

Contents

1	Introduction: Modeling, Learning and Processing of Text-Technological Data Structures	1
	<i>Alexander Mehler, Kai-Uwe Kühnberger, Henning Lobin, Harald Lungen, Angelika Storrer, Andreas Witt</i>	
1.1	Textual Units as Data Structures	1
1.2	Overview of the Book	3
1.2.1	Text Parsing: Data Structures, Architecture and Evaluation	3
1.2.2	Measuring Semantic Distance: Methods, Resources, and Applications	4
1.2.3	From Textual Data to Ontologies, from Ontologies to Textual Data	5
1.2.4	Multidimensional Representations: Solutions for Complex Markup	6
1.2.5	Document Structure Learning	8
1.2.6	Interfacing Textual Data, Ontological Resources and Document Parsing	9
	References	11

Part I: Text Parsing: Data Structures, Architecture and Evaluation

2	The MOTS Workbench	15
	<i>Manfred Stede, Heike Bieler</i>	
2.1	Introduction and Overview	15
2.2	Background: Natural Language Processing Frameworks	17
2.3	The PAULA XML Format	19
2.3.1	PAULA: Logical Structure	20
2.3.2	PAULA: Physical Structure	20

2.4	Processing Pipeline	22
2.4.1	Normalized Input Format: LDOC	23
2.4.2	Tokenization	25
2.4.3	Conversion to PAULA	25
2.4.4	Integrating Analysis Components	25
2.5	Analysis Components	26
2.6	User Interface	28
2.6.1	XML Inline Representation	29
2.6.2	Visualization	30
2.7	Current Developments and Conclusion	31
	References	33
3	Processing Text-Technological Resources in Discourse Parsing	35
	<i>Henning Lobin, Harald Lungen, Mirco Hilbert, Maja Bärenfänger</i>	
3.1	Introduction	35
3.2	Corpus	37
3.3	Architecture	39
3.4	Parser	40
3.4.1	Initialisation	40
3.4.2	Cascade Step	41
3.4.3	Generic Annotation Parser (GAP)	43
3.4.4	Traversing the Chart	52
3.5	Evaluation	52
3.6	Conclusion	54
	References	56

Part II: Measuring Semantic Distance: Methods, Resources, and Applications

4	Semantic Distance Measures with Distributional Profiles of Coarse-Grained Concepts	61
	<i>Graeme Hirst, Saif Mohammad</i>	
4.1	Semantic Distance	61
4.2	Measures of Semantic Distance	63
4.3	A Hybrid Method for Semantic Distance Measures	65
4.4	Evaluation in Monolingual Applications	66
4.5	Extension to Cross-Lingual Applications	69
4.5.1	Method	69
4.5.2	Evaluation	69
4.6	Antonymy and Word Opposition	72
4.6.1	Contrasting Categories	74
4.6.2	Degree of Antonymy	75
4.6.3	Evaluation	75
4.7	Conclusion	76
	References	77

5	Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools	81
	<i>Sonya Nikolova, Jordan Boyd-Graber, Christiane Fellbaum</i>	
5.1	Background and Motivation	81
5.1.1	Aphasia	82
5.1.2	ViVA: Visual Vocabulary for Aphasia	82
5.2	The Design of ViVA	83
5.3	The Organization of Words	84
5.3.1	WORDNET and Evocation	85
5.4	Building the Visual Vocabulary for Aphasia	86
5.4.1	Collecting Inexpensive Ratings from Untrained Annotators	87
5.4.2	Method	87
5.5	Results	88
5.6	Discussion	90
	References	91

Part III: From Textual Data to Ontologies, from Ontologies to Textual Data

6	An Introduction to Hybrid Semantics: The Role of Cognition in Semantic Resources	97
	<i>Alessandro Oltramari</i>	
6.1	Introduction	97
6.2	Statics and Dynamics of Cognition	98
6.3	From Conceptualization to Specification: The Ontological Level	100
6.4	Building 'Hybrid' Semantic Technologies	102
6.5	The Project of a Collaborative Hybrid Semantic Resource	103
6.6	Conclusion	107
	References	108
7	Modal Logic Foundations of Markup Structures in Annotation Systems	111
	<i>Marcus Kracht</i>	
7.1	Introduction	111
7.2	Some Elements of Modal Logic	112
7.3	Classes of Models	115
7.4	Modal Logic and DOMs	117
7.5	XPath	120
7.6	Paths in Dynamic Logic	123
7.7	Conclusion	126
	References	126

8	Adaptation of Ontological Knowledge from Structured Textual Data	129
	<i>Tonio Wandmacher, Ekaterina Ovchinnikova, Uwe Mönnich, Jens Michaelis, Kai-Uwe Kühnberger</i>	
8.1	Introduction	129
8.1.1	Project Context	131
8.2	Theoretical Background	131
8.2.1	Annotation Graphs and Their Logical Representation	131
8.2.2	Ontologies and Description Logics	134
8.3	Automatic Extraction of Ontological Knowledge from Texts	135
8.3.1	Existing Approaches in Ontology Learning	135
8.4	Our Proposal	138
8.5	Axiom Rewriting Procedure	141
8.5.1	Transforming Textual Input into OWL Axioms	141
8.5.2	Discussion	144
8.6	Adaptivity	145
8.6.1	Terminological Inconsistency	146
8.6.2	Adaptation Procedure	147
8.7	Conclusions and Future Work	150
	References	151

Part IV: Multidimensional Representations: Solutions for Complex Markup

9	Ten Problems in the Interpretation of XML Documents	157
	<i>C.M. Sperberg-McQueen, Claus Huitfeldt</i>	
9.1	Background	157
9.1.1	Derivation of Inferences	158
9.1.2	An Example	160
9.2	The Ten Problems	161
9.2.1	The Arity of Statements	161
9.2.2	The Form of Inference Rules	162
9.2.3	Deixis	163
9.2.4	Inheritance	164
9.2.5	Overriding	165
9.2.6	Conflict and Union	167
9.2.7	Milestones	168
9.2.8	The Universe of Discourse	168
9.2.9	Definite Descriptions and Multiple References to the Same Individual	170
9.2.10	Certainty and Responsibility	172
	References	173

10 Markup Infrastructure for the Anaphoric Bank: Supporting Web Collaboration	175
<i>Massimo Poesio, Nils Diewald, Maik Stührenberg, Jon Chamberlain, Daniel Jettka, Daniela Goecke, Udo Kruschwitz</i>	
10.1 Introduction	175
10.2 How Data Is Added to the Anaphoric Bank	177
10.2.1 Filtering Criteria	177
10.2.2 Data That Has Already Been Annotated	178
10.2.3 Using the Expert Annotation Tool	178
10.2.4 Using a Non-expert Annotation Game	181
10.3 Architecture for the Anaphoric Bank	184
10.3.1 SGF – The XML Architecture Part	184
10.3.2 Database Format	190
10.4 Conclusion	192
References	193
11 Integrated Linguistic Annotation Models and Their Application in the Domain of Antecedent Detection	197
<i>Andreas Witt, Maik Stührenberg, Daniela Goecke, Dieter Metzger</i>	
11.1 Introduction	197
11.2 Anaphora Resolution	200
11.3 Logical Document Structure	201
11.3.1 What Is Logical Document Structure?	201
11.3.2 Application of Logical Document Structure for Linguistic Tasks	202
11.3.3 XML-Annotation of Logical Document Structure	204
11.4 Integration of Resources	204
11.4.1 Representation Formats	205
11.4.2 Sekimo Generic Format and XStandoff	207
11.4.3 Antecedent Candidate List	209
11.5 Results of a Corpus Study	210
11.6 Conclusion	214
References	215

Part V: Document Structure Learning

12 Machine Learning for Document Structure Recognition	221
<i>Gerhard Paaß, Iuliu Konya</i>	
12.1 Introduction	221
12.2 Document Analysis for Large-Scale Processing	222
12.2.1 Geometric Layout Analysis	224
12.2.2 Logical Layout Analysis	227
12.2.3 Minimum Spanning Tree-Based Logical Layout Analysis	229
12.2.4 Evaluation	231

12.3	Estimating Document Structure by Conditional Random Fields	236
12.3.1	Basic Model	236
12.3.2	Application of Linear-Chain CRFs to Structure Information Extraction	239
12.3.3	Discriminative Parsing Models	240
12.3.4	Graph-Structured Model	241
12.4	Conclusion	244
	References	245
13	Corpus-Based Structure Mapping of XML Document Corpora: A Reinforcement Learning Based Model	249
	<i>Francis Maes, Ludovic Denoyer, Patrick Gallinari</i>	
13.1	Introduction and Motivation	249
13.2	Related Work	252
13.3	Structure Mapping Task	253
13.3.1	Task Description	253
13.3.2	Notations	254
13.3.3	Complexity of Inferring and Learning Mappings	254
13.4	Reinforcement Learning Based Model	255
13.4.1	Informal Description	255
13.4.2	Deterministic Markov Decision Process	256
13.4.3	Modeling the Policy	258
13.4.4	Evaluating and Learning the Policy	259
13.5	Experiments	260
13.5.1	Datasets	260
13.5.2	Evaluation Measures	262
13.5.3	Comparison to Baseline Models	263
13.5.4	Results	263
13.6	Conclusion	265
	References	265
14	Learning Methods for Graph Models of Document Structure	267
	<i>Peter Geibel, Alexander Mehler, Kai-Uwe Kühnberger</i>	
14.1	Introduction	268
14.2	Directed Generalized Trees	269
14.3	Quantitative Structure Analysis by Example of Websites	272
14.4	Kernel Methods	273
14.4.1	The Soft Tree Kernel for GTs	274
14.4.2	The Soft GT Kernel	278
14.4.3	The INDIGO Context and Label Sequence Kernels	280
14.5	Experiments	286
14.5.1	The Corpus	286
14.5.2	Quantitative Structure Analysis	287
14.5.3	Kernel Methods	291

14.6	Conclusion	293
	References	294
15	Integrating Content and Structure Learning: A Model of Hypertext Zoning and Sounding	299
	<i>Alexander Mehler, Ulli Waltinger</i>	
15.1	Introduction	299
15.1.1	Webgenre Learning in a Two-Level Perspective	300
15.1.2	Thematic-Generic Tracking, Zoning and Sounding	303
15.2	A Two-Level Model of Logical Web Document Structures	305
15.2.1	Hypertext Stage Classifier	306
15.2.2	Hypertext Stage Grammar and Type Classifier	310
15.2.3	Experiments	310
15.3	Thematic-Generic Sounding in the Web	314
15.4	Bounds of Thematic Sounding in Wikipedia	318
15.4.1	Dominator, Successor and Trigger Sets	320
15.4.2	Statistical Moments of Trigger and Dominator Sets	324
15.5	Conclusion	325
	References	326
 Part VI: Interfacing Textual Data, Ontological Resources and Document Parsing		
16	Learning Semantic Relations from Text	333
	<i>Gerhard Heyer</i>	
16.1	Introduction	333
16.2	How Is It Possible to Automatically Process <i>Meaning</i> ?	334
16.3	Some Filters for Semantic Relations	340
16.4	An Architecture for Learning Semantic Relations	342
	References	345
17	Modelling and Processing Wordnets in OWL	347
	<i>Harald Lüngen, Michael Beißwenger, Bianca Selzam, Angelika Storrer</i>	
17.1	Research Context and Motivation	348
17.2	Resources	349
17.2.1	GermaNet	349
17.2.2	TermNet	350
17.2.3	GermaTermNet	351
17.3	Modelling Wordnet-Like Resources in OWL	352
17.3.1	Basic Options	352
17.3.2	Related Work	354
17.3.3	OWL Models for GermaNet, TermNet, and GermaTermNet	355
17.4	Processing WordNet-Like Resources in OWL	363

17.4.1	Processing the OWL Models of TermNet	363
17.4.2	Processing the OWL Full Version of GermaTermNet ...	369
17.5	Conclusion	372
	References	374
18	Exploring Resources for Lexical Chaining: A Comparison of Automated Semantic Relatedness Measures and Human Judgments	377
	<i>Irene Cramer, Tonio Wandmacher, Ulli Waltinger</i>	
18.1	Motivation	377
18.2	Lexical Chaining	378
18.3	Related Work	379
18.3.1	Budanitsky and Hirst	380
18.3.2	Boyd-Graber et al.	381
18.4	Semantic Relatedness Measures	382
18.4.1	Net-Based Measures	382
18.4.2	Distributional Measures	385
18.4.3	Wikipedia-Based Measures	387
18.5	Evaluation	387
18.5.1	Method	387
18.5.2	Results	389
18.5.3	Meta-level Evaluation	392
18.6	Conclusions and Future Work	393
	References	394
	Author Index	397