The Importance of Focused Evaluations: a Case Study of TREC and DUC

Donna Harman
National Institute of Standards and Technology

Evaluation has always been an important part of scientific research. A simplistic breakdown of research such as that done in language technology could be as follows: perform preliminary investigations in a particular area of interest, develop hypotheses about some issue in this area, devise a method of evaluating those hypotheses, perform the necessary experiments and evaluations, analyze the results, feedback those results into further experiments, and, at some point, determine a stopping point and report results in a scientific paper. The evaluation piece of this breakdown is critical because it determines what new investigations need to be made and also when significant findings are worth reporting.

Using the field of information retrieval as an example, we can see a long history of evaluation in areas such as term weighting, use of phrases, development of stemming algorithms or segmentation, and more recently, retrieval of information across languages. In all these cases there has been extensive evaluation, and that evaluation has often used test collections.

In 1992 a new test collection was built at the National Institute of Standards and Technology (NIST) as part of the TIPSTER project, a large evaluation project for text retrieval and extraction sponsored by the US Government. This test collection was larger than the older collections by a factor of 1000, i.e. instead of 2 megabytes of abstracts, this collection contained 2 gigabytes of documents. In addition to making this collection available to researchers, NIST also created a focused evaluation to use this collection. This evaluation, called the Text REtrieval Conference (TREC), has been running for 10 years now, with over 70 participants in the last round of evaluation.

Whereas test collections, and the standardized recall/precision metrics, had allowed research groups to easily compare within their own systems, it was much more difficult to compare across systems because of slightly different experimental designs. The importance of a cross-system comparison is not to determine the best system but to give researchers some basis for understanding the strengths and weaknesses of the various techniques developed by others. This encourages the transfer of good ideas, and the identification of appropriate performance benchmarks for new approaches to match.

TREC has allowed this cross-system comparison, and as a consequence there has been a significant amount of technology transfer across systems, resulting in a doubling in retrieval effectiveness in the first 5 years of TREC. Techniques such as the OKAPI term weighting algorithms and pseudo-relevance feedback have been widely adopted by both researchers and commercial systems. Benchmarks of performance for the various test collections each year are also of critical importance for those groups working on new ideas in retrieval systems.

An equally important role of focused evaluations is the ability to target specific problems in language technology and design tasks for evaluation such that issues can be studied concurrently by multiple groups. The fact that groups usually take a different approaches to solving the problem allows for a major multiplying factor in what is learned. Specific problems have been investigated in TREC in “tracks”, starting in 1994. The most recent TREC, TREC-9, had 7 tracks: filtering, interactive, cross-language retrieval from English to Chinese, speech retrieval, web, query and question-answering. Each of these tracks demonstrate the importance of focused evaluations.

The question-answering track was run for the second time in TREC-9. The purpose of the track is to encourage research into systems that return actual answers, as opposed to ranked lists of documents. The track used 693 short-answer questions taken from query logs of Encarta and Excite as testing material, and it was guaranteed that each of these questions had a document that answered the question in the 2 gigabyte collection that was to be searched (this was verified by NIST before distributing the questions). An example of a question is ”How tall is the Empire State Building?”. Participants had to return either 50 byte or 250 byte answers (two categories of runs were allowed). There were 28 groups worldwide that participated, with the best system finding correct answers for more than two-thirds of the questions within the top 5 answers (50-bytes) that they returned.

There are several important consequences of the question-answering track. First, it demonstrated that it is possible to answer simple questions automatically and this will encourage new research and commercialization of this research. Second, it allowed many NLP groups working in this area to concentrate on a common, well-defined problem, with the evaluation effort done by an unbiased outside group (NIST). Many interesting research issues

arose, such as how to automatically categorize the questions in order to properly assign algorithms. There were also important evaluation issues, such as how the granularity of the answers affected results. As a final consequence, there was a big spread in effectiveness across the systems, and this is likely to result in the methodologies of the better systems becoming adopted by other groups.

A second example of the importance of focused evaluations is the cross-language English-Chinese track. This track had 25 topics in English, and groups had to return ranked lists of documents taken from approximately 250 megabytes of news articles from the Hong Kong newspapers (in Chinese). There were 16 groups that participated, including several large US companies and many groups from various Chinese-speaking countries. The particular value of this track was the incredible range of experiments that were performed. These ranged from full investigations of the use of different n-gram systems vs word-based systems, to the development of a complete language modeling system. Since the task was so focused, this large set of experiments can be compared across systems, and should serve as a major guidepost for further research and commercialization of Chinese cross-language retrieval.

NIST is now starting a new focused evaluation, called the Document Understanding Conference (DUC). DUC will initially concentrate on two tasks, both involving the automatic summarization of documents. In the first task, participants will be asked to produce 100-word abstracts for each document. This is similar to what has been evaluated in the past and many research organizations have projects in this area. The second task involves multi-document summaries. For this task, groups will be asked to produce four summaries at different compression rates (400, 200, 100, and 50 words) for sets of 10 documents (on average) discussing the same "concept". These concepts could be a single event (such as a set of documents on the Kobe earthquake), or multiple similar events (such as ferry sinkings). Alternatively they could be sets of documents exploring the same subject (Japanese scroll painting), or exploring different opinions on the same subject (such as the latest US election).

The goal of DUC is to focus attention on summarization and explore better ways of doing (and evaluating) this important area. The first conference will be held in the fall of 2001.

For further information on TREC or DUC, see
`http://trec.nist.gov` and `http://www-nlpir.nist.gov/projects/duc`