Coling 2008

22nd International Conference on Computational Linguistics

Proceedings of the 2nd workshop on Information Retrieval for Question Answering

Workshop chair:
Mark A. Greenwood

24 August 2008
Manchester, UK

Order copies of this and other Coling 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 978-1-905593-55-2

Design by Chimney Design, Brighton, UK
Production and manufacture by One Digital, Brighton, UK

Introduction

Open domain question answering (QA) has become a very active research area over the past decade, due in large measure to the stimulus of the TREC Question Answering track (now a track within the recently formed Text Analysis Conference, TAC). This track addresses the task of finding **answers** to natural language questions (e.g. "How tall is the Eiffel Tower?", "Who is Aaron Copland?", "What effect does second-hand smoke have on non-smokers?") from large text collections. This task stands in contrast to the more conventional information retrieval (IR) task of finding **documents** relevant to a query, where the query may be simply a collection of keywords (e.g. "Eiffel Tower", "American composer, born Brooklyn NY 1900, ...").

Finding answers requires processing texts at a level of detail that cannot be carried out at retrieval time for very large text collections. This limitation has led many researchers to rely on, broadly, a two stage approach to the QA task. In stage one a subset of question-relevant texts are selected from the whole collection. In stage two this subset is subjected to detailed processing for answer extraction. Clearly performance at stage two is bounded by performance at stage one, and previous work has shown that, despite the sophistication of standard IR ranking algorithms, they are not well suited to the stage one task of retrieving relevant documents given short natural language questions. It is likely that improvements in this area will come from linguistic insights into why QA focused IR is different from the traditional IR model.

With the continued expansion of QA research into more complex question types and with the speed with which answers are returned becoming an issue, the importance of having good, QA-focused IR techniques is likely to increase. To date this topic has received limited explicit attention despite its obvious importance. This 2nd IR4QA workshop aims to address this situation by continuing to attract the attention of researchers to the specific IR challenges raised by QA.

For this workshop, we solicited papers that addressed any aspect of QA-focused IR, in order to improve overall system performance, , suggesting possible topics such as:

- parameterizations/optimizations of specific IR systems for QA
- studies of query formation strategies suited to QA, e.g. named entity pre-processing of questions
- different uses of IR for different question types (e.g. factoid, list, definition, event, how, ...)
- utility of term matching constraints, e.g. term proximity, for QA
- analyses of differing IR techniques for QA
- impact of IR performance on overall QA performance
- QA-orientated corpus pre-processing, e.g. indexing POS tags, named entities, semantically-tagged entities, relationships, etc. rather than simply tokens
- evaluation measures for assessing IR for QA
- retrieval from semi-structured data - i.e. QA from Wikipedia articles

From the papers submitted, 10 were selected following peer review. These papers are included in this proceedings. The enthusiastic response to this workshop confirms the belief that this is an important area of interest to a significant number of researchers.

Mark A. Greenwood

Organizers:

Mark A. Greenwood, Univeristy of Sheffield

Programme Committee:

Matthew W. Bilotti, Carnegie Mellon University
Gosse Bouma, University of Groningen
Charles Clarke, University of Waterloo
Hoa Dang, NIST
Robert Gaizauskas, University of Sheffield
Eduard Hovy, ISI
Jimmy Lin, University of Maryland
John Prager, IBM
Horacio Saggion, University of Sheffield
Jrg Tiedemann, University of Groningen
Bonnie Webber, University of Edinburgh
Ralph Weischedel, BBN

Table of Contents

Conference Programme

Sunday, August 24, 2008

9:15–9:30 Welcome

9:30–10:00 *Improving Text Retrieval Precision and Answer Accuracy in Question Answering Systems*
Matthew Bilotti and Eric Nyberg

10:00–10:30 *Exact Phrases in Information Retrieval for Question Answering*
Svetlana Stoyanchev, Young Chol Song and William Lahti

10:30–11:00 Coffee Break

11:00–11:30 *Simple is Best: Experiments with Different Document Segmentation Strategies for Passage Retrieval*
Jörg Tiedemann and Jori Mur

11:30–12:00 *Passage Retrieval for Question Answering using Sliding Windows*
Mahboob Khalid and Suzan Verberne

12:00–12:30 *A Data Driven Approach to Query Expansion in Question Answering*
Leon Derczynski, Jun Wang, Robert Gaizauskas and Mark A. Greenwood

12:30–2:00 Lunch

2:00–2:30 *Answer Validation by Information Distance Calculation*
Fangtao Li, Xian Zhang and Xiaoyan Zhu

2:30–3:00 *Using Lexico-Semantic Information for Query Expansion in Passage Retrieval for Question Answering*
Lonneke van der Plas and Jörg Tiedemann

3:00–3:30 *Evaluation of Automatically Reformulated Questions in Question Series*
Richard Shaw, Ben Solway, Robert Gaizauskas and Mark A. Greenwood

3:30–4:00 Coffee Break

4:00–4:30 *Topic Indexing and Retrieval for Factoid QA*
Kisuh Ahn and Bonnie Webber

4:30–5:00 *Indexing on Semantic Roles for Question Answering*
Luiz Augusto Pizzato and Diego Mollá

5:00–5:30 Discussion Session