

Theory and Applications of Natural Language Processing

Series Editors:

Graeme Hirst (Textbooks)

Eduard Hovy (Edited volumes)

Mark Johnson (Monographs)

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

“Theory and Applications of Natural Language Processing” is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

- * Downloadable on your PC, e-reader or iPad
- * Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
- * Available online within an extensive network of academic and corporate R&D libraries worldwide
- * Never out of print thanks to innovative print-on-demand services
- * Competitively priced print editions for eBook customers thanks to MyCopy service <http://www.springer.com/librarians/e-content/mycopy>

For other titles published in this series, go to www.springer.com/series/8899

Aline Villavicencio • Thierry Poibeau
Anna Korhonen • Afra Alishahi
Editors

Cognitive Aspects of Computational Language Acquisition

 Springer

Editors

Aline Villavicencio
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil

Thierry Poibeau
Ecole Normale Supérieure
Université Sorbonne Nouvelle
LATTICE-CNRS, Paris, France

Anna Korhonen
Computer Laboratory
University of Cambridge
Cambridge, UK

Afra Alishahi
Tilburg center for Cognition
and Communication (TiCC)
Tilburg University
Tilburg, The Netherlands

ISSN 2192-032X

ISSN 2192-0338 (electronic)

ISBN 978-3-642-31862-7

ISBN 978-3-642-31863-4 (eBook)

DOI 10.1007/978-3-642-31863-4

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012954240

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgements

We would like to acknowledge the support of the labex (cluster of excellence) TransferS (France), the Royal Society and EPSRC grant EP/F030061/1 (UK), CAPES-COFECUB grant 707/11 and CNPq grants 479824/2009-6, 551964/2011-1 478222/2011-4 and 309569/2009-5 (Brazil).

Contents

Computational Modeling as a Methodology for Studying Human Language Learning	1
Thierry Poibeau, Aline Villavicencio, Anna Korhonen, and Afra Alishahi	
1 Overview.....	1
1.1 Theoretical Accounts of Language Modularity and Learnability.....	2
1.2 Investigations of Linguistic Hypotheses.....	5
2 Computational Models of Language Learning.....	7
2.1 What to Expect from a Model.....	8
2.2 Modeling Frameworks.....	10
2.3 Research Methods.....	13
3 Impact of Computational Modeling on the Study of Language.....	16
4 This Collection.....	17
4.1 Methods and Tools for Investigating Phonetics and Phonology.....	17
4.2 Classifying Words and Mapping Them to Meanings.....	18
4.3 Learning Morphology and Syntax.....	19
4.4 Linking Syntax to Semantics.....	20
5 Concluding Remarks.....	22
References.....	22

Part I Methods and Tools for Investigating Phonetics and Phonology

Phon: A Computational Basis for Phonological Database Building and Model Testing	29
Yvan Rose, Gregory J. Hedlund, Rod Byrne, Todd Wareham, and Brian MacWhinney	
1 Introduction.....	29
2 The PhonBank Project.....	31
2.1 PhonBank.....	31
2.2 Phon.....	32

3	Phon	32
3.1	Project Management	33
3.2	Media Linkage and Segmentation	34
3.3	Data Transcription	34
3.4	Multiple-Blind Transcription and Transcript Validation	35
3.5	Transcribed Utterance Segmentation	36
3.6	Syllabification Algorithm	36
3.7	Alignment Algorithm	38
4	Database Query	40
4.1	Terminology	40
4.2	Executing a Query	41
4.3	Creating a Query	42
4.4	An Illustrative Example	42
4.5	Additional Information	44
5	Future Projects	45
5.1	Interface for Acoustic Data	45
5.2	Extensions of Database Query Functionality	46
6	Discussion	47
	References	48
Language Dynamics in the Framework of Complex Networks:		
A Case Study on Self-Organization of the Consonant Inventories		51
Animesh Mukherjee, Monojit Choudhury, Niloy Ganguly, and Anupam Basu		
1	Introduction	51
2	Phonological Inventories: A Primer	53
3	Network Model of Consonant Inventories	56
3.1	Definition of PlaNet	56
3.2	Construction Methodology	57
4	Topological Properties of PlaNet	58
4.1	Degree Distribution of PlaNet	58
5	The Synthesis Model	61
6	Interpretation of the Synthesis Model	64
6.1	Mathematical Analysis of the Model	64
6.2	Linguistic Interpretation of the Model	66
7	Dynamics of the Language Families	68
8	Conclusion	71
	Appendix	72
	References	76

Part II Classifying Words and Mapping Them to Meanings

From Cues to Categories: A Computational Study of Children’s Early Word Categorization 81
 Fatemeh Torabi Asr, Afsaneh Fazly, and Zohreh Azimifar

1 Introduction 82

2 Related Work 83

 2.1 Experimental and Corpus-Based Studies 84

 2.2 Related Computational Models 84

3 Overview of This Study 85

4 Components of the Categorization Model 87

 4.1 Categorization Algorithm 87

 4.2 Cues Used in Categorization 88

5 Experimental Setup 89

 5.1 Corpus 89

 5.2 Feature Extraction 90

 5.3 Model Parameters 91

6 Discovering Syntactic Categories 91

 6.1 Evaluation Strategy 92

 6.2 Novel Word Categorization 92

7 Word Categorization and Semantic Prediction 95

 7.1 Semantic Feature Prediction 96

 7.2 Simulation of the Brown Experiment 97

8 Conclusions and Future Directions 99

Appendix 102

References 102

In Learning Nouns and Adjectives Remembering Matters: A Cortical Model 105
 Alessio Plebe, Vivian M. De la Cruz, and Marco Mazzone

1 Introduction 105

 1.1 On Learning First Words 106

 1.2 On Learning Nouns 108

 1.3 On Learning Adjectives 109

 1.4 Modeling Noun and Adjective Acquisition 110

2 Description of the Model 111

 2.1 Basic Units of the Model 112

 2.2 The Visual Pathway 113

 2.3 Auditory Pathway 115

 2.4 The Higher Cortical Map 116

3 Nouns and Adjectives Acquisition 118

 3.1 Simulation of Intrinsic and Extrinsic Experience 119

 3.2 Emergence of Organization in the Lower Maps 120

 3.3 Representation of Nouns and Adjectives in Model PFC 120

 3.4 Patterns of Connectivity of Nouns and Adjectives 123

4 Conclusions 125
 References 125

Part III Learning Morphology and Syntax

Treebank Parsing and Knowledge of Language 133
 Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick

1 Introduction: Treebank Parsing and Knowledge of Language 133
 2 Experimental Methods 140
 2.1 Parsing Systems Used 140
 2.2 Training Data, Testing, and Evaluation 140
 3 Case Study: Parsing Wh-Questions and QuestionBank 141
 3.1 Augmenting the Training Data 143
 4 Parsing and Tense: The Case of *Read* 146
 5 Case Study: Parsing Passives by Linguistic Regularization 154
 5.1 Passive Transformations: A Pilot Study 154
 6 Parsing “Unnatural” Languages? 160
 6.1 The Experimental Emulation 162
 6.2 Training, Testing and Results 163
 7 Discussion and Conclusions 167
 References 169

Rethinking the Syntactic Burst in Young Children 173
 Christophe Parisse

1 Introduction 173
 2 Assumptions About Children’s Behavior 174
 3 A Testing Procedure in Three Steps 175
 4 Analysis 1 177
 5 Analysis 2 179
 6 Analysis 3 182
 6.1 Results and Discussion: Question 1 183
 6.2 Results and Discussion: Question 2 184
 7 Analysis 4 187
 8 Discussion 188
 Appendix 192
 References 194

Part IV Linking Syntax to Semantics

**Learning to Interpret Novel Noun-Noun Compounds:
 Evidence from Category Learning Experiments** 199
 Barry J. Devereux* and Fintan J. Costello

1 Introduction 200
 2 An Exemplar-Based Account of Compound Interpretation 201

3	Overview of Experiments	202
4	Experiment 1	204
4.1	Method.....	205
4.2	Results	211
4.3	Discussion	215
5	Experiment 2	216
5.1	Method.....	217
5.2	Results	219
5.3	Modelling Relation Selection in Compound Interpretation	223
6	Experiment 3	225
6.1	Method.....	226
6.2	Results	228
6.3	Modelling Relation Selection	231
7	Conclusions.....	231
	References	232
	Child Acquisition of Multiword Verbs: A Computational Investigation...	235
	Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson	
1	Introduction	236
2	Multiword Lexemes with Basic Verbs	238
3	Linguistic Properties and the Usage-Based Measures	239
3.1	Association of a Verb–Noun Pair.....	240
3.2	Semantic Properties of the Noun	241
3.3	Degree of Syntactic Fixedness	242
4	Evaluating the Statistical Measures	243
4.1	Experimental Setup	244
4.2	Measures in Combination: Clustering	245
4.3	Performance of the Individual Measures.....	246
5	Embedding the Measures into a Word Learning Model	247
5.1	The Original Word Learning Model	248
5.2	Learning the Verb–Noun Multiword Lexemes	249
5.3	Experiments on the Modified Word Learner.....	250
6	Conclusions.....	252
	References	253
	Starting from Scratch in Semantic Role Labeling: Early Indirect Supervision	257
	Michael Connor, Cynthia Fisher, and Dan Roth	
1	Introduction.....	257
1.1	Addressing the Ambiguity of Sentences and Scenes: Semantic and Syntactic Bootstrapping.....	258
1.2	How Could Syntactic Bootstrapping Begin?	260
2	BabySRL and Related Computational Models	263
3	Model of Language Acquisition	267
3.1	CHILDES Training Data	268
3.2	Learning Model	269

- 4 Latent Training 274
 - 4.1 Argument, Predicate and Role Classification 277
- 5 Experimental Evaluation 280
 - 5.1 Ambiguous Semantic Feedback 282
- 6 Recovering Argument Knowledge 284
 - 6.1 Bottom-Up Argument Identification 285
 - 6.2 Integrating into Online Latent Classifier 288
- 7 Conclusion 290
- References 293
- Gradual Acquisition of Verb Selectional Preferences in a Bayesian Model** 297

Afra Alishahi and Suzanne Stevenson

 - 1 Introduction 297
 - 1.1 Verb Selectional Preferences 298
 - 1.2 Related Computational Models 299
 - 1.3 Our Approach 301
 - 2 A Computational Model of Learning Verb Selectional Preferences 302
 - 2.1 Learning as Bayesian Clustering 303
 - 2.2 Probabilities of Semantic Properties 304
 - 2.3 Predicting Semantic Profiles for Verbs 305
 - 2.4 Verb-Argument Compatibility 306
 - 3 Experimental Results 306
 - 3.1 The Training Data 307
 - 3.2 Formation of Semantic Profiles for Verbs 307
 - 3.3 Evolution of Verb Semantic Profiles 310
 - 3.4 Verb-Argument Plausibility Judgments 313
 - 4 Discussion and Future Directions 314
 - References 315
- Index** 317

Contributors

Afra Alishahi Department of Communication and Information Studies, Tilburg University, Tilburg, The Netherlands

Zohreh Azimifar Computer Science and Engineering Department, Shiraz University, Shiraz, Iran

Anupam Basu Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Robert C. Berwick Massachusetts Institute of Technology, Cambridge, MA, USA

Rod Byrne Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

Monojit Choudhury Microsoft Research India, Bangalore, Karnataka, India

Michael Connor Department of Computer Science, University of Illinois, Urbana-Champaign, IL, USA

Fintan J. Costello School of Computer Science and Informatics, University College Dublin, Dublin, Ireland

Vivian M. De la Cruz Department of Cognitive Science, University of Messina, Province of Messina, Italy

Barry J. Devereux Centre for Speech, Language and the Brain, Department of Psychology, University of Cambridge, Cambridge, UK

Afsaneh Fazly School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.

Cynthia Fisher Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA

Sandiway Fong University of Arizona Tuscon, AZ, USA

Niloy Ganguly Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Gregory J. Hedlund Department of Linguistics, Memorial University of Newfoundland, St. John's, NL, Canada

Dr. Anna Korhonen Department of Theoretical and Applied Linguistics (DTAL), University of Cambridge, Computer Laboratory, Cambridge, UK

Brian MacWhinney Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Igor Malioutov Massachusetts Institute of Technology, Cambridge, MA, USA

Marco Mazzone Laboratory of Cognitive Science, University of Catania, Catania, Italy

Animesh Mukherjee Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Aida Nematzadeh Department of Computer Science, University of Toronto, Toronto, ON, Canada

Christophe Parisse MoDyCo-INSERM, CNRS, Paris Ouest Nanterre La Défense University, Nanterre cedex, France

Thierry Poibeau LaTTiCe-CNRS, Ecole Normale Supérieure and Université Sorbonne Nouvelle, Paris, France

Alessio Plebe Department Cognitive Science, University of Messina, Province of Messina, Italy

Yvan Rose Department of Linguistics, Memorial University of Newfoundland, St. John's, NL, Canada

Dan Roth Department of Computer Science, University of Illinois, Urbana-Champaign, IL, USA

Suzanne Stevenson Department of Computer Science, University of Toronto, Toronto, ON, Canada

Fatemeh Torabi Asr Computer Science and Engineering Department, Shiraz University, Shiraz, Iran

Aline Villavicencio Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

Laboratory of Information and Decision Systems, Massachusetts Institute of Technology Cambridge, Cambridge, MA, USA

Todd Wareham Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

Beracah Yankama Massachusetts Institute of Technology, Cambridge, MA, USA