NAACL HLT 2009

Semi-supervised Learning
for
Natural Language
Processing

Proceedings of the Workshop

June 4, 2009
Boulder, Colorado

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

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 978-1-932432-38-1

Introduction

Welcome to the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing!

Will semi-supervised learning (SSL) become the next de-facto standard for building natural language processing (NLP) systems, just as supervised learning has transformed the field in the last decade? Or will it remain as a nice idea that doesn't always work in practice? Semi-supervised learning has become an important topic due to the promise that high-quality labeled data and abundant unlabeled data, if leveraged appropriately, can achieve superior performance at lower cost. As researchers in semi-supervised learning reach critical mass, we believe it is time to take a step back and think broadly about whether we can discover general insights from the various techniques developed for different NLP tasks.

The goal of this workshop is to help build a community of SSL-NLP researchers and foster discussions about insights, speculations, and results (both positive and negative) that may otherwise not appear in a technical paper at a major conference. In our call-for-paper, we posed some open questions:

1. Problem Structure: What are the different classes of NLP problem structures (e.g. sequences, trees, N-best lists) and what algorithms are best suited for each class? For instance, can graph-based algorithms be successfully applied to sequence-to-sequence problems like machine translation, or are self-training and feature-based methods the only reasonable choices for these problems?

2. Background Knowledge: What kinds of NLP-specific background knowledge can we exploit to aid semi-supervised learning? Recent learning paradigms such as constraint-driven learning and prototype learning take advantage of our domain knowledge about particular NLP tasks; they represent a move away from purely data-agnostic methods and are good examples of how linguistic intuition can drive algorithm development.

3. Scalability: NLP data-sets are often large. What are the scalability challenges and solutions for applying existing semi-supervised learning algorithms to NLP data?

4. Evaluation and Negative Results: What can we learn from negative results? Can we make an educated guess as to when semi-supervised learning might outperform supervised or unsupervised learning based on what we know about the NLP problem?

5. To Use or Not To Use: Should semi-supervised learning only be employed in low-resource languages/tasks (i.e. little labeled data, much unlabeled data), or should we expect gains even in high-resource scenarios (i.e. expecting semi-supervised learning to improve on a supervised system that is already more than 95% accurate)?

We received 17 submissions and selected 10 papers after a rigorous review process. These papers cover a variety of tasks, ranging from information extraction to speech recognition. Some introduce new techniques, while others compared existing methods under a variety of situations. We are pleased to present these papers in this volume.

Our workshop will be kicked off with a keynote talk by Jason Eisner (Johns Hopkins University). We will end with a panel discussion on the future of SSL-NLP, which will feature invited position papers from several prominent researchers. (Some are included in this volume; others will be online at the workshop website: http://sites.google.com/site/sslnlp/).

We are especially grateful to the program committee for their hard work and the presenters for their excellent papers. We would also like to thank the following people for their many help and support: Hal Daume, Sajib Dasgupta, Jason Eisner, Nizar Habash, Mark Hasegawa-Johnson, Andrew McCallum, Vincent Ng, Anoop Sarkar, Eric Ringger, and Jerry Zhu.

Best regards,

Qin Iris Wang, Kevin Duh, Dekang Lin
SSL-NLP Workshop Organizers

26 April 2009

Organizers:

Qin Iris Wang, AT&T
Kevin Duh, University of Washington
Dekang Lin, Google Research

Program Committee:

Steven Abney (University of Michigan, USA)
Yasemin Altun (Max Planck Institute for Biological Cybernetics, Germany)
Tim Baldwin (University of Melbourne, Australia)
Shane Bergsma (University of Alberta, Canada)
Antal van den Bosch (Tilburg University, The Netherlands)
John Blitzer (UC Berkeley, USA)
Ming-Wei Chang (UIUC, USA)
Walter Daelemans (University of Antwerp, Belgium)
Hal Daume III (University of Utah, USA)
Kevin Gimpel (Carnegie Mellon University, USA)
Andrew Goldberg (University of Wisconsin, USA)
Liang Huang (Google Research, USA)
Rie Johnson [formerly, Ando] (RJ Research Consulting)
Katrin Kirchhoff (University of Washington, USA)
Percy Liang (UC Berkeley, USA)
Gary Geunbae Lee (POSTECH, Korea)
Gina-Anne Levow (University of Chicago, USA)
Gideon Mann (Google, USA)
David McClotsky (Brown University, USA)
Ray Mooney (UT Austin, USA)
Hwee Tou Ng (National University of Singapore, Singapore)
Vincent Ng (UT Dallas, USA)
Miles Osborne (University of Edinburgh, UK)
Mari Ostendorf (University of Washington, USA)
Chris Pinchak (University of Alberta, Canada)
Dragomir Radev (University of Michigan, USA)
Dan Roth (UIUC, USA)
Anoop Sarkar (Simon Fraser University, Canada)
Dale Schuurmans (University of Alberta, Canada)
Akira Shimazu (JAIST, Japan)
Jun Suzuki (NTT, Japan)
Yee Whye Teh (University College London, UK)
Kristina Toutanova (Microsoft Research, USA)
Jason Weston (NEC, USA)
Tong Zhang (Rutgers University, USA)

Ming Zhou (Microsoft Research Asia, China)
Xiaojin (Jerry) Zhu (University of Wisconsin, USA)

Invited Speaker:

Jason Eisner, Johns Hopkins University

Table of Contents

Conference Program

Thursday, June 4, 2009

8:30–9:00 Coffee Service

9:00–9:10 Opening Remarks

9:10–10:10 Invited Talk by Jason Eisner

10:10–10:30 *Coupling Semi-Supervised Learning of Categories and Relations*
Andrew Carlson, Justin Betteridge, Estevam Rafael Hruschka Junior and Tom M. Mitchell

10:30–11:00 Morning Break

11:00–11:20 *Surrogate Learning - From Feature Independence to Semi-Supervised Classification*
Sriharsha Veeramachaneni and Ravi Kumar Kondadadi

11:25–11:45 *Keepin' It Real: Semi-Supervised Learning with Realistic Tuning*
Andrew B. Goldberg and Xiaojin Zhu

11:50–12:10 *Is Unlabeled Data Suitable for Multiclass SVM-based Web Page Classification?*
Arkaitz Zubiaga, Víctor Fresno and Raquel Martínez

12:15–12:35 *A Comparison of Structural Correspondence Learning and Self-training for Discriminative Parse Selection*
Barbara Plank

12:35–2:00 Lunch Break

2:00–2:20 *Latent Dirichlet Allocation with Topic-in-Set Knowledge*
David Andrzejewski and Xiaojin Zhu

2:25–2:45 *An Analysis of Bootstrapping for the Recognition of Temporal Expressions*
Jordi Poveda, Mihai Surdeanu and Jordi Turmo

2:50–3:10 *A Simple Semi-supervised Algorithm For Named Entity Recognition*
Wenhui Liao and Sriharsha Veeramachaneni

Thursday, June 4, 2009 (continued)

3:15–3:35 *Can One Language Bootstrap the Other: A Case Study on Event Extraction*
Zheng Chen and Heng Ji

3:35–4:00 Afternoon Break

4:00–4:20 *On Semi-Supervised Learning of Gaussian Mixture Models for Phonetic Classification*
Jui-Ting Huang and Mark Hasegawa-Johnson

4:25–5:25 Panel Discusstion

Discriminative Models for Semi-Supervised Natural Language Learning
Sajib Dasgupta and Vincent Ng

5:25–5:40 Workshop Wrap-up