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

Antal van den Bosch • Gosse Bouma
Editors

Interactive Multi-modal Question-Answering

Springer

Editors
Antal van den Bosch
Tilburg center for Cognition and Communication
Tilburg University
School of Humanities
5000 LE Tilburg
The Netherlands
Antal.vdnBosch@uvt.nl

Gosse Bouma
Information Science
University of Groningen
9700 AS Groningen
The Netherlands
g.bouma@rug.nl

ISSN 2192-032X
ISBN 978-3-642-17524-4 e-ISBN 978-3-642-17525-1
DOI 10.1007/978-3-642-17525-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011928374

Copy-editing: CumLingua Language & Communication/www.cumlingua.com

Cover design: deblik, Berlin

Printed on acid-free paper

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

Preface

To start off with a question, what kind of book is this? This book is, in many aspects, a collaborative effort. Its contents evolve from the idea of combining the latest research in fields such as natural language processing, dialogue systems, human-machine interaction, and information extraction, around a single common theme. What is more, the book documents a successful collaborative effort to do just this: the combination of all these fields in one single research area, working towards a common goal—an interactive, multimodal spoken question-answering dialogue system, with which a human user could discuss a particular domain. This concerted effort was known as the "Interactive multimodal Information eXtraction" (IMIX) Programme, a Dutch national research programme funded by the Netherlands Organisation for Scientific Research (NWO), that ran between 2004 and 2009.

During the peak years of IMIX, the researchers in this project together formed a major part of the entire language and speech technology community in the Netherlands. The project also continued to have an influence in the work that followed, for example in the post-doc research tracks of the people who wrote their PhD thesis as part of IMIX. The majority of researchers involved in IMIX can be traced back as authors of the chapters in this book and, although the selection of chapters remains incomplete with respect to all the work performed in IMIX, we are very grateful to the many former IMIX colleagues who were willing to write retrospective overviews of their work and performance in IMIX.

The IMIX project was surrounded by a supportive network of individuals, each of whom deserves a special mention: Alice Dijkstra, Christine Erb, and Brigit van der Pas from NWO were instrumental throughout the project in organising regular IMIX meetings and, more generally, in creating an atmosphere in which everyone was encouraged to maximise performance and collaborate with as few constraints and restrictions as possible. Alice Dijkstra, who herself has a background in computational linguistics, had created a sound foundation by inviting a team of academic strategists, headed by Lou Boves, to write up the most challenging mix of ideas that, within reason, could be expected to be illustrated by the Dutch language and speech technology community, on a limited budget and within the time span of a PhD project. After an intensive assessment proposal period, the IMIX programme committee settled on the project proposals. The end results of most of these projects are covered in this book. The programme committeee appointed a project coordinator, a post occupied from the start by Els den Os, to oversee the creation of an actual running, talking, question-answering piece of demonstrator software. This book also highlights this complicated endeavour, which to some extent resembled putting together a working car from a washing machine, two lawnmowers, and a microwave oven (all ultra modern).

The project was fortunate to have the guidance of an international scientific advisory board: Eduard Hovy, Jon Oberlander, Norbert Reithinger, and Steve Young. From start to finish, their advice gave the project all the right nudges, in the right directions. Their final report now appears as the epilogue of this book. On behalf of the whole IMIX team, we would like to express our gratitude to our international advisers and to everyone who helped make this challenging project possible. For assisting us in the final editorial work, we thank Olga Chiarcos at Springer, and the team of copyeditors from CumLingua.

This book is the final publication from the IMIX project. Its aims to go beyond IMIX; it is intended as a unique look at a mix of state-of-the-art methods, rarely found together in one book (despite the fact that we are all involved in language and speech technology—such is the level of specialisation of most of that we do). But in addition to all the developments and future work born from the ideas orig in IMIX, it gives us tremendous pleasure to have had the privilege to edit this book and provide a final answer to the first question: this book is IMIX.

Tilburg and Groningen Antal van den Bosch and Gosse Bouma

Contents

Part I Introduction to the IMIX Programme

Introduction 3
Antal van den Bosch and Gosse Bouma
1 Interactive Multimodal Question Answering 3
2 The IMIX Project 4
3 Contributions 5
4 Further Reading 7
References 8

The IMIX Demonstrator: An Information Search Assistant for the Medical Domain 11
Dennis Hofs, Boris van Schooten and Rieks op den Akker
1 Introduction 11
2 A Medical Information Search Assistant 12
3 Architecture of the Final Version 14
3.1 The Modules 16
3.2 The DAM State Machine 18
3.3 A Modular Version of the Demonstrator 18
4 Conclusion 19
References 21

Part II Interaction Management

Vidiam: Corpus-based Development of a Dialogue Manager for Multimodal Question Answering 25
Boris van Schooten and Rieks op den Akker
1 Introduction 25
1.1 QA Dialogue System Features 26
2 Overview of Existing Systems 29
2.1 FQ Context Completion Strategies 30
3 The Corpora 35
3.1 The Follow-up Question Corpus 35
3.2 The Multimodal Follow-up Question Corpus 38
3.3 The Ritel Corpus 42
4 The Dialogue Manager 45
4.1 FU Classification Performance 45
4.2 Rewriting and Context Completion Performance 47
4.3 Answering Performance 50
5 Conclusions 53
References 54

Multidimensional Dialogue Management 57
Simon Keizer, Harry Bunt and Volha Petukhova
1 Introduction 57
2 Semantic and Pragmatic Framework: DIT 59
2.1 Dimensions and Communicative Functions 59
3 Multifunctionality 63
3.1 Relations Between Communicative Functions 63
3.2 Types of Multifunctionality in Dialogue Units 66
4 Design of a Multidimensional Dialogue Manager 69
4.1 Context Model 69
4.2 Dialogue Act Agents 70
4.3 Application: Dialogue Management for Interactive QA . . 72
5 Context Specification and Update Mechanisms 74
5.1 Specification of the Context Model 75
5.2 Levels of Processing and Feedback 75
5.3 Grounding 76
5.4 Context Update Model 77
6 Constraints on Generating Combinations of Dialogue Acts 78
6.1 Logical Constraints 78
6.2 Pragmatic Constraints 79
6.3 Constraints for Segment Sequences 80
6.4 Constraints Defining Dialogue Strategies 80
6.5 Evaluation Agent Design 83
7 Conclusion 84
References 85

Part III Fusing Text, Speech, and Images

Experiments in Multimodal Information Presentation 89
Charlotte van Hooijdonk, Wauter Bosma, Emiel Krahmer, Alfons Maes and Mariët Theune
1 Introduction 89
2 Experiment 1: Production of Multimodal Answers 91
2.1 Participants 92
2.2 Stimuli 92
2.3 Coding System and Procedure 92
2.4 Results 93
2.5 Conclusion 96
3 Experiment 2: Evaluation of Multimodal Answers 96
3.1 Participants 96
3.2 Design 97
3.3 Stimuli 97
3.4 Procedure 98
3.5 Results 98
3.6 Conclusion 101
4 Automatic Production of Multimodal Answers 102
4.1 Multimedia Summarization 102
4.2 Automatic Picture Selection 103
5 Experiment 3: Evaluating Automatically Produced Multimodal Answers 104
5.1 Participants 105
5.2 Design 105
5.3 Stimuli 106
5.4 Procedure 107
5.5 Data Processing 108
5.6 Results 108
5.7 Conclusion 112
6 General Discussion 112
References 114

Text-to-Text Generation for Question Answering 117
Wauter Bosma, Erwin Marsi, Emiel Krahmer and Mariët Theune
1 Introduction 117
2 Graph-based Content Selection 119
2.1 Related Work 119
2.2 Task Definition 121
2.3 A Framework for Summarisation 122
2.4 Query-based Summarisation 122
2.5 Results 129
2.6 Validating the Results 130
3 Sentence Fusion 131

3.1 Data Collection and Annotation . . . 133
3.2 Automatic Alignment . . . 137
3.3 Merging and Generation . . . 139
3.4 Discussion . . . 141
4 Conclusion . . . 142
References . . . 143

Part IV Text Analysis for Question Answering

Automatic Extraction of Medical Term Variants from Multilingual Parallel Translations . . . 149
Lonneke van der Plas, Jörg Tiedemann and Ismail Fahmi
1 Introduction . . . 149
2 Alignment-based Methods . . . 153
2.1 Translational Context . . . 153
2.2 Measures for Computing Semantic Similarity . . . 155
2.3 Related Work . . . 156
3 Materials and Methods . . . 158
3.1 The multilingual Parallel Corpus EMEA . . . 159
3.2 Automatic Word Alignment and Phrase Extraction . . . 159
3.3 Selecting Candidate Terms . . . 160
3.4 Comparing Translation Vectors . . . 161
3.5 Post-processing . . . 162
4 Evaluation . . . 162
4.1 Gold Standard . . . 163
4.2 Test Set . . . 163
5 Results and Discussion . . . 163
5.1 Two Methods for Comparison . . . 163
5.2 Results . . . 164
5.3 Error Analysis . . . 166
6 Conclusions . . . 167
References . . . 168

Relation Extraction for Open and Closed Domain Question Answering . . . 171
Gosse Bouma, Ismail Fahmi and Jori Mur
1 Introduction . . . 172
2 Related Work . . . 174
2.1 Relation Extraction for Open Domain QA . . . 174
2.2 Biomedical Relation Extraction . . . 175
2.3 Using Syntactic Patterns . . . 175
3 Dependency Information for Question Answering and Relation Extraction . . . 177
4 Relation Extraction for Open Domain QA . . . 179
4.1 Pattern Induction . . . 180
4.2 Experiment . . . 181
4.3 Evaluation . . . 183

5 Relation Extraction for Medical QA 186
5.1 Multilingual Term Labelling 187
5.2 Learning Patterns 189
5.3 Evaluation 190
5.4 Evaluation in a QA Setting 192
6 Conclusions and Future Work 193
References 195

Constraint-Satisfaction Inference for Entity Recognition 199
Sander Canisius, Antal van den Bosch and Walter Daelemans
1 Introduction 199
2 Sequence Labelling 200
3 Related Work 201
4 A Baseline Approach 202
4.1 Class Trigrams 203
4.2 Memory-based Learning 204
5 Constraint Satisfaction Inference 205
5.1 Solving the CSP 208
6 Sequence Labelling Tasks 208
6.1 Syntactic Chunking 209
6.2 Named-Entity Recognition 210
6.3 Medical Concept Chunking: The IMIX Task 211
7 Experimental Set-up 212
7.1 Evaluation 212
7.2 Constraint Prediction 213
8 Results 214
8.1 Comparison to Alternative Techniques 215
9 Discussion 216
9.1 Other Constraint-based Approaches to Sequence Labelling 217
10 Conclusion 218
References 219

Extraction of Hypernymy Information from Text 223
Erik Tjong Kim Sang, Katja Hofmann and Maarten de Rijke
1 Introduction 223
2 Task and Approach 224
2.1 Task 224
2.2 Natural Language Processing 225
2.3 Collecting Evidence 226
2.4 Evaluation 228
3 Study 1: Comparing Pattern Numbers and Corpus Sizes 229
3.1 Extracting Individual Patterns 229
3.2 Combining Corpus Patterns 230
3.3 Web Query Format 232
3.4 Web Extraction Results 233

3.5 Error Analysis 234
3.6 Discussion 234
4 Study 2: Examining the Effect of Ambiguity 235
4.1 Approach 235
4.2 Experiments and Results 235
4.3 Discussion 237
5 Study 3: Measuring the Effect of Syntactic Processing 238
5.1 Experiments and Results 238
5.2 Result Analysis 240
5.3 Discussion 243
6 Concluding Remarks 243
References 244

Towards a Discourse-driven Taxonomic Inference Model 247
Piroska Lendvai
1 Introduction 247
1.1 Conceptual Taxonomy 249
1.2 Discourse Structure 250
1.3 Semantic Inference 251
1.4 Semi-supervised Harvesting of Lexico-semantic Patterns . . 252
1.5 Related Research in Language Technology 253
2 Exploratory Data 254
2.1 Semantic Annotation Types 254
3 Machine Learning of Taxonomy Identification 256
3.1 Feature Construction 256
3.2 Experimental set-up 256
3.3 Results 257
4 The Taxonomy Inference Model and Textual Entailment 258
5 Extraction of Patterns Involving Medical Concept Types 262
5.1 Masking 262
5.2 Experimental Results 263
6 Closing 265
References 266

Part V Epilogue

IMIX: Good Questions, Promising Answers 271
Eduard Hovy, Jon Oberlander and Norbert Reithinger
1 The Legacy of the IMIX Programme 271
2 Evaluation of the IMIX Programme Work 272
2.1 Technical Evaluation 273
2.2 Programmatic Evaluation 276
2.3 Delivery and Outreach 277
3 Recommendations for the Future 277
References 279