HANDBOOK OF

NATURAL LANGUAGE PROCESSING

SECOND EDITION

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

SERIES EDITORS

Ralf Herbrich and Thore Graepel
Microsoft Research Ltd.
Cambridge, UK

AIMS AND SCOPE

This series reflects the latest advances and applications in machine learning and pattern recognition through the publication of a broad range of reference works, textbooks, and handbooks. The inclusion of concrete examples, applications, and methods is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of machine learning, pattern recognition, computational intelligence, robotics, computational/statistical learning theory, natural language processing, computer vision, game AI, game theory, neural networks, computational neuroscience, and other relevant topics, such as machine learning applied to bioinformatics or cognitive science, which might be proposed by potential contributors.

PUBLISHED TITLES

MACHINE LEARNING: An Algorithmic Perspective
Stephen Marsland

HANDBOOK OF NATURAL LANGUAGE PROCESSING,
Second Edition
Nitin Indurkhya and Fred J. Damerau

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

HANDBOOK OF NATURAL LANGUAGE PROCESSING

SECOND EDITION

Edited by
NITIN INDURKHYA
FRED J. DAMERAU

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor and Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-8593-8 (Ebook-PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Fred Damerau
born December 25, 1931; died January 27, 2009

Some enduring publications:

Damerau, F. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM* 7, 3 (Mar. 1964), 171–176.

Damerau, F. 1971. *Markov Models and Linguistic Theory: An Experimental Study of a Model for English*. The Hague, the Netherlands: Mouton.

Damerau, F. 1985. Problems and some solutions in customization of natural language database front ends. *ACM Trans. Inf. Syst.* 3, 2 (Apr. 1985), 165–184.

Apté, C., Damerau, F., and Weiss, S. 1994. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst.* 12, 3 (Jul. 1994), 233–251.

Weiss, S., Indurkhya, N., Zhang, T., and Damerau, F. 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer.

Contents

List of Figures ix
List of Tables xiii
Editors xv
Board of Reviewers xvii
Contributors xix
Preface xxi

PART I Classical Approaches

1 Classical Approaches to Natural Language Processing *Robert Dale* 3
2 Text Preprocessing *David D. Palmer* 9
3 Lexical Analysis *Andrew Hippisley* 31
4 Syntactic Parsing *Peter Ljunglöf and Mats Wirén* 59
5 Semantic Analysis *Cliff Goddard and Andrea C. Schalley* 93
6 Natural Language Generation *David D. McDonald* 121

PART II Empirical and Statistical Approaches

7 Corpus Creation *Richard Xiao* 147
8 Treebank Annotation *Eva Hajičová, Anne Abeillé, Jan Hajič, Jiří Mírovský, and Zdeňka Urešová* 167
9 Fundamental Statistical Techniques *Tong Zhang* 189
10 Part-of-Speech Tagging *Tunga Güngör* 205
11 Statistical Parsing *Joakim Nivre* 237
12 Multiword Expressions *Timothy Baldwin and Su Nam Kim* 267

13 Normalized Web Distance and Word Similarity *Paul M.B. Vitányi and Rudi L. Cilibrasi* 293

14 Word Sense Disambiguation *David Yarowsky* 315

15 An Overview of Modern Speech Recognition *Xuedong Huang and Li Deng* 339

16 Alignment *Dekai Wu* 367

17 Statistical Machine Translation *Abraham Ittycheriah* 409

PART III Applications

18 Chinese Machine Translation *Pascale Fung* 425

19 Information Retrieval *Jacques Savoy and Eric Gaussier* 455

20 Question Answering *Diego Mollá-Aliod and José-Luis Vicedo* 485

21 Information Extraction *Jerry R. Hobbs and Ellen Riloff* 511

22 Report Generation *Leo Wanner* 533

23 Emerging Applications of Natural Language Generation in Information Visualization, Education, and Health Care *Barbara Di Eugenio and Nancy L. Green* 557

24 Ontology Construction *Philipp Cimiano, Johanna Völker, and Paul Buitelaar* 577

25 BioNLP: Biomedical Text Mining *K. Bretonnel Cohen* 605

26 Sentiment Analysis and Subjectivity *Bing Liu* 627

Index 667

List of Figures

Figure 1.1 The stages of analysis in processing natural language 4
Figure 3.1 A spelling rule FST for *glasses* 34
Figure 3.2 A spelling rule FST for *flies* 34
Figure 3.3 An FST with symbol classes 36
Figure 3.4 Russian nouns classes as an inheritance hierarchy 48
Figure 4.1 Example grammar 62
Figure 4.2 Syntax tree of the sentence "the old man a ship" 62
Figure 4.3 CKY matrix after parsing the sentence "the old man a ship" 67
Figure 4.4 Final chart after bottom-up parsing of the sentence "the old man a ship." The dotted edges are inferred but useless 71
Figure 4.5 Final chart after top-down parsing of the sentence "the old man a ship." The dotted edges are inferred but useless 71
Figure 4.6 Example LR(0) table for the grammar in Figure 4.1 74
Figure 5.1 The lexical representation for the English verb *build* 99
Figure 5.2 UER diagrammatic modeling for transitive verb *wake up* 103
Figure 8.1 A scheme of annotation types (layers) 169
Figure 8.2 Example of a Penn treebank sentence 171
Figure 8.3 A simplified constituency-based tree structure for the sentence *John wants to eat cakes* 171
Figure 8.4 A simplified dependency-based tree structure for the sentence *John wants to eat cakes* 171
Figure 8.5 A sample tree from the PDT for the sentence: *Česká opozice se nijak netají tím, že pokud se dostane k moci, nebude se deficitnímu rozpočtu nijak bránit.* (lit.: Czech opposition Refl. in-no-way keeps-back the-fact that in-so-far-as [it] will-come into power, [it] will-not Refl. deficit budget in-no-way oppose. English translation: *The Czech opposition does not keep back that if they come into power, they will not oppose the deficit budget.*) 173
Figure 8.6 A sample French tree. (English translation: *It is understood that the public functions remain open to all the citizens.*) 174
Figure 8.7 Example from the Tiger corpus: complex syntactic and semantic dependency annotation. (English translation: *It develops and prints packaging materials and labels.*) 174
Figure 9.1 Effect of regularization 191
Figure 9.2 Margin and linear separating hyperplane 192
Figure 9.3 Multi-class linear classifier decision boundary 193
Figure 9.4 Graphical representation of generative model 195
Figure 9.5 Graphical representation of discriminative model 196

Figure 9.6 EM algorithm 199
Figure 9.7 Gaussian mixture model with two mixture components 200
Figure 9.8 Graphical representation of HMM 200
Figure 9.9 Viterbi algorithm 202
Figure 9.10 Graphical representation of discriminative local sequence prediction model 202
Figure 9.11 Graphical representation of discriminative global sequence prediction model 203
Figure 10.1 Transformation-based learning algorithm 210
Figure 10.2 A part of an example HMM for the specialized word *that* 218
Figure 10.3 Morpheme structure of the sentence *na-neun hag-gyo-e gan-da* 227
Figure 11.1 Constituent structure for an English sentence taken from the Penn Treebank 239
Figure 11.2 Dependency structure for an English sentence taken from the Penn Treebank 239
Figure 11.3 PCFG for a fragment of English 241
Figure 11.4 Alternative constituent structure for an English sentence taken from the Penn Treebank (cf. Figure 11.1) 243
Figure 11.5 Constituent structure with parent annotation (cf. Figure 11.1) 247
Figure 12.1 A classification of MWEs 279
Figure 13.1 Colors, numbers, and other terms arranged into a tree based on the NWDs between the terms 305
Figure 13.2 Hierarchical clustering of authors 306
Figure 13.3 Distance matrix of pairwise NWD's 306
Figure 13.4 Names of several Chinese people, political parties, regions, and others. The nodes and *solid lines* constitute a tree constructed by a hierarchical clustering method based on the NWDs between all names. The numbers at the perimeter of the tree represent NWD values between the nodes pointed to by the *dotted lines*. For an explanation of the names, refer to Figure 13.5 307
Figure 13.5 Explanations of the Chinese names used in the experiment that produced Figure 13.4 308
Figure 13.6 NWD–SVM learning of "emergencies" 309
Figure 13.7 Histogram of accuracies over 100 trials of WordNet experiment 311
Figure 14.1 Iterative bootstrapping from two seed words for *plant* 330
Figure 14.2 Illustration of vector clustering and sense partitioning 332
Figure 15.1 A source-channel model for a typical speech-recognition system 340
Figure 15.2 Basic system architecture of a speech-recognition system 340
Figure 15.3 Illustration of a five-state left-to-right HMM. It has two non-emitting states and three emitting states. For each emitting state, the HMM is only allowed to remain at the same state or move to the next state 342
Figure 15.4 An illustration of how to compile a speech-recognition task with finite grammar into a composite HMM 348
Figure 15.5 An simple RTN example with three types of arcs: CAT(x), PUSH(x), and POP 349
Figure 15.6 Ford SYNC highlights the car's speech interface—"You talk. SYNC listens" 353
Figure 15.7 Bing Search highlights speech functions—just say what you're looking for! 353
Figure 15.8 Microsoft's Tellme is integrated into Windows Mobile at the network level 355
Figure 15.9 Microsoft's Response Point phone system designed specifically for small business customers 355
Figure 16.1 Alignment examples at various granularities 369
Figure 16.2 Anchors 374
Figure 16.3 Slack bisegments between anchors 374
Figure 16.4 Banding the slack bisegments using variance 375
Figure 16.5 Banding the slack bisegments using width thresholds 376
Figure 16.6 Guiding based on a previous rough alignment 377

Figure 16.7 Example sentence lengths in an input bitext 378
Figure 16.8 Equivalent stochastic or weighted (a) FST and (b) FSTG notations for a finite-state bisegment generation process. Note that the node transition probability distributions are often tied to be the same for all node/nonterminal types 380
Figure 16.9 Multitoken lexemes of length *m* and *n* must be coupled in awkward ways if constrained to using only 1-to-1 single-token bisegments 390
Figure 16.10 The crossing constraint 391
Figure 16.11 The 24 complete alignments of length four, with ITG parses for 22. All nonterminal and terminal labels are omitted. A horizontal bar under a parse tree node indicates an inverted rule 397
Figure 17.1 Machine translation transfer pyramid 410
Figure 17.2 Arabic parse example 412
Figure 17.3 Source and target parse trees 416
Figure 18.1 Marking word segment, POS, and chunk tags for a sentence 429
Figure 18.2 An example of Chinese word segmentation from alignment of Chinese characters to English words 429
Figure 18.3 (a) A simple transduction grammar and (b) an inverted orientation production 433
Figure 18.4 ITG parse tree 433
Figure 18.5 Example grammar rule extracted with ranks 435
Figure 18.6 Example, showing translations after SMT first pass and after reordering second pass 441
Figure 18.7 DK-vec signals showing similarity between *Government* in English and Chinese, contrasting with *Bill* and *President* 444
Figure 18.8 Parallel sentence and bilingual lexicon extraction from quasi-comparable corpora 446
Figure 20.1 Generic QA system architecture 489
Figure 20.2 Number of TREC 2003 questions that have an answer in the preselected documents using TREC's search engine. The total number of questions was 413 492
Figure 20.3 Method of resolution 496
Figure 20.4 An example of abduction for QA 496
Figure 21.1 Example of an unstructured seminar announcement 514
Figure 21.2 Examples of semi-structured seminar announcements 514
Figure 21.3 Chronology of MUC system performance 526
Figure 22.1 Sample of the input as used by SumTime-Turbine. TTXD-*i* are temperatures (in °C) measured by the exhaust thermocouple *i* 537
Figure 22.2 Lexicalized input structure to Streak 542
Figure 22.3 Sample semantic structure for the message 'The air quality index is 6, which means that the air quality is very poor' produced by MARQUIS as input to the MTT-based linguistic module 542
Figure 22.4 Input and output of the Gossip system 545
Figure 22.5 English report generated by FoG 546
Figure 22.6 LFS output text fragment 546
Figure 22.7 Input and output of PLANDOC 547
Figure 22.8 Summaries generated by ARNS, a report generator based on the same technology as the Project Reporter 548
Figure 22.9 Report generated by SumTime-Mousam 548
Figure 22.10 Medical report as generated by MIAKT 549
Figure 22.11 English AQ report as produced by MARQUIS 549
Figure 22.12 An example summary generated by SumTime-Turbine, from the data set displayed in Figure 22.1 550
Figure 22.13 A BT-45 text 550