NAACL HLT 2009

Active Learning for Natural Language Processing (ALNLP-09)

Proceedings of the Workshop

June 5, 2009
Boulder, Colorado

Production and Manufacturing by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

Endorsed by the following ACL Special Interest Groups:

- SIGNLL, Special Interest Group for Natural Language Learning
- SIGANN, Special Interest Group for Annotation

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 978-1-932432-40-4

Introduction

Welcome to the workshop on Active Learning for Natural Language Processing!

We started organizing this workshop in mid-2008 after strong encouragement in response to some of our own work in the area. As we gathered members of the program committee, the timeliness of the topic resonated with several of them: the growing body of knowledge on active learning and on active learning for NLP in particular makes this topic one worth exploring in a focused workshop rather than in isolated papers in occasional, far-flung conferences.

Labeled data is a prerequisite for many popular algorithms in natural language processing and machine learning. While it is possible to obtain large amounts of annotated data for well-studied languages in well-studied domains and well-studied problems, labeled data are rarely available for less common languages, domains, or problems. Unfortunately, obtaining human annotations for linguistic data is labor-intensive and typically the costliest part of the acquisition of an annotated corpus. It has been shown before that active learning can be employed to reduce annotation costs but not at the expense of quality. While diverse work over the past decade has demonstrated the possible advantages of active learning for corpus annotation and NLP applications, active learning is not widely used in many ongoing data annotation tasks. Much of the machine learning literature on the topic has focused on active learning for classification problems with less attention devoted to the kinds of problems encountered in NLP. Related topics such as distributed "human computation", cost-sensitive machine learning, and semi-supervised learning of all kinds are also growing in number as we search for the best ways to overcome the data acquisition bottleneck.

We were interested in bringing together researchers to explore the challenges and opportunities of active learning for NLP tasks, language acquisition, and language learning, and we have been rewarded with excellent submissions and a promising program. The workshop received sixteen submissions, eight of which are included in the final program. Two of the accepted papers are short papers which address ongoing work and pertinent issues. We hope that this gathering and these proceedings begin to shed more light on active learning for NLP classification tasks, sequence labeling, parsing, semantics, and other more complex tasks. The papers in the program also begin to address issues involving the application of active learning in real annotation projects.

We are especially grateful to the diverse and helpful program committee, whose reviews were careful and thoughtful. We are also grateful to all of the researchers who submitted their work for consideration. For the record, more information about the workshop is available online at http://nlp.cs.byu.edu/alnlp/.

Best regards,

Eric Ringger, Robbie Haertel, and Katrin Tomanek

Organizers:

Eric Ringger, Brigham Young University (USA)
Robbie Hertel, Brigham Young University (USA)
Katrin Tomanek, University of Jena (Germany)

Program Committee:

Shlomo Argamon, Illinois Institute of Technology (USA)
Jason Baldridge, University of Texas at Austin (USA)
Markus Becker, SPSS (UK)
Ken Church, Microsoft Research (USA)
Hal Daume, University of Utah (USA)
Robbie Haertel, Brigham Young University (USA)
Ben Hachey, University of Edinburgh (UK)
Udo Hahn, University of Jena (Germany)
Eric Horvitz, Microsoft Research (USA)
Rebecca Hwa, University of Pittsburgh (USA)
Ashish Kapoor, Microsoft Research (USA)
Mark Liberman, University of Pennsylvania/LDC (USA)
Prem Melville, IBM T.J. Watson Research Center (USA)
Ray Mooney, University of Texas at Austin (USA)
Miles Osborne, University of Edinburgh (UK)
Eric Ringger, Brigham Young University (USA)
Kevin Seppi, Brigham Young University (USA)
Burr Settles, University of Wisconsin (USA)
Victor Sheng, New York University (USA)
Katrin Tomanek, University of Jena (Germany)
Jingbo Zhu, Northeastern University (China)

Invited Speakers:

Burr Settles, University of Wisconsin (USA)
Robbie Haertel, Brigham Young University (USA)

Table of Contents

Conference Program

Friday, June 5, 2009

8:30–9:00 Opening Remarks

9:00–10:00 Invited Talk by Burr Settles

Session 1: Anaphora Resolution

10:00–10:30 *Active Learning for Anaphora Resolution*
Caroline Gasperin

10:30–11:00 Break

Session 2: Multiple Annotators and Cost Considerations

11:00–11:30 *On Proper Unit Selection in Active Learning: Co-Selection Effects for Named Entity Recognition*
Katrin Tomanek, Florian Laws, Udo Hahn and Hinrich Schütze

11:30–12:00 *Estimating Annotation Cost for Active Learning in a Multi-Annotator Environment*
Shilpa Arora, Eric Nyberg and Carolyn P. Rosé

12:00–12:30 *Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria*
Pei-Yun Hsueh, Prem Melville and Vikas Sindhwani

12:30–2:00 Lunch

Friday, June 5, 2009 (continued)

Session 3: Real Annotators and Experts

2:00–2:30 *Evaluating Automation Strategies in Language Documentation*
Alexis Palmer, Taesun Moon and Jason Baldridge

2:30–3:00 *A Web Survey on the Use of Active Learning to Support Annotation of Text Data*
Katrin Tomanek and Fredrik Olsson

3:00–3:30 Invited Talk by Robbie Haertel

3:30–4:00 Break

Session 4: New Methods

4:00–4:30 *Active Dual Supervision: Reducing the Cost of Annotating Examples and Features*
Prem Melville and Vikas Sindhwani

4:30–5:00 *Proactive Learning for Building Machine Translation Systems for Minority Languages*
Vamshi Ambati and Jaime Carbonell

5:00–5:30 Discussion