ACL-08: HLT

Third Workshop on Statistical Machine Translation

Proceedings of the Workshop

19 June 2008
Ohio State University
Columbus, Ohio, USA

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

Introduction

The ACL 2008 Workshop on Statistical Machine Translation (WMT-08) took place on Thursday, June 19 in Columbus, Ohio, United States, immediately following the annual meeting of the Association for Computational Linguistics, which was hosted by the Ohio State University.

This is the third time this workshop has been held. It has its root in the ACL 2005 Workshop on Building and Using Parallel Texts In the following years the Workshop on Statistical Machine Translation was held at HLT-NAACL 2006 in New York City, US, and at ACL 2007 in Prague, Czech Republic.

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 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 shared tasks of previous years in many ways, but also included Hungarian-English and Spanish-German as new language pairs. In addition, we evaluated submitted systems against new test sets from the newswire domain.

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.

Due to the large number of high quality submission for the full paper track, shared task submissions were presented as posters. The poster session was held in the afternoon and gave participants of the shared task the opportunity to present their approaches. The rest of the day was devoted to oral paper presentations and Daniel Marcu's invited talk in the afternoon.

Like in previous years, we have received a far larger number of submission than we could accept for presentation. This year we have received 18 full paper submissions and 26 shared task submissions. In total WMT-08 featured 12 full paper oral presentations and 25 shared task poster presentations. The invited talk was given by Daniel Marcu of the Information Sciences Institute at the University of Southern California.

We would like to thank the members of the Program Committee for their timely reviews. We also would like to thank the participants of the shared task and all the other volunteers who helped with the manual evaluations. We also acknowledge the financial support of the shared task by the EuroMatrix project funded by the European Commission (6th Framework Programme).

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce

Co-Organizers

Organizers:

Chris Callison-Burch (Johns Hopkins University)
Philipp Koehn (University of Edinburgh)
Christof Monz (University of London)
Josh Schroeder (University of Edinburgh)
Cameron Shaw Fordyce

Program Committee:

Lars Ahrenberg (Linköping University)
Yaser Al-Onaizan (IBM Research)
Oliver Bender (RWTH Aachen)
Chris Brockett (Microsoft Research)
Bill Byrne (Cambridge University)
Francisco Casacuberta (University of Valencia)
Colin Cherry (Microsoft Research)
Stephen Clark (Oxford University)
Trevor Cohn (Edinburgh University)
Mona Diab (Columbia University)
Hal Daume (University of Utah)
Chris Dyer (University of Maryland)
Andreas Eisele (University Saarbrücken)
Marcello Federico (ITC-IRST)
George Foster (Canada National Research Council)
Alex Fraser (University of Stuttgart)
Ulrich Germann (University of Toronto)
Nizar Habash (Columbia University)
Jan Hajic (Charles University)
Keith Hall (Google)
John Henderson (MITRE)
Rebecca Hwa (University of Pittsburgh)
Doug Jones (Lincoln Labs MIT)
Damianos Karakos (Johns Hopkins University)
Kevin Knight (ISI/University of Southern California)
Shankar Kumar (Google)
Philippe Langlais (University of Montreal)
Alon Lavie (Carnegie Melon University)
Adam Lopez (Edinburgh University)
Daniel Marcu (ISI/University of Southern California)
Lambert Mathias (Johns Hopkins University)
Arul Menezes (Microsoft Research)
Bob Moore (Microsoft Research)

Miles Osborne (University of Edinburgh)
Kay Peterson (NIST)
Mark Przybocki (NIST)
Chris Quirk (Microsoft Research)
Philip Resnik (University of Maryland)
Michel Simard (National Research Council Canada)
Libin Shen (BBN Technologies)
Wade Shen (Lincoln Labs MIT)
Eiichiro Sumita (NICT/ATR)
David Talbot (Edinburgh University)
Jörg Tiedemann (University of Groningen)
Christoph Tillmann (IBM Research)
Kristina Toutanova (Microsoft Research)
Nicola Ueffing (National Research Council Canada)
Clare Voss (Army Research Labs)
Taro Watanabe (NTT)
Dekai Wu (HKUST)
Richard Zens (Google)

Additional Reviewers:

Mahmoud Ghoneim
Jeffrey Micher

Invited Speaker:

Daniel Marcu (ISI/University of Southern California)

Table of Contents

Conference Program

Thursday, June 19, 2008

8:40–8:50	Opening Remarks
	Session 1: Full Papers
8:50–9:10	*An Empirical Study in Source Word Deletion for Phrase-Based Statistical Machine Translation* Chi-Ho Li, Hailei Zhang, Dongdong Zhang, Mu Li and Ming Zhou
9:10–9:30	*Rich Source-Side Context for Statistical Machine Translation* Kevin Gimpel and Noah A. Smith
9:30–9:50	*Discriminative Word Alignment via Alignment Matrix Modeling* Jan Niehues and Stephan Vogel
9:50–10:10	*Regularization and Search for Minimum Error Rate Training* Daniel Cer, Daniel Jurafsky and Christopher Manning
10:10–10:30	*Learning Performance of a Machine Translation System: a Statistical and Computational Analysis* Marco Turchi, Tijl De Bie and Nello Cristianini
10:30–11:00	Coffee Break
	Session 2: Full Papers
11:00–11:20	*Using Syntax to Improve Word Alignment Precision for Syntax-Based Machine Translation* Victoria Fossum, Kevin Knight and Steven Abney
11:20–11:40	*Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT* Josep M. Crego and Nizar Habash
11:40–12:00	*Improved Tree-to-String Transducer for Machine Translation* Ding Liu and Daniel Gildea
12:00-12:40	Invited Talk by Daniel Marcu

Thursday, June 19, 2008 (continued)

Thursday, June 19, 2008 (continued)

Thursday, June 19, 2008 (continued)

Shared Evaluation Task

The Role of Pseudo References in MT Evaluation
Joshua Albrecht and Rebecca Hwa

Ranking vs. Regression in Machine Translation Evaluation
Kevin Duh

A Smorgasbord of Features for Automatic MT Evaluation
Jesus Gimenez and Lluis Marquez

3:30-4:40 Coffee Break and Poster Session

Session 4: Full Papers

4:40–5:00 *Fast, Easy, and Cheap: Construction of Statistical Machine Translation Models with MapReduce*
Chris Dyer, Aaron Cordova, Alex Mont and Jimmy Lin

5:00–5:20 *Dynamic Model Interpolation for Statistical Machine Translation*
Andrew Finch and Eiichiro Sumita

5:20–5:40 *Improved Statistical Machine Translation by Multiple Chinese Word Segmentation*
Ruiqiang Zhang, Keiji Yasuda and Eiichiro Sumita

5:40–6:00 *Optimizing Chinese Word Segmentation for Machine Translation Performance*
Pi-Chuan Chang, Michel Galley and Christopher Manning