经典译丛

Pearson

人类语言技术

自然语言处理综论

（第二版）

Speech and Language Processing

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

【美】 Daniel Jurafsky
James H. Martin 著

冯志伟 孙乐 译

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

中国工信出版集团

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

经典译丛·人类语言技术

自然语言处理综论

（第二版）

Speech and Language Processing

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

Second Edition

[美] Daniel Jurafsky
James H. Martin 著

冯志伟 孙 乐 译

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

内 容 简 介

本书全面论述了自然语言处理技术。本书在第一版的基础上增加了自然语言处理的最新成就，特别是增加了语音处理和统计技术方面的内容，全书面貌为之一新。本书共分五个部分。第一部分“词汇的计算机处理”，讲述单词的计算机处理，包括单词切分、单词的形态学、最小编辑距离、词类，以及单词计算机处理的各种算法，包括正则表达式、有限状态自动机、有限状态转录机、N 元语法模型、隐马尔可夫模型、最大熵模型等。第二部分“语音的计算机处理”，介绍语音学、语音合成、语音自动识别以及计算音系学。第三部分“句法的计算机处理”，介绍英语的形式语法，讲述句法剖析的主要算法，包括 CKY 剖析算法、Earley 剖析算法、统计剖析，并介绍合一与类型特征结构、Chomsky 层级分类、抽吸引理等分析工具。第四部分“语义和语用的计算机处理”，介绍语义的各种表示方法、计算语义学、词汇语义学、计算词汇语义学，并介绍同指、连贯等计算机话语分析问题。第五部分“应用”，讲述信息抽取、问答系统、自动文摘、对话和会话智能代理、机器翻译等自然语言处理的应用技术。本书写作风格深入浅出，实例丰富，引人入胜。

本书可作为高等学校自然语言处理或计算语言学的本科生和研究生的教材，也可以作为从事人工智能、自然语言处理等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：01-2008-4029

图书在版编目(CIP)数据

自然语言处理综论：第 2 版/(美)朱夫斯凯(Jurafsky,D.)，(美)马丁(Martin,J. H.)著；冯志伟，孙乐译.
北京：电子工业出版社，2018.3
(经典译丛·人类语言技术)
书名原文：Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition
ISBN 978-7-121-25058-3

I. ①自…　II. ①朱…　②马…　③冯…　④孙…　III. ①自然语言处理　IV. ①TP391

中国版本图书馆 CIP 数据核字(2014)第 286322 号

策划编辑：马　岚
责任编辑：葛卉婷
印　　刷：三河市鑫金马印装有限公司
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：51　字数：1372 千字
版　　次：2005 年 6 月第 1 版
　　2018 年 3 月第 2 版
印　　次：2018 年 3 月第 1 次印刷
定　　价：198.00 元

凡所购买电子工业出版社的图书有缺损问题，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)88254888，88258888。

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。

本书咨询联系方式：classic-series-info@phei.com.cn。

译者简介

冯志伟 先后在北京大学和中国科学技术大学研究生院两次研究生毕业，获双硕士学位。1978年至1981年，在法国格勒诺布尔理科医科大学应用数学研究所(IMAG)自动翻译中心(CETA)师从法国著名数学家、国际计算语言学委员会主席B. Vauquois教授，专门研究数理语言学和机器翻译问题。回国后，先后担任中国科学技术信息研究所计算中心机器翻译研究组组长、教育部语言文字应用研究所计算语言学研究室主任、杭州师范大学外国语学院高端特聘教授。1986年至2004年，在德国Fraunhofer研究院(FhG)、Trier大学、Konstanz高等技术学院、韩国Korean Advanced Institute of Science and Technology (KAIST)、英国Birmingham大学担任教授或研究员，长期从事语言学和计算机科学的跨学科研究，是我国计算语言学事业的开拓者之一。在中国，他是中国语文现代化学会副会长、中国应用语言学学会常务理事、中国人工智能学会理事、国家语言文字工作委员会21世纪语言文字规范(标准)审定委员会委员、全国科学技术名词审定委员会委员、全国术语标准化技术委员会委员、中国外语教育研究中心学术委员会委员、《数学辞海》总编辑委员会委员、《中国大百科全书》(《语言文字卷》)编辑委员会成员。在国际上，他是TELRI(Trans-European Language Resources Infrastructure)、LREC(Language Resources and Evaluation Conference)、COLING-2010(Computational Linguistics Conference)的顾问委员会委员，并担任IJCL(International Journal of Corpus Linguistics)、IJCC(International Journal of Chinese and Computing)等重要学术期刊编委以及英国Continuum出版公司系列丛书Research in Corpus and Discourse编委。承担国家自然科学基金项目和国家社会科学基金项目多项，出版专著30余部，发表论文300余篇。

孙 乐 1998年5月毕业于南京理工大学，获博士学位。1998年9月至2000年10月在中国科学院软件研究所从事博士后研究，现为中国科学院软件研究所中文信息处理研究室研究员、博士生导师。曾先后在英国Birmingham大学、加拿大Montreal大学做访问学者。目前主要研究方向：自然语言理解、知识图谱、信息抽取、问答系统等。作为项目负责人承担国家自然科学基金重点项目、国家"863"项目、国际合作项目等30多项，在ACL、SIGIR、EMNLP等重要国际会议和国内核心期刊发表论文50多篇。现为中国中文信息学会副理事长兼秘书长、中文信息学报副主编、国家语委语言文字规范标准审定委员会委员、国际测评NTCIR MOAT中文简体任务的组织者、第23届国际计算语言学大会(COLING 2010)组织委员会联席主席、第13届国际机器翻译峰会(MT Summit 2011)组织委员会联席主席、第53届国际计算语言学年会(ACL2015)组织委员会联席主席。

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 Professor Feng Zhiwei, aided by Dr. Sun Le, undertook the arduous job of translating this book. Prof. Feng 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 Prof. Feng 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，鼓励他们去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，我们这本关于这个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，并在学习过程中一步一步地来感受人类语言的美妙性、复杂性和多样性。因此，当我们了解到冯志伟教授在孙乐研究员的协助下承担了把这本教材翻译成中文的艰辛工作的时候，我们感到无比的荣幸。我们认为，冯志伟教授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，既包括普遍的计算语言学研究，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

直是冯教授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！这个中文译本是英文原著的第一个外文译本，它的出版使我们非常之激动和振奋。众所周知，中国在语言研究方面有着悠久的历史，在新世纪，中国年轻一代的科学工作者在这个领域的一些重要的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，并且希望，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，由于冯志伟教授和孙乐研究员的帮助，拙著也能够为此尽我们的绵薄之力！

Daniel Jurafsky
James H. Martin

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60年代才开始的，60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，NLP)。

我们认为，计算机对自然语言的研究和处理，一般应经过如下4个方面的过程：

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《计算机进展》(Advances in Computers)第47卷的《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4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含义。

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含义的知识。

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。如果我们对计算机发一个口头的指令："Delete file x"（"删除文件 X"），我们要通过自然语言处理系统让计算机理解这个指令的含义，并且执行这个指令，一般来说需要经过如下的处理过程：

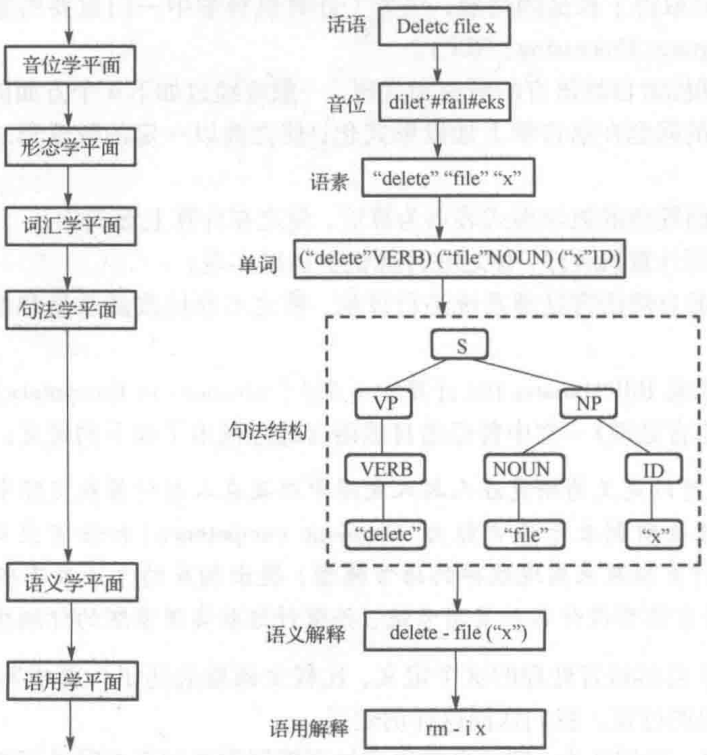

图 0.1　自然语言处理系统中的知识平面

从图 0.1 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" "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，称为 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 个平面的知识；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，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，它涉及的领域如此之广泛，我们翻译的这本《自然语言处理综论》正好满足了这样的要求。

本书的英文原名是：Speech and Language Processing：An Introduction to Natural Language Processing，Computational Linguistics，and Speech Recognition，作者是美国科罗拉多大学的 Daniel Jurafsky 和 James Martin，由 Prentice-Hall，Inc. 出版。

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，在北京的机器翻译研讨会上，电子工业出版社编辑找到我，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，却不像原来想象的那样容易，要把英文的意思表达为确切的中文，下起笔来，总有绠短汲深之感，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，当翻译完第14章（全书的三分之二）的时候，我患了黄斑前膜的眼病，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孙乐研究员表示愿意继续我的工作，与我协作共同完成本书的翻译。孙乐研究员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。孙乐研究员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。本书第一版的中文译文在2005年6月出版。

中文第一版出版后，读者的反响比我预想的热烈。中国传媒大学、北京大学、上海交通大学、解放军外国语学院、大连海事大学都先后采用本书作为自然语言处理或计算语言学课程的教材，受到师生们的一致好评。有的同学对照英文原文，逐词逐句地阅读，反复推敲译文的含义，细心品味原著的内容。有的同学组织起来集体阅读，组织专题讨论，交流学习的心得体会。有的同学写信给我，赞扬本书的译文“既信且达，通顺流畅”；这样的赞扬，对于写信的同学，自然是普通寻常的溢美之词，但对于我这个苦心推敲译文、年逾古稀的译者来说，却是最高的褒奖了。我在这里发自内心地感谢广大读者对于本书的厚爱。

现在本书中文第一版已经销售一空了，很多喜爱自然语言处理的读者想买此书，可是，经常是“一书难求”。

近年来，自然语言处理领域在很多方面有了新的进展，语音和语言技术的应用范围日益扩大，大规模真实书面文本语料库口语语料库的广泛使用，使得自然语言处理技术越来越依赖于统计机器学习的方法。2009 年 Prentice Hall 出版社推出了本书英文版第二版，篇幅由第一版的 21 章增加为 25 章，大大地充实了语音识别、语音合成、统计自然语言处理和统计机器学习方面的内容，更好地反映了这个领域的新进展。

为了满足读者进一步学习的需要，电子工业出版社决定请我和孙乐研究员翻译本书英文版第二版。我们愉快地接受了翻译第二版的任务。我们仍然按照翻译第一版时的分工，由我翻译第 1 ~ 16 章（全书的五分之三），由孙乐研究员翻译第 17 ~ 25 章（全书的五分之二），全书的译文由我统稿。

我在九年前已经步入古稀之年，自从双目出现黄斑前膜之后，视力越来越差，在翻译过程中，我经常要借助于放大镜来阅读英文原著或查询生僻的专业术语，幸好先进的语音合成技术可以把书面的文字转换成口头的语音，使得我能够通过合成的语音来校正中文译文中的差错，省去了我直接用眼力阅读中文译文之苦，我成为了自然语言处理技术的直接受益者，这更加激励我克服重重的困难来完成本书第二版的翻译。我暗暗下定决心，一定要把自己的心血化作火红的宝石，一定要把自己的汗水化作晶莹的珍珠，为这个新的译本增添璀璨的光彩。经过三年多艰辛的工作，在孙乐研究员的积极配合之下，第二版的中文译文终于与读者见面了，这是我最感到欣慰的事。

我研究自然语言处理已经五十多年了，五十多年前，我还是一个不谙世事的十九岁的小青年，现在，我已经是白发苍苍的古稀老人了，我们这一代人正在一天天地变老；然而，我们如痴如醉地钟爱着的自然语言处理事业却是一个新兴的学科，她还非常年轻，充满了青春的活力，尽管她还很不成熟，但是她无疑地有着光辉的发展前景。我们个人的生命是有限的，而科学知识的探讨和研究却是无限的。我们个人渺小的生命与科学事业这棵常青的参天大树相比较，显得多么微不足道，犹如沧海一粟。想到这些，怎不令我们感慨万千！“路漫漫其修远兮，吾将上下而求索”，自然语言处理的探索者任重道远，不论在理论方面还是在应用方面，我们都需要加倍地努力，当前自然语言处理仍然面临诸多的困难，我们还要继续奋战，才能渡过难关，走向一马平川的坦途。谨以这个新的译本献给那些对自然语言处理有兴趣的读者，让我们携起手来，共同来探索自然语言计算机处理的奥秘，并在这样的探索中实现我们个人渺小生命的价值，获取人生的乐趣。

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，本书中文译文第二版的出版能够在我国的自然语言处理的教学和科学研究中，继续产生积极的作用，我们还希望，读者能够喜欢这个新的译本，并给我们提出批评和指正。

本书译者的部分工作得到了国家自然科学基金（编号：61433015）、国家社会科学基金（编号：03BYY019）的资助，特此致谢。

冯志伟

于杭州

序　言

语言学是一门有数百年历史的学科，作为计算机科学的一个组成部分的计算语言学只有50年的历史。然而只是近十年来，由于适用于互联网的信息检索和机器翻译的出现，由于台式计算机上的语音识别逐渐普及，语言的计算机理解才真正成为了一个产业脱颖而出，它牵涉到了成千上万的人。语言信息的表示和计算机处理方面的理论进展，使这样的产业成为可能。

《自然语言处理综论》是第一本全面地论述语言技术的书，这本书的内容涉及了语言技术的各个层面，介绍了语言处理的各种现代技术，并把深入的语言分析和鲁棒的统计方法紧密地结合起来。从层次的角度来看，本书的论述是按照不同的语言层面逐步展开的，首先论述词和词的构成，包括单词序列的性质以及如何说出并且理解它们，接着论述组词成句的方法(句法)，意义形成的方法(语义学)，它们是问答系统、对话系统和语言之间翻译的基础。从技术的角度来看，本书介绍了正则表达式、信息检索、上下文无关语法、合一、一阶谓词演算、隐马尔可夫模型和其他概率模型、修辞结构理论等非常丰富的内容。在此之前，如果你想了解这些知识，你必须读两三本不同的书。本书全面覆盖了这些知识。更重要的是，本书把这些技术彼此联系起来，使读者不仅知道哪些技术是最好用的，并且知道怎样把这些技术结合起来使用。本书的论述风格使读者对于有关的内容始终保持着浓厚的兴趣，乐意去思考各种技术的细节，一步一个脚印地循序渐进而毫无枯燥乏味之感。不论你是从学术的角度还是从产业的角度对于自然语言的计算机处理发生兴趣，本书都可以作为你理想的入门向导和有用的学术参考，它能指引你在将来进一步升堂入室，研究这门引人入胜的学科。

本书第一版自2000年出版以来，这个领域在很多方面有了新的进展。语言技术的应用日益扩大，大规模的语言数据集的使用(不论是书面的还是口头的)使得我们越来越依赖于统计机器学习的方法。本书的第二版从理论和实际两个方面很好地反映了这些新的进展。本书的各个章节之间大都保持着相对的独立性，这样的结构安排也使得读者或教师更容易从中选择一部分来学习。从本书第一版出版以来，尽管在语言处理这个领域出现了一些写得不错的著作，但是，从总体上来说，本书仍然是这个领域中最好的导论性著作。

Peter Norvig & Stuart Russell
Prentice Hall 人工智能丛书主编

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。基于网络的语言技术的开发，基于电话的对话系统的商品化应用，语音合成和语音识别都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著作时，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，本书全面地覆盖了在传统上分别在不同的系和不同的课程中讲授的内容。例如，在电子工程系的语音识别课程的内容；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、机器翻译等内容；在语言学系的计算语言学课程中的计算形态学、计算音系学和计算语用学等内容。本书介绍了这些领域中的基本算法，不论这些算法原来是在语音处理还是在书面语言处理中提出的，不论它们原来是从逻辑的角度还是从统计的角度提出的，我们力求把来自不同领域的算法合在一块统一地加以描述。我们也试图把一些诸如机器翻译、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，使它的覆盖面更加全面。这种广为覆盖的方法的一个潜在问题使得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，尽管这本书写得这么长，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、人工智能、机器学习、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(MHH)到合一算法，从λ运算到对数-线性模型]应用于解决现实世界中遇到的各种重要问题。例如，语音识别、机器翻译、网络上的信息抽取、拼写检查、文本文献检索以及口语对话代理。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，当我们介绍有关自然语言处理的知识的时候，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系统的有组织的评测活动越来越多，这些使得评测得到了越来越多的强调和重视。因此，我们在本书的许多章节中都包括了评测的内容，描述系统评测和错误分析的现代经验方法，例如，训练集和测试集的概念、交叉验证(cross-validation)，以及诸如困惑度(perplexity)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基础上的。这些资源有：语音生语料库

和文本生语料库、标注语料库和树库、标准的标注集等。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，Fisher，CALLHOME，ATIS，TREC，MUC，BNC 等语料库)，并且提供很多有用的标记集的完整的清单以及编码技巧(例如，Penn Treebank，CLAWS 标记集以及 ARPAbet)，不过难以避免会有遗漏。此外，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(http：//www. cs. colorado. edu/ ~ martin/slp. html)，在这个网站上，这些资源可以得到及时的更新。

本书首先可以用作研究生或高年级本科学生的教科书或系列教材。由于本书的覆盖面广，并且有大量的算法，所以，本书也可以用作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

除了序言和书后面的附录之外，本书共分五个部分。第一部分“单词”，讲述与单词和简单的单词序列的计算机处理有关的概念：单词切分，单词的形态学，单词编辑距离，词类，以及单词计算机处理中的各种算法：正则表达式、有限自动机、有限转录机、*N* 元语法模型、隐马尔可夫模型、对数线性模型等。第二部分“语音”，首先介绍语言语音学，然后讲述语音合成、语音识别以及计算音系学中的语言问题。第三部分“句法”，介绍英语的短语结构语法，讲述用于单词之间的句法结构关系的一些主要的算法：CKY 剖析算法、Earley 剖析算法、统计剖析、合一与类型特征结构，以及诸如 Chomsky 层级分类和抽吸引理(pumping lemma)等分析工具。第四部分“语义学和语用学”，介绍一阶谓词演算以及语义的各种表示方法，λ 计算，词汇语义学，诸如 Wordnet，PropBank 和 FrameNet 等词汇语义资源，用于计算单词相似度和词义排歧的词汇语义学的计算模型，以及诸如同指(coreference)和连贯(coherence)等话语分析问题。第五部分“应用”，讲述信息抽取、机器翻译、对话和会话的智能代理等。

本书的使用方法

本书材料丰富，可供一整年的语音处理和语言处理系列教材之用。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. 正则表达式，FSA
4. *N* 元语法	4. *N* 元语法	4. *N* 元语法	3. 形态分析，FST
5. 词类标注	5. 词类标注	5. 词类标注	4. *N* 元语法
12. 上下文无关语法	6. HMM	6. HMM	5. 词类标注
13. 句法剖析	12. 上下文无关语法	8. TTS	13. 句法剖析
14. 统计剖析	13. 句法剖析	9. ASR	14. 统计剖析
19. 词汇语义学	14. 统计剖析	12. 上下文无关语法	15. 计算复杂性
20. 计算词汇语义学	17. 语义学	13. 句法剖析	16. 合一
23. 问答和摘要	18. 计算语义学	14. 统计剖析	20. 计算词汇语义学
25. 机器翻译	19. 词汇语义学	17. 语义学	21. 计算话语学
	20. 计算词汇语义学	19. 词汇语义学	
	21. 计算话语学	20. 计算词汇语义学	
	22. 信息抽取	22. 信息抽取	
	23. 问答和摘要	24. 对话	
	25. 机器翻译	25. 机器翻译	

本书的某些章节也可以选作人工智能、认知科学、信息检索或面向语音处理的电子工程等课程之用。

致谢

Andy Kehler 为本书第一版写了“话语”这一章，在第二版中，我们把这些材料作为写作这一章的起点。这一章仍然保持了 Andy 的文体和结构。与此类似，Nigel Ward 为本书第一版写了“机器翻译”这一章中的大部分材料，在第二版中，我们把这些材料作为写作“机器翻译”这一章的起点。我们保持了这一章中的大部分文字，特别是保持了 25.2 节、25.3 节以及这一章的练习。Kevin Bretonnel Cohen 写了关于生物医学信息抽取的 22.5 节。Keith Vander Linden 写了本书第一版的“生成”这一章。我们还要感谢冯志伟(Feng Zhiwei)教授，他在孙乐(Sun Le)的协助下，把本书第一版翻译成了中文。

科罗拉多大学的所在地博尔德市(Boulder)和斯坦福大学的所在地斯坦福市(Stanford)都是从事语音处理和语言处理的好地方。这里，我们还要感谢在这两个地方的我们的学系、我们的同事们以及我们的学生们，他们给了我们的研究和教学极大的影响。

Daniel Jurafsky 在此还要感谢他的父母，是他们鼓励 Daniel 把每件事都做得尽善尽美，并按时完成。他还要感谢 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“这行吗?”他还要感谢 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?”他还要感谢 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，并教导他要始终重视数据。他还要感谢 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。他还要感谢 Chris Manning，Chris 是他在斯坦福最出色的合作者。他还要感谢过去在博尔德的非常好的所有的同事们。

James H. Martin 在此也要感谢他的父母，是他们给了 James 鼓励，并且容许他走上自然语言处理这条在当时看来似乎有点儿古怪的学术道路。他还要感谢他的博士论文的指导老师 Robert Wilensky，是 Robert 使他有机会在伯克利(Berkeley)开始了自然语言处理的学习。他还要感谢 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。他还要感谢 Rick Alterman，是 Rick 在关键和困难的时刻，给了他鼓励和勇气。他还要感谢 Chuck Fillmore，George Lakoff，Paul Kay 和 Susanna Cumming，因为他们教 James，使他懂得了语言学。他还要感谢 Martha Parmer，Tammy Summer 和 Wayne Ward，他们是 James 在博尔德最好的合作者。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。James 还要感谢他的女儿 Katie，她全身心地等待着本书这个版本的完成。

我们要感谢在本书第一版时给我们巨大帮助的很多人。本书的第二版也得益于我们的很多读者，他们仔细地阅读了本书并且进行了试教。我们特别感谢朋友们对于本书所涉及的广泛领域提出的很有帮助的意见和建议，他们是 Regina Barzilay，Philip Resnik，Emily Bender，Adam Przepiórkowski，我们的编辑 Tracy Dunkelberger，我们的高级编辑经理 Scott Disanno，我们的序列丛书编辑 Peter Norvig 和 Stuart Russell。我们的制作编辑 Jane Bonnell 对于本书的设计和内容也提出了很多有帮助的建议。我们还要感激许多朋友和同事们，他们或者阅读了本书的个别章节，或者在他们的意见和建议中，回答了我们的很多问题。我们还要感激科罗拉多大学和斯坦福大学上这门课的学生们，以及伊利诺依大学厄巴纳-香槟分校

(1999)、麻省理工学院(2005)和斯坦福大学(2007)参加 LSA 暑期学院的学生们。此外，我们还要感谢下面的朋友：

Rieks op den Akker, Kayra Akman, Angelos Alexopoulos, Robin Aly, S. M. Niaz Arifin, Nimar S. Arora, Tsz-Chiu Au, Bai Xiaojing, Ellie Baker, Jason Baldridge, Clay Beckner, Rafi Benjamin, Steven Bethard, G. W. Blackwood, Steven Bills, Jonathan Boiser, Marion Bond, Marco Aldo Piccolino Boniforti, Onn Brandman, Chris Brew, Tore Bruland, Denis Bueno, Sean M. Burke, Dani Byrd, Bill Byrne, Kai-Uwe Carstensen, Alejandro Cdebaca, Dan Cer, Nate Chamber, Pichuan Chang, Grace Chung, Andrew Clausen, Raphael Cohn, Kevin B. Cohen, Frederik Coppens, Stephen Cox, Heriberto Cuayáhuitl, Martin Davidson, Paul Davis, Jon Dehdari, Franz Deuzer, Mike Dillinger, Bonnie Dorr, Jason Eisner, John Eng, Ersin Er, Hakan Erdogan, Gülsen Eryiğit, Barbara Di Eugenio, Christiane Fellbaum, Eric Fosler-Lussier, Olac Fuentes, Mark Gawron, Dale Gerdemann, Dan Gildea, Filip Ginter, Cynthia Girand, Anthony Gitter, John A. Goldsmith, Michelle Gregory, Rocio Guillen, Jeffrey S. Haemer, Adam Hahn, Patrick Hall, Harald Hammarström, Mike Hammond, Eric Hansen, Marti Hearst, Paul Hirschbühler, Julia Hirschberg, Graeme Hirst, Julia Hockenmaier, Jeremy Hoffman, Greg Hullender, Rebecca Hwa, Gaja Jarosz, Eric W. Johnson, Chris Jones, Edwin de Jong, Bernadette Joret, Fred Karlsson, Graham Katz, Stefan Kaufmann, Andy Kehler, Manuer Kirschner, Dan Klein Sheldon Klein, Kevin Knight, Jean-Pierre Koenig, Greg Kondrak, Seleuk Kopru, Kimmo Koskenniemi, Alexander Kostyrkin, Mikoo Kurino, Mike LeBeau, Chia-Ying Lee, Jaeyong Lee, Scott Leishman, Szymon Letowski, Beth Levin, Roger Levy, Liuyang Li, Marc Light, Greger Lind'en, Pierre Lison, Diane Litman, Chao-Lin Liu, Feng Liu, Roussanka Louka, Artyom Lukanin, Jean Ma, Maxim Makatchev, Inderjeet Mani, Chris Manning, Steve Marmon, Marie-Catherine de Marneffe, Hendrik Maryns, Jon May, Dan Melamed, Laura Michaelis, Johanna Moore, Nelson Morgan, Emad Nawfal, Mark-Jan Nederhof, Hwee Tou Ng, John Niekrasz, Rodney Nielsen, Yuri Niyazov, Tom Nurkkala, Kris Nuttycombe, Valerie Nygaard, Mike O'Connell, Robert Oberbreckling, Scott Olsson, Woodley Packard, Gabor Palagyi, Bryan Pellom, Gerald Penn, Rani Pinchuk, Sameer Pradhan, Kathryn Pruitt, Drago Radev, Dan Ramage, William J. Rapaport, Ron Regan, Ehud Reiter, Steve Renals, Chang - han Rhee, Dan Rose, Mike Rosner, Deb Roy, Teodor Rus, William Gregory Sakas, Murat Saraclar, Stefan Schaden, Anna Schapiro, Matt Shannon, Stuart C. Shapiro, Ilya Sherman, Lokesh Shrestha, Nathan Silberman, Noah Smith, Otakar Smrz, Rion Snow, Andeas Stolcke, Niyue Tan, Frank Yung-Fong Tang, Ahmet Cüneyd Tantug, Paul Taylor, Lorne Temes, Rich Thomason, Almer S. Tigelaar, Richard Trahan, Antoine Trux, Clement Wang, Nigel Ward, Wayne Ward, Rachel Weston, Janyce Wiebe, Lauren Wilcox, Ben Wing, Dean Earl Wright III, Dekai Wu, Lei Wu, Eric Yeh, Alan C. Yeung, Margalit Zabludowski, Menno van Zaanen, Zhang Sen, Sam Shaojun Zhao 和 Xingtao Zhao。

还要感谢朋友们允许我们复制下面两个图：一个是图 7.3(© Laszlo Kubinyi 和《科学美国人》)，一个是图 9.14(© Paul Taylor 和剑桥大学出版社)。此外，我们自己作的很多图都是经过改编的，下面的朋友们允许我们改编他们的图(为了简单起见，对于每一个图，我们只列出一位作者)，在此，我们对他们表示感谢。如下 3 个图来自© IEEE 和它们的作者；我们感谢 Esther Levin(图 24.22)和 Lawrence Rabiner(图 6.14 和图 6.15)。我们改编的其他的图还来自如下作者的版权©，我们要感谢计算语言学学会、《计算语言学杂志》以及其他编者 Robert Dale, Regina Barzilay(图 23.19), Michael Collins(图 14.7, 图 14.10,图 14.11) John Goldsmith(图 11.18) Marti

Hearst(图 21.1 和 21.2), Kevin Knight(图 25.35), Philipp Koehn(图 25.25, 图 25.26 和图 25.28), Dekang Lin(图 20.7), Chris Manning(图 14.9), Daniel Marcu(图 23.16), Mehryar Mohri(图 3.10 和图 3.11), Julian Odell(图 10.14), Marilyn Walker(图 24.8, 图 24.14 和图 24.15), David Yarowsky(图 20.4)和 Steve Young(图 10.16)。

Daniel Jurafsky
于加利福尼亚州斯坦福市
James H. Martin
于科罗拉多州博尔德市

目　录