18th Nordic Conference of Computational Linguistics (NODALIDA 2011)

NEALT Proceedings Series Volume 11

Riga, Latvia
11 – 13 May 2011

Editors:

Bolette Sandford Pedersen
Inguna Skadina
Gunta Nespore

ISBN: 978-1-5108-3464-4

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2011) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2017)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Contents

Preface **viii**

Committees **x**

Conference Program **xii**

I Invited Papers **1**

When FrameNet meets a Controlled Natural Language
Guntis Bārzdiņš ***2***

Bare-Bones Dependency Parsing — A Case for Occam's Razor?
Joakim Nivre ***6***

Discourse Structures and Language Technologies
Bonnie Webber ***12***

II Regular papers **17**

Identification of sense selection in regular polysemy using shallow features
Hector Martinez Alonso, Núria Bel and Bolette Sandford Pedersen ***18***

Decision Strategies for Incremental POS Tagging
Niels Beuck, Arne Köhn and Wolfgang Menzel ***26***

A FrameNet for Danish
Eckhard Bick ***34***

Extraction from relative and embedded interrogative clauses in Danish
Anne Bjerre ***42***

The Formal Patterns of the Lithuanian Verb Forms
Loïc Boizou ***50***

Semantic search in literature as an e-Humanities research tool: CONPLISIT — Consumption patterns and life-style in 19th century Swedish literature
Lars Borin, Markus Forsberg and Christer Ahlberger ***58***

Evaluation of terminologies acquired from comparable corpora: an application perspective
Estelle Delpech ***66***

A quantitative and qualitative analysis of Nordic surnames
Eirini Florou and Stasinos Konstantopoulos ***74***

Experiments on Lithuanian Term Extraction
Gintarė Grigonytė, Erika Rimkutė, Andrius Utka and Loïc Boizou ***82***

Fishing in a speech stream, angling for a lexicon
Peter Juel Henrichsen ***90***

The Impact of Part-of-Speech Filtering on Generation of a Swedish-Japanese Dictionary Using English as Pivot Language
Ingemar Hjälmstad, Martin Hassel and Maria Skeppstedt ***98***

A Gold Standard for English–Swedish Word Alignment
Maria Holmqvist and Lars Ahrenberg ***106***

Relevance Prediction in Information Extraction using Discourse and Lexical Features
Silja Huttunen, Arto Vihavainen and Roman Yangarber ***114***

What kind of corpus is a web corpus?
Janne Bondi Johannessen and Emiliano Raúl Guevara ***122***

Morphological analysis of a non-standard language variety
Heiki-Jaan Kaalep and Kadri Muischnek ***130***

Editing Syntax Trees on the Surface
Peter Ljunglöf ***138***

Do wordnets also improve human performance on NLP tasks?
Kristiina Muhonen and Krister Lindén ***146***

Creating Comparable Multimodal Corpora for Nordic Languages
Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen and Patrizia Paggio ***153***

Estimating language relationships from a parallel corpus. A study of the Europarl corpus
Taraka Rama and Lars Borin ***161***

Improving Sentence-level Subjectivity Classification through Readability Measurement
Robert Remus ***168***

Iterative, MT-based Sentence Alignment of Parallel Texts
Rico Sennrich and Martin Volk ***175***

Combining Statistical Models for POS Tagging using Finite-State Calculus
Miikka Silfverberg and Krister Lindén ***183***

Toponym Disambiguation in English-Lithuanian SMT System with Spatial Knowledge
Raivis Skadiņš, Tatiana Gornostay and Valters Šics **191**

Automatic summarization as means of simplifying texts, an evaluation for Swedish
Christian Smith and Arne Jönsson **198**

Using graphical models for PP attachment
Anders Søgaard **206**

Corrective re-synthesis of deviant speech using unit selection
Sofia Strömbergsson **214**

Psycho-acoustically motivated formant feature extraction
Bea Valkenier, Dirkjan Krijnders, Ronald Van Elburg and Tjeerd Andringa **218**

Random Indexing Re-Hashed
Erik Velldal **224**

Evaluating the effect of word frequencies in a probabilistic generative model of morphology
Sami Virpioja, Oskar Kohonen and Krista Lagus **230**

Disambiguation of English Contractions for Machine Translation of TV Subtitles
Martin Volk and Rico Sennrich **238**

Probabilistic Models for Alignment of Etymological Data
Hannes Wettig and Roman Yangarber **246**

Convolution Kernels for Subjectivity Detection
Michael Wiegand and Dietrich Klakow **254**

Explorations on Positionwise Flag Diacritics in Finite-State Morphology
Anssi Yli-Jyrä **262**

III Regular short papers 270

Experiments to investigate the utility of nearest neighbour metrics based on linguistically informed features for detecting textual plagiarism
Per Almquist and Jussi Karlgren **271**

CFG based grammar checker for Latvian
Daiga Deksne and Raivis Skadiņš **275**

Query Constraining Aspects of Knowledge
Ann-Marie Eklund **279**

A categorization scheme for analyzing rules from a handbook of Swedish writing rules
Jody Foo **283**

Something Old, Something New — Applying a Pre-trained Parsing Model to Clinical Swedish
Martin Hassel, Aron Henriksson and Sumithra Velupillai ***287***

Knowledge-free Verb Detection through Sentence Sequence Alignment
Christian Hänig ***291***

"Andre ord" — a wordnet browser for the Danish wordnet, DanNet (DEMO)
Anders Johannsen and Bolette Sandford Pedersen ***295***

Modularisation of Finnish Finite-State Language Description — Towards Wide Collaboration in Open Source Development of a Morphological Analyser
Tommi Pirinen ***299***

A Prague Markup Language profile for the SemTi-Kamols grammar model
Lauma Pretkalniņa, Gunta Nešpore, Kristīne Levāne-Petrova and Baiba Saulīte ***303***

Dialect classification in the Himalayas: a computational approach
Anju Saxena and Lars Borin ***307***

Extraction of Knowledge-Rich Contexts in Russian – A Study in the Automotive Domain
Anne-Kathrin Schumann ***311***

Iterative reordering and word alignment for statistical MT
Sara Stymne ***315***

A double-blind experiment on interannotator agreement: the case of dependency syntax and Finnish
Atro Voutilainen and Tanja Purtonen ***319***

Automatic Question Generation from Swedish Documents as a Tool for Information Extraction
Kenneth Wilhelmsson ***323***

IV Student papers 327

Linguistic Motivation in Automatic Sentence Alignment of Parallel Corpora: the Case of Danish-Bulgarian and English-Bulgarian
Angel Genov and Georgi Iliev ***328***

Finding statistically motivated features influencing subtree alignment performance
Gideon Kotzé ***332***

Evaluating the speech quality of the Norwegian synthetic voice Brage
Marius Olaussen ***336***

A Statistical Part-of-Speech Tagger for Persian
Mojgan Seraji ***340***

Identification of context markers for Russian nouns
Anastasia Shimorina and Maria Grachkova *344*

Author Index **348**