Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation (SedMT 2016)

The 2nd Workshop on Semantics-Driven Machine Translation

Deyi Xiong, Kevin Duh, Eneko Agirre,
Nora Aranberri and Houfeng Wang (editors)

June 16, 2016
San Diego, California, 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 978-1-941643-97-6

Preface

We are very pleased to welcome you to the 2nd Workshop on Semantics-Driven Machine Translation (SedMT) in conjunction with NAACL, held on June 16, 2016 in San Diego, California, USA.

It is widely recognized that machine translation inherently requires semantics to obtain meaning representations of source sentences and to generate meaning-preserving target translations. Recent years have witnessed a resurgent huge interest in exploring semantics for machine translation, e.g., employing lexical semantics for word sense and semantic role disambiguation in machine translation, using compositional semantics for phrasal translation, incorporating discourse semantics into document-level machine translation and so on. The emerging neural machine translation (NMT) is also naturally born as semantic machine translation since it heavily relies on distributional semantic representations.

This workshop seeks to build on the success of its precursor S2MT 2015 (http://hlt.suda.edu.cn/workshop/s2mt/index.html), which was held in conjunction with ACL 2015 in Beijing. S2MT 2015 brought together a large number of researchers from the machine translation and semantics community. Its program included high-quality papers examining and exploring semantics in machine translation from different angles and perspectives. It also featured 4 keynote speeches covering topics that cross boundaries of semantics and machine translation, as well as a thought-provoking panel discussion on the gaps and challenges between semantics and statistical machine translation.

This workshop will continue efforts of promoting the shift of interest from syntax to semantics in machine translation, exploring new horizons and cultivating ideas of cutting-edge models and algorithms for semantic machine translation.

After a rigorous selection, we only accepted 2 high-quality papers to be presented in the workshop program. These two papers study language constructions from a semantic perspective. In particular, Alastair Butler investigates the generation of several typical language constructions of English for machine translation with meaning representation informed decisions. Shili Ge and Rou Song discover 5 different construction patterns for clauses in Chinese-English translation.

In addition to the accepted papers, we also accepted two extended abstracts. Extended abstracts are 2-3 pages long. Since this is the first time to include extended abstracts in the workshop program, we invite contributions from team members of semantic MT projects and semantic MT researchers worldwide. Liangyou Li, Andy Way and Qun Liu incorporate dependency semantics into graph-based machine translation. Kiril Simov, Petya Osenova and Alex Popov examine hybrid MT system with word sense annotations on the source side for Bulgarian-English Translation.

Following the tradition of the former workshop, we invited 4 distinguished keynote speakers from semantics and machine translation to cover topics that cross boundaries of these two areas. Johan Bos (University of Groningen) will give a talk on building a large parallel meaning bank for four languages, i.e., English, Dutch, German and Italian. Jan Hajic (Charles University in Prague) will give a speech on the Chimera-TectoMT architecture of machine translation with deep linguistic analysis. Kyunghyun Cho (New York University) will give a talk on the latest research on neural machine translation without explicit linguistic structures. Martha Palmer (University of Colorado) will give a talk on designing

abstract meaning representations for machine translation. The former two keynote speeches will be in the morning session while the later two talks in the afternoon session.

The workshop will organize a panel discussion on "What semantic phenomena/annotations/forms are most promising for traditional SMT or NMT?" at the end of the program.

We would like to thank the whole Program Committee, the invited keynote and panel speakers and all authors who submitted papers to the workshop. We acknowledge the general support from our sponsors QTLeap project and the National Science Foundation of China and Jiangsu Province (grants No. 61403269 and BK20140355).

Organizers of the SedMT workshop
Deyi Xiong, Kevin Duh, Eneko Agirre, Nora Aranberri and Houfeng Wang

Organizers:

Deyi Xiong, Soochow University
Kevin Duh, Johns Hopkins University Human Language Technology Center of Excellence
Eneko Agirre, University of the Basque Country
Nora Aranberri, University of the Basque Country
Houfeng Wang, Peking University

Program Committee:

Rafael E. Banchs, Institute for Infocomm Research
Johan Bos, University of Groningen
Boxing Chen, National Research Council Canada
David Chiang, University of Notre Dame
Michael Goodman, Nanyang Technological University
Jan Hajic, Charles University in Prague
Zhongjun He, Baidu
Shujian Huang, Nanjing University
Kevin Knight, ISI
Philipp Koehn, Johns Hopkins University
Qun Liu, Dublin City University
Yang Liu, Tsinghua University
Chi-kiu Lo, National Research Council Canada
Wei Lu, Singapore University of Technology and Design
Zhengdong Lu, Noahs Ark Lab, Huawei Technologies
Minh-Thang Luong, Stanford University
Preslav Nakov, Qatar Computing Research Institute
Hwee-Tou Ng, National University of Singapore
Martha Palmer, University of Colorado
Lane Schwartz, University of Illinois
Khalil Sima'an, University of Amsterdam
Jinsong Su, Xiamen University
Hans Uszkoreit, Saarland University
Tong Xiao , Northeastern University
Frances Yung, Nara Institute of Science and Technology
Dongdong Zhang, Microsoft Research Asia
Jiajun Zhang, Institute of Automation, Chinese Academy of Sciences
Yue Zhang, Singapore University of Technology and Design

Invited Speakers:

Johan Bos, University of Groningen
Kyunghyun Cho, New York University
Jan Hajic, Charles University in Prague
Martha Palmer, University of Colorado

Panelists:

Johan Bos, University of Groningen
Kyunghyun Cho, New York University
Jan Hajic, Charles University in Prague
Martha Palmer, University of Colorado

Sponsors:

QTLeap
National Science Foundation of China (grant No. 61403269)
National Science Foundation of Jiangsu Province (grant No. BK20140355)

Proceedings Editor:

Xing Wang, Soochow University

Keynote Speech (I)
Building a Large Parallel Meaning Bank

Johan Bos
University of Groningen
johan.bos@rug.nl

Abstract

Translating from one language into another is a complex task and *meaning*, undoubtly, plays a crucial role (Langeveld, 1986). As anyone knows, there are good and bad translations. A good, faithful translation implies that all nuances of meaning have been preserved in the process of translating from one language to another. Bad translations show unwanted changes in meaning, and are often obtained by "blind" word-by-word substitutions, known to be not necessarily meaning preserving. For instance, literally translating "having a chat" into the Dutch phrase "een praatje hebben" changes the meaning and is therefore considered a bad translation. Paradoxically, often slight differences in meaning *are* allowed in translations: "voetballer" could be translated as "player" in football contexts.

As a matter of fact, meaning-preserving translations are not always possible, because there are 'lexical gaps' between languages, often caused by cultural differences. For example, in English there is a distinction between *cousin* and *nephew*, in Dutch it is just *neef*. There is no single Dutch word for *siblings*, one would say *broers en zussen*. Italian *sentire* generalizes over both the English *to hear*, *to feel* and *to taste*. Many national dishes or habits cannot be translated directly into other languages (e.g. Dutch *stroopwafel* is hard to catch in one Italian word; Italian *buonasera* needs to be translated as *good afternoon* when expressed at 2pm, but as *good evening* when used at 7pm). Some colour terms have no one-to-one correspondence between languages, e.g. *blue*: in Italian *blu* (dark blue), *azzurro* (light blue), or *celeste* (paler blue).

To investigate these phenomena from a semantic perspective on a massive scale one needs (a) a substantial number of translations, (b) meaning representations of these translations, and (c) being able to recognize non-literal interpretations. For a long time, since the interest in research of automated translation, this was simply not possible. But nowadays large parallel corpora exist and (very recently) substantial progress has been made in automated semantic interpretation. The aim of this project is to find out what the conditions are that determine good or bad translations, and investigate what role meaning plays in this process. The key idea is to automatically analyze large collections of translated texts (a technique that drastically improved current machine translation systems in the last decade) and assign cross-lingual formal meaning representations to the translations (Evang and Bos, 2013; Bos, 2014). We hope that this computational approach will offer us a completely new view on the relationship between meaning and translation by building a parallel meaning bank for four languages (English, Dutch, German and Italian). It could also be used to improve machine translation by developing translation quality metrics based on meaning rather than plain words.

Biography

Johan Bos is Professor of Computational Semantics at the University of Groningen. He received his doctorate from the Computational Linguistics Department at the University of the Saarland in 2001. Since then, he held post-doc positions at the University of Edinburgh, working on spoken dialogue systems, and the La Sapienza University of Rome, conducting research on automated question answering. In 2010 he moved to his current position in Groningen, leading the computational semantics group. Bos is the developer of Boxer, a state-of-the-art wide-coverage semantic parser for English, nitiator of the Groningen Meaning Bank, a large semantically-annotated corpus of texts, and inventor of Wordrobe, a game with a purpose for semantic annotation. Bos is PI of an NWO VICI-project "Lost in Translation – Found in Meaning" investigating the role of meaning in human and machine translation. He is also current president of the ACL special interest group in computational semantics (SIGSEM).

References

Johan Bos. 2014. Semantic annotation issues in parallel meaning banking. In Proceedings of the Tenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-10), pages 17–20, Reykjavik, Iceland.

Kilian Evang and Johan Bos. 2013. Using parallel corpora to bootstrap multilingual semantic parsers. In 20 Years of Bitext workshop at EMNLP 2013, Seattle, WA, USA.

Arthur Langeveld. 1986. Vertalen wat er staat. Synthese, De arbeiderspers.

Keynote Speech (II)
The Chimera-TectoMT architecture of Machine Translation with Deep Linguistic Analysis

Jan Hajic
Charles University in Prague
hajic@ufal.mff.cuni.cz

Abstract

The TectoMT system is a result of long-term development which began in the pre-statistical era at Charles University in Prague and continued to include state-of-the-art tools for POS tagging, morphological feature disambiguation, lemmatization parsing, and some aspects of semantic analysis. It follows the usual Analysis-Transfer-Generation workflow, with transfer trained on a large parallel corpus using Hidden Markov Tree Model. Generation is partly rule-based (at the syntax level) and partly statistical (at the inflection/morphology level). Chimera is a hybrid system that uses a specific combination of TectoMT and a standard Phrase-based SMT (Moses), complemented by a "Depfix" automatic post-editing system, which as a whole improves on the individual systems, as documented in the results of the recent WMT Shared tasks. The system has been originally developed for English-Czech and recently transferred to several other languages within the EU QTLeap project (qtleap.eu), where it has been successfully used in the IT domain for both question and answer translation in a Q&A context. Both the TectoMT and Chimera systems will be presented together with a discussion about language (in) dependence of such a hybrid solution.

Biography

Jan Hajic is a full professor of Computational Linguistics at the Institute of Formal and Applied Linguistics, School of Computer Science (Faculty of Mathematics and Physics) at the Charles University in Prague, Czech Republic and has an Adjunct professor appointment at the Computer Science Department at the University of Colorado in Boulder. His interests cover morphology of inflective languages, machine translation, deep language understanding, and the application of statistical methods in natural language processing in general. He has also built a number of richly annotated language resources. His work experience includes both industrial research (IBM Research, Yorktown Heights, 1991-1993) and academia (Charles University, Prague and Johns Hopkins University, Baltimore, MD, USA). He has been the PI or Co-PI of several national and international grants and projects, most notably the large Czech Grant Agency grant for the Center of Computational Linguistics (2005-2011), several EU projects on Machine Translation (EuroMatrix/Plus, Faust, QTLeap, META-NET; currently HimL, QT21 and CRACKER) and the U.S.-based large ITR project "Malach" (coordinated by the Visual History Foundation, Los Angeles, CA, USA). Currently he is head of the large national infrastructure project "LINDAT/CLARIN" (2010-2019) and a deputy director of the Institute; he is the Chairman of the Board of META-NET, European network for multilingual language technology. He has about 110 cited publications, including a book on computational inflectional morphology.

Keynote Speech (III)

Statistical Machine Translation *without* Explicit Linguistic Structures

Kyunghyun Cho
Computer Science and Data Science,
New York University
kyunghyun.cho@nyu.edu

Abstract

Different linguistic structures exist at various levels of natural language. A word is composed of a lexeme and a number of morphemes. A phrase consists of multiple words. Multiple phrases form a sentence. A paragraph is a sequence of more than one sentences, and so on. These structures, be them syntactic or semantic, facilitate our analysis of natural languages. It is however unclear whether these structures are necessary for machine translation. In this talk, I will present some of the latest research in neural machine translation, where nearly no such linguistic structure is being exploited, however, with comparable to, or often better than, many existing machine translation systems. The goal of this talk is not to strongly assert that those linguistic structures are not necessary, but to stimulate active discussion on this issue.

Biography

Kyunghyun Cho is an assistant professor of Computer Science and Data Science at New York University (NYU). Previously, he was a postdoctoral researcher at the University of Montreal under the supervision of Prof. Yoshua Bengio after obtaining a doctorate degree at Aalto University (Finland) in early 2014. Kyunghyun's main research interests include neural networks, generative models and their applications.

Keynote Speech (IV)
Designing Abstract Meaning Representations

Martha Palmer
Departments of Computer Science and Linguistics
University of Colorado
Martha.Palmer@colorado.edu

Abstract

Abstract Meaning Representations (AMRs) provide a single, graph-based semantic representation that abstracts away from the word order and syntactic structure of a sentence, resulting in a more language-neutral representation of its meaning. Current versions of AMRs capture nested predicate argument structures with PropBank-style semantic role labels, Named Entity tags, coreference, discourse relations and explicit interpretations of modality and negation. AMRs implement a simplified, standard neo-Davidsonian semantics. A word in a sentence either maps to a concept or a relation or is omitted if it is already inherent in the representation or it conveys inter-personal attitude (e.g., stance or distancing), (Banarescu et al., 2013). The basis of AMR is the PropBank lexicon of coarse-grained senses of verb, noun and adjective relations as well as the roles associated with each sense (each lexicon entry is a 'roleset') (Palmer et al., 2005). By marking the appropriate roles for each sense, this level of annotation provides information regarding *who* is doing *what* to *whom*. However, unlike PropBank, AMR provides a deeper level of representation of discourse relations, non-relational noun and prepositional phrases, quantities and time expressions (which PropBank largely leaves unanalyzed). Additionally, AMR makes a greater effort to abstract away from language-particular syntactic facts, instead attempting to generalize what can be thought of as different ways of saying the same thing. This talk will explore the differences between PropBank and AMR, the current and future plans for AMR annotation, and the potential of AMR as a basis for machine translation (Xue et al., 2014).

Biography

Martha Palmer is a Professor at the University of Colorado with joint appointments in Linguistics and Computer Science, an Institute of Cognitive Science Faculty Fellow, and the Director of CLEAR, the Computational Language and EducAtion Research Center. She also directed the 2011 Linguistics Institute in Boulder. Professor Palmer's Ph.D. is in Artificial Intelligence from the University of Edinburgh, in 1985. She is an Association of Computational Linguistics (ACL) Fellow, and has won an Outstanding Graduate Advisor 2014 Award and a Boulder Faculty Assembly 2010 Research Award. Her research is focused on capturing elements of the meanings of words that can comprise automatic representations of complex sentences and documents. Supervised machine learning techniques rely on vast amounts of annotated training data so she and her students are engaged in providing data with semantic annotation for English, Chinese, Arabic, Hindi, and Urdu, funded by DARPA and NSF, both manually and automatically. A more recent focus is the application of these methods to biomedical journal articles and

clinical notes, funded by NIH, and also to the Geosciences (ClearEarth), funded by NSF. She co-edits LiLT, Linguistic Issues in Language Technology, and has been a co-editor of the Journal of Natural Language Engineering and on the CLJ Editorial Board. She is a past President of ACL, and past Chair of SIGLEX and of SIGHAN.

Acknowledgments

The content of this talk represents the fruit of a large, long-term collaborative effort that has involved valuable input from all of the co-authors of the papers cited, but which has especially benefited from the thoughts and guidance of Kevin Knight, Daniel Marcu, Ulf Hermjakob, Claire Bonial, Tim O'Gorman, Kira Griffit, Nathan Schneider, Nianwen Xue, Jan Hajic and Zdenka Uresova. We also gratefully acknowledge the support of the National Science Foundation CISE-IISRI-0910992, Richer Representations for Machine Translation, and DARPA DEFT - FA-8750-13-2-0045 (a subcontract from LDC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178–186. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1):71–106.

Nianwen Xue, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of english amrs to chinese and czech. In Proceedings of the Sixth International Conference on Language Resources and Evaluation(LREC).

Table of Contents

Workshop Program

Thursday, June 16, 2016

8:45–9:00 ***Opening Remarks***

9:00–10:30 **Session 1 (Chair: Yvette Graham)**

9:00–10:00 *Keynote Speech (I)*
Building a Large Parallel Meaning Bank
Johan Bos (University of Groningen)

10:00–10:15 *Deterministic natural language generation from meaning representations for machine translation*
Alastair Butler

10:15–10:30 *Extending Phrase-Based Translation with Dependencies by Using Graphs*
Liangyou Li, Andy Way and Qun Liu

10:30–11:00 ***Coffee Break***

11:00–12:00 **Session 2 (Chair: Nora Aranberri)**

11:00–12:00 *Keynote Speech (II)*
The Chimera-TectoMT architecture of Machine Translation with Deep Linguistic Analysis
Jan Hajic (Charles University in Prague)

12:00–14:00 ***Lunch***

Thursday, June 16, 2016 (continued)

14:00–15:30 Session 3 (Chair: Andrew Finch)

14:00–15:00 *Keynote Speech (III)*
Statistical Machine Translation without Explicit Linguistic Structures
Kyunghyun Cho (New York University)

15:00–15:15 *The Naming Sharing Structure and its Cognitive Meaning in Chinese and English*
Shili Ge and Rou Song

15:15–15:30 *Towards Semantic-based Hybrid Machine Translation between Bulgarian and English*
Kiril Simov, Petya Osenova and Alexander Popov

15:30–16:00 ***Coffee Break***

16:00–17:45 Session 4 (Chair: Nora Aranberri)

16:00–17:00 *Keynote Speech (IV)*
Designing Abstract Meaning Representations
Martha Palmer (University of Colorado)

17:00–17:45 *Panel*
What semantic phenomena/annotations/forms are most promising for traditional SMT or NMT?
Panelists: Johan Bos, Kyunghyun Cho, Jan Hajic, Martha Palmer

17:45 ***Closing***