DiscoMT 2013

Discourse in Machine Translation

Proceedings of the Workshop

August 9, 2013
Sofia, Bulgaria

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53704 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 978-1-937284-68-8

Introduction

It is a truism that texts have properties that go beyond those of their individual sentences, including:

- document-wide properties, such as topic mix, style, register, reading level and genre, all of which are manifest in the frequency and distribution of words, word senses, referential forms and syntactic structures;
- patterns of topical or functional sub-structure that show up in localized differences in the frequency and distribution of these elements within documents;
- patterns of discourse coherence, manifest in explicit and implicit relations between sentences (clauses), or between sentences (clauses) and referring forms, or between referring forms themselves;
- common use of reduced expressions that rely on context to convey a lot of information in very few words.

These properties stimulated a good deal of Machine Translation research in the 1990s, aimed at endowing machine–translated target texts with the same document and discourse properties as their source texts, albeit realized differently in source and target languages. This included work on stylistics for Machine Translation (DiMarco & Mah 1994), target language realization of source-language discourse relations (Mitkov 1993) and of referring forms (Bond & Ogura 1998; More et al. 1999; Wada 1990), anaphora resolution for generating appropriate target-language pronouns (Chan and T'sou 1999; Ferrández et al. 1999; Nakaiwa & Ikehara 1992; Nakaiwa 1999), and ellipsis resolution for generating appropriate target-language forms from ellipsed verb-phrases (Balkan 1998). Pointers to much of this work can be found in the *Machine Translation Archive* of conference and workshop papers from the 1990s (see www.mt-archive.info/srch/ling-90.htm).

This early period essentially ended with the 1999 publication of a special issue of the journal *Machine Translation*, edited by Ruslan Mitkov, devoted to anaphora resolution in Machine Translation and multilingual NLP. Only in the past 3–4 years has there been renewed interest in these topics, now from the perspectives of Statistical Machine Translation and Hybrid Machine Translation (Chung & Gildea 2010; Eidelman et al. 2012; Foster et al. 2012; Gong et al. 2011; Guillou 2012; Hardmeier & Federico 2010; Hardmeier et al. 2012; Le Nagard & Koehn 2010; Meyer 2012; Meyer et al. 2012; Voigt & Jurafsky 2012).

With this renewed interest, this ACL Workshop on Discourse in Machine Translation provides a timely forum for the presentation of new approaches to enabling modern systems to produce texts that are not merely sequences of isolated sentences.

Eight submissions have been accepted for the Workshop, on topics that range from multilingual modeling of discourse for machine translation, to actual use of discourse-level features to improve machine translation. From the modeling perspective, the papers presented at the Workshop discuss discourse phenomena such as lexical consistency (Guillou, this volume), lexical cohesion (Beigman Klebanov & Flor, this volume) and implicit connectives (Meyer & Webber, this volume), and "meaning units" with cognitive relevance (Williams et al., this volume). From the perspective of the application to MT, several papers present encouraging results showing that discourse-related features bring measurable improvements to the quality of machine-translated texts. One study uses oracle features, namely connective labels (Meyer & Poláková, this volume), while others use automatically-assigned ones. For instance, the translation of tensed verbs is improved by recognizing whether or not they are conveying

narrative material (Meyer et al., this volume); the translation of the pronoun "it" is improved based on lexical, syntactic and anaphoric features (Novák et al., this volume); and a document-level decoder is used when tuning an SMT system, with a sample of readability-related features (Stymne et al., this volume).

The studies presented at the Workshop provide quantitative data and benchmark scores to which future progress on these tasks should be compared. We hope that the Workshop will stimulate further work in these areas, as well as in the many areas of discourse and Machine Translation that are 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.

Bonnie Webber (chair), Katja Markert, Andrei Popescu-Belis, Jörg Tiedemann (co-chairs)

References

Lorna Balkan (1998). *A Treatment of Verb Phrase Ellipsis for Machine Translation*. PhD thesis, University of Essex.

Beata Beigman Klebanov and Michael Flor (2013). Associative Texture Is Lost In Translation. *This volume*.

Francis Bond and Kentaro Ogura (1998). Reference in Japanese-English Machine Translation. *Machine Translation*, 13:2-3, pp. 107–134.

Marine Carpuat and Michel Simard (2012). The Trouble with SMT Consistency. *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, pp. 442–449.

Samuel Chan and Benjamin T'sou (1999). Semantic Inference for Anaphora Resolution: Toward a Framework in Machine Translation. *Machine Translation*, 14:3-4, pp. 163–190.

Tagyoung Chung and Dan Gildea (2010). Effects of Empty Categories on Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 636–645.

Chrysanne DiMarco and Keith Mah (1994). A Model of Comparative Stylistics for Machine Translation. *Machine Translation*, 9:1, pp. 21–59.

Vladimir Eidelman, Jordan Boyd-Graber and Philip Resnik (2012). Topic Models for Dynamic Translation Model Adaptation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 115–119.

Antonio Ferrández, Manuel Palomar and Lidia Moreno (1999). An Empirical Approach to Spanish Anaphora Resolution. *Machine Translation*, 14:3-4, pp. 191-216.

George Foster, Pierre Isabelle and Roland Kuhn (2010). Translating Structured Documents. *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Anita Gojun (2010). *Null Subjects in Statistical Machine Translation: A Case Study on Aligning English and Italian Verb Phrases with Pronominal subjects*. Diplomarbeit, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart.

Zhengxian Gong, Min Zhang and Guodong Zhou (2011). Cache-based Document-level Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 909–919.

Liane Guillou (2012). Improving Pronoun Translation for Statistical Machine Translation. *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the ACL*, pp. 1–10.

Liane Guillou (2013). Analysing Lexical Consistency in Translation. *This volume*.

Christian Hardmeier and Marcello Federico (2010). Modeling Pronominal Anaphora in Statistical Machine Translation. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 283–289.

Christian Hardmeier, Joakim Nivre and Jörg Tiedemann (2012). Document-wide Decoding for Phrase-based Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1179–1190.

Ronan Le Nagard and Philipp Koehn (2010). Aiding Pronoun Translation with Co-reference Resolution. *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pp. 252–261.

Thomas Meyer (2011). Disambiguating Temporal-contrastive Connectives for Machine Translation. *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the ACL*, pp. 46–51.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui and Andrea Gesmundo (2012). Machine Translation of Labeled Discourse Connectives. *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Thomas Meyer, Cristina Grisot and Andrei Popescu-Belis (2013). Detecting Narrativity to Improve English to French Translation of Simple Past Verbs. *This volume*.

Thomas Meyer and Lucie Poláková (2013). Machine Translation with Many Manually Labeled Discourse Connectives. *This volume*.

Thomas Meyer and Bonnie Webber (2013). Implicitation of Discourse Connectives in (Machine) Translation. *This volume*.

Ruslan Mitkov (1993). How Could Rhetorical Relations Be Used in Machine Translation? (And at least Two Open Questions). *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*, pp.86–89.

Ruslan Mitkov and Johann Haller (1994). Machine Translation, Ten Years on: Discourse Has Yet to Make a Breakthrough. *Proceedings of the International Conference on Machine Translation: Ten Years on*, Cranfield University, England.

Tatsunori Mori, Mamoru Matsuo and Hiroshi Nakawaga (1999). Zero-subject Resolution Using Linguistic Constraints and Defaults: The Case of Japanese Instruction Manuals. *Machine Translation*, 14:3-4, pp. 231-245.

Hiromi Nakaiwa (1999). Automatic Extraction of Rules for Anaphora Resolution of Japanese Zero Pronouns in Japanese-English Machine Translation from Aligned Sentence Pairs. *Machine Translation*, 14:3-4, pp. 247–279.

Hiromi Nakaiwa and Satoru Ikehara (1992). Zero Pronoun Resolution in a Japanese to English Machine Translation System by Using Verbal Semantic Attributes. *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP)*, pp. 201–208.

Michal Novák, Anna Nedoluzhko and Zdenek Zabokrtsky (2013). Translation of "It" in a Deep Syntax Framework. *This volume*.

Sara Stymne, Christian Hardmeier, Jörg Tiedemann and Joakim Nivre (2013). Feature Weight Optimization for Discourse-Level SMT. *This volume*.

Ferhan Ture, Douglas Oard and Philip Resnik (2012). Encouraging Consistent Translation Choices. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 417–426.

Rob Voigt and Dan Jurafsky (2012). Towards a Literary Machine Translation: The Role of Referential Cohesion. *Proceedings of the NAACL-HLT Workshop on Computational Linguistics for Literature*, pp. 18–25.

Hajime Wada (1990). Discourse Processing in MT: Problems in Pronominal Translation. *Proceedings of the 13th International Conference on Computational Linguistics (Coling)*, pp. 73–75.

Jennifer Williams, Rafael Banchs and Haizhou Li (2013). Meaning Unit Segmentation in English and Chinese: a New Approach to Discourse Phenomena. *This volume*.

Organizing Committee

Bonnie Webber, University of Edinburgh (chair)
Ondřej Bojar, Charles University in Prague
Chris Callison-Burch, Johns Hopkins University
Marcello Federico, FBK-IRST, Trento
Pierre Isabelle, National Research Council Canada
Katja Markert, University of Leeds (co-chair)
Andrei Popescu-Belis, Idiap Research Institute, Martigny (co-chair)
Jörg Tiedemann, University of Uppsala (co-chair)

Program Committee

Trevor Cohn, University of Sheffield
George Foster, National Research Council Canada
Dan Gildea, University of Rochester
Liane Guillou, University of Edinburgh
Christian Hardmeier, University of Uppsala
Hitoshi Isahara, Toyohashi University of Technology
Philipp Koehn, University of Edinburgh
Thomas Meyer, Idiap Research Institute, Martigny
Hwee Tou Ng, National University of Singapore
Michal Novák, Charles University in Prague
Maja Popovic, DFKI GmbH, Berlin
Jean Senellart, SYSTRAN, Paris
Lucia Specia, University of Sheffield
Sara Stymne, University of Uppsala
Gregor Thurmair, Linguatec GmbH, Munich
Min Zhang, Institute for Infocomm Research, A*STAR, Singapore
Sandrine Zufferey, Utrecht University

Table of Contents

Conference Program

Friday August 9, 2013

9:00 Introduction by the organizers

9:10 First oral presentation session

9:10 *Meaning Unit Segmentation in English and Chinese: a New Approach to Discourse Phenomena*
Jennifer Williams, Rafael Banchs and Haizhou Li

9:30 *Analysing Lexical Consistency in Translation*
Liane Guillou

9:50 *Implicitation of Discourse Connectives in (Machine) Translation*
Thomas Meyer and Bonnie Webber

10:10 *Associative Texture Is Lost In Translation*
Beata Beigman Klebanov and Michael Flor

10:30 Coffee break

11:00 Poster session, jointly with WMT

In addition to posters from the speakers, posters will also be presented for the following papers:

Detecting Narrativity to Improve English to French Translation of Simple Past Verbs
Thomas Meyer, Cristina Grisot and Andrei Popescu-Belis

Machine Translation with Many Manually Labeled Discourse Connectives
Thomas Meyer and Lucie Poláková

12:30 Lunch break

14:00 Second oral presentation session

14:00 *Translation of "It" in a Deep Syntax Framework*
Michal Novák, Anna Nedoluzhko and Zdeněk Žabokrtský

14:20 *Feature Weight Optimization for Discourse-Level SMT*
Sara Stymne, Christian Hardmeier, Jörg Tiedemann and Joakim Nivre

14:40 Closing discussion

15:30 Coffee break, end of the workshop