# Christian Hardmeier: Discourse in statistical machine translation 2014: abstract

(ActaUniversitatis Upsaliensis – Studia Linguistica Upsaliensis.)

Author: Hardmeier, Christian (Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Languages, Department of Linguistics and Philology) (Datorlingvistik)

Title: Discourse in Statistical Machine Translation
Department: Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Languages, Department of Linguistics and Philology
Publication type: Doctoral thesis, monograph (Other academic)
Language: English
Place of publ.: Uppsala Publisher: Acta Universitatis Upsaliensis
Pages: 185
Series: Studia Linguistica Upsaliensia, ISSN 1652-1366; 15
Year of publ.: 2014
URI: urn:nbn:se:uu:diva-223798
Permanent link: http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-223798
ISBN: 978-91-554-8963-2
Subject category: Language Technology (Computational Linguistics)
Research subject: Computational Linguistics
Keywords(en) : Statistical machine translation, Discourse-level machine translation, Document decoding, Local search, Pronominal anaphora, Pronoun translation, Neural networks

Abstract(en) :

This thesis addresses the technical and linguistic aspects of discourse-level processing in phrase-based statistical machine translation (SMT). Connected texts can have complex text-level linguistic dependencies across sentences that must be preserved in translation. However, the models and algorithms of SMT are pervaded by locality assumptions. In a standard SMT setup, no model has more complex dependencies than an *n*-gram model. The popular stack decoding algorithm exploits this fact to implement efficient search with a dynamic programming technique. This is a serious technical obstacle to discourse-level modelling in SMT.

From a technical viewpoint, the main contribution of our work is the development of a document-level decoder based on stochastic local search that translates a complete document as a single unit. The decoder starts with an initial translation of the document, created randomly or by running a stack decoder, and refines it with a sequence of elementary operations. After each step, the current translation is scored by a set of feature models with access to the full document context and its translation. We demonstrate the viability of this decoding approach for different document-level models.

From a linguistic viewpoint, we focus on the problem of translating pronominal anaphora. After investigating the properties and challenges of the pronoun translation task both theoretically and by studying corpus data, a neural network model for cross-lingual pronoun prediction is presented. This network jointly performs anaphora resolution and pronoun prediction and is trained on bilingual corpus data only, with no need for manual coreference annotations. The network is then integrated as a feature model in the document-level SMT decoder and tested in an English–French SMT system. We show that the pronoun prediction network model more adequately represents discourse-level dependencies for less frequent pronouns than a simpler maximum entropy baseline with separate coreference resolution.

By creating a framework for experimenting with discourse-level features in SMT, this work contributes to a long-term perspective that strives for more thorough modelling of complex linguistic phenomena in translation. Our results on pronoun translation shed new light on a challenging, but essential problem in machine translation that is as yet unsolved.

Public defence: 2014-06-14, Universitetshuset, Sal X, Uppsala, 10:15 (English), Uppsala University
Degree: degree of Doctor of Philosophy
Supervisor: Nivre, Joakim, Professor (Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Languages, Department of Linguistics and Philology)
Tiedemann, Jörg, Professor (Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Languages, Department of Linguistics and Philology)
Federico, Marcello, Professor (Fondazione Bruno Kessler, Trento (Italy))
Opponent: Màrquez, Lluís, Dr. (Qatar Computing Research Institute)
Available from: 2014-05-23
Created: 2014-04-25
Last updated: 2014-05-23