ACL 2015

Proceedings of
NEWS 2015
The Fifth Named Entities Workshop

Xiangyu Duan, Rafael E. Banchs, Min Zhang, Haizhou Li, A. Kumara
(Editors)

July 31, 2015
Beijing, China

Order print-on-demand copies 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 978-1-941643-65-5

Preface

The workshop series, Named Entities WorkShop (NEWS), focus 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; the third one, NEWS 2011, was held as an IJCNLP 2011 workshop in Chiang Mai, Thailand; and the fourth one, NEWS 2012, was held as an ACL 2012 workshop in Jeju, Korea. The current edition, NEWS 2015, was held as an ACL-IJCNLP 2015 workshop in Beijing, China.

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, 5 research papers were submitted, each paper was reviewed by at least 2 reviewers from the program committee. The 5 papers were all chosen for publication, covering named entity recognition and machine transliteration, which applied various new trend methods such as deep neural networks and graph-based semi-supervised learning.

Following the tradition of the NEWS workshop series, NEWS 2015 continued the machine transliteration shared task this year as well. The shared task was first introduced in NEWS 2009 and continued in NEWS 2010, NEWS 2011, and NEWS 2012. In NEWS 2015, 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, while one team withdrew their results at the evaluation phase. Finally, we received 6 teams' submissions. The approaches ranged from traditional learning methods (such as, Phrasal SMT-based, Conditional Random Fields, etc.) to somewhat new approaches (such as, neural network transduction, integration of transliteration mining, hybrid system combination). A concrete study and targeted process between two languages often generate better performances. 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 two members of the program committee to help improve the quality. All the 6 system papers were finally accepted to be published in the workshop proceedings.

We hope that NEWS 2015 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 Organizers:
Min Zhang, Soochow University, China
Haizhou Li, Institute for Infocomm Research, Singapore
Rafael E Banchs, Institute for Infocomm Research, Singapore
A Kumaran, Microsoft Research, India
Xiangyu Duan, Soochow University, China

July 31, 2015
Beijing, China

Organizers:

Min Zhang, Soochow University, China
Haizhou Li, Institute for Infocomm Research, Singapore
Rafael E Banchs, Institute for Infocomm Research, Singapore
A Kumaran, Microsoft Research, India
Xiangyu Duan, Soochow University

Program Committee:

Rafael E. Banchs, Institute for Infocomm Research
Sivaji Bandyopadhyay, Jadavpur University
Marta R. Costa-jussà, Instituto Politécnico Nacional
Xiangyu Duan, Soochow University
Guohong Fu, Heilongjiang University
Sarvnaz Karimi, CSIRO
Mitesh M. Khapra, IBM Research India
Grzegorz Kondrak, University of Alberta
Jong-Hoon Oh, NICT
Richard Sproat, Google
Keh-Yih Su, Institute of Information Science, Academia Sinica
Raghavendra Udupa, Microsoft Research India
Chai Wutiwiwatchai, Intelligent Informatics Research Unit, National Electronics and Computer Technology Center
Deyi Xiong, Soochow University
Muyun Yang, Harbin Institute of Technology
Min Zhang, Soochow University

Table of Contents

Conference Program

Friday, July 31, 2015

9:05–9:15 ***Opening Remarks***

Whitepaper of NEWS 2015 Shared Task on Machine Transliteration
Min Zhang, Haizhou Li, Rafael E. Banchs and A. Kumaran

Report of NEWS 2015 Machine Transliteration Shared Task
Rafael E. Banchs, Min Zhang, Xiangyu Duan, Haizhou Li and A. Kumaran

9:15–10:05 ***Keynote Speech***

How do you spell that? A journey through word representations
Greg Kondrak

10:05–12:15 **Research Papers**

10:05–10:30 *Boosting Named Entity Recognition with Neural Character Embeddings*
Cicero dos Santos and Victor Guimarães

10:30–11:00 ***Coffee Break***

11:00–11:25 *Regularity and Flexibility in English-Chinese Name Transliteration*
Oi Yee Kwong

11:25–11:50 *HAREM and Klue: how to put two tagsets for named entities annotation together*
Livy Real and Alexandre Rademaker

11:50–12:15 *Semi-supervised Learning for Vietnamese Named Entity Recognition using Online Conditional Random Fields*
Quang Hong Pham, Minh-Le Nguyen, Thanh Binh Nguyen and Nguyen Viet Cuong

12:15–13:50 ***Lunch Break***

Friday, July 31, 2015 (continued)

13:50–16:50 System Papers

13:50–14:15 *Boosting English-Chinese Machine Transliteration via High Quality Alignment and Multilingual Resources*
Yan Shao, Jörg Tiedemann and Joakim Nivre

14:15–14:40 *Neural Network Transduction Models in Transliteration Generation*
Andrew Finch, Lemao Liu, Xiaolin Wang and Eiichiro Sumita

14:40–15:05 *A Hybrid Transliteration Model for Chinese/English Named Entities —BJTU-NLP Report for the 5th Named Entities Workshop*
Dandan Wang, Xiaohui Yang, Jinan Xu, Yufeng Chen, Nan Wang, Bojia Liu, Jian Yang and Yujie Zhang

15:05–15:30 *Multiple System Combination for Transliteration*
Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Adam St Arnaud, Ying Xu, Lei Yao and Grzegorz Kondrak

15:30–16:00 *Coffee Break*

16:00–16:25 *Data representation methods and use of mined corpora for Indian language transliteration*
Anoop Kunchukuttan and Pushpak Bhattacharyya

16:25–16:50 *NCU IISR English-Korean and English-Chinese Named Entity Transliteration Using Different Grapheme Segmentation Approaches*
Yu-Chun Wang, Chun-Kai Wu and Richard Tzong-Han Tsai

16:50–17:00 Closing