

Lecture Notes in Artificial Intelligence 7499

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

Petr Sojka Aleš Horák
Ivan Kopeček Karel Pala (Eds.)

Text, Speech and Dialogue

15th International Conference, TSD 2012
Brno, Czech Republic, September 3-7, 2012
Proceedings

 Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Petr Sojka
Masaryk University
Faculty of Informatics
Department of Computer Graphics and Design
Botanická 68a, 602 00, Brno, Czech Republic
E-mail: sojka@fi.muni.cz

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

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-32789-6 e-ISBN 978-3-642-32790-2
DOI 10.1007/978-3-642-32790-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012944853

CR Subject Classification (1998): I.2, H.3-5, J.1, H.2, I.5, F.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

© Springer-Verlag Berlin Heidelberg 2012

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. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

The annual Text, Speech and Dialogue Conference (TSD), which originated in 1998, is in the middle of its second decade. So far more than 1,000 authors from 45 countries have contributed to the proceedings. TSD constitutes a recognized platform for the presentation and discussion of state-of-the-art technology and recent achievements in the field of natural language processing. It has become an interdisciplinary forum, interweaving the themes of speech technology and language processing. The conference attracts researchers not only from Central and Eastern Europe but also from other parts of the world. Indeed, one of its goals has always been to bring together NLP researchers with different interests from different parts of the world and to promote their mutual cooperation. One of the ambitions of the conference is, as its title says, not only to deal with dialogue systems as such, but also to contribute to improving dialogue between researchers in the two areas of NLP, i.e., between text and speech people. In our view, the TSD Conference was successful in this respect in 2012 again.

This volume contains the proceedings of the 15th TSD Conference, held in Brno, Czech Republic, in September 2012. In the review process, 82 papers were accepted out of 173 submitted, an acceptance rate of 47%.

We would like to thank all the authors for the efforts they put into their submissions and the members of the Program Committee and reviewers who did a wonderful job in helping us to select the most appropriate papers. We are also grateful to the invited speakers for their contributions. Their talks provide insight into important current issues, applications, and techniques related to the conference topics.

Last year the workshop on Natural Language Processing of Baltic and Slavonic Languages took place in the frame of the TSD Conference. This year a workshop on Hybrid Machine Translation was organized together with the conference. In part it was related to the EU project PRESEMT (FP7/2007-2013, ICT 248307) in which organizers participate.

Special thanks are due to the members of Local Organizing Committee for their tireless effort in organizing the conference.

The T_EXperts of Petr Sojka resulted in the production of the volume that you are holding in your hands.

We hope that the readers will benefit from the results of this event and disseminate the ideas of the TSD Conference all over the world. Enjoy the proceedings!

July 2012

Aleš Horák
Ivan Kopeček
Karel Pala
Petr Sojka

Organization

TSD 2012 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 <http://www.tsdconference.org/tsd2012/>

Program Committee

Hermansky, Hynek (USA),

General Chair

Agirre, Eneko (Spain)

Černocký, Jan (Czech Republic)

Ferencz, Attila (Romania)

Fišer, Darja (Slovenia)

Garabík, Radovan (Slovakia)

Gelbukh, Alexander (Mexico)

Guthrie, Louise, (UK)

Hajič, Jan (Czech Republic)

Hajičová, Eva (Czech Republic)

Hanks, Patrick (UK)

Hitzenberger, Ludwig (Germany)

Hlaváčová, Jaroslava

(Czech Republic)

Horák, Aleš (Czech Republic)

Hovy, Eduard (USA)

Kopeček, Ivan (Czech Republic)

Krauwer, Steven (The Netherlands)

Kunzmann, Siegfried (Germany)

Loukachevitch, Natalija (Russia)

Matoušek, Václav (Czech Republic)

McCarthy, Diana (UK)

Ney, Hermann (Germany)

Nöth, Elmar (Germany)

Oliva, Karel (Czech Republic)

Pala, Karel (Czech Republic)

Pavesić, Nikola (Slovenia)

Petkevič, Vladimír (Czech Republic)

Pianesi, Fabio (Italy)

Piasecki, Maciej (Poland)

Przepiorkowski, Adam (Poland)

Psutka, Josef (Czech Republic)

Pustejovsky, James (USA)

Rothkrantz, Leon (The Netherlands)

Rusko, Milan (Slovakia)

Skrelin, Pavel (Russia)

Smrž, Pavel (Czech Republic)

Sojka, Petr (Czech Republic)

Steidl, Stefan (Germany)

Stemmer, Georg (Germany)

Tadić, Marko (Croatia)

Varadi, Tamas (Hungary)

Vetulani, Zygmunt (Poland)

Vintsiuk, Taras (Ukraine)

Wiggers, Pascal (The Netherlands)

Wilks, Yorick (UK)

Zakharov, Victor (Russia)

Referees

Alegria, Iñaki

Arregi Uriarte, Olatz

Baisa, Vít

Beňuš, Štefan

Broda, Bartosz

Cerňák, Miloš

Diaz de Ilarraza, Arantza

Evdokimova, Vera

Evgrafova, Karina

Fellbaum, Christiane

Grézl, František

Guthrie, Joe

Hannemann, Mirko
Hlaváčková, Dana
Holub, Martin
Héja, Enikő
Kamshilova, Olga
Karafiát, Martin
Khokhlova, Maria
Kocharov, Daniil
Lopez de Lacalle, Oier
Marcinčuk, Michal
Mareček, David
Materna, Jiří
Matějka, Pavel
Maziarz, Marek
Mihelič, France
Miháltz, Márton
Mikolov, Tomáš
Mitrofanova, Olga

Mráková, Eva
Nedoluzhko, Anna
Němčík, Václav
Nevěřilová, Zuzana
Oravec, Csaba
Otegi, Arantxa
Peterek, Nino
Popel, Martin
Radziszewski, Adam
Rigau Claramunt, German
Růžička, Michal
Sass, Bálint
Sazhok, Mykola
Stemmer, Georg
Szöke, Igor
Veselý, Karel
Wardyński, Adam

Organizing Committee

Dana Hlaváčková (*Administrative Contact*), Aleš Horák (*Co-chairs*),
Dana Komárková (*Secretary*), Ivan Kopeček, Karel Pala (*Co-chair*),
Adam Rambousek (*Web System*), Pavel Rychlý, Petr Sojka (*Proceedings*)

Sponsors and Support

The TSD conference is regularly supported by International Speech Communication Association (ISCA). We would like to express our thanks to the Lexical Computing Ltd., for their kind sponsoring contribution to TSD 2012.

Table of Contents

Part I: Invited Papers

Getting to Know Your Corpus	3
<i>Adam Kilgarriff</i>	
Coreference Resolution: To What Extent Does It Help NLP Applications?	16
<i>Ruslan Mitkov, Richard Evans, Constantin Orăsan, Iustin Dornescu, and Miguel Rios</i>	

Part II: Text

Semantic Similarity Functions in Word Sense Disambiguation	31
<i>Lukasz Kobyliński and Mateusz Kopeć</i>	
Opinion Mining on a German Corpus of a Media Response Analysis . . .	39
<i>Thomas Scholz, Stefan Conrad, and Lutz Hillekamps</i>	
The Soundex Phonetic Algorithm Revisited for SMS Text Representation	47
<i>David Pinto, Darnes Vilariño, Yuridiana Alemán, Helena Gómez, Nahun Loya, and Héctor Jiménez-Salazar</i>	
Sentence Modality Assignment in the Prague Dependency Treebank . . .	56
<i>Magda Ševčíková and Jiří Mírovský</i>	
Literacy Demands and Information to Cancer Patients	64
<i>Dimitrios Kokkinakis, Markus Forsberg, Sofie Johansson Kokkinakis, Frida Smith, and Joakim Öhlen</i>	
Expanding Opinion Attribute Lexicons	72
<i>Aleksander Wawer and Konrad Gołuchowski</i>	
Taggers Gonna Tag: An Argument against Evaluating Disambiguation Capacities of Morphosyntactic Taggers	81
<i>Adam Radziszewski and Szymon Acedański</i>	
An Ambiguity Aware Treebank Search Tool	88
<i>Marcin Woliński and Andrzej Zaborowski</i>	

A New Annotation Tool for Aligned Bilingual Corpora.....	95
<i>Georgios Petasis and Mara Tsoumari</i>	
Optimizing Sentence Boundary Detection for Croatian.....	105
<i>Frane Šarić, Jan Šnajder, and Bojana Dalbelo Bašić</i>	
Mining the Web for Idiomatic Expressions Using Metalinguistic Markers	112
<i>Filip Graliński</i>	
Using Tree Transducers for Detecting Errors in a Treebank of Polish ...	119
<i>Katarzyna Krasnowska, Witold Kieraś, Marcin Woliński, and Adam Przepiórkowski</i>	
Combining Manual and Automatic Annotation of a Learner Corpus	127
<i>Tomáš Jelínek, Barbora Štindlová, Alexandr Rosen, and Jirka Hana</i>	
A Manually Annotated Corpus of Pharmaceutical Patents.....	135
<i>Márton Kiss, Ágoston Nagy, Veronika Vincze, Attila Almási, Zoltán Alexin, and János Csirik</i>	
Large-Scale Experiments with NP Chunking of Polish	143
<i>Adam Radziszewski and Adam Pawlaczek</i>	
Mapping a Lexical Semantic Resource to a Common Framework of Computational Lexicons	150
<i>Milena Slavcheva</i>	
The Rule-Based Approach to Czech Grammaticalized Alternations	158
<i>Václava Kettnerová, Markéta Lopatková, and Zdeňka Uřešová</i>	
Semi-supervised Acquisition of Croatian Sentiment Lexicon	166
<i>Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić</i>	
Towards a Constraint Grammar Based Morphological Tagger for Croatian	174
<i>Hrvoje Peradin and Jan Šnajder</i>	
A Type-Theoretical Wide-Coverage Computational Grammar for Swedish.....	183
<i>Malin Ahlberg and Ramona Enache</i>	
Application of Lemmatization and Summarization Methods in Topic Identification Module for Large Scale Language Modeling Data Filtering	191
<i>Lucie Skorkovská</i>	
Experiments and Results with Diacritics Restoration in Romanian	199
<i>Cristian Grozea</i>	

Supervised Distributional Semantic Relatedness	207
<i>Alistair Kennedy and Stan Szpakowicz</i>	
PSI-Toolkit: How to Turn a Linguist into a Computational Linguist	215
<i>Krzysztof Jassem</i>	
Heterogeneous Named Entity Similarity Function	223
<i>Jan Kocoń and Maciej Piasecki</i>	
Joint Part-of-Speech Tagging and Named Entity Recognition Using Factor Graphs	232
<i>György Móra and Veronika Vincze</i>	
Assigning Deep Lexical Types	240
<i>João Silva and António Branco</i>	
SBFC: An Efficient Feature Frequency-Based Approach to Tackle Cross-Lingual Word Sense Disambiguation	248
<i>Dieter Mourisse, Els Lefever, Nele Verbiest, Yvan Saeys, Martine De Cock, and Chris Cornelis</i>	
Dependency Relations Labeller for Czech	256
<i>Rudolf Rosa and David Mareček</i>	
Preliminary Study on Automatic Induction of Rules for Recognition of Semantic Relations between Proper Names in Polish Texts	264
<i>Michał Marcińczuk and Marcin Ptak</i>	
Actionable Clause Detection from Non-imperative Sentences in Howto Instructions: A Step for Actionable Information Extraction	272
<i>Jihee Ryu, Yuchul Jung, and Sung-Hyon Myaeng</i>	
Authorship Attribution: Comparison of Single-Layer and Double-Layer Machine Learning	282
<i>Jan Rygl and Aleš Horák</i>	
Key Phrase Extraction of Lightly Filtered Broadcast News	290
<i>Luís Marujo, Ricardo Ribeiro, David Martins de Matos, João P. Neto, Anatole Gershman, and Jaime Carbonell</i>	
Using a Double Clustering Approach to Build Extractive Multi-document Summaries	298
<i>Sara Botelho Silveira and António Branco</i>	
A Comparative Study of the Impact of Statistical and Semantic Features in the Framework of Extractive Text Summarization	306
<i>Tatiana Vodolazova, Elena Lloret, Rafael Muñoz, and Manuel Palomar</i>	

Using Dependency-Based Annotations for Authorship Identification	314
<i>Charles Hollingsworth</i>	
A Space-Efficient Phrase Table Implementation Using Minimal Perfect Hash Functions	320
<i>Marcin Junczys-Dowmunt</i>	
Common Sense Inference Using Verb Valency Frames	328
<i>Zuzana Nevřilová and Marek Grác</i>	
A Genetic Programming Experiment in Natural Language Grammar Engineering	336
<i>Marcin Junczys-Dowmunt</i>	
User Modeling for Language Learning in Facebook	345
<i>Maria Virvou, Christos Troussas, Jaime Caro, and Kurt Junshean Espinosa</i>	
Detection of Semantic Compositionality Using Semantic Spaces	353
<i>Lubomír Krčmář, Karel Ježek, and Massimo Poesio</i>	
User Adaptation in a Hybrid MT System: Feeding User Corrections into Synchronous Grammars and System Dictionaries	362
<i>Susanne Preuß, Hajo Keffer, Paul Schmidt, Georgios Goumas, Athanasia Asiki, and Ioannis Konstantinou</i>	
Using Cognates to Improve Lexical Alignment Systems	370
<i>Mirabela Navlea and Amalia Todirascu</i>	
Disambiguating Word Translations with Target Language Models	378
<i>André Lynum, Erwin Marsi, Lars Bungum, and Björn Gambäck</i>	
English-Vietnamese Machine Translation of Proper Names: Error Analysis and Some Proposed Solutions	386
<i>Thi Thanh Thao Phan and Izabella Thomas</i>	
Improved Phrase Translation Modeling Using MAP Adaptation	394
<i>A. Ryan Aminzadeh, Jennifer Drexler, Timothy Anderson, and Wade Shen</i>	

Part III: Speech

A Romanian Language Corpus for a Commercial Text-To-Speech Application	405
<i>Mihai Alexandru Ordean, Andrei Şaupe, Mihaela Ordean, Gheorghe Cosmin Silaghi, and Corina Giurgea</i>	

Making Community and ASR Join Forces in Web Environment	415
<i>Oldřich Krůza and Nino Peterek</i>	
Unsupervised Synchronization of Hidden Subtitles with Audio Track Using Keyword Spotting Algorithm	422
<i>Petr Stanislav, Jan Švec, and Luboš Šmídl</i>	
Did You Say What I Think You Said? Towards a Language-Based Measurement of a Speech Recognizer's Confidence	431
<i>Bernd Ludwig and Ludwig Hitzemberger</i>	
Dealing with Numbers in Grapheme-Based Speech Recognition	438
<i>Miloš Janda, Martin Karafiát, and Jan Černocký</i>	
Discretion of Speech Units for the Text Post-processing Phase of Automatic Transcription (in the Czech Language)	446
<i>Svatava Škodová, Michaela Kuchařová, and Ladislav Šeps</i>	
On the Impact of Annotation Errors on Unit-Selection Speech Synthesis	456
<i>Jindřich Matoušek, Daniel Tihelka, and Luboš Šmídl</i>	
Analysis of the Influence of Speech Corpora in the PLDA Verification in the Task of Speaker Recognition	464
<i>Lukáš Machlica and Zbyněk Zajíc</i>	
Adaptive Language Modeling with a Set of Domain Dependent Models	472
<i>Yangyang Shi, Pascal Wiggers, and Catholijn M. Jonker</i>	
Robust Adaptation Techniques Dealing with Small Amount of Data . . .	480
<i>Zbyněk Zajíc, Lukáš Machlica, and Luděk Müller</i>	
Language Modeling of Nonverbal Vocalizations in Spontaneous Speech	488
<i>Dmytro Prylipko, Bogdan Vlasenko, Andreas Stolcke, and Andreas Wendemuth</i>	
Acoustic Segmentation Using Group Delay Functions and Its Relevance to Spoken Keyword Spotting	496
<i>Srikanth R. Madikeri and Hema A. Murthy</i>	
An In-Car Speech Recognition System for Disabled Drivers	505
<i>Jozef Ivanecký and Stephan Mehlhase</i>	
Captioning of Live TV Programs through Speech Recognition and Re-speaking	513
<i>Aleš Pražák, Zdeněk Loose, Jan Trmal, Josef V. Psutka, and Josef Psutka</i>	

Investigation on Most Frequent Errors in Large-Scale Speech Recognition Applications	520
<i>Marek Boháč, Jan Nouza, and Karel Blavka</i>	
Neural Network Language Model with Cache	528
<i>Daniel Soutner, Zdeněk Loose, Luděk Müller, and Aleš Pražák</i>	
TENOR: A Lexical Normalisation Tool for Spanish Web 2.0 Texts	535
<i>Alejandro Mosquera and Paloma Moreda</i>	
A Bilingual HMM-Based Speech Synthesis System for Closely Related Languages	543
<i>Tadej Justin, Miran Pobar, Ivo Ipšić, France Mihelič, and Janez Žibert</i>	
The Role of Nasal Contexts on Quality of Vowel Concatenations	551
<i>Milan Legát and Radek Skarnitzl</i>	
Analysis and Assessment of State Relevance in HMM-Based Feature Extraction Method	559
<i>Rok Gajšek, Simon Dobrišek, and France Mihelič</i>	
On the Impact of Non-speech Sounds on Speaker Recognition	566
<i>Artur Janicki</i>	
Automatic Rating of Hoarseness by Text-Based Cepstral and Prosodic Evaluation	573
<i>Tino Haderlein, Cornelia Moers, Bernd Möbius, and Elmar Nöth</i>	
Improving the Classification of Healthy and Pathological Continuous Speech	581
<i>Klára Vicsi, Viktor Imre, and Gábor Kiss</i>	

Part IV: Dialogue

Using Foot-Syllable Grammars to Customize Speech Recognizers for Dialogue Systems	591
<i>Daniel Couto Vale and Vivien Mast</i>	
Coupled Pragmatic and Semantic Automata in Spoken Dialogue Management	599
<i>Jolanta Bachan</i>	
Exploration of Metaphor and Affect Sensing Using Semantic Interpretation in an Intelligent Agent	607
<i>Li Zhang</i>	

Sentence Classification with Grammatical Errors and Those Out of Scope of Grammar Assumption for Dialogue-Based CALL Systems	616
<i>Yu Nagai, Tomohisa Senzai, Seiichi Yamamoto, and Masafumi Nishida</i>	
Spoken Dialogue System Design in 3 Weeks	624
<i>Tomáš Valenta, Jan Švec, and Luboš Šmídl</i>	
Integrating Dialogue Systems with Images	632
<i>Ivan Kopeček, Radek Ošlejšek, and Jaromír Plhák</i>	
Natural Language Understanding: From Laboratory Predictions to Real Interactions	640
<i>Pedro Mota, Luísa Coheur, Sérgio Curto, and Pedro Fialho</i>	
Unsupervised Clustering of Prosodic Patterns in Spontaneous Speech . . .	648
<i>András Beke and György Szaszák</i>	
Czech Expressive Speech Synthesis in Limited Domain: Comparison of Unit Selection and HMM-Based Approaches	656
<i>Martin Grüber and Zdeněk Hanzlíček</i>	
Aggression Detection in Speech Using Sensor and Semantic Information	665
<i>Iulia Lefter, Leon J.M. Rothkrantz, and Gertjan J. Burghouts</i>	
Question Classification with Active Learning	673
<i>Domen Marinčič, Tomaž Kompara, and Matjaž Gams</i>	
2B\$ – Testing Past Algorithms in Nowadays Web	681
<i>Hugo Rodrigues and Luísa Coheur</i>	
Morphological Resources for Precise Information Retrieval	689
<i>Anne-Laure Ligozat, Brigitte Grau, and Delphine Tribout</i>	
Author Index	697