ACL-IJCNLP 2009

NEWS 2009

2009 Named Entities Workshop: Shared Task on Transliteration

Proceedings of the Workshop

7 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

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 978-1-932432-57-2 / 1-932432-57-4

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. This is the specific research focus for the Named Entities WorkShop (NEWS), being held as a part of ACL-IJCNLP 2009 conference.

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. Under such broad scope as above, many interesting specific research areas pertaining to the named entities are identified, 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. 17 research papers were submitted, each of which was reviewed by at least 3 reviewers from the program committee. Finally, 9 papers were chosen for publication, covering main research areas, from named entities tagging and extraction, to computational phonology to machine transliteration of named entities. All accepted research papers are published in the workshop proceedings.

An important part of the NEWS workshop is the shared task on Machine Transliteration of named entities. Machine transliteration is a vibrant research area as witnessed by increasing number of publications over the last decade in the Computational Linguistics, Natural Language Processing (ACL, EACL, NAACL, IJCNLP, COLING, HLT, EMNLP, etc.), and Information Retrieval (SIGIR, ECIR, AIRS, etc.) conferences, and primarily in languages that use non-Latin based scripts. However, in spite of its popularity, no meaningful comparison could be possible between the research approaches, as the publications tended to be on different language pairs and different datasets, and on a variety of different metrics. For the first time, we organize a shared task as part of the NEWS workshop to provide a common evaluation platform for benchmarking and calibration of transliteration technologies.

We collected significantly large, hand-crafted parallel named entities corpora in 7 different languages from 6 language families, and made available as common dataset for the shared task. We defined 6 metrics that are language-independent, intuitive and computationally easy to compute. 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 31 teams participated from around the world, including industry, government laboratories and academia. The approaches ranged from traditional unsupervised learning methods (such as, naive-Bayes, Phrasal SMT-based, Conditional Random Fields, etc.) to somewhat unique approaches (such as, sequence prediction models, to Minimum Description Length-based methods, etc.), combined with several model combinations for results re-ranking. While every team submitted *standard runs* that use only the data provided by the NEWS organizers, many teams also submitted *non-standard runs* where they were allowed to use any additional data or language specific modules. In total, about 190 task runs were submitted, covering most approaches comprehensively. 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; 27 of them were finally are accepted to publish in the workshop proceedings.

NEWS 2009 is the first workshop that specifically addresses comprehensively all research avenues concerned with named entities, to the best of our knowledge. Also, the transliteration shared task is the first of its kind, to calibrate such large number of systems using common metrics on common language-specific datasets in a comprehensive set of language pairs.

We hope that NEWS 2009 would provide an exciting and productive forum for researchers working in this research area. The technical programme includes 9 research papers and 27 system papers to be presented in the workshop. Further, we are pleased to have invited Dr Kevin Knight to deliver a keynote speech in the workshop. Dr Knight is a well-known researcher in natural language processing, an Associate Professor at University of Southern California, and the Founder and Chief Scientist of Language Weaver, a human communication solutions company.

We wish to thank all the researchers for their research submission and the enthusiastic participation in the transliteration shared task. We wish to express our gratitude for the data providers (CJK Institute, Institute for Infocomm Research and Microsoft Research India) for the shared task. 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*

7 August 2009

Organizers

Chairs

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

Organizing Committee

Sanjeev Khudanpur, *Johns Hopkins University, USA*
Raghavendra Udupa, *Microsoft Research, India*
Min Zhang, *Institute for Infocomm Research, Singapore*
Monojit Choudhury, *Microsoft Research, 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-jussà, *UPC, Spain*
Gregory Grefenstette, *Exalead, France*
Sanjeev Khudanpur, *Johns Hopkins University, USA*
Kevin Knight, *University of Southern California/ISI, USA*
Greg Kondrak, *University of Alberta, Canada*
A Kumaran, *Microsoft Research, India*
Olivia Kwong, *City University, Hong Kong*
Gina-Anne Levow, *University of Chicago, USA*
Haizhou Li, *Institute for Infocomm Research, Singapore*
Raghavendra Udupa, *Microsoft Research, India*
Arul Menezes, *Microsoft Research, USA*
Jong-Hoon Oh, *NICT, Japan*
Vladimir Pervouchine, *Institute for Infocomm Research, Singapore*
Yan Qu, *Advertising.com, USA*
Sunita Sarawagi, *IIT-Bombay, India*
Sudeshna Sarkar, *IIT-Kharagpur, India*
Richard Sproat, *University of Illinois, Urbana-Champaign, USA*
Keh-Yih Su, *Behavior Design Corporation, Taiwan*
Raghavendra Udupa, *Microsoft Research, India*
Vasudeva Varma, *IIIT-Hyderabad, India*
Min Zhang, *Institute for Infocomm Research, Singapore*

Table of Contents

Conference Program

Friday, August 7, 2009

8:30–9:10 Opening Remarks

Overview of the Shared Tasks

Report of NEWS 2009 Machine Transliteration Shared Task
Haizhou Li, A Kumaran, Vladimir Pervouchine and Min Zhang

Keynote Speech

9:10–10:00 *Automata for Transliteration and Machine Translation*
Kevin Knight

10:00–10:30 Coffee Break

Session 1: Shared Task Paper Presentation

10:30–10:45 *DirecTL: a Language Independent Approach to Transliteration*
Sittichai Jiampojamarn, Aditya Bhargava, Qing Dou, Kenneth Dwyer and Grzegorz Kondrak

10:45–11:00 *Named Entity Transcription with Pair n-Gram Models*
Martin Jansche and Richard Sproat

11:00–11:15 *Machine Transliteration using Target-Language Grapheme and Phoneme: Multi-engine Transliteration Approach*
Jong-Hoon Oh, Kiyotaka Uchimoto and Kentaro Torisawa

11:15–11:30 *A Language-Independent Transliteration Schema Using Character Aligned Models at NEWS 2009*
Praneeth Shishtla, Surya Ganesh V, Sethuramalingam Subramaniam and Vasudeva Varma

11:30–11:45 *Experiences with English-Hindi, English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009*
Manoj Kumar Chinnakotla and Om P. Damani

Friday, August 7, 2009 (continued)

12:00–13:50 Lunch Break

Session 2: Posters

13:50–15:30 Poster Presentation

Testing and Performance Evaluation of Machine Transliteration System for Tamil Language
Kommaluri Vijayanand

Transliteration by Bidirectional Statistical Machine Translation
Andrew Finch and Eiichiro Sumita

Transliteration of Name Entity via Improved Statistical Translation on Character Sequences
Yan Song, Chunyu Kit and Xiao Chen

Learning Multi Character Alignment Rules and Classification of Training Data for Transliteration
Dipankar Bose and Sudeshna Sarkar

Fast Decoding and Easy Implementation: Transliteration as Sequential Labeling
Eiji Aramaki and Takeshi Abekawa

NEWS 2009 Machine Transliteration Shared Task System Description: Transliteration with Letter-to-Phoneme Technology
Colin Cherry and Hisami Suzuki

Combining a Two-step Conditional Random Field Model and a Joint Source Channel Model for Machine Transliteration
Dong Yang, Paul Dixon, Yi-Cheng Pan, Tasuku Oonishi, Masanobu Nakamura and Sadaoki Furui

Phonological Context Approximation and Homophone Treatment for NEWS 2009 English-Chinese Transliteration Shared Task
Oi Yee Kwong

English to Hindi Machine Transliteration System at NEWS 2009
Amitava Das, Asif Ekbal, Tapabrata Mondal and Sivaji Bandyopadhyay

Improving Transliteration Accuracy Using Word-Origin Detection and Lexicon Lookup
Mitesh Khapra and Pushpak Bhattacharyya

Friday, August 7, 2009 (continued)

A Noisy Channel Model for Grapheme-based Machine Transliteration
Jia Yuxiang, Zhu Danqing and Yu Shiwen

Substring-based Transliteration with Conditional Random Fields
Sravana Reddy and Sonjia Waxmonsky

A Syllable-based Name Transliteration System
Xue Jiang, Le Sun and Dakun Zhang

Transliteration System Using Pair HMM with Weighted FSTs
Peter Nabende

English-Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009
Rejwanul Haque, Sandipan Dandapat, Ankit Kumar Srivastava, Sudip Kumar Naskar and Andy Way

A Hybrid Approach to English-Korean Name Transliteration
Gumwon Hong, Min-Jeong Kim, Do-Gil Lee and Hae-Chang Rim

Language Independent Transliteration System Using Phrase-based SMT Approach on Substrings
Sara Noeman

Combining MDL Transliteration Training with Discriminative Modeling
Dmitry Zelenko

ϵ-extension Hidden Markov Models and Weighted Transducers for Machine Transliteration
Balakrishnan Varadarajan and Delip Rao

Modeling Machine Transliteration as a Phrase Based Statistical Machine Translation Problem
Taraka Rama and Karthik Gali

Maximum n-Gram HMM-based Name Transliteration: Experiment in NEWS 2009 on English-Chinese Corpus
Yilu Zhou

Name Transliteration with Bidirectional Perceptron Edit Models
Dayne Freitag and Zhiqiang Wang

Friday, August 7, 2009 (continued)

Bridging Languages by SuperSense Entity Tagging
Davide Picca, Alfio Massimiliano Gliozzo and Simone Campora

Chinese-English Organization Name Translation Based on Correlative Expansion
Feiliang Ren, Muhua Zhu, Huizhen Wang and Jingbo Zhu

Name Matching between Roman and Chinese Scripts: Machine Complements Human
Ken Samuel, Alan Rubenstein, Sherri Condon and Alex Yeh

Analysis and Robust Extraction of Changing Named Entities
Masatoshi Tsuchiya, Shoko Endo and Seiichi Nakagawa

15:30–16:00 Coffee Break

Session 3: Research Paper Presentation

16:00–16:20 *Tag Confidence Measure for Semi-Automatically Updating Named Entity Recognition*
Kuniko Saito and Kenji Imamura

16:20–16:40 *A Hybrid Model for Urdu Hindi Transliteration*
Abbas Malik, Laurent Besacier, Christian Boitet and Pushpak Bhattacharyya

16:40–17:00 *Graphemic Approximation of Phonological Context for English-Chinese Transliteration*
Oi Yee Kwong

17:00–17:20 *Czech Named Entity Corpus and SVM-based Recognizer*
Jana Kravalova and Zdenek Zabokrtsky

17:20–17:40 *Voted NER System using Appropriate Unlabeled Data*
Asif Ekbal and Sivaji Bandyopadhyay