NAACL HLT 2010

Workshop on Active Learning for Natural Language Processing (ALNLP-10)

Proceedings of the Workshop

June 6, 2010
Los Angeles, California

USB memory sticks produced by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

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

Introduction

Labeled training data is often required to achieve state-of-the-art performance in machine learning solutions to natural language tasks. While traditional supervised learning relies on existing labeled data, *active learning* algorithms may select unlabeled data for labeling with the goal of reducing annotation costs, while maintaining high accuracy. Thus, active learning has emerged as a promising framework for NLP applications and annotation projects where unlabeled data is readily available (e.g., web pages, audio recordings, minority language data), but obtaining labels is cost-prohibitive.

The 2010 Workshop on Active Learning for Natural Language Processing (ALNLP) is the sequel to a successful 2009 meeting of the same name—both co-located with the Annual Conference of the North American Chapter of the Association for Computational Linguistics—with the intent of exploring key practical and theoretical aspects of active learning applied to real-world NLP problems. As we assembled the program committee, the topic's timeliness resonated with researchers both in machine learning (who see natural language as an important application domain for active learning), and with those in language technologies (who see maturing active learning methods as an important part of their toolbox). We feel that the topic is one worth exploring in a focused workshop such as this, at the intersection of these research areas, rather than in occasional isolated papers in various international conference venues. Our aim is to foster innovation and discussion that advances our understanding of active learning for NLP.

Our invited speaker, Jaime Carbonell, has a long history of applying active learning to machine translation, rare class discovery, ranking, and realistic labeling scenarios with multiple imperfect annotators. Given the proliferation of machine learning in language applications, and the growing popularity of online "crowd-sourcing" annotation environments, we expect that these topics with play an important role in the future of active learning for NLP. We are grateful to Dr. Carbonell for agreeing to speak about his new work in these areas.

The workshop received ten submissions, five of which were accepted as oral presentations and are included in the final program. These papers represent the fruit of international researchers exploring a variety of challenges and opportunities in active learning for NLP tasks. These include issues pertinent to human-computer interaction, domain adaptation, and remaining robust to rare class labels and other application-specific issues. We hope that this gathering and these proceedings shed more light on active learning for NLP tasks and the real annotation projects that are required to support them.

We are especially grateful to the program committee, whose reviews were thoughtful and constructive. We also thank all of the researchers who submitted their work for consideration. More information about the workshop is archived online at http://active-learning.net/alnlp2010.

Burr Settles, Kevin Small, and Katrin Tomanek
Workshop Organizers

Organizers:

Burr Settles, Carnegie Mellon University (USA)
Kevin Small, Tufts University (USA)
Katrin Tomanek, University of Jena (Germany)

Program Committee:

Markus Becker, SPSS (an IBM company) (UK)
Claire Cardie, Cornell University (USA)
Hal Daume III, University of Utah (USA)
Ben Hachey, Macquarie University (Australia)
Robbie Haertel, Brigham Young University (USA)
Udo Hahn, University of Jena (Germany)
Eric Horvitz, Microsoft Research (USA)
Rebecca Hwa, University of Pittsburgh (USA)
Ashish Kapoor, Microsoft Research (USA)
Prem Melville, IBM T.J. Watson Research Center (USA)
Ray Mooney, University of Texas at Austin (USA)
Fredrik Olsson, SICS (Sweden)
Foster Provost, New York University (USA)
Eric Ringger, Brigham Young University (USA)
Dan Roth, University of Illinois at Urbana-Champaign (USA)
Burr Settles, Carnegie Mellon University (USA)
Kevin Small, Tufts University (USA)
Katrin Tomanek, University of Jena (Germany)

Additional Reviewers:

Piyush Rai, University of Utah (USA)
Avishek Saha, University of Utah (USA)
Byron Wallace, Tufts University (USA)

Invited Speaker:

Jaime Carbonell, Carnegie Mellon University (USA)

Table of Contents

Workshop Program

Sunday, June 6, 2010

1:00-1:15 Introduction by Burr Settles and Kevin Small

Invited Talk

1:15-2:10 *Active and Proactive Machine Learning: From Fundamentals to Applications in Language Technologies and Beyond* by Jaime Carbonell

Research Papers I

2:10–2:35 *Using Variance as a Stopping Criterion for Active Learning of Frame Assignment*
Masood Ghayoomi

2:35–3:00 *Active Semi-Supervised Learning for Improving Word Alignment*
Vamshi Ambati, Stephan Vogel and Jaime Carbonell

3:00-3:30 **Break**

Research Papers II

3:30–3:55 *D-Confidence: An Active Learning Strategy which Efficiently Identifies Small Classes*
Nuno Escudeiro and Alipio Jorge

3:55–4:20 *Domain Adaptation meets Active Learning*
Piyush Rai, Avishek Saha, Hal Daume and Suresh Venkatasubramanian

4:20–4:55 *Parallel Active Learning: Eliminating Wait Time with Minimal Staleness*
Robbie Haertel, Paul Felt, Eric K. Ringger and Kevin Seppi

4:55-5:30 **Discussion**