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
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>viii</td>
</tr>
<tr>
<td>Committees</td>
<td>x</td>
</tr>
<tr>
<td>Conference Program</td>
<td>xii</td>
</tr>
<tr>
<td>I Invited Papers</td>
<td>1</td>
</tr>
<tr>
<td>When FrameNet meets a Controlled Natural Language</td>
<td>2</td>
</tr>
<tr>
<td>Guntis Bārzdīņš</td>
<td></td>
</tr>
<tr>
<td>Bare-Bones Dependency Parsing — A Case for Occam’s Razor?</td>
<td>6</td>
</tr>
<tr>
<td>Joakim Nivre</td>
<td></td>
</tr>
<tr>
<td>Discourse Structures and Language Technologies</td>
<td>12</td>
</tr>
<tr>
<td>Bonnie Webber</td>
<td></td>
</tr>
<tr>
<td>II Regular papers</td>
<td>17</td>
</tr>
<tr>
<td>Identification of sense selection in regular polysemy using shallow features</td>
<td>18</td>
</tr>
<tr>
<td>Hector Martinez Alonso, Núria Bel and Bolette Sandford Pedersen</td>
<td></td>
</tr>
<tr>
<td>Decision Strategies for Incremental POS Tagging</td>
<td>26</td>
</tr>
<tr>
<td>Niels Beuck, Arne Köhn and Wolfgang Menzel</td>
<td></td>
</tr>
<tr>
<td>A FrameNet for Danish</td>
<td>34</td>
</tr>
<tr>
<td>Eckhard Bick</td>
<td></td>
</tr>
<tr>
<td>Extraction from relative and embedded interrogative clauses in Danish</td>
<td>42</td>
</tr>
<tr>
<td>Anne Bjerre</td>
<td></td>
</tr>
<tr>
<td>The Formal Patterns of the Lithuanian Verb Forms</td>
<td>50</td>
</tr>
<tr>
<td>Loïc Boizou</td>
<td></td>
</tr>
<tr>
<td>Semantic search in literature as an e-Humanities research tool: CONPLISIT</td>
<td>58</td>
</tr>
<tr>
<td>— Consumption patterns and life-style in 19th century Swedish literature</td>
<td></td>
</tr>
<tr>
<td>Lars Borin, Markus Forsberg and Christer Ahlberger</td>
<td></td>
</tr>
</tbody>
</table>
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

_Ingemar Hjälmland, 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 Häning 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 Analys...er
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