ACL 2013

Second Workshop on Hybrid Approaches to Translation

Proceedings of the Workshop

August 8, 2013
Sofia, Bulgaria

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53704 USA

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 978-1-937284-63-3

Preface

This second edition of the Workshop on Hybrid Approaches to Translation (HyTra) is co-located with the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013) in Sofia. It further progresses on the findings of the first edition which was held as a joint 2-day event together with the Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012) in Avignon.

The aim of the HyTra workshop series is to bring together and share ideas among MT researchers who combine data-driven statistical approaches with linguistic knowledge models. We open the floor for researchers and groups who develop and improve machine translation systems across different paradigms: rule-based, example-based, statistical or hybrid. The workshop provides a platform for publishing their work, and contributes towards building a research community in the field of hybrid MT, around sharing a common vision, methods, evaluation benchmarks and tools. The uniting focus for this community is a new cross-paradigm view of the area of machine translation, seeing the potential to move the technology beyond the state-of-the-art by combining ideas and models developed in different fields of computational linguistics and artificial intelligence. This workshop gives an opportunity to motivate the cooperation and interaction between them, and to foster innovative combinations between the two main MT paradigms: statistical and rule-based.

The advantages of rule-based MT are that its rules and representations are geared towards human understanding and can be more easily checked, corrected and exploited for applications outside of machine translation such as dictionaries, text understanding and dialog systems. But (pure) rule-based MT has also severe disadvantages, among them slow development cycles, high cost, a lack of robustness in the case of incorrect input, and difficulties in making correct choices with respect to ambiguous words, structures, and transfer equivalents.

The advantages of statistical MT are fast development cycles, low cost, robustness, superior lexical selection and relative fluency due to the use of language models. But (pure) statistical MT has also disadvantages: it needs large amounts of data, which for many language pairs are not available, and are unlikely to become available in the foreseeable future. This problem is especially relevant for under-resourced languages. Recent advances in factored morphological models and syntax-based models in SMT indicate that non-statistical symbolic representations and processing models need to have their proper place in MT research and development, and more research is needed to understand how to develop and integrate these non-statistical models most efficiently.

The translations of statistical systems are often surprisingly good with respect to phrases and short distance collocations, but they often fail when preferences need to be based on more distant words. In contrast, the output of rule-based systems is often surprisingly good if the parser assigns the correct analysis to a sentence. However, it usually leaves something to be desired if the correct analysis cannot be computed, or if there is not enough information for selecting the correct target words when translating ambiguous words and structures.

Given the complementarity of rule-based and statistical MT, it is natural that the boundaries between them have narrowed. The question is what the combined architecture should look like. In the past few years, in the MT scientific community, the interest in hybridization and system combination has significantly increased. This is why a large number of approaches for constructing hybrid MT have already been proposed offering a considerable potential of improving MT quality and efficiency. Mainly, the following hybrid MT systems can be identified: (1) SMT models augmented with morphological, syntactic or semantic information; (2) Rule-based MT systems using parallel and comparable corpora to improve results by enriching their lexicons and grammars and by applying

new methods for disambiguation; (3) MT system combination based on different paradigms (including voting systems); (4) automatic and semi-automatic pre-editing and post-editing approaches, including re-ordering systems.

There is also great potential in expanding hybrid MT systems with techniques, tools and processing resources from other areas of NLP, such as Information Extraction, Information Retrieval, Question Answering, Semantic Web, Automatic Semantic Inferencing.

Given this context, relevant topics for the workshop series include the following:

- ways and techniques of hybridization
- architectures for the rapid development of hybrid MT systems
- applications of hybrid systems
- hybrid systems dealing with under-resourced languages
- hybrid systems dealing with morphologically rich languages
- using linguistic information (morphology, syntax, semantics) to enhance statistical MT (e.g. with hierarchical or factored models)
- using contextual information to enhance statistical MT
- bootstrapping rule-based systems from corpora
- hybrid methods in spoken language translation
- extraction of dictionaries and other large-scale resources for MT from parallel and comparable corpora
- induction of morphological, grammatical, and translation rules from corpora
- machine learning techniques for hybrid MT
- describing structural mappings between languages (e.g. tree-structures using synchronous/ transduction grammars)
- heuristics for limiting the search space in hybrid MT
- alternative methods for the fair evaluation of the output of different types of MT systems (e.g. relying on linguistic criteria)
- system combination approaches such as multi-engine MT (parallel) or automatic post-editing (sequential)
- open source tools and free language resources for hybrid MT

From this range most contributors of the current workshop have chosen to present work about how SMT may be improved by adding linguistic knowledge and representation respectively. For some of the papers this means to add morphological or morpho-syntactic representation levels - and to define the lexicon- and language-models for these representations instead of considering inflected words or chunks of inflected words; for others this (also) means to incorporate pre-processing components for reordering the input (that, possibly, has been morphologically analyzed before). This set of papers where SMT is taken as a basis is complemented by a few papers dedicated to integrating statistical information – mainly about lexical selection and disambiguation - in RBMT systems; and by another few papers

concentrating on extracting information for MT from monolingual resources (including analysis learning for RBMT). A small number of contributions include general considerations about hybrid architectures as such. However, a clear trend in the sense of a convention about hybridity coming into being cannot be entailed from the contributions, not yet. This encourages continuation of the series.

This second HyTra workshop has been supported by the Seventh Framework Programme of the European Commission through the Marie Curie actions HyghTra ("A Hybrid Hygh-Quality Translation System"; grant agreement no.: 251534 - PIAP-GA-2009-251534-HyghTra), IMTraP (Integration of Machine Translation Paradigms, grant agreement no.: 2011-29951), AutoWordNet ("The Automatic Generation of Lexical Databases Analogous to WordNet"; grant agreement no. 254504) and CrossLingMind ("Automated analysis of opinions in a multilingual context"; grant agreement no. 300828). It has also been supported in part by Spanish "Ministerio de Economía y Competitividad", contract TEC2012-38939-C03-02 as well as from the European Regional Development Fund (ERDF/FEDER).

We would like to thank all people who contributed towards making the workshop a success. Our special thanks go to our invited keynote speakers: Hermann Ney (RWTH Aachen), Will Lewis and Chris Quirk (both Microsoft Research); as well as to our above mentioned sponsors, to the members of the program committee who did an excellent job in reviewing the submitted papers despite a very tight schedule, and to the ACL 2013 organizers, in particular the workshop general chairs Aoife Cahill and Qun Liu and the publication team including Roberto Navigli, Jing-Shin Chang, and Stefano Faralli. Last but not least, we would like to thank all authors and participants of the workshop, who have made this second edition of HyTra very successful.

Sofia, Bulgaria, August 2013

Marta R. Costa-jussà, Reinhard Rapp, Patrik Lambert, Kurt Eberle, Rafael E. Banchs, Bogdan Babych

Organizers:

Marta R. Costa-jussà, Institute for Infocomm Research, Singapore
Reinhard Rapp, Universities of Aix-Marseille, France and Mainz, Germany
Patrik Lambert, Barcelona Media Innovation Center, Spain
Kurt Eberle, Lingenio GmbH, Germany
Rafael E. Banchs, Institute for Infocomm Research, Singapore
Bogdan Babych, University of Leeds, UK

Invited Speakers:

Hermann Ney, RWTH Aachen, Germany
Will Lewis and Chris Quirk, Microsoft Research, USA

Program Committee:

Alexey Baytin, Yandex, Moscow, Russia
Núria Bel, Universitat Pompeu Fabra, Barcelona, Spain
Pierrette Bouillon, ISSCO/TIM/ETI, University of Geneva, Switzerland
Michael Carl, Copenhagen Business School, Denmark
Marine Carpuat, National Research Council, Canada
Josep Maria Crego, Systran, Paris, France
Oliver Čulo, University of Mainz, Germany
Andreas Eisele, DGT (European Commission), Luxembourg
Marcello Federico, Fondazione Bruno Kessler, Trento, Italy
Christian Federmann, Language Technology Lab, DFKI, Saarbrücken, Germany
Alexander Fraser, University of Stuttgart, Germany
José A. R. Fonollosa, Universitat Politècnica de Catalunya, Barcelona, Spain
Tony Hartley, Toyohashi University of Technology, Japan, and University of Leeds, UK
Maxim Khalilov, TAUS, Amsterdam, The Netherlands
Kevin Knight, University of Southern California, USA
Philipp Koehn, University of Edimburgh, UK
Udo Kruschwitz, University of Essex, UK
Yanjun Ma, Baidu Inc., Beijing, China
José B. Mariño, Universitat Politècnica de Catalunya, Barcelona, Spain
Maite Melero, Barcelona Media Innovation Center, Barcelona, Spain
Bart Mellebeek, University of Amsterdam, The Netherlands
Haizhou Li, Institute for Infocomm Research, Singapore
Chris Quirk, Microsoft, USA
Paul Schmidt, Institute for Applied Information Science, Saarbrücken, Germany
Anders Søgaard, University of Copenhagen, Denmark
Uta Seewald-Heeg, Anhalt University of Applied Sciences, Köthen, Germany
Nasredine Semmar, CEA LIST, Fontenay-aux-Roses, France
Wade Shen, Massachusetts Institute of Technology, Cambridge, USA
Serge Sharoff, University of Leeds, UK
George Tambouratzis, Institute for Language and Speech Processing, Athens, Greece
Jörg Tiedemann, University of Uppsala, Sweden
Dekai Wu, The Hong Kong University of Science and Technology, Hong Kong, China

Table of Contents

Workshop Program

Thursday, August 8, 2013

8:50–9:00 Workshop Opening

Workshop on Hybrid Approaches to Translation: Overview and Developments
Marta Ruiz Costa-jussà, Rafael Banchs, Reinhard Rapp, Patrik Lambert, Kurt Eberle and Bogdan Babych

9:00–9:50 Keynote Speech 1

Statistical MT Systems Revisited: How much Hybridity do they have?
Hermann Ney

Session 1: Morphology

09:50–10:15 *Hybrid Selection of Language Model Training Data Using Linguistic Information and Perplexity*
Antonio Toral

10:15–10:40 *Machine Learning Disambiguation of Quechua Verb Morphology*
Annette Rios Gonzales and Anne Göhring

10:40–11:00 Coffee Break

Session 2: Syntax I

11:00–11:25 *Improvements to Syntax-based Machine Translation using Ensemble Dependency Parsers*
Nathan Green and Zdeněk Žabokrtský

11:25–11:50 *Using unlabeled dependency parsing for pre-reordering for Chinese-to-Japanese statistical machine translation*
Dan Han, Pascual Martinez-Gomez, Yusuke Miyao, Katsuhito Sudoh and Masaaki NAGATA

Session 3: Syntax II

11:50–12:15 *Reordering rules for English-Hindi SMT*
Raj Nath Patel, Rohit Gupta, Prakash B. Pimpale and Sasikumar M

12:15–12:40 *English to Hungarian Morpheme-based Statistical Machine Translation System with Reordering Rules*
László Laki, Attila Novak and Borbála Siklósi

12:40–14:00 Lunch Break

Thursday, August 8, 2013 (continued)

14:00–14:50 Keynote Speech 2

Controlled Ascent: Imbuing Statistical MT with Linguistic Knowledge
William Lewis and Chris Quirk

14:50–16:00 **Session 4: Poster Session**

14:50–15:15 Poster Booster Presentations (5 minutes per poster)

Unsupervised Transduction Grammar Induction via Minimum Description Length
Markus Saers, Karteek Addanki and Dekai Wu

Integrating morpho-syntactic features in English-Arabic statistical machine translation
Ines Turki Khemakhem, Salma Jamoussi and Abdelmajid Ben Hamadou

Experiments with POS-based restructuring and alignment-based reordering for statistical machine translation
Shuo Li, Derek F. Wong and Lidia S. Chao

Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model
Rahma Boujelbane, Mariem Ellouze khemekhem, Siwar BenAyed and Lamia Hadrich-Belguith

A Hybrid Word Alignment Model for Phrase-Based Statistical Machine Translation
Santanu Pal, Sudip Naskar and Sivaji Bandyopadhyay

15:30–16:00 Coffee Break (to occur concurrently with poster session)

Thursday, August 8, 2013 (continued)

Session 5: Semantics

16:00–16:25 *Lexical Selection for Hybrid MT with Sequence Labeling*
Alex Rudnick and Michael Gasser

16:25–16:50 *Two Approaches to Correcting Homophone Confusions in a Hybrid Machine Translation System*
Pierrette Bouillon, Johanna Gerlach, Ulrich Germann, Barry Haddow and Manny Rayner

Session 6: Multi-level Approaches

16:50–17:15 *Uses of Monolingual In-Domain Corpora for Cross-Domain Adaptation with Hybrid MT Approaches*
An-Chang Hsieh, Hen-Hsen Huang and Hsin-Hsi Chen

17:15–17:40 *Language-independent hybrid MT with PRESEMT*
George Tambouratzis, Sokratis Sofianopoulos and Marina Vassiliou

17:40–17:50 Conclusions and Wrap-up Session