ACL 2010

NLPLING 2010

2010 Workshop on NLP and Linguistics: Finding the Common Ground

Proceedings of the Workshop

16 July 2010
Uppsala University
Uppsala, Sweden

Production and Manufacturing by
Taberg Media Group AB
Box 94, 562 02 Taberg
Sweden

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-932432-75-6 / 1-932432-75-2

Preface

Since early 1990s, with the advancement of machine learning methods and the availability of data resources such as treebanks and parallel corpora, data-driven approaches to NLP have made significant progress. The success of such data-driven approaches has cast doubt on the relevance of linguistics to NLP. Conversely, NLP techniques are rarely used to help linguistics studies. We believe that there is room to expand the involvement of linguistics in NLP, and likewise, NLP in linguistics, and that the cross-pollination of ideas between the disciplines can greatly benefit both fields. We are pleased to present the workshop on *NLP and Linguistics: Finding the Common Ground* in order to focus on some of the work that uses NLP and linguistics for mutual benefit, and discuss future plans for continuing collaborations.

The workshop is intended to spur discussion on how NLP and linguistics can help each other, including new methods in incorporating linguistic knowledge into statistical systems to advance the state of the art of NLP, and the feasibility of using NLP techniques to acquire linguistic knowledge for a large number of languages and to assist linguistic studies. Fifteen papers were submitted and nine were accepted (one later withdrew), and the accepted papers are oriented around the following themes:

- **Research that shows awareness of a particular linguistic phenomenon and its effects on statistical systems:** Caines and Buttery discuss the zero auxiliary construction (*You talking to me?*), awareness of which can improve performance of NLP on spoken English. Samaradžić and Merlo suggest that awareness of different types of light verb constructions could affect word alignment. Su, Huang, and Chen show that the linguistic notion of evidentiality can be used for automatic detection of trustworthiness.

- **New methods in incorporating linguistic knowledge into statistical systems to improve the start of the art:** The papers by Caines and Buttery, Cook and Stevenson, Samaradžić and Merlo, and Su, Huang, and Chen all present a number of linguistic features that can be used for modeling or other corpus-based tasks.

- **Research that demonstrates the feasibility of creating NLP systems to automatically acquire linguistic knowledge for a large number of languages:** Mayer, Rohrdantz, Plank, Bak, Butt, and Keim examine a phonotactic constraint in 3,200 languages. Poornima and Good propose the repurposing of traditional word lists from historical and comparative linguistics to NLP applications.

- **Research that demonstrates the benefits of using NLP techniques to help particular linguistic studies:** This volume is rich with examples of corpus-based techniques shedding light on linguistic phenomena, including the ambiguity of German past participles (Zarrieß, Cahill, Kuhn, and Rohrer), zero auxiliary constructions (Caines and Buttery), light verbs (Samaradžić and Merlo), a paradoxical reading of "no X is too Y to Z" (Cook and Stevenson), the phonotactic constraint of Similar Place Avoidance (Mayer, Rohrdantz, Plank, Bak, Butt, and Keim), and evidentiality (Su, Huang, and Chen).

- **The realtive strengths and weaknesses of corpus-based and rule-based resources:** Plank and van Noord examine the domain portability of rule-based and corpus-trained parsers. Zarrieß, Cahill, Kunh, and Rohrer show that a corpus-based analysis can help reduce ambiguity of German past participles in a rule-based parser.

In addition to the presenters of papers, the workshop includes two panels to discuss the potential contributions of NLP to linguistics and linguistics to NLP. The panelists in the Linguistics-helps-NLP panel have been asked to address the following questions, and the questions for the NLP-helps-Linguistics panel are similar. Three panelists have written a short paper to summarize their positions, and these papers have been included in the proceedings.

1. What kinds of NLP applications could benefit from linguistics? For a particular NLP application, what is the best way of incorporating linguistic knowledge into NLP systems to improve the start of the art. (e.g., as rules in a preprocessing step, as linguistic features in a statistical system, as filters for pruning a search space, as priors in an objective function)?

2. What is the right role for a linguist in developing NLP resources (e.g., recommending features, writing rules, or building resources such as treebanks)?

3. What are the obstacles to using linguistics in NLP and how can they be removed? What do you wish you had available to you but don't?

4. How can we, as a field, encourage more collaborations between NLP researchers and linguists? Are there examples of successful collaborations, and if so, how were these facilitated?

5. What do NLP and linguistic students need to know to engage in these collaborations? How can we get students involved in collaborative research between the two disciplines?

We would like to thank everyone who made this workshop possible: ACL, the program committee, our invited speaker, the panelists, the authors, and workshop participants. Special thanks go to the US National Science Foundation for its support (NSF IIS-1027289).

Fei Xia, William Lewis, and Lori Levin

Organizers:

Fei Xia, University of Washington, USA
William Lewis, Microsoft Research, USA
Lori Levin, Carnegie Mellon University, USA

Program Committee:

Anthony Aristar, LinguistList, USA
Jason Baldridge, University of Texas at Austin, USA
Timothy Baldwin, University of Melbourne, Australia
Dorothee Beermann, NTNU, Norway
Emily M. Bender, University of Washington, USA
Steven Bird, University of Melbourne, Australia
Chris Brew, Ohio State University, USA
Michael Collins, MIT, USA
Michael Cysouw, Max Planck Institute for Evolutionary Anthropology, Germany
Hal Daume III, University of Utah, USA
Markus Dickinson, University of Indiana, USA
Alexis Dimitriadis, Utrecht Institute of Linguistics OTS, The Netherlands
Helen Aristar Dry, LinguistList, USA
Jason Eisner, Johns Hopkins Univ, USA
Erhard Hinrichs, University of Tubingen, Germany
Chu-Ren Huang, The Hong Kong Polytechnic University, Hong Kong, China
Julia Hockenmaier, UIUC, USA
Mark Johnson, Macquarie University, Australia
Kevin Knight, USC/ISI, USA
Mark Liberman, University of Pennsylvania, USA
Dekang Lin, Google, USA
Paola Merlo, University of Geneva, Switzerland
Kathy McKeown, Columbia Univ, USA
Martha Palmer, University of Colorado, USA
Dragomir Radev, University of Michigan, USA
Owen Rambow, Columbia University, USA
Dipti Misra Sharma, IIIT-H, India
Richard Sproat, Oregon Health & Science University, USA
Mark Steedman, Edinburgh, UK
Michael White, Ohio State University, USA
Richard Wicentowski, Swarthmore College, USA
Peter Wittenburg, Max Planck Institute for Psycholinguistics, The Netherlands
Andreas Witt, Institut für Deutsche Sprache, Mannheim, Germany
Nianwen Xue, Brandeis University, USA

Invited Speaker:

Steven Bird, University of Melbourne, Australia

Panelists:

Hal Daume III, University of Utah, USA
Alexis Dimitriadis, Utrecht Institute of Linguistics OTS, The Netherlands
Erhard Hinrichs, University of Tubingen, Germany
Dipti Misra Sharma, IIIT, India

Julia Hockenmaier, UIUC, USA
Eduard Hovy, USC/ISI, USA
Owen Rambow, Columbia University, USA

Sponsor:

US National Science Foundation (Grant IIS-1027289)

Table of Contents

Workshop Program

Friday, July 16, 2010

8:45–8:50 Opening Remarks

8:50–9:50 Invited Talk by Steven Bird: "The Human Language Project: Uniting computational linguistics with documentary linguistics"

Paper Session 1

9:50–10:10 *Modeling and Encoding Traditional Wordlists for Machine Applications*
Shakthi Poornima and Jeff Good

10:10–10:30 *Evidentiality for Text Trustworthiness Detection*
Su Qi, Huang Chu-Ren and Chen Kai-yun

10:30–11:00 Morning break

Panel Session 1: NLP helps Linguistics

11:00–12:00 Presentation and discussion from panelists (Hal Daume, Alexis Dimitriadis, Erhard Hinrichs, and Dipti Misra Sharma)

On the Role of NLP in Linguistics
Dipti Misra Sharma

Matching Needs and Resources: How NLP Can Help Theoretical Linguistics
Alexis Dimitriadis

Friday, July 16, 2010 (continued)

Paper Session 2

12:00–12:20 *Grammar-Driven versus Data-Driven: Which Parsing System Is More Affected by Domain Shifts?*
Barbara Plank and Gertjan van Noord

12:20–12:40 *A Cross-Lingual Induction Technique for German Adverbial Participles*
Sina Zarrieß, Aoife Cahill, Jonas Kuhn and Christian Rohrer

12:40–14:10 Lunch

Paper Session 3

14:10–14:30 *You Talking to Me? A Predictive Model for Zero Auxiliary Constructions*
Andrew Caines and Paula Buttery

14:30–14:50 *Cross-Lingual Variation of Light Verb Constructions: Using Parallel Corpora and Automatic Alignment for Linguistic Research*
Tanja Samardžić and Paola Merlo

14:50–15:10 *No Sentence Is Too Confusing To Ignore*
Paul Cook and Suzanne Stevenson

15:10–15:30 *Consonant Co-Occurrence in Stems across Languages: Automatic Analysis and Visualization of a Phonotactic Constraint*
Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt and Daniel A. Keim

15:30–16:00 Afternoon break

Friday, July 16, 2010 (continued)

Panel Session 2: Linguistics helps NLP

16:00–17:00 Presentation and discussion from panelists (Julia Hockenmeier, Eduard Hovy, and Owen Rambow)

Injecting Linguistics into NLP through Annotation
Eduard Hovy

17:00–17:30 Group discussion and closing