ACL 2012

# Proceedings of
# NEWS 2012
# 2012 Named Entities Workshop

July 12, 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 978-1-937284-40-4

# Preface

The workshop series, Named Entities WorkShop (NEWS), focuses on research on all aspects of the Named Entities, such as, identifying and analyzing named entities, mining, translating and transliterating named entities, etc. The first of the NEWS workshops (NEWS 2009) was held as a part of ACL-IJCNLP 2009 conference in Singapore; the second one, NEWS 2010, was held as an ACL 2010 workshop in Uppsala, Sweden; and the third one, NEWS 2011, was held as an IJCNLP 2011 workshop in Chiang Mai, Thailand. The current edition, NEWS 2012, was held as an ACL 2012 workshop in Jeju, Korea.

The purpose of the NEWS workshop series 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 corpora. 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 modeling, and evaluation methodologies, to name a few. For this year edition, 7 research papers were submitted, each of which was reviewed by 3 reviewers from the program committee. 3 papers were chosen for publication, covering machine transliteration and transliteration mining from comparable corpus and wiki.

Following the tradition of the NEWS workshop series, NEWS 2012 continued the machine transliteration shared task this year as well. The shared task was first introduced in NEWS 2009 and continued in NEWS 2010 and NEWS 2011. In NEWS 2012, by leveraging on the previous success of NEWS workshop series, we released the hand-crafted parallel named entities corpora to include 14 different language pairs from 12 language families, and made them available as the common dataset for the shared task. In total, 7 international teams participated from around the globe. The approaches ranged from traditional learning methods (such as, Phrasal SMT-based, Conditional Random Fields, etc.) to somewhat new approaches (such as, RNN Language Model, Syllable-based Approach (Fine-grained English Segmentation), Two-Stage CRF, Optimization against multiple references and the intermediate representation of Chinese and Arabic). 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 content pages each) describing their approaches, and each of such papers was reviewed by three members of the program committee to help improve the quality. All the 7 system papers were finally accepted to be published in the workshop proceedings.

We hope that NEWS 2012 would provide an exciting and productive forum for researchers working in this research area, and the NEWS-released data continues to serve as a standard dataset for machine transliteration generation and mining. 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, Thailand National Electronics and Computer Technology Centre and The Royal Melbourne Institute of Technology (RMIT)/Sarvnaz Karimi for preparing the data released as a part of the shared tasks. Finally, we thank all the program committee members for reviewing the submissions in spite of the tight schedule.

Workshop Chairs:
Min Zhang, Institute for Infocomm Research, Singapore
Haizhou Li, Institute for Infocomm Research, Singapore
A Kumaran, Microsoft Research, India

12 July 2012
Jeju, Korea

**Organizers:**

Workshop Co-Chair: Min Zhang, Institute for Infocomm Research, Singapore
Workshop Co-Chair: Haizhou Li, Institute for Infocomm Research, Singapore
Workshop Co-Chair: A Kumaran, Microsoft Research, India

**Program Committee:**

Kalika Bali, Microsoft Research, India
Rafael Banchs, Institute for Infocomm Research, Singapore
Sivaji Bandyopadhyay, University of Jadavpur, India
Pushpak Bhattacharyya, IIT-Bombay, India
Monojit Choudhury, Microsoft Research, India
Marta Ruiz Costa-jussa, UPC, Spain
Xiangyu Duan, Institute for Infocomm Research, Singapore
Gregory Grefenstette, Exalead, France
Guohong Fu, Heilongjiang University, China
Sarvnaz Karimi, NICTA and the University of Melbourne, Australia
Mitesh Khapra, IIT-Bombay, India
Greg Kondrak, University of Alberta, Canada
Olivia Kwong, City University, Hong Kong
Ming Liu, Institute for Infocomm Research, Singapore
Jong-Hoon Oh, NICT, Japan
Yan Qu, Advertising.com, USA
Keh-Yih Su, Behavior Design Corporation, Taiwan
Jun Sun, NUS, Singapore
Raghavendra Udupa, Microsoft Research, India
Vasudeva Varma, IIIT-Hyderabad, India
Haifeng Wang, Baidu.com, China
Chai Wutiwiwatchai, NECTEC, Thailand
Deyi Xiong, Institute for Infocomm Research, Singapore
Muyun Yang, HIT, China
Chengqing Zong, Institute of Automation, CAS, China

# Table of Contents

# Conference Program

**Thursday, July 12, 2012**

9:00–9:10 Opening Remarks by Min Zhang, Haizhou Li, A Kumaran and Ming Liu

*Whitepaper of NEWS 2012 Shared Task on Machine Transliteration*
Min Zhang, Haizhou Li, A Kumaran and Ming Liu

*Report of NEWS 2012 Machine Transliteration Shared Task*
Min Zhang, Haizhou Li, A Kumaran and Ming Liu

**9:10–10:30 Session 1: Research Papers**

09:10–09:35 *Accurate Unsupervised Joint Named-Entity Extraction from Unaligned Parallel Text*
Robert Munro and Christopher D. Manning

09:35–10:00 *Latent Semantic Transliteration using Dirichlet Mixture*
Masato Hagiwara and Satoshi Sekine

10:00–10:25 *Automatically generated NE tagged corpora for English and Hungarian*
Eszter Simon and Dávid Márk Nemeskey

10:30–11:00 Morning Break

**11:00–12:30 Session 2: System Papers 1**

11:00–11:25 *Rescoring a Phrase-based Machine Transliteration System with Recurrent Neural Network Language Models*
Andrew Finch, Paul Dixon and Eiichiro Sumita

11:25–11:50 *Syllable-based Machine Transliteration with Extra Phrase Features*
Chunyue Zhang, Tingting Li and Tiejun Zhao

11:50–12:15 *English-Korean Named Entity Transliteration Using Substring Alignment and Re-ranking Methods*
Chun-Kai Wu, Yu-Chun Wang and Richard Tzong-Han Tsai

12:15–13:45 Lunch Break

**Thursday, July 12, 2012 (continued)**

**13:45–15:30 Session 3: System Papers 2**

13:45–14:10 *Applying mpaligner to Machine Transliteration with Japanese-Specific Heuristics*
Yoh Okuno

14:10–14:35 *Transliteration by Sequence Labeling with Lattice Encodings and Reranking*
Waleed Ammar, Chris Dyer and Noah Smith

14:35–15:00 *Transliteration Experiments on Chinese and Arabic*
Grzegorz Kondrak, Xingkai Li and Mohammad Salameh

15:00–15:25 *Cost-benefit Analysis of Two-Stage Conditional Random Fields based English-to-Chinese Machine Transliteration*
Chan-Hung Kuo, Shih-Hung Liu, Mike Tian-Jian Jiang, Cheng-Wei Lee and Wen-Lian Hsu

15:25–15:30 Closing

15:30–16:00 Afternoon Break