

2107  
1000

Foundations  
of  
Statistical  
Natural  
Language  
Processing



E0123734

Christopher D. Manning  
Hinrich Schütze



The MIT Press  
Cambridge, Massachusetts  
London, England

送书

Second printing, 1999  
© 1999 Massachusetts Institute of Technology  
Second printing with corrections, 2000

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

Typeset in 10/13Lucida Bright by the authors using  $\text{\LaTeX} 2_{\epsilon}$ .  
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Information

Manning, Christopher D.  
Foundations of statistical natural language processing / Christopher D.  
Manning, Hinrich Schutze.  
p. cm.  
Includes bibliographical references (p. ) and index.  
ISBN 0-262-13360-1  
1. Computational linguistics-Statistical methods. I. Schutze, Hinrich.  
II. Title.  
P98.5.S83M36 1999  
410'.285—dc21

99-21137  
CIP

1151  
专款图书

# ***Brief Contents***

I Preliminaries	1
1 <i>Introduction</i>	3
2 <i>Mathematical Foundations</i>	39
3 <i>Linguistic Essentials</i>	81
4 <i>Corpus-Based Work</i>	117
II Words	149
5 <i>Collocations</i>	151
6 <i>Statistical Inference: n-gram Models over Sparse Data</i>	191
7 <i>Word Sense Disambiguation</i>	229
8 <i>Lexical Acquisition</i>	265
III Grammar	315
9 <i>Markov Models</i>	317
10 <i>Part-of-Speech Tagging</i>	341
11 <i>Probabilistic Context Free Grammars</i>	381
12 <i>Probabilistic Parsing</i>	407
IV Applications and Techniques	461
13 <i>Statistical Alignment and Machine Translation</i>	463
14 <i>Clustering</i>	495
15 <i>Topics in Information Retrieval</i>	529
16 <i>Text Categorization</i>	575

# *Contents*

**List of Tables** xv

**List of Figures** xxi

**Table of Notations** xxv

**Preface** xxix

**Road Map** xxxv

## **I Preliminaries 1**

### **1 Introduction 3**

- 1.1 Rationalist and Empiricist Approaches to Language 4
- 1.2 Scientific Content 7
  - 1.2.1 Questions that linguistics should answer 8
  - 1.2.2 Non-categorical phenomena in language 11
  - 1.2.3 Language and cognition as probabilistic phenomena 15
- 1.3 The Ambiguity of Language: Why NLP Is Difficult 17
- 1.4 Dirty Hands 19
  - 1.4.1 Lexical resources 19
  - 1.4.2 Word counts 20
  - 1.4.3 Zipf's laws 23
  - 1.4.4 Collocations 29
  - 1.4.5 Concordances 31
- 1.5 Further Reading 34

1.6	Exercises	35
2	<b>Mathematical Foundations</b>	<b>39</b>
2.1	Elementary Probability Theory	40
2.1.1	Probability spaces	40
2.1.2	Conditional probability and independence	42
2.1.3	Bayes' theorem	43
2.1.4	Random variables	45
2.1.5	Expectation and variance	46
2.1.6	Notation	47
2.1.7	Joint and conditional distributions	48
2.1.8	Determining $P$	48
2.1.9	Standard distributions	50
2.1.10	Bayesian statistics	54
2.1.11	Exercises	59
2.2	Essential Information Theory	60
2.2.1	Entropy	61
2.2.2	Joint entropy and conditional entropy	63
2.2.3	Mutual information	66
2.2.4	The noisy channel model	68
2.2.5	Relative entropy or Kullback-Leibler divergence	72
2.2.6	The relation to language: Cross entropy	73
2.2.7	The entropy of English	76
2.2.8	Perplexity	78
2.2.9	Exercises	78
2.3	Further Reading	79
3	<b>Linguistic Essentials</b>	<b>81</b>
3.1	Parts of Speech and Morphology	81
3.1.1	Nouns <del>pronouns</del>	83
3.1.2	Words that accompany nouns: Determiners and adjectives	87
3.1.3	Verbs	88
3.1.4	Other parts of speech	91
3.2	Phrase Structure	93
3.2.1	Phrase <del>structures</del>	96
3.2.2	Dependency: Arguments and adjuncts	101
3.2.3	X' theory	106
3.2.4	Phrase structure ambiguity	107

3.3	Semantics and Pragmatics	109
3.4	Other Areas	112
3.5	Further Reading	113
3.6	Exercises	114
4	<b>Corpus-Based Work</b>	117
4.1	Getting Set Up	118
4.1.1	Computers	118
4.1.2	Corpora	118
4.1.3	Software	120
4.2	Looking at Text	123
4.2.1	Low-level formatting issues	123
4.2.2	Tokenization: What is a word?	124
4.2.3	Morphology	131
4.2.4	Sentences	134
4.3	Marked-up Data	136
4.3.1	Markup schemes	137
4.3.2	Grammatical tagging	139
<b>4.4</b>	Further Reading	145
4.5	Exercises	147
<b>II</b>	<b>Words</b>	<b>149</b>
5	<b>Collocations</b>	151
5.1	Frequency	153
5.2	Mean and Variance	157
5.3	Hypothesis Testing	162
5.3.1	The $t$ test	163
5.3.2	Hypothesis testing of differences	166
5.3.3	Pearson's chi-square test	169
5.3.4	Likelihood ratios	172
5.4	Mutual Information	178
5.5	The Notion of Collocation	183
5.6	Further Reading	187
6	<b>Statistical Inference: <math>n</math>-gram Models over Sparse Data</b>	191
6.1	Bins: Forming Equivalence Classes	192
6.1.1	Reliability vs. discrimination	192
6.1.2	$n$ -gram models	192

6.1.3	Building-gram models	195
6.2	Statistical Estimators	196
6.2.1	Maximum Likelihood Estimation (MLE)	197
6.2.2	Laplace's law, Lidstone's law and the Jeffreys-Perks law	202
6.2.3	Held out estimation	205
6.2.4	Cross-validation (deleted estimation)	210
6.2.5	Good-Turing estimation	212
6.2.6	Briefly noted	216
6.3	Combining Estimators	217
6.3.1	Simple linear interpolation	218
6.3.2	Katz's backing-off	219
6.3.3	General linear interpolation	220
6.3.4	Briefly noted	222
6.3.5	Language models for Austen	223
6.4	Conclusions	224
6.5	Further Reading	225
6.6	Exercises	225
<b>7</b>	<b><i>Word Sense Disambiguation</i></b>	<b>229</b>
<b>7.1</b>	<b>Methodological Preliminaries</b>	<b>232</b>
7.1.1	Supervised and unsupervised learning	232
7.1.2	Pseudowords	233
7.1.3	Upper and lower bounds on performance	233
7.2	Supervised Disambiguation	235
7.2.1	Bayesian classification	235
7.2.2	An information-theoretic approach	239
7.3	Dictionary-Based Disambiguation	241
7.3.1	Disambiguation based on sense definitions	242
7.3.2	Thesaurus-based disambiguation	244
7.3.3	Disambiguation based on translations in a second-language corpus	247
7.3.4	One sense per discourse, one sense per collocation	249
7.4	Unsupervised Disambiguation	252
7.5	What Is a Word Sense?	256
7.6	Further Reading	260
7.7	Exercises	262

<b>8</b>	<b>Lexical Acquisition</b>	<b>265</b>
8.1	Evaluation Measures	267
8.2	Verb Subcategorization	271
8.3	Attachment Ambiguity	278
8.3.1	Hindle and Rooth (1993)	280
8.3.2	General remarks on PP attachment	284
8.4	Selectional Preferences	288
8.5	Semantic Similarity	294
8.5.1	Vector space measures	296
8.5.2	Probabilistic measures	303
8.6	The Role of Lexical Acquisition in Statistical NLP	308
8.7	Further Reading	312
<b>III</b>	<b>Grammar</b>	<b>315</b>
<b>9</b>	<b>Markov Models</b>	<b>317</b>
9.1	Markov Models	318
9.2	Hidden Markov Models	320
9.2.1	Why use HMMs?	322
9.2.2	General form of an HMM	324
9.3	The Three Fundamental Questions for HMMs	325
9.3.1	Finding the probability of an observation	326
9.3.2	Finding the best state sequence	331
9.3.3	The third problem: Parameter estimation	333
9.4	HMMs: Implementation, Properties, and Variants	336
9.4.1	Implementation	336
9.4.2	Variants	337
9.4.3	Multiple input observations	338
9.4.4	Initialization of parameter values	339
9.5	Further Reading	339
<b>10</b>	<b><i>Part-of-Speech</i> Tagging</b>	<b>341</b>
10.1	The Information Sources in Tagging	343
10.2	Markov Model Taggers	345
10.2.1	The probabilistic model	345
10.2.2	The Viterbi algorithm	349
10.2.3	Variations	351
10.3	Hidden Markov Model Taggers	356



10.3.1	Applying HMMs to POS tagging	357	
10.32	The effect of initialization on HMM training		359
10.4	Transformation-Based Learning of Tags	361	
10.4.1	Transformations	362	
10.4.2	The learning algorithm	364	
10.4.3	Relation to other models	365	
10.4.4	Automata	367	
10.4.5	Summary	369	
10.5	Other Methods, Other Languages	370	
10.5.1	Other approaches to tagging	370	
10.5.2	Languages other than English	371	
10.6	Tagging Accuracy and Uses of Taggers		371
10.6.1	Tagging accuracy	371	
10.6.2	Applications of tagging	374	
10.7	Further Reading	377	
10.8	Exercises	379	
<b>11</b>	<b><i>Probabilistic Context Free Grammars</i></b>		<b>381</b>
11.1	Some Features of PCFGs	<b>386</b>	
11.2	Questions for PCFGs	388	
11.3	The Probability of a String	392	
11.3.1	Using inside probabilities	392	
11.3.2	Using outside probabilities	394	
11.3.3	Finding the most likely parse for a sentence		396
11.3.4	Training a PCFG	<b>398</b>	
11.4	Problems with the Inside-Outside Algorithm		401
11.5	Further Reading	402	
11.6	Exercises	404	
<b>12</b>	<b><i>Probabilistic Parsing</i></b>		<b>407</b>
<b>12.1</b>	<b>Some Concepts</b>	408	
12.1.1	Parsing for disambiguation	408	
12.1.2	Treebanks	412	
12.1.3	Parsing models vs. language models	414	
12.1.4	Weakening the independence assumptions of PCFGs	416	
12.1.5	Tree probabilities and derivational probabilities		421
12.1.6	There's more than one way to do it	423	

- 121.7 Phrase structure grammars and dependency grammars 428
- 12.1.8 Evaluation 431
- 12.1.9 Equivalent models 437
- 12.1.10 Building search methods 439
- 12.1.11 Use of the geometric mean 442
- 12.2 Some Approaches 443
  - 12.2.1 Non-lexicalized grammars 443
  - 12.2.2 Lexicalized models using derivational histories 448
  - 12.2.3 Dependency-based models 451
  - 12.2.4 Discussion 454
- 12.3 Further Reading 456
- 12.4 Exercises 458

## IV Applications and Techniques 461

### 13 *Statistical Alignment and Machine Translation* 463

- 13.1 Text Alignment 466
  - 13.1.1 Aligning sentences and paragraphs 467
  - 13.1.2 Length-based methods 471
  - 13.1.3 Offset alignment by signal processing techniques 475
  - 13.1.4 Lexical methods of sentence alignment 478
  - 13.1.5 Summary 484
  - 13.1.6 Exercises 484
- 13.2 Word Alignment 484
- 13.3 Statistical Machine Translation 486
- 13.4 Further Reading 492

### 14 *Clustering* 495

- 14.1 Hierarchical Clustering 500
  - 14.1.1 Single-link and complete-link clustering 503
  - 14.1.2 Group-average agglomerative clustering 507
  - 14.1.3 An application: Improving a language model 509
  - 14.1.4 Top-down clustering 512
- 14.2 Non-Hierarchical Clustering 514
  - 14.2.1 K-means 515
  - 14.2.2 The EM algorithm 518
- 14.3 Further Reading 527