NODALIDA2005

Ling@JoY 1

Ling@JoY :
University of Joensuu electronic publications
in linguistics and language technology;
n:o 1
ISBN 952-458-771-8
ISSN 1796-1114

Proceedings of the 15th NODALIDA conference, Joensuu 2005

Edited by Stefan Werner

JOENSUUN YLIOPISTO
JOENSUU 20006

Stefan Werner (ed.) :
Proceedings of the 15th NODALIDA conference, Joensuu 2005

Ling@JoY :
University of Joensuu electronic publications
in linguistics and language technology;
n:o 1
April 2006 (corrected edition May 2006), University of Joensuu

ISBN 952-458-771-8
ISSN 1796-1114
`http://ling.joensuu.fi/lingjoy/`

Copyright © by the authors.

Typeset with LaTeX2e in Bera 10 pt.

Contents

Preface v

Marita Ailomaa, Vladimír Kadlec, Martin Rajman, Jean-Cédric Chappelier:
Robust stochastic parsing: Comparing and combining two approaches for processing extra-grammatical sentences 1

Ergin Altintas, Elif Karsligil, Vedat Coskun:
A new semantic similarity measure evaluated in word sense disambiguation 8

Lili Aunimo, Reeta Kuuskoski:
Reformulations of Finnish questions for question answering 12

Chris Biemann, Uwe Quasthoff:
Dictionary acquisition using parallel text and co-occurrence statistics 22

Heidi Brøseth:
The influence of written task descriptions in Wizard of Oz experiments 30

Hercules Dalianis:
Improving search engine retrieval using a compound splitter for Swedish 38

Kristin Hagen, Lars Nygaard:
Improbable morphological forms in a computational lexicon 43

Johan Hall, Joakim Nivre:
A generic architecture for data-driven dependency parsing 47

Lars Hellan, Dorothee Beermann:
The 'specifier' in an HPSG grammar implementation of Norwegian 57

Hans Hjelm, Christoph Schwarz:
LiSa - morphological analysis for information retrieval 65

Jon Holmlund, Magnus Sahlgren, Jussi Karlgren:
Creating bilingual lexica using reference wordlists for alignment of monolingual semantic vector spaces 71

Johan Hovold:
Naive Bayes spam filtering using word-position-based attributes and length-sensitive classification thresholds 78

Christer Johansson, Lars G. Johnsen:
An analytical relation between analogical modeling and memory based learning 88

Tuomo Kakkonen:
Dependency treebanks: methods, annotation schemes and tools 94

Viggo Kann, Magnus Rosell:
Free construction of a free Swedish dictionary of synonyms 105

Jussi Karlgren:
Compound terms and their constituent elements in information retrieval 111

Bodil Kyst, Peter Juel Henrichsen:
Synthetic regional Danish 116

Hanne Moa:
Compounds and other oddities in machine translation 124

Jyrki Niemi, Lauri Carlson:
Towards modeling the semantics of calendar expressions as extended regular expressions 133

Kristina Nilsson, Aisha Malmgren:
Towards automatic recognition of product names: an exploratory study of brand names in economic texts 139

Kristina Nilsson, Hans Hjelm, Henrik Oxhammar:
SUiS - cross-language ontology-driven information retrieval in a restricted domain 146

Anders Nøklestad, Christer Johansson:
Detecting reference chains in Norwegian 156

Henrik Oxhammar:
Mapping product descriptions to a large ontology 163

Magnus Rosell, Sumithra Velupillai:
The impact of phrases in document clustering for Swedish 173

Jonas Sjöbergh:
Chunking: an unsupervised method to find errors in text 180

Peter Rossen Skadhauge, Peter Juel Henrichsen:
DanPO - a transcription-based dictionary for Danish speech technology 186

Anders Søgaard, Petter Haugereid:
Functionality in grammar design 193

Heli Uibo:
Optimizing the finite-state description of Estonian morphology 203

Hans Friedrich Witschel, Chris Biemann:
Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization 210

Preface

The 15th Nordic Conference of Computational Linguistics—better known as NODALIDA—was held on May 20 and 21, 2005 at the University of Joensuu in Eastern Finland. Several related events also took place in Joensuu during the days leading up to NODALIDA, most notably the final seminar of the NorFA- (now NordForsk) funded Nordic language technology research program.

Over 120 participants registered for NODALIDA 2005 and every other also presented their work at the conference which offered 31 papers and 29 posters on many different aspects of language and speech technology, in addition to three invited lectures. For more information, please refer to the conference web site which will remain available at `http://ling.joensuu.fi/nodalida2005/`. This electronic proceedings volume contains 27 articles based on NODALIDA presentations.

I would like to express my sincere thanks to all conference participants as well as my special gratitude to those who made it possible at all to have NODALIDA at Joensuu University, in particular Kimmo Koskenniemi and Jussi Niemi. I would also like to take the opportunity to thank the members of our International Program Committee for their very valuable work and their kind assistance. Thanks of course also go to the session chairs and to all local helpers. Finally, let me acknowledge the financial support received from the Academy of Finland.

In the plenary session following Kimmo Koskenniemi's talk on *Past, present and future of language technology in the Nordic area* several interesting ideas for future Nordic and Baltic cooperation were discussed; a short summary can be found at the site of the Language Technology Documentation Centre in Finland (FiLT). One of the decisions made was to accept the University of Tartu's offer to host the 16th NODALIDA in the spring of 2007.

Joensuu, April 13, 2006
Stefan Werner