Information Retrieval for Question Answering a SIGIR 2004 Workshop

Robert Gaizauskas, Mark Hepple and Mark Greenwood
{r.gaizauskas, m.hepple, m.greenwood}@dcs.shef.ac.uk
Department of Computer Science
University of Sheffield
Regent Court, Portobello Road
Sheffield S1 4DP UK

1 Introduction

Open domain question answering has become a very active research area over the past few years, due in large measure to the stimulus of the TREC Question Answering track. This track addresses the task of finding **answers** to natural language (NL) questions (e.g. *How tall is the Eiffel Tower? Who is Aaron Copland?*) from large text collections. This task stands in contrast to the more conventional IR task of retrieving **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 propose, broadly, a two stage approach to the QA task. In stage one a subset of query-relevant texts are selected from the whole collection. In stage two this subset is subjected to detailed processing for answer extraction. To date stage one has received limited explicit attention, despite its obvious importance – performance at stage two is bounded by performance at stage one. The goal of this workshop was to correct this situation, and, hopefully, to draw attention of IR researchers to the specific challenges raised by QA.

A straightforward approach to stage one is to employ a conventional IR engine, using the NL question as the query and with the collection indexed in the standard manner, to retrieve the initial set of candidate answer bearing documents for stage two. However, a number of possibilities arise to optimise this set-up for QA, including:

1. preprocessing the question in creating the IR query;
2. preprocessing the collection to identify significant information that can be included in the indexation for retrieval;
3. adapting the similarity metric used in selecting documents;
4. modifying the form of retrieval return, e.g. to deliver passages rather than whole documents.

For this workshop, we solicited papers that addressed any aspect of how this first, retrieval stage of QA can be adapted 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; different uses of IR for factoid vs. non-factoid questions; utility of term matching constraints, e.g. term proximity, for QA; analyses of passage retrieval vs full document retrieval for QA; analyses of boolean vs ranked retrieval for QA; impact of IR performance on overall QA performance; named entity preprocessing of questions or collections; corpus preprocessing to create corpus-specific thesauri for question expansion; evaluation measures for assessing IR for QA.

In total 16 papers were submitted to the workshop of which 10 were selected for presentation following peer review by three reviewers per paper. Reviewing was conducted by a programme committee consisting of the three organisers plus a further 13 well-known researchers active in the area. The workshop was well attended with approximately 30 participants. The day consisted of ten half hour paper presentations, including questions, and concluded with a lively open discussion session in the final hour.

2 Presentation Summaries

The accepted papers covered a variety of topics within the field. The following summaries should provide a flavour of the research presented. Interested readers can obtain full copies of any of the papers from the workshops website at `http://nlp.shef.ac.uk/ir4qa04/`.

What Works Better for Question Answering: Stemming or Morphological Query Expansion M. Bilotti, B. Katz and J. Lin examine different approaches for handling morphological variation in IR for QA, within a boolean retrieval setting. In particular, they compare use of (Porter) stemming at indexation time, and a no-stemming baseline approach, to an approach using morphological expansion at query time. In the latter, a term is replaced with a disjunction of the term with its morphological variants, where the variants may be given either an equal or lesser weight than the expanded term. Their results indicate that, as compared to the no-stemming baseline, stemming leads to lower recall, whilst expansion gives higher recall, with the approach using differential weighting of terms performing above that with equal weighting.

A Comparative Study on Sentence Retrieval for Definitional Question Answering H. Cui, M-Y. Kan, T-S. Chua and J. Xiao describe a set of detailed experiments on ranking retrieved sentences for use in responding to definition questions, as defined in the TREC QA framework. The utility for ranking of various types of information is considered, including words co-occurring with the definition target in external resources (such as WordNet, general Web pages, and specific definition Web sites) and hard and soft definitional patterns acquired using both supervised and unsupervised techniques. Results show that external resources are helpful (especially task-specific Web sites), that machine-learned definition patterns outperform manually constructed ones, that soft patterns outperform hard ones, and that unsupervised techniques can supplement supervised ones to lead to optimal overall performance.

Using Pertainyms to Improve Passage Retrieval for Questions Requesting Information About a Location In this paper M. Greenwood reports improvements in IR performance for QA when queries constructed from questions containing location nouns are expanded so as to include the adjectival

forms of the location nouns. The adjectival forms are derived using pertainym relationships in WordNet. Interestingly expanding queries containing location adjectives with their corresponding nominal forms appears to have the opposite effect and actually decreases performance.

Minimal Span Weighting Retrieval for Question Answering This paper by C. Monz proposes a novel proximity-based approach to document retrieval for QA called minimal span weighting. The approach uses a parameterized combination of three factors in computing query-document similarity: a conventional global query/document similarity score, the minimum size of document text window that covers all terms common to query and document (a minimal matching span), and the ratio of matching terms in a query-document pair to overall query length. Experimental results show that the approach leads to significant improvements when compared to full document retrieval. The approach also allows the retrieval system to identify short text segments – the minimal matching spans – which are likely to contain an answer to the question and hence provide useful starting points for the answer selection component of a QA system.

Simple Translation Models for Passage Retrieval for QA V. Murdock and B. Croft present an approach to sentence retrieval for QA which is based upon simple translation models. Such models have the advantage of providing a similarity measure, for comparing the question and candidate sentences, that can match terms which are distinct, but which have similar meanings, or are otherwise related. Their results show that their method out-performs an approach using retrieval based on query likelihood.

Sense-Based Blind Relevance Feedback for Question Answering M. Negri presents an approach to query expansion in the QA context based on using relevance feedback to help sense-disambiguate words in the query. Instead of solving the difficult problem of word sense disambiguation (WSD) on short questions, this approach shifts the problem to the easier one of performing WSD within relevant documents which afford a larger context. Terms semantically related to the appropriate sense of each disambiguated word in the query are then used to expand the original query. Preliminary results show this technique improves the number of answer-bearing documents retrieved at rank 50 by 7 percentage points.

Exploring the Performance of Boolean Retrieval Strategies For Open Domain Question Answering H. Saggion, R. Gaizauskas, M. Hepple, I. Roberts and M. Greenwood systematically consider a number of strategies for the construction of boolean queries from natural language questions and the post-hoc ranking of results returned by these queries. These include expanding the query by disjoining Wordnet synonyms or morphological variants of query terms, relaxing the query by progressively discarding query terms with low inverse document frequency, sensitively handling proper names and quoted strings, and experimenting with different sized match windows (the window within which query terms are constrained to match) and passage windows (the window of text which is returned in response to a match). Results show that while the boolean strategies considered do not match the results for ranked retrieval systems, when measured using coverage at rank 200 (percentage of questions for which answer bearing passages are returned within the top 200), they do offer reasonable coverage for much smaller passage sizes, which may mean that overall QA performance is higher depending on the QA system's capability to extract answers from retrieved passages.

Boosting Weak Ranking Functions to Enhance Passage Retrieval For Question Answering N. Usunier, M. Amini and P.Gallinari apply a machine learning approach to the task of re-ranking the passages

returned by a conventional retrieval engine. A series of baseline ranking functions are applied to generate a number of scores for each passage, and the ranking of the passages is based on a weighted combination of these scores. The weights of this combination are learned using a boosting algorithm (*RankBoost*), with the question/answer set from `trec-11` being split to provide the training and test data. Their results show a considerable improvement over the initial ranking produced by the retrieval engine, as measured by coverage.

Seeking an Upper Bound to Sentence Level Retrieval in Question Answering K. White and R. Sutcliffe present the results of a manual investigation in which fifty `trec` factoid questions were compared with sentences in AQUAINT documents that contain their answers, with the aim of identifying the range of morphological and semantic transformations linking query terms to related sentence terms, and their relative importance. Their results indicate the importance of hypernym and co-occurrence mappings to linking questions and supporting sentences, and also suggest that quite a high proportion of questions may have a supported answer that is contained within a single sentence (true for nearly 90% of their fifty questions).

Domain-Specific QA for the Construction Sector Z. Zhang, L. Da Sylva, C. Davidson, G. Lizarralde and J-Y. Nie provide the only paper on domain-specific QA that was included in the workshop; all other papers address open-domain QA. Their research investigates use of domain-specific knowledge to facilitate domain-specific QA. In particular, they use a specialised thesaurus/ontology to identify domain terms (which may be multi-word items) in documents, and then associate them with more general semantic labels (e.g. such as the concept label for the immediate hypernym node). These semantic labels can be included in the indexation of the document collection, allowing retrieval that seeks passages from the collection that contain instances of more general concepts that have been identified within the question. Their results show an improvement in performance, measured by MRR score, as compared to conventional retrieval.

3 The Future of IR for QA

Feedback from workshop participants was exceptionally favourable with the consensus being that the workshop had been highly worthwhile and had hopefully started to promote this important area of QA research. It was generally agreed that another IR4QA workshop should be organised, possibly for Autumn 2005, and again, if possible, in a predominantly IR setting so as to interest both QA and IR researchers.

Acknowledgements

We would like to express our thanks to all the participants in the workshop, especially the authors of the presented papers. We extend a special thank you to the members of the programme committee who did a superb job in reviewing the submitted papers given tight time constraints.