Christina J. Hopfe  Yacine Rezgui
Elisabeth Métais  Alun Preece
Haijiang Li (Eds.)

Natural Language Processing and Information Systems

15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010
Cardiff, UK, June 23-25, 2010
Proceedings

Springer
The 15th International Conference on Applications of Natural Language to Information Systems (NLDB 2010) took place during June 23–25 in Cardiff (UK). Since the first edition in 1995, the NLDB conference has been aiming at bringing together researchers, people working in industry and potential users interested in various applications of natural language in the database and information system area. However, in order to reflect the growing importance of accessing information from a diverse collection of sources (Web, Databases, Sensors, Cloud) in an equally wide range of contexts (including mobile and tethered), the theme of the 15th International Conference on Applications of Natural Language to Information Systems 2010 was "Communicating with Anything, Anywhere in Natural Language."

Natural languages and databases are core components in the development of information systems. Natural language processing (NLP) techniques may substantially enhance most phases of the information system lifecycle, starting with requirement analysis, specification and validation, and going up to conflict resolution, result processing and presentation. Furthermore, natural language-based query languages and user interfaces facilitate the access to information for all and allow for new paradigms in the usage of computerized services.

Hot topics such as information retrieval (IR), software engineering applications, hidden Markov models, natural language interfaces and semantic networks and graphs imply a complete fusion of databases, IR and NLP techniques.

Among an increasing number of submitted papers, the Program Committee selected 14 papers as full papers, thus coming up with an acceptance rate of less than 30%. These proceedings also include 19 short papers that were presented at the conference and two invited talks from industry, one given by James Luke (IBM) and the other one given by Enrique Alfonseca (Google).

This conference was possible thanks to the support of three organizing institutions: Cardiff University, School of Engineering and School of Computer Science and Informatics, UK and Centre National des Arts et Métiers (CNAM), France. We also wish to thank the entire organizing team including secretaries, researchers and students who put their competence, enthusiasm and kindness into making this conference a real success. Special thanks are due to the Welsh Assembly Government (WAG). WAG was the main sponsor of this year’s conference.
Conference Organization

Conference Co-chairs
Yacine Rezgui, Cardiff University, UK
Elisabeth Métais, CNAM, France

Program Co-chairs
Alun Preece, Cardiff University, UK
Christina J. Hopfe, Cardiff University, UK

Conference Publicity Chair
Nadira Lammari, CNAM, France

Program Committee
Akhilesh Bajaj, University of Tulsa, USA
Alexander Gelbukh, Mexican Academy of Sciences, Mexico
Alexander Hinneburg, University of Halle, Germany
Alfredo Cuzzocreaq, University of Calabria, Italy
Andreas Hotho, University of Kassel, Germany
Andrés Montoyo, Universidad de Alicante, Spain
Antje Düsterhöft, Hochschule Wismar, Germany
Bernhard Thalheim, Kiel University, Germany
Cedric du Mouza, CNAM, France
Christian Kop, University of Klagenfurt, Austria
Christian Winkler, University of Klagenfurt, Austria
Christina J. Hopfe, Cardiff University, UK
Deryle Lonsdale, Brigham Young University, USA
Elisabeth Métais, CNAM, France
Epaminondas Kapetanios, University of Westminster, UK
Fabio Rinaldi, University of Zurich, Switzerland
Farid Meziane, Salford University, UK
Frederic Andres, University of Advanced Studies, Japan
Georgia Koutrika, Stanford University, USA
Grigori Sidorov, National Researcher of Mexico, Mexico
Günter Neumann, DFKI, Germany
Günther Fliedl, University of Klagenfurt, Austria
Hae-Chang Rim, Korea University, Korea
Harmain Harman, United Arab Emirates University, UAE
Heinrich C. Mayr, University of Klagenfurt, Austria
Helmut Horacek  
Saarland University, Germany

Hiram Calvo  
National Polytechnic Institute, Mexico

Irena Spasic  
Manchester Centre for Integrative Systems Biology, UK

Isabelle Comyn-Wattiau  
CNAM, France

Jacky Akoka  
CNAM, France

Jana Lewerenz  
Capgemini sd&m Düsseldorf, Germany

Jian-Yun Nie  
Université de Montréal, Canada

Jing Bai  
Yahoo Inc., Canada

Jon Atle Gulla  
Norwegian University of Science and Technology, Norway

Juan Carlos Trujillo  
Universidad de Alicante, Spain

Jürgen Rilling  
Concordia University, Canada

Karin Harbusch  
Universität Koblenz-Landau, Germany

Krishnaprasad Thirunarayan  
Wright State University, USA

Leila Kosseim  
Concordia University, Canada

Luis Alfonso Ureña  
Universidad de Jaén, Spain

Luisa Mich  
University of Trento, Italy

Magdalena Wolska  
Saarland University, Germany

Manolis Koubarakis  
University of Athens, Greece

Markus Schaal  
Bilkent University, Turkey

Max Silberztein  
Université de Franche-Comté, France

Mokrane Bouzeghoub  
Université de Versailles, France

Nadira Lammari  
CNAM, France

Odile Piton  
Université Paris I Panthéon-Sorbonne, France

Panos Vassiliadis  
University of Ioannina, Greece

Paul Johannesson  
Stockholm University, Sweden

Paul McFetridge  
Simon Fraser University, Canada

Philipp Cimiano  
CITEC, University of Bielefeld

Pit Pichappan  
Annamalai University, India

Rafael Muñoz  
Universidad de Alicante, Spain

René Witte  
Concordia University, Canada

Roger Chiang  
University of Cincinnati, USA

Roland Wagner  
University of Linz, Austria

Samia Nefti-Meziani  
University of Salford, UK

Samira Si-Said Cherfi  
CNAM, France

Sophia Ananiadou  
NaCTeM, Manchester Interdisciplinary Biocentre, UK

Stéphane Lopes  
Université de Versailles, France

Udo Hahn  
Friedrich-Schiller-Universität Jena, Germany

Veda Storey  
Georgia State University, USA

Vijay Sugumaran  
Oakland University Rochester, USA

Yacine Rezgui  
University of Salford, UK

Zoubida Kedad  
Université de Versailles, France
Organizing Committee Chair

Christina J. Hopfel	Cardiff University, UK

Organizing Committee

Haijiang Li	Cardiff University, UK
John Miles	Cardiff University, UK
Elisabeth Métais	CNAM, France
Alan Preece	Cardiff University, UK
Yacine Rezgui	Cardiff University, UK
# Table of Contents

## Information Retrieval

An Approach for Adding Noise-Tolerance to Restricted-Domain Information Retrieval ............................................ 1  
*Katia Vila, Josval Díaz, Antonio Fernández, and Antonio Ferrández*

Measuring Tree Similarity for Natural Language Processing Based Information Retrieval ................................. 13  
*Zhiwei Lin, Hui Wang, and Sally McClean*

Sense-Based Biomedical Indexing and Retrieval ..................... 24  
*Duy Dinh and Lynda Tamine*

## Natural Language Processing

Semantic Content Access Using Domain-Independent NLP Ontologies .......................................................... 36  
*René Witte and Ralf Krestel*

Extracting Meronymy Relationships from Domain-Specific, Textual Corporate Databases .................................. 48  
*Ashwin Ittoo, Gosse Bouma, Laura Maruster, and Hans Wortmann*

Automatic Word Sense Disambiguation Using Cooccurrence and Hierarchical Information ............................... 60  
*David Fernandez-Amoros, Ruben Heradio Gil, Jose Antonio Cerrada Somolinos, and Carlos Cerrada Somolinos*

## Software Engineering Applications

Automatic Quality Assessment of Source Code Comments: The JavadocMiner .................................................. 68  
*Ninus Khamis, René Witte, and Juergen Rilling*

Towards Approximating COSMIC Functional Size from User Requirements in Agile Development Processes Using Text Mining ...... 80  
*Ishrar Hussain, Leila Kosseim, and Olga Ormandjieva*
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Enriching of Natural Language Texts with Automatic Thematic Role Annotation</td>
<td>92</td>
</tr>
<tr>
<td>Sven J. Körner and Mathias Landhäußer</td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td></td>
</tr>
<tr>
<td>Adaptive Topic Modeling with Probabilistic Pseudo Feedback in Online Topic Detection</td>
<td>100</td>
</tr>
<tr>
<td>Guoyu Tang and Yunqing Xia</td>
<td></td>
</tr>
<tr>
<td>An Approach to Indexing and Clustering News Stories Using Continuous Language Models</td>
<td>109</td>
</tr>
<tr>
<td>Richard Bache and Fabio Crestani</td>
<td></td>
</tr>
<tr>
<td>Spoken Language Understanding via Supervised Learning and Linguistically Motivated Features</td>
<td>117</td>
</tr>
<tr>
<td>Maria Georgescul, Manny Rayner, and Pierrette Bouillon</td>
<td></td>
</tr>
<tr>
<td>Hidden Markov Models</td>
<td></td>
</tr>
<tr>
<td>Topology Estimation of Hierarchical Hidden Markov Models for Language Models</td>
<td>129</td>
</tr>
<tr>
<td>Kei Wakabayashi and Takao Miura</td>
<td></td>
</tr>
<tr>
<td>Speaker Independent Urdu Speech Recognition Using HMM</td>
<td>140</td>
</tr>
<tr>
<td>Javed Ashraf, Naveed Iqbal, Naveed Sarfraz Khattak, and Ather Mohsin Zaidi</td>
<td></td>
</tr>
<tr>
<td>Second-Order HMM for Event Extraction from Short Message</td>
<td>149</td>
</tr>
<tr>
<td>Huixing Jiang, Xiaojie Wang, and Jilei Tian</td>
<td></td>
</tr>
<tr>
<td>Querying</td>
<td></td>
</tr>
<tr>
<td>Goal Detection from Natural Language Queries</td>
<td>157</td>
</tr>
<tr>
<td>Yulan He</td>
<td></td>
</tr>
<tr>
<td>Parsing Natural Language into Content for Storage and Retrieval in a Content-Addressable Memory</td>
<td>169</td>
</tr>
<tr>
<td>Roland Hausser</td>
<td></td>
</tr>
<tr>
<td>Natural Language Interfaces</td>
<td></td>
</tr>
<tr>
<td>Vague Relations in Spatial Databases</td>
<td>177</td>
</tr>
<tr>
<td>Michael J. Minock</td>
<td></td>
</tr>
<tr>
<td>Conceptual Modeling of Online Entertainment Programming Guide for Natural Language Interface</td>
<td>188</td>
</tr>
<tr>
<td>Harry Chang</td>
<td></td>
</tr>
</tbody>
</table>
Integration of Natural Language Dialogues into the Conceptual Model of Storyboard Design ............................................. 196
Markus Berg, Antje Düsterhöft, and Bernhard Thalheim

Domain Modelling

Autonomous Malicious Activity Inspector – AMAI .................. 204
Umar Manzoor, Samia Nefti, and Yacine Rezgui

Towards Geographic Databases Enrichment ......................... 216
Khaoula Mahmoudi and Sami Faïz

On-Demand Extraction of Domain Concepts and Relationships from Social Tagging Websites ........................................ 224
Vijayan Sugumaran, Sandeep Purao, Veda C. Storey, and Jordi Conesa

Information Extraction

Analysis of Definitions of Verbs in an Explanatory Dictionary for Automatic Extraction of Actants Based on Detection of Patterns ...... 233
Noé Alejandro Castro-Sánchez and Grigori Sidorov

An Automatic Definition Extraction in Arabic Language .............. 240
Omar Trigui, Lamia Hadrich Belguith, and Paolo Rosso

Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus .................................. 248
Alexander Gelbukh, Grigori Sidorov, Eduardo Lavin-Villa, and Liliana Chanona-Hernandez

Weighted Vote Based Classifier Ensemble Selection Using Genetic Algorithm for Named Entity Recognition .......................... 256
Asif Ekbal and Sriparna Saha

Refactoring of Process Model Activity Labels .......................... 268
Henrik Leopold, Sergey Smirnov, and Jan Mendling

Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System ............................................................. 277
Euthymios Drymonas, Kalliopi Zervanou, and Euripides G.M. Petrakis

Semantic Networks & Graphs

Identifying Writers’ Background by Comparing Personal Sense Thesauri ................................................................. 288
Polina Panicheva, John Cardiff, and Paolo Rosso
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval of Similar Electronic Health Records Using UMLS Concept</td>
<td>296</td>
</tr>
<tr>
<td>Graphs</td>
<td></td>
</tr>
<tr>
<td><em>Laura Plaza and Alberto Díaz</em></td>
<td></td>
</tr>
<tr>
<td>Supporting the Abstraction of Clinical Practice Guidelines Using</td>
<td>304</td>
</tr>
<tr>
<td>Information Extraction</td>
<td></td>
</tr>
<tr>
<td><em>Katharina Kaiser and Silvia Miksch</em></td>
<td></td>
</tr>
<tr>
<td><strong>Author Index</strong></td>
<td>313</td>
</tr>
</tbody>
</table>