测等。

第 4 章重点介绍了 IBM 模型。IBM 模型 1 只使用了词汇翻译概率，模型 2 增加了绝对对齐模型，模型 3 增加了繁衍率模型，模型 4 将绝对对齐模型替换为相对对齐模型，模型 5 修正了模型中的缺陷问题，将概率值分配给那些不可能的对齐。

第 5 章介绍了基于短语的统计机器翻译模型，这种模型把短语作为翻译的原子单元。在短语翻译表中，短语之间是一一映射的，也可能存在调序。短语翻译表可以从词对齐中通过机器学习而自动得到，与词对齐一致的所有短语偶对都被添加到短语翻译表中。本章还介绍了一种可以直接从双语平行语料中自动学习短语对齐的替代方法。在短语翻译表中可能存在调序，因此本章也介绍了一个简单的基于距离的调序模型，给出了一个词汇化的调序模型，并使用对数线性模型来融合短语模型中不同的模型组件。在扩展原始的翻译模型时，可以引入额外的模型组件，这些组件包括双向翻译概率、词汇化加权、词惩罚和短语惩罚。

第 6 章介绍了统计机器翻译中的解码算法，对于一个给定的输入句子，使用解码算法进行搜索，就可以找到最有可能的翻译结果。由于搜索空间具有指数级的算法复杂度，需要采用启发式的搜索方法。本章描述了从输入到输出构建翻译的过程，并将其作为搜索算法的动因。在统计机器翻译中，对于给定的输入句子，必须处理很多翻译选项；搜索是建立在一连串的翻译假设上的，从没有翻译任何单词的空假设开始，进行假设扩展，以建立新的假设。翻译假设重组可以减少搜索空间。本章还提出了一种组织栈解码的启发式方法，根据已经翻译过的外语单词的数量，在假设栈里对于翻译假设进行组织。利用剪枝策略对栈空间进行压缩，介绍了直方图剪枝和阈值剪枝两种剪枝方法。本章还介绍了其他一些启发式搜索算法，

例如基于覆盖栈的柱搜索算法，A*搜索算法，贪婪爬山解码算法。本章最后还介绍了有限状态转换机工具包。

第7章介绍了高效使用语言模型的方法。使用语言模型可以为每个给定的英语单词序列计算出一个概率，用于表示该序列在英语中被表达的可能性，从而帮助机器翻译系统产生出流利的译文。可以将语言模型问题分解为一系列利用 n 元文法的统计信息来预测单词的问题。这样的语言模型称为马尔可夫链。在马尔可夫链中，只有前面有限的 $n-1$ 个单词状态会影响当前单词的状态，n 的大小称为语言模型的阶；大小为1、2和3的 n 元文法分别称为一元文法、二元文法和三元文法。由于在有限的训练语料中不能观察到所有可能的 n 元文法，所以必须处理数据稀疏的问题。可以通过平滑经验计数的方法来处理数据稀疏的问题。本章介绍的平滑算法有加1平滑算法、删除估计平滑法、古德-图灵平滑法、插值方法、Witten-Bell平滑法和Kneser-Ney平滑法等。

第8章讲述了如何评测机器翻译系统的性能。由于源语言中的一个句子可能有很多不同的正确翻译，因此译文评测是个很困难的问题。在评测时可以提供一些参考译文，但不能期望机器翻译系统精确地将机器译文与参考译文进行匹配。机器翻译系统的性能评测的依据是忠实度和流利度。忠实度用于评测译文中包含了多少原文要表达的意思，流利度用于评测译文是否流利。不同的人工评测者在评分时会根据自己的标准给译文打分，因此有必要规范这样的评分，使之具有可比性。在评测机器翻译系统时，除了考虑译文质量的指标之外，还要考虑翻译系统的速度、规模、集成性能和领域适应性等。在使用人工评测方法对机器翻译系统的机器译文与参考译文进行对比时，还应考虑单词的准确率和召回率。在评测时，可以采用一种称为Levenshtein距离的指标，Levenshtein距离用于计算将机器译文转换成参考译文时需要的最小编辑次数，以此来衡量词错误率(Word Error Rate，WER)。当前使用最广泛的评测指标是BLEU和METROR。自动评测指标与人工评测结果之间的相关性，可以使用Pearson相关系数来计算。

第三部分讲述统计机器翻译的前沿研究，包括判别式训练的方法及统计机器翻译中综合语言学信息的方法。

第9章介绍判别式训练和对翻译任务进行建模的方法。所谓“判别式训练”就是一种“分辨训练”，对于初步的翻译结果进行重新排序，从中分辨出最优的翻译结果。在进行判别式训练时，要把每个可能的候选译文用一组特征值来表示。每个特征被赋予一个特征权重，表示该特征与优秀译文的相关度。判别式训练使用重排序法，首先利用基线模型产生候选译文，然后再使用额外的特征选择出最佳的译文。把统计机器翻译解码器的搜索图转换为词格，就可以抽取出译文；词格还可以用来产生 n-best 的译文列表。统计机器翻译的判别式训练属于有监督学习。训练时需要准备一组源语言的输入句子和与之对应的候选译文集合，候选译文中至少有一个被标记为正确的。任何一个现代统计机器翻译系统在训练时都包含参数调节过程，用来为重要的系统参数设置最优值，尤其是对数线性模型中的参数权重，用于对相关子模型的分布建模。当前统计机器翻译中的一个富有挑战的研究课题是大规模判别式训练方法。在大规模判别式训练中，概率估计完全被特征和特征值替代，因而在这样的模型中，使用的特征数目达数百万之多。与判别式训练相关的是后验方法，这种后验方法主要研究在最佳候选译文样本集上的概率分布，使用最小贝叶斯风险解码，选择出一个与大多数高概率译文相似的译文。

第10章讲解了如何在统计机器翻译中整合语言学知识的问题。本章试图通过多种方式利用句法标注的语言学信息来提高统计机器翻译的质量，重点阐述如何拓展基于短语翻译的方法，并考虑了如何融入字母翻译、词汇翻译和句子结构等语言学知识的方法。以上提出的

方法中潜在的假设句子是有限集合中的符号序列。如果翻译语言的形态丰富，往往会导致统计机器翻译中词汇容量大和数据稀疏的问题，因此可以使用分解甚至去除语素的方法来简化形态，以尽量避免这样的问题。如果语言在词序方面差别明显，基于短语模型的调序法就不够用了，这时可以使用基于句法的方法来调序。当处理句法树的重构时，可以使用子结点调序限制来降低复杂性，也可以使用重排序方法，在挑选最佳翻译时利用语言的句法特征，检查输入和输出的一致性。因子化翻译模型把附加标记作为因子整合进短语模型中。短语翻译可以分解为映射、生成、翻译三个步骤，这些步骤作为特征函数在对数线性模型中进行建模。

第11章介绍了一种在句法理论中使用最广泛的基于树结构和语法类型的新框架。语言学的句法理论建立在句法树的基础上。然而，前面几章介绍的统计翻译模型是在句子的扁平化序列上处理的，只把句子看成一个词串。由于通过句法树能够挖掘出词和短语之间的句法关系，因此如何以句法树为基础建立机器翻译模型，是一个非常值得探索的课题，这样可以融入更多的语言学知识。本章介绍了统计机器翻译中基于树的翻译模型，把短语结构文法的形式化方法扩展成了同步文法，又进一步把同步文法扩展为同步树替换文法，从而生成非同构的句法树，这样就解决了扁平化上下文无关文法的子结点重排序约束问题。在基于短语模型训练方法的基础上，本章提出了一种同步文法的机器学习方法，这种方法从词对齐信息和句法树入手，首先从中抽取文法规则，对于给定的句法树，需要找到每个短语的管辖结点，以确定规则左部的非终结符。在线图分析算法中，组织数据的结构是包含若干线图条目的线图，这些线图都覆盖了输入句子的一个连续的跨度；从解码的角度来看，如果几个线图条目是等价的，就可以对它们进行重组。而在线图解码算法中，线图条目被压入栈中并进行栈剪枝，剪枝方法可以是直方图剪枝或阈值剪枝。为了降低句法分析解码的复杂度，本章还介绍了几种优化策略。文法规则的右部可能含有一定数目的非终结符，非终结符的数目称为秩，由于秩大的规则会引起计算复杂度高的问题，因此可以对文法规则进行二叉化，从而消除文法规则的右部的非终结符。基于树模型的这些方法和策略都涉及了语言学知识，显示了语言学知识在统计机器翻译中的重要作用。

我通读了全书的英文稿和中文译稿，觉得本书具有如下特点：

- 覆盖全面，内容新颖。本书作者 Philipp Koehn 是英国爱丁堡大学信息学院的讲师，他是欧洲 EuroMatrix 项目的科研协调人（EuroMatrix 现已发展成 EuroMatrixPlus，Philipp Koehn 也参加了这个项目），他与机器翻译领域的知名公司如 Systran 和 Asia Online 等从事过合作研究，有 10 多年的统计机器翻译经验，亲自见证了统计机器翻译的发展过程，他对于统计机器翻译的历史和现状有清楚的了解。本书不仅全面介绍了统计机器翻译的基础知识和核心方法，而且还探讨了统计机器翻译中的一些前沿研究问题，对于当前统计机器翻译发展的新成果进行了系统的总结。书中的参考文献大多数都是 2001 年至 2008 年（本书英文版交稿时间为 2008 年）之间发表的研究成果，有很大的参考价值。
- 深入浅出，通俗易懂。本书作者有丰富的实践经验，他曾经完善了统计机器翻译中广为使用的解码器 Pharaoh，并领导开发了开源统计机器翻译工具 Moses，他对于统计机器翻译的理解是深刻而具体的，因此对统计机器翻译技术所涉及的很多艰深的数学问题，都能用通俗的语言深入浅出地表达出来。本书的第一部分介绍了基本语言学知识和概率论知识等，读者几乎不需要事先具备这些方面的知识就能够阅读本书。本书的每一章都配有难易程度不同（用星号的多少来表示）的练习，有助于读者深入理解有关的内容。

- 实例丰富，图文并茂。本书使用了大量实例来帮助读者理解统计机器翻译的原理，使用了很多具体的图示来展示自然语言处理的运算过程，使读者阅读起来没有枯燥乏味之感。
- 注重评测，实用性强。评测是推动统计机器翻译研究健康发展的重要手段，本书介绍了统计机器翻译评测的各种方法，有助于针对评测中发现的问题不断地改善统计机器翻译系统的实用性能。
- 延伸阅读，别具心裁。本书的各章后面都有延伸阅读，给读者介绍一些有价值的文献和新的研究成果，使读者可以了解到该领域的发展动向。这些别具心裁的处理有助于扩大读者的知识面。

总体而言，本书总结了统计机器翻译多年来的研究成果，对于这些成果进行了系统的整理，通过阅读本书，有助于我们全面了解统计机器翻译的理论和技术。

特别应该说明的是，本书最后两章都用了较大的篇幅，浓墨重彩地讨论了如何在统计机器翻译中整合语言学知识的问题。这个问题是十分重要的，值得我们高度关注。

美国计算语言学家 Kenneth Church 发表过一篇文章 A Pendulum Swung Too Far(钟摆摆得太远了)①，值得我们密切注意。在这篇文章中，Kenneth Church 回顾了上世纪 90 年代在国际计算语言学学会(Association of Computational Linguistics，ACL)中创建数据研究兴趣组(Special Interest Group for Data，SIGDAT)的情形。他说，“当时我们出于实用主义的考虑，背叛了自己老师的理性主义方法的立场，专门建立一个兴趣小组来研究数据。我们认为，既然现在数据可以轻而易举地得到，为什么不能拿来利用一下呢？与其高不成低不就，不如顺水推舟，做一些简单易行的事情。让我们来摘取那些大树的低枝头上的唾手可得的果实吧。”他们采取的技术路线是基于数据的经验主义方法。

当时他们只是想在 ACL 众多的兴趣组中取得一席之地，并没有更大的野心。但几年之后，情况有了很大的变化，计算语言学中的这种基于数据的经验主义方法不仅复苏了，而且取得了很大的成功，以至于成为计算语言学的主流方法。这样，数据就显得特别重要了，Kenneth Church 和 SIGDAT 的同事们率先摘取了那些大树的低枝头上的唾手可得的果实，取得了辉煌的成就，可以看出，他们当初建立 SIGDAT 确实有先见之明。

如果当时 Kenneth Church 等人紧随在他们的老师之后亦步亦趋，不敢越雷池一步，估计就不会有今天这样辉煌的成就了。然而，在这样的成就面前，他们并没有得意忘形，Kenneth Church 清醒地认识到，当前这个经验主义的“钟摆”已经“摆得太远了”。他问道：如果那些低枝头的果实都被摘完之后，谁去摘那些处于大树的高枝头上的果实呢？究竟怎样去摘呢？他在文章中建议他的学生们认真地学习语言学的知识，深入研究语言学中的规律，才有可能摘取高枝头上的果实。

Church 的建议值得我们深思。

Church 的这篇文章发表两年之后，也就是在 2009 年，计算语言学家 Lori Levin 在欧洲计算语言学会(EACL2009)的语言学与计算语言学互动专题讨论上提出了一个发人深省的建议。他建议计算语言学要关注语言学的基础研究，在计算语言学学会(Association for Computational Linguistics，ACL)里设置一个语言学专委会。当 Lori Levin 提出这个问题时，计算语言学的学者们都感到尴尬，觉得这个建议很怪异。他们想：这岂不是有点像在美国儿科学会下

① Kenneth Church，A Pendulum Swung Too Far，*Linguistics Issues in Language Technology—LiLT*，Volume 2，Issue 4，May 2007。

面设立“医学专委会”或者“儿童专委会”一样滑稽可笑吗？然而接下来进一步再想一想，学者们才意识到这个建议的合理性。因为，从本质上讲，在当前的自然语言处理工程里，已经把语言学置于非常次要的地位了，学者们整天考虑的几乎都是程序技术或者算法问题，很少关注自然语言处理工程背景后隐藏着的语言学问题，因此计算语言学事实上已经成为没有语言学支持的语言学科，在计算语言学研究中，语言学在整体上是缺位的！

于是，以色列海法大学计算机科学系高级讲师 Shuly Wintner 发表了文章 What Science Underlies Natural Language Engineering？（什么是自然语言工程的科学支撑？）[①]，强烈呼吁“I want to call for the return of linguistics to computational linguistics”（语言学重新返回到计算语言学中）。她指出，20 多年来，计算语言学界完成了计算语言学研究范式的整体转型。由于语言学知识在数据规模扩张到真实世界的需求后仍然无法应用而带来的沮丧，以及由于形式语言占统治地位的理论带来的沮丧，学者们转向了语言数据，转向了语料库，转向了把语言的使用作为我们知识的潜在源泉。与方法论的转型相伴生的是计算语言学整个行业的目标的微妙变化。在 20 年前，一个计算语言学家或许既对开发自然语言处理的应用系统感兴趣，也对语言学过程的形式化及自动推理感兴趣。而在如今，他们只对开发自然语言处理的应用系统感兴趣，而对语言学过程的形式化及自动推理的研究嗤之以鼻。计算语言学领域的主要会议上的文章，绝大多数都是工程型的，讨论的都是实际问题的工程解决方案，几乎不再有人讨论那些基础性的语言学问题。

Shuly Wintner 认为，并不是说工程性的研究有什么错。因为每个大学都设有工程类的系，其领域之广泛已经涵盖诸如化工、机械工程、航空工程、生物医学工程等等。没有理由说不该在大学里设一个自然语言处理的工程学科。但是，就大多数已经设立的学科来看，工程类的系所进行的研究，都是在科学领域里一个非常成熟的理论分支的指导下进行的：化学工程师研究化学，电气工程师研究物理学，航空工程师研究动力学，生物医学工程师研究生物学、生理学和生命科学，如此等等。但是，自然语言处理的工程师居然不研究语言学，这就是咄咄怪事了！

究竟什么才是给自然语言处理工程作后盾的学科呢？什么才是我们建立应用时所依赖的理论基础支撑呢？当然应当是语言学。自然语言处理的工程师怎么能够不研究语言学呢？

考察一下面向数据的革命以来在自然语言工程领域的重大成果就可以清楚地看到这一点。比如宾州树库，1992 年第一版本问世以来，它的标注体系被用来对众多的词类标注器进行训练。这套标注体系的背后是什么理论呢？在什么意义上这个标注体系是“正确的”？会不会有其他某个标注体系也是同样好的呢？我们凭借什么准则对这样一套资源的质量进行评估？这准则又该嵌入科学的哪一个分支？——显然应当是语言学。

再看半个世纪以来自然语言处理领域的“皇冠上的明珠”——机器翻译。现在的统计机器翻译系统已经足以在很广泛的一类应用场合下使用了，Google 推出了超过 40 种语言的两两之间的免费互译服务。这恐怕算是自然语言处理领域最伟大的成就了吧，但是它到底基于什么学科？受到哪个理论的支撑？——显然也是语言学。

类似的例子不胜枚举。词汇歧义消解、随机句法分析、文本分类、自动问答、语义角色标记、语音识别、知识本体开发，随便哪种令你感兴趣的自然语言处理的应用，都可以同样追问：基于什么学科？受到哪个理论的支撑？它的理论支点在哪里？——显然都应当是语言学。

---

① Shuly Wintner, What Science Underlies Natural Language Engineering? *Computational Linguistics*, Volume 35, Number 4, Association for Computational Linguistics, 2009.

因此，Shuly Wintner 得出结论：没有明确的语言学知识作为基础的自然语言处理系统的应用领域是走不远的。

目前的经费投入机构主要由短期实用目标所驱动，在基础性的研究方面缺乏足够的耐心，对于基础性研究的经费投入不足，也是造成自然语言处理工程忽视语言学的一个原因。

Shuly Wintner 还尖锐地指出了更深层的原因。她认为，语言学作为一个学科，目前正在走向迷失——它只关注句法，而且又以英语的句法为主；语言学的理论变得如此晦涩难懂，如此华而不实，如此自以为是，以至于其他领域的研究者事实上无法跨学科参与进来。相关的语言学文献对于圈外人士来说已经成为了难以逾越的屏障，语言学理论用特殊术语来描述语言，导致计算语言学家很难把它与心理学的其他研究领域中关于认知的成果相联系。因此，计算语言学家们对语言学感到沮丧，在沮丧中他们彻底放弃了语言学，剩下的就只有统计学和概率论了，这样他们就全面地转向了统计学和概率论。

然而，计算语言学肯定不是应用统计学和概率论的一个分支，因为假如真是一个分支，那么自然语言处理系统就和其他非语言的字符串处理系统，比如 DNA 序列、乐谱、棋谱等非语言学的处理系统没有什么区别了。她认为，自然语言处理系统所处理的字符串肯定有某种唯一的特性，有某种可以从理论角度加以概括，在科学意义上加以研究的东西。

Shuly Wintner 最后指出，决定我们的系统的特殊性的，正是在于它处理的是自然语言，而能给我们以指导的唯一的科学领域就是语言学。实际上，在语言学的世界里新东西越多，计算语言学能从中受益的就越多。

Shuly Wintner 是一位计算机背景的计算语言学家，我认为，她的建议是难能可贵的，更是高瞻远瞩的。

看来，在目前统计机器翻译大行其道的时候，我们应当保持清醒的头脑，在机器翻译研究中，把基于规则的理性主义方法和基于统计的经验主义方法结合起来，在统计机器翻译系统中综合并融入语言学知识，使这两种方法互为补充，相得益彰，才能进一步推进机器翻译的发展[①]。

上文已经提到，Shuly Wintner 在她的文章中对语言学提出了批评："语言学的理论变得如此晦涩难懂，如此华而不实，如此自以为是"，这种情况使计算语言学的研究者难以应用语言学知识，以至于"在沮丧中彻底放弃了语言学"。

在我国，语言学研究中的这种令人沮丧的情况也同样存在。曾经有朋友对我说，他认为古汉语的研究很有用，因为这种研究总是力图把别人不明白的问题说明白，从而帮助人们读懂古代汉语的文献；现代汉语研究固然取得了骄人的成绩，但是现代汉语的有些研究往往喜欢把很多明明白白的问题说得让人不明不白，使人如坠五里云雾中，以至于觉得这样的研究是故弄玄虚的智力游戏；令人遗憾的是，做这样研究的专家竟然还大言不惭地宣称，他们的这种研究得出的规则是很严密的，而这正是追求严密的机器翻译所需要的。这位朋友对于现代汉语中这些研究的批评值得我们深思。我以为，用这样的方式研究出来的现代汉语中的这些所谓严密的"规则"，尽管貌似深奥，当专家们坐而论道、纸上谈兵时也许可以大显身手，但在实际的机器翻译系统研制中，充其量也只是"银样蜡枪头"而已，是很难派上用场的。

希望语言学研究彻底改变这种"晦涩难懂""华而不实""自以为是"的局面，针对机器翻

① 冯志伟，用计量方法研究语言，《外语教学与研究》，第 44 卷，第 2 期，p256-269，2012 年。

译的要求，从大规模的真实语料中提取出一些行之有效的规则来，并把这些行之有效的规则融入统计机器翻译系统中，真正把基于规则的理性主义方法和基于统计的经验主义方法结合起来，从而推动我国统计机器翻译的发展。

从总体上讲，目前机器翻译的水平还不高，往往满足不了用户的期望。我们时常会听到用户对于机器翻译质量的抱怨，有的人甚至对机器翻译采取了嘲笑和否定的态度。

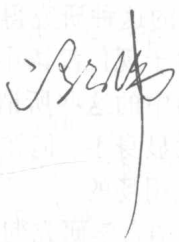
天气预报也有不准确的时候，但公众却能够容忍天气预报的错误。我觉得，在机器翻译译文质量还不理想的时候，我们应当采取对待天气预报的容忍态度来对待机器翻译。天气从来不会是一成不变的，也从来不会有一模一样的天气。由于这样的原因，天气预报的准确性还不高。在很多情况下，一周以上的中长期天气变化很难预测，因而预报的准确性很差，但我们有能力了解和解释各种天气现象，能够辨认出诸如锋面、气流、高压圈等重要的气象特征，一天至两天内的短期天气预报，仍然是有一定的准确性的。可见，尽管我们无法对天气的变化进行完全准确的预测，但天气预报仍然不失为一门真正的科学，天气预报每天都在为公众服务，虽然有时预报得不准确，仍能得到公众的谅解和支持。为什么不能采取对待天气预报的这种态度来对待机器翻译呢？

我们应当对于机器翻译目前的译文质量给予一定的容忍，用善意的态度来鼓励机器翻译的研究，促进机器翻译的发展。

在机器翻译研究中，我们常常需要面对各种界定不清的问题，各种界定不清的环境，甚至于完全不知走向的变化。但是，我们也经常在这些含糊不清的情况下做出决定。实际上，我们是在摸着石头过河，在过河的过程中还要不断改变自己的思想，不断吸取别人的经验，不断尝试以往成功的经验，通过我们不懈的努力，最后也许可能到达河的对岸，体验“山穷水尽疑无路，柳暗花明又一村”的愉悦。

这本书的名字虽然是《统计机器翻译》，但 Philipp Koehn 实际上已经注意到统计方法的局限性，明确地提出要在统计机器翻译中整合语言学知识，这是高瞻远瞩的明智见解。我完全同意 Philipp Koehn 的这种见解，希望读者在学习本书的过程中，自觉地把基于经验主义的统计方法和基于语言学规则的理性主义方法巧妙地结合起来，为我国的机器翻译研究做出新的贡献。

宗成庆和张霄军两位博士在统计机器翻译方面都有丰富的知识和实际经验，他们的英文水平也很高，在忙碌的本职工作之余，利用业余时间把 Philipp Koehn 的这本著作译成中文，我对照英文原著仔细地读了他们的中译本，感觉他们的译文是准确流畅的，如实地表达了作者的原意，在翻译中还校正了原作者的一些不妥和失误之处。这个中译本现在由电子工业出版社出版，我对他们表示祝贺，并写了上面这些不成熟的体会，算是我交给他们的一份读书心得吧！



冯志伟

2012 年 7 月于杭州下沙