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

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 *Mathematical Foundations* 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 *Linguistic Essentials* 81

3.1 Parts of Speech and Morphology 81
3.1.1 Nouns and pronouns 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 structure grammars 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 *Corpus-Based Work* **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
4.4 Further Reading 145
4.5 Exercises 147

II Words 149

5 *Collocations* **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 *Statistical Inference: n-gram Models over Sparse Data* **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 *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 ***Word Sense Disambiguation*** **229**

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

8 *Lexical Acquisition* 265
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

III Grammar 315

9 *Markov Models* 317
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

10 *Part-of-Speech Tagging* 341
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
10.3.2 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

11 *Probabilistic Context Free Grammars* 381

11.1 Some Features of PCFGs 386
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 398
11.4 Problems with the Inside-Outside Algorithm 401
11.5 Further Reading 402
11.6 Exercises 404

12 *Probabilistic Parsing* 407

12.1 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 Building parsers: Search methods 439
12.1.11 Use of the geometric mean 442
12.2 Some Approaches 443
12.2.1 Non-lexicalized treebank 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

14.4 Exercises 528

15 ***Topics in Information Retrieval*** **529**
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

16 ***Text Categorization*** **575**
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

Tiny Statistical Tables **609**

Bibliography **611**

Index **657**

List of Tables

1.1	Common words in *Tom Sawyer.*	21
1.2	Frequency of frequencies of word types in *Tom Sawyer.*	22
1.3	Empirical evaluation of Zipf's law on *Tom Sawyer.*	24
1.4	Commonest bigram collocations in the *New York Times.*	30
1.5	Frequent bigrams 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 *The Economist.*	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

5.4 The nouns *w* occurring most often in the patterns '*strong w*' and '*powerful w*.' 156
5.5 Finding collocations based on mean and variance. 161
5.6 Finding collocations: The *t* test applied to 10 bigrams that occur with frequency 20. 166
5.7 Words that occur significantly more often with *powerful* (the first ten words) and *strong* (the last ten words). 167
5.8 A 2-by-2 table showing the dependence of occurrences of *new* and *companies*. 169
5.9 Correspondence of *vache* and *cow* in an aligned corpus. 171
5.10 Testing for the independence of words in different corpora using χ^2. 171
5.11 How to compute Dunning's likelihood ratio test. 172
5.12 Bigrams of *powerful* with the highest scores according to Dunning's likelihood ratio test. 174
5.13 Damerau's frequency ratio test. 176
5.14 Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information. 178
5.15 Correspondence of *chambre* and *house* and *communes* and *house* in the aligned Hansard corpus. 179
5.16 Problems for Mutual Information from data sparseness. 181
5.17 Different definitions of *mutual information* in (Cover and Thomas 1991) and (Fano 1961). 182
5.18 Collocations in the BBI Combinatory Dictionary of English for the words *strength* and *power*. 185

6.1 Growth in number of parameters for *n*-gram models. 194
6.2 Notation for the statistical estimation chapter. 197
6.3 Probabilities of each successive word for a clause from *Persuasion*. 200
6.4 Estimated frequencies for the AP data from Church and Gale (1991a). 203
6.5 Expected Likelihood Estimation estimates for the word following *was*. 205
6.6 Using the *t* test for comparing the performance of two systems. 209
6.7 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 *Persuasion*. 215
6.10 Back-off language models with Good-Turing estimation tested on *Persuasion*. 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 *drug* 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 *interest* 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 F 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 = (lem, ice_t, 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 Times* 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 $B^{\mathrm{T}}B$. 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 *showed.*	32
1.4	Syntactic frames for *showed* in *Tom Sawyer.*	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: $\mathrm{b}(r;\ 10, 0.7)$ and $\mathrm{b}(r;\ 10, 0.1)$.	52
2.4	Example normal distribution curves: $\mathrm{n}(x;\ 0, 1)$ and $\mathrm{n}(x;\ 1.5, 2)$.	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

5.2 Histograms of the position of *strong* relative to three words. 160

7.1 Bayesian disambiguation. 238
7.2 The Flip-Flop algorithm applied to finding indicators for disambiguation. 240
7.3 Lesk's dictionary-based disambiguation algorithm. 243
7.4 Thesaurus-based disambiguation. 245
7.5 Adaptive thesaurus-based disambiguation. 246
7.6 Disambiguation based on a second-language corpus. 249
7.7 Disambiguation based on "one sense per collocation" and "one sense per discourse." 252
7.8 An EM algorithm for learning a word sense clustering. 254

8.1 A diagram motivating the measures of precision and recall. 268
8.2 Attachments in a complex sentence. 285
8.3 A document-by-word matrix A. 297
8.4 A word-by-word matrix B. 297
8.5 A modifier-by-head matrix C. 297

9.1 A Markov model. 319
9.2 The crazy soft drink machine, showing the states of the machine and the state transition probabilities. 321
9.3 A section of an HMM for a linearly interpolated language model. 323
9.4 A program for a Markov process. 325
9.5 Trellis algorithms. 328
9.6 Trellis algorithms: Closeup of the computation of forward probabilities at one node. 329
9.7 The probability of traversing an arc. 334

10.1 Algorithm for training a Visible Markov Model Tagger. 348
10.2 Algorithm for tagging with a Visible Markov Model Tagger. 350
10.3 The learning algorithm for transformation-based tagging. 364

11.1 The two parse trees, their probabilities, and the sentence probability. 385
11.2 A Probabilistic Regular Grammar (PRG). 390
11.3 Inside and outside probabilities in PCFGs. 391

12.1 A word lattice (simplified). 408

12.2 A Penn Treebank tree. 413
12.3 Two CFG derivations of the same tree. 421
12.4 An LC stack parser. 425
12.5 Decomposing a local tree into dependencies. 430
12.6 An example of the PARSEVAL measures. 433
12.7 The idea of crossing brackets. 434
12.8 Penn trees versus other trees. 436

13.1 Different strategies for Machine Translation. 464
13.2 Alignment and correspondence. 469
13.3 Calculating the cost of alignments. 473
13.4 A sample dot plot. 476
13.5 The pillow-shaped envelope that is searched. 480
13.6 The noisy channel model in machine translation. 486

14.1 A single-link clustering of 22 frequent English words represented as a dendrogram. 496
14.2 Bottom-up hierarchical clustering. 502
14.3 Top-down hierarchical clustering. 502
14.4 A cloud of points in a plane. 504
14.5 Intermediate clustering of the points in figure 14.4. 504
14.6 Single-link clustering of the points in figure 14.4. 505
14.7 Complete-link clustering of the points in figure 14.4. 505
14.8 The K-means clustering algorithm. 516
14.9 One iteration of the K-means algorithm. 517
14.10 An example of using the EM algorithm for soft clustering. 519

15.1 Results of the search '"glass pyramid" Pei Louvre' on an internet search engine. 531
15.2 Two examples of precision-recall curves. 537
15.3 A vector space with two dimensions. 540
15.4 The Poisson distribution. 546
15.5 An example of a term-by-document matrix A. 555
15.6 Dimensionality reduction. 555
15.7 An example of linear regression. 558
15.8 The matrix T of the SVD of the matrix in figure 15.5. 560
15.9 The matrix of singular values of the SVD of the matrix in figure 15.5. 560
15.10 The matrix D^T of the SVD of the matrix in figure 15.5. 561

15.11 The matrix $B_{2\times d} = S_{2\times 2}D^{\mathrm{T}}{}_{2\times d}$ 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