HUMAN LANGUAGE TECHNOLOGY

Proceedings of a Workshop held at
Plainsboro, New Jersey

March 8-11, 1994

Sponsored by:
Advanced Research Projects Agency
Software & Intelligent Systems Technology Office

This document contains copies of reports prepared for the ARPA Human Language Technology Workshop. Included are reports from ARPA sponsored programs and other materials prepared for use at the workshop.

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the United States Government.

Distributed by:
Morgan Kaufmann Publishers, Inc.
340 Pine Street, 6th Floor
San Francisco, CA 94104
ISBN 1-55860-357-3
Printed in the United States of America

TABLE OF CONTENTS

Page

Author Index x

Overview of the ARPA Human Language Technology Workshop
Clifford Weinstein, Chair 3

SESSION 1: Lexicons, Corpora, and Evaluation

Session Summary, George A. Miller, Chair 7

The Comlex Syntax Project: The First Year
Catherine Macleod, Ralph Grishman, Adam Meyers: New York University 8

Lexicons for Human Language Technology
Mark Liberman: Linguistic Data Consortium, University of Pennsylvania 13

Multilingual Text Resources at the Linguistic Data Consortium
David Graff and Rebecca Finch: Linguistic Data Consortium,
University of Pennsylvania 18

Multilingual Speech Databases at LDC
John J. Godfrey: Linguistic Data Consortium, University of Pennsylvania 23

Macrophone: An American English Telephone Speech Corpus
Kelsey Taussig, Jared Bernstein: SRI International 27

Corpus Development Activities at the Center for Spoken Language Understanding
Ron Cole, Mike Noel, Daniel C. Burnett, Mark Fanty, Terri Lander, Beatrice Oshika,
Stephen Sutton: Oregon Graduate Institute of Science and Technology 31

The Hub and Spoke Paradigm for CSR Evaluation
Francis Kubala: BBN Systems and Technologies; Jerome Bellegarda: IBM T.J. Watson
Research Center; Jordan Cohen: Institute for Defense Analyses; David Pallett:
National Institute of Standards and Technology; Doug Paul: MIT Lincoln Laboratory;
Mike Phillips: MIT Laboratory for Computer Science; Raja Rajasekaran: Texas Instruments;
Fred Richardson: Boston University; Michael Riley: AT&T Bell Laboratories;
Roni Rosenfeld: Carnegie Mellon University; Bob Roth: Dragon Systems, Inc.;
Mitch Weintraub: SRI International 37

Expanding the Scope of the ATIS Task: The ATIS-3 Corpus
Deborah A. Dahl, contact: Unisys Corporation; Madeleine Bates, Michael Brown,
William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky,
Elizabeth Shriberg 43

1993 Benchmark Tests for the ARPA Spoken Language Program
David S. Pallett, Jonathan G. Fiscus, William M. Fisher, John S. Garofolo, Bruce A. Lund,
Mark A. Przybocki: National Institute of Standards and Technology 49

Page

SESSION 2: Language Modeling

Session Summary, Xuedong Huang, Chair 75

A Hybrid Approach to Adaptive Statistical Language Modeling
Ronald Rosenfeld: Carnegie Mellon University 76

Language Modeling with Sentence-Level Mixtures
Rukmini Iyer, Mari Ostendorf, J. Robin Rohlicek: Boston University.... 82

Speech Recognition Using a Stochastic Language Model Integrating Local and Global Constraints
Ryosuke Isotani, Shoichi Matsunaga: ATR Interpreting Telecommunications Research Laboratories 88

On Using Written Language Training Data for Spoken Language Modeling
R. Schwartz, L. Nguyen, F. Kubala, G. Chou, G. Zavaliagkos, J. Makhoul:
BBN Systems and Technologies.... 94

SESSION 3: Human Language Evaluation

Session Summary, Lynette Hirschman, Chair 99

Towards Better NLP System Evaluation
Karen Sparck Jones: Cambridge University 102

Automatic Evaluation of Computer Generated Text: A Progress Report on the TextEval Project
Chris Brew, Henry S. Thompson: University of Edinburgh....108

The Penn Treebank: Annotating Predicate Argument Structure
Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies,
Mark Ferguson, Karen Katz, Britta Schasberger: University of Pennsylvania....114

Whither Written Language Evaluation?
Ralph Grishman: New York University....120

Semantic Evaluation for Spoken-Language Systems
Robert C. Moore: SRI International....126

SESSION 4: Machine Translation

Session Summary, Eduard Hovy, Chair.... 133

Evaluation in the ARPA Machine Translation Program: 1993 Methodology
John S. White, Theresa A. O'Connell: PRC Inc....135

Page

Building Japanese-English Dictionary Based on Ontology for Machine Translation
Akitoshi Okumura, Eduard Hovy: USC/Information Sciences Institute 141

Toward Multi-Engine Machine Translation
Sergei Nirenburg, Robert Frederking: Carnegie Mellon University 147

Translating Collocations for Use in Bilingual Lexicons
Frank Smadja, Kathleen McKeown: Columbia University 152

The Candide System for Machine Translation
Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, Lubos Ures: IBM T. J. Watson Research Center 157

The Automatic Component of the LINGSTAT Machine-Aided Translation System
Jonathan Yamron, James Cant, Anne Demedts, Taiko Dietzel, and Yoshiko Ito: Dragon Systems, Inc. 163

SESSION 5: Natural Language, Discourse

Session Summary, Paul S. Jacobs, Chair 169

Issues and Methodology for Template Design for Information Extraction
Boyan Onyshkevych: Department of Defense 171

Principles of Template Design
Jerry Hobbs, David Israel: SRI International 177

Pattern Matching in a Linguistically-Motivated Text Understanding System
Damaris M. Ayuso and the PLUM Research Group: BBN Systems and Technologies 182

Tagging Speech Repairs
Peter A. Heeman, James Allen: University of Rochester 187

Information Based Intonation Synthesis
Scott Prevost, Mark Steedman: University of Pennsylvania 193

SESSION 6: Spoken Language Systems

Session Summary, Madeleine Bates, Chair 199

PEGASUS: A Spoken Language Interface for On-Line Air Travel Planning
Victor Zue, Stephanie Seneff, Joseph Polifroni, Michael Phillips, Christine Pao, David Goddeau, James Glass, Eric Brill: MIT Laboratory for Computer Science 201

Recent Developments in the Experimental "Waxholm" Dialog System
Rolf Carlson: KTH 207

Page

Recent Improvements in the CMU Spoken Language Understanding System
Wayne Ward, Sunil Issar: Carnegie Mellon University 213

Combining Knowledge Sources to Reorder N-Best Speech Hypothesis Lists
Manny Rayner, David Carter, Vassilios Digalakis, Patti Price: SRI International 217

Predicting and Managing Spoken Disfluencies During Human-Computer Interaction
Sharon Oviatt: SRI International 222

Integrating Techniques for Phrase Extraction from Speech
Marie Meteer, J. Robin Rohlicek: BBN Systems and Technologies 228

SESSION 7: Demonstrations

Session Summary, Victor Abrash, Chair 235

A Prototype Reading Coach that Listens: Summary of Project LISTEN
Alex Hauptmann, Jack Mostow, Steven F. Roth, Matthew Kane, Adam Swift:
Carnegie Mellon University 237

SESSION 8: Statistical and Learning Methods I

Session Summary, Fred Jelinek, Chair 239

Using a Semantic Concordance for Sense Identification
George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock,
Robert G. Thomas: Princeton University 240

A New Approach to Word Sense Disambiguation
Rebecca Bruce, Janyce Wiebe: New Mexico State University 244

A Maximum Entropy Model for Prepositional Phrase Attachment
Adwait Ratnaparkhi, Jeff Reynar, Salim Roukos: IBM T. J. Watson Research Center 250

A Report of Recent Progress in Transformation-Based Error-Driven Learning
Eric Brill: MIT Laboratory for Computer Science 256

Weighted Rational Transductions and their Application to Human Language Processing
Fernando Pereira, Michael Riley, Richard Sproat: AT&T Bell Laboratories 262

SESSION 9: Statistical and Learning Methods II

Automatic Grammar Acquisition
Scott Miller, Heidi J. Fox: BBN Systems and Technologies 268

Page

Decision Tree Parsing Using a Hidden Derivation Model
F. Jelinek, J. Lafferty, D. Magerman, R. Mercer, A. Ratnaparkhi, S. Roukos:
IBM T. J. Watson Research Center 272

Statistical Language Processing Using Hidden Understanding Models
Scott Miller, Richard Schwartz, Robert Bobrow, Robert Ingria:
BBN Systems and Technologies 278

Japanese Word Segmentation by Hidden Markov Model
Constantine P. Papageorgiou: BBN Systems and Technologies 283

Phonological Parsing for Bi-directional Letter-to-Sound/Sound-to-Letter Generation
Helen M. Meng, Stephanie Seneff, Victor Zue: MIT Laboratory for Computer Science 289

SESSION 10: Government Panel

Session Summary, Oscar N. Garcia, Chair 295

One DODer's View of ARPA Spoken Language Directions
Calvin Olano: Department of Defense 297

Speech and Human Language Technology at the Naval Research Laboratory
Helen M. Gigley: Naval Research Laboratory 300

Advanced Technologies for Language Learning: Developments at ARI
Melissa Holland: U.S. Army Research Institute 302

Language Processing R&D Programmes Directorate XIII E of the European Commission
Roberto Cencioni, Nino Varile: European Commission 303

SESSION 11: Acoustic Modeling and Robust CSR

Session Summary, Steve Young, Chair 305

Tree-Based State Tying for High Accuracy Acoustic Modelling
S. J. Young, J. J. Odell, P. C. Woodland: Cambridge University 307

High-Accuracy Large-Vocabulary Speech Recognition Using Mixture Tying and Consistency Modeling
Vassilios Digalakis, Hy Murveit: SRI International 313

The LIMSI Continuous Speech Dictation System
J. L. Gauvain, L. F. Lamel, G. Adda, M. Adda-Decker: LIMSI-CNRS 319

Adaptation to New Microphones Using Tied-Mixture Normalization
Anastasios Anastasakos, Francis Kubala, John Makhoul, Richard Schwartz:
BBN Systems and Technologies 325

Page

Signal Processing for Robust Speech Recognition
Fu-Hua Liu, Pedro J. Moreno, Richard M. Stern, Alejandro Acero:
Carnegie Mellon University 330

Microphone-Independent Robust Signal Processing Using Probabilistic Optimum Filtering
Leonardo Neumeyer, Mitchel Weintraub: SRI International 336

Microphone Arrays and Neural Networks for Robust Speech Recognition
C. Che, Q. Lin, J. Pearson*, B. de Vries*, J. Flanagan: Rutgers Univerisity;
*David Sarnoff Research Center 342

SESSION 12: Information Retrieval

Session Summary, Donna Harman, Chair 349

Overview of the Second Text Retrieval Conference (TREC-2)
Donna Harman: National Institute of Standards and Technology 351

Learning from Relevant Documents in Large Scale Routing Retrieval
K. L. Kwok, L. Grunfeld: Queens College 358

Document Representation in Natural Language Text Retrieval
Tomek Strzalkowski: New York University 364

Assessing the Retrieval Effectiveness of a Speech Retrieval System by Simulating Recognition Errors
Peter Schäuble, Ulrike Glavitsch: Swiss Federal Institute of Technology (ETH) 370

Speech-Based Retrieval Using Semantic Co-Occurrence Filtering
Julian Kupiec, Don Kimber, Vijay Balasubramanian:
Xerox Palo Alto Research Center 373

(Almost) Automatic Semantic Feature Extraction from Technical Text
Rajeev Agarwal: Mississippi State University 378

SESSION 13: CSR Search

Session Summary, Richard Schwartz, Chair 385

A Large-Vocabulary Continuous Speech Recognition Algorithm and Its Application to a Multi-Modal Telephone Directory Assistance System
Yasuhiro Minami, Kiyohiro Shikano, Osamu Yoshioka, Satoshi Takahashi,
Tomokazu Yamada, Sadaoki Furui: NTT Human Interface Laboratories 387

Techniques to Achieve an Accurate Real-Time Large-Vocabulary Speech Recognition System
Hy Murveit, Peter Monaco, Vassilios Digalakis, John Butzberger:
SRI International 393

Page

The Lincoln Large-Vocabulary Stack-Decoder Based HMM CSR
Douglas B. Paul: MIT Lincoln Laboratory 399

A One Pass Decoder Design for Large Vocabulary Recognition
J. J. Odell, V. Valtchev, P. C. Woodland, S. J. Young: Cambridge University 405

Is N-Best Dead?
Long Nguyen, Richard Schwartz, Ying Zhao, George Zavaliagkos:
BBN Systems and Technologies 411

SESSION 14: New Directions/Applications

Session Summary, Richard Stern, Chair 415

Advanced Human-Computer Interface and Voice Processing Applications in Space
Julie Payette: Canadian Space Agency 416

Integrated Text and Image Understanding for Document Understanding
Suzanne Liebowitz Taylor, Deborah A. Dahl, Mark Lipshutz, Carl Weir,
Lewis M. Norton, Roslyn Nilson, Marcia Linebarger: Unisys Corporation 421

Use of Lexical and Syntactic Techniques in Recognizing Handwritten Text
Rohini K. Srihari: SUNY at Buffalo 427

On-Line Cursive Handwriting Recognition Using Hidden Markov Models and Statistical Grammars
John Makhoul, Thad Starner, Richard Schwartz, George Chou:
BBN Systems and Technologies 432

Language Identification via Large Vocabulary Speaker Independent Continuous Speech Recognition
Steve Lowe, Anne Demedts, Larry Gillick, Mark Mandel, Barbara Peskin:
Dragon Systems, Inc 437

PROJECT SUMMARIES 443

AUTHOR INDEX

Abrash, V. 235
Acero, A. 330
Adda, G. 319
Adda-Decker, M. 319
Agarwal, R. 378
Allen, J. 187, 477
Anastasakos, A. 325
Ayuso, D. 182
Baker, J. 454
Balasubramanian, V. 373
Bates, M. 43, 199, 447
Bellegarda, J. 37
Berger, A. 157
Bernstein, J. 27, 471, 474
Bies, A. 114
Bobrow, R. 278
Boggess, L. 458
Brew, C. 108
Brill, E. 201, 256
Brown, M. 43
Brown, P. 157
Bruce, R. 244
Burnett, D. 31
Butzberger, J. 393
Cant, J. 163, 453
Carlson, R. 207
Carter, D. 217
Cencioni, R. 303
Che, C. 342
Chodorow, M. 240
Chou, G. 94, 432
Cohen, J. 37
Cohen, M. 472
Cole, R. 31
Cowie, J. 464
Dahl, D. 43, 421
Danielson, D. 474
de Vries, B. 342, 470
Della Pietra, S. 157, 457
Della Pietra, V. 157, 457
Demedts, A. 163, 437, 453
Dietzel, T. 163
Digalakis, V. 217, 313, 393, 473
Fanty, M. 31
Farwell, D. 465
Ferguson, M. 114
Finch, R. 18
Fiscus, J. 49
Fisher, W. 43, 49
Flanagan, J. 342, 470
Fox, H. 268
Frederking, R. 147
Furui, S. 387
Garcia, O. 295
Garofolo, J. 49
Gauvain, J. 319
Gigley, H. 300
Gillett, J. 157
Gillick, L. 437, 454
Glass, J. 201
Glavitsch, U. 370
Goddeau, D. 201
Godfrey, J. 23
Graff, D. 18
Grishman, R. 8, 120, 466
Grunfeld, L. 358

Guthrie, L. 463
Harman, D. 349, 351
Hauptmann, A. 237
Heeman, P. 187
Hirschman, L. 99
Hobbs, J. 177, 475
Hodges, J. 458
Holland, M. 302
Hovy, E. 133, 141, 478
Huang, X. 75
Hufnagel, S. 448
Hunicke-Smith, K. 43, 471
Ingria, R. 278
Isotani, R. 88
Israel, D. 177
Issar, S. 213
Ito, Y. 163, 453
Iyer, R. 82
Jacobs, P. 169, 455
Jelinek, F. 239, 272
Joshi, A. 476
Kane, M. 237
Katz, K. 114
Kim, G. 114
Kimber, D. 373
Kubala, F. 37, 94, 325
Kupiec, J. 373
Kwok, K. 358, 469
Lafferty, J. 157, 272
Lamel, L. 319
Lander, T. 31
Landes, S. 240
Leacock, C. 240
Liberman, M. 13
Liebowitz Taylor, S. 421
Lin, Q. 342, 470
Linebarger, M. 421
Lipshutz, M. 421
Liu, F. 330
Lowe, S. 437
Lund, B. 49
MacIntyre, R. 114
Macleod, C. 8
Magerman, D. 272
Makhoul, J. 94, 325, 432, 445, 447
Mandel, M. 437
Marcinkiewicz, M. 114
Marcus, M. 114, 476
Matsunaga, S. 88
McKeown, K. 152, 452
Meng, H. 289
Mercer, R. 157, 272
Meteer, M. 228
Meyers, A. 8
Miller, G. 7, 240, 468
Miller, S. 268, 278
Minami, Y. 387
Monaco, P. 393, 473
Moore, R. 126, 472
Moreno, P. 330
Mostow, J. 237
Murveit, H. 313, 393, 473
Neumeyer, L. 336
Nguyen, L. 94, 411
Nilson, R. 421
Nirenburg, S. 147, 450
Noel, M. 31
Norton, L. 421
O'Connell, T. 135
Odell, J. 307, 405
Okumura, A. 141
Olano, C. 297

Onyshkevych, B 171
Oshika, B. 31
Ostendorf, M. 82, 448, 449
Oviatt, S. 222
Pallett, D. 37, 43, 49, 461
Pao, C. 43, 201
Papageorgiou, C. 283
Passonneau, R. 452
Paul, D. 37, 399, 459
Payette, J 416
Pearson, J. 342, 470
Pereira, F 262
Peskin, B 437
Phillips, M. 201
Polifroni, J. 201
Prevost, S 193
Price, P. 217, 448
Printz, H. 157
Przybocki, M. 49
Pustejovsky, J. 464
Rajasekaran, R 37
Ratnaparkhi, A. 250, 272
Rayner, M. 217
Reddy, R. 451
Reynar, J. 250
Richardson, F. 37
Riley, M. 37, 262
Rohlicek, J. 82, 228, 449
Rosenfeld, R. 37, 76
Roth, R. 37, 454
Roth, S. 237
Roukos, S. 250, 272, 456
Rudnicky, A 43
Schasberger, B. 114
Schäuble, P 370
Schubert, L. 477
Schwartz, R. 94, 278, 325, 385, 411, 432, 445
Seneff, S. 201, 289
Shikano, K. 387
Shriberg, E. 43
Smadja, F. 152
Sparck Jones, K. 102
Sproat, R 262
Srihari, R. 427
Starner, T. 432
Steedman, M. 193, 476
Stern, R. 330, 415
Strzalkowski, T. 364, 467
Sundheim, B. 462
Sutton, S 31
Swift, A. 237
Takahashi, S. 387
Taussig, K. 27
Thomas, R 240
Thompson, H. 108
Ures, L 157
Valtchev, V 405
Varile, N. 303
Ward, W. 213
Webber, B 476
Weinstein, C. 3, 459
Weintraub, M. 37, 336, 473
Weir, C. 421
Weischedel, R 446
White, J. 135
Wiebe, J. 244
Wilks, Y. 464
Woodland, P 307, 405
Yamada, T. 387
Yamron, J 163, 453
Yoshioka, O 387

Young, S. 305, 307, 405

Zavaliagkos, G. 94, 411

Zhao, Y. 411

Zue, V. 201, 289, 460