Lecture Notes in Artificial Intelligence 2448

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Petr Sojka Ivan Kopeček
Karel Pala (Eds.)

Text, Speech and Dialogue

5th International Conference, TSD 2002
Brno, Czech Republic, September 9-12, 2002
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Petr Sojka
Masaryk University, Faculty of Informatics
Department of Programming Systems and Communication
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: sojka@informatics.muni.cz

Ivan Kopeček
Karel Pala
Masaryk University, Faculty of Informatics
Department of Information Technologies
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {kopecek,pala}@informatics.muni.cz

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Text, speech and dialogue : 5th international conference ; proceedings /
TSD 2002, Brno, Czech Republic, September 9 - 12, 2002. Petr Sojka ... (ed.). -
Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ;
Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2448 : Lecture notes in
artificial intelligence)
ISBN 3-540-44129-8

CR Subject Classification (1998): I.2.7, H.3, H.4, I.7

ISSN 0302-9743
ISBN 3-540-44129-8 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

© Springer-Verlag Berlin Heidelberg 2002
Printed in Germany

Typesetting: Camera-ready by Petr Sojka from source files by respective authors,
data conversion by Petr Sojka and PTP-Berlin, Stefan Sossna e.K.
Printed on acid-free paper SPIN: 10871089 06/3142 5 4 3 2 1 0

Preface

The Text, Speech and Dialogue (TSD) Conference 2002, it should be noticed, is now being held for the fifth time and we are pleased to observe that in its short history it has turned out to be an international forum successfully intertwining the basic fields of NLP. It is our strong hope that the conference contributes to a better understanding between researchers from the various areas and promotes more intensive mutual cooperation. So far the communication between man and computers has displayed a one-way nature, humans have to know how the machines work and only then can they "understand" them. The opposite, however, is still quite far from being real, our understanding of how our "user-friendly" computers can understand us humans is not deep enough yet. A lot of work has to be done both in the near and distant future. Let TSD 2002 be a modest contribution to this goal.

The conference also serves well in its second purpose: to facilitate researchers meeting in the NLP field from Western and Eastern Europe. Moreover, many participants now come from other parts of the world, thus making TSD a real crossroads for researchers in the NLP area. This volume contains the proceedings of this conference held in Brno, September 9–12, 2002. We were honored to have as keynote speakers James Pustejovsky from Brandeis University, and Ronald Cole from the University of Colorado.

We would like to thank all the Program Committee members and external reviewers for their conscientious and diligent reviewing work. This year we had a total of 66 accepted papers out of 134 submitted, altogether contributed by 182 authors. It can be seen that this time the competition was quite tough. We thank all of the authors for their high-quality contributions. Generous thanks are due to all the members of the Organizing Committee for their tireless efforts in organizing the conference. We deeply appreciate the help and cooperation of the Chair of the Program Committee, Hynek Hermansky. We would like also to express our sincere thanks to Dagmar Janoušková and Dana Komárková whose dedicated help contributed substantially to the preparation of the conference, and to Petr Sojka for his TEXpertise which resulted in the nice volume with a subject index which you are holding in your hands just now. We hope that you will be inspired by the conference and the proceedings, and that the program meets your expectations.

July 2002

Ivan Kopeček, Karel Pala
Chairs of the Organizing Committee, TSD 2002

Organization

TSD 2002 was organized by the Faculty of Informatics, Masaryk University, in cooperation with the Faculty of Applied Sciences, University of West Bohemia in Plzeň. The conference webpage is located at `www.fi.muni.cz/tsd2002/`

Program Committee

Jelinek, Frederick (USA), *general chair*
Hermansky, Hynek (USA), *executive chair*
Eneko, Agirre, (Spain)
Baudoin, Geneviève (France)
Ferencz, Attila (Romania)
Hajičová, Eva (Czech Republic)
Hlaváčová, Jaroslava (Czech Republic)
Hovy, Eduard (USA)
Kopeček, Ivan (Czech Republic)
Krauwer, Steven (The Netherlands)
Krishnamurthy, Ramesh (UK)
Matoušek, Václav (Czech Republic)
Nöth, Elmar (Germany)
Oliva, Karel (Czech Republic)
Pala, Karel (Czech Republic)
Pavesić, Nikola (Slovenia)
Petkevič, Vladimír (Czech Republic)
Psutka, Josef (Czech Republic)
Rothkrantz, Leon (The Netherlands)
Schukat-Talamazzini, E. Günter (Germany)
Skrelin, Pavel (Russia)
Vintsiuk, Taras (Ukraine)
Wilks, Yorick (UK)

Organizing Committee

Robert Batůšek, Pavel Cenek, Jan Černocký, Pavel Gaura, Aleš Horák, Dagmar Janoušková, Dana Komárková (Secretary), Ivan Kopeček (Co-chair), Karel Pala (Co-chair), Martin Povolný, Pavel Smrž, Petr Sojka (Proceedings), Marek Veber

Supported by:

International Speech Communication Association

Table of Contents

I Text

A Common Solution for Tokenization and Part-of-Speech Tagging 3
Jorge Graña, Miguel A. Alonso, Manuel Vilares (University La Coruña, Spain)

Rule Parser for Arabic Stemmer 11
Imad A. Al-Sughaiyer, Ibrahim A. Al-Kharashi (Computer and Electronics Research Institute, Riyadh, Saudi Arabia)

Achieving an Almost Correct PoS-Tagged Corpus 19
Pavel Květoň (Charles University, Prague, Czech Republic), Karel Oliva (Austrian Research Institute for Artificial Intelligence, Vienna, Austria)

Evaluation of a Japanese Sentence Compression Method Based on Phrase Significance and Inter-Phrase Dependency 27
Rei Oguro, Hiromi Sekiya, Yuhei Morooka, Kazuyuki Takagi, Kazuhiko Ozeki (The University of Electro-Communications, Tokyo, Japan)

User Query Understanding by the InBASE System as a Source for a Multilingual NL Generation Module 33
Michael V. Boldasov (Moscow State University, Russia), Elena G. Sokolova (Russian Research Institute for Artificial Intelligence, Moscow, Russia), Michael G. Malkovsky (Moscow State University, Russia)

The Role of WSD for Multilingual Natural Language Applications 41
Andrés Montoyo, Rafael Romero, Sonia Vázquez, Carmen Calle, Susana Soler (University Alicante, Spain)

Gibbsian Context-Free Grammar for Parsing 49
Antoine Rozenknop (Swiss Federal Institute of Technology, Lausanne, Switzerland)

Cross-Language Access to Recorded Speech in the MALACH Project 57
Douglas W. Oard, Dina Demner-Fushman (University of Maryland, USA), Jan Hajič (Charles University, Prague, Czech Republic), Bhuvana Ramabhadran (IBM T. J. Watson Research Center, USA), Samuel Gustman (Survivors of the Shoah Visual History Foundation, Los Angeles, USA), William J. Byrne (Johns Hopkins University, Baltimore, USA), Dagobert Soergel, Bonnie Dorr, Philip Resnik (University of Maryland, USA), Michael Picheny (IBM T. J. Watson Research Center, USA)

Using Salient Words to Perform Categorization of Web Sites 65
Marek Trabalka, Mária Bieliková (Slovak University of Technology, Bratislava, Slovakia)

Discourse-Semantic Analysis of Hungarian Sign Language 73
Gábor Alberti (University of Pécs, Hungary), Helga M. Szabó (National Association of the Deaf, Hungary)

Dependency Analyser Configurable by Measures 81
Tomáš Holan (Charles University, Prague, Czech Republic)

The Generation and Use of Layer Information in Multilayered Extended Semantic Networks 89
Sven Hartrumpf, Hermann Helbig (University of Hagen, Germany)

Filtering of Large Numbers of Unstructured Text Documents by the Developed Tool TEA 99
Jan Žižka, Aleš Bourek (Masaryk University in Brno, Czech Republic)

Visualisation Techniques for Analysing Meaning 107
Dominic Widdows, Scott Cederberg, Beate Dorow (Stanford University, California, USA)

Statistical Part-of-Speech Tagging for Classical Chinese 115
Liang Huang, Yinan Peng (Shanghai Jiaotong University, China), Huan Wang (East China Normal University, China), Zhenyu Wu (Shanghai Jiaotong University, China)

Spanish Natural Language Interface for a Relational Database Querying System 123
Rodolfo A. Pazos Rangel (Centro Nacional de Investigación y Desarrollo Tecnológico, Mexico), Alexander Gelbukh (National Polytechnic Institute, Mexico), Juan Javier González Barbosa, Erika Alarcón Ruiz, Alejandro Mendoza Mejía, Ana Patricia Domínguez Sánchez (Instituto Tecnológico de Ciudad Madero, Mexico)

Word Sense vs. Word Domain Disambiguation: A Maximum Entropy Approach 131
Armando Suárez, Manuel Palomar (University Alicante, Spain)

Exploiting Thesauri and Hierarchical Categories in Cross-Language Information Retrieval 139
Fatiha Sadat, Masatoshi Yoshikawa, Shunsuke Uemura (Nara Institute of Science and Technology, Japan)

Valency Lexicon for Czech: From Verbs to Nouns 147
Markéta Lopatková, Veronika Řezníčková, Zdeněk Žabokrtský (Charles University, Prague, Czech Republic)

Term Clustering Using a Corpus-Based Similarity Measure 151
Goran Nenadić, Irena Spasić, Sophia Ananiadou (University of Salford, UK)

Word Sense Discrimination for Czech 155
Robert Král (Masaryk University in Brno, Czech Republic)

Tools for Semi-automatic Assignment of Czech Nouns to Declination Patterns 159
Dita Bartůšková, Radek Sedláček (Masaryk University in Brno, Czech Republic)

II Speech

Automatic Lexical Stress Assignment of Unknown Words for Highly Inflected Slovenian Language 165
Tomaž Šef, Maja Škrjanc, Matjaž Gams (Ljubljana University, Slovenia)

German and Czech Speech Synthesis Using HMM-Based Speech Segment Database . . 173
Jindřich Matoušek, Daniel Tihelka, Josef Psutka, Jana Hesová (University of West Bohemia in Pilsen, Czech Republic)

Comparison and Combination of Confidence Measures 181
Georg Stemmer, Stefan Steidl, Elmar Nöth, Heinrich Niemann, Anton Batliner (University Erlangen-Nürnberg, Germany)

Strategies for Developing a Real-Time Continuous Speech Recognition System for Czech Language 189
Jan Nouza (Technical University of Liberec, Czech Republic)

Comparative Study on Bigram Language Models for Spoken Czech Recognition 197
Dana Nejedlová (Technical University of Liberec, Czech Republic)

Integration of Speech Recognition and Automatic Lip-Reading 205
Pascal Wiggers, Leon J. M. Rothkrantz (Delft University of Technology, The Netherlands)

Utterance Verification Based on the Likelihood Distance to Alternative Paths 213
Gies Bouwman, Lou Boves (University Nijmegen, The Netherlands)

Rejection Technique Based on the Mumble Model 221
Tomáš Bartoš, Luděk Müller (University of West Bohemia in Pilsen, Czech Republic)

Efficient Noise Estimation and Its Application for Robust Speech Recognition 229
Petr Motlíček, Lukáš Burget (Oregon Health & Science University, USA and Brno University of Technology, Czech Republic)

AlfaNum System for Speech Synthesis in Serbian Language 237
Milan Sečujski, Radovan Obradović, Darko Pekar, Ljubomir Jovanov, Vlado Delić (University of Novi Sad, Yugoslavia)

Speech Features Extraction Using Cone-Shaped Kernel Distribution 245
Janez Žibert, France Mihelič, Nikola Pavešić (University of Ljubljana, Slovenia)

Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments 253
Josef Psutka, Pavel Ircing, Josef V. Psutka, Vlasta Radová (University of West Bohemia in Pilsen, Czech Republic), William J. Byrne (Johns Hopkins University, Baltimore, USA), Jan Hajič (Charles University, Prague, Czech Republic), Samuel Gustman (Survivors of the Shoah Visual History Foundation, Los Angeles, USA), Bhuvana Ramabhadran (IBM T.J. Watson Research Laboratory, USA)

On the First Greek-TTS Based on Festival Speech Synthesis 261
P. Zervas, Ilyas Potamitis, Nikos Fakotakis, George Kokkinakis (University of Patras, Greece)

An Analysis of Limited Domains for Speech Synthesis 265
Robert Batůšek (Masaryk University in Brno, Czech Republic)

Advances in Very Low Bit Rate Speech Coding Using Recognition and Synthesis Techniques 269
Geneviève Baudoin (ESIEE Paris, France), François Capman (Thales Communications), Jan Černocký (VUT Brno, Czech Republic), Fadi El Chami (ESIEE Paris, France, University Tripolis, Libanon), Maurice Charbit, Gérard Chollet (ENST Paris, France), Dijana Petrovska-Delacrétaz (University of Fribourg, Switzerland)

A Comparison of Different Approaches to Automatic Speech Segmentation 277
Kris Demuynck, Tom Laureys (University Leuven, Belgium)

Keyword Spotting Using Support Vector Machines 285
Yassine Ben Ayed, Dominique Fohr, Jean Paul Haton (LORIA-CNRS/ INRIA Lorraine, France), Gérard Chollet (ENST, CNRS-LTCI, France)

Automatic Parameter Estimation for a Context- Independet Speech Segmentation Algorithm 293
Guido Aversano (Università Salerno and IIASS, Italy), Anna Esposito (Wright State University, USA and IIASS, Italy)

Phoneme Lattice Based A* Search Algorithm for Speech Recognition 301
Pascal Nocera, Georges Linares, Dominique Massonié, Loïc Lefort (LIA Avignon, France)

Heuristic and Statistical Methods for Speech/Non-speech Detector Design 309
Michal Prcín, Luděk Müller (University of West Bohemia in Pilsen, Czech Republic)

An Analysis of Conditional Responses in Dialogue 317
Elena Karagjosova, Ivana Kruijff-Korbayová (Saarland University, Germany)

Some Like It Gaussian 321
Pavel Matějka (FEEC VUT Brno, Czech Republik), Petr Schwarz, Martin Karafiát, Jan Čocký (FIT VUT Brno, Czech Republic)

Kernel Springy Discriminant Analysis and Its Application to a Phonological Awareness Teaching System 325
András Kocsor (Hungarian Academy of Sciences), Kornél Kovács (University of Szeged, Hungary)

Large Vocabulary Speech Recognition of Slovenian Language Using Data-Driven Morphological Models 329
Tomaž Rotovnik, Mirjam Sepesy Maučec, Bogomir Horvat, Zdravko Kačič (University of Maribor, Slovenia)

Uniform Speech Recognition Platform for Evaluation of New Algorithms 333
Andrej Žgank, Tomaž Rotovnik, Zdravko Kačič, Bogomir Horvat (University of Maribor, Slovenia)

Speech Enhancement Using Mixtures of Gaussians for Speech and Noise 337
Ilyas Potamitis, Nikos Fakotakis, Nikos Liolios, George Kokkinakis (University of Patras, Greece)

Fitting German into N-Gram Language Models 341
Robert Hecht, Jürgen Riedler, Gerhard Backfried (SAIL LABS Technology AG, Vienna, Austria)

Audio Collections of Endangered Arctic Languages in the Russian Federation 347
Marina Lublinskaya (Russian Academy of Science, St. Petersburg, Russia), Tatiana Sherstinova (St. Petersburg State University, Russia)

III Dialogue

Prosodic Classification of Offtalk: First Experiments 357
Anton Batliner, Viktor Zeißler, Elmar Nöth, Heinrich Niemann (University Erlangen-Nürnberg, Germany)

Statistical Decision Making from Text and Dialogue Corpora for Effective Plan Recognition 365
Manolis Maragoudakis, Aristomenis Thanopoulos, Nikos Fakotakis (University of Patras, Greece)

Natural Language Guided Dialogues for Accessing the Web 373
Marta Gatius, Horacio Rodríguez (Technical University of Catalonia, Spain)

Evaluating a Probabilistic Dialogue Model for a Railway Information Task 381
Carlos D. Martínez-Hinarejos, Francisco Casacuberta (València University, Spain)

Applying Dialogue Constraints to the Understanding Process in a Dialogue System . . . 389
Emilio Sanchis, Fernando García, Isabel Galiano, Encarna Segarra (València University, Spain)

Evaluation of Prediction Methods Applied to an Inflected Language 397
Nestor Garay-Vitoria, Julio Abascal, Luis Gardeazabal (University of the Basque Country, Spain)

Knowledge Based Speech Interfacing for Handhelds 405
C.K. Yang, Leon J.M. Rothkrantz (Delft University of Technology, The Netherlands)

Different Approaches to Build Multilingual Conversational Systems 413
Marion Mast, Thomas Roß, Henrik Schulz (IBM Europe, Germany), Heli Harrikari (NOKIA Research Center, Finland)

Strategies to Overcome Problematic Input in a Spanish Dialogue System 421
Victoria Arranz, Núria Castell, Montserrat Civit (Universitat Politècnica de Catalunya, Spain)

Dialogue Systems and Planning 429
Guy Camilleri (University Toulouse, France)

A Flexible Framework for Evaluation of New Algorithms for Dialogue Systems 437
Pavel Cenek (Masaryk University in Brno, Czech Republic)

From HTML to VoiceXML: A First Approach 441
César González Ferreras, David Escudero Mancebo, Valentín Cardeñoso Payo (University Valladolid, Spain)

Voice Chat with a Virtual Character: The Good Soldier Švejk Case Project 445
Jan Nouza, Petr Kolář, Josef Chaloupka (Technical University of Liberec, Czech Republic)

Application of Spoken Dialogue Technology in a Medical Domain 449
I. Azzini, T. Giorgino (University of Pavia, Italy), D. Falavigna, R. Gretter (ITC-irst, Trento, Italy)

A Voice-Driven Web Browser for Blind People 453
Simon Dobrišek, Jerneja Gros, Boštjan Vesnicer, France Mihelič, Nikola Pavešić (University of Ljubljana, Slovenia)

Enhancing Best Analysis Selection and Parser Comparison 461
Aleš Horák, Vladimír Kadlec, Pavel Smrž (Masaryk University in Brno, Czech Republic)

Author Index 467

Subject Index 471