Lecture Notes in Artificial Intelligence 5603

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

Subseries of Lecture Notes in Computer Science

Zygmunt Vetulani Hans Uszkoreit (Eds.)

Human Language Technology

Challenges of the Information Society

Third Language and Technology Conference, LTC 2007
Poznań, Poland, October 5-7, 2007
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 Editors

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

Hans Uszkoreit
German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab
Campus D 3 1, Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
E-mail: uszkoreit@dfki.de

Library of Congress Control Number: 2009933613

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

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-04234-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-04234-8 Springer 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. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2009
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 1274607 06/3180 5 4 3 2 1 0

Preface

Half a century ago not many people had realized that a new epoch in the history of homo sapiens had just started. The term "Information Society Age" seems an appropriate name for this epoch.

Communication was without a doubt a lever of the conquest of the human race over the rest of the animate world. There is little doubt that the human race began when our predecessors started to communicate with each other using language. This highly abstract means of communication was probably one of the major factors contributing to the evolutionary success of the human race within the animal world. Physically weak and imperfect, humans started to dominate the rest of the world through the creation of communication-based societies where individuals communicated initially to satisfy immediate needs, and then to create, accumulate and process knowledge for future use.

The crucial step in the history of humanity was the invention of writing. It is worth noting that writing is a human invention, not a phenomenon resulting from natural evolution. Humans invented writing as a technique for recording speech as well as for storing and facilitating the dissemination of knowledge across the world. Humans continue to be born illiterate, and therefore teaching and conscious supervised learning is necessary to maintain this basic social skill. The invention of writing and the resulting writing-based civilizations based on know-how accumulation stimulated the development of ever more sophisticated technologies created in tune with the laws of nature and based on ever better understanding of these laws. It must be recognized, however, that many of these inventions were accidental and there is no evidence that they were in any way necessary (i.e., they might simply not have happened). The development of technologies and the production of artifacts now seems be beyond the control of any individual, group or organization. This remark is especially valid for the so-called intelligent technologies characteristic of the current epoch which merits the name Information Society Age.

The incipient Information Society is a novel kind of social structure where humans will be surrounded by a new generation of information-rich artifacts and technologies designed to collaborate with human users. Their main role will be to augment the cognitive capabilities of their owners by extending their information resources, memories, knowledge and perception. A central component of this augmented cognition is the preprocessing of huge volumes of information with the goal of exploiting it effectively for improved decision making. Although the new technologies may not be able to fully simulate human intelligence in the foreseeable future, they will certainly exceed human performance in many essential tasks just as they are already outperforming people in memory, search and numeric calculation. Artifacts made by humans "in their own image" will be intended to communicate with them.

This emerging Information Society will probably be characterized by the use of Human Language Technologies which contribute to a world where humans need to communicate not only with each other but also with the artificially created interactive and autonomous technological environment—collaborative, but possibly also hostile.

The history of human civilization tells us that humanity evolves in ways which are barely under control and which are difficult to anticipate. This development is stimulated by the human need to rise to ever greater challenges. We claim that proper development of Human Language Technologies is one of such challenges and we hope that picking up the gauntlet will result in a better, cleverer and happier world.

Human Language Technologies are already in daily use in many areas of human life. Machine translation assists Web users and human translators. Spell and grammar checkers help people to write well-formed texts. Automated telephone dialogues enable information access and transactions outside of business hours. Powerful search engines facilitate the gathering of information from the WWW. The users of these technologies still complain about ridiculous translation errors, communication deadlocks with automated call center agents and incorrect warnings of their grammar checkers.

However, the technologies keep improving at a fast pace. New exciting application areas are explored such as the analysis of opinions in Web fora and discussion groups, the automatic construction of multiple choice questions for educational testing and voice interfaces to all kinds of machinery. We may now already safely predict that in the future every new technology that is put in service will have to master human language—the most powerful and most natural communication medium of mankind.

However, during its evolution and steady adaptation to the needs of its users human language has developed properties that make it well suited for human use but especially unsuited for processing on today's digital computers. The logical complexity of human language, its omnipresent ambiguity, its vagueness and its intricate interplay of regularities and exceptions are serious obstacles to a straightforward formal modeling of the human language proficiency. Do we have to understand all the hidden regularities behind this central cognitive faculty or can we build systems that can learn these principles without understanding any of them such as toddlers, when they acquire their language through interaction?

To make the task even harder, our race does not have only one language. How can we blame technology for not mastering our language if we ourselves cannot speak and understand most of the 2,000 plus languages of the world? The multilingual global society poses another grand challenge to human language technologies. If we want to preserve the multicultural setup of our globalizing world, all surviving languages need to be supported by information technology.

Another central challenge is the different modalities of language (speech, writing, signing) and the integration of language with other media of communication

such as gestures, facial expressions, pictures and films. Each of the media and modi requires different techniques for encoding and processing that need to interact with the processing of language.

A last challenge to be mentioned here is the combination of human language technologies with a new generation of knowledge technologies, which are nowadays often called semantic technologies. The technological avant-garde of the World-Wide Web has gathered behind the powerful vision of a Semantic Web, a Web 3.0 that does connect the contents of the global information infrastructure by their shape or by stipulated hyperlinks but by their true meanings. The resulting Web could answer difficult questions and permit services based on complex inferences. To this end, powerful Web-suited knowledge representation languages have been designed. Language technology is expected to play an essential role in upgrading today's Web to version 3.0 by extracting the knowledge from texts into some structured representation suited for transformation into the Semantic Web languages.

This brief description of the main challenges for Human Language Technologies shows the complexity and diversity of the field. None of the practitioners and users of human language can keep up with the wide scope of developments in the field. This book is a snapshot of this dynamic evolution. It assembles recent results in a variety of subfields that have been selected for wider publication.

In the volume the reader will find 40 papers which are the extended and updated versions of selected papers presented at the Third Language and Technology Conference. The selection was made from the 105 high-quality technical contributions (written by 275 authors) accepted on the basis of anonymous peer reviewing for conference presentation by the international jury. The editors basically followed the ranking established on the ground of reviewers' assessments. We are conscious of the sometimes subjective character of our decisions. We are also conscious that in some cases very interesting and promising papers were left out of the selection, especially where they referred to on-going projects or contributed interesting but partial or over-technical solutions. These 40 selected papers represent the considerable effort of 106 authors from Australia, Estonia, France, Germany, Italy, Japan, Mexico, Poland, Portugal, Romania, Slovenia, Spain, Sweden, Tunisia, Turkey, UK and USA.

For the convenience of the reader we have "structured" this book into thematic chapters. Assigning papers to chapters was a difficult task, as in most cases they addressed more than one thematic issue, so some of our decisions may look arbitrary. This is due to the obvious fact that language processing shares one common feature with language itself: it is complex and multidimensional. In particular there is no natural ordering of the thematic chapters we have proposed except perhaps the order in which a humans appear to process language production in a linguistic communication act, i.e., starting with (spoken) speech analysis, through morphology etc. To follow this natural order we start the volume with the Speech Processing chapter. Within chapters the papers are presented in alphabetical order with respect to the first author name.

Contributions have been grouped into eight chapters. These are:

1. Speech Processing (6)
2. Computational Morphology (3)
3. Parsing and Generation (9)
4. Computational Semantics (5)
5. Digital Language Resources (6)
6. WordNet (3)
7. Information Retrieval/Extraction (5)
8. Machine Translation (3)

The first chapter, Speech Processing, starts with a paper that discusses a solution to the practical problem of automatically detecting the start and end points of broadcast news (Amaral, Trancoso). It is followed by a contribution describing a method of speech feature extraction using morphological signal processing (Drgas, Dabrowski). Next comes an article contributing to the development of a spoken language interface (European Portuguese) for mobile devices (Freitas, Calado, Barros, Sales Dias). The next paper is about a bilingual—Spanish, Valentian—speech recognition system (Luján-Mares, Martínez-Hinarejos, Alabau). The fifth paper in this chapter is a presentation of a project aiming at an annotated speech corpus of spoken dialogues for Polish (Mykowiecka, Marasek, Marciniak, Rabiega-Wiśniewska, Gubrynowicz). The chapter closes with a report on the spoken corpus research resulting with triphone statistics for Polish (Ziółko, Gałka, Manandhar, Wilson, Ziółko).

In the Computational Morphology chapter, the reader will find three papers. The first one is a description of a morphological generator for Romanian word forms out of lemmas and morpho-lexical descriptions, initially designed to serve MT purposes (Irimia). The next paper in this chapter presents a set of tools for the morphological pre-processing of Polish texts in order to obtain "a reasonable input" for further processing (e.g., parsing) (Sagot). Although for most of highly inflected Indo European languages the problem of the formal description of their morphology have been solved and resources and tools exist, the next paper, the last one in this part, presents an alternative approach consisting in application of relational data-base methodology (Woliński).

The Parsing and Generation chapter is dominated by contributions to parsing. It opens with a proposal to use the idea of spreadsheets to represent parse tree forests in a human-readable form (Bień). The next contribution shows how to use a robust deep parser to solve named entity metonymy (Brun, Ehrmann, Jacquet). It is followed by a contribution presenting an application of shallow parsing used as a morphological disambiguation engine (Buczyński, Przepiórkowski). The fourth contribution proposes an object-oriented technique of text generation aimed at NLG designers who are not specialists in computational linguistics (Cullen, O'Neill, Hanna). The fifth paper is a contribution to the study of the relationship between the analogies of form between chunks and corresponding analogies of meanings (for Japanese) (Lepage, Migeot, Guillerm). Another application of shallow parsing as a disfluency detector in transcribed

Estonian speech is described in the next paper (Müürisep, Nigol). The next paper is again an application of a parser, this time to perform lexical corrections (Nicolas, Sagot, Molinero, Farré, de la Clergerie). This is followed by a presentation of two parsers based on recursive transition nets with a string output together with a scrupulous efficiency considerations (Sastre-Martínez, Forcada). Finally the reader will find a description of a deep parser for Polish based on Syntactical Groups Grammar (Suszczańska, Szmal, Simiński).

The first of the five papers in the Computational Semantics chapter is about using supervised learning techniques to augment a semantically complex lexicon of texts with sentimental valence (Argamon, Bloom, Esuli, Sebastiani). The second one deals with the problem of the evaluation of text readability, i.e., automatic identification of easy-to-read fragments in a text corpus, which seems to have important practical potential (vor der Brück, Hartrumpf). Another application with a possible practical importance is the tagger for temporal expressions described in the next paper (Mazur, Dale). The data mining techniques of detecting the sequential structure of texts for text segmentation purposes are described in the fourth paper in this section (Recanati, Rogovski, Bennani). The last paper describes a machine learning-based algorithm of anaphora resolution for Turkish (Yıldırım, Kılıçaslan, Yıldız).

We have included six papers in the Digital Language Resources chapter. These are contributions concerning tools, methodology and resources. The chapter starts with the proposal of a graphical tool to query highly formalized lexical resources observing the LFM standard (Ben Abderrahmen, Gargouri, Jmaiel). The next paper is about the acquisition of new terminology with special attention to parallel corpora as a source (Dura, Gawronska). The third contribution in this chapter presents a system for a text corpus construction using the Internet (Kulików). The fourth paper is about the methodology of building an important grammatical resource for English, namely a lexicon-grammar of predicative nouns (Malik, Royauté). (This is one of the rare examples of an important research topic on a language where English so far has been less profoundly investigated than some other languages—e.g., French.) A large-scale terminological lexicon in biology supporting text mining and information extraction application is presented in the fifth paper (Quochi, Del Gratta, Sassolini, Bartolini, Monachini, Calzolari). The chapter closes with a contribution on extracting Romanian verbo-nominal collocations from large tagged and annotated corpora (Todirascu, Gledhill, Stefanescu).

The WordNet-related chapter has three contributions. The first is about extending English WordNet with morphosemantic links to aid automatic reasoning with Wordnet (Fellbaum, Osherson, Clark). The remaining two contributions in this chapter are about two different approaches for constructing wordnets. In the first, the author proposes a method for the automatic construction of a wordnet for Slovenian through the use of a multilingual parallel corpus (the Orwell Corpus) and several wordnets (Fišer). The second paper presents a completely different approach: an algorithm for the manual construction of a wordnet (designed and first used for Polish but in fact language-independent for a large class

of Indo European languages) using classical monolingual dictionaries (Vetulani, Walkowska, Obrębski, Marciniak, Konieczka, Rzepecki).

The Information Retrieval and/or Extraction chapter opens with a paper on the identification of knowledge about terms in specialized texts using definition extraction techniques (Alarcón, Sierra, Bach) . The next paper describes a hybrid method for extracting pairs of similar terms from Japanese documents (Fukumoto, Suzuki). The next three papers are contributions to name identification in texts, which appears to be an important issue in the IR/IE field. The first focuses on the problem of proper name search on a noisy and difficult data set (Miller, Arehart), the second considers string distance measures for name lemmatization (Piskorski, Sydow, Wieloch), whereas the third is about the identification of co-reference of person names, which may be useful for people search tasks (Popescu, Magnini).

Finally, we have included three papers in the Machine Translation chapter. The first one is a contribution to the problem, often marginalized, of hapaxes in text alignment (Lardilleux, Lepage). The second is about a statistical machine translation from Slovenian to English (Maučec, Brest). In the last paper the authors propose a method of translation of English nominal compounds into Polish and Internet-based validation of translation output (Stępień, Podlejski).

July 2009

Zygmunt Vetulani
Hans Uszkoreit

Committee Listings

Organizing Committee

Zygmunt Vetulani - Conference Chair
Filip Graliński
Paweł Konieczka
Maciej Lison
Jacek Marciniak
Tomasz Obrębski
Jędrzej Osiński
Justyna Walkowska
(All at the Adam Mickiewicz University, Poznań, Poland)

Program Committee

Victoria Arranz
Anja Belz
Janusz S. Bień
Christian Boitet
Leonard Bolc
Nicoletta Calzolari
Nick Campbell
Julie Carson-Berndsen
Khalid Choukri
Adam Dąbrowski
Grażyna Demenko
Elżbieta Dura
Katarzyna Dziubalska-Kołaczyk
Tomaz Erjavec
Cedrick Fairon
Christiane Fellbaum
Maria Gavrilidou
Aleksander Gerd
Dafydd Gibbon
Stefan Grocholewski
Franz Guenthner
Roland Hausser
Wacław Iszkowski
Margaret King
Orest Kossak
Eric Laporte
Gerard Ligozat
Natalia Loukachevitch
Wiesław Lubaszewski
Bente Maegaard
Joseph Mariani
Jacek Martinek
Václav Matousek
Keith J. Miller
Nicholas Ostler
Karel Pala
Pavel S. Pankov
Marcin Paprzycki
Patrick Paroubek
Emil Pływaczewski
Adam Przepiórkowski
Reinhard Rapp
Mike Rosner
Justus Roux
Vasile Rus
Frédérique Ségond
Włodzimierz Sobkowiak
Marek Świdziński
Ryszard Tadeusiewicz
Dan Tufiş
Hans Uszkoreit
Zygmunt Vetulani
Piek Vossen
Tom Wachtel
Jan Węglarz
Richard Zuber

Table of Contents

I Speech Processing

Exploring the Structure of Broadcast News for Topic Segmentation 1
Rui Amaral and Isabel Trancoso

Application of Slope Filtering to Robust Spectral Envelope Extraction for Speech/Speaker Recognition 13
Szymon Drgas and Adam Dabrowski

Spoken Language Interface for Mobile Devices 24
João Freitas, António Calado, Maria João Barros, and Miguel Sales Dias

A Study on Bilingual Speech Recognition Involving a Minority Language 36
Míriam Luján-Mares, Carlos-D. Martínez-Hinarejos, and Vicente Alabau

Annotated Corpus of Polish Spoken Dialogues 50
Agnieszka Mykowiecka, Krzysztof Marasek, Małgorzata Marciniak, Joanna Rabiega-Wiśniewska, and Ryszard Gubrynowicz

Triphone Statistics for Polish Language 63
Bartosz Ziółko, Jakub Gałka, Suresh Manandhar, Richard C. Wilson, and Mariusz Ziółko

II Computational Morphology

ROG – A Paradigmatic Morphological Generator for Romanian 74
Elena Irimia

Building a Morphosyntactic Lexicon and a Pre-syntactic Processing Chain for Polish 85
Benoît Sagot

A Relational Model of Polish Inflection in *Grammatical Dictionary of Polish* 96
Marcin Woliński

III Parsing and Generation

Syntactic Spreadsheets: In Search for a Human-Readable Representation of Parse Tree Forests . . . 107
Janusz S. Bień

A Hybrid System for Named Entity Metonymy Resolution . . . 118
Caroline Brun, Maud Ehrmann, and Guillaume Jacquet

Spejd: A Shallow Processing and Morphological Disambiguation Tool . . . 131
Aleksander Buczyński and Adam Przepiórkowski

Flexible Natural Language Generation in Multiple Contexts . . . 142
Caroline Cullen, Ian O'Neill, and Philip Hanna

A Measure of the Number of True Analogies between Chunks in Japanese . . . 154
Yves Lepage, Julien Migeot, and Erwan Guillerm

Shallow Parsing of Transcribed Speech of Estonian and Disfluency Detection . . . 165
Kaili Müürisep and Helen Nigol

Mining Parsing Results for Lexical Correction: Toward a Complete Correction Process of Wide-Coverage Lexicons . . . 178
Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, and Éric de La Clergerie

Efficient Parsing Using Recursive Transition Networks with Output . . . 192
Javier M. Sastre-Martínez and Mikel L. Forcada

The Deep Parser for Polish . . . 205
Nina Suszczańska, Przemysław Szmal, and Krzysztof Simiński

IV Computational Semantics

Automatically Determining Attitude Type and Force for Sentiment Analysis . . . 218
Shlomo Argamon, Kenneth Bloom, Andrea Esuli, and Fabrizio Sebastiani

A Readability Checker Based on Deep Semantic Indicators . . . 232
Tim vor der Brück and Sven Hartrumpf

The DANTE Temporal Expression Tagger . . . 245
Paweł Mazur and Robert Dale

Hybrid Unsupervised Learning to Uncover Discourse Structure 258
Catherine Recanati, Nicoleta Rogovski, and Younès Bennani

Pronoun Resolution in Turkish Using Decision Tree and Rule-Based Learning Algorithms 270
Savaş Yıldırım, Yılmaz Kılıçaslan, and Tuğba Yıldız

V Digital Language Resources

LMF-QL: A Graphical Tool to Query LMF Databases for NLP and Editorial Use 279
Mehdi Ben Abderrahmen, Bilel Gargouri, and Mohamed Jmaiel

Novelty Extraction from Special and Parallel Corpora 291
Elżbieta Dura and Barbara Gawronska

Construction of Text Corpus of Polish Using the Internet 303
Sławomir Kulików

A Predicate Database for Assisting the Design of a Lexicon-Grammar of Predicative Nouns 312
Mohamed Mahdi Malik and Jean Royauté

A Standard Lexical-Terminological Resource for the Bio Domain 325
Valeria Quochi, Riccardo Del Gratta, Eva Sassolini, Roberto Bartolini, Monica Monachini, and Nicoletta Calzolari

Extracting Collocations in Contexts 336
Amalia Todirascu, Christopher Gledhill, and Dan Stefanescu

VI WordNet

Putting Semantics into WordNet's "Morphosemantic" Links 350
Christiane Fellbaum, Anne Osherson, and Peter E. Clark

Leveraging Parallel Corpora and Existing Wordnets for Automatic Construction of the Slovene Wordnet 359
Darja Fišer

An Algorithm for Building Lexical Semantic Network and Its Application to PolNet - Polish WordNet Project 369
Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Jacek Marciniak, Paweł Konieczka, and Przemysław Rzepecki

VII Information Retrieval/Extraction

ECODE: A Definition Extraction System 382
Rodrigo Alarcón, Gerardo Sierra, and Carme Bach

Using Graph-Based Indexing to Identify Subject-Shift in Topic Tracking 392
Fumiyo Fukumoto and Yoshimi Suzuki

Result Aggregation for Knowledge-Intensive Multicultural Name Matching 405
Keith J. Miller and Mark Arehart

Comparison of String Distance Metrics for Lemmatisation of Named Entities in Polish 413
Jakub Piskorski, Marcin Sydow, and Karol Wieloch

An Iterative Model for Discovering Person Coreferences Using Name Frequency Estimates 428
Octavian Popescu and Bernardo Magnini

VIII Machine Translation

Hapax Legomena: Their Contribution in Number and Efficiency to Word Alignment 440
Adrien Lardilleux and Yves Lepage

Statistical Machine Translation from Slovenian to English Using Reduced Morphology 451
Mirjam Sepesy Maučec and Janez Brest

Semi-automatic Creation of a Dictionary of Nominal Compounds 461
Tomasz Stępień and Bartosz Podlejski

Author Index 471