

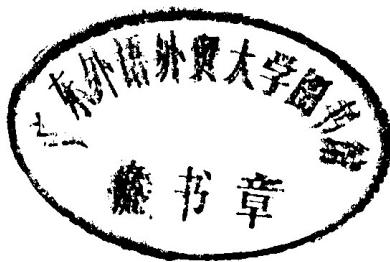
2017  
10.10

# Foundations of Statistical Natural Language Processing



E0123734

Christopher D. Manning  
Hinrich Schütze



The MIT Press  
Cambridge, Massachusetts  
London, England

清华大学图书馆

---

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 10/13 Lucida Bright by the authors using L<sup>A</sup>T<sub>E</sub>X 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

U.S.I  
专款图书

# *Brief Contents*

## I Preliminaries 1

- 1 *Introduction* 3
- 2 *Mathematical Foundations* 39
- 3 *Linguistic Essentials* 81
- 4 *Corpus-Based Work* 117

## II Words 149

- 5 *Collocations* 151
- 6 *Statistical Inference: n-gram Models over Sparse Data* 191
- 7 *Word Sense Disambiguation* 229
- 8 *Lexical Acquisition* 265

## III Grammar 315

- 9 *Markov Models* 317
- 10 *Part-of-Speech Tagging* 341
- 11 *Probabilistic Context Free Grammars* 381
- 12 *Probabilistic Parsing* 407

## IV Applications and Techniques 461

- 13 *Statistical Alignment and Machine Translation* 463
- 14 *Clustering* 495
- 15 *Topics in Information Retrieval* 529
- 16 *Text Categorization* 575

# ***Contents***

<i>List of Tables</i>	<i>xv</i>
<i>List of Figures</i>	<i>xxi</i>
<i>Table of Notations</i>	<i>xxv</i>
<i>Preface</i>	<b><i>xxix</i></b>
<i>Road Map</i>	<b><i>xxxv</i></b>

## **I Preliminaries 1**

1 <i>Introduction</i>	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	<i>Mathematical Foundations</i>	39
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	<i>Linguistic Essentials</i>	81
3.1	Parts of Speech and Morphology	81
3.1.1	Nouns	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 structures	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
<b>4 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 Further Reading</b>	145
4.5 Exercises	147
<b>II Words 149</b>	
<b>5 Collocations</b>	151
5.1 Frequency	153
5.2 Mean and Variance	157
5.3 Hypothesis Testing	162
5.3.1 The <i>t</i> 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
<b>6 Statistical Inference: <i>n</i>-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-grammodels	192

6.1.3	Building n-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
7	<b><i>Word Sense Disambiguation</i></b>	<b>229</b>
7.1	Methodological Preliminaries	232
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 Lexical Acquisition</b>	<b>265</b>
<b>8.1 Evaluation Measures</b>	<b>267</b>
<b>8.2 Verb Subcategorization</b>	<b>271</b>
<b>8.3 Attachment Ambiguity</b>	<b>278</b>
8.3.1 Hindle and Rooth (1993)	280
8.3.2 General remarks on PP attachment	284
<b>8.4 Selectional Preferences</b>	<b>288</b>
<b>8.5 Semantic Similarity</b>	<b>294</b>
8.5.1 Vectorspace measures	296
8.5.2 Probabilistic measures	303
<b>8.6 The Role of Lexical Acquisition in Statistical NLP</b>	<b>308</b>
<b>8.7 Further Reading</b>	<b>312</b>

### **III Grammar      315**

<b>9 Markov Models</b>	<b>317</b>
<b>9.1 Markov Models</b>	<b>318</b>
<b>9.2 Hidden Markov Models</b>	<b>320</b>
9.2.1 Why use HMMs?	322
9.2.2 General form of an HMM	324
<b>9.3 The Three Fundamental Questions for HMMs</b>	<b>325</b>
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
<b>9.4 HMMs: Implementation, Properties, and Variants</b>	<b>336</b>
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
<b>9.5 Further Reading</b>	<b>339</b>

### **10 Part-of-Speech Tagging** **341**

<b>10.1 The Information Sources in Tagging</b>	<b>343</b>
<b>10.2 Markov Model Taggers</b>	<b>345</b>
10.2.1 The probabilistic model	345
10.2.2 The Viterbi algorithm	349
10.2.3 Variations	351
<b>10.3 Hidden Markov Model Taggers</b>	<b>356</b>

---

10.3.1 Applying HMMs to POS tagging	357
10.3.2 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 Probabilistic Context Free Grammars</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 Probabilistic Parsing</b>	<b>407</b>
<b>12.1</b> Some Concepts	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

12.1.7	Phrase structure grammars and dependency grammars	428
12.1.8	Evaluation	431
12.1.9	Equivalent models	437
12.1.10	Buil <del>fmbers</del> Search methods	439
12.1.11	Use of the geometric mean	442
12.2	Some Approaches	443
12.2.1	Non-lexicalized grammar	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

14.4 Exercises	528
<b>15 Topics in Information Retrieval</b>	<b>529</b>
15.1 Some Background on Information Retrieval	530
15.1.1 Common design features of IR systems	532
15.1.2 Evaluation measures	534
15.1.3 The probability ranking principle (PRP)	538
15.2 The Vector Space Model	539
15.2.1 Vector similarity	540
15.2.2 Term weighting	541
15.3 Term Distribution Models	544
15.3.1 The Poisson distribution	545
15.3.2 The two-Poisson model	548
15.3.3 The K mixture	549
15.3.4 Inverse document frequency	551
15.3.5 Residual inverse document frequency	553
15.3.6 Usage of term distribution models	554
15.4 Latent Semantic Indexing	554
15.4.1 Least-squares methods	557
15.4.2 Singular Value Decomposition	558
15.4.3 Latent Semantic Indexing in IR	564
15.5 Discourse Segmentation	566
15.5.1 TextTiling	567
15.6 Further Reading	570
15.7 Exercises	573
<b>16 Text Categorization</b>	<b>575</b>
16.1 Decision Trees	578
16.2 Maximum Entropy Modeling	589
16.2.1 Generalized iterative scaling	591
16.2.2 Application to text categorization	594
16.3 Perceptrons	597
16.4 k Nearest Neighbor Classification	604
16.5 Further Reading	607
<b>Tiny Statistical Tables</b>	<b>609</b>
<b>Bibliography</b>	<b>611</b>
<b>Index</b>	<b>657</b>

## *List of Tables*

1.1	Common words in <i>Tom Sawyer</i> .	21
1.2	Frequency of frequencies of word types in <i>Tom Sawyer</i> .	22
1.3	Empirical evaluation of Zipf's law on <i>Tom Sawyer</i> .	24
1.4	Commonest <b>bigram</b> collocations in the New York <i>Times</i> .	30
1.5	Frequent <b>bigrams</b> after filtering.	32
2.1	Likelihood ratios between two theories.	58
2.2	Statistical NLP problems as decoding problems.	71
3.1	Common inflections of nouns.	84
3.2	Pronoun forms in English.	86
3.3	Features commonly marked on verbs.	90
4.1	Major suppliers of electronic corpora with contact URLs.	119
4.2	Different formats for telephone numbers appearing in an issue of <i>The Economist</i> .	131
4.3	Sentence lengths in newswire text.	137
4.4	Sizes of various tag sets.	140
4.5	Comparison of different tag sets: adjective, adverb, conjunction, determiner, noun, and pronoun tags.	141
4.6	Comparison of different tag sets: Verb, preposition, punctuation and symbol tags.	142
5.1	Finding Collocations: Raw Frequency.	154
5.2	Part of speech tag patterns for collocation filtering.	154
5.3	Finding Collocations: Justeson and Katz' part-of-speech filter.	155

<b>5.4</b>	The nouns <i>w</i> occurring most often in the patterns ' <i>strong w</i> ' and ' <i>powerful w</i> '.	156
<b>5.5</b>	Finding collocations based on mean and variance.	161
<b>5.6</b>	Finding collocations: The <i>t</i> test applied to 10 bigrams that occur with frequency 20.	166
<b>5.7</b>	Words that occur significantly more often with <i>powerful</i> (the first ten words) and <i>strong</i> (the last ten words).	167
<b>5.8</b>	A 2-by-2 table showing the dependence of occurrences of new and <i>companies</i> .	169
<b>5.9</b>	Correspondence of <i>vache</i> and cow in an aligned corpus.	171
<b>5.10</b>	Testing for the independence of words in different corpora using $\chi^2$ .	171
<b>5.11</b>	How to compute Dunning's likelihood ratio test.	172
<b>5.12</b>	Bigrams of <i>powerful</i> with the highest scores according to Dunning's likelihood ratio test.	174
<b>5.13</b>	Damerau's frequency ratio test.	176
<b>5.14</b>	Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information.	178
<b>5.15</b>	Correspondence of <i>chambre</i> and <i>house</i> and <i>communes</i> and <i>house in</i> the aligned Hansard corpus.	179
<b>5.16</b>	Problems for Mutual Information from data sparseness.	181
<b>5.17</b>	Different definitions of <i>mutual information</i> in (Cover and Thomas 1991) and (Fano 1961).	182
<b>5.18</b>	Collocations in the BBI Combinatory Dictionary of English for the words <i>strength</i> and <i>power</i> .	185
<b>6.1</b>	Growth in number of parameters for n-gram models.	194
<b>6.2</b>	Notation for the statistical estimation chapter.	197
<b>6.3</b>	Probabilities of each successive word for a clause from <i>Persuasion</i> .	200
<b>6.4</b>	Estimated frequencies for the AP data from Church and Gale(1991a).	203
<b>6.5</b>	Expected Likelihood Estimation estimates for the word following <i>was</i> .	205
<b>6.6</b>	Using the <i>t</i> test for comparing the performance of two systems.	209
<b>6.7</b>	Extracts from the frequencies of frequencies distribution for bigrams and trigrams in the Austen corpus.	214

6.8	Good-Turing estimates for bigrams: Adjusted frequencies and probabilities.	215
6.9	Good-Turing bigram frequency estimates for the clause from <i>Persuasion</i> .	215
6.10	Back-off language models with Good-Turing estimation tested on <i>Persuasion</i> .	223
6.11	Probability estimates of the test clause according to various language models.	224
7.1	Notational conventions used in this chapter.	235
7.2	Clues for two senses of <i>drug</i> used by a Bayesian classifier.	238
7.3	Highly informative indicators for three ambiguous French words.	239
7.4	Two senses of ash.	243
7.5	Disambiguation of ash with Lesk's algorithm.	243
7.6	Some results of thesaurus-based disambiguation.	247
7.7	How to disambiguate <i>interest</i> using a second-language corpus.	248
7.8	Examples of the one sense per discourse constraint.	250
7.9	Some results of unsupervised disambiguation.	256
8.1	The <i>F</i> measure and accuracy are different objective functions.	270
8.2	Some subcategorization frames with example verbs and sentences.	271
8.3	Some subcategorization frames learned by Manning's system.	276
8.4	An example where the simple model for resolving PP attachment ambiguity fails.	280
8.5	Selectional Preference Strength (SPS).	290
8.6	Association strength distinguishes a verb's plausible and implausible objects.	292
8.7	Similarity measures for binary vectors.	299
8.8	The cosine as a measure of semantic similarity.	302
8.9	Measures of (dis-)similarity between probability distributions.	304
8.10	Types of words occurring in the LOB corpus that were not covered by the OALD dictionary.	310
9.1	Notation used in the HMM chapter.	324
9.2	Variable calculations for $O = (\text{lem}, \text{ice\_t}, \text{cola})$ .	330
10.1	Some part-of-speech tags frequently used for tagging English.	342

10.2	Notational conventions for tagging.	346
10.3	Idealized counts of some tag transitions in the Brown Corpus.	348
10.4	Idealized counts of tags that some words occur within the Brown Corpus.	349
10.5	Table of probabilities for dealing with unknown words in tagging.	352
10.6	Initialization of the parameters of an HMM.	359
10.7	Triggering environments in Brill's transformation-based tagger.	363
10.8	Examples of some transformations learned in transformation-based tagging.	363
10.9	Examples of frequent errors of probabilistic taggers.	374
10.10	A portion of a confusion matrix for part of speech tagging.	375
11.1	Notation for the PCFG chapter.	383
11.2	A simple Probabilistic Context Free Grammar (PCFG).	384
11.3	Calculation of inside probabilities.	394
12.1	Abbreviations for phrasal categories in the Penn Treebank.	413
12.2	Frequency of common subcategorization frames (local trees expanding VP) for selected verbs.	418
12.3	Selected common expansions of NP as Subject vs. Object, ordered by log odds ratio.	420
12.4	Selected common expansions of NP as first and second object inside VP.	420
12.5	Precision and recall evaluation results for PP attachment errors for different styles of phrase structure.	436
12.6	Comparison of some statistical parsing systems.	455
13.1	Sentence alignment papers.	470
14.1	A summary of the attributes of different clustering algorithms.	500
14.2	Symbols used in the clustering chapter.	501
14.3	Similarity functions used in clustering.	503
14.4	An example of K-means clustering.	518
14.5	An example of a Gaussian mixture.	521
15.1	A small stop list for English.	533
15.2	An example of the evaluation of rankings.	535

15.3	Three quantities that are commonly used in term weighting in information retrieval.	542
15.4	Term and document frequencies of two words in an example corpus.	542
15.5	Components of tf.idf weighting schemes.	544
15.6	Document frequency ( $df$ ) and collection frequency ( $cf$ ) for 6 words in the New York <i>Times</i> corpus.	547
15.7	Actual and estimated number of documents with $k$ occurrences for six terms.	550
15.8	Example for exploiting co-occurrence in computing content similarity.	554
15.9	The matrix of document correlations $B^T B$ .	562
16.1	Some examples of classification tasks in NLP.	576
16.2	Contingency table for evaluating a binary classifier.	577
16.3	The representation of document 11, shown in figure 16.3.	581
16.4	An example of information gain as a splitting criterion.	582
16.5	Contingency table for a decision tree for the Reuters category “earnings.”	586
16.6	An example of a maximum entropy distribution in the form of equation (16.4).	593
16.7	An empirical distribution whose corresponding maximum entropy distribution is the one in table 16.6.	594
16.8	Feature weights in maximum entropy modeling for the category “earnings” in Reuters.	595
16.9	Classification results for the distribution corresponding to table 16.8 on the test set.	595
16.10	Perceptron for the “earnings” category.	601
16.11	Classification results for the perceptron in table 16.10 on the test set.	602
16.12	Classification results for an 1NN categorizer for the “earnings” category.	606

## *List of Figures*

1.1	Zipf's law.	26
1.2	Mandelbrot's formula.	27
1.3	Key Word In Context (KWIC) display for the word <i>showed</i> .	32
1.4	Syntactic frames for <i>showed in Tom Sawyer</i> .	33
2.1	A diagram illustrating the calculation of conditional probability $P(A B)$ .	42
2.2	A random variable X for the sum of two dice.	45
2.3	Two examples of binomial distributions: $b(r; 10, 0.7)$ and $b(r; 10, 0.1)$ .	52
2.4	Example normal distribution curves: $n(x; 0, 1)$ and $n(x; 1.5, 2)$ .	53
2.5	The entropy of a weighted coin.	63
2.6	The relationship between mutual information $I$ and entropy $H$ .	67
2.7	The noisy channel model.	69
2.8	A binary symmetric channel.	69
2.9	The noisy channel model in linguistics.	70
3.1	An example of recursive phrase structure expansion.	99
3.2	An example of a prepositional phrase attachment ambiguity.	108
4.1	Heuristic sentence boundary detection algorithm.	135
4.2	A sentence as tagged according to several different tag sets.	140
5.1	Using a three word collocational window to capture bigrams at a distance.	158

<b>5.2</b>	Histograms of the position of strong relative to three words.	<b>160</b>
<b>7.1</b>	Bayesian disambiguation.	<b>238</b>
<b>7.2</b>	The Flip-Flop algorithm applied to finding indicators for disambiguation.	<b>240</b>
<b>7.3</b>	Lesk's dictionary-based disambiguation algorithm.	<b>243</b>
<b>7.4</b>	Thesaurus-based disambiguation.	<b>245</b>
<b>7.5</b>	Adaptive thesaurus-based disambiguation.	<b>246</b>
<b>7.6</b>	Disambiguation based on a second-language corpus.	<b>249</b>
<b>7.7</b>	Disambiguation based on “one sense per collocation” and “one sense per discourse.”	<b>252</b>
<b>7.8</b>	An EM algorithm for learning a word sense clustering.	<b>254</b>
<b>8.1</b>	A diagram motivating the measures of precision and recall.	<b>268</b>
<b>8.2</b>	Attachments in a complex sentence.	<b>285</b>
<b>8.3</b>	A document-by-word matrix $A$ .	<b>297</b>
<b>8.4</b>	A word-by-word matrix $B$ .	<b>297</b>
<b>8.5</b>	A modifier-by-head matrix $C$ .	<b>297</b>
<b>9.1</b>	A Markov model.	<b>319</b>
<b>9.2</b>	The crazy soft drink machine, showing the states of the machine and the state transition probabilities.	<b>321</b>
<b>9.3</b>	A section of an HMM for a linearly interpolated language model.	<b>323</b>
<b>9.4</b>	A program for a Markov process.	<b>325</b>
<b>9.5</b>	Trellis algorithms.	<b>328</b>
<b>9.6</b>	Trellis algorithms: Closeup of the computation of forward probabilities at one node.	<b>329</b>
<b>9.7</b>	The probability of traversing an arc.	<b>334</b>
<b>10.1</b>	Algorithm for training a Visible Markov Model Tagger.	<b>348</b>
<b>10.2</b>	Algorithm for tagging with a Visible Markov Model Tagger.	<b>350</b>
<b>10.3</b>	The learning algorithm for transformation-based tagging.	<b>364</b>
<b>11.1</b>	The two parse trees, their probabilities, and the sentence probability.	<b>385</b>
<b>11.2</b>	A Probabilistic Regular Grammar (PRG).	<b>390</b>
<b>11.3</b>	Inside and outside probabilities in PCFGs.	<b>391</b>
<b>12.1</b>	A word lattice (simplified).	<b>408</b>

<b>12.2</b>	A Penn Treebank tree.	<b>413</b>
<b>12.3</b>	Two <b>CFG</b> derivations of the same tree.	<b>421</b>
<b>12.4</b>	An LC stack parser.	<b>425</b>
<b>12.5</b>	Decomposing a local tree into dependencies.	<b>430</b>
<b>12.6</b>	An example of the <b>PARSEVAL</b> measures.	<b>433</b>
<b>12.7</b>	The idea of crossing brackets.	<b>434</b>
<b>12.8</b>	Penn trees versus other trees.	<b>436</b>
<b>13.1</b>	Different strategies for Machine Translation.	<b>464</b>
<b>13.2</b>	Alignment and correspondence.	<b>469</b>
<b>13.3</b>	Calculating the cost of alignments.	<b>473</b>
<b>13.4</b>	A sample dot plot.	<b>476</b>
<b>13.5</b>	The pillow-shaped envelope that is searched.	<b>480</b>
<b>13.6</b>	The noisy channel model in machine translation.	<b>486</b>
<b>14.1</b>	A single-link clustering of 22 frequent English words represented as a dendrogram.	<b>496</b>
<b>14.2</b>	Bottom-up hierarchical clustering.	<b>502</b>
<b>14.3</b>	Top-down hierarchical clustering.	<b>502</b>
<b>14.4</b>	A cloud of points in a plane.	<b>504</b>
<b>14.5</b>	Intermediate clustering of the points in figure 14.4.	<b>504</b>
<b>14.6</b>	Single-link clustering of the points in figure 14.4.	<b>505</b>
<b>14.7</b>	Complete-link clustering of the points in figure 14.4.	<b>505</b>
<b>14.8</b>	The K-means clustering algorithm.	<b>516</b>
<b>14.9</b>	One iteration of the K-means algorithm.	<b>517</b>
<b>14.10</b>	An example of using the EM algorithm for soft clustering.	<b>519</b>
<b>15.1</b>	Results of the search ‘“glass pyramid” Pei Louvre’ on an internet search engine.	<b>531</b>
<b>15.2</b>	Two examples of precision-recall curves.	<b>537</b>
<b>15.3</b>	A vector space with two dimensions.	<b>540</b>
<b>15.4</b>	The Poisson distribution.	<b>546</b>
<b>15.5</b>	An example of a term-by-document matrix $A$ .	<b>555</b>
<b>15.6</b>	Dimensionality reduction.	<b>555</b>
<b>15.7</b>	An example of linear regression.	<b>558</b>
<b>15.8</b>	The matrix $T$ of the <b>SVD</b> decomposition of the matrix in figure 15.5.	<b>560</b>
<b>15.9</b>	The matrix of singular values of the <b>SVD</b> decomposition of the matrix in figure 15.5.	<b>560</b>

15.10	The matrix $D$ of the SVD decomposition of the matrix in figure 15.5.	561
15.11	The matrix $B = S_{2 \times 2}D_{2 \times n}$ of documents after rescaling with singular values and reduction to two dimensions.	562
15.12	Three constellations of cohesion scores in topic boundary identification.	569
16.1	A decision tree.	578
16.2	Geometric interpretation of part of the tree in figure 16.1.	579
16.3	An example of a Reuters news story in the topic category “earnings.”	580
16.4	Pruning a decision tree.	585
16.5	Classification accuracy depends on the amount of training data available.	587
16.6	An example of how decision trees use data inefficiently from the domain of phonological rule learning.	588
16.7	The Perceptron Learning Algorithm.	598
16.8	One error-correcting step of the perceptron learning algorithm.	600
16.9	Geometric interpretation of a perceptron.	602

## ***Table of Notations***

<b>u</b>	Union of sets
$\cap$	Intersection of sets
$A - B, A \setminus B$	Set difference
$A'$	The complement of set $A$
$\emptyset$	The empty set
$2^A, \mathcal{P}(A)$	The power set of $A$
$ A $	Cardinality of a set
$\sum$	Sum
$\prod$	Product
$p \Rightarrow q$	$p$ implies $q$ (logical inference)
$p \Leftrightarrow q$	$p$ and $q$ are logically equivalent
$\stackrel{\text{def}}{=}$	Defined to be equal to (only used if “=” is ambiguous)
$\mathbb{R}$	The set of real numbers
$\mathbb{N}$	The set of natural numbers
$n!$	The factorial of $n$
$\infty$	Infinity
$ x $	Absolute value of a number
$\ll$	Much smaller than
$\gg$	Much greater than
$f : A - B$	A function $f$ from values in $A$ to $B$
$\max f$	The maximum value of $f$

$\min f$	The minimum value of $f$
$\arg \max f$	The argument for which $f$ has its maximum value
$\arg \min f$	The argument for which $f$ has its minimum value
$\lim_{x \rightarrow \infty} f(x)$	The limit of $f$ as $x$ tends to infinity
$f \propto g$	$f$ is proportional to $g$
$\partial$	Partial derivative
$\int$	Integral
$\log a$	The logarithm of $a$
$\exp(x), e^x$	The exponential function
$[a]$	The smallest integer $i$ s.t. $i \geq a$
$\vec{x}$	A real-valued vector: $\vec{x} \in \mathbb{R}^n$
$ \vec{x} $	Euclidean length of $\vec{x}$
$\vec{x} \cdot \vec{y}$	The dot product of $\vec{x}$ and $\vec{y}$
$\cos(\vec{x}, \vec{y})$	The cosine of the angle between $\vec{x}$ and $\vec{y}$
$c_{ij}$	Element in row $i$ and column $j$ of matrix $C$
$C^T$	Transpose of matrix $C$
$\hat{X}$	Estimate of $X$
$E(X)$	Expectation of $X$
$\text{Var}(X)$	Variance of $X$
$\mu$	Mean
$\sigma$	Standard deviation
$\bar{x}$	Sample mean
$s^2$	Sample variance
$P(A B)$	The probability of $A$ conditional on $B$
$X \sim p(x)$	Random variable $X$ is distributed according to $p$
$b(r; n, p)$	The binomial distribution
$\binom{n}{r}$	Combination or binomial coefficient (the number of ways of choosing $r$ objects from $n$ )
$n(x; \mu, \sigma)$	The normal distribution
$H(X)$	Entropy

$I(X; Y)$	Mutual information
$D(p \parallel q)$	Kullback-Leibler (KL) divergence
$C(\cdot)$	Count of the entity in parentheses
$f_u$	The relative frequency of $u$ .
$w_{ij}, w(i)(j)$	The words $w_i, w_{i+1}, \dots, w_j$
$w_{i,j}$	The same as $w_{ij}$
$w_i, \dots, w_j$	The same as $w_{ij}$
$O(n)$	Time complexity of an algorithm
*	Ungrammatical sentence or phrase or ill-formed word
?	Marginally grammatical sentence or marginally acceptable phrase

Note. Some chapters have separate notation tables for symbols that are used locally: table 6.2 (Statistical Inference), table 7.1 (Word Sense Disambiguation), table 9.1 (Markov Models), table 10.2 (Tagging), table 11.1 (Probabilistic Context-Free Grammars), and table 14.2 (Clustering).

## *Preface*

THE NEED for a thorough textbook for Statistical Natural Language Processing hardly needs to be argued for in the age of on-line information, electronic communication and the World Wide Web. Increasingly, businesses, government agencies and individuals are confronted with large amounts of text that are critical for working and living, but not well enough understood to get the enormous value out of them that they potentially hide.

At the same time, the availability of large text corpora has changed the scientific approach to language in linguistics and cognitive science. Phenomena that were not detectable or seemed uninteresting in studying toy domains and individual sentences have moved into the center field of what is considered important to explain. Whereas as recently as the early 1990s quantitative methods were seen as so inadequate for linguistics that an important textbook for mathematical linguistics did not cover them in any way, they are now increasingly seen as crucial for linguistic theory.

In this book we have tried to achieve a balance between theory and practice, and between intuition and rigor. We attempt to ground approaches in theoretical ideas, both mathematical and linguistic, but simultaneously we try to not let the material get too dry, and try to show how theoretical ideas have been used to solve practical problems. To do this, we first present key concepts in probability theory, statistics, information theory, and linguistics in order to give students the foundations to understand the field and contribute to it. Then we describe the problems that are addressed in Statistical Natural Language Processing (NLP), like tagging and disambiguation, and a selection of important work so

that students are grounded in the advances that have been made and, having understood the special problems that language poses, can move the field forward.

When we designed the basic structure of the book, we had to make a number of decisions about what to include and how to organize the material. A key criterion was to keep the book to a manageable size. (We didn't entirely succeed!) Thus the book is not a complete introduction to probability theory, information theory, statistics, and the many other areas of mathematics that are used in Statistical NLP. We have tried to cover those topics that seem most important in the field, but there will be many occasions when those teaching from the book will need to use supplementary materials for a more in-depth coverage of mathematical foundations that are of particular interest.

We also decided against attempting to present Statistical NLP as homogeneous in terms of the mathematical tools and theories that are used. It is true that a unified underlying mathematical theory would be desirable, but such a theory simply does not exist at this point. This has led to an eclectic mix in some places, but we believe that it is too early to mandate that a particular approach to NLP is right and should be given preference to others.

A perhaps surprising decision is that we do not cover speech recognition. Speech recognition began as a separate field to NLP, mainly growing out of electrical engineering departments, with separate conferences and journals, and many of its own concerns. However, in recent years there has been increasing convergence and overlap. It was research into speech recognition that inspired the revival of statistical methods within NLP, and many of the techniques that we present were developed first for speech and then spread over into NLP. In particular, work on language models within speech recognition greatly overlaps with the discussion of language models in this book. Moreover, one can argue that speech recognition is the area of language processing that currently is the most successful and the one that is most widely used in applications. Nevertheless, there are a number of practical reasons for excluding the area from this book: there are already several good textbooks for speech, it is not an area in which we have worked or are terribly expert, and this book seemed quite long enough without including speech as well. Additionally, while there is overlap, there is also considerable separation: a speech recognition textbook requires thorough coverage of issues in signal analysis and

acoustic modeling which would not generally be of interest or accessible to someone from a computer science or NLP background, while in the reverse direction, most people studying speech would be uninterested in many of the NLP topics on which we focus.

Other related areas that have a somewhat fuzzy boundary with Statistical NLP are machine learning, text categorization, information retrieval, and cognitive science. For all of these areas, one can find examples of work that is not covered and which would fit very well into the book. It was simply a matter of space that we did not include important concepts, methods and problems like minimum description length, back-propagation, the Rocchio algorithm, and the psychological and cognitive-science literature on frequency effects on language processing.

The decisions that were most difficult for us to make are those that concern the boundary between statistical and non-statistical NLP. We believe that, when we started the book, there was a clear dividing line between the two, but this line has become much more fuzzy recently. An increasing number of non-statistical researchers use corpus evidence and incorporate quantitative methods. And it is now generally accepted in Statistical NLP that one needs to start with all the scientific knowledge that is available about a phenomenon when building a probabilistic or other model, rather than closing one's eyes and taking a clean-slate approach.

Many NLP researchers will therefore question the wisdom of writing a separate textbook for the statistical side. And the last thing we would want to do with this textbook is to promote the unfortunate view in some quarters that linguistic theory and symbolic computational work are not relevant to Statistical NLP. However, we believe that there is so much quite complex foundational material to cover that one simply cannot write a textbook of a manageable size that is a satisfactory and comprehensive introduction to all of NLP. Again, other good texts already exist, and we recommend using supplementary material if a more balanced coverage of statistical and non-statistical methods is desired.

A final remark is in order on the title we have chosen for this book. Calling the field Statistical *Natural Language Processing* might seem questionable to someone who takes their definition of a statistical method from a standard introduction to statistics. Statistical NLP as we define it comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra.

While probability theory is the foundation for formal statistical reasoning, we take the basic meaning of the term ‘statistics’ as being broader, encompassing all quantitative approaches to data (a definition which one can quickly confirm in almost any dictionary). Although there is thus some potential for ambiguity, Statistical NLP has been the most widely used term to refer to non-symbolic and non-logical work on NLP over the past decade, and we have decided to keep with this term.

**Acknowledgments.** Over the course of the three years that we were working on this book, a number of colleagues and friends have made comments and suggestions on earlier drafts. We would like to express our gratitude to all of them, in particular, Einat Amitay, Chris Brew, Thorsten Brants, Andreas Eisele, Michael Ernst, Oren Etzioni, Marc Friedman, Eric Gaussier, Eli Hagen, Marti Hearst, Nitin Indurkhya, Michael Inman, Mark Johnson, Rosie Jones, Tom Kalt, Andy Kehler, Julian Kupiec, Michael Littman, Arman Maghbouleh, Amir Najmi, Kris Popat, Fred Popowich, Geoffrey Sampson, Hadar Shemtov, Scott Stoness, David Yarowsky, and Jakub Zavrel. We are particularly indebted to Bob Carpenter, Eugene Charniak, Raymond Mooney, and an anonymous reviewer for MIT Press, who suggested a large number of improvements, both in content and exposition, that we feel have greatly increased the overall quality and usability of the book. We hope that they will sense our gratitude when they notice ideas which we have taken from their comments without proper acknowledgement.

We would like to also thank: Francine Chen, Kris Halvorsen, and Xerox PARC for supporting the second author while writing this book, Jane Manning for her love and support of the first author, Robert Dale and Dikran Karagueuzian for advice on book design, and Amy Brand for her regular help and assistance as our editor.

**Feedback.** While we have tried hard to make the contents of this book understandable, comprehensive, and correct, there are doubtless many places where we could have done better. We welcome feedback to the authors via email to [cmanning@acm.org](mailto:cmanning@acm.org) or [hinrich@hotmail.com](mailto:hinrich@hotmail.com).

In closing, we can only hope that the availability of a book which collects many of the methods used within Statistical NLP and presents them

in an accessible fashion will create excitement in potential students, and help ensure continued rapid progress in the field.

*Christopher Manning*

*Hinrich Schiitze*

*February 1999*

## *Road Map*

IN GENERAL, this book is written to be suitable for a graduate-level semester-long course focusing on Statistical NLP. There is actually rather more material than one could hope to cover in a semester, but that richness gives ample room for the teacher to pick and choose. It is assumed that the student has prior programming experience, and has some familiarity with formal languages and symbolic parsing methods. It is also assumed that the student has a basic grounding in such mathematical concepts as set theory, logarithms, vectors and matrices, summations, and integration - we hope nothing more than an adequate high school education! The student may have already taken a course on symbolic NLP methods, but a lot of background is not assumed. In the directions of probability and statistics, and linguistics, we try to briefly summarize all the necessary background, since in our experience many people wanting to learn about Statistical NLP methods have no prior knowledge in these areas (perhaps this will change over time!). Nevertheless, study of supplementary material in these areas is probably necessary for a student to have an adequate foundation from which to build, and can only be of value to the prospective researcher.

What is the best way to read this book and teach from it? The book is organized into four parts: Preliminaries (part I), Words (part II), Grammar (part III), and Applications and Techniques (part IV).

Part I lays out the mathematical and linguistic foundation that the other parts build on. Concepts and techniques introduced here are referred to throughout the book.

Part II covers word-centered work in Statistical NLP. There is a natural progression from simple to complex linguistic phenomena in its four

chapters on collocations, n-gram models, word sense disambiguation, and lexical acquisition, but each chapter can also be read on its own.

The four chapters in part III, Markov Models, tagging, probabilistic context free grammars, and probabilistic parsing, build on each other, and so they are best presented in sequence. However, the tagging chapter can be read separately with occasional references to the Markov Model chapter.

The topics of part IV are four applications and techniques: statistical alignment and machine translation, clustering, information retrieval, and text categorization. Again, these chapters can be treated separately according to interests and time available, with the few dependencies between them marked appropriately.

Although we have organized the book with a lot of background and foundational material in part I, we would not advise going through all of it carefully at the beginning of a course based on this book. What the authors have generally done is to review the really essential bits of part I in about the first 6 hours of a course. This comprises very basic probability (through section 2.1.8), information theory (through section 2.2.7), and essential practical knowledge – some of which is contained in chapter 4, and some of which is the particulars of what is available at one’s own institution. We have generally left the contents of chapter 3 as a reading assignment for those without much background in linguistics. Some knowledge of linguistic concepts is needed in many chapters, but is particularly relevant to chapter 12, and the instructor may wish to review some syntactic concepts at this point. Other material from the early chapters is then introduced on a “need to know” basis during the course.

The choice of topics in part II was partly driven by a desire to be able to present accessible and interesting topics early in a course, in particular, ones which are also a good basis for student programming projects. We have found collocations (chapter 5), word sense disambiguation (chapter 7), and attachment ambiguities (section 8.3) particularly successful in this regard. Early introduction of attachment ambiguities is also effective in showing that there is a role for linguistic concepts and structures in Statistical NLP. Much of the material in chapter 6 is rather detailed reference material. People interested in applications like speech or optical character recognition may wish to cover all of it, but if n-gram language models are not a particular focus of interest, one may only want to read through section 6.2.3. This is enough to understand the concept of likelihood, maximum likelihood estimates, a couple of simple smoothing methods (usually necessary if students are to be building any

probabilistic models on their own), and good methods for assessing the performance of systems.

In general, we have attempted to provide ample cross-references so that, if desired, an instructor can present most chapters independently with incorporation of prior material where appropriate. In particular, this is the case for the chapters on collocations, lexical acquisition, tagging, and information retrieval.

**Exercises.** There are exercises scattered through or at the end of every chapter. They vary enormously in difficulty and scope. We have tried to provide an elementary classification as follows:

- \* Simple problems that range from text comprehension through to such things as mathematical manipulations, simple proofs, and thinking of examples of something.
- \* \* More substantial problems, many of which involve either programming or corpus investigations. Many would be suitable as an assignment to be done over two weeks.
- ★★★ Large, difficult, or open-ended problems. Many would be suitable as a term project.

**WEBSITE** Finally, we encourage students and teachers to take advantage of the material and the references on the companion *website*. It can be accessed directly at the **URL** <http://www.sultry.arts.usyd.edu.au/fsnlp>, or found through the **MIT** Press website <http://mitpress.mit.edu>, by searching for this book.