NAACL-HLT 2012

The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

Proceedings of the Workshop on Evaluation Metrics and System Comparison for Automatic Summarization

June 8, 2012
Montréal, Canada

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

ISBN 978-1-937284-20-6 / 1-937284-20-4

Introduction

Welcome to the NAACL/HLT 2012 Workshop on Evaluation Metrics and System Comparison for Automatic Summarization. One of the goals of the workshop is to give a retrospective analysis of evaluation methods employed at the Text Analysis Conferences (TAC) and its predecessor, the Document Understanding Conferences (DUC). The other goal is to set plans for the future as we introduce the new task of summarization of scientific articles.

We have planned two invited presentations. Dragomir Radev will talk about his own work on summarization of scientific articles, as well as provide us with some background on related work. Lucy Vanderwende will present the plans for the new summarization task, the evaluation and the time line for future shared tasks. We have reserved ample time for discussion.

In the six regular presentations we will discuss a range of exciting topics in summarization and evaluation. These include task-based evaluations of summarization, assessments of the accuracy of current automatic evaluations, the benefits from using several automatic evaluation measures, case studies of differences between manual and automatic evaluation, cross-lingual summarization and steps towards abstractive summarization.

We anticipate a lively and rewarding workshop. Thank you for your participation!

John M. Conroy
Hoa Trang Dang
Ani Nenkova
Karolina Owczarzak

Organizers:

John M. Conroy, IDA Center for Computing Sciences
Hoa Trang Dang, National Institute for Standards and Technology
Ani Nenkova, University of Pennsylvania
Karolina Owczarzak, National Institute for Standards and Technology

Program Committee:

Enrique Amigo (UNED, Madrid)
Giuseppe Carenini (University of British Columbia)
Katja Filippova (Google Research)
George Giannakopoulos (NCSR Demokritos)
Dan Gillick (University of California at Berkeley)
Min-Yen Kan (National University of Singapore)
Guy Lapalme (University of Montreal)
Yang Liu (University of Texas, Dallas)
Annie Louis (University of Pennsylvania)
Kathy McKeown (Columbia University)
Gabriel Murray (University of British Columbia)
Dianne O'Leary (University of Maryland)
Drago Radev (University of Michigan)
Steve Renals (University of Edinburgh)
Horacio Saggion (Universitat Pompeu Fabra)
Judith Schlesinger (IDA Center for Computing Sciences)
Josef Steinberger (European Commission Joint Research Centre)
Stan Szpakowicz (University of Ottawa)
Lucy Vanderwende (Microsoft Research)
Stephen Wan (CSIRO ICT Centre)
Xiaodan Zhu (National Research Council Canada)

Invited Speakers:

Drago Radev (University of Michigan)
Lucy Vanderwende (Microsoft Research)

Table of Contents

An Assessment of the Accuracy of Automatic Evaluation in Summarization
Karolina Owczarzak, John M. Conroy, Hoa Trang Dang and Ani Nenkova 1

Using the Omega Index for Evaluating Abstractive Community Detection
Gabriel Murray, Giuseppe Carenini and Raymond Ng 10

Machine Translation for Multilingual Summary Content Evaluation
Josef Steinberger and Marco Turchi 19

Ecological Validity and the Evaluation of Speech Summarization Quality
Anthony McCallum, Cosmin Munteanu, Gerald Penn and Xiaodan Zhu 28

The Heterogeneity Principle in Evaluation Measures for Automatic Summarization
Enrique Amigó, Julio Gonzalo and Felisa Verdejo 36

Discrepancy Between Automatic and Manual Evaluation of Summaries
Shamima Mithun, Leila Kosseim and Prasad Perera 44

Conference Program

8:50AM Opening remarks

9:00AM *An Assessment of the Accuracy of Automatic Evaluation in Summarization*
Karolina Owczarzak, John M. Conroy, Hoa Trang Dang and Ani Nenkova

9:30AM *Using the Omega Index for Evaluating Abstractive Community Detection*
Gabriel Murray, Giuseppe Carenini and Raymond Ng

10:00AM *Machine Translation for Multilingual Summary Content Evaluation*
Josef Steinberger and Marco Turchi

10:30AM BREAK

11:00AM A new pilot task: summarization of academic articles by Lucy Vanderwende

11:40AM Discussion and planning for the pilot task

12:30PM LUNCH

2:30PM Generation of surveys of research areas by Dragomir Radev

3:30PM BREAK

4:00PM *Ecological Validity and the Evaluation of Speech Summarization Quality*
Anthony McCallum, Cosmin Munteanu, Gerald Penn and Xiaodan Zhu

4:30PM *The Heterogeneity Principle in Evaluation Measures for Automatic Summarization*
Enrique Amigó, Julio Gonzalo and Felisa Verdejo

5:00PM *Discrepancy Between Automatic and Manual Evaluation of Summaries*
Shamima Mithun, Leila Kosseim and Prasad Perera