Contents

1 **Introduction** 1
 1.1 Scope 1
 1.2 Approach 2
 1.3 Prerequisite Knowledge 3
 1.4 Structure 4
 1.5 Learning Paths 5

Part I Searching for Entities in Text

2 **Entities, Labels, and Surface Forms** 9
 2.1 A First Surface Form: Entity Label 10
 2.2 Experiment — Searching for *Beethoven* 11
 2.3 Looking for Additional Surface Forms 13
 2.4 Categorizing various Surface Forms 15
 2.5 Expanding the Set of Surface Forms with a Generative Approach 17
 2.6 Multilingual Surface Forms 18
 2.7 The Most Ambiguous Surface Form: Pronouns 19
 2.8 In Summary 20
 2.9 Further Reading 21
 2.10 Exercises 21

3 **Searching for Named Entities** 23
 3.1 Entity Types 24
 3.2 Gazetteers 25
 3.3 Capturing Regularities in Named Entities: Regular Expressions 27
 3.4 Experiment — Finding DATE Instances in a Corpus 29
  3.4.1 Gold Standard and Inter-Annotator Agreement 29
  3.4.2 Baseline Algorithm: Simple DATE Regular Expression 32
  3.4.3 Refining the DATE Expressions 34

3.5 Language Dependency of Named Entity Expressions 35
3.6 In Summary 36
3.7 Further Reading 36
3.8 Exercises 37

4 Comparing Surface Forms 39
4.1 Edit Distance — A Dynamic Programming Algorithm 40
4.2 From Distance to Binary Classification — Finding Thresholds 43
4.3 Soundex — A Phonetic Algorithm 47
4.4 Experiment — A Comparative Setting for Misspelling Detection 49
4.4.1 Gold Standard 50
4.4.2 Looking at Results 52
4.5 Soundex and Edit Distance in a Multilingual Setting 52
4.6 In Summary 54
4.7 Further Reading 54
4.8 Exercises 54

Part II Working with Corpora

5 Exploring Corpora 59
5.1 Defining Corpora 60
5.2 Characterizing Documents 61
5.3 Corpora Resources 62
5.4 Surface Form Frequencies in Corpus 64
5.5 Information Content 66
5.6 Experiment — Comparing Domain-Specific Corpora 67
5.6.1 Building Domain-Specific Corpora 68
5.6.2 Tokenizing the Corpora 70
5.6.3 Calculating Frequencies and IC 71
5.6.4 Observing and Comparing Results 72
5.7 Mutual Information and Collocations 74
5.8 Viewing Candidate Collocations through a Concordancer 76
5.9 In Summary 79
5.10 Further Reading 80
5.11 Exercises 82

6 Words in Sequence 85
6.1 Introduction to Language Modeling 86
6.2 Estimating Sequence Probabilities 87
6.3 Gathering N-Gram Probabilities from Corpora 91
6.3.1 Building a Domain-Independent Corpus 91
6.3.2 Calculating N-Gram Probabilities 92
6.3.3 Issues in N-Gram Estimation 95

6.4 Application: Spelling Correction 98
6.5 In Summary 101
6.6 Further Reading 102
6.7 Exercises 102

7 **Bilingual Corpora** 105
7.1 Types of Bilingual Corpora 105
7.2 Exploring Noise in Bilingual Corpora 107
7.3 Language Identification 109
7.3.1 Estimating Letter Trigram Probabilities 110
7.3.2 Testing a Letter Trigram Model for Language Identification 111
7.3.3 Language Identification as Preprocessing Step 112
7.4 Searching for Term Equivalents in Parallel Corpora 114
7.4.1 Obtaining and Indexing a Parallel Corpus 116
7.4.2 Adapting PMI for Term Equivalent Search 117
7.5 Experiment — Term Equivalent Search 119
7.5.1 Dataset and a posteriori Evaluation 119
7.5.2 Inter-Annotator Agreement 120
7.5.3 Result Analysis 122
7.6 In Summary 123
7.7 Further Reading 123
7.8 Exercises 124

Part III Semantic Grounding and Relatedness

8 **Linguistic Roles** 129
8.1 Tokenization 130
8.2 Sentence Splitting 133
8.3 Lemmatization 134
8.4 POS Tagging 136
8.5 Constituency Parsing 138
8.6 Experiment — Classifying Groups of Entities 141
8.6.1 Goal of the Experiment 142
8.6.2 Gold Standard and Evaluation Method 143
8.6.3 Our Method: Classification through POS Tagging 143
8.6.4 Performance Evaluation 144
8.6.5 Result Analysis — Intrinsic versus Extrinsic Evaluation 145
8.7 In Summary 147
8.8 Further Reading 148
8.9 Exercises 149

9 **Definition-Based Grounding** 151
9.1 Word Sense Disambiguation 152
9.2 Entity Linking 154

9.3 Bag-of-Words Representation 156
9.4 Bag-of-Words Content — Looking at Text Cohesion 159
9.5 Bag-of-Words Comparison 161
9.6 Grounding Algorithm: BOW-Match 162
9.7 Experiment — Disambiguating *Beethoven*. 163
9.7.1 Grounding Space, Gold Standard, and Evaluation Method. 164
9.7.2 Testing our BOW-Match Algorithm 165
9.7.3 Result Analysis 168
9.8 In Summary 169
9.9 Further Reading. 169
9.10 Exercises 170

10 Relatedness 173
10.1 Building a Co-occurrence Vector 174
10.1.1 Corpus Size. 177
10.1.2 Word Filtering — Linguistic versus Statistic 178
10.1.3 Window Dependency — Positioning 180
10.1.4 Window Dependency — Size 181
10.2 Measuring Relatedness 183
10.2.1 First-Level Relatedness 183
10.2.2 Second-Level Relatedness 184
10.3 Word Embeddings 185
10.4 Experiment 1 — Relatedness for Semantic Similarity 187
10.4.1 Dataset and Evaluation 187
10.4.2 Testing Relatedness Measures and Analyzing Results 189
10.5 Experiment 2 — Relatedness for Entity Linking 193
10.5.1 Dataset and Evaluation 193
10.5.2 Developing a BOW-Similarity Algorithm 194
10.5.3 Result Analysis 195
10.6 In Summary 197
10.7 Further Reading. 198
10.8 Exercices 199

Part IV Knowledge Acquisition

11 Pattern-Based Relation Extraction 205
11.1 Relation Types and Textual Resources 206
11.2 Lexical Patterns 208
11.2.1 Regular Expressions for Lexical Patterns. 209
11.3 Lexico-Syntactic Patterns 211
11.3.1 Regular Expressions for Lexico-Syntactic Patterns 212

11.4 Experiment — Pattern-Based Synonymy Relation Extraction . . . 213
11.4.1 Development Set . . . 213
11.4.2 Defining a Synonym Search Strategy Using Lexical and Lexico-Syntactic Patterns . . . 215
11.4.3 Defining a Gold Standard and Evaluation Method . . . 216
11.4.4 Testing Lexical and Lexico-Syntactic Patterns . . . 217
11.4.5 Result Analysis . . . 218
11.5 Toward a Semi-automatic Process of Knowledge Acquisition . . . 220
11.6 In Summary . . . 225
11.7 Further Reading . . . 226
11.8 Exercises . . . 226

12 From Syntax to Semantics . . . 231
12.1 Dependency Grammars . . . 232
12.2 Semantic Frames . . . 235
12.3 From Sentences to Semantic Frames . . . 237
12.3.1 Automatic Frame Identification . . . 237
12.3.2 Semantic Role Labeling . . . 238
12.4 Semi-automatic Acquisition of Syntactic Realizations . . . 240
12.4.1 Human Annotation Effort — Writing Prototype Sentences . . . 240
12.4.2 Algorithmic Effort — Processing Prototype Sentences into Syntactic Realizations . . . 242
12.5 Experiment — Semantic Role Labeling of *Cooking* Sentences . . . 244
12.5.1 Gold Standard and Evaluation . . . 244
12.5.2 Define our Semantic Role Labeling Strategy . . . 246
12.5.3 Result Analysis and Discussion . . . 247
12.6 Syntactic Realizations as Patterns for Relation Extraction . . . 248
12.7 In Summary . . . 251
12.8 Further Reading . . . 251
12.9 Exercises . . . 252

13 Semantic Types . . . 255
13.1 Common Semantic Types — Exploring NER Systems . . . 256
13.2 Specific Semantic Types — Exploring Lexico-Semantic Resources . . . 257
13.2.1 Semantic Types in FrameNet . . . 257
13.2.2 Semantic Types in WordNet . . . 259
13.3 Building Gazetteers from Textual Resources . . . 260
13.3.1 Lexico-Syntactic Patterns . . . 261
13.3.2 Dependency Patterns . . . 262
13.3.3 Comparing Approaches . . . 263

13.4 Experiment — Using Semantic Types to Filter Semantic Roles 264
13.4.1 Task Definition 265
13.4.2 Defining a Gold Standard 265
13.4.3 Combining Resources and Algorithms within Voting Strategies 266
13.4.4 Evaluation of Voting Strategies 269
13.4.5 Result Analysis and Discussion 270
13.5 In Summary 271
13.6 Further Reading 272
13.7 Exercises 272

Appendix A: A Look into the Semantic Web 275

Appendix B: NLP Tools, Platforms and Resources 281

Appendix C: Relation Lists 283

Glossary 291

References 313