LAW VI

The 6th Linguistic Annotation Workshop in conjunction with ACL-2012

Proceedings of the Workshop

July 12 - 13, 2012
Jeju, Republic of Korea

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 978-1-937284-32-9

Introduction

The Linguistic Annotation Workshop (The LAW) is organized annually by the Association for Computational Linguistics Special Interest Group for Annotation (ACL SIGANN). It provides a forum to facilitate the exchange and propagation of research results concerned with the annotation, manipulation, and exploitation of corpora; work towards the harmonization and interoperability from the perspective of the increasingly large number of tools and frameworks for annotated language resources; and work towards a consensus on all issues crucial to the advancement of the field of corpus annotation. These proceedings include papers presented at LAW VI, held in Jeju, Korea, on 12-13 July 2012.

This year's call for papers was answered by over 40 submissions. After careful review, the Program Committee accepted 14 long papers, together with nine additional papers to be presented as posters. This year's submissions addressed many topics of interest for resource annotation, with a particularly strong representation of papers describing annotation schemes devised to handle phenomena at a wide range of linguistic levels, from particles in Korean to social actions in discourse. Another topic that received considerable attention concerned strategies to evaluate and improve the reliability of annotations, especially those that are manually produced as well as annotations obtained via crowdsourcing. Annotated written and spoken resources in a variety of languages, including Korean, Urdu, Hindi, and Indonesian, were also represented.

The LAW VI call for papers included a new and special component: a call for submissions to answer *The LAW Challenge*, sponsored by the U.S. National Science Foundation (IIS 0948101 Content of Linguistic Annotation: Standards and Practices (CLASP)) and the ACL Special Interest Group on Annotation (ACL SIGANN). The challenge was established this year to promote the use and collaborative development of open, shared resources, and to identify and promote best practices for annotation interoperability. The evaluation criteria included the following:

- innovative use of linguistic information from different annotation layers;
- demonstrable interoperability with at least one other annotation scheme or format developed by others;
- quality of the annotated resource in terms of scheme design, documentation, tool support, etc.;
- open availability of developed resources for community use;
- usability and reusability of the annotation scheme or annotated resource;
- outstanding contribution to the development of annotation best practices.

The winner of the first LAW Challenge was *Who Did What to Whom? A Contrastive Study of Syntacto-Semantic Dependencies*, which examines interoperability among a broad range of common annotation schemes for syntacto-semantic dependencies within the LinGO Redwoods Treebank project. The strengths of the project were seen to be its design and focus on interoperability, as well as its potential to promote work on interoperability that will help the community to develop larger, richer representations to train various linguistic tools. The winning paper received a monetary award to cover the authors' travel expenses and workshop registration. The selection process for the winner of the first LAW Challenge was extremely difficult, and therefore, the committee decided to acknowledge a strong runner-up, entitled *Prague Markup Language Framework*, which was recognized for the extensive influence of the described scheme on the community and the extent to which the scheme and tools have been applied to other languages.

We would like to thank SIGANN for its continuing organization of the LAW workshops, as well as the support of the ACL 2012 workshop committee chairs, Massimo Poesio and Satoshi Sekine. Most of all, we would like to thank all the authors for submitting their papers to the workshop and our program committee members and reviewers for their dedication and informative reviews.

Nancy Ide and Fei Xia, Program Committee Co-chairs

Program Committee Chairs:

Nancy Ide (Vassar College)
Fei Xia (University of Washington)

Program Committee:

Collin Baker (ICSI/UC Berkeley)
Emily Bender (University of Washington)
Nicoletta Calzolari (ILC/CNR)
Steve Cassidy (Macquarie University)
Christopher Cieri (LDC/University of Pennsylvania)
Stefanie Dipper (Ruhr-Universitaet Bochum)
Tomaz Erjavec (Josef Stefan Institute)
Alex Chengyu Fang (City University of Hong Kong)
Christiane Fellbaum (Princeton University)
Dan Flickinger (Stanford University)
Udo Hahn (Friedrich Schiller Universität Jena)
Chu-Ren Huang (Hong Kong Polytechnic)
Aravind Joshi (University of Pennsylvania)
Adam Meyers (New York University)
Antonio Pareja Lora (UCM / ATLAS-UNED)
Martha Palmer (University of Colorado)
Massimo Poesio (University of Trento)
Christopher Potts (Stanford University)
Sameer Pradhan (BBN Technologies)
James Pustejovsky (Brandeis University)
Owen Rambow (Columbia University)
Manfred Stede (Universitat Potsdam)
Mihai Surdeanu (Yahoo! Research, Barcelona)
Katrin Tomanek (Universitaet Dordrecht)
Theresa Wilson (University of Edinburgh)
Andreas Witt (IDS Mannheim)
Nianwen Xue (Brandeis University)

Table of Contents

The Role of Linguistic Models and Language Annotation in Feature Selection for Machine Learning
James Pustejovsky 1

Who Did What to Whom? A Contrastive Study of Syntacto-Semantic Dependencies
Angelina Ivanova, Stephan Oepen, Lilja Øvrelid and Dan Flickinger 2

Prague Markup Language Framework
Jirka Hana and Jan Štěpánek 12

Exploiting naive vs expert discourse annotations: an experiment using lexical cohesion to predict Elaboration / Entity-Elaboration confusions
Clémentine Adam and Marianne Vergez-Couret 22

Pair Annotation: Adaption of Pair Programming to Corpus Annotation
Isin Demirsahin, Ihsan Yalcinkaya and Deniz Zeyrek 31

Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers
Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn and Pierre Zweigenbaum 40

Intra-Chunk Dependency Annotation : Expanding Hindi Inter-Chunk Annotated Treebank
Prudhvi Kosaraju, Bharat Ram Ambati, Samar Husain, Dipti Misra Sharma and Rajeev Sangal 49

A Model for Linguistic Resource Description
Nancy Ide and Keith Suderman 57

A GrAF-compliant Indonesian Speech Recognition Web Service on the Language Grid for Transcription Crowdsourcing
Bayu Distiawan and Ruli Manurung 67

Towards Adaptation of Linguistic Annotations to Scholarly Annotation Formalisms on the Semantic Web
Karin Verspoor and Kevin Livingston 75

Intonosyntactic Data Structures: The Rhapsodie Treebank of Spoken French
Kim Gerdes, Sylvain Kahane, Anne Lacheret, Paola Pietandrea and Arthur Truong 85

Annotation Schemes to Encode Domain Knowledge in Medical Narratives
Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff B. Pelz, Pengcheng Shi and Anne Haake 95

Usability Recommendations for Annotation Tools
Manuel Burghardt 104

Search Result Diversification Methods to Assist Lexicographers
Lars Borin, Markus Forsberg, Karin Friberg Heppin, Richard Johansson and Annika Kjellandsson 113

Simultaneous error detection at two levels of syntactic annotation
Adam Przepiórkowski and Michał Lenart 118

Exploring Temporal Vagueness with Mechanical Turk
Yuping Zhou and Nianwen Xue 124

Developing Learner Corpus Annotation for Korean Particle Errors
Sun-Hee Lee, Markus Dickinson and Ross Israel 129

Annotating Archaeological Texts: An Example of Domain-Specific Annotation in the Humanities
Francesca Bonin, Fabio Cavulli, Aronne Noriller, Massimo Poesio and Egon W. Stemle 134

Annotating Preferences in Chats for Strategic Games
Anais Cadilhac, Nicholas Asher and Farah Benamara 139

Morpheme Segmentation in the METU-Sabancı Turkish Treebank
Ruket Cakici 144

AlvisAE: a collaborative Web text annotation editor for knowledge acquisition
Frédéric Papazian, Robert Bossy and Claire Nédellec 149

CSAF - a community-sourcing annotation framework
Jin-Dong Kim and Yue Wang 153

Dependency Treebank of Urdu and its Evaluation
Riyaz Ahmad Bhat and Dr. Dipti Misra Sharma 157

Annotating Coordination in the Penn Treebank
Wolfgang Maier, Sandra Kübler, Erhard Hinrichs and Julia Kriwanek 166

Annotating Particle Realization and Ellipsis in Korean
Sun-Hee Lee and Jae-Young Song 175

Annotation of Adversarial and Collegial Social Actions in Discourse
David Bracewell, Marc Tomlinson, Mary Brunson, Jesse Plymale, Jiajun Bracewell and Daniel Boerger 184

Workshop Program

Thursday, July 12, 2012

8:45–9:00 Opening Remarks

Invited talk

9:00–9:35 *The Role of Linguistic Models and Language Annotation in Feature Selection for Machine Learning*
James Pustejovsky

Special Session: The LAW Challenge

9:35–9:40 Presentation of LAW Challenge Award

9:40–10:05 Challenge Winner: *Who Did What to Whom? A Contrastive Study of Syntacto-Semantic Dependencies*
Angelina Ivanova, Stephan Oepen, Lilja Øvrelid and Dan Flickinger

10:05–10:30 Special Recognition: *Prague Markup Language Framework*
Jirka Hana and Jan Štěpánek

10:30–11:00 Morning coffee break

Paper Session 1

11:00–11:25 *Exploiting naive vs expert discourse annotations: an experiment using lexical cohesion to predict Elaboration / Entity-Elaboration confusions*
Clémentine Adam and Marianne Vergez-Couret

11:25–11:50 *Pair Annotation: Adaption of Pair Programming to Corpus Annotation*
Isin Demirsahin, Ihsan Yalcinkaya and Deniz Zeyrek

11:50–12:15 *Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers*
Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn and Pierre Zweigenbaum

12:15–12:40 *Intra-Chunk Dependency Annotation : Expanding Hindi Inter-Chunk Annotated Treebank*
Prudhvi Kosaraju, Bharat Ram Ambati, Samar Husain, Dipti Misra Sharma and Rajeev Sangal

12:40–14:15 Lunch

Thursday, July 12, 2012 (continued)

Paper Session 2

14:15–14:40 *A Model for Linguistic Resource Description*
Nancy Ide and Keith Suderman

14:40–15:05 *A GrAF-compliant Indonesian Speech Recognition Web Service on the Language Grid for Transcription Crowdsourcing*
Bayu Distiawan and Ruli Manurung

15:05–15:30 *Towards Adaptation of Linguistic Annotations to Scholarly Annotation Formalisms on the Semantic Web*
Karin Verspoor and Kevin Livingston

15:30–16:00 Afternoon coffee break

Paper Session 3

16:00–16:25 *Intonosyntactic Data Structures: The Rhapsodie Treebank of Spoken French*
Kim Gerdes, Sylvain Kahane, Anne Lacheret, Paola Pietandrea and Arthur Truong

16:25–16:50 *Annotation Schemes to Encode Domain Knowledge in Medical Narratives*
Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff B. Pelz, Pengcheng Shi and Anne Haake

16:50–17:15 *Usability Recommendations for Annotation Tools*
Manuel Burghardt

17:15–17:30 SIGANN business meeting

Friday, July 13, 2012

Poster Session (9:00-10:05am)

Search Result Diversification Methods to Assist Lexicographers
Lars Borin, Markus Forsberg, Karin Friberg Heppin, Richard Johansson and Annika Kjellandsson

Simultaneous error detection at two levels of syntactic annotation
Adam Przepiórkowski and Michał Lenart

Exploring Temporal Vagueness with Mechanical Turk
Yuping Zhou and Nianwen Xue

Developing Learner Corpus Annotation for Korean Particle Errors
Sun-Hee Lee, Markus Dickinson and Ross Israel

Annotating Archaeological Texts: An Example of Domain-Specific Annotation in the Humanities
Francesca Bonin, Fabio Cavulli, Aronne Noriller, Massimo Poesio and Egon W. Stemle

Annotating Preferences in Chats for Strategic Games
Anais Cadilhac, Nicholas Asher and Farah Benamara

Morpheme Segmentation in the METU-Sabancı Turkish Treebank
Ruket Cakici

AlvisAE: a collaborative Web text annotation editor for knowledge acquisition
Frédéric Papazian, Robert Bossy and Claire Nédellec

CSAF - a community-sourcing annotation framework
Jin-Dong Kim and Yue Wang

Friday, July 13, 2012 (continued)

Paper Session 4

10:05–10:30 *Dependency Treebank of Urdu and its Evaluation*
Riyaz Ahmad Bhat and Dr. Dipti Misra Sharma

10:30-11:30: Morning coffee break

Paper Session 5

11:00–11:25 *Annotating Coordination in the Penn Treebank*
Wolfgang Maier, Sandra Kübler, Erhard Hinrichs and Julia Kriwanek

11:25–11:50 *Annotating Particle Realization and Ellipsis in Korean*
Sun-Hee Lee and Jae-Young Song

11:50–12:15 *Annotation of Adversarial and Collegial Social Actions in Discourse*
David Bracewell, Marc Tomlinson, Mary Brunson, Jesse Plymale, Jiajun Bracewell and Daniel Boerger

12:15–12:30 Closing