Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization

Proceedings of the ACL-05 Workshop

Organizing Committee
Jade Goldstein, US Department of Defense
Alon Lavie, CMU Language Technologies Institute
Chin-Yew Lin, USC Information Sciences Institute
Clare Voss, US Army Research Laboratory

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

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

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

Preface

This workshop is the first meeting to focus on the challenges that the machine translation (MT) and summarization communities face in developing valid and useful evaluation measures. Our aim is to bring these two communities together to learn from each other's approaches.

Prior ACL workshops on evaluation have had as their central focus a core computational task (e.g., word sense disambiguation, parsing), a genre (e.g., dialogue, multi-modal interfaces), a computational technique (e.g., unsupervised learning, finite state models), a resource (e.g., parallel texts, WordNet), or a process (e.g., reading comprehension, question-answering). This workshop, in clear contrast, has as its central focus the examination of evaluation measures, or "meta-evaluation" as Dan Melamed has noted.

The initial impetus for this workshop came at the biennial meeting of the Association for Machine Translation in the Americas (AMTA) held at Georgetown University in September 2004, when the following question arose in a discussion session: "Why isn't recall a part of MT evaluation the way that it is for summarization evaluation?" Several of us continued this discussion afterwards and proposed to convene together again more formally to address this question and other evaluation challenges that both the MT and summarization communities have been tackling.

We wish to thank Bonnie Dorr and Ed Hovy, in particular, for their encouragement and contributions in shaping the initial workshop proposal and the subsequent call for papers. Boyan Onyshkevych, Barb Wheatley, Donna Harmon, and Judith Klavans also provided insightful comments in the proposal writing phase of the workshop that helped guide and focus the topics we chose to address.

We also would like also to thank several others. Charles Wayne, Joe Olive, Donna Harmon, Hoa Dang, Lori Buckland, and Chris Cieri were critical in helping make the datasets available to workshop participants. Jason Eisner and Philipp Koehn, ACL publications chairs, provided us invaluable assistance in preparing the proceedings.

Many thanks also to the Program Committee and additional reviewers who graciously spent time with a short schedule to review submitted papers and provide valuable feedback. We have an exciting program for which we thank the many authors who submitted their research papers.

Jade Goldstein, Alon Lavie, Chin-Yew Lin, Clare Voss
June 2005

Excerpts from Call for Papers

This one-day workshop will focus on the challenges that the MT and summarization communities face in developing valid and useful evaluation measures. Our aim is to bring these two communities together to learn from each other's approaches.

In the past few years, we have witnessed—in both MT and summarization evaluation—the innovation of ngram-based intrinsic metrics that automatically score system-outputs against human-produced reference documents (e.g., IBM's BLEU and ISI/USC's counterpart ROUGE). Similarly, there has been renewed interest in user applications and task-based extrinsic measures in both communities (e.g., DUC'05 and TIDES'04). Most recently, evaluation efforts have tested for correlations to cross-validate independently derived intrinsic and extrinsic assessments of system-outputs with each other and with human judgments on output, such as accuracy and fluency.

The concrete questions that we hope to see addressed in this workshop include, but are not limited to:

- How adequately do intrinsic measures capture the variation between system-outputs and human-generated reference documents (summaries or translations)? What methods exist for calibrating and controlling the variation in linguistic complexity and content differences in input test-sets and reference sets? How much variation exists within these constructed sets? How does that variation affect different intrinsic measures? How many reference documents are needed for effective scoring?
- How can intrinsic measures go beyond simple n-gram matching, to quantify the similarity between system-output and human-references? What other features and weighting alternatives lead to better metrics for both MT and summarization? How can intrinsic measures capture fluency and adequacy? Which types of new intrinsic metrics are needed to adequately evaluate non-extractive summaries and paraphrasing (e.g.,interlingual) translations?
- How effectively do extrinsic (or proxy extrinsic) measures capture the quality of system output, as needed for downstream use in human tasks, such as triage (document relevance judgments), extraction (factual question answering), and report writing; and in automated tasks, such as filtering, information extraction, and question-answering? For example, when is an MT system good enough that a summarization system benefits from the additional information available in the MT output?
- How should metrics for MT and summarization be assessed and compared? What characteristics should a good metric possess? When is one evaluation method better than another? What are the most effective ways of assessing the correlation testing and statistical modeling that seek to predict human task performance or human notions of output quality (e.g., fluency and adequacy) from "cheaper" automatic metrics? How reliable are human judgments?

Anyone with an interest in MT or summarization evaluation research or in issues pertaining to the combination of MT and summarization is encouraged to participate in the workshop. We are looking for research papers on the aforementioned topics, as well as position papers that identify limitations in current approaches and describe promising future research directions.

To faciliate the comparison of different measures during the workshop, we will be making available data sets in advance for workshop participants to test their approaches to evaluation. Although the shared data sets are separated, we would encourage participants to apply their automatic metrics on both data sets and report comparative results in the workshop.

Shared Data Sets

Shared Data Set for MT Evaluation:

The shared data set consists of the 2003 TIDES MT-Eval Test Data for both Chinese-to-English and Arabic-to-English MT. For each of these two language-pair data sets, the following is provided:

- The set of test sentences in the original source language (Chinese or Arabic)
- MT system output for the set of sentences for 7 different MT systems
- A collection of 4 reference translations (human translated) into English
- Human judgments of MT quality (adequacy and fluency) for the various MT system translations of every sentence. Each sentence was judged by two subjects, each of which assigned both an adequacy score and a fluency score, in the integer range of [1-5].

Shared Data Set for Summarization Evaluation:

The summarization shared data set consists of four years' worth of data from past Document Understanding Conferences (DUC) including:

- Documents
- Summaries, results, etc.
  - Manually created summaries
  - Automatically created baseline summaries
  - Submitted summaries created by the participating groups' systems
  - Tables with the evaluation results
  - Additional supporting data and software

Organizers:

Jade Goldstein	US Department of Defense, USA
Alon Lavie	CMU Language Technologies Institute, USA
Chin-Yew Lin	USC Information Sciences Institute, USA
Clare Voss	Army Research Laboratory, USA

Program Committee:

Yasuhiro Akiba	ATR, Japan
Leslie Barrett	TransClick, USA
Bonnie Dorr	University of Maryland, USA
Tony Hartley	University of Leeds, UK
John Henderson	MITRE, USA
Chiori Hori	CMU Language Technologies Institue, USA
Eduard Hovy	USC Information Sciences Insititute, USA
Doug Jones	MIT Lincoln Laboratory, USA
Philipp Koehn	University of Edinburgh, UK
Marie-Francine Moens	Katholieke Universiteit, Leuven, Belgium
Hermann Ney	RWTH Aachen, Germany
Franz Och	Google, USA
Rebecca Passonneau	Columbia University, USA
Andrei Popescu-Belis	University of Geneva ISSCO/TIM/ETI, Switzerland
Dragomir Radev	University Michigan, USA
Karen Sparck Jones	University of Cambridge Computer Laboratory, UK
Simone Teufel	University of Cambridge Computer Laboratory, UK
Nicola Ueffing	RWTH Aachen, Germany
Hans van Halteren	University of Nijmegen, The Netherlands
Michelle Vanni	Army Research Laboratory, USA
Dekai Wu	HKUST, Hong Kong

Additional Reviewers:

Chad Langley	US Department of Defense, USA
Gregor Leusch	RWTH Aachen, Germany
Klaus Macherey	Google,USA
Wolfang Macherey	Google,USA
Christof Monz	University of Maryland, USA
Judith Schlesinger	IDA Center for Computing Sciences, USA

Table of Contents

Workshop Program

Wednesday, June 29, 2005

8:45–8:50 Opening Remarks

Session 1: Summarization Metrics I

8:50–9:15 *A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?*
Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz and David Zajic

9:15–9:40 *On the Subjectivity of Human Authored Summaries*
BalaKrishna Kolluru and Yoshihiko Gotoh

Session 2: MT Metrics I

9:40–10:05 *Preprocessing and Normalization for Automatic Evaluation of Machine Translation*
Gregor Leusch, Nicola Ueffing, David Vilar and Hermann Ney

10:05–10:30 *Syntactic Features for Evaluation of Machine Translation*
Ding Liu and Daniel Gildea

10:30–11:00 Break

Session 3: Invited Talk

11:00–12:00 *Results of the Multilingual Summarization Evaluation*, Kathy McKeown

Session 4: Student Session - Work in Progress

12:00–12:15 *Evaluation of Sentence Selection on Spoken Dialogue*, Xiaodan Zhu

12:15–12:30 *Toward a Predictive Statistical Model of Task-based Performance*, Calandra R. Tate

12:30–2:15 Lunch

Session 5: Summarization Metrics II

2:15–2:40 *Evaluating Automatic Summaries of Meeting Recordings*
Gabriel Murray, Steve Renals, Jean Carletta and Johanna Moore

2:40–3:05 *Evaluating Summaries and Answers: Two Sides of the Same Coin?*
Jimmy Lin and Dina Demner-Fushman

3:05–3:30 *Evaluating DUC 2004 Tasks with the QARLA Framework*
Enrique Amigó, Julio Gonzalo, Anselmo Peñas and Felisa Verdejo

Session 6: MT Metrics II

4:00–4:25 *On Some Pitfalls in Automatic Evaluation and Significance Testing for MT*
Stefan Riezler and John T. Maxwell

4:25–4:50 *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*
Satanjeev Banerjee and Alon Lavie

Session 7: Panel Discussion and Open Forum on Future Plans

4:50–5:50 Panel Discussion

5:50–6:00 Future Plans