LAW X

The 10th Linguistic Annotation Workshop held in conjunction with ACL 2016

Workshop Proceedings

August 11, 2016
Berlin, Germany

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 978-1-945626-05-0
Proceedings of the 10th Linguistic Annotation Workshop (LAW-X)
Annemarie Friedrich and Katrin Tomanek (eds.)

Introduction to the Workshop

The Linguistic Annotation Workshop (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 harmonisation 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. The series is now in its tenth year, with these proceedings including papers that were presented at LAW X, held in conjunction with the annual meeting of the Association for Computational Linguistics (ACL) in Berlin, Germany, on August 11, 2016.

In 2016, the LAW celebrates its 10th anniversary – the first workshop took place in 2007 at the ACL in Prague. Since then, the LAW has been held every year, consistently drawing substantial participation (both in terms of paper/poster submissions and participation in the actual workshop) providing evidence that the LAW's overall focus continues to be an important area of interest in the field.

This year's LAW has received 50 submissions, out of which 19 long papers and 2 short papers have been accepted to be presented at the workshop, 7 as talks and 14 as posters. In addition to oral paper presentations, LAW X also features an invited talk by Marie-Catherine de Marneffe and a special theme session both dedicated to this year's special theme "Evaluation of Annotation Quality". The special theme session includes a short tutorial on the advantages of using item-response models by Dirk Hovy as well as a general discussion.

Our thanks go to SIGANN, our organizing committee, for its continuing organization of the LAW workshops, and to the ACL 2016 workshop chairs for their support. Also, we thank the ACL 2016 publication chairs for their help with these proceedings. Most of all, we would like to thank all the authors for submitting their papers to the workshop, and our program committee members for their dedication and their thoughtful reviews. We also thank our sponsor, the Cluster of Excellence "Multimodal Computing and Interaction" (MMCI) at Saarland University.

Special Theme: Evaluation of Annotation Quality

This special theme considers current practice in evaluation of linguistic annotations and its successes and failures by asking questions such as: How are we as a community measuring inter-annotator agreement to date, and are there sounder ways to measure it? How can we estimate the annotation quality of existing resources, and what can be done to document annotated data to help others assess its reliability?

1. How agreement is measured in various (new or existing) annotation projects, and what the different scores tell us in each case.
2. Good acceptance thresholds for different annotation tasks and metrics, and/or how to determine them.
3. Previously proposed but not widely used measures for agreement or annotation quality.
4. Proposals for quantitative or qualitative methods to measure agreement or annotation quality.
5. Proposals for documentation of published resources to support their evaluation, means and methods to achieve community evaluation of linguistically-annotated resources, etc.

Annemarie Friedrich and Katrin Tomanek

Invited Talk: Marie-Catherine de Marneffe

Assessing the Consistency and Use of "Common Sense" and Dependency Annotations

In this talk, I will discuss my work on two types of annotations: "common sense" annotations obtained through crowdsourcing techniques as well as specific linguistic annotations by experts.

First, I will talk about "common sense" annotations gathered on Mechanical Turk. I focus on two datasets, the Internet Argument Corpus, which contains annotation of agreement in online debate (Walker et al., 2012), and the PragBank corpus, which provides veridicality annotations – whether events described in a text are viewed as actual, non-actual or uncertain (de Marneffe et al., 2012). I will review the quality of the annotations of these corpora and how the corpora have been used in research. I will suggest that since judgments of agreement and veridicality are not always categorical, they should be modeled as distributions, in line with Passonneau and Carpenter (2014).

Second, I will turn to annotations of specific linguistic representations, mainly dependency annotations where experts are annotating grammatical relations between words of a sentence, and investigate how we can assess the consistency of these annotations within a corpus. I will present preliminary results of our assessment of how much consistency is found in some of the Universal Dependency corpora using the Boyd et al. (2008)'s technique for identifying errors in dependency annotations.

References:

Adriane Boyd, Markus Dickinson and Detmar Meurers. 2008. *On detecting errors in dependency treebanks.* In Research on Language and Computation 6(2): 113–137.

Marie-Catherine de Marneffe, Christopher D. Manning and Christopher Potts. 2012. *Did it happen? The pragmatic complexity of veridicality assessment.* In Computational Linguistics 38(2): 301-333.

Rebecca J. Passonneau and Bob Carpenter. 2014. *The benefits of a model of annotation.* In Transactions of the Association for Computational Linguistics 2: 311-326.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. *A corpus for research on deliberation and debate.* In Proceedings of the 8th Language Resources and Evaluation Conference: 812–817.

Bio. Marie-Catherine de Marneffe is an assistant professor in Linguistics at The Ohio State University. She received her PhD from Stanford University in December 2012 under the supervision of Christopher D. Manning. She is developing computational linguistic methods that capture what is conveyed by speakers beyond the literal meaning of the words they say. Primarily she wants to ground meanings in corpus data, and show how such meanings can drive pragmatic inference. She has also worked on Recognizing Textual Entailment and contributed to defining the Stanford Dependencies and the Universal Dependencies representations, which are practical representations of grammatical relations and predicate argument structure. She serves as a member of the NAACL board and the Computational Linguistics editorial board.

Invited Tutorial: Dirk Hovy

How Item-Response Models Can Help us Take the Headache out of Annotation Projects

In annotation projects, we are usually interested in three questions (to varying degrees):

1. how do I aggregate my scores to get the "correct" answer?
2. how much can I trust the annotators?
3. how difficult is the task/individual items?

The traditional approach to answer these has been through inter-annotator agreement (IAA) scores, such as Cohen's Kappa, which can give us weights for each annotator, or simply by raw agreement and majority voting. However, there have been known problems with both Kappa (overestimating chance agreement when one label is prevalent, Feinstein and Cicchetti, 1990) and majority voting (unreliable annotators can swamp the result) that negatively affect questions 1 and 2 (see also Artstein and Poesio, 2008). In addition, neither of these measures tell us how difficult the task is. IAAs are thus only a proxy for the answers we really want.

Recently, Passonneau and Carpenter (2014) have suggested probabilistic item-response models (IRT) as an alternative. These models have several advantages, since thet can directly answer the above questions via

- annotator scores
- distributions over labels
- entropy scores for the task and individual items.

Despite this promise, IRTs are not yet in wide use, possibly because they can seem complex, unintuitive, and complicated to use. In this hands-on tutorial, I want to therefore introduce an available IRT (MACE: Hovy et al., 2013) and show in examples how we can easily get the answers we want from the data, plus a host of other information. The code is freely available, it is easy to use, and it can help us answer all the relevant questions for an annotation task.

References:

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. *Learning Whom to Trust with MACE.* In Proceedings of NAACL HLT.

Rebecca J Passonneau and Bob Carpenter. 2014. *The benefits of a model of annotation.* In Transactions of the Association for Computational Linguistics.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. *High agreement but low kappa: I. the problems of two paradoxes.* In Journal of Clinical Epidemiology.

Ron Artstein and Massimo Poesio. 2008. *Inter-coder agreement for computational linguistics.* In Computational Linguistics.

Bio. Dirk Hovy is an associate professor in natural language processing at the University of Copenhagen. His research focuses on the interaction of statistical models, language, and demographic factors. He received his PhD in Computer Science from the University of Southern California, and holds an MA in sociolinguistics from the University of Marburg, Germany. Dirk has authored papers on a variety of NLP topics, including semantic and syntactic analysis, domain adaptation, and information extraction. All of these involved annotation at some point, and the associate problems have led to the development of MACE. Outside of research, Dirk enjoys cooking, tango, and leather-crafting, as well as picking up heavy things and putting them back down. You can find an updated biography and more at `http://dirkhovy.com/`.

LAW Co-chairs

Annemarie Friedrich, Saarland University
Katrin Tomanek, OpenTable

Organizing Committee:

Stefanie Dipper, Ruhr University Bochum
Chu-Ren Huang, The Hong Kong Polytechnic University
Nancy Ide, Vassar College
Lori Levin, Carnegie Mellon University
Adam Meyers, New York University
Antonio Pareja-Lora, SIC & ILSA, UCM / ATLAS, UNED
Massimo Poesio, University of Trento
Sameer Pradhan, Cemantix.org and Boulder Learning, Inc.
Ines Rehbein, Leibniz ScienceCampus
Manfred Stede, University of Potsdam
Fei Xia, University of Washington
Nianwen Xue, Brandeis University
Heike Zinsmeister, University of Hamburg

Program Committee:

Adam Meyers, New York University
Alexis Palmer, Heidelberg University
Andreas Witt, Institut für Deutsche Sprache
Ani Nenkova, University of Pennsylvania
Ann Bies, Linguistic Data Consortium
Anna Nedoluzhko, Charles University Prague
Antonio Pareja-Lora, Universidad Complutense de Madrid
Aravind Joshi, University of Pennsylvania
Archna Bhatia, Florida Institute for Human and Machine Cognition
Barbara Plank, University of Groningen
Bonnie Webber, University of Edinburgh
Caroline Sporleder, University of Göttingen
Christian Chiarcos Goethe University Frankfurt
Christiane Fellbaum, Princeton University
Chu-Ren Huang, The Hong Kong Polytechnic University
Collin Baker, University of California, Berkeley
Dirk Hovy, University of Copenhagen
Djamé Seddah, University Paris-Sorbonne
Els Lefever, Ghent University
Fei Xia, University of Washington
Heike Zinsmeister, Hamburg University
Ines Rehbein, Heidelberg University
Joel Tetreault, Yahoo!
James Pustejovsky, Brandeis University
Josef Ruppenhofer, Heidelberg University
Kim Gerdes, University Paris-Sorbonne

Lori Levin, Carnegie Mellon University
Manfred Pinkal, Saarland University
Manfred Stede, University of Potsdam
Markus Dickinson, Indiana University
Martha Palmer, University of Colorado Boulder
Massimo Poesio, University of Essex
Nancy Ide, Vassar College
Nathan Schneider, University of Edinburgh
Nianwen Xue, Brandeis University
Nicoletta Calzolari, Italian National Research Council
Omri Abend, University of Jerusalem
Özlem Çetinoğlu, University of Stuttgart
Sameer Pradhan, Cemantix.org and Boulder Learning, Inc.
Sandra Kübler, Indiana University, Bloomington
Stefanie Dipper, Ruhr University Bochum
Tomaž Erjavec, Jožef Stefan Institute, Ljubljana
Udo Hahn, University of Jena
Valia Kordoni, Humboldt University of Berlin

Invited Speakers:

Marie-Catherine de Marneffe, The Ohio State University
Dirk Hovy, University of Copenhagen

Table of Contents

Building a Cross-document Event-Event Relation Corpus
Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji and Martha Palmer 1

Annotating the Little Prince with Chinese AMRs
Bin Li, Yuan Wen, Weiguang QU, Lijun Bu and Nianwen Xue 7

Converting SynTagRus Dependency Treebank into Penn Treebank Style
Alex Luu, Sophia A. Malamud and Nianwen Xue 16

A Discourse-Annotated Corpus of Conjoined VPs
Bonnie Webber, Rashmi Prasad, Alan Lee and Aravind Joshi 22

Annotating Spelling Errors in German Texts Produced by Primary School Children
Ronja Laarmann-Quante, Lukas Knichel, Stefanie Dipper and Carina Betken 32

Supersense tagging with inter-annotator disagreement
Héctor Martínez Alonso, Anders Johannsen and Barbara Plank 43

Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task
Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis and Bonnie Webber 49

Comparison of Annotating Methods for Named Entity Corpora
Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki and Hiroyuki Shinnou 59

Different Flavors of GUM: Evaluating Genre and Sentence Type Effects on Multilayer Corpus Annotation Quality
Amir Zeldes and Dan Simonson 68

Addressing Annotation Complexity: The Case of Annotating Ideological Perspective in Egyptian Social Media
Heba Elfardy and Mona Diab 79

Evaluating Inter-Annotator Agreement on Historical Spelling Normalization
Marcel Bollmann, Stefanie Dipper and Florian Petran 89

A Corpus of Preposition Supersenses
Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O'Gorman and Martha Palmer 99

Focus Annotation of Task-based Data: Establishing the Quality of Crowd Annotation
Kordula De Kuthy, Ramon Ziai and Detmar Meurers 110

Part of Speech Annotation of a Turkish-German Code-Switching Corpus
Özlem Çetinoğlu and Çağrı Çöltekin 120

Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies
Kim Gerdes and Sylvain Kahane 131

Conversion from Paninian Karakas to Universal Dependencies for Hindi Dependency Treebank
Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat and Dipti Sharma 141

Phrase Generalization: a Corpus Study in Multi-Document Abstracts and Original News Alignments
Ariani Di-Felippo and Ani Nenkova 151

Generating Disambiguating Paraphrases for Structurally Ambiguous Sentences
Manjuan Duan, Ethan Hill and Michael White 160

Applying Universal Dependency to the Arapaho Language
Irina Wagner, Andrew Cowell and Jena D. Hwang 171

Annotating the discourse and dialogue structure of SMS message conversations
Nianwen Xue, Qishen Su and Sooyoung Jeong 180

Creating a Novel Geolocation Corpus from Historical Texts
Grant DeLozier, Ben Wing, Jason Baldridge and Scott Nesbit 188

Workshop Program

9:00 – 10:30	**Session 1: Opening and Invited Talk**
9:00 – 9:10	Opening Remarks
9:10 – 10:05	Invited talk: *Assessing the Consistency and Use of "Common Sense" and Dependency Annotations.* Marie-Catherine de Marneffe
10:05 – 10:30	*Generating Disambiguating Paraphrases for Structurally Ambiguous Sentences* Manjuan Duan, Ethan Hill and Michael White
10:30 – 11:00	**Coffee Break**
11:00 – 12:40	**Session 2: Dependency Annotation and Discourse**
11:00 – 11:25	*Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies* Kim Gerdes and Sylvain Kahane
11:25 – 11:50	*Conversion from Paninian Karakas to Universal Dependencies for Hindi Dependency Treebank* Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat and Dipti Sharma
11:50 – 12:15	*Different Flavors of GUM: Evaluating Genre and Sentence Type Effects on Multilayer Corpus Annotation Quality* Amir Zeldes and Dan Simonson
12:15 – 12:40	*Annotating the Discourse and Dialogue Structure of SMS Message Conversations* Nianwen Xue, Qishen Su and Sooyoung Jeong
12:40 – 14:00	**Lunch Break**
14:00 – 14:50	**Session 3: Evaluation of Agreement (Special Theme)**
14:00 - 14:25	*Evaluating Inter-Annotator Agreement on Historical Spelling Normalization* Marcel Bollmann, Stefanie Dipper and Florian Petran
14:25 – 14:50	*Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task* Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis and Bonnie Webber
14:50 – 16:00	**Session 4: Poster Presentations**
14:50 – 15:05	Poster boasters
15:05 – 16:00	Poster presentation and coffee
16:00 – 17:30	**Session 5: Invited Tutorial and Discussion (Special Theme)**
16:00 – 16:30	Invited tutorial: *How Item-Response Models Can Help us Take the Headache out of Annotation Projects.* Dirk Hovy
16:30 – 17:15	Discussion
17:15 – 17:30	Closing remarks: 10 years of LAW