Contents

Abstract v

Kurzfassung vii

1 Introduction 1

I Fundamentals 7

2 Language Communication 11
2.1 Interpersonal Communication 11
2.2 Human-Computer Interaction 13
2.3 Word Meaning 15
2.3.1 Semantic Prototypes and Primitives 16
2.3.2 Word Senses and Concepts 18
2.3.3 Lexical Gaps and Compound Words 20
2.4 Conclusions 21

3 Language Resources 23
3.1 Written Language Corpora 23
3.2 Lexical Resources 24
3.2.1 WordNet 27
3.2.2 EuroWordNet (EWN) 27
3.2.3 MultiWordNet 30
3.2.4 WordNet Criticism 32
3.3 Conclusions 34

4 Text Retrieval 37
4.1 Short History 39
4.2 Query and Document Preprocessing 41
4.2.1 Spelling Correction 42
4.2.2 Stemming 43
4.2.3 Named-Entity Recognition 44
4.2.4 Query Expansion 45
4.2.5 Text Indexing 46
4.3 Retrieval Models 46
4.3.1 Vector Space Model 47

Bibliografische Informationen
http://d-nb.info/990194914

digitalisiert durch DEUTSCHE NATIONAL BIBLIOTHEK

4.3.2 Document Encoding 48
4.4 Performance Measures 49
4.5 Classification and Clustering 50
4.6 Conclusions 54

5 Word Sense Disambiguation **57**
5.1 Short History 57
5.2 Word Sense Granularity 58
5.3 Word Sense Clustering 58
5.4 Word Sense Disambiguation Approaches 59
5.4.1 WSD Corpus-based Approaches 60
5.4.2 WSD Knowledge-based Approaches 61
5.4.3 Evaluating WSD 65
5.5 Word Sense Disambiguation and Information Retrieval 67
5.6 Conclusions 69

II Multilingual Text Retrieval **71**

6 Motivation and Challenges **75**
6.1 Needs for Multilingual Text Retrieval 76
6.2 Query and Document Translation 78
6.3 Interactive Multilingual Information Retrieval 79
6.4 Conclusions 81

7 Multilingual Information Retrieval Systems **83**
7.1 Machine Translation 84
7.2 Corpus-based Approach 86
7.3 Knowledge-based Approach 88
7.3.1 Concept-based Information Retrieval 88
7.3.2 The Dictionary-based Retrieval 91
7.3.3 Thesaurus- and Controlled vocabulary-based Approach 93
7.3.4 Combination of Dictionary-based and Corpus-based Approaches . . 95
7.4 Interactive Multilingual Information Retrieval Systems 96
7.5 Discussion 100

III The Multilingual Sense Folder Approach **103**

8 Sense Folder Disambiguation **107**
8.1 Query and Document Pre-Processing 110
8.1.1 Spelling Correction (Multispell) 110
8.1.2 Stemming (Snowball) 113
8.1.3 Named-Entity Recognition (Stanford NER) 113
8.2 The Search Engine Framework (CARSA) 113
8.3 Semantic Prototype Engine 116

8.4 Sense Folder Classification . . . 118
8.4.1 Indexing the Results Set . . . 119
8.4.2 Disambiguating the Meaning of the Search Terms . . . 119
8.4.3 Multilinguality in Sense Folder Classification . . . 120
8.5 Sense Folder Clustering Methods . . . 121
8.5.1 k-Means Clustering . . . 121
8.5.2 Expectation-Maximization Clustering . . . 122
8.5.3 Density-based Clustering . . . 123
8.6 Merging Word Senses and Merging Methods . . . 124
8.6.1 Merging Word Senses . . . 125
8.6.2 Merging Methods . . . 126
8.7 Conclusions . . . 127

9 Evaluation **129**
9.1 Evaluation of Hand-Tagged Data . . . 129
9.2 BankSearch Data Set Collection Evaluation . . . 132
9.2.1 "rule" and "operation" BankSearch Subsets . . . 133
9.2.2 Linguistic Parameters and Manual Merging Methods . . . 136
9.3 Sense Folder Corpus Evaluation . . . 140
9.3.1 Corpora Analysis . . . 141
9.3.1.1 Senseval and Semcor Data Sets . . . 142
9.3.1.2 IR-SemCor Corpus . . . 143
9.3.1.3 Building the Sense Folder Corpus . . . 143
9.3.2 Baselines . . . 144
9.3.3 Evaluation of the Linguistic Parameters . . . 149
9.3.4 Merging Methods Evaluation . . . 158
9.3.5 Concluding Remarks . . . 163

10 Multilingual User Interfaces **167**
10.1 Multilingual Sense Folder User Interface . . . 167
10.1.1 Interacting with the Multilingual User Interface . . . 168
10.1.2 Search Configuration . . . 170
10.2 The Lexical Restructuring Tool (LexiRes) . . . 172
10.2.1 Related Work . . . 174
10.2.2 User Interface and Interaction . . . 174
10.3 MultiLexExplorer: Combining Multilingual Web Search with Multilingual Lexical Resources . . . 176
10.3.1 Language Learning Software . . . 178
10.3.2 Navigating Multilingual Information . . . 179
10.3.3 Related Work: Tools for Visualization and Navigation of Linguistic Ontologies . . . 180
10.3.4 System Architecture . . . 181
10.4 A Mobile User Interface . . . 185

11 Concluding Remarks **189**

11.1 Open Problems 189
11.2 RDF/OWL EuroWordNet Representation 191

Appendix **193**

A Sense Folders Creation **195**
A.1 Overview 195
A.2 Example for term "network" and its linguistic relations 196

B Merging Evaluation **199**
B.1 Overview 199
B.2 Best Performing Merging Methods and Thresholds 201
B.3 Other Tested Merging Methods and Thresholds 209
B.4 Combination of Best Performing Merging Methods and Thresholds 221