The TREC-8 Question Answering Track Evaluation

Ellen M. Voorhees, Dawn M. Tice
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The TREC-8 Question Answering track was the first large-scale evaluation of systems that return answers, as opposed to lists of documents, in response to a question. As a first evaluation, it is important to examine the evaluation methodology itself to understand any limits on the conclusions that can be drawn from the evaluation and possibly to find ways to improve subsequent evaluations. This paper has two main goals: to describe in detail how the evaluation was implemented, and to examine the consequences of the methodology on the comparative performance of the systems participating in the evaluation. The examination uncovered no serious flaws in the methodology, supporting its continued use for question answering evaluation. Nonetheless, redefining the specific task to be performed so that it more closely matches an actual user task does appear warranted.

1 Introduction

The Text REtrieval Conference (TREC) is a series of workshops designed to advance the state-of-the-art in text retrieval by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. Evaluating competing technologies on a common test set has had the desired effect of increasing text retrieval system effectiveness as demonstrated, for example, by the doubling of performance of the SMART system since the beginning of TREC [1]. However, users generally would prefer to receive *answers* in response to their questions, as opposed to the document lists traditionally returned by text retrieval systems. The TREC-8 Question Answering Track is an initial effort to bring the benefits of large-scale evaluation to bear on the question answering task.

Of course, in general "question answering" is a wide field ranging from simple yes/no answers for true-false questions to the presentation of complex results synthesized from multiple data sources. Designing an evaluation entailed defining a specific task to be performed, including how correct responses are recognized and results scored. The track coordinators, Amit Singhal and Tomek Strzalkowski, came to the TREC-7 conference with a proposed track definition, which was discussed and revised during a TREC-7 track planning workshop. The task was further refined through discussions by participants and other interested parties on the track mailing list. The track was advertised by posting to natural language processing related mailing lists with potential participants referred to the track web site at `http://www.research.att.com/~singhal/qa-track.html`.

The following specification of the track eventually emerged. 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 strings 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.

The results of the evaluation are reported elsewhere in this proceedings in the Question Answering Track Overview paper. The current paper examines the evaluation methodology itself. That is, the paper explores the appropriateness of the evaluation design and the validity of the conclusions that can be drawn. The next section describes how the design was implemented: how the test questions were selected, the instructions given to the human assessors, and an analysis of how the assessors perceived their task. One of the main conclusions of this analysis is that even for these highly constrained questions, answers depend on context and different assessors have legitimate differences of opinion as to whether a particular answer string is correct. In light of this finding, section 3 examines the effect of different judgments on the comparative performance of the answering systems. As is true with differences in relevance judgments for document retrieval, the differences in answer judgments do affect absolute scores, but relative scores across different runs remain stable. Using assessor judgments for the question answering task is thus a valid way to compare answering system quality.

2 Implementation of the TREC-8 QA Evaluation

The description of the QA task given above is complete as far as it goes, yet a number of additional decisions must be made to actually implement that description. Experience with TREC document retrieval tasks has demonstrated that seemingly minor details in the implementation of a task can occasionally have far-reaching effects on the evaluation results. As an example, the introduction of the three best content words as the "Title" field of TREC topic descriptions in TREC-6 has altered the nature of the TREC topic statements. In this section, therefore, we present a detailed description of how the QA task was implemented, including how the questions were selected, how the assessors were trained, and the implications of the mean reciprocal rank scoring metric. We also include a qualitative analysis of how the assessors perceived their task.

2.1 Creating the question set

The QA task as defined required a test set of 200 short-answer questions. Our goal was to have the test set represent a wide spectrum of subjects and question types while meeting this general specification. To accomplish the 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 FAQFinder system. 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 FAQFinder logs, on the other hand, were submitted to the FAQFinder system by undergraduate students who were genuinely interested in the answers to the questions[1]. Appendix A lists the set of questions in the final test set. The table at the end of the appendix gives the source of each question.

The FAQFinder 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. Starting from the top of that list, Dawn Tice used NIST's PRISE search engine to look for documents containing answers to the current question, stopping when 24 answers had been found. Sadly, the FAQFinder question "Where did the Voorhees family ancestors immigrate from?" had to be eliminated at this step.

The other three sources 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 QA 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. For instance, question 151, "Where did Dylan Thomas die?" was extracted from document FT934-10120 which reads, "DYLAN Thomas died in New York 40 years ago next Tuesday."

[1]The FAQFinder question logs were given to NIST by Claire Cardie of Cornell University, with permission of Robin Burke, the creator of the FAQFinder system who is now at the University of California, Irvine.

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.

Finally, 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.

The 337 candidate questions from sources other than the FAQFinder logs were then filtered by the NIST team to select the final test set. Our main goal for the first running of the track was to create a set of clean, straightforward questions and answers. We eliminated any 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 question 16, "What two US biochemists won the Nobel Prize in medicine in 1992?" required the names of both biochemists.

Once the set of 200 questions was selected, NIST checked the document ids and answer strings submitted by the source to ensure the answer was really there. To do so, we entered the supplied answer texts as search strings in PRISE. We found some instances of incorrect document numbers or answer strings for the questions as submitted, but were eventually able to find correct [answer string, document] pairs for all 200 questions (or so we thought).

Despite the care we took to select questions with straightforward, 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." However, the assessors were unsure whether "domestic market" referred to Japan or the United States, and refused to accept 42% as the largest percentage without further proof. Question 184, "When was Queen Victoria born?" was a FAQFinder question. Document FT924-6257 contains "Queen Victoria (1837-1901)," so we assumed 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.

Part of the reason we did not anticipate the variety of different answers for these questions was with the way we checked the answer strings. By searching for the answer text as supplied by the source, we were immediately put in the same mindset as the question author. Had we started from scratch not knowing the answer, we probably would have found many of the other answers. Of course, this would have substantially increased the cost of what was already a labor-intensive question selection process. Furthermore, it would not have made the questions any more obvious, or less ambiguous, but simply would have made us aware of the complexities at an earlier stage.

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 no FAQFinder questions were included in the development set (since we had not yet verified any FAQFinder questions at the time the set was released). 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.

2.2 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 [2]. Accordingly, the assessors who performed the QA task received special training developed specifically for the QA 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.

To minimize any confusion between tasks, the assessors were required to finish their assessments for the TREC ad hoc task before being trained on the QA task. The assessors generally finished the ad hoc task at different times, so most QA training was on a one-on-one basis. At most, we had two assessors being trained simultaneously on the QA task. Fifteen different assessors were trained on the task.

2.2.1 The QA assessment system

Since the QA task has different requirements from document relevance judging, a new assessment system was created especially for the QA task, though the new assessment system was based on the relevance judging system the assessors are very familiar with. Given a question number, the QA assessment system displays the text of the question and each answer string to be judged. The system also displays the document id associated with each answer string,[2] with answer strings sorted by document id so all strings associated with the same document are adjacent to one another. To judge a string, assessors click on either the "yes" or "no" radio button located next to the answer string. Clicking on either the document id or the answer string displays the corresponding document in a separate window. To assist them in targeting the relevant areas of a document, assessors can enter a search string that is then highlighted in the document. The search used is an exact (case-insensitive) string match on the entire string. This is different from the document assessing system in which the assessors enter a set of search terms and all words that conflate to the same stem as any of the search terms are highlighted in the text. When we developed the QA assessment system we thought the exact string match was better-suited to the QA task than was the set of search terms. However, the assessors found this difference between the two assessment systems difficult, and generally searched using only one word at a time in the QA task.

The QA training session took approximately two hours, and consisted of a general introduction that motivated the task, system-based training from a visual training manual of the QA assessment system, and task-based training using four sample questions with small answer pools concocted by the NIST TREC team. As a first step, the assessor was asked to read the written instructions reproduced in Appendix B. Next, we taught the mechanics of the QA assessment system by having the assessor follow the steps in the training manual. The visual training manual consists of screen shots of the system annotated with explanations of the system's features.

2.2.2 QA task training

The answer pools for the four sample questions used in the task-based training 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 (described in detail below) were ordered roughly by difficulty of judging. First the assessor was given the following scenario to use for judging answer strings:

[2]Participants were required to submit a document id with each answer string.

> 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 straightforward, 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. But document context then becomes an issue. We judged as incorrect answer strings that contained "Vasquez" when the document associated with the response was about Lupe Vasquez or Ruben Vasquez (individuals who are completely unrelated to Lou Vasquez). Another 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." This string was also judged 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 the sort of "interference" referred to in the written instructions to the assessors and what was meant by insisting that the answer string be responsive to the question. If answer strings contained multiple entities that were of the same semantic category as the correct answer, but did not indicate which of those entities was the actual answer, the response was 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. When we inserted the inauguration document into the sample answer pool we assumed we were getting overly convoluted in our examples. But an equivalent issue did arise in the actual pools. The pool for question 147, "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. However, 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 incorrect.

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. Once again this issue was seen in the set of test questions. For test question 73, "Where is the Taj Mahal?" we accepted only Agra, India, not the Taj Mahal casino in Atlantic City, New Jersey, nor the Taj Mahal Hotel in Bombay. Similarly, 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.

2.3 Judging the test set

Our original intention was that each test question would be judged by one assessor. However, once the assessing started it became clear that the assessors could judge an entire question much faster than we had originally planned for, and that judging correctness was much less cut-and-dried than we had hoped for. As a result, we decided to have each question be judged independently by three assessors. This would allow us to build a high-quality judgment set and also to gain insight into the effect of human assessors on question answering evaluation.

On average, an assessor took approximately a half-hour to judge one question. To judge a question, the assessor needed to judge each answer string that was in that question's answer pool. The answer pool consisted of each distinct [doc id, answer string] pair in the set of 45 runs submitted to the QA track. 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). As expected, there was very little overlap in strings across runs.

In all, fifteen different assessors judged some QA question. Because the assessors started at different times (depending on when they finished their ad hoc assessing) and worked at different rates, different assessors judged different numbers of questions and judged questions in different orders. The only invariants in assigning questions to assessors were that each question was judged three times and that no assessor judged a question more than once. The three sets of judgments for one question were mostly independent of each other, though there were occasional discussions among the assessors about particularly interesting assessing situations.

2.3.1 Assessors perception of their task

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 (i.e., the simplest form of the 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. An observer (Tice) recorded the comments as the assessor judged the strings. Interruptions by the observer were kept to a minimum, although assessors were occasionally reminded to think aloud. 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.

We had two goals for the information we gathered from the assessors. First, we hoped to gain a better understanding of how the assessors perceived the task and how they actually judged the answer strings (as opposed to how they were guided to judge the answer strings). Second, we wanted to discover if there were specific aspects of the task that could be improved for future evaluations.

Our observations suggest that the assessors understood their task and could do it. Several commented on how enjoyable they found it. Since this was the first time they had done this type of assessing, they were sometimes surprised (or amused or frustrated) by what the systems returned. For some questions, most of the answer strings were associated with the document that contained the answer, but none or few of the strings actually contained the answer. For example no one successfully extracted Marlon Brando as the actor who played the part of the Godfather in the movie "The Godfather" (question 77). Frequently strings would be truncated immediately before the answer. The answer pool for question 1, "Who is the author of the book, 'The Iron Lady: A Biography of Margaret Thatcher'?" contained not only the string "The Iron Lady; A Biography of Margaret Thatcher by" but also responses of Ronald Reagan, Giroux, Deirdre Bair, Alfred A. Knopf, Lady Dorothy Neville, and Samuel Beckett. The systems' lack of true understanding of the text sometimes led to amusing responses. One response to question 21, "Who was the first American in space?" was Jerry Brown, taken from document LA110190-0188 which says "As for Wilson himself, he became a senator by defeating Jerry Brown, who has been called the first American in space." (The answer string was marked as incorrect.) A similar response was returned for question 196, "Who wrote 'Hamlet'?": "'Hamlet,' directed by Franco Zeffirelli and written by . . . well, you know." (This response was also judged as incorrect.)

Because we told the assessors the NIST TREC team created the answer pools for the training questions, many assessors believed we had a hand in creating the responses to the test questions as well, and didn't really seem to believe our claims of complete ignorance as to how the systems had created their responses. The assessors often picked apart the questions word for word looking for "tricks." They were unhappy with question 93, "Who first circumnavigated the globe?" because their research outside of NIST showed that Magellan died before the trip was completed. We had them accept Magellan as the answer anyway. They also objected to questions 131 and 184, which we subsequently removed from the test set.

The assessors generally followed the assessing guidelines, though we did find some common patterns of mistakes. As alluded to earlier, 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 question 146, "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 persisted.

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 question 193, "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 *hated* this, and found it very difficult to do. Giving credit for both a right answer when it was not supported by the document and a wrong answer when it was supported by the document is too weird and does not reflect a real user task. This became one of the largest sources of inconsistency among the assessors' judgments as some assessors stopped checking document context altogether and others failed to mark such an answer string as correct.

2.3.2 Differences among assessors

Many differences among assessors were not caused by mistakes, however, 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 question 54, "When did Nixon die?" but the other assessor required the year as well. Year-only is almost always acceptable for historical questions, and even decade- or century-only is acceptable if the event in question is ancient enough. For question 160, "When did French revolutionaries storm the Bastille?", "July 14" and "1789" (as well as "July 14, 1789") were all considered acceptable for some assessors. Similar issues arise with locations. For question 191, "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, for question 40, "Who won the Nobel Peace Prize in 1991?" the full name of the recipient is Aung San Suu Kyi. Some assessors accepted all of "Aung San Suu Kyi," "Suu Kyi," "San Suu Kyi," and "Kyi."

These examples illustrate the myth of the obvious answer. It is pointless to try to create a set of rules that specify exactly what is acceptable or unacceptable in all cases, since the granularity of an acceptable response really does depend on the question and on the person receiving the answer. Even if it were possible to get assessors to judge exactly the same way all the time, that would defeat the purpose of the evaluation. Eventual end-users of the technology will have different opinions and expectations, and the technology will have to be able to accommodate those differences to be useful.

2.4 Scoring the results

With three sets of judgments for each question, we were able to form a high-quality judgment set as the final result of the assessment process. For each question, the three sets of assessor judgments were compared, and any [string, document] pair that had two different judgments was reviewed by an adjudicator (Voorhees). The adjudicator's role was not to provide a fourth judgment, but rather to decide if the differences in judgments were caused by differences of opinions or misapplication of the assessing guidelines. If a difference was a matter of opinion, the judgment of the majority of the assessors was used, even if the adjudicator would have judged it differently. If the difference was caused by an incorrect application of the judging guidelines, or caused the judgments to be inconsistent across the set of strings in the pool, the adjudicator overruled the majority opinion. Appendix C gives the total number of pairs in the pool, the number disagreed on, and the number of times the majority opinion was overruled by the adjudicator for each question.

On average, 6% of the answer strings that were judged were disagreed on, and 16% of the disagreements had the majority opinion overruled by the adjudicator. 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 [3], 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. The table in Appendix C also gives the overlap for each question.

The QA track runs were scored using the adjudicated judgment set. Recall that a run consisted of a ranked list of up to five [answer string, document] pairs for each question, and that every pair from every run was judged. For each run, the scoring routine read the answer pairs for the current question in order. The current answer pair was located in the adjudicated judgment file. If the pair was judged as correct, the reciprocal of the current rank was computed and added to a running sum of reciprocal ranks for this run; the remaining responses for this question were ignored. If no correct pair was found for the question, the reciprocal was set to 0. After all questions were processed, the mean reciprocal rank was computed from the running sum, and both the mean and the number of questions for which no correct answer was found were written out.

The reciprocal rank has several advantages as a scoring metric. It is closely related to the average precision measure used extensively in document retrieval. It is bounded between 0 and 1, inclusive, and averages well. A run is penalized for not retrieving any correct answer for a question, but not unduly so. The measure also has some drawbacks that perhaps should be addressed in future evaluations. The score for an individual question can take on only six values (0, .2, .25, .33, .5, 1), so it is unlikely that parametric statistical significance tests would be appropriate for this task. Question answering systems are given no credit for retrieving multiple (different) correct answers. Also, since the track required at least one response for each question, systems could receive no credit for realizing they did not know the answer.

3 The Reliability of System Comparisons

One of the primary ways TREC has been successful in improving document retrieval performance is by creating appropriate test collections for researchers to use when developing their systems. Unfortunately, the TREC-8 QA track did not create a comparable QA test collection. The unit that was judged for correctness was the entire answer string. This is not comparable to judging documents in document retrieval test collections because different question answering runs almost never return exactly the same answer strings. Developing a true equivalent of document retrieval's relevance judgment sets for question answering is a high priority research problem, but for now it remains unsolved.

A second test collection issue is the reliability of the comparisons between judged runs. Since the preceding section makes it clear that assessor opinions do differ even for the simple questions with "obvious" answers that were used as test questions, there is little hope that more carefully defined assessor instructions or more proscribed question selection procedures will eliminate inconsistencies in judgments among assessors. We must therefore ensure that the relative effectiveness of two question answering strategies is insensitive to modest changes in the judgment set since no one judgment set represents a gold standard answer key. There is reason to believe that this may be the case. Relevance judgments for documents are also known to vary across different assessors [4], yet relative retrieval effectiveness is stable despite the differences [6]. This section investigates whether the stability of document retrieval system evaluation is also true for question answering system evaluation.

3.1 Defining different judgment sets

As described earlier, the judgment set used to score the TREC-8 question answering systems was created such that each individual question was judged by three different assessors. Any differences in the judgments among the three assessors were reviewed by an adjudicator who let the majority's judgment stand unless judgments were inconsistent across different answer strings or the assessing guidelines were not followed. This process reduces the number of blunders in the final judgment set and increases the likelihood that the stated assessing guidelines were actually followed, though it also more than triples the cost of creating a judgment set as compared to using a single assessor's judgments for each question. Since different judgments are available for each question, it is possible to directly measure the effect different judgments have on the systems' scores. Two questions need to be addressed: whether judgment sets that use a single assessor for a question ("one-judge qrels") are equivalent to one another, and if so, whether a one-judge qrels is an adequate substitute for the adjudicated qrels.

The procedure used to measure the effects of different judgment sets on final scores was identical to the one used to gauge the effect of differences in relevance judgments in document retrieval system evaluation [6], namely quantifying the changes in *system rankings* when different qrels are used to score the runs. A system ranking is a list of the systems under consideration sorted by decreasing mean reciprocal rank. We used a correlation based on Kendall's tau [5] as the measure of association between two rankings. Kendall's tau computes the distance between two rankings as the minimum number of pairwise adjacent swaps to turn one ranking into the other. The distance is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0, the correlation between a ranking and its perfect inverse is −1.0, and the expected correlation of two rankings chosen at random is 0.0.

With three judgments for each of 198 questions, we can form 3^{198} different one-judge qrels for this QA task. We generated a sample of 100,003 of these one-judge qrels[3] by randomly selecting one of the assessors for each question and combining the selected judgments into one qrels. We then scored 41 of the QA runs using each of the 100,003 qrels (the remaining four runs were clear outliers and we removed them for this part of the investigation). We calculated the sample mean and standard deviation of the mean reciprocal rank for each run. The means are plotted in Figure 1 where the runs are sorted by decreasing mean. The error bars in Figure 1 indicate the minimum and the maximum mean reciprocal rank obtained for that run over the sample. Values for the means, standard deviations, and minimum and maximum scores qrels are given in the second column of Table 1.

In addition to the 100,003 one-judge qrels, we created four multiple-judge qrels. The first of these is the adjudicated qrels described above. The second is a straight majority opinion qrels set. This differs from the adjudicated set in that there was no overruling of the majority opinion when assessments differed. The remaining two qrels sets are the union and intersection sets. In the union qrels a response is considered to be correct if any assessor judged it correct; in the intersection qrels a response is considered to be correct if all three assessors judged it as correct.

The mean reciprocal rank scores for each of the runs for the four multiple-judge qrels are plotted along with the sample means in Figure 1, and are given in the first column of Table 1. These points demonstrate how the system ranking changes for a particular qrels versus the ranking by the mean: a run with a symbol higher than the corresponding symbol of a run to its left would be ranked differently in the particular qrels ranking. For example, the first two runs (SMUNLP2 and textract9908) would switch positions when evaluated by the adjudicated qrels set.

As is true for document retrieval evaluations, the absolute values of the scores do change when different qrels are used to evaluate the runs. The final column of Table 1 gives the number of questions whose score changes for that run depending on which individual assessor's judgments are used. However, we are interested in the effect on relative scores, which means we need to look at how the system rankings change when different qrels are used. We computed the mean of the Kendall correlations among the system rankings in two ways. In the first case, we took the mean of all pair-wise correlations in a random sample of 1000 of the one-judge rankings. In the second case, we took the mean of the Kendall's correlation between the adjudicated qrels and all 100,003 one-judge rankings. Finally we computed the correlation between the adjudicated ranking and each of the other multiple-judge rankings. The correlations are given in Table 2. The numbers in parentheses show the number of pairwise adjacent swaps a correlation represents given that there are 41 different runs being ranked. Since any two one-judge qrels are likely to contain exactly the same judgments for 1/3 of the questions on average, the qrels are not independent of one another. Thus the Kendall correlation shown may be slightly higher than it would be with completely independent qrels.

The correlations in the top part of Table 2 show that QA system rankings produced from one-judge qrels are at least as stable as document retrieval system rankings in the face of changes in judgments. There are minor differences in the rankings, but most of those differences are caused by runs whose mean reciprocal rank scores are very close. This answers our first question in the affirmative. One-judge rankings are essentially equivalent with one another for the purpose of comparative evaluation of QA systems.

The second half of Table 2 suggests that one-judge qrels are also equivalent to the expensive adjudicated qrels. As can be seen from Figure 1, the adjudicated score for a run always lies within the boundaries of the minimum and maximum scores obtained on the sample of one-judge qrels. This is not true for the union and intersection qrels, which is a difference between QA evaluation and document retrieval evaluation.

[3] In the document retrieval study, three of the qrels sets were special cases. They are not special cases for the QA task, but existing code computed them, so they were left in.

Table 1: Variation in mean reciprocal rank by judgment set ("qrels"). The runs are ordered by decreasing mean reciprocal rank using the official adjudicated qrels (Adj). The next three columns give the score obtained using the Majority (Maj), Union (Union), and Intersection (Inter) qrels. The next four columns give the distribution of the mean reciprocal rank over the set of 100,003 single-judge qrels sets: the sample mean (Mean), the sample standard deviation (σ), the minimum (Min), and the maximum (Max). The final column gives the number of questions whose score varies depending on which single judge assessments are used.

	Qrels				Over 100,003				#
Run	Adj	Maj	Union	Inter	Mean	σ	Min	Max	Qs
textract9908	.660	.622	.705	.518	.617	.013	.564	.676	48
SMUNLP2	.646	.648	.667	.566	.627	.010	.583	.662	28
SMUNLP1	.555	.530	.586	.395	.504	.014	.446	.557	49
attqa250p	.545	.549	.596	.485	.543	.010	.502	.582	32
GePenn	.510	.501	.541	.446	.496	.009	.458	.529	32
attqa250e	.483	.478	.528	.434	.480	.009	.439	.517	31
uwmt8qa1	.471	.468	.504	.410	.462	.008	.431	.492	37
mds08q1	.453	.452	.495	.405	.452	.008	.418	.486	33
xeroxQA8lC	.453	.450	.493	.374	.440	.010	.396	.480	38
nttd8ql1	.439	.444	.497	.380	.441	.010	.398	.478	43
MTR99250	.434	.429	.484	.327	.415	.012	.367	.465	44
IBMDR992	.430	.438	.461	.388	.429	.008	.394	.459	21
IBMVS992	.395	.402	.435	.339	.393	.009	.355	.429	30
INQ635	.383	.380	.430	.297	.369	.010	.326	.408	46
nttd8ql4	.371	.364	.415	.276	.353	.011	.305	.402	45
attqa50e	.356	.360	.387	.315	.355	.008	.323	.384	25
LimsiLC	.341	.338	.384	.298	.340	.008	.308	.376	28
INQ639	.336	.328	.377	.281	.330	.009	.295	.365	31
CRDBASE250	.319	.319	.347	.265	.310	.008	.277	.343	28
IBMDR995	.319	.307	.345	.215	.288	.011	.233	.334	40
xeroxQA8sC	.317	.314	.363	.232	.303	.011	.257	.347	39
umdqa	.298	.291	.344	.239	.293	.009	.257	.330	35
clr99s	.281	.272	.322	.222	.273	.009	.238	.309	34
MTR99050	.281	.253	.319	.191	.257	.010	.217	.303	38
IBMVS995	.280	.283	.313	.212	.269	.010	.231	.306	31
nttd8qs1	.273	.264	.306	.200	.257	.009	.211	.293	36
CRL250	.268	.259	.292	.199	.250	.009	.211	.290	29
UIowaQA1	.267	.264	.300	.245	.270	.007	.246	.298	17
attqa50p	.261	.228	.277	.149	.218	.011	.174	.262	35
nttd8qs2	.259	.252	.283	.176	.238	.009	.197	.276	36
CRL50	.220	.205	.233	.147	.195	.008	.160	.226	29
INQ634	.191	.179	.220	.131	.178	.008	.145	.212	29
CRDBASE050	.158	.146	.195	.114	.152	.008	.122	.186	31
INQ638	.126	.124	.150	.094	.123	.006	.098	.146	22
Scai8QnA	.121	.119	.138	.087	.114	.006	.089	.137	20
shefinq250	.111	.106	.121	.071	.099	.007	.071	.121	10
shefatt250	.096	.086	.101	.076	.088	.005	.076	.101	5
NTU99	.087	.088	.101	.057	.083	.006	.060	.101	13
shefinq50	.081	.061	.086	.040	.063	.007	.040	.086	9
shefatt50	.071	.066	.076	.056	.066	.005	.056	.076	4
UIowaQA2	.060	.056	.082	.041	.059	.006	.041	.082	14

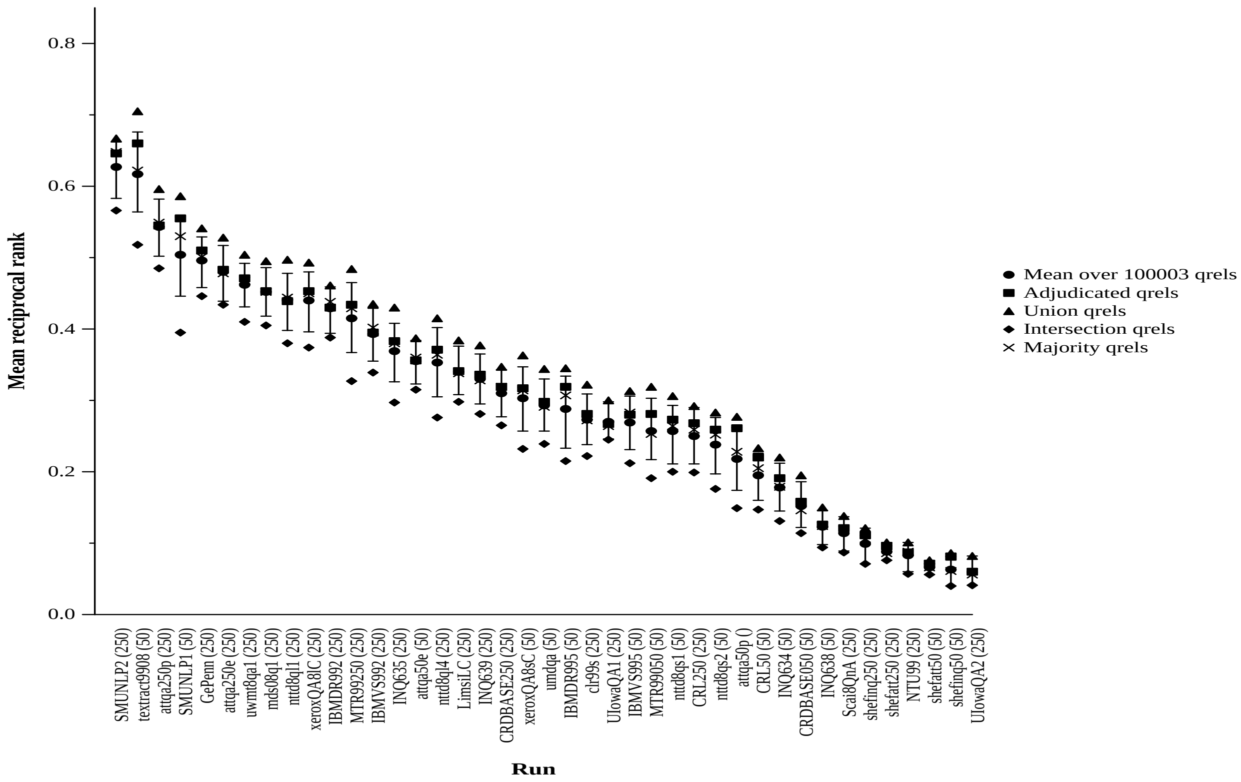

Figure 1: Sample mean, min, and max of the mean reciprocal rank computed for QA runs over a sample of 100,003 one-judge qrels. Also plotted are the mean reciprocal rank for the adjudicated, majority, union, and intersection qrels. Runs are labeled as either 50 byte limit (50) or 250 byte limit (250).

Table 2: Kendall correlation (τ) of system rankings and corresponding number of pairwise adjacent swaps produced by different qrels sets. With 41 systems, there is a maximum of 820 possible pairwise adjacent swaps.

	Mean τ		Min τ		Max τ	
in subsample	.9632	(15.1)	.9171	(34)	.9976	(1)
with adjudicated	.9563	(17.9)	.9146	(35)	.9878	(5)

a) correlations for one-judge rankings

	τ	
majority	.9683	(13)
union	.9780	(9)
intersection	.9146	(35)
a 1-judge qrels	.9683	(13)

b) correlations with the adjudicated ranking

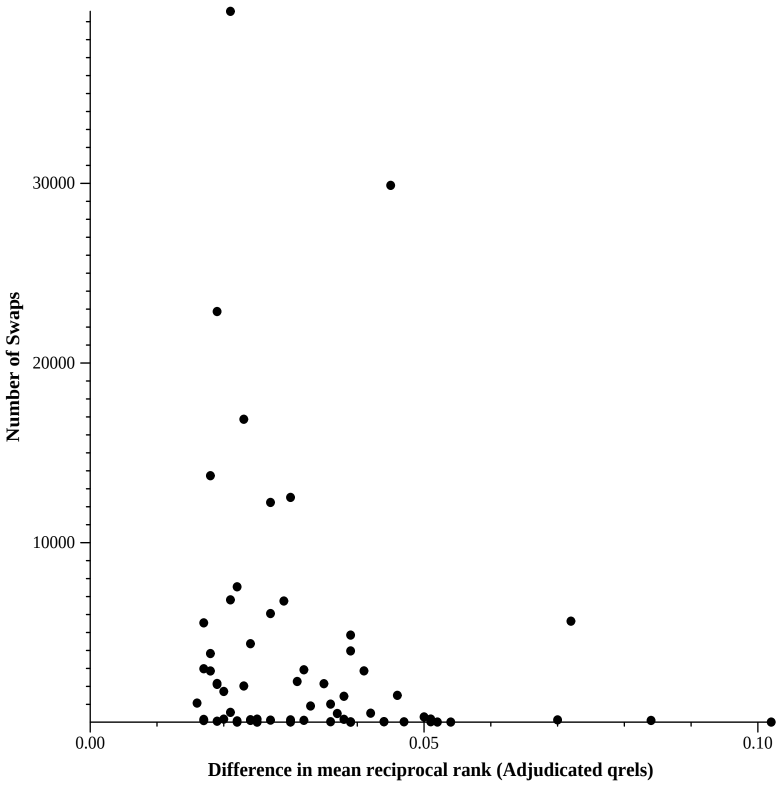

Figure 2: Number of swaps versus difference in mean reciprocal rank as computed using the Adjudicated qrels. Only pairs of systems that have a difference greater than .015 and that swapped more than 10 times are plotted.

3.2 Estimating the likelihood of a swap

While statements regarding the average stability of a set of runs are nice, researchers are often more interested in knowing the likelihood that any two particular runs will be ranked in a different order if the judgment sets change. Define a *swap* to be the situation in which one qrels ranks run i before run j and a second qrels ranks run j before rank i. We can use the sample of 100,003 one-judge qrels to count how often each pair of runs swaps.

Let $B[i, j]$ be the number of qrels that cause system i to evaluate as better then system j. Then the number of swaps for i and j is the smaller of $B[i, j]$ and $B[j, i]$ (assuming the larger number represents the "true" ranking). For the 820 pairs of runs under consideration, 123 had at least one swap. For pairs that had a difference greater than .015 in their mean reciprocal ranks according to the adjudicated qrels, only 85 ever swapped. As Table 1 shows, the standard deviation of the sample mean for a single run approaches .015 when qrels sets vary, so runs with smaller differences must be regarded as equivalent.

Figure 2 plots the number of swaps for a run pair against the difference in their mean reciprocal ranks according to the adjudicated qrels. Pairs with differences of less than .015 are not plotted, nor are pairs that swapped fewer than 10 times, leaving 66 pairs plotted. Points plotted furthest from the origin of the graph are of interest because they represent pairs of systems that either swap often or have large differences in the mean reciprocal rank. For example, the extreme point on the x-axis represents a difference of .102 in the mean reciprocal ranks between the SMUNLP1 run and the mds08q1 run. This pair of runs swapped only 14 times, however, thus having an estimated probability of a swap of only .0003 (14/50,000). At the other extreme, the umdqa and IBMDR995 runs swapped 39,567 times (estimated probability of .791) with a difference of .021 in the reciprocal ranks. While the difference in scores is (inversely) correlated with the probability that a pair of runs will swap, it is clearly not the only factor. Run pairs with much smaller differences between their scores swapped many fewer times than these two runs.

3.3 Question set size

It is important to remember that the system rankings used as a basis of this analysis were computed using the mean score over 198 questions. Using averages over a sufficient number of questions is vital to obtaining a stable evaluation.

What is sufficient? From a stability viewpoint, more questions in a test set is always better than fewer questions. But a test set with more questions is also more expensive to build than a set with fewer questions. With so little experience with the task, it is premature to set a final figure, though the TREC-8 evaluation does provide interesting data points. All of our analysis thus far has shown that 200 (or 198) is sufficient. To set a lower bound, note the last column in Table 1 that shows the number of questions whose score changed when the assessor changed for each run. Since some runs had almost 50 questions that were affected by judgment differences, a test set should probably have at least 100 questions.

4 Conclusion

The first running of any TREC track is more a test of the evaluation methodology used in the track than of the participating systems. This is particularly true with the TREC-8 QA track, which used an evaluation methodology based on human assessors for the question answering task. This paper validated the methodology used by showing it was both appropriate and effective.

The general question answering task was deliberately simplified for the TREC-8 track by constraining the questions to be fact-based, short-answer questions. The results of the track make it clear, however, that even for this highly-constrained version of the task legitimate differences of opinions exist as to whether a supplied answer string actually answers the question. Assessors differ on how much of a name is required, and on the granularity of times and locations. Having the evaluation accommodate differences of opinion in the answer keys reflects a requirement of the real problem; if assessors have different opinions on what constitutes an answer then eventual end-users of the technology will have different opinions as well. The technology must be able to accommodate user differences to be useful.

An evaluation is effective if the conclusions that can be drawn from it are meaningful and valid. The purpose of the TREC-8 evaluation was to compare different technologies for (a limited version of) the question answering task. The QA evaluation produced a nice spread of scores, and those runs that intuitively seemed better got higher scores. The comparisons are valid to the extent that they are stable under changes in the judgments that produce the scores. Indeed, our analysis suggests that the expensive adjudicated judgment set used in the track can be replaced with a single-opinion judgment, since the system rankings produced by single-opinion judgment sets are equivalent to one another and to the ranking produced by the adjudicated judgment set.

While on the whole the TREC-8 QA evaluation was sound, there was at least one problem with this implementation of the track. Since the track guidelines required a document be returned but allowed answer generation (as opposed to simple extraction), NIST made the attempt to judge as correct both strings that contained a wrong answer that had been correctly extracted from a mistaken document, and strings that contained a correct answer even though that answer could not be obtained from the information contained within the cited document. In retrospect, this was a bad decision. Taking document context into account only for certain situations was difficult for the assessors to do, and was one of the largest contributors to differences in judgments among the assessors. It also does not correspond well to any real user task. In future evaluations the task should either be a true question answering task wherein the system generates a response from any resource available to it and is evaluated strictly on whether the returned string contains a correct answer[4] or an information extraction task in which the system extracts a response from text and is evaluated on successful extraction.

Another issue that must eventually be addressed is the fact that the current methodology does not create a true test collection. Currently judgments are based on entire answer strings because it is not yet clear how to map specific answer strings into more general answers. Unfortunately, this renders the judgments unsuitable for anything other than comparing the specific runs used to build the answer pools. The judgments do not constitute a generic judgment file for the QA task because the amount of overlap in answer strings

[4]Some mechanism whereby the assessors learn the set of correct answers is required.

across runs is quite small. The benefits of large-scale evaluation for question answering technology will not be fully realized until true test collections can be devised.

References

[1] Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. SMART high precision: TREC 7. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 285–298, August 1999. NIST Special Publication 500-242. Electronic version available at http://trec.nist.gov/pubs.html.

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

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

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

[5] Alan Stuart. Kendall's tau. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 367–369. John Wiley & Sons, 1983.

[6] Ellen M. Voorhees. 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, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, Melbourne, Australia, August 1998. ACM Press, New York.

A The Test Set of 200 Questions

1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Qintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?
9. How far is Yaroslavl from Moscow?
10. Name the designer of the shoe that spawned millions of plastic imitations, known as "jellies".
11. Who was President Cleveland's wife?
12. How much did Manchester United spend on players in 1993?
13. How much could you rent a Volkswagen bug for in 1966?
14. What country is the biggest producer of tungsten?
15. When was London's Docklands Light Railway constructed?
16. What two US biochemists won the Nobel Prize in medicine in 1992?
17. How long did the Charles Manson murder trial last?
18. Who was the first Taiwanese President?
19. Who was the leader of the Branch Davidian Cult confronted by the FBI in Waco, Texas in 1993?
20. Where is Inoco based?
21. Who was the first American in space?
22. When did the Jurassic Period end?
23. When did Spain and Korea start ambassadorial relations?
24. When did Nixon visit China?
25. Who was the lead actress in the movie "Sleepless in Seattle"?
26. What is the name of the "female" counterpart to El Nino, which results in cooling temperatures and very dry weather?
27. Where did the 6th annual meeting of Indonesia-Malaysia forest experts take place?
28. Who may be best known for breaking the color line in baseball?
29. What is the brightest star visible from Earth?
30. What are the Valdez Principles?
31. Where was Ulysses S. Grant born?
32. Who received the Will Rogers Award in 1989?
33. What is the largest city in Germany?
34. Where is the actress, Marion Davies, buried?
35. What is the name of the highest mountain in Africa?
36. In 1990, what day of the week did Christmas fall on?
37. What was the name of the US helicopter pilot shot down over North Korea?
38. Where was George Washington born?
39. Who was chosen to be the first black chairman of the military Joint Chiefs of Staff?
40. Who won the Nobel Peace Prize in 1991?
41. What is the legal blood alcohol limit for the state of California?

42. What was the target rate for M3 growth in 1992?
43. What costume designer decided that Michael Jackson should only wear one glove?
44. Who is the director of the international group called the Human Genome Organization (HUGO) that is trying to coordinate gene-mapping research worldwide?
45. When did Lucelly Garcia, a former ambassador of Columbia to Honduras, die?
46. Who is the mayor of Marbella?
47. What company is the largest Japanese ship builder?
48. Where is the massive North Korean nuclear complex located?
49. Who fired Maria Ybarra from her position in San Diego council?
50. When was Dubai's first concrete house built?
51. Who is the president of Stanford University?
52. Who invented the road traffic cone?
53. Who was the first doctor to successfully transplant a liver?
54. When did Nixon die?
55. Where is Microsoft's corporate headquarters located?
56. How many calories are there in a Big Mac?
57. What is the acronym for the rating system for air conditioner efficiency?
58. Name a film that has won the Golden Bear in the Berlin Film Festival?
59. Who was President of Costa Rica in 1994?
60. What is the fare cost for the round trip between New York and London on Concorde?
61. What brand of white rum is still made in Cuba?
62. What is the name of the chronic neurological autoimmune disease which attacks the protein sheath that surrounds nerve cells causing a gradual loss of movement in the body?
63. What nuclear-powered Russian submarine sank in the Norwegian Sea on April 7, 1989?
64. Who is the voice of Miss Piggy?
65. Name a country that is developing a magnetic levitation railway system?
66. Name the first private citizen to fly in space.
67. What is the longest river in the United States?
68. What does El Nino mean in spanish?
69. Who came up with the name, El Nino?
70. How many lives were lost in the China Airlines' crash in Nagoya, Japan?
71. In what year did Joe DiMaggio compile his 56-game hitting streak?
72. When did the original Howdy Doody show go off the air?
73. Where is the Taj Mahal?
74. Who leads the star ship Enterprise in Star Trek?
75. What cancer is commonly associated with AIDS?
76. In which year was New Zealand excluded from the ANZUS alliance?
77. Who played the part of the Godfather in the movie, "The Godfather"?
78. Which large U.S. city had the highest murder rate for 1988?
79. What did Shostakovich write for Rostropovich?
80. What is the name of the promising anticancer compound derived from the pacific yew tree?
81. How many inhabitants live in the town of Ushuaia?
82. How many consecutive baseball games did Lou Gehrig play?
83. What is the tallest building in Japan?

84. Which country is Australia's largest export market?
85. Which former Ku Klux Klan member won an elected office in the U.S.?
86. Who won two gold medals in skiing in the Olympic Games in Calgary?
87. Who followed Willy Brandt as chancellor of the Federal Republic of Germany?
88. What is Grenada's main commodity export?
89. At what age did Rossini stop writing opera?
90. Who is the founder of Scientology?
91. Which city in China has the largest number of foreign financial companies?
92. Who released the Internet worm in the late 1980s?
93. Who first circumnavigated the globe?
94. Who wrote the song, "Stardust"?
95. What country is the worlds leading supplier of cannabis?
96. What time of day did Emperor Hirohito die?
97. How large is the Arctic refuge to preserve unique wildlife and wilderness value on Alaska's north coast?
98. Where is the highest point in Japan?
99. What is the term for the sum of all genetic material in a given organism?
100. What is considered the costliest disaster the insurance industry has ever faced?
101. How many people live in the Falklands?
102. Who is the Voyager project manager?
103. How many people died when the Estonia sank in 1994?
104. What language is most commonly used in Bombay?
105. How many people does Honda employ in the U.S.?
106. What is the second highest mountain peak in the world?
107. When was China's first nuclear test?
108. Which company created the Internet browser Mosaic?
109. Where does Buzz Aldrin want to build a permanent, manned space station?
110. Who killed Lee Harvey Oswald?
111. How long does it take to travel from Tokyo to Niigata?
112. Who is the President of Ghana?
113. What is the name of the medical condition in which a baby is born without a brain?
114. How much stronger is the new vitreous carbon material invented by the Tokyo Institute of Technology compared with the material made from cellulose?
115. What is Head Start?
116. Which team won the Super Bowl in 1968?
117. What two researchers discovered the double-helix structure of DNA in 1953?
118. What percentage of the world's plant and animal species can be found in the Amazon forests?
119. What Nobel laureate was expelled from the Philippines before the conference on East Timor?
120. Who held the endurance record for women pilots in 1929?
121. Who won the first general election for President held in Malawi in May 1994?
122. Who is section manager for guidance and control systems at JPL?
123. How many Vietnamese were there in the Soviet Union?
124. What was Agent Orange used for during the Vietnam War?
125. In what city is the US Declaration of Independence located?
126. When did Israel begin turning the Gaza Strip and Jericho over to the PLO?

127. Which city has the oldest relationship as a sister-city with Los Angeles?
128. Who was the second man to walk on the moon?
129. How many times was pitcher, Warren Spahn, a 20-game winner in his 21 major league seasons?
130. When was Yemen reunified?
131. Which Japanese car maker had its biggest percentage of sale in the domestic market?
132. What is the capital of Uruguay?
133. What is the name for the technique of growing certain plants in soils contaminated with toxic metals, wherein the plants take up the toxic metals, are harvested, and the metals recovered for recycling?
134. Where is it planned to berth the merchant ship, Lane Victory, which Merchant Marine veterans are converting into a floating museum?
135. What famous communist leader died in Mexico City?
136. Who is the Queen of Holland?
137. Who is the president of the Spanish government?
138. What is the name of the normal process in all living things, including humans, in which cells are programmed to "commit suicide"?
139. How many people did the United Nations commit to help restore order and distribute humanitarian relief in Somalia in September 1992?
140. How many people on the ground were killed from the bombing of Pan Am Flight 103 over Lockerbie, Scotland, December 21, 1988?
141. What is the duration of the trip from Bristol to London by rail?
142. What is the population of Ulan Bator, capital of Mongolia?
143. Where does most of the marijuana entering the United States come from?
144. How many megawatts will the power project in Indonesia, built by a consortium headed by Mission Energy of US, produce?
145. What did John Hinckley do to impress Jodie Foster?
146. In what year did Ireland elect its first woman president?
147. Who is the prime minister of Japan?
148. How many soldiers were involved in the last Panama invasion by the United States of America?
149. Where is the Bulls basketball team based?
150. What is the length of border between the Ukraine and Russia?
151. Where did Dylan Thomas die?
152. How many people live in Tokyo?
153. What is the capital of California?
154. How many Grand Slam titles did Bjorn Borg win?
155. Who was the Democratic nominee in the American presidential election?
156. When was General Manuel Noriega ousted as the leader of Panama and turned over to U.S. authorities?
157. Where is Dartmouth College?
158. How many mines can still be found in the Falklands after the war ended?
159. Why are electric cars less efficient in the north-east than in California?
160. When did French revolutionaries storm the Bastille?
161. How rich is Bill Gates?
162. What is the capital of Kosovo?
163. What state does Charles Robb represent?
164. Who is the leading competitor of Trans Union Company?
165. Which type of submarine was bought recently by South Korea?

166. When did communist control end in Hungary?
167. What nationality is Pope John Paul II?
168. Who was the captain of the tanker, Exxon Valdez, involved in the oil spill in Prince William Sound, Alaska, 1989?
169. Whom did the Chicago Bulls beat in the 1993 championship?
170. Who was President of Afghanistan in 1994?
171. Who is the director of intergovernmental affairs for the San Diego county?
172. Where is the Keck telescope?
173. How many moons does Jupiter have?
174. When did Jaco Pastorius die?
175. When did beethoven die?
176. How many people in Tucson?
177. How tall is Mt. Everest?
178. What is the capital of Congo?
179. What is the capital of Italy?
180. What is the capital of Sri Lanka?
181. What novel inspired the movie BladeRunner?
182. What was the first Gilbert and Sullivan opera?
183. What was the name of the computer in "2001: A Space Odyssey"?
184. When was Queen Victoria born?
185. When was the battle of the Somme fought?
186. Where did the Battle of the Bulge take place?
187. Where was Lincoln assassinated?
188. When was the women's suffrage amendment ratified?
189. Where is Qatar?
190. Where is South Bend?
191. Where was Harry Truman born?
192. Who was Secretary of State during the Nixon administration?
193. Who was the 16th President of the United States?
194. Who wrote "The Pines of Rome"?
195. Who wrote "Dubliners"?
196. Who wrote "Hamlet"?
197. What did Richard Feynman say upon hearing he would receive the Nobel Prize in Physics?
198. How did Socrates die?
199. How tall is the Matterhorn?
200. How tall is the replica of the Matterhorn at Disneyland?

Table 3: Sources of the 198 final test set questions. Two questions (131 and 184) were dropped from the evaluation after assessors determined there was no answer in the document collection. “Qid” is the question identifier. The remaining columns indicate the source of the corresponding question: “FAQ” for questions taken from the logs of the FAQFinder system, “Part” for questions submitted by TREC participants or the NIST TREC team, and “Assess” for questions submitted by the NIST assessors.

Qid	FAQ	Part	Assess	Qid	FAQ	Part	Assess	Qid	FAQ	Part	Assess	Qid	FAQ	Part	Assess
1			X	51		X		101		X		152		X	
2			X	52		X		102		X		153		X	
3		X		53			X	103		X		154		X	
4		X		54		X		104		X		155		X	
5		X		55		X		105		X		156			X
6		X		56		X		106		X		157		X	
7		X		57			X	107		X		158		X	
8			X	58		X		108		X		159			X
9		X		59			X	109		X		160		X	
10			X	60		X		110		X		161		X	
11		X		61		X		111		X		162		X	
12		X		62			X	112			X	163		X	
13		X		63			X	113			X	164		X	
14			X	64		X		114		X		165		X	
15		X		65			X	115		X		166			X
16		X		66			X	116			X	167		X	
17			X	67		X		117			X	168			X
18		X		68		X		118		X		169		X	
19			X	69		X		119		X		170			X
20		X		70		X		120		X		171		X	
21		X		71			X	121			X	172		X	
22		X		72		X		122		X		173	X		
23		X		73		X		123		X		174	X		
24		X		74		X		124		X		175	X		
25		X		75		X		125		X		176	X		
26		X		76		X		126			X	177	X		
27		X		77			X	127		X		178	X		
28			X	78			X	128		X		179	X		
29		X		79		X		129			X	180	X		
30		X		80			X	130		X		181	X		
31		X		81		X		132		X		182	X		
32		X		82			X	133			X	183	X		
33		X		83		X		134			X	185	X		
34			X	84		X		135		X		186	X		
35			X	85		X		136		X		187	X		
36		X		86		X		137		X		188	X		
37		X		87			X	138			X	189	X		
38		X		88			X	139		X		190	X		
39		X		89		X		140			X	191	X		
40		X		90		X		141		X		192	X		
41		X		91		X		142			X	193	X		
42		X		92		X		143		X		194	X		
43		X		93		X		144		X		195	X		
44			X	94			X	145		X		196	X		
45		X		95		X		146		X		197		X	
46		X		96		X		147		X		198		X	
47			X	97		X		148		X		199		X	
48			X	98		X		149		X		200		X	
49		X		99			X	150		X					
50			X	100			X	151			X				

B Written Instructions to the Assessors

Today's search systems take a question and return a list of documents likely to contain an answer to the question. The user of the system must then read the documents to find the desired answer within them, if it's there. This can be a very tedious, time-consuming process, and frustrating process.

It would be better if the system returned smaller pieces of text—a few words or a sentence or at most a paragraph believed to contain the answer. Then the user would have less reading to do in order to see if any of the pieces contained an answer.

Such improved systems exist today in experimental versions. In order to know how good a job such improved search systems are doing we need to judge whether, given a question, the systems return pieces of text that are responsive (i.e., you can recognize the answer in the piece) to the question. Your task will be to make these judgments.

For each question you will be given several pieces of text varying in size from a few words to a paragraph. The experimental search systems returned these pieces of text in response to this question. Your job is to decide whether each piece of text really contains some words that answer the question. Here is how you should proceed.

For each question:

1. Read the question carefully, then
2. Find the answer by skimming through the document and answer strings. It may be the case that no one retrieved the answer. If this happens, see Dawn.
3. For each answer string, read the piece of text and judge whether it contains a valid answer to the question.
   - If the answer string is the answer, judge it correct (yes).
   - If the answer string contains the answer plus supporting text, judge it correct (yes).
   - If the answer string contains the answer plus miscellaneous other stuff, judge it correct (yes).
   - If the answer string contains the answer plus other text that interferes with recognizing the answer, then you should decide how much interference there is, but it is probably incorrect.
   - If the answer string does not contain the answer, then judge it incorrect (no).

Notes:

- It's possible that there may be more than one answer. Check the document to see whether an answer string contains a valid answer.
- Judge the answer in the context of the document – even if the document gives an answer that you believe is wrong, judge the answer string on the basis of what the document says anyway.
- It is up the the assessor to decide if "partial answers" are responsive. Example: Last name only may be acceptable for "who" questions.

C Assessor Agreement per Question

Table 4: Number of answer strings judged (J); number of answers that were judged differently by different assessors (D); number of majority opinions overruled in adjudication (O); and overlap of the correct sets across assessors (OL). The overlap is undefined if no answers are judged as correct.

Qid	J	D	O	OL	Qid	J	D	O	OL	Qid	J	D	O	OL	Qid	J	D	O	OL
1	189	17	3	.66	51	200	12	1	.83	101	199	1	0	.94	152	200	0	0	1.0
2	175	0	0	1.0	52	197	7	0	.78	102	202	31	1	0.0	153	200	7	3	.22
3	195	47	2	.49	53	202	3	1	.88	103	197	22	1	.49	154	196	15	1	.53
4	189	1	1	.89	54	189	28	1	0.0	104	184	1	1	.86	155	193	42	2	.43
5	202	1	0	.97	55	185	23	3	.28	105	202	11	0	.27	156	198	16	0	0.0
6	185	9	0	.64	56	192	1	0	.94	106	191	14	2	.62	157	197	17	4	.77
7	191	6	1	.70	57	196	20	2	.50	107	188	3	0	.77	158	194	3	0	.87
8	191	19	0	.10	58	183	14	8	.85	108	203	27	10	0.0	159	176	2	0	.90
9	185	2	0	.80	59	202	19	2	.60	109	198	12	3	.43	160	187	26	2	.62
10	180	9	0	.70	60	194	12	4	.64	110	195	15	4	.75	161	175	17	11	.48
11	200	0	0	1.0	61	183	4	2	.73	111	188	5	0	.62	162	202	11	5	.82
12	197	0	0	1.0	62	186	8	0	.73	112	205	9	0	.90	163	199	31	12	.61
13	198	4	1	.85	63	195	3	0	.97	113	199	5	0	.77	164	195	30	14	.29
14	191	37	1	.31	64	194	1	0	.95	114	184	1	0	.92	165	182	1	0	.89
15	187	4	0	.73	65	187	30	15	.35	115	179	18	2	.10	166	193	7	0	.30
16	189	1	0	.95	66	178	1	0	.96	116	199	12	0	0.0	167	195	8	2	0.0
17	197	2	1	.87	67	192	6	1	.40	117	193	32	4	.53	168	202	4	0	.87
18	190	64	0	.36	68	174	3	0	.90	118	177	13	1	.61	169	198	0	0	—
19	192	4	0	.92	69	191	4	1	.60	119	187	4	2	.64	170	197	15	1	.83
20	172	10	0	.44	70	199	11	0	.83	120	194	13	1	.64	171	195	2	2	.94
21	196	1	0	.90	71	197	7	0	.78	121	202	11	1	.74	172	198	18	0	.76
22	184	1	0	.89	72	185	0	0	1.0	122	191	1	0	.97	173	199	26	2	0.0
23	186	0	0	1.0	73	207	69	1	.10	123	197	1	0	.94	174	175	12	0	.50
24	194	37	2	.54	74	198	20	0	.50	124	192	30	3	.47	175	176	4	0	.56
25	190	4	1	.75	75	194	9	1	.47	125	189	0	0	—	176	195	0	0	1.0
26	195	1	1	.96	76	196	26	0	.50	126	196	20	1	0.0	177	196	4	0	.87
27	183	11	0	.45	77	196	0	0	—	127	186	10	2	.72	178	204	20	5	.76
28	190	10	0	.84	78	190	24	14	.43	128	201	15	0	.73	179	181	12	7	.66
29	201	12	2	.64	79	189	17	0	.39	129	197	1	0	.96	180	199	13	1	.85
30	182	4	1	.88	80	186	1	0	.99	130	192	3	0	.89	181	195	0	0	—
31	198	6	0	.79	81	194	1	0	.95	132	197	22	1	.64	182	186	4	1	.20
32	191	3	3	.67	82	195	7	1	.93	133	173	0	0	1.0	183	195	5	0	.86
33	198	10	3	.70	83	195	13	0	.68	134	185	34	0	.52	185	184	8	0	.81
34	176	5	0	.75	84	187	0	0	1.0	135	198	6	0	.50	186	175	14	3	.46
35	190	5	0	.90	85	187	34	1	.65	136	193	0	0	—	187	178	8	2	.62
36	193	2	0	.83	86	193	10	2	.57	137	184	6	3	.85	188	193	18	0	.36
37	192	6	2	.86	87	191	2	2	.60	138	202	3	0	.90	189	200	10	5	.17
38	193	1	1	.96	88	195	6	0	.82	139	189	24	0	0.0	190	193	16	5	.70
39	193	26	7	.73	89	192	5	0	.29	140	183	4	0	.93	191	194	10	4	.60
40	197	8	0	.85	90	182	7	1	.90	141	184	5	0	.44	192	190	30	4	.29
41	196	19	1	.59	91	205	11	6	.45	142	188	0	0	1.0	193	185	20	14	.09
42	189	12	1	.78	92	187	7	0	.67	143	198	18	18	.17	194	190	7	5	.63
43	198	7	0	.46	93	196	4	0	.71	144	192	32	1	.41	195	196	18	3	.57
44	199	0	0	1.0	94	197	0	0	1.0	145	177	16	3	.62	196	200	9	0	.40
45	185	42	0	.09	95	195	43	20	.16	146	197	6	1	0.0	197	191	5	0	.55
46	197	6	0	.86	96	184	2	0	.33	147	188	18	1	.84	198	169	14	0	.12
47	191	7	1	.61	97	192	12	3	0.0	148	196	3	0	.62	199	192	11	2	.76
48	185	6	2	.92	98	182	4	1	.64	149	203	10	4	.80	200	188	13	3	.50
49	193	23	8	.18	99	177	3	1	.93	150	191	2	1	.78					
50	197	4	2	.83	100	191	26	1	.65	151	190	1	0	.96					