Lecture Notes in Artificial Intelligence 6562

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Zygmunt Vetulani (Ed.)

Human Language Technology

Challenges for Computer Science and Linguistics

4th Language and Technology Conference, LTC 2009
Poznan, Poland, November 6-8, 2009
Revised Selected Papers

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 Editor

Zygmunt Vetulani
Adam Mickiewicz University in Poznan
Faculty of Mathematics and Computer Science
ul. Umultowska 87, 61614 Poznan, Poland
E-mail: vetulani@amu.edu.pl

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-20094-6 e-ISBN 978-3-642-20095-3
DOI 10.1007/978-3-642-20095-3
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011923534

CR Subject Classification (1998): I.2.1, I.2.7, I.2, H.2.5, H.5.2, F.4.2, I.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

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

Human language technologies emerged in late twentieth century as a natural consequence of the technological progress of the human race. Since the industrial revolution in the eighteenth century, through the nineteenth and twentieth centuries, humans considerably transformed the world and definitely entered into the role of masters of the game. In an environment dominated by today's technology, humans effectively took over the totality of the animate world. The key to this success has consisted, since ancient times, in having mastered energy: fire, gunpowder, steam, coal, water power, electricity, nuclear power. The result was the creation of ever more sophisticated tools and artifacts at the high price of the destruction of large areas of the natural environment and of traditional social structures. Humans have become slaves of the technologies they invented and the subjects of new complex social structures. A new kind of resource, with roots dating back to Gutenberg and beyond, was focussed on at the beginning of the twenty-first century: information. In the information-rich technological environment, a new sociopolitical concept of the Information Society became a paradigm[1]. The Information Society implies a novel kind of social relation in the world in which people are surrounded by a mass of information-rich artifacts and technologies designed to be collaborative aids for them. This idea, explicitly addressed by EU policies of the last 20 years, gave a new stimulus for the development of technologies involving (or depending on) various forms of Natural Language Processing and Speech Technology, i.e., the so-called Human Language Technologies[2].

In the preface to the LTC 2007 Revised Selected Papers[3] the editors mentioned a number of challenges which inspire researchers and language engineers in this field across the world:

- Human languages evolved in such a way that made it well suited for humans but especially unsuited for processing on today's digital computers. Development of appropriate methodology to face the logical complexity of the human language remains a major challenge both for linguists and computer science engineers, although an essential progress has been achieved during the last 50 years.

[1] Information Society Technologies was one of thematic priorities under the European Six Framework Programme for the period 2002-2006.

[2] This term first(?) appeared in the name of the ARPA Human Language Technology (HLT) Workshops in 1993 (Former DARPA Speech and Natural Language Workshops).

[3] Zygmunt Vetulani and Hans Uszkoreit (Eds.)(2009): Human Language Technology, Challenges of the Information Society, Third Language and Technology Conference, LTC 2007, Poznan, Poland, October 2007, Revised Selected Papers. LNAI 5603. Springer-Verlag, Berlin, Heidelberg.

- The multilingual global society poses another grand technological challenge because in order to preserve the multicultural setup of the globalizing world all surviving languages need to be supported by language technologies.
- Another challenge consists in the integration of language with other media of communication such as gesture, facial expression, pictures and films, each of them requiring different techniques for encoding and processing that need to interact with language processing.
- Combination of human language technologies with a new generation of knowledge technologies which are nowadays called semantic technologies (cf. the Semantic Web).

All four of the challenges cited above continue to be valid and stimulating. But this list should not be considered closed. Technologies evolve at a high speed and the day approaches when our technological environment will be populated with devices which are equipped with artificial but human-like and human-friendly language and speech competences. To provide equal access to this environment for all members of the Information Society, independently of their mother tongue and thus avoiding their technological exclusion, will be a major challenge in the future. The time to start facing this challenge is now. This idea was behind the LTC-FLaReNet joint LRL Workshop "Getting Less-Resourced Languages on-Board!"[4] aiming at the international promotion of the development of language resources and technologies for under-resourced languages. The goal of this session was first to illustrate the various dimensions of that topic for various technologies (both written and spoken language processing) and for various languages (Indian languages (including Sanskrit), Celtic languages (Welsh, Irish, Scottish Gaelic, Manx Gaelic, Cornish and Breton), Amharic, Luxembourgish, Romani, Basque, Catalan, Galician, Sami and the Ga African language). A final panel session allowed for a general discussion and the drafting of a list of recommendations for at a better language coverage in Language Resources, and therefore Language Technologies, as it appears in the final report[5]. Some texts contained in this volume were presented at the LRL session.

In the present volume the reader will find the revised and in many cases substantially extended versions of 52 selected papers presented at the 4th Language and Technology Conference. This selection was made from a total of 103 high-quality contributions written by 250 authors qualified for conference presentation by an international jury on the basis of blind reviews. As a rule these assessments were taken into account for the selection to this volume. Still, we are aware of the partly subjective character of this selection. In a small number

[4] The Less-Resourced Languages Workshop (LRL) was proposed and set-up by Khalid Choukri, Joseph Mariani and Zygmunt Vetulani.

[5] J. Mariani, K. Choukri and Z. Vetulani, "Report on the Special joint LTC-FLaReNet session 'Getting Less-Resourced Languages On-Board !' at LTC'09 Conference". Cf. http://www.flarenet.eu/sites/default/files/LREC2010_International_Cooperation_Workshop_Mariani-Choukri-Vetulani.pdf
(cf. also http://www.ltc.amu.edu.pl).

of cases some interesting papers which presented partial or preliminary results of on-going research and development did not qualify for this book, but it is hoped that more complete results will be presented soon.

The selection of revised papers well illustrates the complexity and diversity of the field of Human Language Technologies. The papers collected in this book report on many man-years of hard work by 130 authors representing research institutions from 21 countries[6]: Belgium, Canada, Czech Republic, Finland, France, Germany, Greece, India, Italy, Iran, Ireland, Japan, Lithuania, Poland, Portugal, Romania, Russia, Spain, Switzerland, UK, Ukraine[7].

The reader will find the papers structured into thematic chapters. Clustering papers was a difficult task as in most cases the contributions addressed more than one thematic area so that our decisions should be considered as approximative. In particular, their attributions to chapters do not necessarily correspond to their attribution to the LTC thematic sessions and also may not correspond to the authors' first choice.

These chapters are:

1. Speech Processing (9)
2. Computational Morphology/Lexicography (4)
3. Parsing (4)
4. Computational Semantics (6)
5. Entailment (2)
6. Dialogue Modeling and Processing (4)
7. Digital Language Resources (9)
8. WordNet (3)
9. Document Processing (2)
10. Information Processing (IR,IE,other) (7)
11. Machine Translation (2).

The ordering of six initial chapters follows the natural order which humans use to process language understanding in NL-communication: starting with speech, and proceeding through morphology, syntax and semantics to dialogue. The next two chapters focus on resources, and the last three on multi-aspectual language engineering tasks. The idea to close this selection with machine translation (MT) papers symbolizes our opinion that machine translation, being (with the Warren Weaver's Memorandum, 1947)[8] the first large-scale program in language

[6] Against 250 authors from 38 countries participating in LTC 2009.

[7] In fact the geographical/language coverage is larger then what may look like from the presented data, as we know only the present affiliation of the data. Also, language coverage is larger than what may be inferred from the list of 21 countries. For example, languages such as Almaric, Bulgarian, Luxembourgish or Sanskrit do not correspond to the affiliations of the authors of respective papers.

[8] Weaver, W. (1949): 'Translation'. Repr. in: Locke, W.N. and Booth, A.D. (eds.) Machine translation of languages: fourteen essays (Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology, 1955), pp. 15-23.

engineering, will probably be the last of these research and development programs to reach a fully satisfactory result. There is no significant ordering within chapters, where papers are presented in alphabetical order with respect to the first author's family name.

The first chapter, "Speech Processing," contains ten contributions. It starts with a text on the evaluation of automatic speech-to-phoneme alignment systems (Baghai-Ravary, Kochanski, Coleman).The next one is a detailed presentation of a speech corpus of European Portuguese for text-to-speech systems (Barros, Möbius). It is followed by a paper on the quality improvement of a Japanese speech recognition system for the (noisy) car environment designed on the basis of weighted finite-state transducers (Betkowska Cavalcante, Shinoda, Furui). What follows is a contribution about TTS-oriented corpus design for Bulgarian, a South Slavic language with a relatively low number of speakers (9 million) by national language standards, but with a strong tradition in language engineering, particularly MT (Chalamandaris, Tsiakoulis, Raptis, Karabetsos). Phonetic similarity based on recognized observed properties (in automatic speech recognition, ASR) is the focus of the next paper. Conclusions concerning pronunciation variation modeling are presented (Kane, Mauclair, Carson-Berndsen). This paper is followed by a work on the detection of errors occurring in ASR output considered as a post-processing stage of ASR (Pellegrini, Trancoso). The last of the five challenges mentioned above is directly addressed in the study of pronunciation and writing variants in Luxembourgish, an under-resourced language spoken by approximately 300,000 speakers (Snoeren, Adda-Decker, Adda). The reader will then find a paper on morpheme-based language modeling for speech recognition for Amharic, another less-resourced language represented in this book (Tachbelie, Abate, Menzel). The next paper is on multilevel annotation software for speech corpora, in which the authors present a generic and corpus-independent toolkit supporting de facto standards and commonly used annotation formats (Wilson, Carson-Berndsen). The chapter closes with a report on time duration of phonemes (in Polish). This study is oriented to the development of speech/speaker recognition systems (B. Ziółko, M. Ziółko).

In the "Computational Morphology/Lexicography" chapter we present three papers. The first is about the classification of Japanese polysemous verbs using advanced mathematical methods (Fukumoto, Suzuki, Yamashita). The next paper deals with a problem of the similarity or typological proximity of languages. It proposes to measure the proximity between languages in terms of their vocabulary structure (Lepage, Gosme, Lardilleux). A tool for lexicographic description of multi-word units is presented in the last paper of this chapter (Marciniak, Savary, Sikora, Woliński).

Four contributions were accepted for the chapter concerned with "Parsing," which opens with a presentation of a parsing algorithm for context-free and probabilistic context-free (CFG and PCFG) grammars. This contribution is language independent and therefore of general interest for those who use CFG/PCFG for NL investgations (Hulden). The next article proposes a parsing algorithm described for the Czech language but presented as representative of the family of

Slavic languages. In order to deal with problems caused by word-order-related phenomena (Slavic languages have a relatively free word order) the authors propose a pattern-matching-based algorithm (Kovář, Horák, Jakubíček). Sentence segmentation as a pre-processing stage for higher-level processing (parsing) is considered in the next paper; the SRX standard for sentence segmentation was applied for English and Polish and the results have been compared (Miłkowski, Lipski). The last contribution in the parsing chapter is about using lexicon grammar entries for French verbs in a large-coverage parser (Tolone, Sagot).

Various issues of computational semantics are of first interest in the next chapter of six contributions. This chapter opens with the problem of overt pronouns resolution (Fukumoto, Suzuki). The second one addresses a methodological question of whether sentiment intensity may be considered to be a good summary indicator. This problem, relevant for summarization tasks, is given a negative answer (Kabadjov, Balahur, Boldrini). The third paper contributes to deep semantic analysis which often requires temporal information processing. Temporal information may be inferred from the textual descriptions of events. The paper contributes to classification of temporal relations (Mirroshandel, Khayyamian, Ghassem-Sani). Semantic disambiguation is the focus of the next paper where the authors present a platform (applied to Polish) for testing various word sense disambiguation technologies (Młodzki, Przepiórkowski). The next paper is about a semantic analyzer used in the translation of natural language texts into the Polish Sign Language (Romaniuk, Suszczańska, Szmal). The last article of this chapter presents a system which outputs semantic representations (logical forms or discourse representation structures) on top of dependency relations produced by a statistical parser (Zouaq, Gagnon, Ozell).

A chapter with two papers on text entailment follows. In the first one a system for recognizing sentential entailment is presented (Bédaride, Gardent), whereas a formal logical framework for definition and combination of specialized entailment engines is discussed in the second (Cabrio, Magnini).

The next chapter is on "Dialogue Modeling and Processing". The chapter starts with a paper on Wizard-of-Oz experiments for natural dialogue collection. The collected dialogues were processed and the results were used at the design stage of an artificial companion project and re-used for creation of an expressive speech corpus for Czech (Grůber, Legát, Ircing, Romportl, Psutka). Dialogue turn analysis for the purpose of automatic summarization of voice conversations was the theme of the second paper (Pallotta, Delmonte, Bristot). The next two papers cover various aspects of a man–machine dialogue system (POLINT-112-SMS) designed to understand short messages (SMS) in natural language. The first of them covers both methodological aspects of project development and the architecture of the resulting system (Vetulani, Marciniak). The second one focusses on dialogue control and on solutions of several hard problems like anaphora, overt pronouns etc. (Walkowska).

Nine papers are collected in the "Digital Language Resources" chapter. It opens with a research report on the development strategy and the HLT resources obtained so far for the Basque language for which the "survival"

program supported by the local administration in the Basque Country is explicitly based on language technologies (Alegria, Aranzabe, Arregi, Artola, Díaz de Ilarraza, Mayor, Sarasola). Methodological issues related to the construction of morphosyntactic resources with special attention to the comparison of natural languages were investigated by the authors of the second paper of the chapter. An important potential output of this research is a practical indication of how to predict the costs of development of morphosyntactic resources (Blancafort, De Loupy). The next paper is a presentation of a tool for corpora visualization, especially with respect to different types of frequency information (Culy, Lyding). Automatic acquisition of entries for bilingual lexica is a concern of the fourth paper, in which the authors propose an acquisition method consisting in exploration of keyword lists attached to bilingual documents (Graliński, Jassem, Kurc). An experiment in annotating Sanskrit, the oldest documented still spoken Indo-European language, is reported in the next paper (Jha, Gopal, Mishra). Readers concerned with dialectology resources may have interest in the paper on authorizing procedures for e-learning courses on dialectical phonetics (Kedrova, Yegorov, Volkova). The next paper presents a corpus collection exercise whose aim was modeling of user language competence with particular interest in describing spatial relations (Osiński). The following overview of a number of XML standards for multilevel corpus annotation is a contribution to the general problem of standards development, as too many standards mean no standards at all (Przepiórkowski, Bański). The last contribution in the LR chapter is a report on a project aiming at a national-scale corpus of academic Lithuanian, a language for which the existing corpora are not sufficient to cover the whole range of scientific discourse (Usoniene, Butenas, Ryvityte, Sinkuniene, Jasionyte, Juozapavicius).

Although many papers in this book refer to WordNet-based methodologies (technologies), WordNet occupies a central position in only three of them. In the first of these the author shows how WordNet, independently of the internal organization of its data, may be applied as a tool to enrich a valence dictionary of Polish verbs by adding semantic information (Hajnicz). Using a Princeton WordNet-based sense disambiguation algorithm to evaluate the degree of semantic relatedness between words is the main concern of the second paper (Ion, Ştefănescu). The last one in this chapter presents an interface to Polish WordNet (PolNet) and its application within an NL understanding system which uses PolNet as ontology (Kubis).

Two papers may be classified as directly contributing to the "Document Processing" field. The first of the two presents a Web-accessible tool to support diplomatic transcriptions of historical language data (i.e., transcriptions free of any kind of interpretation involved in the transcribing process) (Dipper, Schnurrenberger). The second one proposes an algorithm of authorship attribution for short texts (Nawrot).

Several (7) papers in this selection deal with various aspects of "Information Processing in form of Retrieval, Extraction and Other". Looking for speculative sentences in scientific texts (biology) is proposed as a tool for biologists interested

in finding new hypotheses published in scientific literature (J. Desclés, Alrahabi, J.-P. Desclés). The second paper presents two Arabic summarization systems based on the extraction of sentences that best match words in the query (El-Haj, Kruschwitz, Fox). Opinion extraction consisting in identification of subjectivity expressions is the theme of the third paper (Esuli, Sebastiani). The next paper in the chapter presents a method of analyzing the structure of the titles of research papers by means of information-extraction techniques (Kondo, Nanba, Takezawa, Okumura). The fifth article presents a system for the extraction and presentation of quotations found in French newswire transmissions. This tool is of direct practical interest, e.g., for press agencies (de La Clergerie, Sagot, Stern, Denis, Recourcé, Mignot). The next contribution shows the reader how to improve the precision of contextual advertising with the help of language technologies (Pak). The closing article of this chapter presents a comparison of four unsupervised algorithms for automatically extracting keywords from the multimedia archive of Belga News Archive (Palomino, Wuytack).

Finally, the volume ends with two papers classified as contributions to "Machine Translation." Both papers address the issue of MT quality evaluation. The first one explores the use of paraphrases for the refinement of traditional methods for text evaluation (valid also for summarization) (Hirahara, Nanba, Takezawa, Okumura). The last paper in the book describes the usage of normalized compression distance as a language-independent machine translation quality evaluation tool (Kettunen).

January 2011

Zygmunt Vetulani
Joseph Mariani

Organization

Organizing Committee

Zygmunt Vetulani - Conference Chair
Marek Kubis
Piotr Kuszyk
Jacek Marciniak
Tomasz Obrębski
Jędrzej Osiński
Justyna Walkowska

(All at the Adam Mickiewicz University, Poznań, Poland)

LTC Program Committee

Victoria Arranz
Anja Belz
Janusz S. Bień
Krzysztof Bogacki
Christian Boitet
Leonard Bolc
Lynne Bowker
Nicoletta Calzolari
Nick Campbell
Julie Carson-Berndsen
Khalid Choukri
Adam Dąbrowski
Elżbieta Dura
Katarzyna Dziubalska-Kołaczyk
Tomaz Erjavec
Cedrick Fairon
Christiane Fellbaum
Maria Gavrilidou
Dafydd Gibbon
Stefan Grocholewski
Franz Guenthner
Hans Guesgen
Eva Hajičová
Roland Hausser
Steven Krauwer
Eric Laporte
Yves Lepage
Gerard Ligozat
Natalia Loukachevitch
Wiesław Lubaszewski
Bente Maegaard
Bernardo Magnini
Joseph Mariani
Jacek Martinek
Gayrat Matlatipov
Keith J. Miller
Nicholas Ostler
Karel Pala
Pavel S. Pankov
Patrick Paroubek
Stelios Piperidis
Emil Pływaczewski
Gabor Proszeky
Adam Przepiórkowski
Reinhard Rapp
Zbigniew Rau
Mike Rosner
Justus Roux
Vasile Rus
Rafał Rzepka
Frédérique Ségond
Zhongzhi Shi
Włodzimierz Sobkowiak
Hanna Szafrańska
Marek Świdzinski
Ryszard Tadeusiewicz
Dan Tufiş
Hans Uszkoreit
Zygmunt Vetulani - Chair
Piek Vossen
Tom Wachtel
Jan Węglarz
Mariusz Ziółko
Richard Zuber

LRL Workshop Program Committee

LRL Co-chairs: Joseph Mariani, Khalid Choukri, Zygmunt Vetulani

Núria Bel
Gerhard Budin
Nicoletta Calzolari
Dafydd Gibbon
Marko Grobelnik
Jan Hajič
Alfed Majewicz
Asunción Moreno
Jan Odijk
Nicholas Ostler
Stellios Piperidis
Gabor Proszeky
Mohsen Rashwan
Kepa Sarasola
Marko Tadić
Dan Tufiş
Cristina Vertan
Briony Williams

Invited Reviewers

Xabier Arregi
Richard Beaufort
Olivier Blanc
Dragos Burileanu
Tommaso Caselli
Louise-Amélie Cougnon
Jolanta Cybulka
Damir Ćavar
Arantza Díaz de Ilarraza
Paweł Dybała
Nerea Ezeiza
Byron Georgantopoulos
Filip Graliński
Krzysztof Jassem
Marcin Junczys-Dowmunt
Besim Kabashi
Natalia Kotsyba
Anna Kupść
Rafał L. Górski
Penny Labropoulou
Kevers Laurent
Maciej Lison
Jacek Marciniak
Montserat Marimon
Agnieszka Mykowiecka
Tomasz Obrębski
Maciej Piasecki
Dawid Pietrala
Thomas Proisi
Prokopis Prokopidis
Michał Ptaszyński
Valeria Quochi
Flo Reeder
Paweł Rydzewski
Agata Savary
Grażyna Vetulani
Marta Villegas
Jorge Vivaldi
Alina Wróblewska
Dapeng Zhang
Bartosz Ziółko

The reviewing process was effected by the members of Program Committees and Invited Reviewers recommended by PC members.

Table of Contents

Speech Processing

Data-Driven Approaches to Objective Evaluation of Phoneme Alignment Systems 1
Ladan Baghai-Ravary, Greg Kochanski, and John Coleman

Phonetically Transcribed Speech Corpus Designed for Context Based European Portuguese TTS 12
Maria Barros and Bernd Möbius

Robust Speech Recognition in the Car Environment 24
Agnieszka Betkowska Cavalcante, Koichi Shinoda, and Sadaoki Furui

Corpus Design for a Unit Selection TtS System with Application to Bulgarian 35
Aimilios Chalamandaris, Pirros Tsiakoulis, Spyros Raptis, and Sotiris Karabetsos

Automatic Identification of Phonetic Similarity Based on Underspecification 47
Mark Kane, Julie Mauclair, and Julie Carson-Berndsen

Error Detection in Broadcast News ASR Using Markov Chains 59
Thomas Pellegrini and Isabel Trancoso

Pronunciation and Writing Variants in an Under-Resourced Language: The Case of Luxembourgish Mobile N-Deletion 70
Natalie D. Snoeren, Martine Adda-Decker, and Gilles Adda

Morpheme-Based and Factored Language Modeling for Amharic Speech Recognition 82
Martha Yifiru Tachbelie, Solomon Teferra Abate, and Wolfgang Menzel

The Corpus Analysis Toolkit - Analysing Multilevel Annotations 94
Stephen Wilson and Julie Carson-Berndsen

Computational Morphology/Lexicography

Time Durations of Phonemes in Polish Language for Speech and Speaker Recognition 105
Bartosz Ziółko and Mariusz Ziółko

Polysemous Verb Classification Using Subcategorization Acquisition and Graph-Based Clustering 115
Fumiyo Fukumoto, Yoshimi Suzuki, and Kazuyuki Yamashita

Estimating the Proximity between Languages by Their Commonality in Vocabulary Structures 127
Yves Lepage, Julien Gosme, and Adrien Lardilleux

Toposław – A Lexicographic Framework for Multi-word Units 139
Małgorzata Marciniak, Agata Savary, Piotr Sikora, and Marcin Woliński

Parsing

Parsing CFGs and PCFGs with a Chomsky-Schützenberger Representation 151
Mans Hulden

Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech 161
Vojtěch Kovář, Aleš Horák, and Miloš Jakubíček

Using SRX Standard for Sentence Segmentation 172
Marcin Miłkowski and Jarosław Lipski

Using Lexicon-Grammar Tables for French Verbs in a Large-Coverage Parser 183
Elsa Tolone and Benoît Sagot

Computational Semantics

Effect of Overt Pronoun Resolution in Topic Tracking 192
Fumiyo Fukumoto and Yoshimi Suzuki

Sentiment Intensity: Is It a Good Summary Indicator? 203
Mijail Kabadjov, Alexandra Balahur, and Ester Boldrini

Syntactic Tree Kernels for Event-Time Temporal Relation Learning 213
Seyed Abolghasem Mirroshandel, Mahdy Khayyamian, and Gholamreza Ghassem-Sani

The WSD Development Environment 224
Rafał Młodzki and Adam Przepiórkowski

Semantic Analyzer in the Thetos-3 System 234
Julia Romaniuk, Nina Suszczańska, and Przemysław Szmal

Unsupervised and Open Ontology-Based Semantic Analysis 245
Amal Zouaq, Michel Gagnon, and Benoît Ozell

Entailment

Non Compositional Semantics Using Rewriting 257
Paul Bédaride and Claire Gardent

Defining Specialized Entailment Engines Using Natural Logic Relations 268
Elena Cabrio and Bernardo Magnini

Dialogue Modeling and Processing

Czech Senior COMPANION: Wizard of Oz Data Collection and Expressive Speech Corpus Recording and Annotation 280
Martin Grůber, Milan Legát, Pavel Ircing, Jan Romportl, and Josef Psutka

Abstractive Summarization of Voice Communications 291
Vincenzo Pallotta, Rodolfo Delmonte, and Antonella Bristot

Natural Language Based Communication between Human Users and the Emergency Center: POLINT-112-SMS 303
Zygmunt Vetulani and Jacek Marciniak

Dialogue Organization in Polint-112-SMS 315
Justyna Walkowska

Digital Language Resources

Valuable Language Resources and Applications Supporting the Use of Basque 327
Iñaki Alegria, Maxux Aranzabe, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza, Aingeru Mayor, and Kepa Sarasola

Clues to Compare Languages for Morphosyntactic Analysis: A Study Run on Parallel Corpora and Morphosyntactic Lexicons 339
Helena Blancafort and Claude de Loupy

Corpus Clouds - Facilitating Text Analysis by Means of Visualizations 351
Chris Culy and Verena Lyding

Acquiring Bilingual Lexica from Keyword Listings 361
Filip Graliński, Krzysztof Jassem, and Roman Kurc

Annotating Sanskrit Corpus: Adapting IL-POSTS 371
Girish Nath Jha, Madhav Gopal, and Diwakar Mishra

Effective Authoring Procedure for E-learning Courses' Development in Philological Curriculum Based on LOs Ideology 380
Galina Kedrova, Anatoly Yegorov, and Maria Volkova

Acquisition of Spatial Relations from an Experimental Corpus 388
Jędrzej Osiński

Which XML Standards for Multilevel Corpus Annotation? 400
Adam Przepiórkowski and Piotr Bański

Corpus Academicum Lithuanicum: Design Criteria, Methodology, Application 412
Aurelija Usoniene, Linas Butenas, Birute Ryvityte, Jolanta Sinkuniene, Erika Jasionyte, and Algimantas Juozapavicius

WordNet

The EM-Based Wordnet Synsets Annotation of NP/PP Heads 423
Elżbieta Hajnicz

Unsupervised Word Sense Disambiguation with Lexical Chains and Graph-Based Context Formalization 435
Radu Ion and Dan Ştefănescu

An Access Layer to PolNet – Polish WordNet 444
Marek Kubis

Document Processing

OTTO: A Tool for Diplomatic Transcription of Historical Texts 456
Stefanie Dipper and Martin Schnurrenberger

Automatic Author Attribution for Short Text Documents 468
Monika Nawrot

BioExcom: Detection and Categorization of Speculative Sentences in Biomedical Literature 478
Julien Desclés, Motasem Alrahabi, and Jean-Pierre Desclés

Experimenting with Automatic Text Summarisation for Arabic 490
Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox

Enhancing Opinion Extraction by Automatically Annotated Lexical Resources (Extended Version) 500
Andrea Esuli and Fabrizio Sebastiani

Technical Trend Analysis by Analyzing Research Papers' Titles 512
Tomoki Kondo, Hidetsugu Nanba, Toshiyuki Takezawa, and Manabu Okumura

Information Processing (IR, IE, other)

Extracting and Visualizing Quotations from News Wires 522
Éric de La Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, and Victor Mignot

Using Wikipedia to Improve Precision of Contextual Advertising 533
Alexander Pak

Unsupervised Extraction of Keywords from News Archives 544
Marco A. Palomino and Tom Wuytack

Machine Translation

Automatic Evaluation of Texts by Using Paraphrases 556
Kazuho Hirahara, Hidetsugu Nanba, Toshiyuki Takezawa, and Manabu Okumura

Packing It All Up in Search for a Language Independent MT Quality Measure Tool – Part Two 567
Kimmo Kettunen

Author Index 577