2006

COLING • ACL

COLING • ACL 2006

Task-Focused Summarization and Question Answering

Proceedings of the Workshop

Chairs:
Tat-Seng Chua, Jade Goldstein,
Simone Teufel and Lucy Vanderwende

23 July 2006
Sydney, Australia

Production and Manufacturing by
BPA Digital
11 Evans St
Burwood VIC 3125
AUSTRALIA

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 1-932432-79-5

Table of Contents

Preface

The *Task-Focused Summarization and Question Answering* workshop, to be held on July 23, 2006 in Sydney, aims to bring together the two communities of summarization and question answering by examining how to create output that is directed to a user's needs, i.e., how to create task-focused output. The user scenarios that are described in the accepted papers include the medical and computer domain, readers of short stories and also traditional multidocument news collections, some with interesting, and different, evaluation methodologies. By focusing on the benefits that summarization and question answering can have for users, we hope to contribute to the discussion of the evaluation in both areas.

We included the call for papers in these proceedings. Of the fourteen papers submitted, we accepted seven to be presented at the workshop. We want to thank all the members of the program committee for their thoughtful and in depth reviews. All the reviews were completed on time, despite very tight deadlines.

We wanted to invite a speaker who is deeply involved in both the question answering and summarization communities and who can help bring the communities further together. We thank Sanda Harabagiu for her talk, as well as for her support in this area. Furthermore, we hope that convening a panel to bring together researchers engaged in evaluation of summarization and question answering from around the world will increase our understanding of the current state of the art in evaluation and provide opportunties to share our understanding.

Finally, we thank the workshop participants for sharing their current work at this workshop, and for sharing with us their views on the utility of summarization and question answering to users' needs.

Tat-Seng Chua, Jade Goldstein, Simone Teufel, Lucy Vanderwende

Excerpts from Call for Papers

This one-day workshop will focus on the challenges that the Summarization and QA communities face in developing useful systems and in developing evaluation measures. Our aim is to bring these two communities together to discuss the current challenges and to learn from each other's approaches, following the success of a similar workshop held at ACL-05, which brought together the Machine Translation and Summarization communities.

A previous summarization workshop (*Text Summarization Branches Out*, ACL-04) targeted the exploration of different scenarios for summarization, such as small mobile devices, legal texts, speech, dialog, email and other genres. We encourage a deeper analysis of these, and other, user scenarios, focusing on the utility of summarization and question answering for such scenarios and genres, including cross-lingual ones.

By focusing on the measurable benefits that summarization and question answering has for users, we hope one of the outcomes of this workshop will be to better motivate research and focus areas for summarization and question answering, and to establish task-appropriate evaluation methods. Given a user scenario, it would ideally be possible to demonstrate that a given evaluation method predicts greater/lesser utility for users. We especially encourage papers describing intrinsic and extrinsic evaluation metrics in the context of these user scenarios.

Both summarization and QA have a long history of evaluations: Summarization since 1998 (SUMMAC) and QA since 1999 (TREC). The importance of summarization evaluation is evidenced by the many DUC workshops; in DUC-05, extensive discussions were held regarding the use of ROUGE, ROUGE-BE, and the pyramid method, a semantic-unit based approach, for evaluating summarization systems. The QA community has related evaluation issues for answers to complex questions such as the TREC definition questions. Some common considerations in both communities include what constitutes a good answer/response to an information request, and how does one determine whether a "complex" answer is sufficient? In both communities, as well as in the distillation component of the 2005 DARPA program GALE, researchers are exploring how to capture semantic equivalence among components of different answers (nuggets, factoids or SCUs). There also have been efforts to design new automatic scoring measures, such as ROUGE-BE and POURPRE. We encourage papers discussing these and other metrics that report on how well the metric correlates with human judgments and/or predicts effectiveness in task-focused scenarios for summarization and QA.

This workshop is a continuation of ACL 2005 for the summarization community, In which those interested in evaluation measures participated in a joint Workshop on evaluation for summarization and MT. As a sequel to the ACL 2005 workshop, in which the results of the first Multilingual multidocument summarization evaluation (MSE) were presented, we plan to report and discuss the results of the 2006 MSE evaluation.

In summary, we solicit papers on any or all of the following three topics:

- Task-based user scenarios requiring question answering (beyond factoids/lists) and/or summarization, across genres and languages
- Extrinsic and intrinsic evaluations, correlating extrinsic measures with outcome of task completion and/or intrinsic measures with human judgments previously obtained.
- The 2006 Multilingual Multidocument Summarization Evaluation

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

Multilingual Summarization Evaluation 2006

The 2nd Multilingual Summarization Evaluation will be held in conjunction with the COLING//ACL 2006 Workshop *Task-Focused Summarization and Question Answering* and the results of the evaluation will be reported during the COLING/ACL Workshop. This evaluation repeats the first Multilingual Summarization Evaluation held in 2005 as part of the ACL workshop *Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Task Description:

Given a cluster of documents on the same event, some in English, some translated from Arabic (Arabic source is also available), generate a 100-word summary of the event. Clusters contain on average 10 documents per cluster. The distribution between Arabic and English varies between clusters.

Data:

25 clusters from the Multilingual Summarziation Evaluation 2005 are available for training, as well as the DUC2004 data for Task 4, available at http://duc.nist.gov.

25 clusters will be used for testing. These clusters were created by running a clustering algorithm developed by Columbia over the the TDT4 corpus, which contains 41,728 Arabic documents and 23,602 English documents. ISI's MT system was used to translate the Arabic data. Both source and translation are available in the cluster. Human annotators at the LDC sorted through the automatically created clusters, to select 50 (25 clusters for last year, 25 clusters for this year) that were good to use, editing the clusters as needed. Four humans wrote a 100-word summary for each cluster. Thus, there are 4 model summaries per cluster.

Organizers:

Jade Goldstein, US Department of Defense
Lucy Vanderwende, Microsoft Research
Liang Zhou, USC/ISI

Participants:

Lehmam Abderrafih, Pertinence Mining
John M. Conroy, Dianne P. O'Leary, Judith D. Schlesinger, IDA/CCS and University of Maryland Angelo Dalli, University of Sheffield
David Kirk Evans, Japanese National Institute of Information
Maher Jaoua, MIRACL Laboratory for Computer Sciences, University of Sfax, Tunisia
Wenjie Li, Department of Computing, The Hong Kong Polytechnic University
Prasad Pingali, Jagadeesh J, Vasudeva Varma, IIIT, Hyderabad
Wei Xu, Tsinghua University, Beijing, China
David Zajic, University of Maryland and BBN Technologies (UMD/BBN)

Organizers

Chairs:

Tat-Seng Chua, National University of Singapore (Singapore)
Jade Goldstein, US Department of Defense (USA)
Simone Teufel, Cambridge University (UK)
Lucy Vanderwende, Microsoft Research (USA)

Program Committee:

Regina Barzilay, MIT (USA)
Sabine Bergler, Concordia University (Canada)
Silviu Cucerzan, Microsoft Research (USA)
Hang Cui, National University of Singapore (Singapore)
Krzysztof Czuba, Google (USA)
Hal Daume III, USC/ISI (USA)
Hans van Halteren, Radboud University, Nijmegen (Netherlands)
Sanda Harabagiu, University of Texas, Dallas (USA)
Chiori Hori, Carnegie Mellon University (USA)
Eduard Hovy, USC/ISI (USA)
Hongyan Jing, IBM Research (USA)
Guy Lapalme, University of Montreal (Canada)
Geunbae (Gary) Lee, Postech Univ (Korea)
Chin-Yew Lin, Microsoft Research Asia (China)
Inderjeet Mani, MITRE (USA)
Marie-France Moens, Katholieke Universiteit Leuven (Belgium)
Ani Nenkova, Columbia University (USA)
Manabu Okumura, Tokyo Institute of Technology (Japan)
John Prager, IBM Research (USA)
Horacio Saggion, University of Sheffield (UK)
Judith Schlesinger, IDA/CCS (USA)
Karen Sparck Jones, University of Cambridge (UK)
Nicola Stokes, University of Melbourne (Australia)
Beth Sundheim, SPAWAR Systems Center (USA)
Tomek Strzalkowski, University at Albany (USA)
Ralph Weischedel, BBN (USA)

Invited Speaker:

Sanda Harabagiu, Language Computer Corporation (USA)

Panelists:

Hoa Trang Dang, NIST (USA)
Eduard Hovy, USC/ISI (USA)
Noriko Kando, NTCIR (Japan)

Scenario Based Question Answering

Sanda Harabagiu

When faced with a task described by a complex scenario, users ask questions that are motivated by the need to explore complex relationships. These questions test the capabilities of Q/A systems to (1) tackle complex requests; (2) take into account the scenario context; and (3) enable a coherent dialogue with the user.

In this talk I shall describe our experience with Ferret, our interactive Q/A system, within several experiments that involved multiple scenarios and a varied number of users. I shall present the lessons learned and focus on the most challenging problems.

Workshop Program

Sunday, 23 July 2006

8:30–8:40 Opening Remarks

Session 1: Summarization

8:40–9:05 *Dimensionality Reduction Aids Term Co-Occurrence Based Multi-Document Summarization*
Ben Hachey, Gabriel Murray and David Reitter

9:05–9:30 *Challenges in Evaluating Summaries of Short Stories*
Anna Kazantseva and Stan Szpakowicz

Session 2: Invited Talk

9:30–10:30 Invited Talk, *Scenario-based Question Answering* by Sanda Harabagiu

10:30-10:50 Break

Session 3: Question Answering

10:50–11:15 *Question Pre-Processing in a QA System on Internet Discussion Groups*
Chuan-Jie Lin and Chun-Hung Cho

11:15–11:40 *Situated Question Answering in the Clinical Domain: Selecting the Best Drug Treatment for Diseases*
Dina Demner-Fushman and Jimmy Lin

11:40–12:05 *Using Scenario Knowledge in Automatic Question Answering*
Sanda Harabagiu and Andrew Hickl

12:05–12:30 *Automating Help-desk Responses: A Comparative Study of Information-gathering Approaches*
Yuval Marom and Ingrid Zukerman

12:30-2:00 Lunch

Session 4: Evaluation

2:00–2:25 *DUC 2005: Evaluation of Question-Focused Summarization Systems*
Hoa Trang Dang

2:25–3:30 Panel, Evaluation Programs: Hoa Trang Dang, Eduard Hovy, Noriko Kando

3:30–4:00 Break

Session 5: Multilingual Summarization Evaluation (MSE) 2006

4:00-4:30 *Overview of MSE and Evaluation Results* by John Conroy

4:30-5:10 Invited Presentations from MSE participants

5:10-5:30 Discussion of Multilingual Summarization