Semantics, Discourse and Statistical Machine Translation

Deyi Xiong and **Min Zhang**
Provincial Key Laboratory for Computer Information Processing Technology
Soochow University, Suzhou, China 215006
{dyxiong, minzhang}@suda.edu.cn

1 Description

In the past decade, statistical machine translation (SMT) has been advanced from word-based SMT to phrase- and syntax-based SMT. Although this advancement produces significant improvements in BLEU scores, crucial meaning errors and lack of cross-sentence connections at discourse level still hurt the quality of SMT-generated translations. More recently, we have witnessed two active movements in SMT research: one towards combining semantics and SMT in attempt to generate not only grammatical but also meaning-preserved translations, and the other towards exploring discourse knowledge for document-level machine translation in order to capture inter-sentence dependencies.

The emergence of semantic SMT are due to the combination of two factors: the necessity of semantic modeling in SMT and the renewed interest of designing models tailored to relevant NLP/SMT applications in the semantics community. The former is represented by recent numerous studies on exploring word sense disambiguation, semantic role labeling, bilingual semantic representations as well as semantic evaluation for SMT. The latter is reflected in CoNLL shared tasks, SemEval and SenEval exercises in recent years.

The need of capturing cross-sentence dependencies for document-level SMT triggers the resurgent interest of modeling translation from the perspective of discourse. Discourse phenomena, such as coherent relations, discourse topics, lexical cohesion that are beyond the scope of conventional sentence-level n-grams, have been recently considered and explored in the context of SMT.

This tutorial aims at providing a timely and combined introduction of such recent work along these two trends as discourse is inherently connected with semantics. The tutorial has three parts. The first part critically reviews the phrase- and syntax-based SMT. The second part is devoted to the lines of research oriented to semantic SMT, including a brief introduction of semantics, lexical and shallow semantics tailored to SMT, semantic representations in SMT, semantically motivated evaluation as well as advanced topics on deep semantic learning for SMT. The third part is dedicated to recent work on SMT with discourse, including a brief review on discourse studies from linguistics and computational viewpoints, discourse research from monolingual to multilingual, discourse-based SMT and a few advanced topics.

The tutorial is targeted for researchers in the SMT, semantics and discourse communities. In particular, the expected audience comes from two groups: 1) Researchers and students in the SMT community who want to design cutting-edge models and algorithms for semantic SMT with various semantic knowledge and representations, and who would like to advance SMT from sentence-by-sentence translation to document-level translation with discourse information; 2) Researchers and students from the semantics and discourse community who are interested in developing models and methods and adapting them to SMT.

2 Outline

1. SMT Overall Review (30 minutes)
   - SMT architecture
   - phrase- and syntax-based SMT
2. Semantics and SMT (1 hour and 15 minutes)
   - Brief introduction of semantics
   - Lexical semantics for SMT
   - Semantic representations in SMT
   - Semantically Motivated Evaluation
   - Advanced topics: deep semantic learning for SMT
   - Future directions

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials, pages 11–12,
Baltimore, Maryland, USA, 22 June 2014.

3. Discourse and SMT (1 hour and 15 minutes)
    - Introduction of discourse: linguistics, computational and bilingual discourse
    - Discourse-based SMT: modeling, training, decoding and evaluation
    - Future directions

3 Bios of Presenters

Dr. Deyi Xiong is a professor at Sochoow University. His research interests are in the area of natural language processing, particularly statistical machine translation and parsing. Previously he was a research scientist at the Institute for Infocomm Research of Singapore. He received the B.Sc degree from China University of Geosciences (Wuhan, China) in 2002, the Ph.D.degree from the Institute of Computing Technology (Beijing, China) in 2007, both in computer science. He has published papers in prestigious journals and conferences on statistical machine translation, including Computational Linguistics, IEEE TASLP, JAIR, NLE, ACL, EMNLP, AAAI and IJCAI. He was the program co-chair of IALP 2012 and CLIA workshop 2011.

Dr. Min Zhang, a distinguished professor and Director of the Research Center of Human Language Technology at Soochow University (China), received his Bachelor degree and Ph.D. degree in computer science from Harbin Institute of Technology in 1991 and 1997, respectively. From 1997 to 1999, he worked as a postdoctoral research fellow in Korean Advanced Institute of Science and Technology in South Korea. He began his academic and industrial career as a researcher at Lernout & Hauspie Asia Pacific (Singapore) in Sep. 1999. He joined Infotalk Technology (Singapore) as a researcher in 2001 and became a senior research manager in 2002. He joined the Institute for Infocomm Research (Singapore) as a research scientist in Dec. 2003. He joined the Soochow University as a distinguished professor in 2012.

His current research interests include machine translation, natural language processing, information extraction, social network computing and Internet intelligence. He has co-authored more than 150 papers in leading journals and conferences, and co-edited 10 books/proceedings published by Springer and IEEE. He was the recipient of several awards in China and oversea. He is the vice president of COLIPS (2011-2013), the elected vice chair of SIGHAN/ACL (2014-2015), a steering committee member of PACLIC (2011-now), an executive member of AFNLP (2013-2014) and a member of ACL (since 2006). He supervises Ph.D students at National University of Singapore, Harbin Institute of Technology and Soochow University.