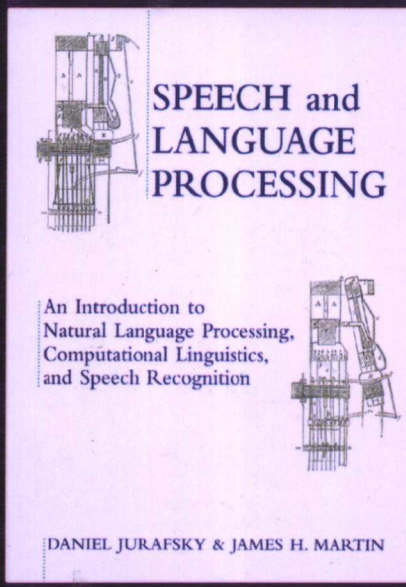
国外计算机科学教材系列

自然语言处理综论

Speech and Language Processing

An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition



[美] Daniel Jurafsky
James H. Martin 著

冯志伟 孙 乐 译

PEARSON Prentice Hall

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

内容简介

本书是一本全面系统地讲述计算机自然语言处理的优秀教材。本书英文版出版之后好评如潮，国外许多著名大学纷纷把本书选为自然语言处理和计算语言学课程的主要教材，该书被誉为该领域教材的“黄金标准”。本书包含的内容十分丰富，分为四个部分，共21章，深入细致地探讨了计算机处理自然语言的词汇、句法、语义、语用等各个方面的问题，介绍了自然语言处理的各种现代技术。从层次的角度看，本书的论述是按照自然语言的不同层面逐步展开的，首先论述单词的自动形态分析，接着论述自动句法分析，然后论述各种语言单位的自动语义分析，最后论述连贯文本的自动分析、对话与会话的智能代理以及自然语言生成。从技术的角度看，本书介绍了正则表达式、有限状态自动机、文本－语音转换、发音与拼写的概率模型、词类自动标注、*N*元语法、隐马尔可夫模型、上下文无关语法、特征与合一、词汇化剖析与概率剖析、一阶谓词演算、词义排歧、修辞结构理论、机器翻译等非常广泛的内容。本书具有“覆盖全面、注重实用、强调评测、语料为本”四大特色。在本书的配套网站上，还提供了相关的资源和工具，便于读者在实践中进一步提高。

本书不仅可以作为高等学校自然语言处理和计算语言学等课程的本科生和研究生教材，而且也是从事自然语言处理相关领域的研究人员和技术人员的必备参考。

版权贸易合同登记号　图字：01-2003-0357

图书在版编目（CIP）数据

自然语言处理综论 /（美）朱夫斯凯（Jurafsky, D.）等著；冯志伟，孙乐译.
北京：电子工业出版社，2005.6
（国外计算机科学教材系列）
书名原文：Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
ISBN 7-121-00776-2

Ⅰ. 自...　Ⅱ. ①朱... ②冯... ③孙...　Ⅲ. 自然语言处理－高等学校－教材　Ⅳ. TP391

中国版本图书馆CIP数据核字（2005）第047102号

责任编辑：马　岚　　特约编辑：马爱文
印　　刷：北京市天竺颖华印刷厂
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
经　　销：各地新华书店
开　　本：787 × 1092　1/16　　印张：38.25　字数：1079千字
印　　次：2005年6月第1次印刷
定　　价：78.00元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系电话：（010）68279077。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。

出 版 说 明

21世纪初的5至10年是我国国民经济和社会发展的重要时期，也是信息产业快速发展的关键时期。在我国加入WTO后的今天，培养一支适应国际化竞争的一流IT人才队伍是我国高等教育的重要任务之一。信息科学和技术方面人才的优劣与多寡，是我国面对国际竞争时成败的关键因素。

当前，正值我国高等教育特别是信息科学领域的教育调整、变革的重大时期，为使我国教育体制与国际化接轨，有条件的高等院校正在为某些信息学科和技术课程使用国外优秀教材和优秀原版教材，以使我国在计算机教学上尽快赶上国际先进水平。

电子工业出版社秉承多年来引进国外优秀图书的经验，翻译出版了“国外计算机科学教材系列”丛书，这套教材覆盖学科范围广、领域宽、层次多，既有本科专业课程教材，也有研究生课程教材，以适应不同院系、不同专业、不同层次的师生对教材的需求，广大师生可自由选择和自由组合使用。这些教材涉及的学科方向包括网络与通信、操作系统、计算机组织与结构、算法与数据结构、数据库与信息处理、编程语言、图形图像与多媒体、软件工程等。同时，我们也适当引进了一些优秀英文原版教材，本着翻译版本和英文原版并重的原则，对重点图书既提供英文原版又提供相应的翻译版本。

在图书选题上，我们大都选择国外著名出版公司出版的高校教材，如Pearson Education培生教育出版集团、麦格劳－希尔教育出版集团、麻省理工学院出版社、剑桥大学出版社等。撰写教材的许多作者都是蜚声世界的教授、学者，如道格拉斯·科默（Douglas E. Comer）、威廉·斯托林斯（William Stallings）、哈维·戴特尔（Harvey M. Deitel）、尤利斯·布莱克（Uyless Black）等。

为确保教材的选题质量和翻译质量，我们约请了清华大学、北京大学、北京航空航天大学、复旦大学、上海交通大学、南京大学、浙江大学、哈尔滨工业大学、华中科技大学、西安交通大学、国防科学技术大学、解放军理工大学等著名高校的教授和骨干教师参与了本系列教材的选题、翻译和审校工作。他们中既有讲授同类教材的骨干教师、博士，也有积累了几十年教学经验的老教授和博士生导师。

在该系列教材的选题、翻译和编辑加工过程中，为提高教材质量，我们做了大量细致的工作，包括对所选教材进行全面论证；选择编辑时力求达到专业对口；对排版、印制质量进行严格把关。对于英文教材中出现的错误，我们通过与作者联络和网上下载勘误表等方式，逐一进行了修订。

此外，我们还将与国外著名出版公司合作，提供一些教材的教学支持资料，希望能为授课老师提供帮助。今后，我们将继续加强与各高校教师的密切联系，为广大师生引进更多的国外优秀教材和参考书，为我国计算机科学教学体系与国际教学体系的接轨做出努力。

电子工业出版社

教材出版委员会

中文版序言

The goal of a textbook author is the same as the goal of any teacher: passing on our love for our field to a new generation of students, encouraging them to do innovative and creative new work, and helping them to advance the state of human knowledge. For a textbook in the interdisciplinary area of speech and language processing, there are the additional goals of enabling students from differing backgrounds (computer science, linguistics, electrical engineering) to acquire the knowledge and tools of the new interdisciplinary field, and to develop an appreciation for the beauty and complexity and variety of human language. We therefore feel extremely lucky that Feng Zhiwei Laoshi, aided by Dr. Sun Le, undertook the arduous job of translating this book. Feng Laoshi is the perfect scholar for the job of translating such a book, because of his long experience in our field, his wide breadth of research interests throughout computational linguistics in general and Chinese computational linguistics specifically, his remarkable familiarity with the state of our field across the world, from China to France, from Korea to Germany, and of course his expertise on translation as a research area! We are also very excited that this translation into Chinese is the first translation of our book out of English. China's long history of the study of language is of course well known, and in this new century the young scientists of China are already playing a key role in the important scientific advances of our field. We look forward to even more amazing contributions from China and hope that our small book, now with the help of Feng Laoshi and Dr. Sun, can provide a small aide in the great role that Chinese scientists are playing on the world scientific stage!

Daniel Jurafsky and James H. Martin
Palo Alto, California, and Boulder, Colorado

——译文——

教材的作者与所有教师有着相同的目标，即把我们对于本专业的热爱传达给新一代的学生，鼓励他们进行创新性的研究和探索，帮助他们把人类知识进一步向前推进。由于语音和语言的计算机处理属于交叉学科的领域，所以这本关于该交叉学科领域的教材还有其特定的目标。这些特定的目标就是使来自不同知识背景（计算机科学、语言学和电子工程）的学生掌握这门新的交叉学科的基本知识和工具，并在学习过程中循序渐进地感受人类语言的美妙性、复杂性和多样性。因此，当了解到冯志伟老师在孙乐博士的协助下承担了把这本教材翻译成中文的艰辛工作时，我们感到无比荣幸。我们认为，冯志伟老师是翻译这本教材的最理想的学者，因为他在这个专业领域具有多年的经验；他的研究兴趣涉及面广，既包括普通的计算语言学研究，也包括具体的汉语计算语言学的研

究；他对于这个学科在全世界的情况了如指掌，从中国到法国，从韩国到德国，他都亲身参与了这些国家的计算语言学研究工作；并且，翻译一直是冯老师长期从事的一个研究领域，他当然也是精研通达的翻译内行！这个中译本是英文原著的第一个外文译本，它的出版使我们非常激动和振奋。众所周知，中国在语言研究方面有着悠久的历史，在21世纪，中国年轻一代的科学工作者在这个领域的一些重要科学进展方面已经起着关键性的作用。我们期待着中国在这个领域里进一步做出更加出色的贡献。我们也希望，在中国科学工作者为全世界的科学进步事业所发挥的巨大作用中，由于冯老师和孙乐博士的帮助，拙著也能够为此尽我们的绵薄之力！

译 者 序

采用计算机技术来研究和处理自然语言是20世纪40年代末期和20世纪50年代才开始的，50多年来，这项研究取得了长足的进展，成为了计算机科学中一门重要的新兴学科——自然语言处理（Natural Language Processing，简称NLP）。

我们认为，计算机对自然语言的研究和处理，一般应经过以下4方面的过程：

1. 把需要研究的问题在语言学上加以形式化，使之能以一定的数学形式，严密而规整地表示出来；
2. 把这种严密而规整的数学形式表示为算法，使之在计算上形式化；
3. 根据算法编写计算机程序，使之在计算机上加以实现；
4. 对于所建立的自然语言处理系统进行评测，使之不断地改进质量和性能，以满足用户的要求。

美国计算机科学家Bill Manaris在*Advanced in Computers*（《计算机进展》）第47卷的Natural language processing: A human-computer interaction perspective（《从人机交互的角度看自然语言处理》）一文中曾经给自然语言处理提出了如下的定义：

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。

Bill Manaris关于自然语言处理的这个定义，比较全面地表达了计算机对自然语言的研究和处理的上述四个方面的过程。我们认同这样的定义。

根据这样的定义，我们认为，建立自然语言处理模型需要如下不同平面的知识：

1. 声学和韵律学的知识：描述语言的节奏、语调和声调的规律，说明语音怎样形成音位。
2. 音位学的知识：描述音位的结合规律，说明音位怎样形成语素。
3. 形态学的知识：描述语素的结合规律，说明语素怎样形成单词。
4. 词汇学的知识：描述词汇系统的规律，说明单词本身固有的语义特性和语法特性。
5. 句法学的知识：描述单词（或词组）之间的结构规则，说明单词（或词组）怎样形成句子。
6. 语义学的知识：描述句子中各个成分之间的语义关系，这样的语义关系是与情景无关的，说明怎样从构成句子的各个成分推导出整个句子的语义。
7. 话语分析的知识：描述句子与句子之间的结构规律，说明怎样由句子形成话语或对话。
8. 语用学的知识：描述与情景有关的情景语义，说明怎样推导出句子具有的与周围话语有关的各种涵义。
9. 外界世界的常识性知识：描述关于语言使用者和语言使用环境的一般性常识，例如，语言使用者的信念和目的，说明怎样推导出这样的信念和目的内在的结构。

当然，关于自然语言处理所涉及的知识平面还有不同的看法，不过，一般而言，大多数的自然语言处理研究人员都认为，这些语言学知识至少可以分为词汇学知识、句法学知识、语义学知识和

语用学知识等平面。每一个平面传达信息的方式各不相同。例如，词汇学平面可能涉及具体的单词的构成成分（如语素）以及它们的屈折变化形式的知识；句法学平面可能涉及在具体的语言中单词或词组怎样结合成句子的知识；语义学平面可能涉及怎样给具体的单词或句子指派意义的知识；语用学平面可能涉及在对话中话语焦点的转移以及在给定的上下文中怎样解释句子的涵义的知识。

下面我们具体说明在自然语言处理中这些知识平面的一般情况。如果对计算机发一个口头的指令：Delete file x（删除文件 X），为了通过自然语言处理系统让计算机理解这个指令的涵义，并且执行这个指令，一般来说需要经过如图 0.1 所示的处理过程。

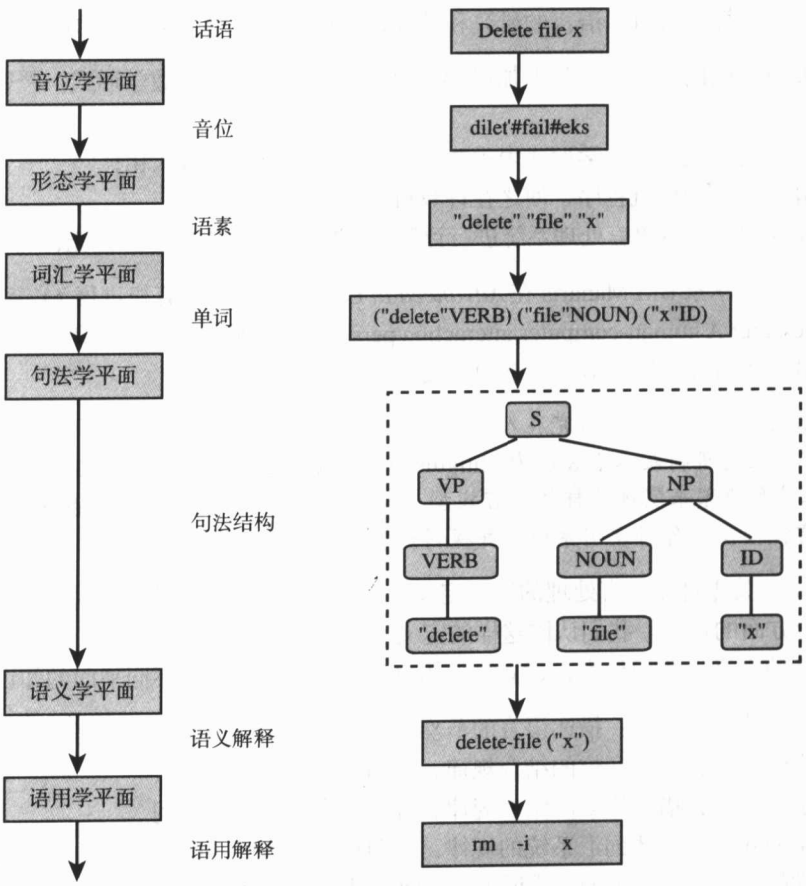


图 0.1　自然语言处理系统中的知识平面

从图中可以看出，自然语言处理系统首先把指令 Delete file x 在音位学平面转化成音位系列 dilet'#fail#eks，然后在形态学平面把这个音位系列转化为语素系列 delete，file 和 x，接着在词汇学平面把这个语素系列转化为单词系列并标注相应的词性：("delete"VERB) ("file"NOUN) ("x"ID)，在句法学平面进行句法分析，得到这个单词系列的句法结构，用树形图表示，在语义学平面得到这个句法结构的语义解释 delete-file ("x")，在语用学平面得到这个指令的语用解释 rm -i x，最后让计算机执行这个指令。

这个例子来自美国自然语言处理学者 Wilensky 为 UNIX 设计的一个语音理解界面，叫做 UNIX Consultant。这个语音理解界面使用了上述的第 1 个至第 6 个平面的知识，得到口头指令 Delete file x

的语义解释：delete-file ("x")。然后，使用第 8 个平面的语用学知识把这个语义解释转化为计算机的指令语言 rm -i x，让计算机执行这个指令，这样便可以使用口头指令来指挥计算机的运行了。

不同的自然语言处理系统需要的知识平面可能与UNIX Consultant不一样，根据实际应用的不同要求，很多自然语言处理系统只需要使用上述9个平面中的部分平面的知识就行了。例如，书面语言的机器翻译系统只需要第3个至第7个平面的知识，个别的机器翻译系统还需要第8个方面的知识；语音识别系统只需要第 1 个至第 5 个平面的知识。

上述9个平面的知识主要涉及的是语言学知识，由于自然语言处理是一个多边缘的交叉学科，除了语言学之外，它还涉及如下的知识领域：

- **计算机科学**：给自然语言处理提供模型表示、算法设计和计算机实现的技术。
- **数学**：给自然语言处理提供形式化的数学模型和形式化的数学方法。
- **心理学**：给自然语言处理提供人类言语行为的心理模型和理论。
- **哲学**：给自然语言处理提供关于人类的思维和语言的更深层次的理论。
- **统计学**：给自然语言处理提供基于样本数据来预测统计事件的技术。
- **电子工程**：给自然语言处理提供信息论的理论基础和语言信号处理技术。
- **生物学**：给自然语言处理提供大脑中人类语言行为机制的理论。

自然语言处理需要的知识如此之丰富，涉及的领域如此之广泛，而我们翻译的美国科罗拉多大学 Daniel Jurafsky 和 James Martin 的这本著作正好满足了这样的要求。

几年前我从韩国到新加坡参加国际会议时，在书店发现此书，马上就被它丰富的内容和流畅的表达吸引住了。会议结束回到韩国之后，我就开始认真阅读此书，我发现此书覆盖面非常广泛，理论分析十分深入，而且强调实用性和注重评测技术，几乎所有的例子都来自真实的语料库，此书的内容不仅覆盖了我们在上面所述的9个平面的语言学知识和外在世界的常识性知识，而且还涉及到计算机科学、数学、心理学、哲学、统计学、电子工程和生物学等领域的知识，我怀着极大的兴趣前后通读了两遍。当时我在韩国科学技术院电子工程与计算机科学系担任访问教授，在我给该系博士研究生开的“自然语言处理 -II”（NLP-II）的课程中，使用了该书的部分内容，效果良好。我觉得这确实是一本很优秀的自然语言处理的教材。我常常想，如果我们能够把这本优秀的教材翻译成中文，让国内的年轻学子们也能学习本书，那该是多么好的事情！

后来，在一次机器翻译研讨会上，电子工业出版社的编辑找到我，告诉我说他们打算翻译出版此书。当时电子工业出版社已经进行过调查，目前国外绝大多数大学的计算机科学系都采用此书作为“自然语言处理”课程的研究生教材，他们希望我来翻译这本书，与电子工业出版社配合，推出高质量的中文译本。我们双方的想法不谋而合，于是，我欣然接受了本书的翻译任务，开始进行本书的翻译。

我虽然已经通读过本书两遍，对于本书应该说是有一定的理解了，但是，亲自动手翻译起来，却不像原来想像的那样容易，要把英文的意思表达为确切的中文，下起笔来，总有汲深绠短之感，大量的新术语如何用中文来表达，也是颇费周折令人踌躇的难题。我利用了全部的业余时间来进行翻译，连续工作了 11 个月，当翻译完 14 章（全书的三分之二）的时候，我患了眼病，视力出现障碍，难于继续翻译工作，还剩下 7 章（全书的三分之一）没有翻译，“行百里者半九十”，这 7 章的翻译工作究竟如何来完成呢？正当我束手无策一筹莫展的时候，中国科学院软件研究所副研究员孙乐博士表示愿意继续我的工作，与我协作共同完成本书的翻译。孙乐博士有很好的自然语言处理的基础，我们又是忘年之交的好朋友，由他来继续我的翻译工作是最理想不过的了，电子工业出版社也同意孙乐参与本书的翻译。孙乐博士的翻译工作十分认真，他每翻译一章，就交给我审校，遇到

疑难问题时我们共同切磋，反复推敲，他顺利地完成了第15章到第21章的翻译，现在，在我们两人的通力合作下，全书的翻译总算大功告成了。原书每章都有习题，由于这些习题涉及的语言背景不适合中国读者，因此在中文版中未将习题包括在内。

正如本书作者指出的，本书具有“覆盖全面，强调实用，注重评测，语料为本”的特点，我们希望，这个中文译本能够在我国的自然语言处理的教学和科学研究中，产生积极的作用，我们还希望，读者能够喜欢这个译本，并给我们提出批评和指正。

本书译者的部分工作得到国家自然科学基金（编号：60203007）和北京市科技新星计划（编号：H020820790130）的资助，特此致谢。

冯志伟

前　言

现在，语音和语言的计算机处理进入了一个令人振奋的时期。在这个时期，历史上彼此不同的研究部门（自然语言处理、语音识别、计算语言学、计算心理语言学）开始融合在一起。语音识别研究的商品化以及对于基于互联网的语言处理技术的需求，有力地推动了各种实用的自然语言处理系统的开发。由于使用大规模的联机语料库，使得在从语音到话语的各个不同的层面都可以使用统计方法。我们在设计这本既可作为教学之用又可作为参考书之用的专著时，试图描绘出各个不同学科开始融合在一起的这种情景。本书具有如下的特点：

1. **覆盖全面**　为了统一地描述语音处理和语言处理，本书涵盖了传统上分别在不同的系和不同课程中讲授的内容。例如，在电子工程系的语音识别课程中的内容；在计算机科学系的自然语言处理课程中的自动句法分析、语义解释、语用学等内容；在语言学系的计算语言学课程中的计算形态学和计算音系学等内容。本书介绍了这些领域中的基本算法，无论这些算法原来是在语音处理还是在语言处理中提出的，无论它们原来是从逻辑的角度还是从统计的角度提出的，我们力求将来自不同领域的算法合在一起统一描述。我们也试图把一些诸如拼写检查、信息检索和信息抽取这样的应用领域的内容包括在本书中，使其覆盖得更全面。这种广为覆盖的方法的一个潜在问题是，我们只好把每个领域中的一些概论性的材料也包括到本书中。因此，在阅读本书时，语言学家可以跳过有关发音语音学方面的章节，计算机科学家可以跳过有关正则表达式的章节，电子工程师可以跳过有关信号处理的章节。当然，尽管这本书写得这样长，我们也不可能做到包罗万象。正因为如此，本书不能替代语言学、自动机和形式语言理论，特别是关于统计学和信息论的各种专门著作，这些著作显然是非常重要的。
2. **注重实用**　理论联系实际是非常重要的。在本书中，我们始终注意把自然语言处理的算法和技术（从隐马尔可夫模型到合一算法，从λ运算到基于转换的学习）应用于解决现实世界中遇到的各种重要问题，例如拼写检查、文本文献检索、语音识别、网页信息处理、词类标注、机器翻译、口语对话代理等。为了达到这样的目的，我们在每一章中都要讲授一些关于自然语言处理的应用问题。这种方法的好处是，当介绍有关自然语言处理的知识时，可以给学生们提供一个背景来理解和模拟特定领域中的应用问题。
3. **强调评测**　近年来，在自然语言处理中统计算法越来越受到重视，语音处理和语言处理的有组织的评测系统越来越多，这些都使得评测得到了越来越多的强调和重视。因此，我们在本书中许多领域设立了“方法论探讨”（Methodology Box），具体讲述怎样评测一个系统。例如，我们介绍了训练集和测试集的概念，交叉确认（cross-validation）以及诸如困惑度（perplexity）这样的信息论评测指标。
4. **语料为本**　现代的语音处理和语言处理很多是建立在公共资源的基础上的。这些资源包括：语音生语料库和文本生语料库，标注语料库和树库，用于语音标记、词类标记、自动句法分析、词义以及对话层面的现象的标准标注集等。我们力图在全书中介绍很多这样的重要语言资源（例如，Brown, Switchboard, callhome, ATIS, TREC, MUC, BNC 等语料库），并且提供

很多有用的标记集的完整清单以及编码技巧（例如 Penn Treebank, CLAWS C5 和 C7，以及 ARPAbet），不过难免会有遗漏。此外，除了在本书中直接包括了很多资源的 URL（Uniform Resource Locator）之外，我们还把这些资源放在本书的网站上，这样即可使这些得到及时的更新。

本书首先可以用做研究生或高年级本科生的教材或系列教材。由于本书的覆盖面广，并且有大量的算法，所以，本书也可以用做语音处理和语言处理的各个领域中的大学生和专业人员的参考书。

本书概览

除了前言和书后面的附录之外，本书共分为四个部分。第一部分是“词汇的计算机处理”，讲述与词汇的计算机处理有关的语音学、音系学、形态学的基本概念，介绍语音和词汇计算机处理中的各种算法，如有限自动机、有限转录机、加权转录机、*N* 元模型、隐马尔可夫模型等。第二部分是“句法的计算机处理”，介绍英语的词类和短语的结构语法，讲述用于词类处理和结构处理的一些主要的算法，如基于 HMM 的词类标注、基于转换的学习、CYK 分析算法、Earley 分析算法、合一与类型特征结构、词汇化剖析和概率剖析以及诸如 Chomsky 层级分类和抽吸引理（pumping lemma）等分析工具。第三部分是“语义的计算机处理”，介绍一阶谓词演算以及语义的各种表示方法，组合语义分析的各种方法、信息抽取、言语理解和机器翻译。第四部分是“语用的计算机处理”，讲述所指判定（reference resolution）、话语的结构和连贯性、口语对话的现象、对话和言语行为模式、对话管理以及机器翻译和自然语言生成中的各种处理方法。

本书使用方法

本书材料丰富，可作为两学期的语音处理和语言处理系列教材。本书也可以作为各种不同用途的一个学期的教材使用。

自然语言处理 一个季度	自然语言处理 一个学期	语音与语言处理 一个学期	计算语言学 一个季度
1. 导论	1. 导论	1. 导论	1. 导论
2. 正则表达式，FSA	2. 正则表达式，FSA	2. 正则表达式，FSA	2. 正则表式，FSA
8. 词性标注	3. 形态学，FST	3. 形态学，FST	3. 形态学，FST
9. 上下文无关文法	6. *N* 元语法	4. 计算音系学	4. 计算音系学
10. 剖析	8. 词性标注	5. 发音的概率模型	10. 剖析
11. 合一	9. 上下文无关文法	6. *N* 元语法	11. 合一
14. 语义学	10. 剖析	7. HMM 与语音识别	13. 复杂性
15. 语义分析	11. 合一	8. 词性标注	16. 词汇语义学
18. 话语	12. 概率剖析	9. 上下文无关文法	18. 话语
20. 生成	14. 语义学	10. 剖析	19. 对话
	15. 语义分析	12. 概率剖析	
	16. 词汇语义学	14. 语义学	
	17. WSD 与信息检索	15. 语义分析	
	18. 话语	19. 对话	
	20. 生成	21. 机器翻译	
	21. 机器翻译		

本书的某些章节也可以选用于人工智能、认知科学或者信息检索等课程。

致谢

有三位作者对本书做出了贡献，他们协助我们写了有关章节。他们是Andy Kehler, Keith Vander Linden 和 Nigel Ward。Andy Kehler 写了第 18 章（话语分析），Keith Vander Linden 写了第 20 章（生成），Nigel Ward 写了第 21 章（机器翻译）。Andy Kehler 还写了 19.4 节。此外，Paul Taylor 帮助我们写了 4.7 节和 7.8 节。

Daniel Jurafsky 在此要感谢他的父母，是他们鼓励 Daniel 把每件事都做得尽善尽美，按时完成，并且抽时间到体育馆去锻炼身体。感谢 Nelson Morgan，因为 Morgan 引导他从事语音识别的研究，并且教导他对任何事情都要问一个“这样行吗？”感谢 Jerry Feldman，因为 Jerry 经常帮助他寻找问题的正确答案，教导他对于任何事情都要问一问：“这确实是重要的吗？”感谢 Chuck Fillmore，因为 Chuck 是他的第一个咨询人，和他分享对于语言的爱好，特别是对于研究论元结构的爱好，教导他要始终重视数据，并且要用实例来说明问题，只有当数据真正有价值时，才值得花时间和精力去研究。感谢 Robert Wilensky，Robert 是他的博士论文的指导教师，Robert 教导他懂得了合作共事以及团队精神的重要性。感谢 Doris 和 Cary, Elaine 和 Eric, Irene 和 Sam, Susan 和 Richard, Lisa 和 Mike, Mike 和 Fia, Erin 和 Chris, Eric 和 Beth, Pearl 和 Tristan, Bruce 和 Peggy, Ramon 和 Rebecca, Adele 和 Ali, Terry, Kevin, Becky, Temmy, Lil, Lin, Ron 和 David, Mike, Jessica 和 Bill，以及他们的家庭，因为他们给了他热情的支持，并且给他提供停留的地方以便他进行写作。

James H. Martin 在此也要感谢他的父母，是他们给了 James 鼓励，并且允许他走上自然语言处理这条在当时看来似乎很古怪的学术道路。感谢他的博士论文指导老师 Robert Wilensky，是 Robert 给他机会在 Berkeley 开始学习自然语言处理。感谢 Peter Norvig，是 Peter 给他提供了许多正面的例子，并且指引他找到正确的途径。感谢 Rick Alterman，是 Rick 在关键和困难的时刻，给了他鼓励和勇气。感谢 Chuck Fillmore, George Lakoff, Paul Kay 和 Susanna Cumming，因为他们教过 James，使他懂得了语言学。感谢 Michael Main，因为当他不能担负系里的许多工作时，是 Michael 替他做了弥补。最后，James 还要感谢他的妻子 Linda，正是由于她多年的支持和耐心，James 才能够完成本书的写作。

科罗拉多大学的所在地 Boulder 是从事语音处理和语言处理的好地方。这里，我们要感谢我们在科罗拉多大学的同事们，他们的合作共事对我们的研究和教学有极大的影响。他们是：语言学系的 Alan Bell, Barbara Fox, Laura Michaelis 和 Lise Menn；计算机科学系的 Clayton Lewis, Gerhard Fischer, Mike Eisenberg, Mike Mozer, Liz Jessup 和 Andrzej Ehrenfeucht；心理学系的 Walter Kintsch, Tom Landauer 和 Alice Healy；口语理解中心的 Ron Cole, John Hansen 和 Wayne Ward；我们还要感谢我们在计算机科学系和语言学系的现在的和过去的学生 Marion Bond, Noah Coccaro, Michelle Gregory, Keith Herold, Michael Jones, Patrick Juola, Keith Vander Linden, Laura Mather, Taimi Metzler, Douglas Roland 和 Patrick Schone。

许多朋友仔细地阅读了本书并且进行了试教，他们提出了很多有益的建议。我们的同事也在百忙中抽出时间来阅读本书，提出了宝贵的意见和建议。这些意见和建议使得本书的很多部分都得到了改进，增色不少。我们应该深深地感激他们。他们是：Alan Bell, Bob Carpenter, Jan Daciuk, Graeme Hirst, Andy Kehler, Kemal Oflazer, Andreas Stolcke 和 Nigel Ward。我们的编辑 Alan Apt，系列丛书编辑 Peter Norvig 和 Stuart Russell，以及制作编辑 Scott DiSanno，都对于本书的设计和内容提出了很多有益的建议。我们还要感激许多朋友和同事们，他们阅读了本书的个别章节。对于他们的意见和建议中我们不明白的地方，他们还回答了我们很多的问题。我们还要感激科罗拉多大学 Boulder 校

部选这门课的学生们，加利福尼亚大学 Berkeley 分校 Daniel Jurafsky 班上的同学们，以及伊利诺伊大学 Urbana-Champaign 分校 LSA 暑期学院的学生们。此外还有：

Yoshi Asano, Todd M. Bailey, John Bateman, Giulia Bencini, Lois Boggess, Michael Braverman, Nancy Chang, Jennifer Chu-Carroll, Noah Coccaro, Gary Cottrell, Gary Dell, Jeff Elman, Robert Dale, Dan Fass, Bill Fisher, Eric Fosler-Lussier, James Garnett, Susan Garnsey, Dale Gerdemann, Dan Gildea, Michelle Gregory, Nizar Habash, Jeffrey Haemer, Jorge Hankamer, Keith Herold, Beth Heywood, Derrick Higgins, Erhard Hinrichs, Julia Hieschberg, Jerry Hobbs, Fred Jelinek, Liz Jessup, Aravind Joshi, Terry Kleeman, Jean-Pierre Koenig, Kevin Knight, Shalom Lappin, Julie Larson, Stephen Levinson, Jim Magnuson, Jim Mayfield, Lise Menn, Laura Michaelis, Corey Miller, Nelson Morgan, Christine Nakatani, Mike Neufeld, Peter Norvig, Mike O'Connell, Mick O'Donnell, Rob Oberbreckling, Martha Palmer, Dragomir Radev, Terry Regier, Ehud Reiter, Phil Resnik, Klaus Ries, Ellen Riloff, Mike Rosner, Dan Roth, Patrick Schone, Liz Shriberg, Richard Sproat, Subhashini Srinivasin, Paul Taylor, Wayne Ward, Pauline Welby, Dekai Wu 和 Victor Zue。

我们要感谢认知科学研究所以及计算机科学系和语言学系，谢谢他们对于我们多年的支持。我们还要感谢国家科学基金会，Daniel Jurafsky在写作本书期间还部分地得到了NSF CAREER IIS-9733067 的资助，Andy Kehler 也部分地得到了 NSF IIS-9619126 的资助。

目　录