



MORGAN & CLAYPOOL PUBLISHERS

Linguistic Structure Prediction

49c24dbf57a9a262d0eb858701dff4a3
ebrary

Noah A. Smith

49c24dbf57a9a262d0eb858701dff4a3
ebrary

***SYNTHESIS LECTURES ON
HUMAN LANGUAGE TECHNOLOGIES***

Graeme Hirst, *Series Editor*

49c24dbf57a9a262d0eb858701dff4a3
ebruary

49c24dbf57a9a262d0eb858701dff4a3
ebruary

49c24dbf57a9a262d0eb858701dff4a3
ebruary

49c24dbf57a9a262d0eb858701dff4a3
ebruary

Linguistic Structure Prediction

Synthesis Lectures on Human Language Technologies

Editor

Graeme Hirst, *University of Toronto*

49c24dbf57a9a262d0eb858701dff4a3
ebrary

The series consists of 50- to 150-page monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Linguistic Structure Prediction

Noah A. Smith

2011

Learning to Rank for Information Retrieval and Natural Language Processing

Hang Li

2011

Computational Modeling of Human Language Acquisition

Afra Alishahi

2010

Introduction to Arabic Natural Language Processing

Nizar Y. Habash

2010

Cross-Language Information Retrieval

Jian-Yun Nie

2010

Automated Grammatical Error Detection for Language Learners

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault

2010

Data-Intensive Text Processing with MapReduce

Jimmy Lin and Chris Dyer

2010

49c24dbf57a9a262d0eb858701dff4a3
ebrary

49c24dbf57a9a262d0eb858701dff4a3
ebrary

Semantic Role Labeling

Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems

Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics

Graham Wilcock
2009

Dependency Parsing

Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval

ChengXiang Zhai
2008

Copyright © 2011 by Morgan & Claypool

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Linguistic Structure Prediction

Noah A. Smith

www.morganclaypool.com

ISBN: 9781608454051 paperback

ISBN: 9781608454068 ebook

DOI 10.2200/S00361ED1V01Y201105HLT013

A Publication in the Morgan & Claypool Publishers series

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #13

Series Editor: Graeme Hirst, *University of Toronto*

Series ISSN

Synthesis Lectures on Human Language Technologies

Print 1947-4040 Electronic 1947-4059

Linguistic Structure Prediction

49c24dbf57a9a262d0eb858701dff4a3
ebruary

Noah A. Smith
Carnegie Mellon University

49c24dbf57a9a262d0eb858701dff4a3
ebruary

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #13



MORGAN & CLAYPOOL PUBLISHERS

49c24dbf57a9a262d0eb858701dff4a3
ebruary

ABSTRACT

A major part of natural language processing now depends on the use of text data to build linguistic analyzers. We consider statistical, computational approaches to modeling linguistic structure. We seek to unify across many approaches and many kinds of linguistic structures. Assuming a basic understanding of natural language processing and/or machine learning, we seek to bridge the gap between the two fields. Approaches to decoding (i.e., carrying out linguistic structure prediction) and supervised and unsupervised learning of models that predict discrete structures as outputs are the focus. We also survey natural language processing problems to which these methods are being applied, and we address related topics in probabilistic inference, optimization, and experimental methodology.

49c24dbf57a9a262d0eb858701dff4a3
ebrary

KEYWORDS

natural language processing, computational linguistics, machine learning, decoding, supervised learning, unsupervised learning, structured prediction, probabilistic inference, statistical modeling

Contents

Preface	xiii
Acknowledgments	xix
1 Representations and Linguistic Data	1
1.1 Sequential Prediction	2
1.2 Sequence Segmentation	4
1.3 Word Classes and Sequence Labeling	5
1.3.1 Morphological Disambiguation	6
1.3.2 Chunking	7
1.4 Syntax	9
1.5 Semantics	11
1.6 Coreference Resolution	14
1.7 Sentiment Analysis	16
1.8 Discourse	16
1.9 Alignment	17
1.10 Text-to-Text Transformations	18
1.11 Types	18
1.12 Why Linguistic Structure is a Moving Target	19
1.13 Conclusion	20
2 Decoding: Making Predictions	23
2.1 Definitions	23
2.2 Five Views of Decoding	24
2.2.1 Probabilistic Graphical Models	27
2.2.2 Polytopes	31
2.2.3 Parsing with Grammars	37
2.2.4 Graphs and Hypergraphs	39
2.2.5 Weighted Logic Programs	41
2.3 Dynamic Programming	44
2.3.1 Shortest or Minimum-Cost Path	45

2.3.2	Semirings	48
2.3.3	DP as Logical Deduction	50
2.3.4	Solving DPs	55
2.3.5	Approximate Search	61
2.3.6	Reranking and Coarse-to-Fine Decoding	64
2.4	Specialized Graph Algorithms	64
2.4.1	Bipartite Matchings	65
2.4.2	Spanning Trees	65
2.4.3	Maximum Flow and Minimum Cut	66
2.5	Conclusion	67
3	Learning Structure from Annotated Data	69
3.1	Annotated Data	69
3.2	Generic Formulation of Learning	70
3.3	Generative Models	71
3.3.1	Decoding Rule	73
3.3.2	Multinomial-Based Models	73
3.3.3	Hidden Markov Models	74
3.3.4	Probabilistic Context-Free Grammars	78
3.3.5	Other Generative Multinomial-Based Models	78
3.3.6	Maximum Likelihood Estimation By Counting	79
3.3.7	Maximum <i>A Posteriori</i> Estimation	81
3.3.8	Alternative Parameterization: Log-Linear Models	83
3.3.9	Comments	85
3.4	Conditional Models	86
3.5	Globally Normalized Conditional Log-Linear Models	88
3.5.1	Logistic Regression	88
3.5.2	Conditional Random Fields	89
3.5.3	Feature Choice	91
3.5.4	Maximum Likelihood Estimation	92
3.5.5	Maximum <i>A Posteriori</i> Estimation	94
3.5.6	Pseudolikelihood	97
3.5.7	Toward Discriminative Learning	98
3.6	Large Margin Methods	99
3.6.1	Binary Classification	99
3.6.2	Perceptron	101
3.6.3	Multi-class Support Vector Machines	103

3.6.4	Structural SVM	104
3.6.5	Optimization	105
3.6.6	Discussion	106
3.7	Conclusion	106
4	Learning Structure from Incomplete Data	109
4.1	Unsupervised Generative Models	110
4.1.1	Expectation Maximization	111
4.1.2	Word Clustering	112
4.1.3	Hard and Soft K -Means	115
4.1.4	The Structured Case	117
4.1.5	Hidden Markov Models	119
4.1.6	EM Iterations Improve Likelihood	120
4.1.7	Extensions and Improvements	122
4.1.8	Log-Linear EM	123
4.1.9	Contrastive Estimation	124
4.2	Bayesian Unsupervised Learning	125
4.2.1	Empirical Bayes	127
4.2.2	Latent Dirichlet Allocation	127
4.2.3	EM in the Empirical Bayesian Setting	129
4.2.4	Inference	129
4.2.5	Nonparametric Bayesian Methods	134
4.2.6	Discussion	139
4.3	Hidden Variable Learning	140
4.3.1	Generative Models with Hidden Variables	141
4.3.2	Conditional Log-Linear Models with Hidden Variables	142
4.3.3	Large Margin Methods with Hidden Variables	143
4.4	Conclusion	145
5	Beyond Decoding: Inference	147
5.1	Partition Functions: Summing over \mathcal{Y}	148
5.1.1	Summing by Dynamic Programming	148
5.1.2	Other Summing Algorithms	150
5.2	Feature Expectations	150
5.2.1	Reverse DPs	152
5.2.2	Another Interpretation of Reverse Values	155
5.2.3	From Reverse Values to Expectations	157

5.2.4	Deriving the Reverse DP	159
5.2.5	Non-DP Expectations	160
5.3	Minimum Bayes Risk Decoding	163
5.4	Cost-Augmented Decoding	165
5.5	Decoding with Hidden Variables	165
5.6	Conclusion	167
A	Numerical Optimization	169
A.1	The Hill-Climbing Analogy	170
A.2	Coordinate Ascent	171
A.3	Gradient Ascent	172
A.3.1	Subgradient Methods	174
A.3.2	Stochastic Gradient Ascent	175
A.4	Conjugate Gradient and Quasi-Newton Methods	175
A.4.1	Conjugate Gradient	176
A.4.2	Newton's Method	176
A.4.3	Limited Memory BFGS	176
A.5	"Aggressive" Online Learners	177
A.6	Improved Iterative Scaling	178
B	Experimentation	181
B.1	Methodology	181
B.1.1	Training, Development, and Testing	182
B.1.2	Cross-Validation	183
B.1.3	Comparison without Replication	183
B.1.4	Oracles and Upper Bounds	184
B.2	Hypothesis Testing and Related Topics	184
B.2.1	Terminology	185
B.2.2	Standard Error	186
B.2.3	Beyond Standard Error for Sample Means	187
B.2.4	Confidence Intervals	188
B.2.5	Hypothesis Tests	189
B.2.6	Closing Notes	197
C	Maximum Entropy	199

D Locally Normalized Conditional Models 203

D.1 Probabilistic Finite-State Automata 203

D.2 Maximum Entropy Markov Models 204

D.3 Directional Effects 205

D.4 Comparison to Globally Normalized Models 206

D.5 Decoding 207

D.6 Theory vs. Practice 208

Bibliography 209

Author's Biography 241

Index 243

49c24dbf57a9a262d0eb858701dff4a3
ebruary

49c24dbf57a9a262d0eb858701dff4a3
ebruary

49c24dbf57a9a262d0eb858701dff4a3
ebruary

49c24dbf57a9a262d0eb858701dff4a3
ebruary