ACL 2010

NEWS 2010

2010 Named Entities Workshop

Proceedings of the Workshop

16 July 2010
Uppsala University
Uppsala, Sweden

Production and Manufacturing by
Taberg Media Group AB
Box 94, 562 02 Taberg
Sweden

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 978-1-932432-78-7 / 1-932432-78-7

Preface

Named Entities play a significant role in Natural Language Processing and Information Retrieval. While identifying and analyzing named entities in a given natural language is a challenging research problem by itself, the phenomenal growth in the Internet user population, especially among the non-English speaking parts of the world, has extended this problem to the crosslingual arena. We specifically focus on research on all aspects of the Named Entities in our workshop series, Named Entities WorkShop (NEWS). The first of the NEWS workshops (NEWS 2009) was held as a part of ACL-IJCNLP 2009 conference in Singapore, and the current edition (NEWS 2010) is being held as a part of ACL 2010, in Uppsala, Sweden.

The purpose of the NEWS workshop is to bring together researchers across the world interested in identification, analysis, extraction, mining and transformation of named entities in monolingual or multilingual natural language text. The workshop scope includes many interesting specific research areas pertaining to the named entities, such as, orthographic and phonetic characteristics, corpus analysis, unsupervised and supervised named entities extraction in monolingual or multilingual corpus, transliteration modelling, and evaluation methodologies, to name a few. For this years edition, 11 research papers were submitted, each of which was reviewed by at least 3 reviewers from the program committee. 7 papers were chosen for publication, covering main research areas, from named entities recognition, extraction and categorization, to distributional characteristics of named entities, and finally a novel evaluation metrics for co-reference resolution. All accepted research papers are published in the workshop proceedings.

This year, as parts of the NEWS workshop, we organized two shared tasks: one on Machine Transliteration Generation, and another on Machine Transliteration Mining, participated by research teams from around the world, including industry, government laboratories and academia.

The transliteration generation task was introduced in NEWS 2009. While the focus of the 2009 shared task was on establishing the quality metrics and on baselining the transliteration quality based on those metrics, the 2010 shared task expanded the scope of the transliteration generation task to about dozen languages, and explored the quality depending on the direction of transliteration, between the languages. We collected significantly large, hand-crafted parallel named entities corpora in dozen different languages from 8 language families, and made available as common dataset for the shared task. We published the details of the shared task and the training and development data six months ahead of the conference that attracted an overwhelming response from the research community. Totally 7 teams participated in the transliteration generation task. The approaches ranged from traditional unsupervised learning methods (such as, Phrasal SMT-based, Conditional Random Fields, etc.) to somewhat unique approaches (such as, DirectTL approach), combined with several model combinations for results re-ranking. A report of the shared task that summarizes all submissions and the original whitepaper are also included in the proceedings, and will be presented in the workshop. The participants in the shared task were asked to submit short system papers (4 pages each) describing their approach, and each of such papers was reviewed by at least two members of the program committee to help improve the quality of the content and presentation of the papers. 6 of them were finally accepted to be published in the workshop proceedings (one participating team did not submit their system paper in time).

NEWS 2010 also featured a second shared task this year, on Transliteration Mining; in this shared task we focus specifically on mining transliterations from the commonly available resource Wikipedia titles. The objective of this shared task is to identify transliterations from linked Wikipedia titles between English and another language in a non-Latin script. 5 teams participated in the mining task, each participating in multiple languages. The shared task was conducted in 5 language pairs, and the paired

Wikipedia titles between English and each of the languages was provided to the participants. The participating systems output was measured using three specific metrics. All the results are reported in the shared task report.

We hope that NEWS 2010 would provide an exciting and productive forum for researchers working in this research area. The technical programme includes 7 research papers and 9 system papers (3 as oral papers, and 6 as poster papers) to be presented in the workshop. Further, we are pleased to have Dr Dan Roth, Professor at University of Illinois and The Beckman Institute, delivering the keynote speech at the workshop.

We wish to thank all the researchers for their research submission and the enthusiastic participation in the transliteration shared tasks. We wish to express our gratitude to CJK Institute, Institute for Infocomm Research, Microsoft Research India, Cairo Microsoft Innovation Centre and Thailand National Electronics and Computer Technology Centre for preparing the data released as a part of the shared tasks. Finally, we thank all the programme committee members for reviewing the submissions in spite of the tight schedule.

Workshop Chairs

Haizhou Li, Institute for Infocomm Research, Singapore
A Kumaran, Microsoft Research, India

16 July 2010
Uppsala, Sweden

Organizers:

Haizhou Li, Institute for Infocomm Research (Singapore)
A Kumaran, Microsoft Research (India)

Workshop Organizing Committee:

Haizhou Li, Institute for Infocomm Research (Singapore)
A Kumaran, Microsoft Research (India)
Kevin Knight, ISI (USA)
Grzegorz Kondrak, University of Alberta (Canada)
Min Zhang, Institute for Infocomm Research (Singapore)

Shared Task Organizing Committee - Transliteration Generation:

Haizhou Li, Institute for Infocomm Research (Singapore)
A Kumaran, Microsoft Research (India)
Min Zhang, Institute for Infocomm Research (Singapore)
Vladimir Pervouchine, Institute for Infocomm Research (Singapore)

Shared Task Organizing Committee - Transliteration Mining:

A Kumaran, Microsoft Research (India)
Haizhou Li, Institute for Infocomm Research (Singapore)
Mitesh M. Khapra, Indian Institute of Technology Bombay (India)

Program Committee:

Kalika Bali, Microsoft Research (India)
Rafael Banchs, BarcelonaMedia (Spain)
Sivaji Bandyopadhyay, University of Jadavpur (India)
Pushpak Bhattacharyya, IIT-Bombay (India)
Monojit Choudhury, Microsoft Research (India)
Marta Ruiz Costa-jussa, UPC (Spain)
Gregory Grefenstette, Exalead (France)
Mitesh M. Khapra, Indian Institute of Technology Bombay, (India)
Sanjeev Khudanpur, Johns Hopkins University (USA)
Kevin Knight, ISI (USA)
Grzegorz Kondrak, University of Alberta (Canada)
A Kumaran, Microsoft Research (India)
Olivia Kwong, City University (Hong Kong)
Gina-Anne Levow, University of Manchester (UK)
Haizhou Li, Institute for Infocomm Research (Singapore)
Andrew McCallum, University of Massachusetts Amherst (USA)
Arul Menezes, Microsoft Research (USA)
Jong-Hoon Oh, NICT (Japan)
Vladimir Pervouchine, Institute for Infocomm Research (Singapore)
Yan Qu, Advertising.com (USA)
Satoshi Sekine, New York University (USA)
Sunita Sarawagi, IIT-Bombay (India)

Sudeshna Sarkar, IIT-Kharagpur (India)
Richard Sproat, University of Illinois at Urbana-Champaign (USA)
Keh-Yih Su, Behavior Design Corporation (Taiwan)
Raghavendra Udupa, Microsoft Research (India)
Vasudeva Varma, IIIT-Hyderabad (India)
Haifeng Wang, Baidu.com Inc. (China)
Shuly Wintner, University of Haifa (Israel)
Chai Wutiwiwatchai, NECTEC (Thailand)
Min Zhang, Institute for Infocomm Research (Singapore)

Table of Contents

Conference Program

Friday, 16 July 2010

Session 1: Oral

9:00–9:15 Opening Remarks
A. Kumaran and Haizhou Li

9:15–10:00 Keynote Speech
Dan Roth

10:00–10:30 *Transliteration Generation and Mining with Limited Training Resources*
Sittichai Jiampojamarn, Kenneth Dwyer, Shane Bergsma, Aditya Bhargava, Qing Dou, Mi-Young Kim and Grzegorz Kondrak

10:30–11:00 Morning Break

Session 2: Oral

11:00–11:20 *Transliteration Using a Phrase-Based Statistical Machine Translation System to Re-Score the Output of a Joint Multigram Model*
Andrew Finch and Eiichiro Sumita

11:20–11:40 *Transliteration Mining with Phonetic Conflation and Iterative Training*
Kareem Darwish

Friday, 16 July 2010 (continued)

Session 3: Poster

11:40–12:40 Poster Presentation

Language Independent Transliteration Mining System Using Finite State Automata Framework
Sara Noeman and Amgad Madkour

Reranking with Multiple Features for Better Transliteration
Yan Song, Chunyu Kit and Hai Zhao

Syllable-Based Thai-English Machine Transliteration
Chai Wutiwiwatchai and Ausdang Thangthai

English to Indian Languages Machine Transliteration System at NEWS 2010
Amitava Das, Tanik Saikh, Tapabrata Mondal, Asif Ekbal and Sivaji Bandyopadhyay

Mining Transliterations from Wikipedia Using Pair HMMs
Peter Nabende

Phrase-Based Transliteration with Simple Heuristics
Avinesh PVS and Ankur Parikh

12:40–14:00 Lunch Break

Session 4: Oral

14:00–14:20 *Classifying Wikipedia Articles into NE's Using SVM's with Threshold Adjustment*
Iman Saleh, Kareem Darwish and Aly Fahmy

14:20–14:40 *Assessing the Challenge of Fine-Grained Named Entity Recognition and Classification*
Asif Ekbal, Eva Sourjikova, Anette Frank and Simone Paolo Ponzetto

14:40–15:00 *Using Deep Belief Nets for Chinese Named Entity Categorization*
Yu Chen, You Ouyang, Wenjie Li, Dequan Zheng and Tiejun Zhao

15:00–15:20 *Simplified Feature Set for Arabic Named Entity Recognition*
Ahmed Abdul Hamid and Kareem Darwish

Friday, 16 July 2010 (continued)

15:20–16:00 Lunch Break

Session 5: Oral

16:00–16:20 *Think Globally, Apply Locally: Using Distributional Characteristics for Hindi Named Entity Identification*
Shalini Gupta and Pushpak Bhattacharyya

16:20–16:40 *Rule-Based Named Entity Recognition in Urdu*
Kashif Riaz

16:40–17:00 *CONE: Metrics for Automatic Evaluation of Named Entity Co-Reference Resolution*
Bo Lin, Rushin Shah, Robert Frederking and Anatole Gershman

17:00–17:10 Closing