PROCEEDINGS of the 16th Nordic Conference of Computational Linguistics

NODALIDA-2007

Editors: Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit

University of Tartu
Tartu 2007

Editors: Joakim Nivre, Heiki-Jaan Kaalep,
Kadri Muischnek and Mare Koit
J. Liivi 2, 50409 Tartu, Estonia

Published by the University of Tartu
University of Tartu, Ülikooli 18, 50090 Tartu, Estonia
Tartu, 2007

ISBN 978-9985-4-0514-7 (CD-ROM)
ISBN 978-9985-4-0513-0 (online)

Contents

Preface

Language technology research in Northern Europe is thriving. One clear sign of this is the fact that the NODALIDA conference (also known as the Nordic Conference of Computational Linguistics) has grown to the point where it is no longer manageable to let the local organization committee do all the work in putting together the program. Hence, the need for a separate program committee.

When trusted with the responsibility of chairing the program committee, the program chair made an informal survey of expectations in the community through an e-mail questionnaire sent out on the NODALI list. (Thanks to everyone who responded.) Four things emerged clearly from this survey: People wanted a high-quality technical track with review of full papers. However, people also wanted an opportunity for students to present their work, and an opportunity to get feedback on work in progress in the form of posters. Last but not least, people wanted workshops. Hence, we give you regular paper sessions, student paper sessions, a poster session, and a full day of workshops. As a bonus, we also give you two distinguished keynote speakers, Diana McCarthy and Walter Daelemans, the first business meeting of the newly established Northern European Association for Language Technology (NEALT), and a tutorial on the Estonian language, hoping to establish a new NODALIDA tradition of local language tutorials. (Thanks to Koenraad de Smedt for this great idea.)

When issuing the call for workshops, regular papers, student papers, and posters, we were unsure whether there would be enough work going on in Northern Europe and the rest of the world to fill all the categories. The community response surpassed all our expectations. We received 41 regular paper submissions, 24 student paper submissions, 32 poster submissions, and 7 workshop proposals, for a total of 104 submissions, an all time record for NODALIDA. (Thanks to everyone who submitted their work.) Moreover, a fair share of these submissions came from countries outside our region, clearly showing that NODALIDA, while remaining a conference with a strong regional character, is also being recognized in the rest of the world. In the final program, there are 26 regular papers, 12 student papers, 26 posters, and 4 workshops. We want to thank the program committee and all our 69 reviewers for their hard work in putting together the program.

Finally, NODALIDA 2007 is special not only for being the biggest ever in terms of submissions, but also for being the first NODALIDA held in Estonia, in the beautiful city of Tartu, at one of the oldest universities in the region, founded

in 1632. We want to thank the local organization committee for all their hard work to welcome the NODALIDA participants in Tartu.

We wish you all an enjoyable NODALIDA 2007!

Joakim Nivre
Program Chair
NODALIDA 2007

Mare Koit and Tiit Roosmaa
Local Co-Chairs
NODALIDA 2007

Committees

Program Committee

Joakim Nivre (chair), Växjö University and Uppsala University
Helena Ahonen-Myka, University of Helsinki
Daniel Hardt, Copenhagen Business School
Kristiina Jokinen, University of Helsinki and University of Tartu
Pierre Nugues, Lund University
Stephan Oepen, University of Oslo, NTNU Trondheim and Stanford University
Patrizia Paggio, University of Copenhagen
Torbjørn Svendsen, NTNU Trondheim

Local Organization Committee

Mare Koit (co-chair), University of Tartu
Tiit Roosmaa (co-chair), University of Tartu
Urve Talvik, University of Tartu
Heli Uibo, University of Tartu
Kadri Vider, University of Tartu

Reviewers

Lars Ahrenberg
Ingunn Amdal
Beáta Bandmann Megyesi
Francis Bond
Janne Bondi Johannessen
Lars Borin
Matthias Buch-Kromann
Rolf Carlson
Mathias Creutz
Antoine Doucet
Laila Dybkjær
Helge Dyvik
Eva Ejerhed
Björn Gambäck
Barbara Gawronska
Jerneja Zganec Gros
Nina Grønnum
Petter Haugereid
Peter Juel Henrichsen
Merle Horne
Hannes Högni Vilhjálmsson
Richard Johansson
Magne H. Johnsen
Arne Jönsson
Viggo Kann
Jussi Karlgren
Sabine Kirchmeier-Andersen
Ola Knutsson
Mare Koit
Jacques Koreman
Kimmo Koskenniemi
Mikko Kurimo
Leena Kuure
Knut Kvale
Juha-Pertti Laaksonen
Torbjörn Lager
Birger Larsen
Krister Lindén
Jan Tore Lønning
Ramón López-Cózar Delgado
Bodil Nistrup Madsen
Bente Maegaard
Jean-Claude Martin
Bilyana Martinovski
Michael McTear
Tor Andre Myrvoll
Costanza Navarretta
Anders Nøklestad
Torbjørn Nordgård
Bjarne Ørsnes
Maria Teresa Pazienza
Bolette Pedersen
Jussi Piitulainen
Ari Pirkola
Aarne Ranta
Victoria Rosén
Eiríkur Rögnvaldsson
Rune Sætre
Anders Søgaard
Markku Turunen
Wim van Dommelen
Martin Volk
Jürgen Wedekind
Stefan Werner
Mats Wirén
Roman Yangarber
Zhang Yi
Anssi Yli-Jyrä
Fabio Massimo Zanzotto

Conference program
NODALIDA-2007 Main Conference

Friday, May 25, 2007

9.00–10.30 **Opening Session**
9.00– 9.30 Opening
9.30–10.30 Invited Talk
Evaluating Automatic Approaches for Word Meaning Discovery and Disambiguation using Lexical Substitution
Diana F. McCarthy, University of Sussex
10.30–11.00 **Coffee Break**
11.00–12.30 **Parallel Paper Sessions**

	Parsing	**Multilingual Resources and Translation**	**Speech Technology**
11.00–11.30	*IceParser: An Incremental Finite-State Parser for Icelandic* Hrafn Loftsson and Eiríkur Rögnvaldsson	*Using Danish as a CG Interlingua: A Wide-Coverage Norwegian-English Machine Translation System* Eckhard Bick and Lars Nygaard	*Automatic Compound Word Reconstruction for Speech Recognition of Compounding Languages* Tanel Alumäe
11.30–12.00	*Combining Contexts in Lexicon Learning for Semantic Parsing* Richard Socher, Chris Biemann and Rainer Osswald	*Identifying Cross Language Term Equivalents Using Statistical Machine Translation and Distributional Association Measures* Hans Hjelm	*Multivariate Cepstral Feature Compensation on Band-limited Data for Robust Speech Recognition* Nicolas Morales, John H. L. Hansen, Doroteo T. Toledano and Javier Garrido
12.00–12.30	*Polynomial Charts for Totally Unordered Languages* Anders Søgaard	*The Swedish-Turkish Parallel Corpus and Tools for its Creation* Beata B. Megyesi and Bengt Dahlqvist	*Development of Text-To-Speech System for Latvian* Kārlis Goba and Andrejs Vasiļjevs

12.30–14.00 **Lunch**

14.00–15.30 **Parallel Paper Sessions**

	Treebanks	**Information Extraction and Summarization**	**Phonetics and Speech**
14.00–14.30	*Theoretically Motivated Treebank Coverage* Victoria Rosén and Koenraad De Smedt	*Time Extraction from Real-time Generated Football Reports* Markus Borg	*Utterance-initial Duration of Finnish Nonplosive Consonants* Tuomo Saarni, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho and Tapio Salakoski
14.30–15.00	*Extended Constituent-to-Dependency Conversion for English* Richard Johansson and Pierre Nugues	*The Extraction of Trajectories from Real Texts Based on Linear Classification* Hanjing Li, Tiejun Zhao, Sheng Li, Jiyuan Zhao	*An Advanced Speech Corpus for Norwegian* Janne Bondi Johannessen, Kristin Hagen, Joel James Priestley and Lars Nygaard
15.00–15.30	*Clausal Coordinate Ellipsis in German: The TIGER Treebank as a Source of Evidence* Karin Harbusch and Gerard Kempen	*Widening the HolSum Search Scope* Martin Hassel and Jonas Sjöbergh	*Spoken Document Retrieval in a Highly Inflectional Language* Inger Ekman and Kalervo Järvelin

15.30–16.00 **Coffee Break**

16.00–17.00 **Poster Session 1**

LinES: An English-Swedish Parallel Treebank
Lars Ahrenberg

Posterior Probability Based Confidence Measures Applied to a Children's Speech Reading Tracking System
Daniel Bolanos and Wayne H. Ward

Estonian-English Statistical Machine Translation: the First Results
Mark Fishel, Heiki-Jaan Kaalep and Kadri Muischnek

A Hybrid Constituency-Dependency Parser for Swedish
Johan Hall, Joakim Nivre and Jens Nilsson

Íslenskur Orðasjóður – Building a Large Icelandic Corpus
Erla Hallsteinsdóttir, Thomas Eckart, Chris Biemann, Uwe Quasthoff and Matthias Richter

A Survey and Classification of Methods for (Mostly) Unsupervised Learning

Harald Hammarström

Marvina – A Norwegian Speech-Centric, Multimodal Visitor Guide
Ole Hartvigsen, Erik Harborg, Tore Amble and Magne H. Johnsen

A Norwegian Letter-to-Sound Engine with Danish as a Catalyst
Peter Juel Henrichsen

Dialogue Simulation and Context Dynamics for Dialogue Management
Simon Keizer and Roser Morante

Managing Keyword Variation with Frequency Based Generation of Word Forms in IR
Kimmo Kettunen

Developing and Evaluating a Searchable Swedish-Thai Lexicon
Wanwisa Khanaraksombat and Jonas Sjöbergh

Identification of Entity References in Hospital Discharge Letters
Dimitrios Kokkinakis and Anders Thurin

Lexical Parameters, Based on Corpus Analysis of English and Swedish Cancer Data, of Relevance for NLG
Dimitrios Kokkinakis, Maria Toporowska Gronostaj, Catalina Hallett and David Hardcastle

17.00–18.00 **Poster Session 2**

Anatomy of an XML-based Text Corpus Server
Mikko Lounela

Perceptual Assessment of the Degree of Russian Accent
Lya Meister

Terminology Extraction and Term Ranking for Standardizing Term Banks
Magnus Merkel and Jody Foo

Representing Calendar Expressions with Finite-State Transducers that Bracket Periods of Time on a Hierachical Timeline
Jyrki Niemi and Kimmo Koskenniemi

Parsing Manually Detected and Normalized Disfluencies in Spoken Estonian
Helen Nigol

Designing a Speech Corpus for Estonian Unit Selection Synthesis
Liisi Piits, Meelis Mihkla, Tõnis Nurk and Indrek Kiissel

Evaluating Evaluation Measures
Ines Rehbein and Josef van Genabith

Development of a Modern Greek Broadcast-News Corpus and Speech Recognition System
Jürgen Riedler and Sergios Katsikas

Role of Different Spectral Attributes in Vowel Categorization: the Case of Udmurt
Janne Savela, Stina Ojala, Olli Aaltonen and Tapio Salakoski

Recreating Humorous Split Compound Errors in Swedish by Using Grammaticality
Jonas Sjöbergh and Kenji Araki

A Re-examination of Question Classification
Håkan Sundblad

Interpretation of Yes/No Questions as Metaphor Recognition
Tarmo Truu, Haldur Õim and Mare Koit

Rule-based Logical Forms Extraction
Cenny Wenner

19.00– Conference Dinner

Saturday, May 26, 2007

9.00–10.00 **Plenary Session**
Invited Talk
Text Analysis and Machine Learning for Stylometrics and Stylogenetics
Walter Daelemans, University of Antwerp

10.00–10.30 **Coffee Break**

10.30–12.30 **Parallel Student Sessions**

	Spoken Language Processing	**Multilingual Resources and Translation**	**Natural Language Processing**
10.30–11.00	*Clause Boundary Detection in Transcribed Spoken Language* Fredrik Jørgensen	*Memory-Based Learning of Word Translation* Maria Holmqvist	*Unmediated Data-Oriented Generation* Dave Cochran
11.00–11.30	*The Effects of Disfluency Detection in Parsing Spoken Language* Fredrik Jørgensen	*Using Parallel Corpora to Create a Greek-English Dictionary with UPLUG* Konstantinos Charitakis	*Linguistically Fuelled Text Similarity* Björn Andrist and Martin Hassel

11.30–12.00	*Tagging a Norwegian Speech Corpus* Anders Nøklestad and Åshild Søfteland	*Using a Wizard of Oz as a Baseline to Determine Which System Architecture Is the Best for a Spoken Language Translation System* Marianne Starlander	*Decomposing Swedish Compounds Using Memory-Based Learning* Karin Friberg
12.00-12.30	*Initial Experiments with Estonian Speech Recognition* Anton Ragni	*Grammar Sharing Techniques for Rule-based Multilingual NLP Systems* Marianne Santaholma	*A Method for Resolution of Temporal Expressions in Estonian Natural Language Dialogue Systems* Margus Treumuth

12.30–14.00 **Lunch**
14.00–15.30 **Parallel Paper Sessions**

	Parsing and Translation	**Machine Learning**	**Dialogue**
14.00–14.30	*Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order* Guntis Bārzdiņs, Normunds Grūzītis, Gunta Nešpore and Baiba Saulīte	*Comparison of the Self-Organizing Map and Multidimensional Scaling in Analysis of Estonian Emotion Concepts* Toomas Kirt and Ene Vainik	*Interview and Delivery: Dialogue Strategies for Conversational Recommender Systems* Pontus Wärnestål, Lars Degerstedt and Arne Jönsson
14.30–15.00	*Comparing French PP-attachment to English, German and Swedish* Martin Volk and Frida Tidström	*Evaluating Stages of Development in Second Language French: A Machine-Learning Approach* Jonas Granfeldt and Pierre Nugues	*Achieving Goals in Collaboration: Analysis of Estonian Institutional Calls* Olga Gerassimenko, Mare Koit, Andriela Rääbis and Krista Strandson

15.00–15.30	*Comprehension Assistant for Languages of Baltic States* Inguna Skadiņa, Andrejs Vasiļjevs, Daiga Deksne, Raivis Skadiņš and Linda Goldberga	*Inducing Baseform Models from a Swedish Vocabulary Pool* Eva Forsbom	

15.30–16.00 **Coffee Break**
16.00–17.30 **Closing Session**
16.00–17.00 Business Meeting of the Northern European Association for Language Technology (NEALT)
17.00–17.30 Closing