ACL 2012

50th Annual Meeting of the
Association for Computational Linguistics

Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries

July 10, 2012
Jeju Island, Korea

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 978-1-937284-29-9

Preface

Fifty years of Computational Linguistics are nothing when compared to the long history of human language. However, those same fifty years of Computational Linguistics constitute a lot in terms of achievements and advances towards better understanding one of the most natural yet complex human phenomena. Fifty years of Computational Linguistics are, indeed, a lifetime of endeavours, successes and failures, but they just represent the very beginning of an interesting journey over a vast sea of undiscovered knowledge.

In this first 50th anniversary of the Association for Computational Linguistics, we just take this brief pause to review our history and project our future, to ensure that our current legacy will endure the indifference of time and that future generations can step on the shoulders of the many pioneers of this new wonderful discipline, in which language is clearly showing its reluctance to being tamed by mathematics.

Welcome to the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries!

Rafael E. Banchs
Jeju, Korea, July 10th, 2012

Workshop Objectives

This workshop is intended to commemorate the 50th anniversary of the Association for Computational Linguistics (ACL) by creating a new space for debating and discussing about specific issues related to preserving, analysing and exploiting the scientific heritage of the ACL, as well as to envisage future trends, applications and research in Computational Linguistics.

The main objective of the workshop has been to gather contributions about the history, the evolution and the future of research in Computational Linguistics. Although the call for papers was open to any kind of technical contribution that was relevant to the main objective of the workshop, we specially encouraged the submission of research work related to the application of natural language processing and text mining techniques to the ACL Anthology Reference Corpus (ACL ARC), which is publicly available from the ACL ARC project website.

In addition to the technical program, the workshop introduces a new contributed task, in the spirit of a crowd-sourcing activity, for augmenting and improving the current status of the ACL Anthology Reference Corpus. The goal of the contributed task is to provide a high quality version of the textual content of the ACL Anthology as a corpus. Besides the more accurate text extraction, the rich text markup can be also an important source of information for corpus-based applications such as summarization, scientific discourse analysis, citation analysis, citation classification, question answering, textual entailment, taxonomy, ontology, information extraction, parsing, coreference resolution, semantic search and many more.

Acknowledgments

This special workshop has been possible thanks to the effort of many people...

Special thanks to the Steering Committee and the Contributed Task Committee for their timely advice and suggestions.

Special thanks also to all Program Committee members for their devoted work and recommendations during the peer reviewing process, as well as to Publicity and Technical Committee members for their invaluable help during the workshop planning and preparation.

Finally, and most important of all; special thanks to all authors and co-authors who had contributed with their work to put together such an interesting technical program.

Workshop Organizer:

Rafael E. Banchs, Institute for Infocomm Research (Singapore)

Steering Committee:

Steven Bird, University of Melbourne (Australia)
Robert Dale, Macquarie University (Australia)
Min Yen Kan, National University of Singapore (Singapore)
Haizhou Li, Institute for Infocomm Research (Singapore)
Dragomir Radev, University of Michigan (USA)
Ulrich Schäfer, German Research Center for Artificial Intelligence (Germany)

Contributed Task Committee:

Jonathon Read, University of Oslo (Norway)
Stephan Oepen, University of Oslo (Norway)
Ulrich Schäfer, German Research Center for Artificial Intelligence (Germany)
Tze Yuang Chong, Nanyang Technological University (Singapore)

Program Committee:

Toni Badia, Barcelona Media Innovation Centre (Spain)
Timothy Baldwin, University of Melbourne (Australia)
Sivaji Bandyopadhyay, Jadavpur University (India)
Emily M. Bender, University of Washington (USA)
Kenneth Church, Johns Hopkins University (USA)
Marta R. Costa-jussa, Barcelona Media Innovation Centre (Spain)
Iryna Gurevych, Technische Universität Darmstadt (Germany)
Carlos Henriquez, Universitat Politecnica de Catalunya (Spain)
Daniel Jurafsky, Stanford University (USA)
Min Yen Kan, National University of Singapore (Singapore)
Kevin Knight, Information Sciences Institute (USA)
Philipp Koehn, University of Edinburgh (UK)
Haizhou Li, Institute for Infocomm Research (Singapore)
Bing Liu, University of Illinois at Chicago (USA)
Yang Liu, National University of Singapore (Singapore)
Yuji Matsumoto, Nara Institute of Science and Technology (Japan)
Kathleen McKeown, Columbia University (USA)
Rada Mihalcea, University of North Texas (USA)
Hwee Tou Ng, National University of Singapore (Singapore)
Joakim Nivre, Uppsala University (Sweden)
Stephan Oepen, University of Oslo (Norway)
Dragomir Radev, University of Michigan (USA)

Jonathon Read, University of Oslo (Norway)
Paolo Rosso, Universidad Politecnica de Valencia (Spain)
Horacio Saggion, Universitat Pompeu Fabra (Spain)
Ulrich Schäfer, German Research Center for Artificial Intelligence (Germany)
Fabrizio Silvestri, Instituto di Scienza e Tecnologie dell'Informazione (Italy)
Eiichiro Sumita, National Institute of Information and Communications Technology (Japan)
Simone Teufel, University of Cambridge (UK)
Junichi Tsujii, Microsoft Research Asia (China)
Anita de Waard, Elsevier Labs (The Netherlands)
Haifeng Wang, Baidu (China)
Magdalena Wolska, Saarland University (Germany)
Deyi Xiong, Institute for Infocomm Research (Singapore)
Min Zhang, Institute for Infocomm Research (Singapore)
Ming Zhou, Microsoft Research Asia (China)

Publicity Committee:

Marta R. Costa-jussa, Barcelona Media Innovation Centre (Spain)
Seokhwan Kim, Institute for Infocomm Research (Singapore)

Technical Committee:

Ming Liu, Institute for Infocomm Research (Singapore)

Table of Contents

Special Workshop Program

Tuesday, July 10th

(11:00-12:30) Session 1: The People (ACL Session 4a)

11:00 *Rediscovering ACL Discoveries Through the Lens of ACL Anthology Network Citing Sentences*
Dragomir Radev and Amjad Abu-Jbara

11:20 *Towards a Computational History of the ACL: 1980-2008*
Ashton Anderson, Dan Jurafsky and Daniel A. McFarland

11:40 *Discovering Factions in the Computational Linguistics Community*
Yanchuan Sim, Noah A. Smith and David A. Smith

12:00 *He Said, She Said: Gender in the ACL Anthology*
Adam Vogel and Dan Jurafsky

(14:00-15:30) Session 2: The Contents (ACL Session 5a)

14:00 *Discourse Structure and Computation: Past, Present and Future*
Bonnie Webber and Aravind Joshi

14:20 *Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis*
Melanie Reiplinger, Ulrich Schäfer and Magdalena Wolska

14:40 *Applying Collocation Segmentation to the ACL Anthology Reference Corpus*
Vidas Daudaravicius

15:00 *Text Reuse with ACL: (Upward) Trends*
Parth Gupta and Paolo Rosso

Tuesday, July 10th (continued)

(16:00-17:30) Session 3: The Anthology (ACL Session 6a)

16:00 *Integrating User-Generated Content in the ACL Anthology*
Praveen Bysani and Min-Yen Kan

16:20 *Towards an ACL Anthology Corpus with Logical Document Structure. An Overview of the ACL 2012 Contributed Task*
Ulrich Schäfer, Jonathon Read and Stephan Oepen

16:40 *Towards High-Quality Text Stream Extraction from PDF. Technical Background to the ACL 2012 Contributed Task*
Øyvind Raddum Berg, Stephan Oepen and Jonathon Read

17:00 *Combining OCR Outputs for Logical Document Structure Markup. Technical Background to the ACL 2012 Contributed Task*
Ulrich Schäfer and Benjamin Weitz

17:20 *Linking Citations to their Bibliographic references*
Huy Do Hoang Nhat and Praveen Bysani