IJCNLP 2011

Proceedings of
NEWS 2011
2011 Named Entities Workshop

November 12, 2011
Shangri-La Hotel
Chiang Mai, Thailand

IJCNLP 2011

NEWS 2011
2011 Named Entities Workshop

November 12, 2011
Chiang Mai, Thailand

We wish to thank our sponsors

Gold Sponsors

www.google.com

www.baidu.com

The Office of Naval Research (ONR)

The Asian Office of Aerospace Research and Development (AOARD)

Department of Systems Engineering and Engineering Managment, The Chinese University of Hong Kong

Silver Sponsors

Microsoft Corporation

Bronze Sponsors

Chinese and Oriental Languages Information Processing Society (COLIPS)

Supporter

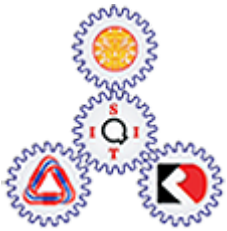
Thailand Convention and Exhibition Bureau (TCEB)

We wish to thank our sponsors

Organizers

Asian Federation of Natural Language Processing (AFNLP)

National Electronics and Computer Technology Center (NECTEC), Thailand



Sirindhorn International Institute of Technology (SIIT), Thailand

Rajamangala University of Technology Lanna (RMUTL), Thailand

Maejo University, Thailand

Chiang Mai University (CMU), Thailand

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, and the second one, NEWS 2010, was held as an ACL 2010 workshop in Uppsala, Sweden. The current edition, NEWS 2011, was held as an IJCNLP 2011 workshop, in Chiang Mai, Thailand.

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 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, 8 research papers were submitted, each of which was reviewed by at least 3 reviewers from the program committee. 5 papers were chosen for publication, covering main research areas, from named entities identification, classification, to machine transliteration and transliteration mining from comparable corpus and wiki. All accepted research papers are published in the workshop proceedings.

Following the tradition of the NEWS workshop series, NEWS 2011 continued the machine transliteration shared task this year as well. The shared task was first introduced in NEWS 2009 and continued in NEWS 2010. In NEWS 2011, by leveraging on the previous success of NEWS 2009 and NEWS 2011, we significantly increased the hand-crafted parallel named entities corpora to include 14 different language pairs from 11 language families, and made them available as the common dataset for the shared task. We published the details of the shared task and the training and development data several months ahead of the conference that attracted an overwhelming response from the research community. In total, 10 international teams participated from around the globe. The approaches ranged from traditional unsupervised learning methods (such as, Phrasal SMT-based, Conditional Random Fields, etc.) to somewhat new approaches (such as, Non-Parametric Bayesian Co-segmentation, Multi-to-Multi Joint Source Channel Model and Leveraging Transliterations from Multiple Languages), in addition to several teams resorting to model/system 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 content pages each) describing their approaches, and each of such papers was reviewed by at least three members of the program committee to help improve the quality. All the 10 system papers were finally accepted to be published in the workshop proceedings.

It is heartening for us to report that the previous year's NEWS datasets are being regularly requested by research groups throughout the year outside the NEWS shared tasks, for calibration of new approaches by groups that had not previously participated in the shared tasks. We expect such trend to continue, establishing the NEWS parallel names corpora as a standard dataset, and NEWS metrics as a standard measure for future machine transliteration research.

We hope that NEWS 2011 would provide an exciting and productive forum for researchers working in this research area. 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
Haizhou Li, Institute for Infocomm Research, Singapore
A Kumaran, Microsoft Research, India
Min Zhang, Institute for Infocomm Research, Singapore

12 November 2011,
Chiang Mai, Thailand

Organizers:

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

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
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
Sudeshna Sarkar, IIT-Kharagpur, India
Keh-Yih Su, Behavior Design Corporation, Taiwan
Raghavendra Udupa, Microsoft Research, India
Vasudeva Varma, IIIT-Hyderabad, India
Haifeng Wang, Baidu.com, China
Chai Wutiwiwatchai, NECTEC, Thailand
Chengqing Zong, Institute of Automation, CAS, China

Table of Contents

Conference Program

November 12, 2011

8:30-10:00 Session 1

8:30–8:40 Opening Remarks by Haizhou Li, A Kumaran, Min Zhang and Ming Liu

8:40–9:00 *Integrating Models Derived from non-Parametric Bayesian Co-segmentation into a Statistical Machine Transliteration System*
Andrew Finch, Paul Dixon and Eiichiro Sumita

9:00–9:20 *Simple Discriminative Training for Machine Transliteration*
Canasai Kruengkrai, Thatsanee Charoenporn and Virach Sornlertlamvanich

9:20–9:40 *English-Korean Named Entity Transliteration Using Statistical Substring-based and Rule-based Approaches*
Yu-Chun Wang and Richard Tzong-Han Tsai

9:40–10:00 *Leveraging Transliterations from Multiple Languages*
Aditya Bhargava, Bradley Hauer and Grzegorz Kondrak

10:00-10:30 Morning Break

November 12, 2011 (continued)

10:00-12:00 Session 2

10:00–10:30 *Comparative Evaluation of Spanish Segmentation Strategies for Spanish-Chinese Transliteration*
Rafael E. Banchs

10:30–11:00 *Using Features from a Bilingual Alignment Model in Transliteration Mining*
Takaaki Fukunishi, Andrew Finch, Seiichi Yamamoto and Eiichiro Sumita

11:00–11:30 *Product Name Identification and Classification in Thai Economic News*
Nattadaporn Lertcheva and Wirote Aroonmanakun

11:30–12:00 *Mining Multi-word Named Entity Equivalents from Comparable Corpora*
Abhijit Bhole, Goutham Tholpadi and Raghavendra Udupa

12:00–14:00 Lunch Break

14:00-15:30 Session 3

14:00–14:30 *An Unsupervised Alignment Model for Sequence Labeling: Application to Name Transliteration*
Najmeh Mousavi Nejad and Shahram Khadivi

14:30–14:50 *Forward-backward Machine Transliteration between English and Chinese Based on Combined CRFs*
Ying Qin and GuoHua Chen

14:50–15:10 *English-to-Chinese Machine Transliteration using Accessor Variety Features of Source Graphemes*
Mike Tian-Jian Jiang, Chan-Hung Kuo and Wen-Lian Hsu

15:10–15:30 *The Amirkabir Machine Transliteration System for NEWS 2011: Farsi-to-English Task*
Najmeh Mousavi Nejad, Shahram Khadivi and Kaveh Taghipour

15:30–16:00 Afternoon Break

November 12, 2011 (continued)

16:00-17:00 Session 4

16:00–16:20 *English-Chinese Personal Name Transliteration by Syllable-Based Maximum Matching*
Oi Yee Kwong

16:20–16:40 *Statistical Machine Transliteration with Multi-to-Multi Joint Source Channel Model*
Yu Chen, Rui Wang and Yi Zhang

16:40–17:00 *Named Entity Transliteration Generation Leveraging Statistical Machine Translation Technology*
Pradeep Dasigi and Mona Diab

17:00-17:10 Closing