EACL 2014

14th Conference of the European Chapter of the Association for Computational Linguistics

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)

April 27, 2014
Gothenburg, Sweden

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

Gothenburg, April 2014 ISBN 978-1-937284-89-3

Introduction

The Third Workshop on Hybrid Approaches to Translation (HyTra) intends to further progress on the findings from the second HyTra, held at ACL 2013, and first HyTra which was held (together with the ESIRMT workshop) as a joint 2-day EACL 2012 workshop. The first editions of HyTra brought together researchers working on diverse aspects of hybrid machine translation. HyTra proceedings put together high-quality papers experimenting with current topics including statistical approaches integrating morphological, syntactic, semantic and rule-based information.

Machine Translation (MT) is a highly interdisciplinary and multidisciplinary field since it is approached from the point of view of human translators, engineers, computer scientists, mathematicians and linguists. This workshop aims at motivating the cooperation and interaction between them, and to foster innovative combinations between the two main MT paradigms: statistical and rule-based.

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 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 translations of statistical systems are often surprisingly good with respect to phrases and short distance collocations, but they often fail when selectional 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 statistical and rule-based MT, it is natural that the boundaries among 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. 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. The aim of the proposed workshop is to bring together and share ideas among researchers developing statistical, example-based, or rule-based translation systems and who enhance MT systems with elements from the other approaches. Hereby a focus will be on effectively combining linguistic and data driven approaches (rule-based and statistical MT).

Organizers:

Rafael E. Banchs (Institute for Infocomm Research, Singapore)
Marta R. Costa-jussà (Institute for Infocomm Research, Singapore)
Reinhard Rapp (Universities of Aix-Marseille and Mainz)
Patrik Lambert (Pompeu Fabra University, Barcelona)
Kurt Eberle (Lingenio GmbH, Heidelberg)
Bogdan Babych (University of Leeds)

Invited Speakers:

Hans Uszkoreit (Saarland University and DFKI, Germany) Abstract.
Joakim Nivre (Uppsala University, Sweden)

Program Committee:

Ahmet Aker, University of Sheffield, UK
Bogdan Babych, University of Leeds, UK
Rafael E. Banchs, Institute for Infocomm Research, Singapore
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
Marta R. Costa-jussa, Institute for Infocomm Research, Singapore
Oliver Culo, University of Mainz, Germany
Kurt Eberle, Lingenio GmbH, Heidelberg, Germany
Andreas Eisele, DGT (European Commission), Luxembourg
Marcello Federico, Fondazione Bruno Kessler, Trento, Italy
Christian Federmann, Language Technology Lab, DFKI, Saarbrücken, Germany
José A. R. Fonollosa, Universitat Politècnica de Catalunya, Barcelona, Spain
Maxim Khalilov, TAUS, Amsterdam, The Netherlands
Patrik Lambert, Pompeu Fabra University, Barcelona, Spain
Udo Kruschwitz, University of Essex, UK
Yanjun Ma, Baidu Inc., Beijing, China
José B. Mariño, Universitat Politècnica de Catalunya, Barcelona, Spain
Bart Mellebeek, University of Amsterdam, The Netherlands
Hermann Ney, RWTH Aachen, Germany
Reinhard Rapp, Universities of Aix-Marseille, France, and Mainz, Germany
Anders Søgaard, University of Copenhagen, Denmark
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

Table of Contents

Workshop Program

09:00-10:30 **Session 1**

09:00-09:45 Invited Talk: *Analytical Approaches to Combining MT Technologies*
Hans Uszkoreit

09:45-10:00 *Using Hypothesis Selection Based Features for Confusion Network MT System Combination*
Sahar Ghannay and Loïc Barrault

10:00-10:15 *Comparing CRF and template-matching in phrasing tasks within a Hybrid MT system*
George Tambouratzis

10:15-10:30 *Controlled Authoring In A Hybrid Russian-English Machine Translation System*
Svetlana Sheremetyeva

10:30-11:00 **Coffee Break**

11:00-12:45 **Session 2**

11:00-11:15 *Using Feature Structures to Improve Verb Translation in English-to-German Statistical MT*
Philip Williams and Philipp Koehn

11:15-11:30 *Building a Spanish-German Dictionary for Hybrid MT*
Anne Göhring

11:30-11:45 *An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese*
Giancarlo Salton, Robert Ross and John Kelleher

11:45-12:00 *Resumptive Pronoun Detection for Modern Standard Arabic to English MT*
Stephen Tratz, Clare Voss and Jamal Laoudi

12:00-12:15 *Automatic Building and Using Parallel Resources for SMT from Comparable Corpora*
Santanu Pal, Partha Pakray and Sudip Kumar Naskar

12:15-12:30 *Improving the precision of automatically constructed human-oriented translation dictionaries*
Alexandra Antonova and Alexey Misyurev

12:45-14:00 **Lunch Break**

14:00-14:45 Session 3

14:00-14:45 Invited Talk: *Adventures in Multilingual Parsing*
Joakim Nivre

15:00-15:30 Industry Session: Added value of hybrid methods in Machine Translation from a commercial perspective - Part 1

15:00-15:30 Maxim Khalilov, bmmt GmbH
Machine translation for LSPs: strategy and implementation

15:30-16:00 Coffee Break with Poster Session

A Principled Approach to Context-Aware Machine Translation
Rafael E. Banchs

Deriving de/het gender classification for Dutch nouns for rule-based MT generation tasks
Bogdan Babych, Jonathan Geiger, Mireia Ginestí Rosell and Kurt Eberle

Chinese-to-Spanish rule-based machine translation system
Jordi Centelles and Marta R. Costa-jussà

Extracting Multiword Translations from Aligned Comparable Documents
Reinhard Rapp and Serge Sharoff

16:00-18:00 Industry Session: Added value of hybrid methods in Machine Translation from a commercial perspective - Part 2

16:00-16:30 Adrià de Gispert, SDL Research
SDL Research: bringing research in MT from the lab to the product

16:30-17:00 Josep M. Crego, SYSTRAN
tba

17:00-17:30 Andrejy Vasiljevs, Tilde
How to overtake Google in MT quality - the Baltic case

17:30-18:00 Kurt Eberle, Lingenio GmbH
Hybrid Strategies for better products and shorter time-to-market

18:00-18:15 Concluding Remarks and Discussion