

Proposal for the Third Workshop on Discourse in Machine Translation (DiscoMT 2017)

1 Description and justification

When translating entire texts, as opposed to individual sentences in isolation, one cannot ignore text-level properties such as:

- 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 across a document as a whole;
- patterns of topical or functional sub-structure that show up in local differences in frequency and distribution;
- patterns of discourse coherence, manifest in explicit and implicit relations between sentences or clauses, or between sentences or clauses and referring forms, or between referring forms themselves (which include noun phrases and pronouns);
- use of reduced context-sensitive expressions such as ellipsed verb phrases (e.g., “I do too.”) that exploit context to convey what is intended.

In 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. This included work on stylistics for MT, target language realization of source-language discourse relations and of referring forms, anaphora resolution for generating appropriate target-language pronouns, and ellipsis resolution for generating appropriate target-language forms from ellipsed verb-phrases. Pointers to much of this work can be found in the *Machine Translation Archive* of conference and workshop papers from the 1990s (<http://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 multi-lingual NLP. A period of ten years then elapsed before interest resumed in these topics, now from the perspectives of Statistical and/or Hybrid Machine Translation.

Since 2010, there has been considerably more work on discourse and machine translation, by an increasing number of people and groups. Moreover, SMT has itself evolved in ways that allow more access to needed linguistic knowledge, through the feature-rich statistical models that have recently become available. The branch of *document-level machine translation* is now well established, thanks in particular to the first two editions of the DiscoMT (in 2013 and 2015), as well as a 2016 Shared Task within the Workshop on Machine Translation (WMT).

Considerable effort has been expended recently on document-level MT, such that now, several individuals and/or groups are working on similar or overlapping problems. For example, papers presented at DiscoMT 2015 demonstrate active research on: pronoun translation between languages which differ in pronoun usage; explicitation/implication in translating discourse connectives; context-aware translation of ambiguous terms; assessing document-level properties of MT output, including coherence; preserving document-level properties characteristic of register, genre, and other types of text variation; and difficulties in preserving them in a purely alignment-based MT framework.

A critical mass of researchers for the DiscoMT 2015 and the WMT 2016 shared tasks on pronoun-focused translation has been reached. The advent of neural machine translation in the past two years has

opened new opportunities for taking account of larger contexts with deep neural networks, and we expect to see a number of results presented and discussed at DiscoMT 2017.

2 Previous editions of DiscoMT: 2013 and 2015

The growing interest in discourse and MT led to the first *ACL Workshop on Discourse in Machine Translation (DiscoMT)* in 2013, held in Sofia, Bulgaria.¹ The workshop featured eight papers on such topics as lexical consistency in human and machine translation, lexical tightness as a whole-document measure of machine-translated targets, improving the translation of tensed verbs by recognizing whether or not they are conveying narrative, and document-level decoding. For the first edition, there were 12 submissions, 6 of which were accepted as oral presentations and 2 as posters only. Oral presenters were also encouraged to make posters describing their work, and all resulting posters were presented in joint poster session with WMT 2013, thus offering DiscoMT an increased visibility.

To allow for sufficient novelty and volume of research contributions, the organizers decided to wait two years before holding a second DiscoMT workshop. This second *Workshop on Discourse in Machine Translation (DiscoMT 2015)* was held in Lisbon, in conjunction with EMNLP.² DiscoMT 2015 saw an increase in the number of both submitted and accepted papers, while also hosting a shared task and maintaining a joint poster session with WMT. The workshop volume also contained an extended overview of the two sub-tasks within the shared task.

3 Shared tasks: DiscoMT 2015, WMT 2016, and DiscoMT 2017

Pronoun translation is a problem that has been independently approached by different researchers in the last few years, which led to the organization of two shared tasks, in 2015 and 2016. Incorrect pronoun translations can have a large impact on the readability of a text, and the overuse of masculine default pronouns in MT output is not only a linguistic, but also a gender-political issue. Pronoun use is governed by strong grammatical rules, so it is relatively straightforward to evaluate system performance objectively. Finally, the importance of the problem is easy to demonstrate, even to an audience that is ignorant of discourse.

The DiscoMT 2015 shared task³ included two subtasks, relevant to both the machine translation and the discourse communities: (i) pronoun-focused translation, a practical MT task, and (ii) cross-lingual pronoun prediction, a classification task that required no specific MT expertise and is interesting as a machine learning task in its own right. The task focused on EN/FR, for which MT output is generally of high quality, but has visible issues with pronoun translation due to differences in the pronoun systems of the two languages. Six groups participated in the pronoun-focused translation task and eight groups in the cross-lingual pronoun prediction task.

¹Bonnie Webber and Andrei Popescu-Belis and Katja Markert and Jörg Tiedemann. (2013). *Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT)*. Sofia, Bulgaria. Association for Computational Linguistics. <http://www.aclweb.org/anthology/W13-33>.

²Webber, B., Carpuat, M., Popescu-Belis, A., & Hardmeier, C. (2015) *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT 2015)*. Lisbon, Portugal. Association for Computational Linguistics. <https://aclweb.org/anthology/W/W15/#2500>

³Christian 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. *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT)*, pages 1–16, Lisbon, Portugal, 17 September 2015. Association for Computational Linguistics.

The WMT 2016 shared task on cross-lingual pronoun prediction⁴ was a classification task in which participants were asked to provide predictions on what pronoun class label should replace a placeholder value in the target-language text, provided in lemmatised and PoS-tagged form. Four subtasks were offered, for EN/FR and EN/DE language pairs, in both directions. Eleven teams participated in the shared task. Most of the submissions outperformed two strong language-model-based baseline systems, with systems using deep recurrent neural networks outperforming those using other architectures for most language pairs.

We foresee the organization of a shared task on pronoun translation at DiscoMT 2017. Minutes have been taken after the concluding discussion at the WMT 2016 shared task, and plans have been made for a 2017 shared task. The idea is to maintain a sub-task on EN/FR/DE pronoun prediction (given the source text and a lemmatized (or “impoverished”) version of the target text stripped of its pronouns), but to consider again the possibility of a pronoun-focused translation task, while attempting to solve the evaluation issues (trade-off between cost and accuracy) related to it.

4 Timeline

We are proposing to co-locate DiscoMT 2017 with EMNLP, and thus to follow the same timeline. As in the past years, an appropriate schedule will be defined for shared task submissions. As the time needed to prepare and check the datasets and adapt the evaluation tools is significant, we prefer to run the workshop in connection with a conference in the early or mid-Fall.

5 Organizational Information

The proposed workshop has a small set of **Co-Chairs**, with responsibility for scientific and administrative matters involved in delivering a workshop. The members of the **Programme Committee** will mainly be responsible for reviewing submissions, but they will also be consulted for matters related to the organization and intellectual content of the workshop. A separate group will have responsibility for the Shared Task, and liaise with the Co-Chairs regarding in particular the quality of the system papers. The responsible person for the shared task will be a co-editor of the proceedings volume as well.

Co-Chairs:

- Bonnie Webber <bonnie.webber@ed.ac.uk>
School of Informatics, University of Edinburgh, 10 Crichton St, Edinburgh UK EH8 9AB
+44 131 650 4190
<http://homepages.inf.ed.ac.uk/bonnie/>
Expertise/Interests: Bonnie is a professor at the University of Edinburgh, and co-developer of the Penn Discourse TreeBank. She has worked on a range of discourse phenomena, including discourse anaphora, VP ellipsis, patterns of tense in discourse, the use of deictic reference in discourse, and the expression of discourse relations in text. Her current interest is in how these phenomena can be exploited to improve SMT.
- Andrei Popescu-Belis <andrei.popescu-belis@idiap.ch>
Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, 1920 Martigny,

⁴Guillou, Liane; Hardmeier, Christian; Nakov, Preslav; Stymne, Sara; Tiedemann, Jörg; Versley, Yannick; Cettolo, Mauro; Webber, Bonnie; Popescu-Belis, Andrei. (2016). Findings of the 2016 WMT shared task on cross-lingual pronoun prediction, *Proceedings of the First Conference on Machine Translation (WMT16)*, pages 522-539. Berlin, Germany. Association for Computational Linguistics. <https://aclweb.org/anthology/W/W16/W16-2345.pdf>

Switzerland

<http://www.idiap.ch/~apbelis/>

Phone: +41 27 721 7729

Expertise/Interests: Andrei's research focuses on language processing at the discourse-level, in particular reference and anaphora resolution, discourse markers and connectives, and dialogue acts; the evaluation of NLP and interactive systems, including machine translation systems and meeting browsers; and language and multimodal resources.

- Jörg Tiedemann <jorg.tiedemann@helsinki.fi> Department of Modern Languages, University of Helsinki

<http://www.helsinki.fi/~tiedeman/>

Jörg is a Professor of language technology at the Department of Modern Languages at the University of Helsinki. His main research interests are in cross-lingual NLP and machine translation. He has been working in MT for the past 15 years, and is the co-author of the monograph "Bitext Alignment", part of Morgan & Claypool's Synthesis Lectures on HLT. He has worked on document-level translation, and is currently co-supervising research on discourse-level approaches to Machine Translation.

Tentative Programme Committee

- Mauro Cettolo, Fondazione Bruno Kessler, Trento, Italy (confirmed)
- Liane Guillou, University of München, Germany
- Filip Ginter, University of Turku, Finland (confirmed)
- Christian Hardmeier, University of Uppsala, Sweden (confirmed)
- Shafiq Joty, Qatar Computing Research Institute, Doha, Qatar (confirmed)
- Lori Levin, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA (confirmed)
- Sharid Loaiciga Sanchez, University of Geneva, Switzerland (confirmed)
- Ekaterina Lapshinova-Koltunski, Saarland University, DE (confirmed)
- Ngoc-Quang Luong, Idiap Research Institute, Switzerland
- Thomas Meyer, Google, Zurich, Switzerland (confirmed)
- Preslav Nakov, Qatar Computing Research Institute, Doha, Qatar (confirmed)
- Michal Novak, Charles University, Prague CZ (confirmed)
- Maja Popovic, DFKI, Berlin DE (confirmed)
- Annette Rios, University of Zurich, Switzerland (confirmed)
- Lucia Specia, University of Sheffield, UK (confirmed)
- Sara Stymne, University of Uppsala, Uppsala, Sweden
- Yannick Versley, University of Heidelberg, Heidelberg, Germany
- Martin Volk, University of Zurich, Switzerland (confirmed)
- Min Zhang, Soochow University, Suzhou, China (confirmed)
- Sandrine Zufferey, University of Bern, Switzerland (confirmed)

6 Other Information

- **Proposed workshop length:** One day (Ideally, this would coincide with the second day of the Workshop on Machine Translation, so that the two workshops could share a poster session.)
- **Estimated number of attendees:** 25-30 people
- **Special requirements or technical needs:** None
- **Preferred venue:** EMNLP 2017 (Copenhagen) This are two reasons for this preference. First, the majority of people working in the area are in Europe, though the area has begun to attract a growing number of researchers from China. Copenhagen is thus the venue that will keep transportation costs down for the majority of expected attendees, as per the ACL General Policy on Workshops. Secondly, having the workshop in connection with EMNLP will give us the most time to prepare the resources and tools that we want to release in support of the shared task. If WMT is also held in Copenhagen, organizers of both workshops have already to have a shared poster session, since this was so successful in 2013 and 2015.