COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT

Workshop chairs:
Josef van Genabith, Toni Badia, Christian Federmann,
Maite Melero, Marta R. Costa-jussà and Tsuyoshi Okita

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT
Josef van Genabith, Toni Badia, Christian Federmann, Maite Melero, Marta R. Costa-jussà and Tsuyoshi Okita (eds.)
Revised preprint edition, 2012

Published by The COLING 2012 Organizing Committee
Indian Institute of Technology Bombay,
Powai,
Mumbai-400076
India
Phone: 91-22-25764729
Fax: 91-22-2572 0022
Email: pb@cse.iitb.ac.in

Also available online in the ACL Anthology at http://aclweb.org

Message from the Workshop organisers

We are delighted to welcome you to the of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT and associated Shared Task (ML4HMT-2012) in Mumbai.

The Shared Task is an effort to trigger systematic investigation on improving state-of-the-art Hybrid MT, using advanced machine-learning (ML) methodologies. Its main focus is trying to answer the following question: *Can Hybrid/System Combination MT techniques benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved?*

Participants to the challenge are requested to build hybrid translations by combining the output of several MT systems of different types. Five participating combination systems, each following a different solution strategy, have been submitted to the shared task.

The Workshop will be composed of two parts. In the first part we will have an invited talk and the presentation of three research papers. In the second part, participants to the shared task will describe their systems and results. At the end of this part, there will be a presentation of the joint evaluation, followed by a discussion panel.

We are looking forward to an interesting workshop and want to thank all authors, presenters and attendees for making this a successful workshop.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.:249119).

We thank the organisers of COLING 2012 for their support.

Organisation committee

Prof. Josef van Genabith, Dublin City University (DCU) and Centre for Next Generation Localisation (CNGL)

Prof. Toni Badia, Universitat Pompeu Fabra and Barcelona Media (BM)

Christian Federmann, German Research Center for Artificial Intelligence (DFKI), contact person: cfedermann@dfki.de

Dr. Maite Melero, Barcelona Media (BM)

Dr. Marta R. Costa-jussà, Barcelona Media (BM)

Dr. Tsuyoshi Okita, Dublin City University (DCU)

The ML4HMT-2012 workshop is supported by META NET

Organizers:

Prof. Josef van Genabith (Dublin City University (DCU) and Centre for Next Generation Localisation (CNGL))
Prof. Toni Badia (Universitat Pompeu Fabra and Barcelona Media (BM))
Christian Federmann (German Research Center for Artificial Intelligence (DFKI))
Dr. Maite Melero (Barcelona Media (BM))
Dr. Marta R. Costa-jussà (Barcelona Media (BM))
Dr. Tsuyoshi Okita (Dublin City University (DCU))

Programme Committee:

Eleftherios Avramidis (German Research Center for Artificial Intelligence, Germany)
Prof. Sivaji Bandyopadhyay (Jadavpur University, India)
Dr. Rafael Banchs (Institute for Infocomm Research I2R, Singapore)
Prof. Loïc Barrault (LIUM University of Le Mans, France)
Prof. Antal van den Bosch (Centre for Language Studies, Radboud University Nijmegen, Netherlands)
Dr. Grzegorz Chrupala (Saarland University, Saarbrücken, Germany)
Prof. Jinhua Du (Xi'an University of Technology (XAUT), China)
Dr. Andreas Eisele (DirectorateGeneral for Translation (DGT), Luxembourg)
Dr. Cristina EspañaBonet (Technical University of Catalonia, TALP, Barcelona)
Dr. Declan Groves (Center for Next Generation Localisation, Dublin City University, Ireland)
Prof. Jan Hajic (Institute of Formal and Applied Linguistics, Charles University in Prague)
Prof. Timo Honkela (Aalto University, Finland)
Dr. Patrick Lambert (LIUM University of Le Mans, France)
Prof. Qun Liu (Institute of Computing Technology, Chinese Academy of Sciences, China)
Dr. Maite Melero (Barcelona Media Innovation Center, Spain)
Dr. Tsuyoshi Okita (Dublin City University, Ireland)
Prof. Pavel Pecina (Institute of Formal and Applied Linguistics, Charles University in Prague)
Dr. Marta R. Costajussà (Barcelona Media Innovation Center, Spain)
Dr. Felipe Sanchez Martinez (Escuela Politecnica Superior, Universidad de Alicante, Spain)
Dr. Nicolas Stroppa (Google, Zurich, Switzerland)
Prof. Hans Uszkoreit (German Research Center for Artificial Intelligence, Germany)
Dr. David Vilar (German Research Center for Artificial Intelligence, Germany)

Table of Contents

Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT

Program

Saturday, 15 December 2012

09:00–09:15 Josef van Genabith — Welcome and introductory remarks

09:15–09:40 *Hybrid Adaptation of Named Entity Recognition for Statistical Machine Translation*
Vassilina Nikoulina, Agnes Sandor and Marc Dymetman

09:40–10:05 *Confusion Network Based System Combination for Chinese Translation Output: Word-Level or Character-Level?*
Maoxi Li and MingWen Wang

10:05–10:30 *Using Cross-Lingual Explicit Semantic Analysis for Improving Ontology Translation*
Kartik Asooja, Jorge Gracia, Nitish Aggarwal and Asunción Gómez Pérez

10:30–10:50 *System Combination with Extra Alignment Information*
Xiaofeng Wu, Tsuyoshi Okita, Josef van Genabith and Qun Liu

10:50–11:10 *Topic Modeling-based Domain Adaptation for System Combination*
Tsuyoshi Okita, Antonio Toral and Josef van Genabith

11:10–11:30 *Sentence-Level Quality Estimation for MT System Combination*
Tsuyoshi Okita, Raphaël Rubino and Josef van Genabith

11:30–11:45 Tea break

11:45–12:05 *Neural Probabilistic Language Model for System Combination*
Tsuyoshi Okita

12:05–12:25 *System Combination Using Joint, Binarised Feature Vectors*
Christian Federmann

12:25–12:30 *Results from the ML4HMT-12 Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation*
Christian Federmann, Tsuyoshi Okita, Maite Melero, Marta R. Costa-Jussa, Toni Badia and Josef van Genabith

12:30–12:50 **Discussion Panel**
Panelists: Marc Dymetman (TBC), Jan Hajič, Qun Liu (TBC), Hans Uszkoreit, Josef van Genabith
Topics include:

- The Future of Hybrid MT: is there a single-paradigm winner?
- Will we see increasing usage of additional, potentially highly sparse, features?
- Will research efforts in Machine Translation and Machine Learning converge?
- How do we evaluate progress in terms of translation quality for Hybrid MT?
- What are the baselines? Can Human Judgment be integrated?

12:50–13:30 **Invited talk:**
Deep Linguistic Information in Hybrid Machine Translation
Jan Hajič, Institute of Formal and Applied Linguistics, Charles University in Prague