Proceedings of

SSST-4

Fourth Workshop on

Syntax and Structure in Statistical Translation

Dekai Wu (editor)

COLING 2010 / SIGMT Workshop
23rd International Conference on Computational Linguistics
Beijing, China
28 August 2010

Produced by
Chinese Information Processing Society of China

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Introduction

The Fourth Workshop on Syntax and Structure in Statistical Translation (SSST-4) was held on 28 August 2010 following the Coling 2010 conference in Beijing. Like the first three SSST workshops in 2007, 2008, and 2009, it aimed to bring together researchers from different communities working in the rapidly growing field of statistical, tree-structured models of natural language translation.

We were honored to have Martin Kay deliver this year's invited keynote talk. This field is indebted to Martin Kay for not one but *two* of the classic cornerstone ideas that inspired bilingual tree-structured models for statistical machine translation: first, chart parsing, and second, parallel text alignment.

Tabular approaches to parsing, using dynamic programming and/or memoization, were heavily influenced by Kay's (1980) charts (or forests, packed forests, well-formed substring tables, or WFSTs). Today's biparsing models—the bilingual generalizations of this influential work—lie at the heart of numerous alignment and training algorithms for inversion transduction grammars or ITGs—including all syntax-directed transduction grammars or SDTGs (or synchronous CFGs) of binary rank or ternary rank, such as those learned by hierarchical phrase-based translation approaches.

At the same time, Kay and Röscheisen's (1988) seminal work on alignment of parallel texts led the way in statistical machine translation's basic paradigm of integrating the simultaneous learning of translation lexicons with aligning parallel texts. Today's biparsing models generalize this by simultaneously learning tree structures as well. Once again, cross-pollination of ideas across different areas and disciplines, empirical and theoretical, has provided mutual inspiration.

We selected 15 papers for this year's workshop. Studies emphasizing formal and algorithmic aspects include a method for intersecting synchronous/transduction grammars (S/TGs) with finite-state transducers (Dymetman and Cancedda), and a comparison of linear transduction grammars (LTGs) with ITGs (Saers, Nivre and Wu). Experiments on using syntactic features and constraints within flat phrase-based translation models include studies by Jiang, Du and Way; by Cao, Finch and Sumita; by Kolachina, Venkatapathy, Bangalore, Kolachina and PVS; and by Zhechev and van Genabith. Dependency constraints are also used to improve HMM word alignments for both flat phrase-based as well as S/TG based translation models (Ma and Way). Extensions to the features and parameterizations in two S/TG based translation models, as well as methods for merging models, are studied by Zollman and Vogel. The potential of incorporating LFG-style deep syntax within S/TGs is explored by Graham and van Genabith. A tree transduction based approach is presented by Khalilov and Sima'an. Meanwhile, Lo and Wu empirically compare n-gram, syntactic, and semantic structure based MT evaluation approaches. An encouraging trend is an uptick in work on low-resource language pairs and from underrepresented regions, including English-Persian (Mohaghegh, Sarrafzadeh and Moir), Manipuri-English (Singh and Bandyopadhyay), Tunisia (Khemakhem, Jamoussi and Ben hamadou), and English-Hindi (Venkatapathy, Sangal, Joshi, and Gali).

Once again this year, thanks are due to our authors and our Program Committee for making the SSST workshop another success.

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. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) 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

Program Committee:

Srinivas BANGALORE, AT&T Research, USA
Marine CARPUAT, Hong Kong University of Science and Technology (HKUST), Hong Kong
David CHIANG, USC Information Sciences Institute, USA
Pascale FUNG, Hong Kong University of Science and Technology (HKUST), Hong Kong
Daniel GILDEA, University of Rochester, USA
Dan KLEIN, University of California at Berkeley, USA
Kevin KNIGHT, USC Information Sciences Institute, USA
Jonas KUHN, University of Potsdam, Germany
Yang LIU, Institute of Computing Technology, Chinese Academy of Sciences, China
Yanjun MA, Dublin City University, Ireland
Daniel MARCU, USC Information Sciences Institute, USA
Yuji MATSUMOTO, Nara Institute of Science and Technology, Japan
Hermann NEY, RWTH Aachen, Germany
Owen RAMBOW, Columbia University, USA
Philip RESNIK, University of Maryland, USA
Stefan RIEZLER, Google Inc., USA
Libin SHEN, BBN Technologies, USA
Christoph TILLMANN, IBM T. J. Watson Research Center, USA
Stephan VOGEL, Carnegie Mellon University, USA
Taro WATANABE, NTT Communication Science Laboratories, Japan
Andy WAY, Dublin City University, Ireland
Yuk-Wah WONG, Google Inc., USA
Richard ZENS, Google Inc., USA
Chengqing ZONG, Institute of Automation, Chinese Academy of Sciences, China

Table of Contents

Conference Program

Saturday, August 28, 2010

9:00–10:40 *Open Tutorial: Tree-Structured and Syntactic SMT*
Dekai Wu

10:40–11:05 Coffee break

11:05–12:05 *Invited Keynote*
Martin Kay

12:05–12:25 *Intersecting Hierarchical and Phrase-Based Models of Translation: Formal Aspects and Algorithms*
Marc Dymetman and Nicola Cancedda

12:25–12:55 *A Systematic Comparison between Inversion Transduction Grammar and Linear Transduction Grammar for Word Alignment*
Markus Saers, Joakim Nivre and Dekai Wu

12:55–14:00 Lunch break

14:00–14:20 *Source-side Syntactic Reordering Patterns with Functional Words for Improved Phrase-based SMT*
Jie Jiang, Jinhua Du and Andy Way

14:20–14:40 *Syntactic Constraints on Phrase Extraction for Phrase-Based Machine Translation*
Hailong Cao, Andrew Finch and Eiichiro Sumita

14:40–15:00 *Phrase Based Decoding using a Discriminative Model*
Prasanth Kolachina, Sriram Venkatapathy, Srinivas Bangalore, Sudheer Kolachina and Avinesh PVS

15:00–15:20 *Seeding Statistical Machine Translation with Translation Memory Output through Tree-Based Structural Alignment*
Ventsislav Zhechev and Josef van Genabith

15:20–15:40 *Semantic vs. Syntactic vs. N-gram Structure for Machine Translation Evaluation*
Chi-kiu Lo and Dekai Wu

15:40–16:05 Posters / Coffee break

Arabic morpho-syntactic feature disambiguation in a translation context
Ines Turki Khemakhem, Salma Jamoussi and Abdelmajid Ben hamadou

Saturday, August 28, 2010 (continued)

A Discriminative Approach for Dependency Based Statistical Machine Translation
Sriram Venkatapathy, Rajeev Sangal, Aravind Joshi and Karthik Gali

Improved Language Modeling for English-Persian Statistical Machine Translation
Mahsa Mohaghegh, Abdolhossein Sarrafzadeh and Tom Moir

Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations
Thoudam Doren Singh and Sivaji Bandyopadhyay

16:05–16:25 *A Discriminative Syntactic Model for Source Permutation via Tree Transduction*
Maxim Khalilov and Khalil Sima'an

16:25–16:45 *HMM Word-to-Phrase Alignment with Dependency Constraints*
Yanjun Ma and Andy Way

16:45–17:05 *New Parameterizations and Features for PSCFG-Based Machine Translation*
Andreas Zollmann and Stephan Vogel

17:05–17:25 *Deep Syntax Language Models and Statistical Machine Translation*
Yvette Graham and Josef van Genabith

17:25–17:40 Discussion