2012 AMTA 20 Years

The Tenth Biennial Conference of the
Association for Machine Translation in the Americas

Monolingual Machine Translation

Tsuyoshi Okita
Dublin City University

Artem Sokolov
LIMSI

Taro Watanabe
National Institute of Information
and Communications Technology

SAN DIEGO, CA
OCTOBER 28- NOVEMBER 1, 2012

Proceedings of the

Monolingual Machine Translation-2012 Workshop

Collocated with the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2012)

Edited by

Tsuyoshi Okita

and

Artem Sokolov

and

Taro Watanabe

1 November 2011

San Diego, USA

Preface to the MONOMT-2012 Workshop Proceedings

On behalf of the program committee of the Monolingual Machine Translation workshop (MONOMT-2012), it is our pleasure to present you this proceedings. The MONOMT-2012 workshop was held on November 1 2012, co-located with the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012) that took place from October 29 to 31, 2012 in San Diego, USA.

This workshop was the first attempt to showcase various monolingual machine translation subtasks which we frequently encounter in the recent MT research. Such subtasks include MT for morphologically rich languages [Bojar, 08], system combination strategy [Matusov et al., 05], statistical post-editing [Dugast et al., 07; Simard et al., 07], domain adaptation [Daume III, 07], two-step approach to a long-range reordering strategy (SVO and SOV) [Isozaki et al., 10], MERT process [Arun et al., 10], and translation memory and MT integration strategy [Ma et al., 11]. If we permit human-aided translation instead of MT, such as error identification and voting with independent monolingual crowdsources [Hu et al., 11] and monolingual Machine Translator [Koehn, 10], this would further enlarge the area of monolingual translation subtasks. With this showcasing, our intention was to preferably go further to obtain algorithmic tools: the modest goal would be in the form of monolingual MT tools such as MBR decoding, monolingual word alignment (based on TER and METEOR), language model learnt by its representation of data, and machine learning strategies, while the ultimate goal would be to build a general-purpose device *Monolingual Machine Translation system*. Note that we did not intend to replace the current available SMT systems, such systems would complement them.

We received 8 submissions, of which 3 were accepted as long presentations while 4 were accepted as short presentations. Submissions came from different countries: China, Czech Republic / Indonesia, Germany, India, Japan, Spain, UK and US. All the papers concerned various aspects of monolingual MT subtasks. Three papers proposed interesting ways to use monolingual resources via automatically created multiple references, domain clustering using huge web crawled data, and morphology prediction system. Regretfully, no paper concerned *generic monolingual MT tools*, probably because of the short preparation period. In this sense, our principal goal largely remains untouched until, we hope, near future; although some of our invited speakers might kindly mention them in their talks. We would like to thank invited speakers, who accepted to give a talk at this workshop: Prof. Marcello Federico, Prof. Philipp Koehn, Prof. Qun Liu, Dr. Evgeny Matusov, Dr. Taro Watanabe (At the time of writing, this list was not complete).

We note that this workshop has close links to two previous workshops: MTML 2011 workshop and ML4HMT 2011 workshop. An MTML 2011 workshop (Machine Translation and Morphologically-Rich Languages, Research Workshop of the Israel Science Foundation) was held in January 2011 in Haifa. This workshop concerned machine translation from morphologically poor into morphologically rich languages. Among its presentations, there were two that had mentioned a two-step approach for MT of morphologically rich languages. An ML4HMT 2011 workshop (Machine Learning Techniques to Optimizing the Division of Labour in Hybrid MT) was held in November 2011 in Barcelona. This workshop focused on incorporating of semantic / syntactic meta knowledge into system combination strategy. Typically, such strategy handles only monolingual MT and

we regard this as our starting point.

This workshop would not be possible without the valued help of the AMTA General Chair Prof. Alon Lavie and the AMTA workshop organizer Dr. Désillets Alain. We particularly thank Prof. Lavie for inspiring us to organize this workshop through the above two workshops. We would also like to thank all the members of the program committee half of them overlapping with the above two workshops, who have dedicated their valuable time to review the papers. Finally, we gratefully acknowledge the sponsorship of Prof. Josef van Genabith, CNGL / DCU (Center for Next Generation Localisation / Dublin City University).

1st November 2012
Tsuyoshi Okita, Artem Sokolov, and Taro Watanabe
MONOMT-2012 Workshop

Workshop Chairs

Tsuyoshi Okita (Dublin City University, Ireland)
Artem Sokolov (LIMSI, France)
Taro Watanabe (National Institute of Information and Communications Technology, Japan)

Program Committee

Bogdan Babych (University of Leeds, UK)
Loic Barrault (LIUM, Universite du Maine, France)
Nicola Bertoldi (FBK, Italy)
Ergun Bicici (CNGL, Dublin City University, Ireland)
Ondrej Bojar (Charles University, Czech)
Boxing Chen (NRC Institute for Information Technology, Canada)
Trevor Cohn (University of Sheffield, UK)
Marta Ruiz Costa-jussa (Barcelona Media, Spain)
Josep M. Crego (SYSTRAN, France)
John DeNero (Google, USA)
Jinhua Du (Xi'an University of Technology, China)
Kevin Duh (Nara Institute of Science and Technology, Japan)
Chris Dyer (CMU, USA)
Christian Federmann (DFKI, Germany)
Yvette Graham (University of Melbourne, Australia)
Barry Haddow (University of Edinburgh, UK)
Xiadong He (Microsoft, USA)
Jagadeesh Jagarlamudi (University of Maryland, USA)
Jie Jiang (Applied Language Solutions, UK)
Philipp Koehn (University of Edinburgh, UK)
Shankar Kumar (Google, USA)
Alon Lavie (CMU, USA)
Yanjun Ma (Baidu, China)
Aurelien Max (LIMSI, University Paris Sud, France)
Maite Melero (Barcelona Media, Spain)
Philip Resnik (University of Maryland, USA)
Stefan Riezler (University of Heidelberg, Germany)
Lucia Specia (University of Sheffield, UK)
Marco Turchi (JRC, Italy)
Antal van den Bosch (Radboud University Nijmegen, Netherlands)
Xianchao Wu (Baidu, Japan)
Dekai Wu (HKUST, Hong Kong)
Francois Yvon (LIMSI, University Paris Sud, France)

Table of Contents

Long Presentations

Short Presentations