

计算语言学
与语言科技
原文丛书

CAMBRIDGE

INTRODUCING SPEECH
AND LANGUAGE PROCESSING
语音语言处理导论

作者 John Coleman
导读 常宝宝



北京大学出版社
PEKING UNIVERSITY PRESS

Introducing Speech and Language Processing 语音语言处理导论

作者 John Coleman

导读 常宝宝



北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记 图字 01-2009-4177

图书在版编目(CIP)数据

语音语言处理导论: Introducing Speech and Language Processing/克勒曼
(Coleman, J.)著. —北京:北京大学出版社, 2010. 8

(计算语言学与语言科技原文丛书)

ISBN 978-7-301-17153-0

I. 语… II. 克… III. 自然语言处理—高等学校—教材—英文 IV. TP391

中国版本图书馆 CIP 数据核字(2010)第 077685 号

Introducing Speech and Language Processing, first edition (ISBN: 978-0-521-53069-5) by Coleman first published by Cambridge University Press 2005.

All rights reserved.

This reprint edition for the People's Republic of China is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press & Peking University Press 2010

This book is in copyright. No reproduction of any part may take place without the written permission of Cambridge University Press and Peking University Press.

This edition is for sale in the People's Republic of China (excluding Hong Kong SAR, Macau SAR and Taiwan Province) only.

此版本仅限在中华人民共和国境内(不包括香港、澳门特别行政区及台湾地区)销售。

书 名: 语音语言处理导论

著作责任者: John Coleman 著

责任编辑: 白雪 李凌

标准书号: ISBN 978-7-301-17153-0/H·2495

出版发行: 北京大学出版社

地 址: 北京市海淀区成府路 205 号 100871

网 址: <http://www.pup.cn>

电子邮箱: zpup@pup.pku.edu.cn

电 话: 邮购部 62752015 发行部 62750672 编辑部 62753334

出版部 62754962

印 刷 者: 三河市北燕印装有限公司

经 销 者: 新华书店

787 毫米×1092 毫米 16 开本 20.75 印张 469 千字

2010 年 8 月第 1 版 2010 年 8 月第 1 次印刷

定 价: 46.00 元(含 CD-ROM 1 张)

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究 举报电话: 010-62752024

电子邮箱: fd@pup.pku.edu.cn

《计算语言学与语言科技原文丛书》由北京大学—香港理工大学汉语语言学研究、北京大学计算语言学研究所(由 973 课题“文本内容理解的数据基础”支持)和北京大学出版社合作推出

学术委员会 Academic Advisory Committee

主 任:

黄居仁(香港)

委 员:

陈克健(台北)

Chris Manning (Stanford)

董振东(北京)

Harold Somers (Dublin)

李宇明(北京)

陆俭明(北京)

Maarten de Rijke (Amsterdam)

沈 阳(北京)

石定栩(香港)

苏克毅(台北)

Suzanne Stevenson (Toronto)

王逢鑫(北京)

王厚峰(北京)

王士元(香港)

谢清俊(台北)

俞士汶(北京)

松本裕治(奈良)

郑锦全(Urbana-Champaign)

邹嘉彦(香港)

编委会

Editorial Committee

主 编：

黄居仁教授(香港)

编 委：

冯志伟教授(北京)

顾曰国教授(北京)

黄伟道教授(Singapore)

黄莹菁教授(上海)

姬东鸿教授(武汉)

陆 勤教授(香港)

蒙美玲教授(香港)

苏新春教授(厦门)

孙茂松教授(北京)

陶红印教授(Los Angeles)

徐飞玉教授(Saarbrücken)

薛念文教授(Waltham)

杨立范编审(北京)

俞士汶教授(北京)

曾淑娟副研究员(台北)

詹卫东副教授(北京)

赵铁军教授(哈尔滨)

周 明研究员(北京)

宗成庆研究员(北京)

常宝宝副教授(执行秘书)(北京)

丛书前言

计算语言学(Computational Linguistics, CL)在语言科学与信息科学的研究领域扮演关键性的角色。语言学理论寻求对语言现象规律性的揭示与完整的解释。计算语言学正好提供了验证与应用这些规律与解释的大好机会。作为语言学、信息科学乃至至于心理学与认知科学结合的交叉学科,计算语言学更提供了语言学基础研究与应用研究的绝佳界面。事实上,计算语言学与人类语言科技(Human Language Technology, HLT)可以视为一体两面,不可分割。

计算语言学研究滥觞于上世纪五六十年代的机器翻译研究。中文的相关研究也几乎同步开始,1960年起在柏克莱加州大学研究室,王士元、邹嘉彦、C. Y. Dougherty 等人已开始研究中英、中俄机器翻译。他们的中文计算语言学研究,可说是与世界最尖端科技同步的。中国国内中俄翻译研究也不遑多让,大约在上世纪 50 年代中期便已开始。可惜的是,这些中文相关早期机器翻译研究,由于硬件与软件的限制,没能延续下来。中文计算语言学研究比较有系统的进展,还要等到 1986 年;海峡两岸在同一年成立了两个致力于中文计算语言学基础架构建立的研究群。北京大学的计算语言学研究所在朱德熙先生倡导下成立,随后一段时间由陆俭明、俞士汶主持。而台湾“中研院”的中文词知识库小组,由谢清俊创立,陈克健主持,黄居仁 1987 年返台后加入。

中文计算语言学的研究,20 余年来已累积了相当可观的成绩,重要研究领域与议题中都有可观的研究成果,华人计算语言学者也渐渐在国际学术界崭露头角。随着世界经济转向知识密集产业,跨语言跨文化沟通与知识整合成为知识产业的关键,语言科技的发展日渐成为国际主流。在这个有利发展的大环境下,我们相信,中文计算语言学与华人计算语言学学者的成绩,将会百尺竿头更进一步,进入计算语言学学术核心,并产生把握学科动态、引领学术走向的大师。

回顾计算语言学研究在过去二十年的蓬勃发展,统计模式的引入应该是最主要的原因之一。但二十年后学界也开始看到了统计模式的局限,因此最近几届 ACL 终身成就奖得主,不约而同地大力提倡结合语言学理论与概率

模型的研究,来提升计算语言研究的层次,以寻求新的突破。

回顾中国国内的计算语言学发展,来自计算机科学的贡献多于语言学的贡献。这在理论与概率模型整合研究的大趋势下,不免令人忧心。这也许可以部分归咎于英文研究专著获得不易。国内较易取得期刊或会议论文,但由于篇幅的限制,往往无法对理论做深入完整的阐述,因此也导致国内年轻学者,长于运算而拙于理据。因此,藉由英文专书来弥补不足,巩固研究理据,进而开拓研究视野,是非常重要的一步。

剑桥大学计算语言学原版书系列的引进,就是在上述背景下产生的。本人忝为 Cambridge University Press 所出版的 Studies in Natural Language Processing 系列编辑委员之一,并将于 2010 接任主编。能够将此系列中较重要的几部著作引进国内,责无旁贷。引进原版,不是难事;要真正搭建知识的桥梁,使国内学者与学生开拓研究视野,将原文著作的理论精髓,更多应用于中文研究,则需另加努力。因此,本丛书的特色,是在保留原版的基础上,每本书都邀请一位专家撰写中文导读,其着力点有三:

其一,全书内容简介。导读作者长年浸淫于该领域,对原著能提纲挈领,切中肯綮,并提供相关研究背景。可助读者更准确地掌握并吸收该书的内容。其二,中文相关研究。原作不一定会提到相关的中文研究。由导读专家补充介绍,能搭起理论与中文相关应用的桥梁,从而能够使读者掌握在这个议题进入中文研究的最佳切入点,让相关中文研究的开拓者获得理论的参照和指导。其三,补充原书出版后该领域研究的新发展。现代科技发展迅速,任何经典著作出版后,几乎马上有新的相关研究。因此,在理论架构的脉络中,加上新近发展,使读者能更贴切地掌握研究脉动。全书摘要通常采用文字叙述。而中文相关研究及最新研究发展则分别以文字叙述及延伸阅读书目的方式呈现。延伸阅读书目,使读者可以很快上手,进入相关研究领域,也是本丛书策划者的苦心所在。可以说导读是本丛书的亮点,不特为原书增色,亦且增加了不少附加价值。

本丛书的出版,是多方协作的结果。在规划出版的漫长过程中,北大计算语言学研究所俞士汶老师及常宝宝老师提供了无私无悔的支持。香港理工大学,特别是北大一理大汉语语言学与陈瑞端、石定栩、沈阳几位在关键时刻的挹注,也起到了关键作用。当然,整个系列能够顺利出版,离不开有学术眼光和胸襟的北大出版社的支持,而剑桥出版社主管编辑 Helen Barton 从中斡旋,使合约能顺利签订,是必不可少的一环。最后,我要感谢本丛书的国内编委,特别是此次担任导读的各位主笔的辛勤付出,他们为读者搭建了进入学术殿堂的台阶。本丛书的出版,适逢 2010 COLING 国际计算

语言学会议在北京举办之际,正象征着国内计算语言学研究与国际的接轨;国内学者风云际会,大展身手,跻身计算语言学的国际舞台,将指日可待。

丛书主编

黄居仁

谨志于香港红磡

二零一零年元月

导 读

常 宝 宝

1. 学科背景及本书的定位

语音和语言处理都以人类语言为处理对象,但处理的层面却不相同。语音处理关注的是利用计算机对语音信号进行分析和综合,从而实现合成和识别人类语音。而语言处理更多关注对人类语言的句法、语义进行建模,终极目标是实现人类语言的机器理解。

尽管语音处理和语言处理的对象都是人类语言,不过传统上这是两个相互独立的学科,这两个领域的研究和教学也是独立进行的,教科书内容通常也不交叉。不过近年来,随着这两个学科的进一步发展,它们之间的关系变得日益密切,两个学科之间的联系越来越受到人们的重视。人们越来越意识到,对于语言理解而言,语音特征绝非没有价值。反过来,语音处理研究也需要关注语言处理的进展,将语言处理的成果应用到语音识别和语音合成研究,有利于提升语音处理的性能。此外,这两个学科在方法论方面也有共性,相互影响和借鉴的情况也时有发生。例如,在 20 世纪 80 年代,隐马尔可夫模型被用来解决词性自动标注问题,就是受到该模型应用在语音识别研究中获得成功的启发,尤其值得一提的是,隐马尔可夫模型引入语言处理领域所产生的影响决不是局部性的,它带动了以语料库为基础的经验主义方法在语言处理领域的全面复苏,有着十分重要的意义。这些事例说明,语音和语言处理正在日趋交叉和融合,未来的相关研究和开发需要对这两个领域都有相当理解和把握的人才。

语音和语言处理又均是对技术背景要求很高的交叉学科,搞清语音的发生机制,需要对人类的发声器官以及声学知识有足够的了解,也需要为语言和语音建立数学模型,并运用计算机的手段去实现语音和语言的计算处理。语音和语言处理的研究人员主要来自理工科专业。因此,大多数关于语音和语言处理的教材都假定读者有较强的理工科背景,教科书中常常充满了数学公式和推导,这样的教科书放在计算机系和电子工程系这样的系科使用是适

当的,但对于传统上属于人文学科的语言学系的学生而言,则无疑会对他们造成很多理解上的困扰。作为语言学专业的学生,他们一方面可能成为语音和语言处理研究的参与者,另一方面他们也是语言语音处理技术的使用者和实践者,借助语言和语音处理工具,他们可以更好地观察和分析语言,从而推动语言学本身的研究。可是要有效参与语音语言处理研究或正确使用语音语言处理工具,只能在对语言和语音处理技术有相当理解的基础上才有可能。但是面向理工背景编写的教材却与来自文科背景的阅读需求存在着明显的脱节。

《语音语言处理导论》是一本写作较为成功的关于语音语言处理技术的入门级教材,本书可以说打破了两个界限,因而非常值得推荐。首先本书打破了语音和语言处理的界限,在一本教科书中综合讲授了这两个领域的基本概念和方法;二是本书也打破了文理学科的界限,巧妙的选材和灵活的讲授方式,让语言学专业的学生也能免受背景知识的制约,轻易得窥语音语言处理技术的精髓。

2. 内容提要

本书共分为九章,第一章是导引,简要介绍了教材的主要内容、结构、目标读者以及体例安排。其余八章分别介绍了语音信号的数字表示、数字滤波器及其在语音合成中的应用、语音信号的倒谱分析和线性预测编码、有限状态技术及其在语音语言处理中的应用、语音识别技术、概率化有限状态技术和隐马尔可夫模型、句法分析、概率化上下文无关文法。在这八个部分中,有四个部分属于语音处理技术,分别是第二、三、四和六章,第二、三、四章介绍语音在信号层面的处理方法,在夯实语音处理的技术基础同时,简要介绍了语音合成技术,第六章则是介绍了语音识别技术。第五、七章内容跨越语音和语言处理两个领域,因为有限状态技术在两个领域中都是基本的处理技术。第八、九章是语言处理技术的代表,介绍自然语言的句法分析。纵观全书,语音处理和语言处理这两条主线既分且合,内容安排自然有序。

3. 章节内容介绍

第一章 导引

介绍了本书的定位和编写目标,开宗明义地指出,本书是语音和语音处理方面的初级教材,面向的读者是缺乏程序设计经验和背景知识的读者,主

要面向语言学等文科领域的学生,指出他们需要学习语言语音处理知识的理由。还概要介绍了本书的主要内容和章节安排,实验实践所需要的硬件配置、计算机应用技能等。

第二章 声音和数字

本章以余弦波为例通俗地解释了声音信号的数字表示方式,并讲授了模拟声音信号如何通过采样、量化方式转换为数字声音信号以及正确数字化音频信号对采样频率的要求(即采样定律)。由于考虑到读者的语言学背景,在这一章中,除讲授语音的数字表示问题外,穿插了必要的计算机程序设计知识和三角知识的讲解,涵盖的基础概念包括 C 语言程序、目标代码、机器代码、编译、链接、C 语言中的数值类型、数组的定义和存储、程序结构、控制结构、弧度的数学定义等。本章还以习题的方式讲解了如何运用 GCC 编译器完成 C 语言程序的编译和链接,以及如何使用声音处理工具 CoolEdit 2000 观察声音的波形显示等。

第三章 数字滤波器和共鸣器

本章介绍了如何运用计算机对数字语音信号进行加工和处理。从简单入手,首先介绍如何计算语音信号的幅度均值和均方根(RMS)幅度值,此后逐渐转向深入,引入滑动均值及数字滤波器的概念,解释了滤波器的截止频率和滑动窗口大小之间的关系,并在此基础上引入无限脉冲响应滤波器(IIR)的概念并给出了程序实现。还以表格的形式给出了具有不同截止频率的数字滤波器的系数设置,在这些系数以及所附程序的基础上,相信读者可以轻松实现各种低通、高通以及带通数字滤波器。作为滤波器的应用,作者以 Klatt 共振峰合成器为例,简要介绍了应用各种不同参数的滤波器进行语音信号合成的原理。

第四章 频率分析和线性预测编码

本章主要介绍了语音信号的分析手段以及语音参数的提取问题。首先介绍了建立在快速傅立叶变换基础上的声谱分析及程序实现,根据书中提供的程序代码,读者可以对语音信号进行声谱分析,并可利用 Excel 提供的图表功能生成声音信号指定样本处的声谱图。接着作者说明了声谱图中的不同峰值类型,指出声谱图不能很直观地用于基频提取,并由此引入倒谱分析法。接下来作者介绍了利用逆向快速傅立叶变换对语音信号进行倒谱分析的原理并给出其程序实现,基于倒谱分析的基频提取原理及其代码实现。除基于

倒谱分析的基频提取方法外,作者还介绍了基于自相关原理的基频提取方法并给出了程序实现。为了屏蔽语音信号中清音段对基频跟踪过程的影响,作者又给出了一种浊音检测方法并透过程序代码和例子对其进行了解释。线性预测编码(LPC)是本章的另一个主题,其工作原理在于通过样本点线性组合的方式将语音样本转换为 LPC 系数和预测误差,作者给出了基于 LPC 技术对语音信号进行分析和重构的代码实现,在完成对这些代码的详细说明之后,作者又简单探讨了 LPC 编/解码在语音压缩、语音合成、语音信号参数化方面的作用。

第五章 有限状态机

本章主要介绍了有限状态技术,有限状态技术在语音和语言处理两个领域中都有很多成功的应用。在讲解有限状态技术的有关概念和原理时,作者仍然选择使用了大量的例子。通过这些例子,作者首先向读者讲解了有限状态自动机的概念、有限状态自动机的确定性和非确定性、非确定性有限状态自动机中的回溯策略等,然后作者给出了有限状态自动机的 Prolog 代码实现,并结合例子讲述了有限状态自动机如何可被用来进行字符串的生成和识别。在此基础上,作者又结合具体实例讲述了有限状态转换机的概念和工作原理,并给出了程序实现。本章所提到的例子主要有英语单音节词音位串的识别和生成、单音节词的拼写形式和音位串形式之间的转换、语音信号到音位串的转变、英文句子的句法处理等,涵盖了语音和语言处理两个领域。在程序实现方面,本章不再使用 C 语言,而使用 Prolog 语言,因此本章也穿插介绍了 Prolog 语言及其与 C 语言的不同。

第六章 语音识别技术导引

本章概要介绍了语音识别技术,特点是没有深入技术细节,只在概念层面进行讨论。同其他章节不同,本章也没有提供可供实验的程序实例。作者首先介绍了研制语音识别系统所采用的两种典型方法:基于知识的方法和基于模式匹配的方法。接下来分别介绍了这两种方法的架构和一般过程。对于前者重点探讨了特征提取和知识集成问题。对于模式匹配方法,则在概要介绍之后引出本章另一重要议题——动态时间规整(DTW)算法。本章以浅显生动的方式介绍了 DTW 算法的原理和工作过程,介绍了该算法在单词语音识别中的应用。最后本章概要探讨了给语音识别造成困难的一些语音变化问题,如方言/口音、语音的性别差异问题等,但没有谈及应对策略。

第七章 概率化有限状态模型

本章主要介绍概率化有限状态模型的基本原理及应用。针对语言中存在的不确定性问题,概率化方法具有特别的优势。而所谓概率化的有限状态模型,实际上主要指马尔可夫模型和隐马尔可夫模型,作为马尔可夫模型在语言建模问题中的应用,作者重点讲解了 n 元语言模型的建模及参数估计方法,简单讨论了不完全的训练数据会造成数据稀疏问题,给出了平滑和回退的基本概念。对于隐马尔可夫模型,罗列了模型所涉及的三个核心问题,然后简单讲述了该模型在词性标注和单词语音识别问题中的应用。概率化方法在 20 世纪 50 年代曾在语言处理领域中被提出,但遭到了乔姆斯基及其学生的反对,因此作者在本章最后一节反驳了乔姆斯基的一些论点。

总体而言,本章讲述依然通俗易懂、虽无具体的程序代码,但实例丰富,还以适当的方式补充了概率的基本概念。但作者的部分讲述方式有欠缺,尤其是对于隐马尔可夫模型和马尔可夫模型间的联系和区别讲解的不太到位。

第八章 句法分析

本章主要讲授如何自动划分句子的句法结构,即自动句法分析。自动句法分析是语言处理中的一个核心问题。由于定子句文法(DCG)与 Prolog 语言之间的紧密联系,本章采用定子句文法作为应用实例的基础文法。作者首先描述了句法分析的直观过程,而后介绍了基于递归下降策略的句法分析技术,分别作为自底而上和自顶向下的分析策略的具体示例。这些讲解逐层递进,清晰易懂。接着作者给出了句法分析的部分 Prolog 实现。基本的句法分析程序只能判定句子的合法性,而不能给出其所对应的句法树,因此作者又对程序进行了扩展以使其能够输出句子的句法树。除了传统的句法分析实例,作者还将句法分析方法用于分析单词音节结构,并附有程序实现。此后作者在简单提及其他句法分析策略之后,转入了确定性问题的讨论,说明句法分析中的非确定性问题会造成语言片段的重复分析以及处理效率低下,随后引入线图作为保存中间分析结果的数据结构,提升分析效率。最后,作者进一步讨论了句法分析中的确定性问题,介绍了旨在避免回溯的句法分析方法——确定性分析方法,提及了展望符号的概念。在本章结尾,作者还以学生提问的方式提出了并行算法的问题,指出并行策略并不能应对句法分析这种具有组合爆炸性质的问题。

第九章 概率化语法

本章介绍了概率上下文无关文法,仍着眼于句法分析问题,是第八章内

容的扩展。作者首先探讨在文法中引入概率因素的动因,接着通过实例介绍了什么是概率上下文无关文法(PCFG)以及其中的概率归一约束。然后作者讲述了 PCFG 的训练问题,指出存在两种训练方法,基于手工标记语料库的方法和基于原始语料库自动推导的方法,即一般所说的有指导方法和无指导方法。作者接着给出了训练程序和分析程序的 Prolog 代码。但需要提醒读者注意的是,作者这里给出的代码并非真正的无指导 PCFG 训练方法(即 IO 算法),因为它暗含着一个前提,即假定文法没有歧义,对语言处理而言,这是做不到的,这样的假设也与引入概率上下文无关文法的初衷不符。在本章中,作者也对概率上下文无关文法的局限性进行了说明,但这些说明不够全面,似没有触及核心。本章最后,作者简单讨论了树邻接文法(TAG)和面向数据的句法分析(DOP),前者实际针对的是上下文无关文法的不足,本身并没有概率因素,书中没有提到树邻接文法也有概率化扩展。

4. 本书的特色和不足

本书特色在于其目标定位十分明确,无论从内容组织、讲授方式还是体例安排无不围绕面向文科背景学生这一目标展开。向文科学生讲授理工内容以及为此编写教材不是新鲜事物,但有相当难度,尺度极难把握,要么对背景知识要求过多,造成学生理解上的困难,要么流于泛泛,虽然容易理解,但读后难有实质收获。难能可贵的是,本书作者 John Coleman 恰当地把握了这个尺度,真正做到了既易于理解,又不乏实质内容。在体例安排方面,本书也有许多可以称道之处,首先在每章开始处,有一个章节预览(preview)和关键术语(key terms)表,章节预览通常不足百字,简单扼要地对本章内容进行引导。关键术语表则列出了本章需要重点关注的基本概念。在每章结尾,作者又会给出一个章节摘要(summary),对本章内容做简要回顾。在每章的最后,还有两个小节的内容不应忽视,一是学完本章之后,为了加深理解,作者推荐的深入阅读材料(further reading),另一小节则是为了学习后续章节,作者推荐的准备性材料(reading in preparation for next topic)。本书的习题安排也与一般的教材有所不同,并不是所有的习题都放在章节结尾部分,而有可能散列在正文之中,甚至还可能作为引出正文话题的手段。在正文中最具特色的安排是提问回答版块,在这些与正文底色不同的区域中,作者首先以学生的身份提出问题,然后针对这些问题进行解答。这些体例上的特殊安排丰富了作者的表达手段,增加了表达的自由度和灵活性,使得讲解和叙述不必拘泥于正文逻辑的局限,有助于把内容表达的更加清楚。

在涉及技术和程序问题时,本书没有采取完全回避的态度,一般面向文科背景的教材,因为担心读者没有知识储备,从而回避技术细节和程序代码,造成内容空洞和操作性不强的问题,读者或许了解了几个干巴巴的概念,但对他们具体工作却缺乏帮助。John Coleman 在假设读者没有技术背景和程序设计经验的前提下,却在教材中使用了大量程序实例。他对这个问题的解决办法是,结合这些程序例子,同时进行程序设计和技术背景的普及。这些介绍不同于一般的程序设计教程,不对某种程序设计语言进行事无巨细的全面介绍,而只针对正确理解和使用示例程序所需要的部分知识,例如他对机器代码、字节存放顺序、数据类型、数组等内容的介绍。之所以介绍这些概念,是因为不理解这些概念,理解示例程序就会有困难。程序中没有用到的语言特色则都不会涉及,例如 C 语言中的结构类型等、Prolog 中合一概念等就不会专门介绍。对相关数学概念的介绍也是如此,例如只介绍了弧度和角度的关系、介绍了概率的简单定义等。一定程度上看,作者在讲解有关内容时,实际上始终把他的讲授对象放在心中,对他们可能遇到的技术问题都预先逐一化解。对于一些更为深入的知识,作者会采用“黑箱”的讲授方式,十分明确地告诉读者,读者无需掌握技术细节,在了解其输入输出的前提下,不会影响使用,教材中没有讲解傅立叶变换的具体算法,也没有讲解 LPC 系数计算的具体细节,但不妨碍在程序代码中使用快速傅立叶变化完成相应的处理任务。由于设想的读者是语言学专业的学生,关于语音学、语言学的一些基本概念就避开不讲,因为这些内容对于语言学背景的学生而言,不会构成障碍。不过,这样的写作是否完全成功,大概还要听取语言学专业读者的意见,但无疑这样的努力降低了他们接受这些技术话题的门槛。

本书还提供了两个外围资源,一是本书的配套光盘,光盘中除了有正文中的所有程序代码和示例数据,还有 C 语言编译器、Prolog 解释器以及音频分析工具软件 CoolEdit 2000 的试用版,避免了读者可能由于缺少这些工具无法有效实践。此外教材还有一个配套网站,读者可以在这里找到其他软件工具的链接信息、教材上练习题目的参考答案、教材内容的勘误信息等。从内容上看,本书议题不算丰富,因为跨越了两个领域,其中又穿插了许多技术背景的补充性介绍,所以本书对语音和语言两个领域的介绍都非常基础,不具有系统性,内容虽谈不上落后,却也并不新颖。在语音处理领域,语音合成和识别都是简单提及,所提到的方法也不是目前采用的主流方法。在语言处理领域,只涉及了句法分析这个话题,对词法分析、语义分析几乎没有涉及,即使对于句法分析而言,也只是介绍了基于上下无关文法的句法分析和概率上下文无关文法,而对诸如句法分析中的词汇化思想则完全没有提及。当然这与

本书的定位是有关系的,本书只是一本入门教程,内容虽然不是最新最近,但却是进一步学习前沿最新发展的基础。

本书也有不足之处,例如书中使用了两种程序设计语言,一种是 C 语言,一种是 Prolog 语言,对于语音信号处理使用 C 语言,对于自动机技术、句法分析使用 Prolog 语言,在一本教材中同时使用两种编程语言,一定会增加读者的学习负担,此外 Prolog 语言并不是一种常用的程序设计语言,其程序设计理念也与 C 语言有较大差异,目前在语言处理领域中也很少使用,不如在语言处理技术部分继续使用 C 语言。

可能是由于作者来自语音处理领域的原因,语言处理技术部分的讲解似有欠缺,如对于 PCFG 的训练问题,书中指出存在有指导和无指导两种训练方法,并在举例时采用了无指导训练方法,但所给出例子有较强的误导性,因为例子所涉及的文法是没有歧义的特殊文法,针对自然语言而言,文法是无法没有歧义的,因此书中的训练方法对自然语言处理没有实际意义。再如,隐马尔可夫模型是马尔可夫模型的扩展,二者有联系但也有关键性的区别,书中似乎没有把二者之间的界限介绍清楚,相信会对读者造成一些阅读困扰。

全书中还有一些错误,本书的配套网站上刊登了一个勘误表,列出了一些错误,读者需要加以注意。不过显然该勘误表没有能罗列出所有错误,例如,书中很多 C 语言程序中存在“悬空指针赋值”的错误,具体而言,在许多程序中,都有下面代码片段:

```
int *length, ...;  
...  
x = signal_in(infile, length);
```

这里 length 被声明为一个整型指针,程序均没有令 length 指向预先分配的整型数据单元。这种做法是危险的,C 语言是弱类型检查语言,这种错误常常不能被编译器发现和报告,这会造成不可预期的运行错误。书中还有一些表述也不适当,在介绍 PCFG 的参数估计时,为了指出数据稀疏造成的问题,作者说明语料库中出现过的规则或结构的概率会被过高估计,这当然是正确的,但接下来,作者说明训练导致“语料库中所有句子的概率之和为 1,这相当于说(模型)不允许(语言中)出现其他(未在语料库中出现过)的句子”,这显然是不准确的,即使训练受制于所使用的训练语料,使得训练语料中出现过的规则和结构的概率偏高,但由于文法的递归性,通常也不可能导致训练语料中所有句子的概率和为 1。不过瑕不掩瑜,对语言语音处理领域的初学者而言,这是一本编写十分成功的教材,尤其是对于语言学专业的读者而言,也是值得大力推荐的一部教材。