Proceedings of

SSST-3

Third Workshop on

Syntax and Structure in Statistical Translation

Dekai Wu and David Chiang (editors)

NAACL HLT 2009
Boulder, Colorado
June 5, 2009

Production and Manufacturing by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

©2009 The Association for Computational Linguistics

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 978-1-932432-39-8

Introduction

The Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) was held on 5 June 2009 following the NAACL-HLT 2009 conference hosted by the University of Colorado at Boulder. Like the first two SSST workshops in 2007 and 2008, 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 Alfred V. Aho deliver this year's invited keynote talk. Along with Lewis and Stearns' (1968) seminal "Syntax-directed transduction" introducing syntax-directed transduction grammars or SDTGs—to which synchronous CFGs are also equivalent—the classic pair of Aho and Ullman's (1969) articles "Syntax-directed translations and the pushdown assembler" and "Properties of syntax-directed translations" established the foundations of formal transduction approaches to translation. The motivation behind their formal language theory work was compiler translation of programming languages. But much of the current work at SSST reflects the evolution of those ideas into today's state-of-the-art approaches to modeling of syntax and structure in statistical machine translation of human languages.

Nowhere is it better demonstrated that research is healthily driven by the pollination of ideas across these disciplines. These formal language pioneers foresaw clearly the importance of formalizing compositional models of transduction. But without empirical research across various natural languages, formal language theorists would not perhaps have anticipated the wide applicability in human language translation of an intermediate restricted class of transductions between the syntax-directed transductions that can be described by SDTGs (or synchronous CFGs) at one extreme, and the very restricted finite-state transductions that can be described by FSTs at the other—like the broad equivalence class of inversion transductions that can be described by ITGs (which include synchronous/transduction grammars whose rules are all binary rank, ternary rank, or monotonically straight or inverted in reordering permutation). Nor might they have foreseen the success of the rich variety of statistical machine learning techniques that have been developed to induce such synchronous/transduction grammars, such as the techniques introduced by the hierarchical phrase-based translation approach. As has so often happened over the years of cross-fertilization cycles between formal language and natural language research, the theoretical and empirical lines of research provide mutual inspiration.

We selected ten papers for this year's workshop. Studies on alignment ranged from the theoretical (Søgaard) to data analysis (Nakazawa and Kurohashi; Søgaard and Kuhn; Jiang, Li, Yang and Zhao), to empirical impact on actual translation performance (Saers and Wu; Hashimoto, Yamamoto, Okuma, Sumita and Tokuda). New contributions to translation decoding included purely unsupervised methods leveraging compositional structure constraints (Saers and Wu), methods using explicit syntactic information (Chang, Tseng, Jurafsky and Manning; Khalilov, Fonollosa and Dras), as well as methods attempting to blend the two (Hashimoto, Yamamoto, Okuma, Sumita and Tokuda; Hanneman and Lavie). The program was rounded out by a paper considering the use of explicit syntax in automatic evaluation (Wang, Zhao, Yang and Li).

We would like to thank our authors and our Program Committee for making this year's SSST workshop another success.

Dekai Wu and David Chiang

Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract Nos. HR0011-06-C-0023, subcontract SRI International (Dekai Wu) and HR0011-06-C-0022, subcontract BBN Technologies 9500008412 (David Chiang). 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
David CHIANG, USC Information Sciences Institute, USA

Program Committee:

Srinivas BANGALORE, AT&T Research, USA
Marine CARPUAT, Hong Kong University of Science and Technology (HKUST), Hong Kong
Pascale FUNG, Hong Kong University of Science and Technology (HKUST), Hong Kong
Daniel GILDEA, University of Rochester, 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
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

Table of Contents

Decoding with Syntactic and Non-Syntactic Phrases in a Syntax-Based Machine Translation System
Greg HANNEMAN and Alon LAVIE 1

Statistical Phrase Alignment Model Using Dependency Relation Probability
Toshiaki NAKAZAWA and Sadao KUROHASHI 10

Empirical Lower Bounds on Aligment Error Rates in Syntax-Based Machine Translation
Anders SØGAARD and Jonas KUHN 19

Improving Phrase-Based Translation via Word Alignments from Stochastic Inversion Transduction Grammars
Markus SAERS and Dekai WU 28

References Extension for the Automatic Evaluation of MT by Syntactic Hybridization
Bo WANG, Tiejun ZHAO, Muyun YANG and Sheng LI 37

A Study of Translation Rule Classification for Syntax-based Statistical Machine Translation
Hongfei JIANG, Sheng LI, Muyun YANG and Tiejun ZHAO 45

Discriminative Reordering with Chinese Grammatical Relations Features
Pi-Chuan CHANG, Huihsin TSENG, Dan JURAFSKY and Christopher D. MANNING 51

On the Complexity of Alignment Problems in Two Synchronous Grammar Formalisms
Anders SØGAARD 60

Reordering Model Using Syntactic Information of a Source Tree for Statistical Machine Translation
Kei HASHIMOTO, Hirohumi YAMAMOTO, Hideo OKUMA, Eiichiro SUMITA and Keiichi TOKUDA
69

Coupling Hierarchical Word Reordering and Decoding in Phrase-Based Statistical Machine Translation
Maxim KHALILOV, José A. R. FONOLLOSA and Mark DRAS 78

Conference Program

Friday, June 5, 2009

9:00–9:15 Opening

9:15–9:40 *Decoding with Syntactic and Non-Syntactic Phrases in a Syntax-Based Machine Translation System*
Greg HANNEMAN and Alon LAVIE

9:40–10:05 *Statistical Phrase Alignment Model Using Dependency Relation Probability*
Toshiaki NAKAZAWA and Sadao KUROHASHI

10:05–10:30 *Empirical Lower Bounds on Aligment Error Rates in Syntax-Based Machine Translation*
Anders SØGAARD and Jonas KUHN

10:30–11:00 Coffee Break

11:00–11:25 *Improving Phrase-Based Translation via Word Alignments from Stochastic Inversion Transduction Grammars*
Markus SAERS and Dekai WU

11:25–12:30 Invited Talk by Alfred V. AHO: "Unnatural Language Processing"

12:30–14:15 Lunch

14:15–14:40 *References Extension for the Automatic Evaluation of MT by Syntactic Hybridization*
Bo WANG, Tiejun ZHAO, Muyun YANG and Sheng LI

14:40–15:05 *A Study of Translation Rule Classification for Syntax-based Statistical Machine Translation*
Hongfei JIANG, Sheng LI, Muyun YANG and Tiejun ZHAO

15:05–15:30 *Discriminative Reordering with Chinese Grammatical Relations Features*
Pi-Chuan CHANG, Huihsin TSENG, Dan JURAFSKY and Christopher D. MANNING

15:30–16:00 Coffee Break

16:00–16:25 *On the Complexity of Alignment Problems in Two Synchronous Grammar Formalisms*
Anders SØGAARD

Friday, June 5, 2009 (continued)

16:25–16:50 *Reordering Model Using Syntactic Information of a Source Tree for Statistical Machine Translation*
Kei HASHIMOTO, Hirohumi YAMAMOTO, Hideo OKUMA, Eiichiro SUMITA and Keiichi TOKUDA

16:50–17:15 *Coupling Hierarchical Word Reordering and Decoding in Phrase-Based Statistical Machine Translation*
Maxim KHALILOV, José A. R. FONOLLOSA and Mark DRAS

17:15–17:30 Concluding Discussion