ACL HLT 2011

Fifth Workshop on

Syntax, Semantics and Structure in Statistical Translation

SSST-5

Proceedings of the Workshop

Dekai Wu, Marianna Apidianaki,
Marine Carpuat and Lucia Specia (editors)

23 June, 2011
Portland, Oregon, USA

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 978-1-932432-99-2

Introduction

The Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5) was held on 23 June 2011 following the ACL HLT 2011 conference in Portland, Oregon. Like the first four SSST workshops in 2007, 2008, 2009, and 2010, it aimed to bring together researchers from different communities working in the rapidly growing field of structured statistical models of natural language translation.

During these past five years, statistical machine translation research has seen a movement toward not only tree-structured and syntactic models incorporating stochastic synchronous/transduction grammars, but also increasingly semantic models. There is no doubt that issues of deep syntax and shallow semantics are closely linked, and this encouraging trend has been reflected at recent SSST workshops. Semantic SMT research now includes context-dependent WSD (word sense disambiguation) for SMT (Carpuat and Wu 2007, 2008; Chan, Ng and Chiang 2007; Giménez and Màrquez 2007); SRL (semantic role labeling) for SMT (Wu and Fung 2009); and SRL for MT evaluation (Lo and Wu 2010, 2011).

In order to emphasize structure and representation at semantic and not only syntactic levels, “Semantics” has been explicitly added to the name of this year’s Workshop (the acronym remains SSST), and is a special workshop theme.

We selected 15 papers for this year’s workshop. Many either directly fall under the special theme of Semantics in SMT, or span the area between deep syntax and shallow semantics, illustrating the variety of semantic representations and models that are relevant to current statistical MT.

SRL predicate-argument structure clearly emerges as a useful representation for many aspects of SMT and MT evaluation. Wu and Palmer show that it is possible to automatically learn accurate cross-lingual SRL mappings between Chinese and English SRL annotated bitext. Input-side SRL is used to define reordering rules for Chinese-English word alignment (Meyers, Kosaka, Liao and Xue), and to improve pairwise translation hypothesis ranking (Pighin and Màrquez). Output-side SRL informs rule extraction in hierarchichal phrase-based SMT (Gao and Vogel), and provides structure for meaningfully comparing translation hypotheses and references in MT evaluation (Lo and Wu).

WSD also emerges as a prominent research direction with semantically richer SMT models designed to address ambiguity in translation lexical choice. Banchs and Costa-jussa use Latent Semantic Indexing to build a context-dependent phrase-based SMT model. Jiang, Du and Way integrate input paraphrases into SMT via confusion networks. Lefever and Hoste show that dedicated classifiers learned on parallel corpora outperform phrase-based SMT on a cross-lingual WSD task. SMT can also be seen as a tool to enrich semantic resources: McCrae, Espinoza, Ponsoda, Aguado-de-Cea and Cimiano propose several strategies for automatically translating ontologies and taxonomies, leveraging their rich semantic structure to compensate for the weakness of standard text translation methods.

A rich range of syntactic and tree-based approaches for learning translation rules is also seen. Attardi, Chanev and Miceli Barone learn reordering rules for a decoding approach drivenby a input-side dependency parser to guide reordering. Hanneman and Lavie describe a method for inducing nonterminals in synchronous/transduction grammars, by clustering nonterminal-pairs across input and output languages. Na and Lee propose a method for encoding alternative binarizations of a single input-side dependency tree into a forest by merging vertices before extracting translation rules. Hanneman, Burroughs and Lavie extract synchronous/transduction grammar rules combining input-side and output-side parse tree information with the highly lexicalized approach of hierarchical phrase-based methods. Input-side parse features are incorporated within a maximum-entropy reordering approach by Xiang, Ge and Ittycheriah. On the formal side, Saers and Wu show how to simplify calculation of rule expectations for expectation-maximization training of transduction grammars as well as monolingual grammars, by reifying rules directly into the hypergraph representation of a deductive system so that a rule becomes an extra child rather than meta-information of a hyperedge.

Thanks once again this year are due to our authors and our Program Committee for making the SSST workshop another success.

Dekai Wu, Marianna Apidianaki, Marine Carpuat, and Lucia Specia

Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract Nos. HR0011-06-C-0022, subcontract BBN Technologies and HR0011-06-C-0023, subcontract SRI International, and by the Hong Kong Research Grants Council (RGC) research grant GRF621008 (Dekai Wu); Alpage INRIA (Marianna Apidianaki); the National Research Council Institute for Information Technology (Marine Carpuat); and the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. HR0011-08-C-0110, subcontract IBM (Lucia Specia). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Organizers:

Dekai WU, Hong Kong University of Science and Technology (HKUST), Hong Kong

Co-chairs for special theme on Semantics in SMT:

Marianna APIDIANAKI, Alpage, INRIA and University Paris 7, France
Marine CARPUAT, National Research Council (NRC), Canada
Lucia SPECIA, University of Wolverhampton, UK

Program Committee:

Eneko AGIRRE, University of the Basque Country, Spain
Colin CHERRY, National Research Council (NRC), Canada
Marc DYMETMAN, Xerox Research Center Europe, France
Hieu HOANG, University of Edinburgh, UK
Philipp KOEHN, University of Edinburgh, UK
Philippe LANGLAIS, University of Montreal, Canada
Aurélien MAX, Université Paris Sud 11, France
Diana MCCARTHY, Lexical Computing, UK
Sudip Kumar NASKAR, Dublin City University, Ireland
Roberto NAVIGLI, University of Rome “La Sapienza”, Italy
Hwee Tou NG, National University of Singapore, Singapore
Sebastian PADO, Universität Heidelberg, Germany
Martha PALMER, University of Colorado, USA
Ted PEDERSEN, University of Minnesota, USA
Markus SAERS, Hong Kong University of Science and Technology (HKUST), Hong Kong
Matthew SNOVER, City University of New York, USA
Nicolas STROPPA, Google, Switzerland
François YVON, Université Paris Sud 11, France

Table of Contents

Conference Program

Session 1

09:00 Opening Remarks

09:15 *Automatic Projection of Semantic Structures: an Application to Pairwise Translation Ranking*
Daniele Pighin and Lluís Màrquez

09:40 *Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation*
Chi-kiu Lo and Dekai Wu

10:05 *Semantic Mapping Using Automatic Word Alignment and Semantic Role Labeling*
Shumin Wu and Martha Palmer

10:30 Coffee Break / Poster Session

Incorporating Source-Language Paraphrases into Phrase-Based SMT with Confusion Networks
Jie Jiang, Jinhua Du and Andy Way

Multi-Word Unit Dependency Forest-based Translation Rule Extraction
Hwidong Na and Jong-Hyeok Lee

An Evaluation and Possible Improvement Path for Current SMT Behavior on Ambiguous Nouns
Els Lefever and Véronique Hoste

Improving Reordering for Statistical Machine Translation with Smoothed Priors and Syntactic Features
Bing Xiang, Niyu Ge and Abraham Ittycheriah

Session 2

11:00 *Reestimation of Reified Rules in Semiring Parsing and Biparsing*
Markus Saers and Dekai Wu

11:25 *A Dependency Based Statistical Translation Model*
Giuseppe Attardi, Atanas Chanev and Antonio Valerio Miceli Barone

11:50 *Improving MT Word Alignment Using Aligned Multi-Stage Parses*
Adam Meyers, Michiko Kosaka, Shasha Liao and Nianwen Xue

12:15 Lunch

(continued)

Session 3

13:50 *Automatic Category Label Coarsening for Syntax-Based Machine Translation*
Greg Hanneman and Alon Lavie

14:15 *Utilizing Target-Side Semantic Role Labels to Assist Hierarchical Phrase-based Machine Translation*
Qin Gao and Stephan Vogel

14:40 *Combining statistical and semantic approaches to the translation of ontologies and taxonomies*
John McCrae, Mauricio Espinoza, Elena Montiel-Ponsoda, Guadalupe Aguado-de-Cea and Philipp Cimiano

15:05 *A Semantic Feature for Statistical Machine Translation*
Rafael E. Banchs and Marta R. Costa-jussa

15:30 Coffee Break / Poster Session

Session 4

16:00 *A General-Purpose Rule Extractor for SCFG-Based Machine Translation*
Greg Hanneman, Michelle Burroughs and Alon Lavie

16:25 Panel Discussion