MODERN: Modeling Discourse Entities and Relations for Coherent Machine Translation*

A. POPESCU-BELIS[1], J. EVERS-VERMEUL[4], M. FISHEL[3], C. GRISOT[2], M.GROEN[4], J. HOEK[4], S. LOAICIGA[2], N.Q. LUONG[1], L. MASCARELL[3], T. MEYER[1], L. MICULICICH[1], J. MOESCHLER[2], X. PU[1], A. RIOS[3], T. SANDERS[4], M.VOLK[3], S. ZUFFEREY[4]

[1] Idiap Research Institute, 1920 Martigny, Switzerland
[2] University of Geneva, Department of Linguistics, 1211 Genève 4, Switzerland
[3] University of Zürich, Institute of Computational Linguistics, 8050 Zürich, Switzerland
[4] Utrecht University, Utrecht Institute of Linguistics, 3512 JK Utrecht, The Netherlands

andrei.popescu-belis@idiap.ch

Abstract. The MODERN project addresses coherence issues in sentence-by-sentence statistical MT, by propagating across sentences discourse-level information regarding discourse connectives, verb tenses, noun phrases and pronouns.

The goal of the MODERN project is to model and detect word dependencies across sentences, and to study their use by MT systems (MT), in order to demonstrate improvements in translation quality over state-of-the-art statistical MT, which still operates on a sentence-by-sentence basis. Three types of text-level dependencies are studied: referring expressions such as noun phrases and pronouns [4], discourse relations signaled by discourse connectives or implicit ones [1, 3], and verb tenses [2].

The overall approach of MODERN is to study the discourse-level phenomena from a theoretical perspective, but also using corpus-based approaches, in order to derive acceptable labels, features for automatic labeling, and training/test data. The automatic labeling systems are designed and combined with phrase-based statistical MT systems using factored models. The improvement in translating the respective phenomena is evaluated using specific automatic metrics or human evaluators.

MODERN started in 2013, building upon the COMTIS project started in 2010, with studies on English, French, German, Italian, Dutch, Arabic and Chinese. The main corpora used are Europarl, WIT3 (transcripts of TED talks), and Text+Berg (Swiss Alpine Club yearbooks).

References

1. Hoek J. et al., "The role of expectedness in the implicitation and explicitation of discourse relations", *Proc. of the 2nd DiscoMT workshop*, Lisbon, 2015.
2. Grisot C., *Temporal reference: empirical and theoretical perspectives*, PhD, UniGe, 2015.
3. Meyer T., *Discourse-level features for statistical machine translation*, PhD, EPFL, 2015.
4. Pu X., Mascarell L. et al., "Leveraging Compounds to Improve Noun Phrase Translation from Chinese and German", *Proc. of the ACL Student Session*, Beijing, 2015.

* MODERN is supported by the Swiss NSF, see www.idiap.ch/project/modern/.