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

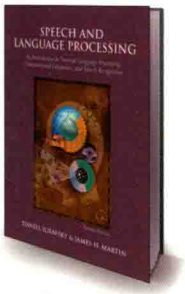
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

经典译丛

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



本书自第一版出版以来，一直好评如潮，被国外许多大学选作自然语言处理或计算语言学的教材。

本书第一版综合了自然语言处理、计算语言学和语音识别的内容，全面论述计算机自然语言处理，深入探讨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。该版对于第一版做了全面的改写，增加了大量反映自然语言处理新成就的内容，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。

本书四大特色：**覆盖全面　强调实用　注重评测　语料为本**

本书配套网站：**http：//www.prenhall.com/jurafsky-martin**

作者简介

Daniel Jurafsky 在美国加利福尼亚大学伯克利分校于 1983 年获语言学学士学位，于 1992 年获计算机科学博士学位。现任斯坦福大学语言学系和计算机科学系副教授，主要研究方向为语言的概率模型和语音信息处理。他在语音和语言处理领域发表了 90 多篇论文，并在 1998 年获得美国国家基金会 CAREER 奖，在 2002 年获得 Mac-Arthur 奖。

James H. Martin 于 1981 年在哥伦比亚大学获计算机科学学士学位，1988 年在美国加利福尼亚大学伯克利分校获计算机科学博士学位。现任美国科罗拉多大学博尔德分校语言学系、计算机科学系教授，认知科学研究所研究员，主要研究方向为计算语义学、机器学习和信息检索。他发表过 70 多篇有关计算机科学的专著，出版了 A Computational Model of Metaphor Interpretation 一书。

译者简介

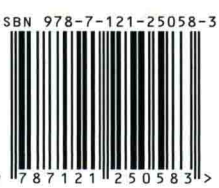
冯志伟　国家教育部语言文字应用研究所研究员、博士生导师。先后在北京大学和中国科学技术大学获双硕士学位，在语音和语言的计算机处理领域具有多年的研究经验，曾在多个国家参与研究和教学工作，主要研究方向为自然语言处理、计算语言学和机器翻译，主要著作有《自然语言的计算机处理》《数理语言学》等 18 部。

孙　乐　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。1998 年在南京理工大学获博士学位，后在中国科学院软件研究所从事博士后研究。曾先后在英国 Birmingham 大学、加拿大 Montreal 大学做访问学者。主要研究方向为自然语言理解、知识图谱、信息抽取、问答系统等。作为项目负责人完成国家级项目 30 多个，发表论文 50 多篇。

策划编辑：马　岚
责任编辑：葛卉婷
责任美编：孙焱津

Pearson
www.pearson.com

ISBN 978-7-121-25058-3



定价：198.00 元

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：

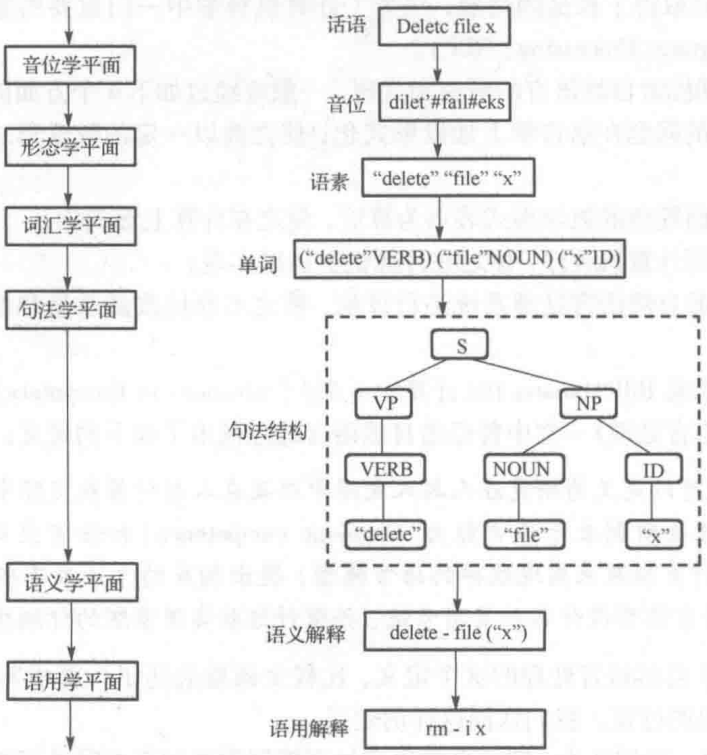


图 0.1　自然语言处理系统中的知识平面

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。