SPOKEN LANGUAGE UNDERSTANDING
SYSTEMS FOR EXTRACTING SEMANTIC INFORMATION FROM SPEECH

Gokhan Tur
Microsoft Speech Labs, Microsoft Research, USA

Renato De Mori
McGill University, Montreal, Canada and University of Avignon, France
Contents

List of Contributors xvii
Foreword xxv
Preface xxix

1 Introduction 1
Gokhan Tur and Renato De Mori
1.1 A Brief History of Spoken Language Understanding 1
1.2 Organization of the Book 4

PART 1 SPOKEN LANGUAGE UNDERSTANDING FOR HUMAN/MACHINE INTERACTIONS

2 History of Knowledge and Processes for Spoken Language Understanding 11
Renato De Mori
2.1 Introduction 11
2.2 Meaning Representation and Sentence Interpretation 12
2.2.1 Meaning Representation Languages 12
2.2.2 Meaning Extraction from Sentences 16
2.3 Knowledge Fragments and Semantic Composition 18
2.3.1 Concept Tags and Knowledge Fragments 19
2.3.2 Composition by Fusion of Fragments 21
2.3.3 Composition by Attachment 23
2.3.4 Composition by Attachment and Inference 24
2.4 Probabilistic Interpretation in SLU Systems 25
2.5 Interpretation with Partial Syntactic Analysis 26
2.6 Classification Models for Interpretation 28
2.7 Advanced Methods and Resources for Semantic Modeling and Interpretation 30
2.8 Recent Systems 32
2.9 Conclusions 35
References 36

References 7
### 3 Semantic Frame-based Spoken Language Understanding 41

*Ye-Yi Wang, Li Deng and Alex Acero*

3.1 Background 41

3.1.1 History of the Frame-based SLU 41

3.1.2 Semantic Representation and Semantic Frame 43

3.1.3 Technical Challenges 45

3.1.4 Standard Data Sets 47

3.1.5 Evaluation Metrics 47

3.2 Knowledge-based Solutions 49

3.2.1 Semantically Enhanced Syntactic Grammars 49

3.2.2 Semantic Grammars 51

3.2.3 Knowledge-based Solutions in Commercial Applications 52

3.3 Data-driven Approaches 54

3.3.1 Generative Models 55

3.3.2 Integrating Knowledge in Statistical Models – A Case Study of the Generative HMM/CFG Composite Model 65

3.3.3 Use of Generative Understanding Models in Speech Recognition 71

3.3.4 Conditional Models 74

3.3.5 Other Data-driven Approaches to SLU 84

3.3.6 Frame-based SLU in Context 86

3.4 Summary 87

References 88

### 4 Intent Determination and Spoken Utterance Classification 93

*Gokhan Tur and Li Deng*

4.1 Background 93

4.2 Task Description 96

4.3 Technical Challenges 97

4.4 Benchmark Data Sets 98

4.5 Evaluation Metrics 98

4.5.1 Direct Metrics 98

4.5.2 Indirect Metrics 99

4.6 Technical Approaches 99

4.6.1 Semantic Representations 100

4.6.2 The HMIHY Way: Using Salient Phrases 101

4.6.3 Vector-state Model 103

4.6.4 Using Discriminative Classifiers 103

4.6.5 Using Prior Knowledge 105

4.6.6 Beyond ASR 1-Best: Using Word Confusion Networks 106

4.6.7 Conditional Understanding Models Used for Discriminative Training of Language Models 108

4.6.8 Phone-based Call Classification 115

4.7 Discussion and Conclusions 115

References 117
5 Voice Search
Ye-Yi Wang, Dong Yu, Yun-Cheng Ju and Alex Acero
5.1 Background
  5.1.1 Voice Search Compared with the Other Spoken Dialogue Technologies
  5.1.2 History of Voice Search
  5.1.3 Technical Challenges
  5.1.4 Data Sets
  5.1.5 Evaluation Metrics
5.2 Technology Review
  5.2.1 Speech Recognition
  5.2.2 Spoken Language Understanding/Search
  5.2.3 Dialogue Management
  5.2.4 Closing the Feedback Loop
5.3 Summary
References

6 Spoken Question Answering
Sophie Rosset, Olivier Galibert and Lori Lamel
6.1 Introduction
6.2 Specific Aspects of Handling Speech in QA Systems
6.3 QA Evaluation Campaigns
  6.3.1 General Presentation
  6.3.2 Question Answering on Speech Transcripts: Evaluation Campaigns
6.4 Question-answering Systems
  6.4.1 General Overview
  6.4.2 Approaches Used in the QAst Campaigns
  6.4.3 QAst Campaign Results
6.5 Projects Integrating Spoken Requests and Question Answering
6.6 Conclusions
References

7 SLU in Commercial and Research Spoken Dialogue Systems
David Suendermann and Roberto Pieraccini
7.1 Why Spoken Dialogue Systems do not have to Understand
7.2 Approaches to SLU for Dialogue Systems
  7.2.1 Rule-based Semantic Grammars
  7.2.2 Statistical SLU
  7.2.3 Dealing with Deficiencies of Speech Recognition and SLU in Dialogue Systems
  7.2.4 Robust Interaction Design and Multiple Levels of Confidence Thresholds
  7.2.5 N-best Lists
  7.2.6 One-step Correction and Mixed Initiative
  7.2.7 Belief Systems
7.3 From Call Flow to POMDP: How Dialogue Management Integrates with SLU
   7.3.1 Rule-based Approaches: Call Flow, Form-filling, Agenda, Call-routing, Inference
   7.3.2 Statistical Dialogue Management: Reinforcement Learning, MDP, POMDP
7.4 Benchmark Projects and Data Sets
   7.4.1 ATIS
   7.4.2 Communicator
   7.4.3 Let's Go!
   7.4.4 Datasets in Commercial Dialogue Systems
7.5 Time is Money: The Relationship between SLU and Overall Dialogue System Performance
   7.5.1 Automation Rate
   7.5.2 Average Handling Time
   7.5.3 Retry Rate and Speech Errors
7.6 Conclusion
References

8 Active Learning
Dilek Hakkani-Tür and Giuseppe Riccardi
8.1 Introduction
8.2 Motivation
   8.2.1 Language Variability
   8.2.2 The Domain Concept Variability
   8.2.3 Noisy Annotation
   8.2.4 The Data Overflow
8.3 Learning Architectures
   8.3.1 Passive Learning
   8.3.2 Active Learning
8.4 Active Learning Methods
   8.4.1 The Statistical Framework
   8.4.2 Certainty-based Active Learning Methods
   8.4.3 Committee-based Active Learning
   8.4.4 Density-based Active Learning
   8.4.5 Stopping Criteria for Active Learning
8.5 Combining Active Learning with Semi-supervised Learning
8.6 Applications
   8.6.1 Automatic Speech Recognition
   8.6.2 Intent Determination
   8.6.3 Concept Segmentation/Labeling
   8.6.4 Dialogue Act Tagging
8.7 Evaluation of Active Learning Methods
8.8 Discussion and Conclusions
References
PART 2  SPOKEN LANGUAGE UNDERSTANDING FOR HUMAN/HUMAN CONVERSATIONS

9 Human/Human Conversation Understanding 227
Gokhan Tur and Dilek Hakkani-Tür
  9.1 Background 227
  9.2 Human/Human Conversation Understanding Tasks 229
  9.3 Dialogue Act Segmentation and Tagging 231
    9.3.1 Annotation Schema 232
    9.3.2 Modeling Dialogue Act Tagging 236
    9.3.3 Dialogue Act Segmentation 237
    9.3.4 Joint Modeling of Dialogue Act Segmentation and Tagging 239
  9.4 Action Item and Decision Detection 240
  9.5 Addressee Detection and Co-reference Resolution 242
  9.6 Hot Spot Detection 244
  9.7 Subjectivity, Sentiment, and Opinion Detection 244
  9.8 Speaker Role Detection 245
  9.9 Modeling Dominance 247
  9.10 Argument Diagramming 247
  9.11 Discussion and Conclusions 250
  References 251

10 Named Entity Recognition 257
Frédéric Béchet
  10.1 Task Description 258
    10.1.1 What is a Named Entity? 258
    10.1.2 What are the Main Issues in the NER Task? 260
    10.1.3 Applicative Frameworks of NER in Speech 261
  10.2 Challenges Using Speech Input 263
  10.3 Benchmark Data Sets, Applications 265
    10.3.1 NER as an IE Task 265
    10.3.2 NER as an SLU Task in a Spoken Dialogue Context 266
  10.4 Evaluation Metrics 266
    10.4.1 Aligning the Reference and Hypothesis NE Annotations 267
    10.4.2 Scoring 267
  10.5 Main Approaches for Extracting NEs from Text 269
    10.5.1 Rules and Grammars 269
    10.5.2 NER as a Word Tagging Problem 270
    10.5.3 Hidden Markov Model 271
    10.5.4 Maximum Entropy 273
    10.5.5 Conditional Random Field 274
    10.5.6 Sample Classification Methods 275
    10.5.7 Conclusions on the Methods for NER from Text 276
  10.6 Comparative Methods for NER from Speech 277
    10.6.1 Adapting NER Systems to ASR Output 277
    10.6.2 Integrating ASR and NER Processes 281
10.7 New Trends in NER from Speech
   10.7.1 Adapting the ASR Lexicon 284
   10.7.2 Collecting Data on the ASR Lexicon 285
   10.7.3 Toward an Open-vocabulary ASR System for NER from Speech 286
10.8 Conclusions 287

References 287

11 Topic Segmentation
Matthew Purver

11.1 Task Description 291
   11.1.1 Introduction 291
   11.1.2 What is a Topic? 292
   11.1.3 Linear versus Hierarchical Segmentation 292
11.2 Basic Approaches, and the Challenge of Speech 293
   11.2.1 Changes in Content 293
   11.2.2 Distinctive Boundary Features 294
   11.2.3 Monologue 294
   11.2.4 Dialogue 295
11.3 Applications and Benchmark Datasets 295
   11.3.1 Monologue 296
   11.3.2 Dialogue 296
11.4 Evaluation Metrics 297
   11.4.1 Classification-based 297
   11.4.2 Segmentation-based 298
   11.4.3 Content-based 302
11.5 Technical Approaches 302
   11.5.1 Changes in Lexical Similarity 302
   11.5.2 Similarity-based Clustering 305
   11.5.3 Generative Models 306
   11.5.4 Discriminative Boundary Detection 310
   11.5.5 Combined Approaches, and the State of the Art 310
11.6 New Trends and Future Directions 313
   11.6.1 Multi-modality 313
   11.6.2 Topic Identification and Adaptation 313

References 314

12 Topic Identification
Timothy J. Hazen

12.1 Task Description 319
   12.1.1 What is Topic Identification? 319
   12.1.2 What are Topics? 320
   12.1.3 How is Topic Relevancy Defined? 321
   12.1.4 Characterizing the Constraints on Topic ID Tasks 321
   12.1.5 Text-based Topic Identification 323
12.2 Challenges Using Speech Input 323
  12.2.1 The Naive Approach to Speech-based Topic ID 323
  12.2.2 Challenges of Extemporaneous Speech 323
  12.2.3 Challenges of Imperfect Speech Recognition 324
  12.2.4 Challenges of Unconstrained Domains 325
12.3 Applications and Benchmark Tasks 326
  12.3.1 The TDT Project 326
  12.3.2 The Switchboard and Fisher Corpora 327
  12.3.3 Customer Service/Call Routing Applications 327
12.4 Evaluation Metrics 328
  12.4.1 Topic Scoring 328
  12.4.2 Classification Error Rate 328
  12.4.3 Detection-based Evaluation Metrics 328
12.5 Technical Approaches 333
  12.5.1 Topic ID System Overview 333
  12.5.2 Automatic Speech Recognition 333
  12.5.3 Feature Extraction 334
  12.5.4 Feature Selection and Transformation 335
  12.5.5 Latent Concept Modeling 340
  12.5.6 Topic ID Classification and Detection 343
  12.5.7 Example Topic ID Results on the Fisher Corpus 346
  12.5.8 Novel Topic Detection 350
  12.5.9 Topic Clustering 350
12.6 New Trends and Future Directions 352
References 353

13 Speech Summarization 357
  Yang Liu and Dilek Hakkani-Tür 357

13.1 Task Description 357
  13.1.1 General Definition of Summarization 357
  13.1.2 Speech Summarization 359
  13.1.3 Applications 361
13.2 Challenges when Using Speech Input 362
  13.2.1 Automatic Speech Recognition Errors 363
  13.2.2 Speaker Turns 363
  13.2.3 Sentence Boundaries 363
  13.2.4 Disfluencies and Ungrammatical Utterances 364
  13.2.5 Other Style and Structural Information 365
13.3 Data Sets 366
  13.3.1 Broadcast News (BN) 367
  13.3.2 Lectures 368
  13.3.3 Multi-party Conversational Speech 369
  13.3.4 Voice Mail 371
13.4 Evaluation Metrics 371
  13.4.1 Recall, Precision, and F-measure 372
  13.4.2 ROUGE 372
13.4.3 The Pyramid Method 373
13.4.4 Weighted Precision 374
13.4.5 SumACCY and Weighted SumACCY 374
13.4.6 Human Evaluation 375
13.4.7 Issues and Discussions 375
13.5 General Approaches 375
  13.5.1 Extractive Summarization: Unsupervised Methods 376
  13.5.2 Extractive Summarization: Supervised Learning Methods 381
  13.5.3 Moving Beyond Generic Extractive Summarization 385
  13.5.4 Summary 386
13.6 More Discussions on Speech versus Text Summarization 386
  13.6.1 Speech Recognition Errors 386
  13.6.2 Sentence Segmentation 388
  13.6.3 Disfluencies 389
  13.6.4 Acoustic/Prosodic and Other Speech Features 390
13.7 Conclusions 391
References 392

14 Speech Analytics 397
  I. Dan Melamed and Mazin Gilbert
  14.1 Introduction 397
  14.2 System Architecture 398
  14.3 Speech Transcription 401
  14.4 Text Feature Extraction 402
  14.5 Acoustic Feature Extraction 403
  14.6 Relational Feature Extraction 405
  14.7 DBMS 405
  14.8 Media Server and Player 408
  14.9 Trend Analysis 409
  14.10 Alerting System 413
  14.11 Conclusion 414
References 415

15 Speech Retrieval 417
  Ciprian Chelba, Timothy J. Hazen, Bhuvana Ramabhadran and Murat Saraçlar
  15.1 Task Description 417
    15.1.1 Spoken Document Retrieval 417
    15.1.2 Spoken Utterance Retrieval 418
    15.1.3 Spoken Term Detection 418
    15.1.4 Browsing 418
  15.2 Applications 418
    15.2.1 Broadcast News 419
    15.2.2 Academic Lectures 419
    15.2.3 Sign Language Video 419
    15.2.4 Historical Interviews 420
    15.2.5 General Web Video 420
15.3 Challenges Using Speech Input 420
  15.3.1 Overview 420
  15.3.2 Coping with ASR Errors Using Lattices 421
  15.3.3 Out-of-vocabulary Words 422
  15.3.4 Morphologically Rich Languages 423
  15.3.5 Resource-limited Languages and Dialects 423
15.4 Evaluation Metrics 424
15.5 Benchmark Data Sets 425
  15.5.1 TREC 425
  15.5.2 NIST STD 426
15.6 Approaches 426
  15.6.1 Basic SDR Approaches 426
  15.6.2 Basic STD Approaches 428
  15.6.3 Using Sub-word Units 430
  15.6.4 Using Lattices 432
  15.6.5 Hybrid and Combination Methods 434
  15.6.6 Determining Thresholds 435
  15.6.7 Presentation and Browsing 437
  15.6.8 Other Previous Work 438
15.7 New Trends 439
  15.7.1 Indexing and Retrieval for very Large Corpora 439
  15.7.2 Query by Example 441
  15.7.3 Optimizing Evaluation Performance 442
  15.7.4 Multilingual Speech Retrieval 443
15.8 Discussion and Conclusions 443
References 444
Index 447