Frontiers in Artificial Intelligence and Applications

FAIA covers all aspects of theoretical and applied artificial intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes. The FAIA series contains several sub-series, including “Information Modelling and Knowledge Bases” and “Knowledge-Based Intelligent Engineering Systems”. It also includes the biennial ECAI, the European Conference on Artificial Intelligence, proceedings volumes, and other ECCAI – the European Coordinating Committee on Artificial Intelligence – sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.

Series Editors:
J. Breuker, N. Guarino, J.N. Kok, J. Liu, R. López de Mántaras,
R. Mizoguchi, M. Musen, S.K. Pal and N. Zhong

Volume 247

Recently published in this series

Vol. 244. S. Scheider, Grounding Geographic Information in Perceptual Operations
Vol. 240. M. Virvou and S. Matsuura (Eds.), Knowledge-Based Software Engineering – Proceedings of the Tenth Joint Conference on Knowledge-Based Software Engineering
Vol. 238. A. Respicio and F. Burstein (Eds.), Fusing Decision Support Systems into the Fabric of the Context
Vol. 236. M.A. Biasiotti and S. Faro (Eds.), From Information to Knowledge – Online Access to Legal Information: Methodologies, Trends and Perspectives

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)
Preface

This volume contains papers presented at the Fifth International Conference “Human Language Technologies – The Baltic Perspective” (Baltic HLT 2012), held in Tartu, Estonia on 4–5 October 2012.

Since its first edition in 2004, Baltic HLT has served as a special venue for new and ongoing work in computational linguistics and related disciplines in the Baltic states as well as in a broader geographical perspective.

The main aim of this conference is to provide a forum for the sharing of new ideas and recent advances in human language processing and to promote cooperation between the research communities of computer science and linguistics from the Baltic countries and the rest of the world. The conference brings together scientists, developers, providers and users to discuss state-of-the-art of human language technologies in the Baltic countries, to exchange information and to discuss problems, to find new synergies and to promote initiatives for international cooperation.

The call for papers for the fifth Baltic HLT laid special emphasis on multilinguality in language resources and on applications of human language technology, while also encouraging the potential authors to submit papers on other subfields of computational linguistics and related disciplines.

51 submissions were received; each submission was evaluated by at least two reviewers.

The Programme Committee consisted of 25 members from 13 different countries. Based on their scores and the comments they provided on the content and quality of the papers, 20 long papers and 20 posters or demos were accepted for presentation and publication.

The accepted submissions cover a wide range of topics: morphological disambiguation, dependency syntax and valency, computational semantics, named entities, dialogue modeling, terminology extraction and management, machine translation, corpus and parallel corpus compiling, speech modeling and multimodal communication. A few papers give a general overview of the state of the art of the human language technology and/or language resources in the Baltic states.

Completing the programme are the invited lectures by Lori Lamel “Multilingual Speech Processing Activities in Quaero: application to multimedia search in unstructured data” and Bente Maegaard “A Multilingual Research Infrastructure”.

We wish to express our gratitude to the members of the Programme Committee who worked hard to review all submissions.

We also want to thank the organizers and supporters of this conference: Institute of Computer Science, University of Tartu and Estonian Ministry of Education and Research as funder of National Programme for Estonian Language Technology. The conference is also supported by the CLARIN and META-NORD projects.

Arvi Tavast
Kadri Muischnek
Kadri Vider
Mare Koit
The publication of these Proceedings was supported by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS.
Programme Committee

Arvi Tavast (chair), Institute of the Estonian Language, Estonia
Tanel Alumäe, Institute of Cybernetics, Tallinn University of Technology, Estonia
Eckhard Bick, University of Southern Denmark, Denmark
Koenraad De Smedt, University of Bergen, Norway
Mark Fishel, University of Zurich, Switzerland
Markus Forsberg, University of Gothenburg, Sweden
Gintarė Grigonytė, Vytautas Magnus University, Lithuania
Karin Harbusch, University of Koblenz-Landau, Germany
Heiki-Jaan Kaalep, University of Tartu, Estonia
Kaarel Kaljurand, University of Zurich, Switzerland
Adam Kilgarriff, Lexical Computing Ltd, UK
Ramón López-Cózar Delgado, University of Granada, Spain
Bente Maegaard, University of Copenhagen, Denmark
Rūta Marcinkevičienė, Vytautas Magnus University, Lithuania
Beata Megyesi, Uppsala University, Sweden
Costanza Navarretta, University of Copenhagen, Denmark
Joakim Nivre, Uppsala University, Sweden
Bolette Sandford Pedersen, University of Copenhagen, Denmark
Fabio Rinaldi, University of Zurich, Switzerland
Eiríkur Rögnvaldsson, University of Iceland, Iceland
Inguna Skadiņa, IMCS/Tilde, Latvia
Jörg Tiedemann, Uppsala University, Sweden
Martin Volk, University of Zurich, Switzerland
Peter Wittenburg, Max Planck Institute for Psycholinguistics, Netherlands
Anssi Yli-Jyrä, University of Helsinki, Finland
Contents

Preface v

Arvi Tavast, Kadri Muischnek, Kadri Vider and Mare Koit

Programme Committee vii

Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data 1

Lori Lamel

A Multilingual Research Infrastructure 9

Bente Maegaard

Transcription System for Semi-Spontaneous Estonian Speech 10

Tanel Alumäe

Towards the Automatic Extraction of Term-Defining Contexts in Lithuanian 18

Agnė Bielinskienė, Loïc Boïzou, Jolanta Kovalevskaitė and Andrius Utka

Automatic Inference of Base Forms for Multiword Terms in Lithuanian 27

Loïc Boïzou, Gintarė Grigonytė, Erika Rimkutė and Andrius Utka

Data Pre-Processing to Train a Better Lithuanian-English MT System 36

Daiga Deksne and Raivis Skadiņš

Perception of Russian Vowels in Singing 42

Karina Evgrafova and Vera Evdokimova

In-Domain Data FTW 50

Mark Fishel

Improving SMT by Using Parallel Data of a Closely Related Language 58

Petra Galuščáková and Ondřej Bojar

Terminology Extraction from Comparable Corpora for Latvian 66

Tatiana Gornostay, Anita Ramm, Ulrich Heid, Emmanuel Morin, Rima Harastani and Emmanuel Planas

Change of Biomedical Domain Terminology Over Time 74

Gintarė Grigonytė, Fabio Rinaldi and Martin Volk

A Trivial Method for Choosing the Right Lemma 82

Heiki-Jaan Kaalep, Riin Kirt and Kadri Muischnek

Geoinformational Database of Lithuanian Toponyms 90

Dalia Kačinaitė-Vrubliauskienė

Cross-Linking Experience of Estonian WordNet 96

Neeme Kahusk, Heili Orav and Kadri Vare
Automatic Generation of Specialized Dictionaries Using the Dictionary Writing System EELEX
   Jelena Kallas and Margit Langemets

Managing Word Form Variation of Text Retrieval in Practice – Why Five Character Truncation Takes It All?
   Kimmo Kettunen

Towards Automatic Recognition of the Structure of Estonian Directory Inquiries
   Mare Kõit

Adaptation of Morpheme-Based Speech Recognition for Foreign Entity Names
   André Mansikkaniemi and Mikko Kurimo

Towards Audiovisual TTS in Estonian
   Einar Meister, Sascha Fagel and Rainer Metsvahi

Multimodal Corpus of Speech Production: Work in Progress
   Einar Meister and Lya Meister

Towards a Latvian Valency Lexicon
   Gunta Nešpore, Baiba Saulīte, Normunds Grūzītis and Ginta Garkāje

Creation of HMM-Based Speech Model for Estonian Text-to-Speech Synthesis
   Tõnis Nurk

Towards Named Entity Annotation of Latvian National Library Corpus
   Peteris Paikens, Ilze Auzina, Ginta Garkāje and Madara Paegle

MT Adaptation for Under-Resourced Domains – What Works and What Not
   Mārcis Pinnis and Raivis Skadiņš

Syntactic Issues Identified Developing the Latvian Treebank
   Lauma Pretkalniņa and Laura Rituma

How Does the Choice of Morphological Analyser Influence the Quality of Syntactical Analysis?
   Tiina Puolakainen

Knowledge Acquisition Tool for Dialogue Systems
   Raul Sirel

Dynamic User Interfaces for Synchronous Encoding and Linguistic Uniforming of Textual Clinical Data
   Raul Sirel

Noisy-Channel Spelling Correction Models for Estonian Learner Language Corpus Lemmatisation
   Kairit Sirts

Linguistically Motivated Evaluation of English-Latvian Statistical Machine Translation
   Inguna Skadiņa, Kristīne Levāne-Petrova and Guna Rābante
Simplifying Email Management
Tõnu Tamme
230

Intensity of Estonian Emotional Speech
Kairi Tamuri
238

Multimodal Feedback Signaling in Finnish
Emmi Toivio and Kristiina Jokinen
247

Adjusting to User’s Style in Dialogue Systems
Margus Treumuth and Margus Ojamurme
256

Lithuanian-Latvian-Lithuanian Parallel Corpus
Andrius Utka, Kristīne Levāne-Petrova, Agnė Bielinskienė, Jolanta Kovalevskaitė, Erika Rimkutė and Daira Vēvere
260

Latvian Language Resources and Tools: Assessment, Description and Sharing
Andrejs Vasiljevs and Inguna Skadiņa
265

Strategic Importance of Language Technology in Estonia
Kadri Vider, Krista Liin and Neeme Kahusk
273

Ontological Features of Entities in Motion Events and Their Role in the Semantic Analysis of Sentences
Haldur Õim
280

Prosodic Patterns of Estonian Words: A Corpus-Based Description Using Spontaneous Speech
Rena Nemoto and Martine Adda-Decker
286

Subject Index
295

Author Index
297