DiscoMT 2015

DISCOURSE IN MACHINE TRANSLATION

Proceedings of the Workshop

17 September 2015
Lisbon, Portugal

Order print-on-demand copies from:

Curran Associates
57 Morehouse Lane
Red Hook, New York 12571 USA
Tel: +1-845-758-0400
Fax: +1-845-758-2633
curran@proceedings.com

ISBN: 978-1-941643-32-7

Preface

It is well-known that texts have properties that go beyond those of their individual sentences and that reveal themselves in the frequency and distribution of words, word senses, referential forms and syntactic structures, including:

- document-wide properties, such as style, register, reading level and genre;
- patterns of topical or functional sub-structure;
- patterns of discourse coherence, as realized through explicit and/or implicit relations between sentences, clauses or referring forms;
- anaphoric and elliptic expressions, in which speakers exploit the previous discourse context to convey subsequent information very succinctly.

By the end of the 1990s, these properties had stimulated considerable research in Machine Translation, aimed at endowing machine-translated texts with similar document and discourse properties as their source texts. A period of ten years then elapsed before interest resumed in these topics, now from the perspectives of Statistical and/or Hybrid Machine Translation. This led to the *First Workshop on Discourse in Machine Translation* (DiscoMT) in 2013, held in Sofia, Bulgaria, in connection with the annual ACL conference.

Since then, SMT has itself evolved in ways that reflect more interest in and provide more access to needed linguistic knowledge. This evolution is charted in this *Second Workshop on Discourse in Machine Translation (DiscoMT 2015)*, held in Lisbon, Portugal, in connection with EMNLP. Part of this evolution has been the growth of interest in one particular problem: the translation of pronouns whose form in the target language may be constrained in challenging ways by their context. This shared interest has created an environment in which a shared task on pronoun translation or prediction from English–to–French was able to stimulate responses from groups in China, the Czech Republic, Malta, Sweden, Switzerland, and the UK.

In addition to nine papers describing shared task submissions and an overview of the shared task, the submitted systems and the findings (Hardmeier et al., 2015), twelve submissions were accepted for presentation (five as long papers, three as short papers, and four as posters). The papers and posters span the topics of: pronoun translation between languages which differ in pronoun usage (Novák et al., 2015; Guillou and Webber, 2015); explicitation/implicitation in translating discourse connectives (Hoek et al., 2015; Yung et al., 2015); context-aware translation of ambiguous terms (Mascarell et al., 2015; Zhang and Ittycheriah, 2015); assessing document-level properties of MT output, including coherence (Sim Smith et al., 2015; Gong et al., 2015); preserving document-level properties characteristic of register, genre, and other types of text variation (Lapshinova-Koltunski and Vela, 2015; van der Wees et al., 2015; Lapshinova-Koltunski, 2015); and difficulties in preserving them in a purely alignment-based MT framework (Hardmeier, 2015). We hope that workshops such as this one will continue to stimulate work on these aspects of Discourse and Machine Translation, as well as in the many areas not yet represented.

We would like to thank all the authors who submitted papers to the workshop, as well as all the members of the Program Committee who reviewed the submissions and delivered thoughtful, informative reviews.

The Organizers
August 15, 2015

References

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal. Association for Computational Linguistics.

Liane Guillou and Bonnie Webber. 2015. Analysing ParCor and its translations by state-of-the-art SMT systems. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 24–32, Lisbon, Portugal. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Christian Hardmeier. 2015. On statistical machine translation and translation theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon, Portugal. Association for Computational Linguistics.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2015. The role of expectedness in the implicitation and explicitation of discourse relations. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 41–46, Lisbon, Portugal. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski and Mihaela Vela. 2015. Measuring 'registerness' in human and machine translation: A text classification approach. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 122–131, Lisbon, Portugal. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski. 2015. Exploration of inter- and intralingual variation of discourse phenomena. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 158–167, Lisbon, Portugal. Association for Computational Linguistics.

Laura Mascarell, Mark Fishel, and Martin Volk. 2015. Detecting document-level context triggers to resolve translation ambiguity. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 47–51, Lisbon, Portugal. Association for Computational Linguistics.

Michal Novák, Dieke Oele, and Gertjan van Noord. 2015. Comparison of coreference resolvers for deep syntax translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 17–23, Lisbon, Portugal. Association for Computational Linguistics.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2015. A proposal for a coherence corpus in machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 52–58, Lisbon, Portugal. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2015. Translation model adaptation using genre-revealing text features. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 132–141, Lisbon, Portugal. Association for Computational Linguistics.

Frances Yung, Kevin Duh, and Yuji Matsumoto. 2015. Crosslingual annotation and analysis of implicit discourse connectives for machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 142–152, Lisbon, Portugal. Association for Computational Linguistics.

Rong Zhang and Abraham Ittycheriah. 2015. Novel document level features for statistical machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 153–157, Lisbon, Portugal. Association for Computational Linguistics.

Organizing Committee

Bonnie Webber, University of Edinburgh (chair)
Marine Carpuat, University of Maryland (co-chair)
Andrei Popescu-Belis, Idiap Research Institute, Martigny (co-chair)

Mark Fishel, University of Zurich
Christian Hardmeier, Uppsala University
Lori Levin, Carnegie Mellon University
Preslav Nakov, Qatar Computing Research Institute
Ani Nenkova, University of Pennsylvania
Lucia Specia, University of Sheffield
Jörg Tiedemann, Uppsala University
Min Zhang, Soochow University

Program Committee

Beata Beigman Klebanov, Educational Testing Service
Liane Guillou, University of Edinburgh
Francisco Guzmán, Qatar Computing Research Institute
Shafiq Joty, Qatar Computing Research Institute
Thomas Meyer, Google, Zürich
Michal Novák, Charles University in Prague
Lucie Poláková, Charles University in Prague
Maja Popovic, DFKI, Berlin
Sara Stymne, Uppsala University
Yannick Versley, Heidelberg University
Marion Weller, University of Stuttgart

Shared Task Organizers

Christian Hardmeier, Uppsala University
Preslav Nakov, Qatar Computing Research Institute
Sara Stymne, Uppsala University
Jörg Tiedemann, Uppsala University
Yannick Versley, Heidelberg University

Table of Contents

Workshop Program

Thursday, September 17, 2015

09:00–10:30 Session 1

09:00–09:05 ***Introduction***

09:05–09:35 *Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation*
Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley and Mauro Cettolo

09:35–09:50 *Comparison of Coreference Resolvers for Deep Syntax Translation*
Michal Novák, Dieke Oele and Gertjan van Noord

09:50–10:15 *Analysing ParCor and its Translations by State-of-the-art SMT Systems*
Liane Guillou and Bonnie Webber

10:15–10:30 ***Poster Boaster***

10:30–11:00 ***Coffee Break***

11:00–12:30 Session 2a: Regular Track Posters

Document-Level Machine Translation Evaluation with Gist Consistency and Text Cohesion
Zhengxian Gong, Min Zhang and Guodong Zhou

The Role of Expectedness in the Implicitation and Explicitation of Discourse Relations
Jet Hoek, Jacqueline Evers-Vermeul and Ted J.M. Sanders

Detecting Document-level Context Triggers to Resolve Translation Ambiguity
Laura Mascarell, Mark Fishel and Martin Volk

A Proposal for a Coherence Corpus in Machine Translation
Karin Sim Smith, Wilker Aziz and Lucia Specia

Thursday, September 17, 2015 (continued)

11:00–12:30 Session 2b: Posters Related to Oral Presentations

On Statistical Machine Translation and Translation Theory
Christian Hardmeier

Exploration of Inter- and Intralingual Variation of Discourse Phenomena
Ekaterina Lapshinova-Koltunski

Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach
Ekaterina Lapshinova-Koltunski and Mihaela Vela

Translation Model Adaptation Using Genre-Revealing Text Features
Marlies van der Wees, Arianna Bisazza, Christof Monz

Crosslingual Annotation and Analysis of Implicit Discourse Connectives for Machine Translation
Frances Yung, Kevin Duh, Yuji Matsumoto

11:00–12:30 Session 2c: Shared Task Posters

Part-of-Speech Driven Cross-Lingual Pronoun Prediction with Feed-Forward Neural Networks
Jimmy Callin, Christian Hardmeier and Jörg Tiedemann

Automatic Post-Editing for the DiscoMT Pronoun Translation Task
Liane Guillou

A Document-Level SMT System with Integrated Pronoun Prediction
Christian Hardmeier

Predicting Pronoun Translation Using Syntactic, Morphological and Contextual Features from Parallel Data
Sharid Loáiciga

Rule-Based Pronominal Anaphora Treatment for Machine Translation
Sharid Loáiciga and Eric Wehrli

Pronoun Translation and Prediction with or without Coreference Links
Ngoc Quang Luong, Lesly Miculicich Werlen and Andrei Popescu-Belis

Thursday, September 17, 2015 (continued)

Predicting Pronouns across Languages with Continuous Word Spaces
Ngoc-Quan Pham and Lonneke van der Plas

Baseline Models for Pronoun Prediction and Pronoun-Aware Translation
Jörg Tiedemann

A Maximum Entropy Classifier for Cross-Lingual Pronoun Prediction
Dominikus Wetzel, Adam Lopez and Bonnie Webber

12:30–14:00 ***Lunch Break***

14:00–15:30 **Session 3**

14:00–14:25 *Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach*
Ekaterina Lapshinova-Koltunski and Mihaela Vela

14:25–14:50 *Translation Model Adaptation Using Genre-Revealing Text Features*
Marlies van der Wees, Arianna Bisazza and Christof Monz

14:50–15:15 *Crosslingual Annotation and Analysis of Implicit Discourse Connectives for Machine Translation*
Frances Yung, Kevin Duh and Yuji Matsumoto

15:15–15:30 *Novel Document Level Features for Statistical Machine Translation*
Rong Zhang and Abraham Ittycheriah

15:30–16:00 ***Coffee Break***

16:00–17:30 **Session 4**

16:00–16:25 *Exploration of Inter- and Intralingual Variation of Discourse Phenomena*
Ekaterina Lapshinova-Koltunski

16:25–16:40 *On Statistical Machine Translation and Translation Theory*
Christian Hardmeier

16:40–17:30 ***Final Discussions and Conclusions***