2006

COLING • ACL

COLING • ACL 2006

CLIIR
How Can Computational Linguistics Improve Information Retrieval?

Proceedings of the Workshop

Chairs:
John Tait and Michael Oakes

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 1-932432-80-9

Table of Contents

Preface

There has been a long standing interest in using various forms of deep natural language processing to improve information or document retrieval. We have a Cambridge Language Research (CLRU) memo from 1964 by Yorick Wilks which describes an application to text searching of a clear precursor of his later well-known machine translation system. We are also aware of even earlier work in the CLRU on information retrieval by Karen Sparck Jones and Margaret Masterman.

This interest has continued right up to the present day, but successes have been few and far between. In general search engines are based on statistical modeling of documents which lacks at least transparent and visible knowledge of language in any conventional sense. Although many continue to believe search engines which do not, for example, recognise that words have multiple senses, cannot do a good job of the task of matching queries and documents, the fact is that most of the time most users of Google find enough relevant documents in the first page or two of hits without such linguistic sophistication.

Computational Linguistics has progressed enormously in the past few years. CL has made significant contributions to the specialised areas of information retrieval, most notably question answering. However, the dominant use model for information retrieval remains the classic search engine task, in which a short key word query is used to generate a ranked list from a pre-indexed heterogeneous collection of documents, and very little work from computational linguistics has been used in the development of these engines.

This workshop will provide a forum to discuss why this is the case, and how to achieve a better take up of what computational linguistic technology within the search engine community.

We would like to thank our two invited speakers, Jamie Callan and Cécile Paris, in particular Jamie who traveled from the US to Australia especially to take part in the workshop, all the authors (whether their papers were accepted or not) and our program committee. The workshop could not have happened without your efforts!

We would like to acknowledge the kind sponsorship of the Cambridge University Press.

John Tait and Michael Oakes
June 2006

Organizers

Chair:

John Tait, University of Sunderland, UK

Co-Chair:

Michael Oakes, University of Sunderland, UK

Program Committee:

Branimir Boguraev, IBM, USA
Stephen Clark, University of Oxford, UK
Bruce Croft, UMass Amherst, USA
Hang Cui, National University of Singapore
Gael Dias, University of Beira Interior, Portugal
Rob Gaizauskas, University of Sheffield, UK
Alexander Gelbukh, National Polytechnic Institute, Mexico
Rosie Jones, Yahoo, USA
Noriko Kando, NII, Japan
Mirella Lapata, University of Edinburgh, UK
Liz Liddy, Syracuse University, USA
Lucia Rino, UFSCAR, Brazil
Mark Sanderson, University of Sheffield, UK
Karen Sparck Jones, University of Cambridge, UK
Chris Stokoe, University of Sunderland, UK
Tomek Strzalkowski, University at Albany, USA
Simone Teufel, University of Cambridge, UK
Olga Vechtomova, University of Waterloo, Canada

Invited Speakers:

Jamie Callan, Carnegie Mellon University, USA
Cécile Paris, CSIRO, Sydney, Australia

Workshop Program

Sunday, 23 July 2006

7:45–8:45 Registration

8:45–9:00 Opening Remarks

9:00–10:00 Invited Talk by Jamie Callan

Session 1: Accepted Paper

10:00–10:30 *Indonesian-Japanese CLIR Using Only Limited Resource*
Ayu Purwarianti, Masatoshi Tsuchiya and Seiichi Nakagawa

10:30–11:00 Coffee

11:00–12:00 Invited Talk by Cécile Paris

Session 2: Accepted Paper

12:00–12:30 *Hybrid Systems for Information Extraction and Question Answering*
Rodolfo Delmonte

12:30–14:00 Lunch

Session 3: Accepted Papers

14:00–14:30 *Extracting Key Phrases to Disambiguate Personal Name Queries in Web Search*
Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka

14:30–15:00 *How to Find Better Index Terms Through Citations*
Anna Ritchie, Simone Teufel and Stephen Robertson

15:00–15:30 *Exploring Semantic Constraints for Document Retrieval*
Hua Cheng, Yan Qu, Jesse Montgomery and David A. Evans

15:30–16:00 Coffee

16:00–17:00 Discussion led by John Tait