

Foundations of Statistical Natural Language Processing

Christopher D. Manning

Hinrich Schütze

The MIT Press
Cambridge, Massachusetts
London, England

Contents

<i>List of Tables</i>	xv
-----------------------	----

<i>List of Figures</i>	xxi
------------------------	-----

<i>Table of Notations</i>	xxv
---------------------------	-----

<i>Preface</i>	xxix
----------------	------

<i>Road Map</i>	xxxv
-----------------	------

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	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	<i>Corpus-Based Work</i>	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	<i>Collocations</i>	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	<i>Statistical Inference: n-gram Models over Sparse Data</i>	191
6.1	Bins: Forming Equivalence Classes	192
6.1.1	Reliability vs. discrimination	192
6.1.2	n -gram models	192

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
<i>Tiny Statistical Tables</i>	609
<i>Bibliography</i>	611
<i>Index</i>	657