Question Answering in TREC

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899
ellen.voorhees@nist.gov

Traditional text retrieval systems return a ranked list of documents in response to a user's request. While a ranked list of documents can be an appropriate response for the user, frequently it is not. Usually it would be better for the system to provide the answer itself instead of requiring the user to search for the answer in a set of documents. The Text REtrieval Conference (TREC) is sponsoring a question answering "track" to foster research on the problem of retrieving answers rather than document lists.

TREC is a workshop series sponsored by the National Institute of Standards and Technology and the U.S. Department of Defense [7]. The purpose of the conference series is to encourage research on text retrieval for realistic applications by providing large test collections, uniform scoring procedures, and a forum for organizations interested in comparing results. The conference has focused primarily on the traditional IR problem of retrieving a ranked list of documents in response to a statement of information need, but has also included other tasks, called tracks, that focus on new areas or particularly difficult aspects of information retrieval. A question answering track was introduced in TREC-8 (1999). The track has generated wide-spread interest in the QA problem [2, 3, 4], and has documented significant improvements in question answering system effectiveness in its two-year history.

This paper provides a brief summary of the findings of the TREC question answering track to date and discusses the future directions of the track. The paper is extracted from a fuller description of the track given in "The TREC Question Answering Track" [8]. Complete details about the TREC question answering track can be found in the TREC proceedings.

1. The TREC Question Answering Task

"Question answering" covers a broad range of activities from simple yes/no responses for true-false questions to the presentation of complex results synthesized from multiple data sources. The specific task in the TREC track was to return text snippets drawn from a large corpus of newspaper articles in response to fact-based, short-answer questions such as *How many calories are there in a Big Mac?* and *Where is the Taj Mahal?*. The TREC task was restricted in that only closed-class questions were used, yet the subject domain was essentially unconstrained since the document set was newspaper articles.

Participants were given the document collection and a test set of questions. Each question was guaranteed to have at least one document in the collection that explicitly answered it. Participants returned a ranked list of five [*document-id*, *answer-string*] pairs per question 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, and could either be extracted from the corresponding document or automatically generated from information contained in the document. Human assessors read each string and decided whether the string actually did contain an answer to the question in the context provided by the document. Given a set of judgments for the strings, the score computed for a submission was the mean reciprocal rank. An individual question 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 submission was then the mean of the individual questions' reciprocal ranks.

Table 1 shows the data that was used in the two years of the track. The TREC-9 track used both a larger document set and a larger test set of questions than the TREC-8 track. A more substantive difference was the source of the questions used in the different years. The majority of the questions uses in the TREC-8 track were developed specifically for the track. These questions were often back-formulations of statements in the documents, which made the questions somewhat unnatural and also made the task easier since the target document contained most of the question words. For the TREC-9 track, NIST obtained two query logs (one an Encarta log from Microsoft and the other an Excite log) and used those as a source of questions. NIST assessors checked whether each candidate question had an answer in the document collection, and a candidate question was discarded if no answer was found.

In many evaluations of natural language processing tasks, application experts create a gold-standard answer key that is assumed to contain all possible correct responses. An absolute score for a system's response is computed by measuring the difference between the response and the answer key. For text retrieval, however, different people are known to have

	TREC-8	TREC-9
number of documents	528,000	979,000
megabytes of document text	1904	3033
document sources	TREC disks 4–5	news from TREC disks 1–5
number of questions	200	693
question sources	FAQ Finder log, assessors, participants	Encarta log, Excite log

Table 1: Data used in the two TREC question answering tracks.

different opinions about whether or not a given document should be retrieved for a query [5], so a single list of relevant documents cannot be created. Instead, the list of relevant documents produced by one person (the assessor) is used as an example of a correct response, and systems are evaluated on the sample. While the absolute scores of systems change when different assessors are used, relative scores generally remain stable [9].

A sub-goal of the TREC-8 QA track was to investigate whether different people have different opinions as to what constitutes an acceptable answer, and, if so, how those differences affect QA evaluation. To accomplish this goal, 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.

The results of this assessment process demonstrated that assessors do have legitimate differences of opinion as to what constitutes 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 [10]. 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, but this is unavoidable. If 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.

2. Track Results

Twenty different organizations participated in the TREC-8 question answering track, submitting 20 runs using the 50-byte limit and 25 runs using the 250-byte limit. For TREC-9, 28 organizations participated, submitting 34 runs using the 50-byte limit and 44 runs using the 250-byte limit.

Not surprisingly, allowing 250 bytes in a response was an easier task than limiting responses to 50 bytes: for every organization that submitted runs of both lengths, the 250-byte limit run had a higher mean reciprocal rank. The relatively simple bag-of-words approaches that are successfully used in text retrieval are also useful for retrieving answer passages as long as 250 bytes, but are not sufficient for extracting specific, fact-based answers [6].

For the 50-byte limit runs, the best performing systems were able to answer about 70% of the questions in TREC-8 and about 65% of the questions in TREC-9. While the 65% score is a slightly worse result than the TREC-8 score in absolute terms, it represents a very significant improvement in question answering systems since the TREC-9 question set was much more difficult. Recall that the TREC-9 questions were actual users' questions rather than questions constructed specifically for the track. The motivation for using "real" questions was the belief that constructed questions are easier for QA systems because the question and answer document share the same vocabulary. While this is true, the difference between the TREC-8 and TREC-9 question sets was larger than just vocabulary issues. Real users ask vague questions such as *Who is Colin Powell?* and *Where do lobsters like to live?* that are substantially more difficult for the systems to answer.

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 its answer type as suggested by its question word. For example, 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. Improvements in TREC-9 systems generally resulted from doing a better job of classifying questions as to the expected answer type, and using a wider variety of methods for finding the entailed answer types in retrieved passages.

3. Question Answering Test Collections

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. One of the key goals of the QA track was to build a reusable QA test collection—that is, to devise a means to evaluate a QA run that uses the same document and question sets but was not among the runs judged by the assessors. Unfortunately, the judgment sets produced by the assessors for the TREC QA track do not constitute a reusable test collection because the unit that is judged is the entire answer string. 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.

As an approximate solution to this problem, NIST created a set of Perl string-matching patterns from the set of strings that the assessors judged correct [10]. An answer string that matches any pattern for its question is marked correct, and is marked incorrect otherwise. The patterns have been created such that almost all strings that were judged correct would be marked correct, sometimes at the expense of marking as correct strings that were judged incorrect. Patterns are constrained to match at word boundaries and case is ignored. An average of 1.7 patterns per question was created for the TREC-8 test set, with 65% of the questions having a single pattern. The TREC-9 set averaged 3.5 patterns per question with only 45% of the questions having a single pattern. The increase in the number of patterns per question for the TREC-9 set is another indication that the TREC-9 test set was more difficult.

Using the patterns to evaluate the TREC-8 runs produced differences in the relative scores of different systems that were comparable to the differences caused by using different human assessors, though the differences in relative scores for TREC-9 were somewhat larger. Furthermore, unlike the different judgments among human assessors, patterns misjudge broad classes of responses—classes that are precisely the cases that are difficult for the original systems being evaluated. Thus, the patterns are not a true solution to the problem of building a reusable test collection for question answering. Nonetheless, the patterns can be useful for providing quick feedback as to the relative quality of alternative question answering techniques.

4. The Future

Evaluating competing technologies on a common problem set is a powerful way to improve the state of the art and hasten technology transfer. As the first large-scale evaluation of domain-independent question answering systems, the TREC question answering track looks to bring the benefits of large-scale evaluation to bear on the question answering task.

A roadmap for question answering research was recently developed under the auspices of the DARPA TIDES project [1]. The roadmap describes a highly ambitious program to increase the complexity of the types of questions that can be answered, the diversity of sources from which the answers can be drawn, and the means by which answers are displayed. The roadmap also includes a five year plan for introducing aspects of these research areas as subtasks of the TREC QA track.

The QA track in TREC 2001 (TREC-10) will include the first steps of the roadmap. The main task in the track will be similar to the task used in TRECs 8 and 9, but there will be no guarantee that an answer is actually contained in the corpus. Recognizing that the answer is not available is challenging, 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 track will also contain a list subtask in which each question asks for a specific number of instances to be returned. The questions have been constructed such that retrieving the target number of instances requires collating information from more than one document. For example, a list question such as *Name the countries the Pope visited in 1994.* requires finding multiple documents that describe the Pope's visits and extracting the country from each. The system must also detect duplicate reports of the same visit so that countries are listed only once per visit.

5. REFERENCES

[1] Sanda Harabagiu, John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. Issues, tasks, and program structures to roadmap research in question & answering (Q&A), October 2000. `http://www-nlpir.nist.gov/projects/duc/roadmapping.html`.

[2] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the Association for Computational Linguistics*, pages 274–281, July 2001.

[3] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi, and Richard J. Mammone. Question answering using maximum entropy components. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 33–39, June 2001.

[4] John Prager, Eric Brown, Anni Coden, and Dragomir Radev. Question-answering by predictive annotation. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 184–191, July 2000.

[5] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

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

[7] The Text REtrieval Conference web site. `http://trec.nist.gov`.

[8] Ellen M. Voorhees. The TREC question answering track. *Journal of Natural Language Engineering*. To appear.

[9] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[10] Ellen M. Voorhees and Dawn M. Tice. 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 2000.