MM 2012

First Workshop on Multilingual Modeling

Proceedings of the Workshop

July 13, 2012
Jeju, Republic of Korea

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 978-1-937284-35-0

Introduction

The burgeoning community of multilingual users poses variety of new problems and also enables new opportunities. The large number of multilingual corpora requires effective and scalable ways for organizing them. This additional data in different languages provides a different perspective. Resource poor languages can utilize the training data available in other languages and improve the accuracies of monolingual applications.

Recently, we have seen an increasing number of researchers working on multilingual problems varying from mining comparable corpora from the web to multilingual part-of-speech tagging. It is encouraging to see how the abundant training data in a resource rich languages (such as English) is used along with very little training data in the target language to solve problems in resource-poor languages. In addition, resource rich languages have been used successfully to bridge the language barrier between two resource poor languages. This workshop is aimed to bring researchers working on different aspects of multilingualism to a common ground to share their experiences so that the entire community can benefit.

We received a total of 13 submissions. After a rigorous review process we selected 4 papers for presentation at the workshop. We would like to thank the members of the Program Committee for their excellent work — the reviews were all very thorough, carefully written, and detailed, and helped the authors to improve their papers.

This workshop features a mix of equal number of Invited Talks (IT), Invited Papers (IP) and Contribution Talks (CT). We are experimenting with this format to improve the quality of the discussions among the participants. We spent a considerable amount of time in selecting the IPs. These are by invitation only and are not be included in the workshop proceedings.

Organizers:

Jagadeesh Jagarlamudi (University of Maryland, USA)
Sujith Ravi (Google, USA)
Xiaojun Wan (Peking University, China)
Hal Daumé III (University of Maryland, USA)

Program Committee:

Kumaran A (Microsoft Research, India)
Pushpak Bhattacharyya (Indian Institute of Technology, India)
Srinivas Bangalore (AT&T Labs-Research, USA)
Hal Daumé III (University of Maryland, USA)
Kareen Darwish (Qatar Computing Research Institute, Qatar)
Dipanjan Das (Carnegie Mellon University, USA)
Marcello Federico (FBK – Fondazione Bruno Kessler, Tirento, Italy)
Anna Feldman (Montclair State University, USA)
Wei Gao (Qatar Computing Research Institute, Qatar)
Jagadeesh Jagarlamudi (University of Maryland, USA)
Heng Ji (City University of New York)
Mitesh Khapra (Indian Institute of Technology, India)
Alexandre Klementiev (Saarland University, USA)
Kevin Knight (USC/ISI, USA)
Yang Liu (Tsinghua University, China)
Paul McNamee (Johns Hopkins University, USA)
Rada Mihalcea (University of North Texas, USA)
Xiaochuan Ni (Microsoft)
Doug Oard (University of Maryland, USA)
Reinhard Rapp (Johannes Gutenberg-Universität Mainz, Germany)
Ari Rappoport (The Hebrew University, Israel)
Sujith Ravi (Google, USA)
Benjamin Snyder (University of Wisconsin-Madison, USA)
Benno Stein (Bauhaus-Universität Weimar, Germany)
Sebastian Stüker (Karlsruhe Institute of Technology, Germany)
Jun'ichi Tsujii (Microsoft Research Asia)
Kentaro Torisawa (NICT, Japan)
Raghavendra Udupa (Microsoft Research, India)
Xiaojun Wan (Peking University, China)
Mausam (University of Washington, USA)

Invited Speakers:

Slav Petrov, Google
Reinhard Rapp, University of Leeds
Benjamin Snyder, University of Wisconsin-Madison

Table of Contents

Workshop Program

Friday, July 13, 2012

9:00 Invited Talk by Reinhard Rapp
Bilingual Lexicon Extraction Using Parallel and Comparable Corpora

9:40 *Implementing a Language-Independent MT Methodology*
Sokratis Sofianopoulos, Marina Vassiliou and George Tambouratzis

10:05 Bilingual Lexicon Extraction from Comparable Corpora Using Label Propagation
Akihiro Tamura, Taro Watanabe and Eiichiro Sumita (Invited Paper)

10:30 Coffee break

11:00 Invited Talk by Benjamin Snyder
Multilingual Modeling: Current Work and Future Frontiers

11:40 *The Study of Effect of Length in Morphological Segmentation of Agglutinative Languages*
Loganathan Ramasamy, Zdeněk Žabokrtský and Sowmya Vajjala

12:05 Unsupervised Structure Prediction with Non-Parallel Multilingual Guidance
Shay B. Cohen, Dipanjan Das and Noah A. Smith (Invited Paper)

12:30 Lunch break

2:00 Invited Talk by Slav Petrov
Multilingual Syntactic Analysis

2:40 *Language Independent Named Entity Identification using Wikipedia*
Mahathi Bhagavatula, Santosh GSK and Vasudeva Varma

3:05 Cross-Lingual Parse Disambiguation based on Semantic Correspondence
Lea Frermann and Francis Bond (Invited Paper)

3:30 Coffee break

4:00 Learning Discriminative Projections for Text Similarity Measures
Wen-tau Yih, Kristina Toutanova, John Platt, and Chris Meek (Invited Paper)

4:25 Untangling the Cross-Lingual Link Structure of Wikipedia
Gerard de Melo and Gerhard Weikum (Invited Paper)

Friday, July 13, 2012 (continued)

4:50 *A Comparable Corpus Based on Aligned Multilingual Ontologies*
Roger Granada, Lucelene Lopes, Carlos Ramisch, Cassia Trojahn, Renata Vieira and Aline Villavicencio

5:15 Discussion