The Evaluation of Question Answering Systems: Lessons Learned from the TREC QA Track

Ellen M. Voorhees

National Institute of Standards and Technology
100 Bureau Dr. STOP 8940
Gaithersburg, MD 20899-8940
ellen.voorhees@nist.gov

Abstract

The TREC question answering (QA) track was the first large-scale evaluation of open-domain question answering systems. In addition to successfully fostering research on the QA task, the track has also been used to investigate appropriate evaluation methodologies for question answering systems. This paper gives a brief history of the TREC QA track, motivating the decisions made in its implementation and summarizing the results. The lessons learned from the track will be used to evolve new QA evaluations for both the track and the ARDA AQUAINT program.

1. The TREC QA Task

TREC is a workshop series designed to provide the infrastructure required for large-scale evaluation of text retrieval and related technologies (National Institute of Standards and Technology, 2002). A "track" for the investigation of question answering systems was introduced into TREC-8 in 1999, and has been run each year since then for a total of three times to date.

The original motivation for the track was to foster research that would move retrieval systems closer to *information* retrieval systems rather than *document* retrieval systems. Document retrieval systems' ability to work in any domain was considered an important feature to maintain. At the same time, the technology that had been developed by the information extraction community appeared ready to exploit. Thus the task for the TREC-8 QA track was defined such that both the information retrieval and the information extraction communities could work on a common problem. The task was very similar to that used in the MURAX system (Kupiec, 1993), which used an on-line encyclopedia as a source of answers for closed-class questions, except that the answers were to be found in a large corpus of documents rather than an encyclopedia. Since the documents consisted mostly of newswire and newspaper articles, the domain was essentially unconstrained. However, only closed-class questions were used, so answers were generally entities familiar to information extraction systems.

Participants were given a document collection and a test set of questions. The questions were fact-based, short-answer questions such as *How many calories are there in a Big Mac?* and *Where is the Taj Mahal?*. Each question was guaranteed to have at least one document in the collection that answered it. For each question, participants returned a ranked list of five [*document-id*, *answer-string*] pairs such that each answer string was believed to contain an answer to the question. Answer strings were limited to either 50 or 250 bytes depending on the run type. Human assessors read each string and made a decision as to whether or not the string contained an answer to the question in the context provided by the document. Individual questions received a score equal to the reciprocal of the rank at which the first correct response was returned (or 0 if none of the five responses contained a correct answer). The score for a run was the mean of the individual questions' reciprocal ranks.

2. Evaluation

The TREC QA evaluations have been based on the assumption that different people will have different ideas of what constitutes a correct answer. This assumption was demonstrated to be true during the TREC-8 evaluation. For TREC-8, each question was independently judged by three different assessors. The separate judgments were combined into a single judgment set through adjudication for the official track evaluation, but the individual judgments were used to measure the effect of differences in judgments on systems' scores. Assessors had legitimate differences of opinion as to what constituted an acceptable answer even for the deliberately constrained questions used in the track. Two prime examples of where such differences arise are the completeness of names and the granularity of dates and locations.

Fortunately, as with document retrieval evaluation, the relative scores between QA systems remain stable despite differences in the judgments used to evaluate them (Voorhees and Tice, 2000). The lack of a definitive answer key does mean that evaluation scores are only meaningful in relation to other scores on the same data set. Absolute scores *do* change if you use a different set of judges, or a different set of questions. However, this is an unavoidable characteristic of QA evaluation. Since assessors' opinions of correctness differ, the eventual end users of the QA systems will have similar differences of opinion, and an evaluation of the technology must accommodate these differences.

A [*document-id*, *answer-string*] pair was judged correct if, in the opinion of the NIST assessor, the answer-string contained an answer to the question, the answer-string was responsive to the question, and the document supported the answer. If the answer-string was responsive and contained a correct answer, but the document did not support that answer, the pair was judged "Not supported" (except in TREC-8 where it was marked correct). Otherwise, the pair was judged incorrect. Requiring that the answer string be responsive to the question addressed a variety of issues. Answer strings that contained multiple entities of the same semantic category as the correct answer but did not indicate which of those entities was the actual answer (e.g., a list of names in response to a who question) were judged as incorrect. Certain punctuation and units were also required. Thus "5 5 billion" was not an acceptable substitute for "5.5 billion", nor was "500" acceptable when the correct answer was "$500". Finally, unless the question specifically stated otherwise, correct responses for questions about a famous entity had to refer to *the* famous entity and not to imitations, copies, etc. For example, two TREC-8 questions asked for the height of the Matterhorn (i.e., the Alp) and the replica of the Matterhorn at Disneyland. Correct responses for one of these questions were incorrect for the other.

One of the problems of judging entire strings for correctness is that the resulting judgments do not create a reusable test collection. The primary way TREC has been successful in improving document retrieval performance is by creating appropriate test collections for researchers to use when developing their systems. While creating a large collection can be time-consuming and expensive, once it is created researchers can automatically evaluate the effectiveness of a retrieval run. Unfortunately, different QA runs very seldom return exactly the same answer strings, and it is quite difficult to determine automatically whether the difference between a new string and a judged string is significant with respect to the correctness of the answer. Word recall (Breck et al., 2000) and answer patterns (Voorhees and Tice, 2000) have been suggested as ways of approximating a reusable test collection. These approximations have been well-correlated with human judgments in tests to date, but they mis-judge broad classes of responses. Since the mis-judged classes are frequently the cases that are difficult for the original systems being evaluated, the approximations are likely to be less useful as QA systems continue to improve. Nonetheless, they are currently helpful for providing quick feedback as to the relative quality of alternate question answering techniques.

3. Retrieval Results

The most accurate of the TREC-8 systems were able to answer more than 2/3 of the questions. When an answer was found at all, it was likely to be highly ranked. Not surprisingly, allowing 250 bytes in a response is an easier task than limiting responses to 50 bytes. Indeed, traditional passage retrieval techniques are effective when a response as long as 250 bytes is acceptable (Singhal et al., 2000).

Most participants used a version of the following general approach to the question answering problem. The system first attempted to classify a question according to the type of its answer as suggested by its question word. For example, a question that begins with "who" implies a person or an organization is being sought, and a question beginning with "when" implies a time designation is needed. Next, the system retrieved a small portion of the document collection using standard text retrieval technology and the question as the query. The system performed a shallow parse of the returned documents to detect entities of the same type as the answer. If an entity of the required type was found sufficiently close to the question's words, the system returned that entity as the response. If no appropriate answer type was found, the system fell back to best-matching-passage techniques.

The absolute value of the scores for TREC-9 systems was lower than for TREC-8, but in fact the systems were significantly improved (the TREC-9 task was much more difficult as described below). The improvement in QA systems came from refinements to the individual steps of the general strategy described above rather than an entirely new approach. TREC-9 systems were better at classifying questions as to the expected answer type, and used a wider variety of methods for finding the entailed answer types in retrieved passages. Many systems used WordNet (Fellbaum, 1998) as a source of related words for the initial query and as a means of determining whether an entity extracted from a passage matched the required answer type.

Many systems continued to refine this approach in the TREC 2001 track. However, the TREC 2001 track also saw a resurgence of approaches that relied on simpler pattern matching methods using very large corpora (generally the web) rather than sophisticated language processing. The idea exploited in the massive data approach is the fact that in a large enough data source a correct answer will usually be repeated often enough to distinguish it from the noise that happens to occasionally match simple patterns.

4. Creating a Question Set

The manner in which the test set of questions was assembled has had a big effect on the results of the QA evaluations. In TREC-8, the majority of the questions were created expressly for the track, and thus tended to be back-formulations of a statement in a document. In TREC-9, the questions were selected from an Encarta log that contained actual questions, and a raw Excite log. Since the raw Excite log did not contain many grammatically well-formed questions, NIST staff used the Excite log as a source of ideas for actual questions. All the questions were created without looking at any documents. The resulting test set of questions was much more difficult than the TREC-8 set, mainly because the TREC-9 set contained many more high-level questions such as *Who is Colin Powell?*. For the TREC 2001 track, the source of the questions was again web logs, this time from Microsoft and AskJeeves who automatically filtered their raw logs to select queries containing question words. NIST did additional human filtering of the logs, selecting a final set of 500 questions. Except for some tweaking of the spelling and punctuation, the questions were as they appeared in the log.

NIST has made no attempt to control the relative number of different types of questions in the test set from year to year. Instead, the distribution of question types in the final test set has reflected the distribution in the source of questions. The TREC 2001 test set contained a dramatically greater proportion of definition questions than the previous years. While a large fraction of definition questions is "real" in that the filtered MSNSearch and AskJeeves logs contain many definition questions, there are easier ways to find the definitions of terms than searching for a concise definition in a corpus of news articles. As a result, NIST intends to exert somewhat more control over the distribution of question types in future tracks.

5. Other Tasks

Each of the TREC QA tracks have differed slightly from one another in ways other than the manner in which the test set of questions was assembled. To investigate whether QA systems are robust to the variety of different ways a question can be phrased, the TREC-9 question set contained 500 questions drawn from the logs, plus an additional 193 questions that were syntactic variants of an original question. For example, the test set contained four variants for the question *What is the tallest mountain?*: *What is the world's highest peak?*, *What is the highest mountain in the world?*, *Name the highest mountain.*, and *What is the name of the tallest mountain in the world?*. Systems that parsed questions into a common representation generally had fewer differences in their responses to question variants than did systems that relied on templates to classify questions by answer types. Overall, however, most variant sets showed little variability in the average score obtained by the different participants, indicating that the difficulty of obtaining the underlying information being sought dominated the results. For the few variant sets that did have a wide range of average scores, the difference was usually caused by different word choices in the variants. For example, the original question *Where was Poe born?* had a much higher average score than any of the variants that all asked for Poe's birthplace.

The TREC 2001 track contained three tasks, the main task, the list task, and the context task. The main task was similar to the previous tracks except questions were not guaranteed to have an answer in the document collection. Recognizing that there is no answer is a challenging task, but it is an important ability for operational systems to possess since returning an incorrect answer is usually worse than not returning an answer at all. The majority of the systems did not attempt to do no-answer processing.

The list task was designed to require systems to assemble an answer from information located in multiple documents. Such questions are harder to answer than the questions used in the main task since information duplicated in the documents must be detected and reported only once. The test set of questions consisted of 25 questions constructed by NIST assessors, each of which specified a target number of instances of a particular kind of information to be retrieved. For example, *What are 9 novels written by John Updike?* was one of the question used in the task. Systems returned an unordered list of [*document-id*, *answer-string*] pairs where each pair represented a single instance. The list could contain no more than the target number of instances. Each individual instance was judged as in the main task. The evaluation metric used was average accuracy, where the accuracy for a single question was the number of distinct correct instances retrieved divided by the target number of instances. The best performing system had an average accuracy of 76%, suggesting that the list task as defined is feasible with current technology.

The context task was intended to test systems' ability to track discourse objects (context) through a short series of questions. However, system performance was so dominated by whether the system could answer the particular type of question posed that differences in ability to track context were not detectable. More research is needed to create an evaluation that actually measures a system's ability to track context.

6. Future Evaluations

The TREC QA track has stimulated research on open-domain question answering and has created a foundation on which future evaluations can build. The data used in the TREC tracks, including questions, answer patterns, sentences containing answers, and evaluation scripts are available on the TREC web site (National Institute of Standards and Technology, 2002).

To date, the TREC QA track has used only factoid questions. This allows the evaluation of the answers to be judged using a binary decision of correct/incorrect. While assessors' opinions as to correctness differ even for this basic question type, evaluation is at least stable in that the relative quality of different QA systems is not materially affected by such differences in opinion. Answers to other types of questions require a more fined-grained scoring procedure: answers that are explanations or summaries or biographies or comparative evaluations cannot be meaningfully rated as simply right or wrong. The appropriate dimensions along which such answers should be judged, scoring mechanisms that reflect quality in those dimensions, and the stability of evaluations using those scoring mechanisms all need to be investigated.

The impact the way in which the test set of questions was assembled has had on system effectiveness in TREC illustrates the balancing of tensions required to create an effective test. One the one hand, careful selection of questions allows specific features of QA systems to be tested, enabling crisper conclusions to be drawn. On the other hand, such selection generally reduces the realism of the test. Designed tests usually lack the diversity of subject matter, vocabulary, and sentence constructions that are represented in large samples of naturally occurring questions. Such diversity can be particularly important to include in initial evaluations when the features that affect performance on the task are not well understood.

The TREC track will continue, with the goal of increasing the kinds and difficulty of the questions that systems can answer. The main task in the TREC 2002 will focus on having systems retrieve the *exact* answer. In past tracks, responses could contain extraneous information and still be judged correct provided the extraneous information was not distracting. Such fuzziness in the definition of correct was used in the first track when it was unclear what the systems' abilities were, and it has remained. However, the fuzziness is masking true differences in systems in the final scores. Forcing system to be precise will not only allow scores to better distinguish among technologies, but also improve QA technology.

An evaluation effort related to the TREC QA track is the new AQUAINT (Advanced QUestion and Answering for INTelligence) program sponsored by ARDA (Advanced Research and Development Activity), a research center within the U.S. Department of Defense (see `http://www.ic-arda.org/`). The main focus of AQUAINT is to move beyond factoid questions, including the investigation of scoring mechanisms for complex answer types. Within the first year of AQUAINT (2002), AQUAINT contractors and NIST will run pilot studies to experiment with different measures.

7. References

Eric Breck, John Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light, and Inderjeet Mani. 2000. How to evaluate your question answering system every day ... and still get real work done. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, volume 3, pages 1495–1500.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Julian Kupiec. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In Robert Korfage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 181–190. Special issue of the SIGIR FORUM.

National Institute of Standards and Technology. 2002. The Text REtrieval Conference web site. `http://trec.nist.gov`.

Amit Singhal, Steve Abney, Michiel Bacciani, Michael Collins, Donald Hindle, and Fernando Pereira. 2000. AT&T at TREC-8. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. NIST Special Publication 500-246. Electronic version available at `http://trec.nist.gov/pubs.html`.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, July.