TURING 图灵原版计算机科学系列

PEARSON

Speech and Language Processing

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

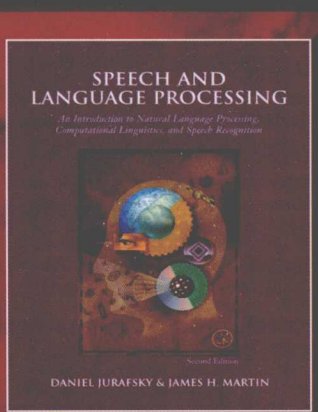
Second Edition

语音与语言处理

自然语言处理、计算语言学和语音识别导论

(英文版·第2版)

[美] Daniel Jurafsky
James H. Martin 著



人民邮电出版社
POSTS & TELECOM PRESS

Regular Expression Syntax

Expression	Description	Examples and Expansions
	Single character expressions	
`.`	any single character	`spi.e` matches "spice", "spike", etc.
`\char`	for a nonalphanumeric *char*, matches *char* literally	`\*` matches "*"
`\n`	newline character	
`\r`	carriage return character	
`\t`	tab character	
`[...]`	any single character listed in the brackets	`[abc]` matches "a", "b", or "c"
`[...-...]`	any single character in the range	`[0-9]` matches "0" or "1" ... or "9"
`[^...]`	any single character not listed	`[^sS]` matches one character that is neither "s" nor "S"
`[^...-...]`	any single character not in the range	`[^A-Z]` matches one character that is not an uppercase letter
	Anchors/Expressions which match positions	
`^`	beginning of line	
`$`	end of line	
`\b`	word boundary	`nt\b` matches "nt" in "paint" but not in "pants"
`\B`	word non-boundary	`all\B` matches "all" in "ally" but not in "wall"
	Counters/Expressions which quantify previous expressions	
`*`	zero or more of previous r.e.	`a*` matches "", "a", "aa","aaa",...
`+`	one or more of previous r.e.	`a+` matches "a", "aa", "aaa",...
`?`	exactly one or zero of previous r.e.	`colou?r` matches "color" or "colour"
`{n}`	*n* of previous r.e.	`a{4}` matches "aaaa"
`{n,m}`	from *n* to *m* of previous r.e.	
`{n,}`	at least *n* of previous r.e.	
`.*`	any string of characters	
`(...)`	grouping for precedence and memory for backreference	
`...|...`	matches either of neighbor r.e.s	`(dog)|(cat)` matches "dog" or "cat"
	Shortcuts	
`\d`	any digit	`[0-9]`
`\D`	any non-digit	`[^0-9]`
`\w`	any alphanumeric/underscore	`[a-zA-Z0-9_]`
`\W`	any non-alphanumeric	`[^a-zA-Z0-9_]`
`\s`	whitespace (space, tab)	`[␣\r\t\n\f]`
`\S`	non-whitespace	`[^␣\r\t\n\f]`

Penn Treebank Part-of-Speech Tags

Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	**SYM**	symbol	+,%, &
CD	cardinal number	*one, two, three*	**TO**	"to"	*to*
DT	determiner	*a, the*	**UH**	interjection	*ah, oops*
EX	existential "there"	*there*	**VB**	verb, base form	*eat*
FW	foreign word	*mea culpa*	**VBD**	verb, preterite (past tense)	*ate*
IN	preposition or subordinating conjunction	*of, in, by*	**VBG**	verb, gerund	*eating*
JJ	adjective	*yellow*	**VBN**	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	**VBP**	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	**VBZ**	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	**WDT**	wh-determiner	*which, that*
MD	modal	*can, should*	**WP**	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama, snow*	**WP$**	possessive wh-	*whose*
NNS	noun, plural	*llamas*	**WRB**	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	**$**	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	**#**	pound sign	#
PDT	predeterminer	*all, both*	**“**	left quote	‘ or “
POS	possessive ending	*’s*	**”**	right quote	’ or ”
PRP	personal pronoun	*I, you, he*	**(**	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one’s*	**)**	right parenthesis	],), }, >
RB	adverb	*quickly, never*	**,**	comma	,
RBR	adverb, comparative	*faster*	**.**	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	**:**	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

TURING 图灵原版计算机科学系列

Speech and Language Processing

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

Second Edition

语音与语言处理

自然语言处理、计算语言学和语音识别导论

(英文版·第2版)

[美] Daniel Jurafsky
James H. Martin 著

人民邮电出版社
北京

图书在版编目（CIP）数据

语音与语言处理 : 自然语言处理、计算语言学和语音识别导论 : 第2版 : 英文 / (美) 朱拉斯凯(Jurafsky,D.), (美) 马丁 (Martin,J.H.) 著. -- 北京 : 人民邮电出版社, 2010.12
(图灵原版计算机科学系列)
书名原文: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
ISBN 978-7-115-23892-4

Ⅰ. ①语… Ⅱ. ①朱… ②马… Ⅲ. ①自然语言处理—研究—英文②数理语言学—研究—英文③语音识别—研究—英文 Ⅳ. ①TP391②H087③TN912.3

中国版本图书馆CIP数据核字(2010)第189044号

内 容 提 要

本书全面系统地介绍了计算机自然语言处理。全书分为5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、语法、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。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。

图灵原版计算机科学系列

语音与语言处理：自然语言处理、计算语言学和语音识别导论（英文版·第2版）

◆ 著 [美] Daniel Jurafsky, James H. Martin
责任编辑 杨海玲

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子函件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京艺辉印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：64.25
字数：1234千字 2010年12月第1版
印数：1－2 000册 2010年12月北京第1次印刷

著作权合同登记号 图字：01-2010-2363号

ISBN 978-7-115-23892-4

定价：138.00元

读者服务热线：(010)51095186 印装质量热线：(010)67129223
反盗版热线：(010)67171154

版 权 声 明

For my parents, Ruth and Al Jurafsky — D.J.

For Linda and Katie — J.M.

序

语言学作为一个科学分支已经有 100 年的历史，计算语言学作为计算机科学的一部分也有了 50 年的历史。但是只是在过去的 10 年中，随着信息检索和机器翻译在因特网上投入使用，以及语音识别在桌面计算机上越来越流行，语言理解才作为一个行业惠及百万大众。语言信息的表示和处理方面的理论进步激活了这个行业。

本书是第一本从各个层面全面介绍语言技术的书，涉及了所有的现代技术。本书将深入的语言分析与健壮的统计方法结合起来。从层次的角度看，本书从单词及其组成成分开始，介绍了单词序列的属性及其表达和理解方式，然后介绍单词组合的方法（语法），表达含义的方法（语义），以及单词作为问答、对话和语言翻译的基础的方法。从技术的角度看，本书介绍了正则表达式、信息检索、上下文无关文法、合一、一阶谓词演算、隐马尔可夫及其他概率模型、修辞结构理论等。以前你可能需要通过两三本书才能看到所有这些内容，而本书却将所有这些内容融合在一本书中，把各种技术相互联系起来，让读者了解怎样才能最佳地利用每种技术，又怎样才能将各种技术结合起来使用。本书写作风格引人入胜，使人兴趣盎然，深入技术细节而又不让人感觉枯燥。无论从科研视角还是产业视角，本书都是对这一迷人领域开展进一步研究的理想导读、参考书和指南。

自从这本书的第 1 版于 2000 年出版以来，这一领域已经在很多方面有了长足发展，语言技术应用也更多更广。大的语言数据集（无论是书面的还是语音的）的出现已经使人们更加依赖统计机器学习方法。第 2 版涵盖了这些理论和实践的新发展。本书的组织方法使读者和指导教师更易于选择学习的内容——各章的依赖关系很小。尽管自上一版出版以来，语言处理领域已经有了一些出色的著作，但本书仍然是整个领域中最好的入门教程。

Peter Norvig，Stuart Russell

Prentice Hall 人工智能丛书主编

前　言

对于从事语音和语言处理的人来说，这是一个激动人心的时期。以往彼此不同的研究领域（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。基于 Web 的语言技术迅猛发展，基于电话的对话系统投入商用，还有语音合成和语音识别技术，一起为现实系统的开发提供了强劲的动力。由于可以使用大规模的联机语料库，所以在从发音到话语的各个不同层面都可以使用语言的统计模型。在筹划这本既可作为教材又可作为参考书的著作时，我们力图描绘出这种欣欣向荣的状态。

1. 覆盖全面

为了全面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系和不同课程中讲授的内容，例如，电子工程系的语音识别，计算机科学系的自然语言处理课程中的语法分析、语义解释及机器翻译，语言学系的计算语言学课程中的计算形态学、音系学和语用学等。本书介绍了这些领域中的基本算法，无论这些算法原来是为口语语言提出的还是为书面语言提出的，无论它们原来是从逻辑的角度提出的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的描述合在一起。我们也试图把一些机器翻译、拼写检查、信息检索和信息提取这样的应用以及认知建模领域的内容包括在本书中，使其覆盖更加全面。这种面面俱到的方法有一个潜在问题，即要求我们概述性介绍每个领域。因此，在阅读本书时，语言学家可以跳过与发音语音学相关的章节，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，尽管这本书篇幅很大，但仍不可能包罗万象。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、统计学和信息论等重要相关课程的各种专门著作。

2. 注重实际应用

我们认为，很重要的一点是要说明与语言相关的算法和技术（从隐马尔可夫模型到合一算法，从 λ 演算到对数线性模型）如何应用于重要的现实问题——语音识别、机器翻译、Web 的信息提取、拼写检查、文本文档搜索和口语对话。我们试图在每一章中都讲解一些语言处理应用的描述。这种方法的好处是，让学生有一点儿语言学的背景知识，便于理解特定领

域并开展建模。

3. 强调评测

近年来，在语言处理中统计算法越来越受到重视，语音和语言处理方面有组织的评测越来越多，因此本书特别加强了评测方面的内容。书中许多章都有专门的评测小节，具体描述评测系统和分析错误的现代经验方法，包括训练集和测试集、交叉确认以及像困惑度这样的信息论评测指标之类的概念。

4. 描述广泛可用的语言处理资源

现代的语音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、标准标注集等。我们力图在全书中介绍很多这样的重要资源（如 Brown、Switchboard、Fisher、CALLHOME、ATIS、TREC、MUC、BNC 等语料库），并且列出了很多有用的标注集以及编码方案（如 Penn Treebank、CLAWS 标注集、ARPAbet），不过难免会有遗漏。此外，本书中没有直接包括很多资源的 URL，而是把它们放在本书的网站上，这样即可得到及时的更新，网址是

http://www.cs.colorado.edu/~martin/slp.html

本书主要用于研究生或高年级本科生课程。因为本书的覆盖面广，并且有大量的算法，所以也可以作为语音和语言处理的各个领域中的大学生和专业人员的参考书。

全书概览

本书主要内容分为 5 个部分，此外还有第 1 章绪论和正文后的一些内容。第一部分是“词”，讲述与词及简单的词序列的处理有关的基本概念——单词分割、单词词法、单词编辑距离、语音部件，以及用于处理单词的算法——正则表达式、有限自动机、加权转录机、N 元文法、隐马尔可夫模型和对数线性模型。第二部分是“语音”，介绍英语语音学上的发音，然后是语音合成、语音识别和计算音韵学这些语言学主题。第三部分是“语法”，介绍英语的短语结构文法，并给出了处理单词之间结构化语法关系的基本算法：用于语法分析、统计语法分析的 CKY 算法和 Earley 算法，合一和分类的特征结构，以及 Chomsky 层次结构和泵引理（pumping lemma）等分析工具。第四部分是“语义和语用”，介绍含义表示的一阶逻辑以及其他方法、λ 演算、词法语义、词法语义资源（如 WordNet、PropBank 和 FrameNet）和用于单词相似度、单词歧义性、话语主题（如话语共指、话语连贯）词法语义的计算模型。第五部分是“应用”，讲解

信息提取、机器翻译以及对话和会话代理。

本书用法

本书内容丰富，可作为两个学期的语音与语言处理课程的教材。本书也可以作为各种不同用途的一学期课程的教材使用。

自然语言处理 一学季	自然语言处理 一学期	语音 + 自然语言处理 一学期	计算机语言学 一学季
第 1 章	第 1 章	第 1 章	第 1 章
第 2 章	第 2 章	第 2 章	第 2 章
第 4 章	第 4 章	第 4 章	第 3 章
第 5 章	第 5 章	第 5 章	第 4 章
第 12 章	第 6 章	第 6 章	第 5 章
第 13 章	第 12 章	第 8 章	第 13 章
第 14 章	第 13 章	第 9 章	第 14 章
第 19 章	第 14 章	第 12 章	第 15 章
第 20 章	第 17 章	第 13 章	第 16 章
第 23 章	第 18 章	第 14 章	第 20 章
第 25 章	第 19 章	第 17 章	第 21 章
	第 20 章	第 19 章	
	第 21 章	第 20 章	
	第 22 章	第 22 章	
	第 23 章	第 24 章	
	第 25 章	第 25 章	

还可以节选书中某些章节增加人工智能、认知科学、信息检索或面向电子工程的语音处理课程的内容。

致谢

Andy Kehler 编写了第 1 版的 Discourse 一章，我们把它作为第 2 版对应章的起点。Andy 的许多陈述和结构仍然在这一章随处可见。类似地，Nigel Ward 写了第 1 版的 Machine Translation 一章的大部分内容，我们把它作为第 2 版对应章的起点，他的许多文字仍然保留，特别是在 25.2 节、25.3 节和练习中。Kevin Bretonnel Cohen 写了关于生物信息提取的 22.5 节，Keith Vander Linden 写了第 1 版的 Generation 一章，冯志伟和孙乐将本书的第 1 版翻译为中文，Gideon Mann 和 Richard Wicentowski 为第 2 版的国际版编写了不少习题。

科罗拉多大学博尔德分校和斯坦福大学对我们在语音和语言处理方面的工作提供了很好的待遇，我们要感谢我们所在的系、系里的同事以及我们的学生的协助，这对我们的研究和教学都有很大的影响。

Dan 想感谢他的父母一直激励他做正确的事，并赶早不赶晚。他还要感谢 Nelson Morgan 将它引入语音识别领域，并教会他多问“这样是否可行”；感谢 Jerry Feldman 影响他对正确结果孜孜以求，并教会他多思考“这是否真的重要”；感谢他的第一位指导老师 Chuck Fillmore 引导他爱上语言，并教他多看数据；感谢他的论文指导老师 Robert Wilensky 让他明白协作和团队精神在研究领域的重要性；感谢 Chris Manning 作为在斯坦福大学极好的合作者，当然还要感谢他在科罗拉多大学博尔德分校的所有好同事。

Jim 想要感谢他的父母鼓励并允许他走当时看起来有点奇怪的道路。他还要感谢他的论文指导老师 Robert Wilensky 在伯克利带他进入自然语言处理的大门，感谢 Peter Norvig 给了他很多正面引导的例子，感谢 Rick Alterman 在关键时刻的鼓励和启发，感谢 Chuck Fillmore、George Lakoff、Paul Kay 和 Susanna Cumming 让他明白自己对语言学的了解还太少，感谢 Martha Palmer、Tammy Sumner 和 Wayne Ward 在科罗拉多大学博尔德分校的成功合作。最后，Jim 要感谢妻子 Linda 这么多年以来的支持和耐心，感谢女儿 Katie 全心等待这一版的完成。

我们非常感激很多人对本书第 1 版给予的极大帮助，第 2 版也得益于很多读者和他们试用本教材。特别感谢如下几位人士全面研读本书并提出了很有帮助的意见，他们是 Regina Barzilay、Philip Resnik、Emily Bender 和 Adam Przepiórkowski。本书的编辑 Tracy Dunkelberger、编辑部主任 Scott Disanno、丛书主编 Peter Norvig 和 Stuart Russell、制作编辑 Jane Bonnell 对本书的内容和设计给出了很多建议。我们还要感谢很多朋友和同事，他们分别阅读了书的某些章节并回答了我们的很多问题，包括在科罗拉多大学、斯坦福大学和伊利诺伊大学厄巴纳-香槟分校（1999）、MIT（2005）和斯坦福大学（2007）LSA 暑期学院上我们的课的学生，以 及 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 Chambers、Pichuan Chang、Grace Chung、Andrew Clausen、Raphael Cohn、Kevin B. Cohen、Frederik

Coppens、Stephen Cox、Heriberto Cuayáhuitl、Martin Davidsson、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、Manuel Kirschner、Dan Klein Sheldon Klein、Kevin Knight、Jean-Pierre Koenig、Greg Kondrak、Selcuk Kopru、Kimmo Koskenniemi、Alexander Kostyrkin、Mikko Kurimo、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、Andreas Stolcke、Niyue Tan、Frank Yung-Fong Tang、Ahmet Cüneyd Tantuğ、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。

我们还要感谢允许我使用图 7.3（©Laszlo Kubinyi 和《科学美国人》）和图 9.14（©Paul Taylor 和剑桥大学出版社）的授权人。此外，我们还在一些原图的基础上创建了我们的很多图，我们要感谢下面这些人允许我们改编他们的图。以下三张图的版权是 IEEE 和作者的，感谢 Esther Levin（图 24.22）和 Lawrence Rabiner（图 6.14 和图 6.15）；其他改编的图版权是作者的，感谢计算语言学学会（Association of Computational Linguistics）、《计算语言学期刊》（*Journal of Computational Linguistics*）及其编辑 Robert Dale，还要感谢 Regina Barzilay（图 23.19）、Michael Collins（图 14.7、图 14.10 和图 14.11）、John Goldsmith（图 11.18）、Marti Hearst（图 21.1 和图 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