ACL-05

# Building and Using Parallel Texts:

# Data-Driven Machine Translation and Beyond

## Proceedings of the Workshop

29-30 June 2005
University of Michigan
Ann Arbor, Michigan, USA

Production and Manufacturing by
*Omnipress Inc.*
*Post Office Box 7214*
*Madison, WI 53707-7214*

©2005 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

# Introduction

The ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, took place on Wednesday, June 29 and Thursday, June 30 in Ann Arbor Michigan, immediately following the 43rd Annual Meeting of the Association for Computational Linguistics.

This workshop represented a merger of two workshops that were originally proposed as independent events. Joel Martin, Rada Mihalcea, and Ted Pedersen had proposed a workshop on *Building and Using Parallel Texts for Languages with Scarce Resources*, which was intended as a follow-up event to the NAACL 2003 Workshop on Parallel Text that had been organized by Mihalcea and Pedersen. At the same time, Philipp Koehn and Christof Monz had proposed a workshop on *Exploiting Parallel Texts for Statistical Machine Translation*, featuring a shared task on Phrase Based Machine Translation.

Given the close relationship between the two proposed topics, the idea of a merger was quickly embraced by all concerned. It was agreed that the workshop would have two tracks, one regarding Parallel Texts for Languages with Scarce Resources (Track 1), and the other focused on Statistical Machine Translation (Track 2).

Prior to the workshop, in addition to soliciting relevant papers for review and possible presentation, the organizers of both tracks conducted shared tasks that brought together systems for an evaluation on previously unseen data. Track 1 featured a Word Alignment shared task, where the object was to align parallel text in one or more of the following langauge pairs: Inuktitut–English, Romanian–English, and Hindi–English. Track 2 carried out a shared task on Phrase Based Statistical Machine Translation, where eleven participating teams competed to build machine translation systems for French–English, Spanish–English, German–English, and Finnish–English.

The results of the shared tasks were announced at the workshop, and these proceedings also include an overview paper for each 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 for each shared task that describe their underlying system in some detail.

Wednesday, June 29 was dedicated to Track 1. It featured an invited talk by Mike Maxwell of the Linguistic Data Consortium, eight long paper presentations relevant to the topic of building and using parallel texts for languages with scarce resources, six short paper presentations describing systems that participated in the Word Alignment shared task (four additional short papers are included in the proceedings), a shared task overview, and a panel discussion about lessons learned from the shared task.

Track 2 was featured on Thursday, June 30. It included an invited talk by Franz Josef Och of Google, six long paper presentations, a shared task overview, and nine shared task system descriptions.

We would like to thank the members of the Program Committee for their timely reviews.

Philipp Koehn, Joel Martin, Rada Mihalcea, Christof Monz, and Ted Pedersen
Co–Organizers

**Organizers:**

Philipp Koehn (University of Edinburgh)
Joel Martin (National Research Council of Canada)
Rada Mihalcea (University of North Texas)
Christof Monz (University of Maryland)
Ted Pedersen (University of Minnesota, Duluth)

**Invited Speakers:**

Mike Maxwell (Linguistic Data Consortium, University of Pennsylvania)
Franz Josef Och (Google)

**Program Committee:**

Lars Ahrenberg (Linköping University)
Bill Byrne (University of Cambridge, Johns Hopkins University)
Chris Callison-Burch (University of Edinburgh)
Nicoletta Calzolari (Istituto di Linguistica Computazionale del CNR, Pisa)
Francisco Casacuberta (Universitat Politècnica de València)
David Chiang (University of Maryland)
Mona Diab (Columbia University)
George Foster (National Research Council of Canada)
Alexander Fraser (ISI/University of Southern California)
Pascale Fung (Hong Kong University of Science and Technology)
Rob Gaizauskas (University of Sheffield)
Ulrich Germann (University of Toronto)
Dan Gildea (University of Rochester)
Jan Hajic (Charles University)
Andrew Hardie (University of Lancaster)
Rebecca Hwa (University of Pittsburgh)
Nancy Ide (Vassar College)
Kevin Knight (ISI/University of Southern California)
Greg Kondrak (University of Alberta)
Roland Kuhn (National Research Council of Canada)
Shankar Kumar (Johns Hopkins University)
Philippe Langlais (University of Montreal)
Alon Lavie (Carnegie Mellon University)
Lori Levin (Carnegie Mellon University)
Daniel Marcu (ISI/University of Southern California)
Tony McEnery (University of Lancaster)
Bridget McInnes (University of Minnesota, Twin Cities)
Magnus Merkel (Linköping University)
Bob Moore (Microsoft Research)

Herman Ney (RWTH Aachen)
Maria das Graças Volpe Nunes (University of São Paulo)
Franz Josef Och (Google)
Kemal Oflazer (Sabancı University)
Miles Osborne (University of Edinburgh)
Andrei Popescu-Belis (University of Geneva)
Katharina Probst (Carnegie Mellon University)
Amruta Purandare (University of Pittsburgh)
Florence Reeder (MITRE)
Philip Resnik (University of Maryland)
Antonio Ribeiro (European Commission, Joint Research Centre)
Michel Simard (Xerox Research Centre Europe)
Kevin Scannell (St. Louis University)
Libin Shen (University of Pennsylvania)
Eiichiro Sumita (ATR Spoken Language Communication Research Laboratories)
Joerg Tiedemann (University of Groningen)
Christoph Tillmann (IBM)
Hajime Tsukada (NTT Communication Science Laboratories)
Dan Tufiş (Research Institute for AI of the Romanian Academy)
Jean Véronis (Université de Provence)
Michelle Vanni (Army Research Lab)
Stephan Vogel (Carnegie Mellon University)
Clare Voss (Army Research Lab)
Taro Watanabe (ATR Spoken Language Translation Research Laboratories)
Dekai Wu (Hong Kong University of Science and Technology)

**Additional Reviewers:**

Colin Cherry (University of Alberta)
Behrang Mohit (University of Pittsburgh)

# Table of Contents

*Association-Based Bilingual Word Alignment*
Robert C. Moore . . . . . 1

*Cross Language Text Categorization by Acquiring Multilingual Domain Models from Comparable Corpora*
Alfio Gliozzo and Carlo Strapparava . . . . . 9

*Parsing Word-Aligned Parallel Corpora in a Grammar Induction Context*
Jonas Kuhn . . . . . 17

*Bilingual Word Spectral Clustering for Statistical Machine Translation*
Bing Zhao, Eric P. Xing and Alex Waibel . . . . . 25

*Revealing Phonological Similarities between Related Languages from Automatically Generated Parallel Corpora*
Karin Müller . . . . . 33

*Augmenting a Small Parallel Text with Morpho-Syntactic Language*
Maja Popović, David Vilar, Hermann Ney, Slobodan Jovičić and Zoran Šarić . . . . . 41

*Induction of Fine-Grained Part-of-Speech Taggers via Classifier Combination and Crosslingual Projection*
Elliott Drábek and David Yarowsky . . . . . 49

*A Hybrid Approach to Align Sentences and Words in English-Hindi Parallel Corpora*
Niraj Aswani and Robert Gaizauskas . . . . . 57

*Word Alignment for Languages with Scarce Resources*
Joel Martin, Rada Mihalcea and Ted Pedersen . . . . . 65

*NUKTI: English-Inuktitut Word Alignment System Description*
Philippe Langlais, Fabrizio Gotti and Guihong Cao . . . . . 75

*Models for Inuktitut-English Word Alignment*
Charles Schafer and Elliott Drábek . . . . . 79

*Improved HMM Alignment Models for Languages with Scarce Resources*
Adam Lopez and Philip Resnik . . . . . 83

*Symmetric Probabilistic Alignment*
Ralf D. Brown, Jae Dong Kim, Peter J. Jansen and Jaime G. Carbonell . . . . . 87

*ISI's Participation in the Romanian-English Alignment Task*
Alexander Fraser and Daniel Marcu . . . . . 91

*Experiments Using MAR for Aligning Corpora*
Juan Miguel Vilar . . . . . 95

*Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs*
Anil Kumar Singh and Samar Husain . . . . . 99

*Combined Word Alignments*
Dan Tufis, Radu Ion, Alexandru Ceausu and Dan Stefanescu . . . . . 107

*LIHLA: Shared Task System Description*
Helena M. Caseli, Maria G. V. Nunes and Mikel L. Forcada . . . . . 111

*Aligning words in English-Hindi Parallel Corpora*
Niraj Aswani and Robert Gaizauskas . . . . . 115

*Shared Task: Statistical Machine Translation between European Languages*
Philipp Koehn and Christof Monz . . . . . 119

*Improved Language Modeling for Statistical Machine Translation*
Katrin Kirchhoff and Mei Yang . . . . . 125

*PORTAGE: A Phrase-Based Machine Translation System*
Fatiha Sadat, Howard Johnson, Akakpo Agbago, George Foster, Roland Kuhn, Joel Martin and Aaron Tikuisis . . . . . 129

*Statistical Machine Translation of Euparl Data by using Bilingual N-grams*
Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert and José B. Mariño . . . . 133

*RALI: SMT Shared Task System Description*
Philippe Langlais, Guihong Cao and Fabrizio Gotti . . . . . 137

*A Generalized Alignment-Free Phrase Extraction*
Bing Zhao and Stephan Vogel . . . . . 141

*Combining Linguistic Data Views for Phrase-based SMT*
Jesús Giménez and Lluís Màrquez . . . . . 145

*Improving Phrase-Based Statistical Translation by Modifying Phrase Extraction and Including Several Features*
Marta Ruiz Costa-jussà and José A. R. Fonollosa . . . . . 149

*First Steps towards Multi-Engine Machine Translation*
Andreas Eisele . . . . . 155

*Competetive Grouping in Integrated Phrase Segmentation and Alignment Model*
Ying Zhang and Stephan Vogel . . . . . 159

*Deploying Part-of-Speech Patterns to Enhance Statistical Phrase-Based Machine Translation Resources*
Christina Lioma and Iadh Ounis . . . . . 163

*Novel Reordering Approaches in Phrase-Based Statistical Machine Translation*
Stephan Kanthak, David Vilar, Evgeny Matusov, Richard Zens and Hermann Ney . . . . . 167

*Gaming Fluency: Evaluating the Bounds and Expectations of Segment-based Translation Memory*
John Henderson and William Morgan . . . . . 175

*Hybrid Example-Based SMT: the Best of Both Worlds?*
Declan Groves and Andy Way . . . . . 183

*Word Graphs for Statistical Machine Translation*
Richard Zens and Hermann Ney . . . . . 191

*A Recursive Statistical Translation Model*
Juan Miguel Vilar and Enrique Vidal . . . . . 199

*Training and Evaluating Error Minimization Decision Rules for Statistical Machine Translation*
Ashish Venugopal, Andreas Zollmann and Alex Waibel . . . . . 208

# Conference Program

**Wednesday, June 29, 2005**

8:45–9:00 Welcome

**Invited Talk**

9:00–10:00 Mike Maxwell *So many languages, so few resources: How to bridge the gap?*

**Session 1: Long Papers**

10:00–10:20 *Association-Based Bilingual Word Alignment*
Robert C. Moore

10:20–11:00 Break

**Session 2: Long Papers (continued)**

11:00–11:20 *Cross Language Text Categorization by Acquiring Multilingual Domain Models from Comparable Corpora*
Alfio Gliozzo and Carlo Strapparava

11:20–11:40 *Parsing Word-Aligned Parallel Corpora in a Grammar Induction Context*
Jonas Kuhn

11:40–12:00 *Bilingual Word Spectral Clustering for Statistical Machine Translation*
Bing Zhao, Eric P. Xing and Alex Waibel

12:00–12:20 *Revealing Phonological Similarities between Related Languages from Automatically Generated Parallel Corpora*
Karin Müller

12:20–12:40 *Augmenting a Small Parallel Text with Morpho-Syntactic Language*
Maja Popović, David Vilar, Hermann Ney, Slobodan Jovičić and Zoran Šarić

12:40–2:00 Lunch

**Wednesday, June 29, 2005 (continued)**

**Session 3: Long Papers (continued)**

2:00–2:20 *Induction of Fine-Grained Part-of-Speech Taggers via Classifier Combination and Crosslingual Projection*
Elliott Drábek and David Yarowsky

2:20–2:40 *A Hybrid Approach to Align Sentences and Words in English-Hindi Parallel Corpora*
Niraj Aswani and Robert Gaizauskas

**Shared Task I Overview**

2:40–3:00 *Word Alignment for Languages with Scarce Resources*
Joel Martin, Rada Mihalcea and Ted Pedersen

**Session 4: Shared Task I Papers**

3:00–3:15 *NUKTI: English-Inuktitut Word Alignment System Description*
Philippe Langlais, Fabrizio Gotti and Guihong Cao

3:15–3:30 *Models for Inuktitut-English Word Alignment*
Charles Schafer and Elliott Drábek

3:30–4:00 Break

**Session 5: Shared Task I Papers (continued)**

4:00–4:15 *Improved HMM Alignment Models for Languages with Scarce Resources*
Adam Lopez and Philip Resnik

4:15–4:30 *Symmetric Probabilistic Alignment*
Ralf D. Brown, Jae Dong Kim, Peter J. Jansen and Jaime G. Carbonell

4:30–4:45 *ISI's Participation in the Romanian-English Alignment Task*
Alexander Fraser and Daniel Marcu

4:45–5:00 *Experiments Using MAR for Aligning Corpora*
Juan Miguel Vilar

**Wednesday, June 29, 2005 (continued)**

**Shared Task I Papers without Presentations**

*Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs*
Anil Kumar Singh and Samar Husain

*Combined Word Alignments*
Dan Tufis, Radu Ion, Alexandru Ceausu and Dan Stefanescu

*LIHLA: Shared Task System Description*
Helena M. Caseli, Maria G. V. Nunes and Mikel L. Forcada

*Aligning words in English-Hindi Parallel Corpora*
Niraj Aswani and Robert Gaizauskas

**Shared Task I Panel Discussion**

5:00–6:00 TBA *Lessons Learned, and Future Directions*

**Thursday, June 30, 2005**

**Shared Task II Overview**

9:15–9:30 *Shared Task: Statistical Machine Translation between European Languages*
Philipp Koehn and Christof Monz

**Session 6: Shared Task II Papers**

9:30–9:45 *Improved Language Modeling for Statistical Machine Translation*
Katrin Kirchhoff and Mei Yang

9:45–10:00 *PORTAGE: A Phrase-Based Machine Translation System*
Fatiha Sadat, Howard Johnson, Akakpo Agbago, George Foster, Roland Kuhn, Joel Martin and Aaron Tikuisis

10:00–10:15 *Statistical Machine Translation of Euparl Data by using Bilingual N-grams*
Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert and José B. Mariño

**Thursday, June 30, 2005 (continued)**

10:15–11:00 Break

**Session 7: Shared Task II Papers (continued)**

11:00–11:15 *RALI: SMT Shared Task System Description*
Philippe Langlais, Guihong Cao and Fabrizio Gotti

11:15–11:30 *A Generalized Alignment-Free Phrase Extraction*
Bing Zhao and Stephan Vogel

11:30–11:45 *Combining Linguistic Data Views for Phrase-based SMT*
Jesús Giménez and Lluís Màrquez

11:45–12:00 *Improving Phrase-Based Statistical Translation by Modifying Phrase Extraction and Including Several Features*
Marta Ruiz Costa-jussà and José A. R. Fonollosa

12:00–12:15 *First Steps towards Multi-Engine Machine Translation*
Andreas Eisele

12:15–12:30 *Competetive Grouping in Integrated Phrase Segmentation and Alignment Model*
Ying Zhang and Stephan Vogel

**Shared Task II Paper without Presentation**

*Deploying Part-of-Speech Patterns to Enhance Statistical Phrase-Based Machine Translation Resources*
Christina Lioma and Iadh Ounis

12:30–2:00 Lunch

**Thursday, June 30, 2005 (continued)**

**Invited Talk**

2:00–3:00 Franz Och *Statistical Machine Translation: The Fabulous Present and Future*

**Session 7: Long Papers**

3:10–3:30 *Novel Reordering Approaches in Phrase-Based Statistical Machine Translation*
Stephan Kanthak, David Vilar, Evgeny Matusov, Richard Zens and Hermann Ney

3:30–4:00 Break

**Session 8: Long Papers (continued)**

4:00–4:20 *Gaming Fluency: Evaluating the Bounds and Expectations of Segment-based Translation Memory*
John Henderson and William Morgan

4:20–4:40 *Hybrid Example-Based SMT: the Best of Both Worlds?*
Declan Groves and Andy Way

4:40–5:00 *Word Graphs for Statistical Machine Translation*
Richard Zens and Hermann Ney

5:00–5:20 *A Recursive Statistical Translation Model*
Juan Miguel Vilar and Enrique Vidal

5:20–5:40 *Training and Evaluating Error Minimization Decision Rules for Statistical Machine Translation*
Ashish Venugopal, Andreas Zollmann and Alex Waibel