Implementing a Question Answering Evaluation

Ellen M. Voorhees, Dawn M. Tice

National Institute of Standards and Technology
Gaithersburg, MD 20899
{ellen.voorhees,dawn.tice}@nist.gov

Abstract

The most recent TREC workshop contained a track for evaluating domain-independent question answering (QA) systems. In addition to fostering research on the QA task, the track was used to investigate whether the evaluation methodology used for document retrieval is appropriate for a different natural language processing task. This paper describes the implementation of the track, focusing on how the test questions were selected and the responses judged. By using multiple assessors to independently judge the responses, we verified that assessors do have legitimate differences of opinion as to the correctness of a response, even for the fact-based, short-answer questions used in the track.

The Text REtrieval Conference (TREC) is a workshop series organized by the U.S. National Institute of Standards and Technology (NIST) designed to advance the state-of-the-art in text retrieval (Voorhees, 2000). The most recent TREC, held in November 1999, included a "track" on question answering where the goal was to evaluate technology for finding answers to fact-based questions in unrestricted domains.

One of NIST's goals for the question answering track was to investigate an appropriate evaluation paradigm for the question answering task. 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 different opinions about whether or not a given document should be retrieved for a query (Schamber, 1994), so a single correct list of documents cannot be created. Instead, the list of documents produced by one person (known as 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' opinions are used, relative scores generally remain stable, so scores computed using sample judgments are valid for comparing different retrieval techniques (Lesk and Salton, 1969; Voorhees, 1998).

We wished to investigate whether different people have different opinions as to what constitutes an acceptable answer, and, if so, how those differences affect evaluation scores in the question answering task. Accordingly, we used three different assessors to independently judge each question. The results demonstrate that different assessors do have legitimate differences of opinion as to what constitutes an acceptable answer. In particular, assessors differ on how much of a name is required, and on the granularity of times and locations. The context the assessor assumes is also important. For example, some assessors accepted Atlantic City (home of the Taj Mahal casino) as an acceptable response to *Where is the Taj Mahal?* since they could imagine contexts in which Atlantic City would be the preferred answer.

Fortunately, as with text retrieval evaluation, relative scores between question answering systems are stable despite differences in the underlying judgments. For the official track evaluation, we combined the three different judgment sets per question into a single, adjudicated judgment set. Subsequent analysis demonstrated that the adjudicated judgments can be replaced by a judgment set such that each question is judged by only one assessor without losing accuracy in comparative evaluations.

This paper describes the implementation of the Question Answering (QA) track. It focuses on how the test questions were selected, the training given to the assessors, and an analysis of how the assessors perceived their task. Retrieval methods and results for individual systems are not given in this paper; see the participants' papers in the TREC-8 proceedings for these results (Voorhees and Harman, 2000). In addition, the stability of comparative results in the presence of different judgments for the QA task is treated only cursorily here. More thorough analyses are given in (Voorhees and Tice, 2000a; Voorhees and Tice, 2000b).

1. The TREC-8 Question Answering Track Task

Designing an evaluation entails defining a specific task to be performed, including how correct responses are recognized and results scored. For the QA track, participants received a large collection of documents and 200 fact-based, short-answer questions such as *How many calories are there in a Big Mac?*. Each question was guaranteed to have at least one document in the collection that answered the question. Participants were to return a ranked list of five [doc id, answer string] pairs per question such that each string was believed to contain an answer to the question. Depending on the run type, answer strings were limited to either 50 or 250 bytes. Human assessors read each string and made a binary decision as to whether or not the string actually did contain an answer to the question. 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. Creating the Question Set

We wanted the test set of 200 questions to represent a wide spectrum of subjects and question types while meeting the general specification of fact-based, short-answer questions. To accomplish this goal we collected a pool of 1,837 candidate questions from four different sources: TREC QA participants, the NIST TREC team, the NIST assessors, and question logs from the FAQ Finder system (Burke et al., 1997). Our intention was that these different sources would provide different kinds of questions. The TREC participants have detailed knowledge about how their systems work and might have used that knowledge to select questions that would stress the technology. NIST team members created questions mostly to investigate how to teach assessors to create questions, but also have technical knowledge of question answering systems. The assessors have limited technical knowledge regarding question answering systems, and so represent a general user's point of view. Nonetheless, the assessors created their questions from the test document collection specifically for the track, and thus their questions do not represent natural information-seeking behavior. The questions taken from the FAQ Finder logs, on the other hand, were submitted to the FAQ Finder system by undergraduate students who were genuinely interested in the answers to the questions.

The FAQ Finder logs contained 1500 questions. A subset of approximately 100 questions was selected by first eliminating entries that were not in the form of a question or that asked about subject matter deemed inappropriate for a government-sponsored evaluation, and then selecting those questions most likely to have an answer in the test document collection. The document collection was searched using NIST's PRISE retrieval system to verify which of these 100 questions actually did have answers in the collection.

The other three sources of questions supplied answer strings and document ids with the candidate questions. Seven NIST assessors created 70 candidate questions (10 questions each). The assessors used PRISE to search the test document collection. Their methodology entailed thinking of a topic of interest, entering key search terms, and reading the text of the document to form a question. As a result, many of the assessors' questions are back-formulations of sentences from the texts.

Four members of the NIST TREC Team submitted a total of 25 candidate questions. One of the team members used web search engines to create his questions, which were then verified as having answers in the QA document collection. The others in the team searched the QA document database in much the same manner as the assessors.

A total of 242 candidate questions were submitted by 23 groups that signed up to participate in the track (though in the end not all of these groups were able to submit runs). NIST does not know what methods participants used to create the questions, but the range of question types suggests a variety of methods was used.

Our main goal for the first running of the track was to create a set of clean, straight-forward questions and answers. We eliminated any candidate question that a member of the team thought was ambiguous, that had a list of three or more items for an answer, that had an answer string greater than 50 bytes, or that were much too obscure or contrived (i.e., extreme examples of back-formulations). Since we were unsure how difficult (or easy) the task would be for existing systems, we tried to select questions with a range of difficulty, including some questions that we felt would challenge the systems. For example, while most questions that required compound answers were eliminated, we kept a few questions that had compound answers by rewording the questions to indicate that a compound answer was required. Thus *What two US biochemists won the Nobel Prize in medicine in 1992?*, required the names of both biochemists. The final test set of 200 questions contained 127 questions originally submitted by participants or NIST staff, 49 questions from the assessors, and 24 questions from the FAQ Finder logs.

Despite the care we took to select questions with straight-forward, obvious answers and to ensure that all questions had answers in the document collection, once assessing began it became clear that there is no such thing as a question with an obvious answer. Not only did most questions have more different answers than we anticipated, but the assessors determined that two of the 200 questions had no clear answer. Question 131, *Which Japanese car maker had its biggest percentage of sale in the domestic market?*, was submitted by a participant who supplied the answer of "Toyota" with a document that states "Toyota had 42% of the domestic market". But the assessors were unsure whether "domestic market" referred to Japan or the U.S., and they refused to accept 42% as the largest percentage without further proof. Question 184, *When was Queen Victoria born?*, was a FAQ Finder question. We found a document containing "Queen Victoria (1837-1901)", so assumed that the answer was 1837. Unfortunately, a closer reading of the document makes it clear that 1837 was the beginning of her reign, not of her life. Due to these problems, we eliminated questions 131 and 184 from the evaluation results.

Prior to the release of the test set of questions, NIST released a development set of 38 questions. These questions came from the same sources as the test set except that no FAQ Finder questions were included in the development set. The development set included all of the different types of questions as in the test set, but we made no attempt to keep the proportion of questions of a given type the same in the two sets. None of the development set questions was included in the test set.

3. Assessor Training

The rationale for using human assessors to evaluate a task is to incorporate the perceptions of the end-users of the technology into the evaluation to the greatest extent possible. This argument suggests we should give assessors minimal or even no training so we receive their natural reactions. However, we must also be able to interpret the results of an evaluation, and this argues for ensuring that different assessors have the same basic understanding of what their task is. Our experience with the document relevance judging task has demonstrated the importance of adequate training for the assessors (Downey and Tice, 1999). Thus, the assessors who performed the QA task received special training developed specifically for this task. The purpose of the training was not to drill the assessors on a specific set of assessment rules, but rather to motivate their task and provide guidance on the sorts of issues that might arise while they were assessing the test questions.

Fifteen different assessors were trained for the QA task. The QA training session took approximately two hours, and consisted of a general introduction that motivated the task, system-based training for the QA assessment system developed for the track, and task-based training using four sample questions with small answer pools. As a first step, the assessor was asked to read a set of written instructions (the instructions are reproduced in (Voorhees and Tice, 2000c)). Next, we taught the mechanics of the QA assessment system by having the assessor follow the steps in a visual training manual that consisted of screen shots of the system annotated with explanations of the system's features. The assessors then used the system to judge the sample answer pools.

The answer pools for the four sample questions were concocted by the NIST TREC team to illustrate the types of issues the assessors would face when judging actual test questions. Since we did not have previous experience with this task, we were not certain exactly what issues would arise, but made our best guess from the discussions on the track mailing list and the categorization of questions we developed while selecting the final test set of questions. In the end, all of the issues included in the training, plus more, actually occurred in judging the test set questions.

The four training questions were ordered roughly by difficulty of judging. First the assessor was given the following scenario to use for judging answer strings:

> Assume there is a user who trusts the answering system completely, and therefore does not require that the system provide justification in its answer strings. Your job is to take each answer string in turn and judge if this answer string alone were returned to the trustful user, would the user be able to get the correct answer to the question from the string.

The assessor then began judging the answer strings for the first training question. Assessors were asked to hold their questions until they had judged all the answer strings for a given training question. Once they finished the first question, we reviewed the judgments with them. We paid particular attention to the reasons for their judgments as we discussed the training questions.

The first training question was *Who was Johnny Mathis's track coach?*, and the correct answer is Lou Vasquez. This question is relatively straight-forward, and we used it to introduce the fundamentals of QA judging to the assessors: that the answer strings would contain snippets of text that were not necessarily grammatically correct and might even contain word fragments; that the answer string did not need to contain justification to be counted as correct; that the assessors were to judge the *string* not the document from which the string was drawn (after eight years of judging documents, this lesson was sometimes a hard one to learn); that the document context must be taken into account; and that the answer string had to be responsive to the question. The pool also illustrated a problem specific to "who" questions, i.e., whether first name only, or last name only is sufficient for a correct response. In the case when only part of the name is given, we told the assessors that they should use their own judgment, though we did suggest that first name only was probably insufficient while last name only was probably sufficient.

One of the answer strings in this pool contained a list of names extracted from the document that contained the correct answer, "Lou Vasquez, O.J. Simpson, Ollie Matson and Johnny Mathis". We trained the assessors to judge this string as *incorrect* despite the fact that the correct answer is contained within it. The reasoning behind judging this string as incorrect is that the user given just this answer string still does not know who the track coach is, though admittedly the field is narrowed significantly. This is part of what we meant by insisting that the answer string be responsive to the question. Answer strings that contained multiple entities of the same semantic category as the correct answer, but that did not indicate which of those entities was the actual answer, were judged as incorrect.

The second training question, *Who is the President of the United States?*, demonstrated a question for which the correct answer changes over time thus making document context vital. Responses to questions phrased in the present tense were judged as correct or incorrect based on the time of the document associated with the response. In the sample answer pool we had strings containing the names of former presidents extracted from a document that stated Clinton was president; these responses were incorrect. We also had strings where the same presidents were named but the strings were associated with documents when they were president; these responses were correct. Another answer string was "Bush" taken from a document written when George Bush was president but dealing exclusively with shrubs. All the assessors judged that string as incorrect (and we agreed). We also used the string "Bush" associated with a document that was written on the eve of Bush's inauguration. Since Bush was not yet officially president at the time of the document, we judged this response as incorrect. We suspected we were getting overly convoluted in our examples, but an equivalent issue did arise in the actual pools. The pool for *Who is the prime minister of Japan?* contained many responses associated with documents that announced the resignation of Prime Minister Hosokawa.

The third training question, *What is the world's population?*, illustrated questions whose answers change over time and can be reported to different levels of accuracy or in different units. Once again, the document context was used to determine whether a particular figure was correct. Example answer strings of "world population of 5.5 billion", "5.4bn", and "5.7bn" were all judged as correct since the corresponding documents gave these figures for the current population of the world. The answer string that gave a figure for the world's population extracted from a document that was discussing a historical number (the population at the time of WWII for a document written in the 1980s) was judged as incorrect. Similarly, predictions for the future population were also judged as incorrect. Assessors were warned to watch for unfortunate truncation in answer strings. For example, the answer string ".4bn" extracted from a document that gives the world's population as 5.4bn is incorrect. There were similar issues that were not covered in the training material but arose during the actual assessing. Some systems removed punctuation from the document source before extracting answers, and occasionally suffered for it. "5 5 billion" was not an acceptable substitute for "5.5 billion". Money units disappeared this way, too. A response of "500" was not acceptable when the correct answer was "$500". In general, answer strings that did not contain a unit of measure for questions that required a quantity as a response were judged incorrect (because they were unresponsive).

The final training question was *How tall is the Statue of Liberty?*. This question was one of the questions in the development set, and one of the track participants pointed out that there were a number of documents that talked about various replicas of the Statue of Liberty, each of which was a different height. NIST decided that a string that correctly extracted the height of a replica from a document that was clearly discussing a replica would be judged as *incorrect*, arguing once again that such an answer was not responsive to the question. Unless the question specifically stated otherwise, we assumed that any question regarding a famous entity was asking about *the* famous entity and not about imitations, copies, etc.

This issue also occurred in the set of test questions. Questions 199 and 200 asked for the height of the Matterhorn (i.e., the Alp) and the replica of the Matterhorn at Disneyland, respectively. Correct responses for one of these questions were incorrect for the other. The test question *Where is the Taj Mahal?* was more of a challenge to judge because some assessors accepted the Taj Mahal casino in Atlantic City as a famous entity in its own right (though the judgments used to score the runs accepted only Agra, India).

4. Assessors' Perception of Their Task

Once their training was finished, the assessors judged the submissions to the track. For each question, we created an answer pool consisting of each distinct [doc id, answer string] pair for that question in the set of 45 runs submitted to the QA track. To judge a question, the assessor needed to make a binary decision as to whether the answer string was a correct response in the context of the document for every pair in the pool. The mean size of an answer pool was 191.6 pairs (minimum pool size was 169 pairs, maximum pool size was 207 pairs), and the pools contained a mean of 55.3 distinct documents (min 28, max 93). On average, an assessor took approximately a half hour to judge one question.

Each question was independently judged by three different assessors. Because the assessors started at different times and worked at different rates, different assessors judged different numbers of questions and judged questions in different orders.

During the assessing process, the assessors interacted freely with the NIST TREC team members, asking for clarification of the assessment guidelines and verifying their application of the guidelines to particular cases. The interaction provided an informal mechanism for us to learn how the assessors perceived their task. We also instituted two more formal methods for gathering this information. Every time an assessor started a question, he or she was given a sheet of paper to record the canonical answers to the question. We asked that they write any other comments they had about the question on the same sheet, and most sheets were returned with comments. The most detailed information came from a series of "think-aloud" observations of assessors judging an entire question. During a think-aloud session, the assessor was asked to think aloud as he or she considered each answer string in the answer pool, and an observer recorded the comments. Eight think-aloud sessions were held, one each with five different assessors on five different questions plus all three assessors on a sixth question. To eliminate any "start-up" effects in the observations, each of the assessors judged at least three questions before being observed. The sessions were performed in a separate room with only the assessor and the observer present.

The data gathered from the assessors suggest that the assessors understood their task and could do it. The assessors generally followed the assessing guidelines, though we did find some common patterns of mistakes. Some assessors needed reminding to judge an answer string based on what the string itself contained rather than what the associated document contained. For example, for the question *In what year did Ireland elect its first woman president?*, one assessor was observed marking strings that contained no year as correct because the document contained the year. Another pattern was to mark strings as incorrect because they did not contain supporting evidence for the correctness of the answer. This was not so much a problem when the answer string contained only the answer (*What is the capital of Kosovo?*, answer string "Pristina") but when the answer string contained random other information (answer string "Arkan Calls For Expulsion of 700,000 Albanians AU0305195294 Pristina KOSOVA DAILY REPORT Nr. 347 in English 3 May 94 AU0 305195294 Pristina KOSOVA DAILY REPORT Nr. 347"). Of course, there were also just plain blunders: times when the assessor hit the wrong button or whatever. Frequently the assessors would catch the blunders and correct them, but inevitably there were some blunders that persist.

There was one aspect of the judging task that caused the assessors significant difficulty—an aspect related to the way in which the QA task itself was defined. The track guidelines required participants to return a document with the answer string and allowed answer strings to be generated (i.e., the answer strings did not have to be extracted from the document returned). The document was returned to provide the context for judging the answer string. The context was used not only to provide a frame of reference for questions whose answer changes over time, but also to give credit to systems that correctly extract information from a document that is in error. There were a number of instances, however, when an answer string contained the correct answer but that answer could not possibly have been determined from the document returned. For example, the correct answer for *Who is the 16th President of the United States?* is Abraham Lincoln. One of the answer strings returned contained Abraham Lincoln, but the associated document discussed Lincoln's Gettysburg Address. The document does not even mention that Lincoln was president, let alone that he was the 16th president. Because generated answers were allowed in the track, we instructed the assessors to judge these strings as correct. The assessors found this very difficult to do since giving credit for both a right answer when it is not supported by the document and a wrong answer when it is supported by the document does not reflect any real user task.

5. Validity of Results

Many differences among assessors were not caused by mistakes, but by legitimate differences of opinion as to what constitutes an acceptable answer. Two prime examples of where such differences arise are the completeness of names and the granularity of dates and locations. For example, two assessors accepted “April 22” as a correct response to *When did Nixon die?*, but the other assessor required the year as well. Year-only was always acceptable for historical questions. Similar issues arise with locations. For *Where was Harry Truman born?*, some assessors accepted only Lamar, Missouri, while others accepted just Missouri. No assessor accepted just USA, though for other questions country-only designations were judged as acceptable. People are addressed in a variety of ways as well. The assessor training suggested that surname-only is usually acceptable while first-name-only seldom is. Besides obvious exceptions such as Cher or Madonna, there are the different forms of address in other cultures. For example, the full name of the recipient of the 1991 Nobel Pease Prize is Aung San Suu Kyi. Some assessors accepted all of “Aung San Suu Kyi”, “Suu Kyi”, “San Suu Kyi”, and “Kyi”.

On average, 6% of the answer strings that were judged were disagreed on. Looking at the total percentage of answer strings that had disagreements is somewhat misleading, though, since a large percentage of the answer strings are obviously wrong and assessors agree on those. Following the document relevance judgment literature (Lesk and Salton, 1969), we can compute the *overlap* in the sets of strings that were judged correct. Overlap is defined as the size of the intersection of the sets of strings judged correct divided by the size of the union of the sets of strings judged correct. Thus, an overlap of 1.0 means perfect agreement and an overlap of 0.0 means the sets of strings judged as correct were disjoint. The mean overlap across all three judges for the 193 test questions that had at least 1 correct string found was .641.

This overlap is greater than the overlap generally reported for relevance judgments (which is not surprising given that only questions with “obvious” answers were selected to be in the test), but is still small enough to demonstrate the futility of trying to devise a single authoritative list of correct answers. Even if assessors could reach a consensus on such a list, the resulting list would be of little use since the eventual end users of the systems will have similar differences of opinion. To drive research in useful directions, an evaluation paradigm must capture as much of the actual environment the systems will be deployed in as possible. In this case, that means recognizing that different people have different definitions of correct.

But how can systems be evaluated if there is no universal definition of correct behavior? As with text retrieval evaluations, comparative evaluations are valid because the results are stable despite differences in judgments. We computed the correlation among the system rankings produced by evaluating runs using different judgment sets. The mean correlation between two rankings was greater than .95, and the minimum correlation was greater than .91. The mean correlation represents an average of 18 pairwise differences between the two rankings out of a maximum possible 820 pairwise differences. Almost all of the differences were from runs whose evaluation scores were very close to one another (less than a 5% difference) when they were evaluated on one judgment set.

6. Summary

NIST used the TREC-8 QA track to investigate whether the evaluation paradigm used for text retrieval—where the opinion of a single human judge is the answer standard—can be used for another natural language processing task. The results of the track confirm that the paradigm is an appropriate choice for the question answering task. Assessors both understand and can do their task, provided the judgments they are to make mimic some real-world task sufficiently closely. Assessors disagree as to whether a particular response is a correct answer frequently enough to make the quest for a single gold-standard answer key pointless. Fortunately, the relative scores between QA systems remain stable despite differences in the judgments used to evaluate them. Indeed, for the purpose of a comparative evaluation, judgment sets that consist of a single judge's opinion for each question are not only equivalent to one another, but are also equivalent to judgment sets that use a combination of assessors' opinions for each question. Using a single opinion per question is an important cost savings since combination judgment sets are much more expensive to produce.

7. References

Burke, Robin D., Kristian J. Hammond, Vladimir A. Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg, 1997. Questions answering from frequently-asked question files: Experiences with the FAQ Finder system. Technical Report TR-97-05, The University of Chicago, Computer Science Department.

Downey, Laura L. and Dawn M. Tice, 1999. A usability case study using TREC and ZPRISE. *Information Processing and Management*, 35(5):589–603.

Lesk, M.E. and G. Salton, 1969. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4:343–359.

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

Voorhees, Ellen M., 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (eds.), *Proceedings of* *the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Melbourne, Australia: ACM Press, New York.

Voorhees, Ellen M., 2000. Special issue: The sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1).

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

Voorhees, Ellen M. and Dawn M. Tice, 2000b. The TREC-8 Question Answering Track. In *Proceedings of the Second international conference on Language Resources and Evaluation (LREC-2000)*. To appear.

Voorhees, Ellen M. and Dawn M. Tice, 2000c. The TREC-8 question answering track evaluation. In (Voorhees and Harman, 2000).

Voorhees, E.M. and D.K. Harman (eds.), 2000. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. Electronic version available at http://trec.nist.gov/pubs.html.