ROBUS 2011

Proceedings of
Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing
(at RANLP 2011)

15 September, 2011
Hissar, Bulgaria

INTERNATIONAL WORKSHOP
ROBUST UNSUPERVISED AND SEMI-SUPERVISED METHODS
IN NATURAL LANGUAGE PROCESSING

PROCEEDINGS

Hissar, Bulgaria
15 September 2011

ISBN 978-954-452-017-5

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Introduction

In natural language processing (NLP), supervised learning scenarios are more frequently explored than unsupervised or semi-supervised ones. Unfortunately, labeled data are often highly domain-dependent and short in supply. It has therefore become increasingly important to leverage both labeled and unlabeled data to achieve the best performance in challenging NLP problems that involve learning of structured variables.

Until recently most results in semi-supervised learning of structured variables in NLP were negative, but today the best part-of-speech taggers, named entity recognizers, and dependency parsers exploit mixtures of labeled and unlabeled data. Unsupervised and minimally unsupervised NLP also sees rapid growth.

The most commonly used semi-supervised learning algorithms in NLP are feature-based methods and EM, self- or co-training. Mixture models have also been successfully used. While feature-based methods seem relatively robust, self-training and co-training are very parameter-sensitive, and parameter tuning has therefore become an important research topic. This is not only a concern in NLP, but also in other areas such as face recognition. Parameter-sensitivity is even more dramatic in unsupervised learning of structured variables, e.g. unsupervised part-of-speech tagging and grammar induction.

The aim of this workshop was to bring together researchers dedicated to designing and evaluating robust unsupervised or semi-supervised learning algorithms for NLP problems. We received 11 papers, but accepted only six. Shane Bergsma gave an invited talk on feature-based methods.

The organizers would like to thank the review committee for their thorough high-quality reviews and their timeliness, and the RANLP 2011 organizers for their assistance.

Organizers:

Chris Biemann, TU Darmstadt
Anders Søgaard, University of Copenhagen

Program Committee:

Steven Abney, University of Michigan
Stefan Bordag, ExB Research & Development
Eugenie Giesbrecht, FZI Karlsruhe
Katja Filippova, Google
Florian Holz, University of Leipzig
Jonas Kuhn, University of Stuttgart
Vivi Nastase, HITS Heidelberg
Reinhard Rapp, JG University of Mainz
Lucia Specia, University of Wolverhampton
Valentin Spitkovsky, Stanford University
Sven Teresniak, University of Leipzig
Dekai Wu, HKUST
Torsten Zesch, TU Darmstadt
Jerry Zhu, University of Wisconsin-Madison

Invited Speaker:

Shane Bergsma, Johns Hopkins University

Table of Contents

Workshop Program

Thursday, 15 September, 2011

Chair: Chris Biemann

10:00–10:30 *Gibbs Sampling with Treeness Constraint in Unsupervised Dependency Parsing*
David Mareček and Zdeněk Žabokrtský

10:30–11:00 Coffee Break

11:00–11:30 *Guided Self Training for Sentiment Classification*
Brett Drury, Luis Torgo and Jose Joao Almeida

11:30–12:00 *Investigating the Applicability of current Machine-Learning based Subjectivity Detection Algorithms on German Texts*
Malik Atalla, Christian Scheel, Ernesto William De Luca and Sahin Albayrak

12:00–14:15 Lunch

14:15–15:15 Invited talk: Simple, Effective, Robust Semi-Supervised Learning, Thanks To Google N-grams, Shane Bergsma

15:15–16:00 Coffee Break

16:00–16:30 *Learning Protein Protein Interaction Extraction using Distant Supervision*
Philippe Thomas, Illés Solt, Roman Klinger and Ulf Leser

16:30–17:00 *Topic Models with Logical Constraints on Words*
Hayato Kobayashi, Hiromi Wakaki, Tomohiro Yamasaki and Masaru Suzuki

17:00–17:30 *Investigation of Co-training Views and Variations for Semantic Role Labeling*
Rasoul Samad Zadeh Kaljahi and Mohd Sapiyan Baba

17:30–18:00 Closing Remarks