HLT-NAACL 06

Statistical Machine Translation

Proceedings of the Workshop

8-9 June 2006
New York City, USA

Production and Manufacturing by
Omnipress Inc.
2600 Andersen Street
Madison, WI 53704

©2006 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

Introduction

The HLT-NAACL 2006 Workshop on Statistical Machine Translation (WMT-06) took place on Thursday, June 8 and Friday, June 9 in New York City, immediately following the *Human Language Technology Conference — North American Chapter of the Association for Computational Linguistics Annual Meeting*, which was hosted by New York University.

This is the second time that this workshop has been held. The first time was last year as part of the *ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, which was a merger of two workshops that were originally proposed as independent events.

The focus of our workshop was to use parallel corpora for machine translation. Recent experimentation has shown that the performance of SMT systems varies greatly with the source and target language. In this workshop we encouraged researchers to investigate ways to improve the performance of SMT systems for diverse languages, including morphologically more complex languages and languages with partial free word order.

Prior to the workshop, in addition to soliciting relevant papers for review and possible presentation, we conducted a shared task that brought together machine translation systems for an evaluation on previously unseen data. This year's task resembled the one from last year's in many ways, but also included a manual evaluation of MT system output and focused on translation *from* English into other languages, whereas most other evaluations focus on translation *into* English.

The results of the shared task were announced at the workshop, and these proceedings also include an overview paper for the shared task that summarizes the results, as well as provides information about the data used and any procedures that were followed in conducting or scoring the task. In addition, there are short papers from each participating team that describe their underlying system in some detail.

The first day of the workshop, Thursday, June 8 was dedicated to full paper presentations, whereas the second day, Friday June 9 was mainly dedicated to system descriptions and discussions from teams that have participated in the shared task.

The workshop attracted a considerably larger number of submissions compared to last year's workshop. In total, WMT-06 featured 13 full paper oral presentations and 12 shared task presentations. The invited talk was given by Kevin Knight of the Information Sciences Institute/University of Southern California.

We would like to thank the members of the Program Committee for their timely reviews. We are also indebted to the many volunteers who served as judges in the manual evaluation of the shared task.

Philipp Koehn and Christof Monz

Co-Chairs

Organizers:

Philipp Koehn, Univeristy of Edinburgh, UK
Christof Monz, Queens Mary, Univeristy of London, UK

Program Committee:

Yaser Al-Onaizan, IBM, USA
Bill Byrne, University of Cambridge, UK
Chris Callison-Burch, University of Edinburgh, UK
Francisco Casacuberta, University of Valencia, Spain
David Chiang, ISI/University of Southern California, UK
Stephen Clark, Oxford University, UK
Marcello Federico, ITC-IRST, Italy
George Foster, Canada National Research Council, Canada
Alexander Fraser, ISI/University of Southern California, USA
Ulrich Germann, University of Toronto, Canada
Jan Hajic, Charles University, Czech Republic
Kevin Knight, ISI/University of Southern California, USA
Greg Kondrak, University of Alberta, Canada
Shankar Kumar, Google, USA
Philippe Langlais, University of Montreal, Canada
Daniel Marcu, ISI/University of Southern California, USA
Dan Melamed, New York University, USA
Franz-Josef Och, Google, USA
Miles Osborne, University of Edinburgh, UK
Philip Resnik, University of Maryland, USA
Libin Shen, University of Pennsylvania, USA
Wade Shen, MIT-Lincoln Labs, USA
Michel Simard, Canada National Research Council, Canada
Eiichiro Sumita, ATR Spoken Language Translation Research Laboratories, Japan
Joerg Tiedemann, University of Groningen, Netherlands
Christoph Tillmann, IBM, USA
Taro Watanabe, NTT, Japan
Dekai Wu, HKUST, China
Richard Zens, RWTH Aachen, Germany

Additional Reviewers:

Colin Cherry, University of Alberta, Canada
Fatiha Sadat, Canada National Research Council, Canada
Tarek Sherif, University of Alberta, Canada

Invited Speaker:

Kevin Knight, ISI/University of Southern California, USA

Table of Contents

Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output
Maja Popovic, Adrià de Gispert, Deepa Gupta, Patrik Lambert, Hermann Ney, José B. Mariño, Marcello Federico and Rafael Banchs 1

Initial Explorations in English to Turkish Statistical Machine Translation
ilknur Durgar El-Kahlout and Kemal Oflazer 7

Morpho-syntactic Arabic Preprocessing for Arabic to English Statistical Machine Translation
Anas El Isbihani, Shahram Khadivi, Oliver Bender and Hermann Ney 15

Quasi-Synchronous Grammars: Alignment by Soft Projection of Syntactic Dependencies
David Smith and Jason Eisner 23

Why Generative Phrase Models Underperform Surface Heuristics
John DeNero, Dan Gillick, James Zhang and Dan Klein 31

Phrase-Based SMT with Shallow Tree-Phrases
Philippe Langlais and Fabrizio Gotti 39

Searching for alignments in SMT. A novel approach based on an Estimation of Distribution Algorithm
Luis Rodríguez, Ismael García-Varea and Jose A. Gámez 47

Discriminative Reordering Models for Statistical Machine Translation
Richard Zens and Hermann Ney 55

Generalized Stack Decoding Algorithms for Statistical Machine Translation
Daniel Ortiz-Martínez, Ismael García-Varea and Francisco Casacuberta 64

N-Gram Posterior Probabilities for Statistical Machine Translation
Richard Zens and Hermann Ney 72

Partitioning Parallel Documents Using Binary Segmentation
Jia Xu, Richard Zens and Hermann Ney 78

Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation
Karolina Owczarzak, Declan Groves, Josef Van Genabith and Andy Way 86

How Many Bits Are Needed To Store Probabilities for Phrase-Based Translation?
Marcello Federico and Nicola Bertoldi 94

Manual and Automatic Evaluation of Machine Translation between European Languages
Philipp Koehn and Christof Monz 102

NTT System Description for the WMT2006 Shared Task
Taro Watanabe, Hajime Tsukada and Hideki Isozaki 122

Mood at work: Ramses versus Pharaoh
Alexandre Patry, Fabrizio Gotti and Philippe Langlais 126

Stochastic Inversion Transduction Grammars for Obtaining Word Phrases for Phrase-based Statistical Machine Translation
Joan Andreu Sánchez and José Miguel Benedí 130

PORTAGE: with Smoothed Phrase Tables and Segment Choice Models
Howard Johnson, Fatiha Sadat, George Foster, Roland Kuhn, Michel Simard, Eric Joanis and Samuel Larkin 134

Syntax Augmented Machine Translation via Chart Parsing
Andreas Zollmann and Ashish Venugopal 138

TALP Phrase-based statistical translation system for European language pairs
Marta R. Costa-jussà, Josep M. Crego, Adrià de Gispert, Patrik Lambert, Maxim Khalilov, José B. Mariño, José A. R. Fonollosa and Rafael Banchs 142

Phramer - An Open Source Statistical Phrase-Based Translator
Marian Olteanu, Chris Davis, Ionut Volosen and Dan Moldovan 146

Language Models and Reranking for Machine Translation
Marian Olteanu, Pasin Suriyentrakorn and Dan Moldovan 150

Constraining the Phrase-Based, Joint Probability Statistical Translation Model
Alexandra Birch, Chris Callison-Burch, Miles Osborne and Philipp Koehn 154

Microsoft Research Treelet Translation System: NAACL 2006 Europarl Evaluation
Arul Menezes, Kristina Toutanova and Chris Quirk 158

N-gram-based SMT System Enhanced with Reordering Patterns
Josep M. Crego, Adrià de Gispert, Patrik Lambert, Marta R. Costa-jussà, Maxim Khalilov, Rafael Banchs, José B. Mariño and José A. R. Fonollosa 162

The LDV-COMBO system for SMT
Jesús Giménez and Lluís Màrquez 166

Conference Program

Thursday, June 8, 2006

8:45–9:00 Opening Remarks

Session 1: Paper Presentations

9:00–9:30 *Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output*
Maja Popovic, Adrià de Gispert, Deepa Gupta, Patrik Lambert, Hermann Ney, José B. Mariño, Marcello Federico and Rafael Banchs

9:30–10:00 *Initial Explorations in English to Turkish Statistical Machine Translation*
ilknur Durgar El-Kahlout and Kemal Oflazer

10:00–10:30 *Morpho-syntactic Arabic Preprocessing for Arabic to English Statistical Machine Translation*
Anas El Isbihani, Shahram Khadivi, Oliver Bender and Hermann Ney

10:30–11:00 Coffee Break

Session 2: Paper Presentations

11:00–11:30 *Quasi-Synchronous Grammars: Alignment by Soft Projection of Syntactic Dependencies*
David Smith and Jason Eisner

11:30–12:00 *Why Generative Phrase Models Underperform Surface Heuristics*
John DeNero, Dan Gillick, James Zhang and Dan Klein

12:00–12:30 *Phrase-Based SMT with Shallow Tree-Phrases*
Philippe Langlais and Fabrizio Gotti

12:30–14:00 Lunch

Thursday, June 8, 2006 (continued)

Session 3: Paper Presentations

14:00–14:30 *Searching for alignments in SMT. A novel approach based on an Estimation of Distribution Algorithm*
Luis Rodríguez, Ismael García-Varea and Jose A. Gámez

14:30–15:30 Invited Talk by Kevin Knight

15:30–16:00 Coffee Break

Session 4: Paper Presentations

16:00–16:30 *Discriminative Reordering Models for Statistical Machine Translation*
Richard Zens and Hermann Ney

16:30–17:00 *Generalized Stack Decoding Algorithms for Statistical Machine Translation*
Daniel Ortiz-Martínez, Ismael García-Varea and Francisco Casacuberta

17:00–17:30 *N-Gram Posterior Probabilities for Statistical Machine Translation*
Richard Zens and Hermann Ney

Friday, June 9, 2006

Session 5: Paper Presentations

9:00–9:30 *Partitioning Parallel Documents Using Binary Segmentation*
Jia Xu, Richard Zens and Hermann Ney

9:30–10:00 *Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation*
Karolina Owczarzak, Declan Groves, Josef Van Genabith and Andy Way

10:00–10:30 *How Many Bits Are Needed To Store Probabilities for Phrase-Based Translation?*
Marcello Federico and Nicola Bertoldi

10:30–11:00 Coffee Break

Friday, June 9, 2006 (continued)

Session 6: Shared Task

11:00–11:30 *Manual and Automatic Evaluation of Machine Translation between European Languages*
Philipp Koehn and Christof Monz

11:30–11:45 *NTT System Description for the WMT2006 Shared Task*
Taro Watanabe, Hajime Tsukada and Hideki Isozaki

11:45–12:00 *Mood at work: Ramses versus Pharaoh*
Alexandre Patry, Fabrizio Gotti and Philippe Langlais

12:00–14:00 Lunch

Session 7: Shared Task

14:00–14:15 *Stochastic Inversion Transduction Grammars for Obtaining Word Phrases for Phrase-based Statistical Machine Translation*
Joan Andreu Sánchez and José Miguel Benedí

14:15–14:30 *PORTAGE: with Smoothed Phrase Tables and Segment Choice Models*
Howard Johnson, Fatiha Sadat, George Foster, Roland Kuhn, Michel Simard, Eric Joanis and Samuel Larkin

14:30–14:45 *Syntax Augmented Machine Translation via Chart Parsing*
Andreas Zollmann and Ashish Venugopal

14:45–15:00 *TALP Phrase-based statistical translation system for European language pairs*
Marta R. Costa-jussà, Josep M. Crego, Adrià de Gispert, Patrik Lambert, Maxim Khalilov, José B. Mariño, José A. R. Fonollosa and Rafael Banchs

15:00–15:15 *Phramer - An Open Source Statistical Phrase-Based Translator*
Marian Olteanu, Chris Davis, Ionut Volosen and Dan Moldovan

15:15–15:30 *Language Models and Reranking for Machine Translation*
Marian Olteanu, Pasin Suriyentrakorn and Dan Moldovan

15:30–16:00 Coffee Break

Friday, June 9, 2006 (continued)

Session 8: Shared Task

16:00–16:15 *Constraining the Phrase-Based, Joint Probability Statistical Translation Model*
Alexandra Birch, Chris Callison-Burch, Miles Osborne and Philipp Koehn

16:15–16:30 *Microsoft Research Treelet Translation System: NAACL 2006 Europarl Evaluation*
Arul Menezes, Kristina Toutanova and Chris Quirk

16:30–16:45 *N-gram-based SMT System Enhanced with Reordering Patterns*
Josep M. Crego, Adrià de Gispert, Patrik Lambert, Marta R. Costa-jussà, Maxim Khalilov, Rafael Banchs, José B. Mariño and José A. R. Fonollosa

16:45–17:00 *The LDV-COMBO system for SMT*
Jesús Giménez and Lluís Màrquez

17:00–18:00 Panel Discussion