Proceedings of S^2MT 2015

The 1st Workshop on Semantics-Driven Statistical Machine Translation

Deyi Xiong, Kevin Duh,
Christian Hardmeier and Roberto Navigli (editors)

ACL-IJCNLP 2015 Workshop
July 30, 2015
Beijing, China

Order 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 978-1-941643-61-7

Preface

We are very pleased to welcome you to the 1st Workshop on Semantics-Driven Statistical Machine Translation (S^2MT) in conjunction with ACL, held on July 30, 2015 at Beijing, China.

Over the last two decades, statistical machine translation (SMT) has made a substantial progress from word-based to phrase and syntax-based SMT. Recently the progress curve has reached a stage where translation quality increases more slowly even if we use sophisticated syntactic forest-based models for translation. On the other hand, crucial meaning errors, such as incorrect translations of word senses and semantic roles, are still pervasive in SMT-generated translation hypotheses. These errors sometimes make the meanings of target translations significantly drift from the original meanings of source sentences. With an eye on the current dilemma of SMT, one might ask questions: Does SMT reach the maturity stage of its lifespan? Or is it time for us to find a new direction for SMT in order to catalyse next breakthroughs?

Semantics-driven SMT may be one of these breaking points. Semantics at different levels may enable SMT to generate not only grammatical but also meaning-preserving translations. Lexical semantics provides useful information for sense and semantic role disambiguation during translation. Compositional semantics allows SMT to generate target phrase and sentence translations by means of semantic composition. Discourse semantics captures inter-sentence dependencies for document-level machine translation. Large-scale semantic knowledge bases such as WordNet, YAGO and BabelNet, can provide external semantic knowledge for SMT. Semantics-driven SMT allows us to gradually shift from syntax to semantics and offers insights on how meaning is correctly conveyed during translation.

The goals of this workshop are to identify key challenges of exploring semantics in SMT, to discuss how semantics can help SMT and how SMT can benefit from rapid developments of semantic technologies theoretically and practically, and to find new opportunities emerging from the combination of semantics and SMT. Our key interest is to provide insights into semantics-driven SMT. Specifically, the motivations of this workshop are:

- To bring researchers in the SMT and semantics community together and to cultivate new ideas for cutting-edge models and algorithms of semantic SMT.
- To theoretically examine what semantics can provide for SMT and how SMT can benefit from semantics from a broad perspective.
- To explore new research horizons for semantics-driven SMT in practice.

We received 8 submissions from Asia, Europe and USA. After a rigorous selection, we only accepted 4 high-quality papers in the workshop program. The accepted papers examine and explore semantics in machine translation from different angles and perspectives. Alastair Butler studies the round-trip transformations between parsed sentences and meaning representations. Elior Sulem, Omri Abend and Ari Rappoport investigate semantic annotations in contrast to syntactic annotations using French-English language pair as a case study. Jinan Xu, Jiangming Liu, Yufeng Chen, Yujie Zhang, Fang Ming and Shaotong Li incorporate case frames into hierarchical phrase-based Japanese-Chinese translation. Aleš Tamchyna, Chris Quirk and Michel Galley present an abstract meaning representation to string translation model in a discriminative framework.

In addition to the accepted papers, we are very delighted to invite 4 distinguished keynote speakers from semantics and machine translation to cover topics that cross boundaries of these two areas. Percy Liang (Stanford University) and Gerard de Melo (Tsinghua University) will give talks in the morning session,

which connect semantic parsing and multilingual semantics to machine translation. Quoc V. Le (Google) will give a talk on neural language understanding in the afternoon session. Finally, António Branco (University of Lisbon) will present high-quality translation via deep language engineering approaches.

The workshop also features a panel on "Semantics and Statistical Machine Translation: Gaps and Challenges" at the end of the program. We invite Eduard Hovy, Percy Liang, Antonio Branco, Quoc V. Le and Chris Quirk as our panel speakers. Semantics-driven machine translation is an emerging and inter-disciplinary direction, which is still in its infancy. The panel discussion will shed light on the future practices and roadmap of semantics-driven machine translation research.

This is the first time that the workshop is held. The success of the workshop relies on a plenty of colleagues involved in this event. We would like to thank the whole Program Committee (30 members) for their invaluable and generous efforts on reviewing the papers this year. We are also very grateful to our invited keynote and panel speakers. Special thanks goes to Prof. Eduard Hovy who suggested the topic of the panel discussion. Additionally, we would like to thank all authors who submitted papers to the workshop. Finally, we acknowledge the general support from our sponsors NiuTrans and the National Science Foundation of China and Jiangsu Province (grants No. 61403269 and BK20140355).

Organizers of the S^2MT workshop
Deyi Xiong, Kevin Duh, Christian Hardmeier and Roberto Navigli

Organizers:

Deyi Xiong (Soochow University)
Kevin Duh (Nara Institute of Science and Technology)
Christian Hardmeier (Uppsala University)
Roberto Navigli (Sapienza University of Rome)

Program Committee:

Francis Bond (Nanyang Technological University)
Johan Bos (University of Groningen)
Ondřej Bojar (Charles University)
Rafael E. Banchs (Institute for Infocomm Research)
Boxing Chen (National Research Council Canada)
JiaJun Chen (Nanjing University)
David Chiang (University of Notre Dame)
Chris Dyer (Carnegie Mellon University)
Spence Green (Stanford University)
Kevin Knight (ISI)
Khalil Sima'an (University of Amsterdam)
Alon Lavie (Carnegie Mellon University)
Quoc V. Le (Google)
Qun Liu (Dublin City University)
Shujie Liu (Microsoft Research Asia)
Yang Liu (Tsinghua University)
Wei Lu (Singapore University of Technology and Design)
Preslav Nakov (Qatar Computing Research Institute)
Martha Palmer (University of Colorado)
Lane Schwartz (University of Illinois)
Xiaodong Shi (Xiamen University)
Linfeng Song (University of Rochester)
Jinsong Su (Xiamen University)
Frances Yung (Nara Institute of Science and Technology)
Jiajun Zhang (Institute of Automation, Chinese Academy of Sciences)
Yue Zhang (Singapore University of Technology and Design)
Tiejun Zhao (Harbin Institute of Technology)
Jinbo Zhu (Northeastern University)
Will Zou (Stanford University)

Additional Reviewers:

Yang Ding (National University of Singapore)

Keynote Speakers:

António Branco (University of Lisbon)
Quoc V. Le (Google)
Percy Liang (Stanford University)

Gerard de Melo (Tsinghua University)

Panelists:

António Branco (University of Lisbon)
Eduard Hovy (Carnegie Mellon University)
Quoc V. Le (Google)
Percy Liang (Stanford University)
Chris Quirk (Microsoft)

Sponsors:

NiuTrans (*http://www.zjyatuo.com/english/*)
National Science Foundation of China (grant No. 61403269)
National Science Foundation of Jiangsu Province (grant No. BK20140355)

Keynote Speech (I)
Semantic Parsing as, via, and for Machine Translation

Percy Liang
Stanford University
pliang@cs.stanford.edu

Abstract

Semantic parsing, the task of mapping natural language sentences to logical forms, has recently played an important role in building natural language interfaces and question answering systems. In this talk, I will present three ways in which semantic parsing relates to machine translation: First, semantic parsing can be viewed *__as__* a translation task with many of the familiar issues, e.g., divergent hierarchical structures. Second, I discuss recent work in which semantic parsing is tackled *__via__* translation (more accurately, paraphrasing) techniques, where original sentences are mapped into canonical sentences encoding the logical form. Finally, I will discuss ways in which semantic parsing could be useful *__for__* translation. Hopefully, this talk will open a deeper dialogue between the semantic parsing and machine translation communities and generate some fresh perspectives on semantics and translation.

Biography

Percy Liang is an Assistant Professor of Computer Science at Stanford University (B.S. from MIT, 2004; Ph.D. from UC Berkeley, 2011). His research interests include (i) modeling natural language semantics, (ii) developing machine learning methods that infer rich latent structures from limited supervision, (iii) and studying the tradeoff between statistical and computational efficiency. He is a 2015 Sloan Research Fellow, 2014 Microsoft Research Faculty Fellow, a 2010 Siebel Scholar, and won the best student paper at the International Conference on Machine Learning in 2008.

Keynote Speech (II)
Learning Multilingual Semantics from Big Data on the Web

Gerard de Melo
IIIS
Tsinghua University
Beijing, China
gdm@demelo.org

Abstract

Over the years, statistical machine translation has gradually shifted from surface form projections to more sophisticated syntactically and to some extent also semantically informed transformations. Still, high-quality semantic analysis of text has to date been a rather elusive goal. Fortunately, unprecedented amounts of Big Data are now readily available via the Web. While genuine semantic interpretation remains challenging, these large quantities of data enable us to develop systems that are more robust and cover a much wider range of concepts and phenomena than those of the past.

Expanding on this idea, I present a series of results on how we can develop systems that learn from Big Data in order to derive better semantic analyses, which in turn have the potential to improve machine translation. These show that it is possible to learn representations that inherit some of the benefits of language-neutral interlingua-like forms, yet preserve language-specific subtleties.

One notable example is UWN (de Melo and Weikum, 2009), a highly multilingual lexical resource allowing us to better cope with lexical gaps and generalize from observed translations. Another one is MENTA, a multilingual knowledge graph describing millions of names and words in over 200 languages in a semantic hierarchy.

The WebChild project (Tandon et al., 2014) mines large amounts of common-sense knowledge from the Web, for instance, that salad is edible and that dogs are capable of barking.

This sort of knowledge extracted from text can additionally be injected into word2vec-style distributed vector representations of words (Chen and de Melo, 2015).

Finally, efforts such as FrameBase (Rouces et al., 2015) harmonize different ways of expressing relationships both in knowledge bases and in text (Čulo and de Melo, 2012).

Biography

Gerard de Melo is a Tenure-Track Assistant Professor at Tsinghua University, Beijing, where he is heading the Web Mining and Language Technology group. He has published over 50 research papers in these areas, being awarded Best Paper awards at CIKM 2010, ICGL 2008, and the NAACL 2015 Workshop on Vector Space Modeling, as well as an ACL 2014 Best Paper Honorable Mention, a Best Student Paper Award nomination at ESWC 2015, and the WWW 2011 Best Demonstration Award, among others. Prior to joining Tsinghua, de Melo had

spent two years as a Visiting Scholar at UC Berkeley, working in ICSI's AI/FrameNet group. He received his doctoral degree at the Max Planck Institute for Informatics in Germany. He serves on the Editorial Boards of IEEE Collective Intelligence and of the Language Science Press TMNLP book series. For more information, please refer to his home page at `http://gerard.demelo.org`.

References

Jiaqiang Chen and Gerard de Melo. 2015. Semantic information extraction for improved word embeddings. In Proceedings of the NAACL Workshop on Vector Space Modeling for NLP.

Gerard de Melo and Gerhard Weikum. 2009. Towards a Universal Wordnet by learning from combined evidence. In Proceedings of CIKM 2009.

Jacobo Rouces, Gerard de Melo, and Katja Hose. 2015. Framebase: Representing n-ary relations using semantic frames. In Proceedings of ESWC 2015.

Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In Proceedings of ACM WSDM 2014.

Oliver Čulo and Gerard de Melo. 2012. Source-Path- Goal: Investigating the cross-linguistic potential of frame-semantic text analysis. it - Information Technology, 54(3).

Keynote Speech (III)
Sequence to Sequence Learning for Language Understanding

Quoc V. Le
Google
qvl@google.com

Abstract

Most language understanding problems can be formulated as a variable-length input and variable-length output prediction problem. In this talk, I will present a neural network framework to deal with this problem. Our framework makes use of recurrent networks to read in the input sequence of word vectors and predict the output sequence, one token at a time. On our benchmark with WMT'14 our method is as good as with state-of-art phrase based translation methods. I will also present results applying this method to model conversations and generate captions for images.

Biography

Quoc V. Le is one of leading scientists in Deep Learning and Artificial Intelligence, currently working at Google Brain. Quoc obtained his PhD at Stanford, undergraduate degree with First Class Honours and Distinguished Scholar at the Australian National University. He was a researcher at National ICT Australia, Microsoft Research and Max Planck Institute of Biological Cybernetics. Quoc was named one of the innovators under 35 by the MIT Tech Review.

Keynote Speech (IV)
Machine Translation and Deep Language Engineering Approaches

António Branco
University of Lisbon
antonio.branco@di.fc.ul.pt

Abstract

The deeper the processing of utterances the less language-specific differences should remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

In this talk, I will be reporting on the collective work done in the QTLeap project. This is a project that explores novel ways for attaining machine translation of higher quality that we believe are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

Biography

António Branco is the Director of the Portuguese node of the CLARIN research infrastructure. He is a professor of language science and technology at the University of Lisbon, where he was the founder and is the head of research of the Natural Language and Speech Group (NLX Group) of the Department of Informatics. He is the (co-)author of over 150 publications in the area of language science and technology and has participated and coordinated several national and international R&D projects. He was the coordinator of the European project METANET4U, integrating the R&D network of excellence META-NET. He is a member of the META-NET Executive Board and he is the first author of the White Paper on the Portuguese Language in the Digital Age.

António Branco is coordinating the QTLeap project (qtleap.eu), an European research project on quality machine translation by deep language engineering approaches.

Table of Contents

Workshop Program

Thursday, July 30, 2015

8:45–9:00 ***Opening Remarks***

9:00–10:30 **Session 1**

9:00–10:00 *Keynote Speech (I)*
Semantic Parsing as, via, and for Machine Translation
Percy Liang (Stanford University)

10:00–10:30 *Round trips with meaning stopovers*
Alastair Butler

10:30–11:00 ***Coffee Break***

11:00–12:30 **Session 2**

11:00–12:00 *Keynote Speech (II)*
Learning Multilingual Semantics from Big Data on the Web
Gerard de Melo (Tsinghua University)

12:00–12:30 *Conceptual Annotations Preserve Structure Across Translations: A French-English Case Study*
Elior Sulem, Omri Abend and Ari Rappoport

12:30–13:30 ***Lunch***

Thursday, July 30, 2015 (continued)

13:30–15:30 Session 3

13:30–14:30 *Keynote Speech (III)*
Sequence to Sequence Learning for Language Understanding
Quoc V. Le (Google)

14:30–15:00 *Integrating Case Frame into Japanese to Chinese Hierarchical Phrase-based Translation Model*
Jinan Xu, Jiangming Liu, Yufeng Chen, YUJIE ZHANG, Fang Ming and Shaotong Li

15:00–15:30 *A Discriminative Model for Semantics-to-String Translation*
Aleš Tamchyna, Chris Quirk and Michel Galley

15:30–16:00 *Coffee Break*

16:00–17:45 Session 4

16:00–17:00 *Keynote Speech (IV)*
Machine Translation and Deep Language Engineering Approaches
António Branco (University of Lisbon)

17:00–17:45 *Panel*
Semantics and Statistical Machine Translation: Gaps and Challenges
Panel Chair: Chris Quirk
Panelists: Eduard Hovy, Percy Liang, António Branco, Quoc V. Le

17:45 *Closing*