DiscoMT 2017

Discourse in Machine Translation

Proceedings of the Workshop

September 8, 2017
Copenhagen, Denmark

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 978-1-945626-87-6

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 in 2013 to the *First Workshop on Discourse in Machine Translation (DiscoMT)*, held in Sofia, Bulgaria, in conjunction to the annual ACL conference.

The evolution of Statistical MT, in ways that reflected more interest in and provided more access to needed linguistic knowledge was charted in the *Second Workshop on Discourse in Machine Translation (DiscoMT 2015)*, held in Lisbon, Portugal, in conjunction to 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 several research groups.

The shared task in pronoun prediction has been continued as one of the shared tasks of the First Conference on Machine Translation (WMT 2016), and then again at this year's *Third Workshop on Discourse in Machine Translation (DiscoMT 2017)*, held in Copenhagen, Denmark, in conjunction to EMNLP. As observed with systems presented at previous shared tasks, and confirmed by several papers at DiscoMT 2017, the neural turn in MT has started having a significant impact on discourse-level or document-level translation, with neural networks being adapted to consider wider contexts when generating translations.

We hope that workshops such as this one will continue to stimulate work on Discourse and Machine Translation, in a wide range of discourse phenomena and MT architectures.

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 Chairs
July 21, 2017

Chairs

Bonnie Webber, University of Edinburgh, UK
Andrei Popescu-Belis, Idiap Research Institute, Martigny, Switzerland
Jörg Tiedemann, University of Helsinki, Finland

Program Committee

Mauro Cettolo, Fondazione Bruno Kessler, Trento, Italy
Filip Ginter, University of Turku, Finland
Liane Guillou, Brainnwave, Edinburgh, UK
Christian Hardmeier, Uppsala University, Sweden
Shafiq Joty, Qatar Computing Research Institute, Doha, Qatar
Lori Levin, Carnegie Mellon University, Pittsburgh, PA, USA
Ekaterina Lapshinova-Koltunski, Saarland University, Germany
Ngoc-Quang Luong, Nuance Communications, Belgium
Thomas Meyer, Google, Zurich, Switzerland
Preslav Nakov, Qatar Computing Research Institute, Doha, Qatar
Michal Novak, Charles University, Prague, Czech Republic
Maja Popovic, DFKI, Berlin, Germany
Annette Rios, University of Zurich, Switzerland
Rico Sennrich, University of Edinburgh, UK
Lucia Specia, University of Sheffield, UK
Sara Stymne, Uppsala University, Sweden
Yannick Versley, LangTec, Hamburg, Germany
Martin Volk, University of Zurich, Switzerland
Min Zhang, Soochow University, Suzhou, China
Sandrine Zufferey, University of Bern, Switzerland

Shared Task Organizers

Sharid Loáiciga Sánchez, Uppsala University, Sweden, *coordinator*
Christian Hardmeier, Uppsala University, Sweden
Preslav Nakov, Qatar Computing Research Institute, Doha, Qatar
Sara Stymne, Uppsala University, Sweden
Jörg Tiedemann, University of Helsinki, Finland
Yannick Versley, LangTec, Hamburg, Germany

Table of Contents

Conference Program

Friday, September 8, 2017

09:00–10:30 Session 1

09:00–09:10 ***Introduction***

09:10–09:40 *Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction*
Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo and Yannick Versley

09:40–10:10 *Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT)*
Lesly Miculicich Werlen and Andrei Popescu-Belis

10:10–10:30 ***Poster Boaster***

10:30–11:00 ***Coffee Break***

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

Using a Graph-based Coherence Model in Document-Level Machine Translation
Leo Born, Mohsen Mesgar and Michael Strube

Treatment of Markup in Statistical Machine Translation
Mathias Müller

Friday, September 8, 2017 (continued)

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

A BiLSTM-based System for Cross-lingual Pronoun Prediction
Sara Stymne, Sharid Loáiciga and Fabienne Cap

Neural Machine Translation for Cross-Lingual Pronoun Prediction
Sébastien Jean, Stanislas Lauly, Orhan Firat and Kyunghyun Cho

Predicting Pronouns with a Convolutional Network and an N-gram Model
Christian Hardmeier

Cross-Lingual Pronoun Prediction with Deep Recurrent Neural Networks v2.0
Juhani Luotolahti, Jenna Kanerva and Filip Ginter

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

Combining the output of two coreference resolution systems for two source languages to improve annotation projection
Yulia Grishina

Discovery of Discourse-Related Language Contrasts through Alignment Discrepancies in English-German Translation
Ekaterina Lapshinova-Koltunski and Christian Hardmeier

Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction
Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo and Yannick Versley

Neural Machine Translation with Extended Context
Jörg Tiedemann and Yves Scherrer

Translating Implicit Discourse Connectives Based on Cross-lingual Annotation and Alignment
Hongzheng Li, Philippe Langlais and Yaohong Jin

Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT)
Lesly Miculicich Werlen and Andrei Popescu-Belis

12:30–14:00 ***Lunch Break***

Friday, September 8, 2017 (continued)

14:00–15:30 Session 3

14:00–14:30 *Neural Machine Translation with Extended Context*
Jörg Tiedemann and Yves Scherrer

14:30–14:50 *Discovery of Discourse-Related Language Contrasts through Alignment Discrepancies in English-German Translation*
Ekaterina Lapshinova-Koltunski and Christian Hardmeier

14:50–15:10 *Translating Implicit Discourse Connectives Based on Cross-lingual Annotation and Alignment*
Hongzheng Li, Philippe Langlais and Yaohong Jin

15:10–15:50 *Combining the output of two coreference resolution systems for two source languages to improve annotation projection*
Yulia Grishina

15:30–16:00 ***Coffee Break***

16:00–17:30 Session 4

16:00–16:30 *Lexical Chains meet Word Embeddings in Document-level Statistical Machine Translation*
Laura Mascarell

16:30–16:50 *On Integrating Discourse in Machine Translation*
Karin Sim Smith

16:50–17:30 ***Final Discussion and Conclusion***