Possible Approaches to Evaluating Adaptive Question Answering Systems for Mobile Environments

Peter Bailey and George Ferizis

CSIRO ICT Centre, GPO Box 664, Canberra ACT 2601 AUSTRALIA
{Peter.Bailey, George.Ferizis}@csiro.au

Introduction

The CSIRO ICT Centre has recently constructed a question answering (QA) system – My Instant Expert™ – designed for mobile phones. The client-server system supports asking open domain natural language questions and attempts to find answers from the (English) Wikipedia. Due to the small display of mobile phone devices, the space available for both question entry and answer display is limited (e.g. 240x320 pixels).

Adaptive IR and delivery techniques appeal as methods to maximize the use of this limited display and minimise the use of the costly and low network bandwidth. QA systems typically are constructed from a mixture of information retrieval (IR) and computational linguistics technologies. Adaptive approaches to IR are posited as being more likely to improve overall user satisfaction with performance. It is unclear that the existing format of QA test collections will work effectively for evaluation.

Experience

There is substantial related work in the area of building QA test collections, for example [5]. In the iterative prototyping development of the system, we faced the problem of not having a reliable baseline to benchmark our work against. Our compromise was to use a large selection of questions from TREC QA track topics, then manually verify that identical answers to these questions existed within the English Wikipedia corpus. TREC QA track-style answer patterns were used to identify whether retrieved answers contained matches.

The system uses a fairly standard approach to pipelining a sequence of IR and computational linguistic components. Performance at each component's output stage was measured using standard metrics. Representativeness of these TREC questions was an issue, especially with respect to having few questions with numeric/scale answer types. This approach provided us with some measure of server-side performance of the system, but specific to the exact TREC question set.

A significant issue with deploying a QA system on a mobile phone is the user experience of interaction, including answer presentation and answer-in-context display. In other words, the whole client-side of the equation is important to consider.

Possible approaches to evaluation

Our experiences and the additional challenges of adaptivity lead us to support the directions recently articulated by Sparck Jones's [3]. Her analysis framework which captures input, purpose, and output factors is more appropriate for adaptive systems.

Specifically, we believe it is vital to consider, model and assess output factors such as format and brevity when the interaction device is a mobile phone. Similarly, input factors such as the form of the source and subject type (e.g. news articles vs Wikipedia articles) play a part in understanding how users will interact with and trust the information. Most importantly, purpose factors such as audience and use (e.g. answering trivia questions is different from answering current stock prices) are essential for evaluating the quality of a QA system.

A comparative system evaluation approach. When the purpose of evaluation is to improve system performance and/or user satisfaction, not to compare it to past performance of other systems, then test collections need not be reusable. The approach of Thomas and Hawking involving side-by-side comparative judging in context of result displays is appropriate here [4].

To provide support for repeated evaluations, elements of standard test collections can be valuable. These include a set of queries, preferably a larger set of queries than usual, and they should be real user queries. To support this, we intend to provide a substantial query log from the My Instant Expert™ system in the coming months. Similarly a fixed corpus such as the INEX Wikipedia corpus [1] is preferred.

The classic test collection approach, updated. If there are many groups working on the same aspect of adaptive QA, then the creation of a reusable QA test collection becomes more valuable. The approach of [2] to address the creation of reusable answer judgements for specific sub-problems (e.g. factoid questions) could be adopted, with appropriate sampling and search-mediated judging. Making the test collection reusable will be far more complex, as it will entail modeling the additional factors appropriately. For example, if the query is “how many players are there in a football team?”, and if input factor locality is “UK” then the answer is 11 (soccer); if input factor locality is “New Zealand” then the answer is 15 (rugby union).

References

1. Denoyer, L. and Gallinari, P. 2006. The Wikipedia XML Corpus. *SIGIR Forum*
2. Lin, J. and Katz, B. 2006. Building a reusable test collection for question answering. *J. Am. Soc. Inf. Sci. Technol. 57*, 7, 851-861
3. Sparck Jones, K. 2006. What's the value of TREC: is there a gap to jump or a chasm to bridge? *SIGIR Forum 40*, 1, 10-20.
4. Thomas, P. and Hawking, D. 2006. Evaluation by comparing result sets in context. In *ACM Fifteenth Conference on Information and Knowledge Management* (November 2006).
5. Voorhees, E. M. and Tice, D. M. 2000. Building a question answering test collection. In *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2000), pp. 200-207.