The TREC-8 Question Answering Track

Ellen M. Voorhees, Dawn M. Tice

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

Abstract

The TREC-8 Question Answering track was the first large-scale evaluation of domain-independent question answering systems. This paper summarizes the results of the track, including both an overview of the approaches taken to the problem and an analysis of the evaluation methodology. Retrieval results for the more stringent condition in which system responses were limited to 50 bytes showed that explicit linguistic processing was more effective than the bag-of-words approaches that are effective for document retrieval. The use of multiple human assessors to judge the correctness of the systems' responses demonstrated that assessors have legitimate differences of opinion as to correctness even for fact-based, short-answer questions. Evaluations of question answering technology will need to accommodate these differences since eventual end-users of the technology will have similar differences.

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 (Buckley et al., 1999). 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.

This paper summarizes the results of the Question Answering track. The specific task that was used in the track is defined in the next section. The task was necessarily a constrained version of the general question answering task to make it both feasible for current technology and amenable to evaluation. Section 3 provides a brief overview of the techniques used to answer the questions and the effectiveness of those techniques. The most accurate systems found a correct response for more than 2/3 of the questions. Relatively simple bag-of-words approaches were adequate for finding answers when responses could be as long as a paragraph (250 bytes), but more sophisticated processing was necessary for more direct responses (50 bytes). Finally, Section 4 discusses the evaluation methodology used in the track, which included using human assessors to judge the correctness of the responses. While human assessors are part of the standard evaluation methodology for document retrieval, they have not generally been used to evaluate other kinds of natural language processing tasks. One of the main findings of the track is that assessors have legitimate differences of opinions as to whether a response contains a correct answer even for the constrained questions used in the track. If assessors have these differences, then eventual end-users of the technology will as well, and evaluations of question answering technology must accommodate these differences to be useful.

- How many calories are there in a Big Mac?
- What two US biochemists won the Nobel Prize in medicine in 1992?
- Who was the first American in space?
- Who is the voice of Miss Piggy?
- Where is the Taj Mahal?
- What costume designer decided that Michael Jackson should only wear one glove?
- In what year did Joe DiMaggio compile his 56-gam hitting streak?
- What language is commonly used in Bombay?
- How many Grand Slam titles did Bjorn Borg win?
- Who was the 16th President of the United States?

Figure 1: Example questions used in the question answering track.

2. The Task

A successful evaluation requires a task that is neither too easy nor too difficult for the current technology. If the task is too simple, all systems do very well and nothing is learned. Similarly, if the task is too difficult, all systems do very poorly and again nothing is learned. Accordingly, we chose a constrained version of the general question answering problem as the focus of the track.

Participants received a large collection of documents (approximately 1.5 gigabytes of text) and 200 fact-based, short-answer questions. Examples of the questions are shown in Figure 1. The documents consisted mostly of newspaper articles and thus contained information on a wide variety of subjects. Each question was guaranteed to

have at least one document in the collection that explicitly answered the question.

Participants returned a ranked list of five [*document-id*, *answer-string*] pairs per question such that each answer string was believed to contain an answer to the question. Answer strings were limited to either 50 or 250 bytes depending on the run type, and could either be extracted from the corresponding document or automatically generated from information contained in the document. Human assessors read each string and made a binary decision as to whether or not the string actually did contain an answer to the question in the context provided by the document. Taking document context into account allowed a system that correctly derived a response from a document that was incorrect to be given full credit for its response.

Given a set of judgments for the strings, the score computed for a submission was the mean reciprocal rank. An individual question received a score equal to the reciprocal of the rank at which the first correct response was returned, or 0 if none of the five responses contained a correct answer. The score for a submission was then the mean of the individual questions' reciprocal ranks. 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. However, the measure also has some drawbacks. 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. Question Answering Results

Twenty different organizations participated in the Question Answering track. The participants are listed in Figure 2. A total of 45 runs were submitted, 20 runs using the 50-byte limit and 25 runs using the 250-byte limit. Table 1 gives both the mean reciprocal rank and the number of questions for which no answer was found for each run. (Four submissions that contained errors are omitted from the table.) The scores are computed over the 198 questions that comprised the official test set. The table is split between the 50-byte and the 250-byte runs and is sorted by decreasing mean reciprocal rank within run type.

The number of questions for which no answer was found shows that the most accurate systems were able to find an answer for more than 2/3 of the questions. Furthermore, when the answer was found at all it was usually ranked first, as shown by the fact that the mean reciprocal rank is also close to 2/3 for these systems.

While the run with the highest mean reciprocal rank score was a 50-byte run, a direct comparison between 50- and 250-byte submissions from the same participant shows that the 50-byte task is more difficult. For every organization that submitted runs of both lengths, the 250-byte limit run had a higher mean reciprocal rank. This is not a surprising result—a system has a greater chance of including a correct response in a longer string—but it was not a guaranteed result. That is, longer strings that include a correct response were not always a correct response themselves. As described below, response strings that contained multiple entities of the same semantic type as the answer and did not specifically indicate which if the entities was the answer were marked as incorrect. Thus, for the question *What is the capital of Kosovo?* the 50-byte response of

```
0 miles northwest of Pristina,
five demonstrators
```

was judged correct, while the 250-byte response of

```
protesters called for military
intervention to end "the Al-
banian uprising." </P> <P> At
Vucitrn, 20 miles northwest of
Pristina, five demonstrators
were reported injured, appar-
ently in clashes with police.
</P> <P> Violent clashes were
also repo
```

was judged incorrect since it is unclear whether the capital is Vucitrn or Pristina.

The submissions from AT&T Research Labs demonstrate that existing passage-retrieval techniques can be successful for 250-byte runs, but are not suitable for 50-byte runs (Singhal et al., 2000). Their question answering system used a traditional vector-based retrieval system to select 50 documents and then scored each sentence within those documents by the number of question words in the surrounding context. For the passage-based runs (`attqa50p` and `attqa250p`), the highest scoring sentences were returned as the response. For their "entity-based" runs (`attqa50e` and `attqa250e`), high scoring sentences were further processed by a linguistic module. The passage-based method was very competitive for the 250-byte limit, but was not nearly as successful when restricted to just 50 bytes. These results suggest that the relatively simple bag-of-words approaches that are successfully used in text retrieval are not sufficient for extracting specific, fact-based answers.

Most participants used a variant of the following general approach to the question answering problem; please see the individual participants' papers in the TREC-8 proceedings for their particular implementations of this strategy (Voorhees and Harman, 2000). In the approach, the system first attempts to classify a question according to the type of its answer as suggested by its question word. For example, a question that begins with "who" (*Who is the prime minister of Japan?*) implies a person or an organization is being sought, and a question beginning with "when" (*When did the Jurassic Period end?*) implies a time designation is needed. Next, the system retrieves a small portion of the document collection using standard text retrieval technology and the question as the query. The system performs a shallow parse of the returned documents to detect entities of the same type as the answer. If an entity of the required type is found sufficiently close to the question's

AT&T Labs Research	MultText Project	U. of Iowa
CL Research	New Mexico State U.	U. of Maryland, College Park
Cymfony, Inc.	NTT DATA Corp.	U. of Massachusetts
GE/U. of Pennsylvania	National Taiwan U.	U. of Ottawa
IBM Research	Royal Melbourne Inst. Technology	U. of Sheffield
LIMSI-CNRS	Seoul National U.	Xerox Research Centre Europe
MITRE	Southern Methodist U.	

Figure 2: Participants in the Question Answering track.

Run Name	Participant	MRR	# not found
textract9908	Cymfony, Inc.	.660	54
SMUNLP1	Southern Methodist U.	.555	63
attqa50e	AT&T Research	.356	109
IBMDR995	IBM	.319	110
xeroxQA8sC	Xerox Research Centre Europe	.317	111
umdqa	U. of Maryland	.298	118
MTR99050	MITRE	.281	118
IBMVS995	IBM	.280	120
nttd8qs1	NTT Data Corp.	.273	121
attqa50p	AT&T Research	.261	121
nttd8qs2	NTT Data	.259	120
CRL50	New Mexico State U.	.220	130
INQ634	U. of Massachusetts	.191	140
CRDBASE050	GE/U. of Pennsylvania	.158	148
INQ638	U. of Massachusetts	.126	158
shefinq50	U. of Sheffield	.081	182
shefatt50	U. of Sheffield	.071	184

a) Runs with a 50-byte limit on the length of the response.

SMUNLP2	Southern Methodist U.	.646	44
attqa250p	AT&T Research	.545	63
GePenn	GE/U. of Pennsylvania	.510	72
attqa250e	AT&T Research	.483	78
uwmt9qa1	MultiText Project	.471	74
mds08q1	Royal Melbourne Inst. Tech	.453	77
xeroxQA8lC	Xerox Research Centre Europe	.453	83
nttd8ql1	NTT Data Corp.	.439	79
MTR99250	MITRE	.434	86
IBMDR992	IBM	.430	89
IBMVS992	IBM	.395	95
INQ635	U. of Massachusetts	.383	95
nttd8ql4	NTT Data Corp.	.371	93
LimsiLC	LIMSI-CNRS	.341	110
INQ639	U. of Massachusetts	.336	104
CRDBASE250	GE/U. of Pennsylvania	.319	111
clr99s	CL Research	.281	115
CRL250	New Mexico State University	.268	122
UIowaQA1	U. of Iowa	.267	117
Scai8QnA	Seoul National U.	.121	154
shefinq250	U. of Sheffield	.111	176
shefatt250	U. of Sheffield	.096	179
NTU99	National Taiwan U.	.087	173
UIowaQA2	U. of Iowa	.060	175

b) Runs with a 250-byte limit on the length of the response.

Table 1: Mean reciprocal rank (MRR) and number of questions for which no correct response was found (# not found) for Question Answering track submissions.

words, the system returns that entity as the response. If no appropriate answer type is found, the system falls back to best-matching-passage techniques.

This approach works well provided the query types recognized by the system have broad enough coverage and the system can classify questions sufficiently accurately. Most systems could answer questions that began with "who" very accurately. However, questions that sought a person but did not actually begin with "who" (*Name the first private citizen to fly in space. What Nobel laureate was expelled from the Philippines before the conference on East Timor?*) were much more difficult. More difficult still were questions whose answers were not an entity of a specific type (*What is Head Start? Why did David Koresh ask the FBI for a word processor?*). Of course, pattern matching on expected answer types is not fool-proof even when "good" matches are found. One response to the question *Who was the first American in space?* was Jerry Brown, taken from a document that says

```
As for Wilson himself, he became
a senator by defeating Jerry
Brown, who has been called the
first American in space.
```

A similar response was returned for the question *Who wrote `Hamlet'?*:

```
`Hamlet,' directed by Franco
Zeffirelli and written
by...well, you know.
```

4. Evaluation Methodology

Our experience with TREC document retrieval tasks has demonstrated that seemingly minor details in the implementation of an evaluation 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 a TREC topic statement altered the nature of the topic statements by guaranteeing that the best words were repeated at least twice. In this section we examine the impact of two features of the track design on questions answering (QA) results: the effect of the process by which test questions were selected, and the use of human assessors to judge answer strings.

4.1. Selecting Test Questions

Our goal in creating the test set of questions was to include a wide variety of subjects and question types while respecting the general restriction to fact-based, short-answer questions. In addition, we made a special effort to select only "straight-forward" questions; that is, we deliberately avoided questions that we felt were unclear or tricky. To this end, we collected a pool of candidate questions from three different sources: TREC QA participants and NIST staff, the TREC assessors (i.e., the people who create relevance judgments for the TREC text retrieval tasks and who judged the answer strings in the QA track), and question logs from the FAQFinder system. Our assumption 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. 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].

NIST staff filtered the pool of candidate questions to obtain the final set of 200 test questions. Many of the FAQFinder questions did not have answers in the document collection so could not be used. Questions that were extremely obvious back-formations of a document statement were removed, as was any question a staff member thought was fuzzy, ambiguous, or unclear. Most questions whose answer was a list were removed, though a few questions that required two responses were retained after making the request for two answers explicit in the question (*What two US biochemists won the Nobel Prize in medicine in 1992?*). The final test set contained 127 questions originally submitted by participants or NIST staff, 49 questions from the assessors, and 24 questions from the FAQFinder logs.

Despite the care taken 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 anticipated, but the assessors determined that two of the 200 questions had no clear answer. The question *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 U.S., and refused to accept 42% as the largest percentage without further proof. The question *When was Queen Victoria born?* was a FAQFinder question. Unfortunately the document thought to have given the years of her life actually gave the years of her reign. These two questions were thus eliminated from the evaluation results, resulting in an actual test set of 198 questions.

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. 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.

Given the processing strategy outlined above that was

[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.

used by most participants, the specific proportion of different types of questions in a test set will affect the evaluation results. A test set with a relatively high proportion of "who" questions, for example, will produce higher scores than a test set with relatively many "what" questions. In the absence of any information regarding the relative frequency of question types in particular environments, continuing to use a large sample of questions picked from many sources is the best alternative. The fact that the vast majority of questions were constructed from a document that contained the answer is likely to have made the task somewhat easier. We anticipate using a greater number of spontaneous questions (extracted from various search engine logs) in future runnings of the track.

4.2. Judging Answer Strings

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, so a single correct list of documents cannot be created. Instead, the list of documents produced by one person (the assessor) is used as an example of a correct response, and systems are evaluated on the sample. While the absolute scores of systems change when different assessors' opinions are used, relative scores generally remain stable, so scores computed using sample judgments are valid for comparing different retrieval techniques.

We wanted to investigate whether different people have different opinions as to what constitutes an acceptable answer, and, if so, how those differences affect QA evaluation. To accomplish this goal, each question was independently judged by three different assessors. The separate judgments were combined into a single judgment set through adjudication for the official track evaluation, but the individual judgments were used to measure the effect of differences in judgments on systems' scores. Judging a test question entailed making a binary decision as to whether a response string was an acceptable answer to the question for every response in the participants' submissions. Questions had an average of 191.6 distinct response strings, which took an assessor approximately a half hour to judge.

4.2.1. Assessor training

Assessors were trained for the QA task before they did any judging. The purpose of the training was to motivate the assessors' task and provide general guidance on the issues that would arise during assessing rather than to drill the assessors on a specific set of assessment rules. To begin, each assessor was given the following instructions.

> 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.

Assessors then judged four sample questions whose response strings were concocted by NIST staff to illustrate various fundamentals of QA judging:

- that answer strings would contain snippets of text that were not necessarily grammatically correct;
- that the answer string did not need to contain justification;
- that the assessor was to judge the *string*, not the document from which the string was drawn;
- that document context must be taken into account; and
- that the string must be responsive to the question.

Document context was vital for questions whose answers change over time. For example, responses to questions phrased in the present tense (*Who is the prime minister of Japan?*) were judged as correct or incorrect based on the time of the document associated with the response. Requiring that the answer string be responsive to the question addressed a variety of issues. As mentioned above, answer strings that contained multiple entities of the same semantic category as the correct answer but did not indicate which of those entities was the actual answer were judged as incorrect. Certain punctuation and units were also required. Thus "5 5 billion" was not an acceptable substitute for "5.5 billion", nor was "500" acceptable when the correct answer was "$500". Finally, unless the question specifically stated otherwise, correct responses for questions about a famous entity had to refer to *the* famous entity and not to imitations, copies, etc. For example, two separate questions asked for the height of the Matterhorn (i.e., the Alp) and the replica of the Matterhorn at Disneyland. Correct responses for one of these questions were incorrect for the other.

4.2.2. Differences among assessors

We had several mechanisms for gathering feedback from the assessors as they judged the test questions. First, the assessors interacted freely with NIST staff during the assessing, asking for clarification of the assessment guidelines and verifying their application of the guidelines to specific cases. In addition, assessors were asked to record the canonical answer(s) for each question as they judged it, and written comments about the question were solicited at the same time. 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 recorded the comments as the assessor judged the strings. Eight think-aloud sessions were held, one each with five different assessors on five different questions plus all three assessors on a sixth question.

This feedback from the assessors confirmed that the assessors understood their task and were able to do it. They 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; after eight years of judging documents, this habit was sometimes hard to break. Another pattern was marking 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, but when the answer string contained random other information. That is, for the question *What is the capital of Kosovo?* the assessors did not have a problem with the answer string `Pristina` but sometimes had problems with answer strings such as

```
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.

Most differences among the assessors were not caused by mistakes, however, but represented 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 the question *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 the question *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 *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 Peace 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 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.

Given that assessor opinions regarding the correctness of an answer differ even for the simple questions with "obvious" answers that were used as test questions, eventual end-users of QA technology will also have differences of opinions regarding what constitutes an acceptable answer. It is pointless to try and force unanimous agreement among assessors in an evaluation since the systems being tested will need to accommodate the varied expectations of different users. However, if we cannot assume there is a single correct answer, then we must ensure that the relative effectiveness of two QA strategies is insensitive to modest changes in the judgment set to have a valid evaluation. The stability of comparative evaluation has been established for text retrieval (Lesk and Salton, 1969; Voorhees, 1998), and we can use the same procedure to examine the stability of QA system comparisons.

4.2.3. Stability of QA evaluation

The procedure used to measure the stability of comparative evaluations quantifies changes in *system rankings* when different judgment sets are used to score runs. For question answering evaluation, a system ranking is a list of the systems under consideration sorted by decreasing mean reciprocal rank. We use a correlation based on Kendall's tau (Stuart, 1983) 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.

Call the judgments for a set of questions a *qrels*. We can define two basic types of qrels: a single-judge qrels in which each question's judgments are the opinions of one assessor, and a multiple-judge qrels in which each question's judgments are some function of individual assessor opinions.

With three judgment sets for each of 198 questions, we can form 3^{198} different single-judge qrels for this QA task. We generated a sample of 100,000 of these single-judge qrels by randomly selecting one of the three assessors who judged the question for each question, and combining the selected judgments into one qrels. We then scored the QA runs using each of the 100,000 qrels, and calculated the sample mean of the mean reciprocal rank for each run. The means are plotted in Figure 3 where the runs are sorted by decreasing mean. The error bars in Figure 3 indicate the minimum and the maximum mean reciprocal rank obtained for that run over the sample of 100,000 qrels.

We also created four multiple-judge qrels. The first of these is the adjudicated qrels used to produce the official evaluation scores. The second is a simple majority opinion qrels in which the majority opinion of the three assessors

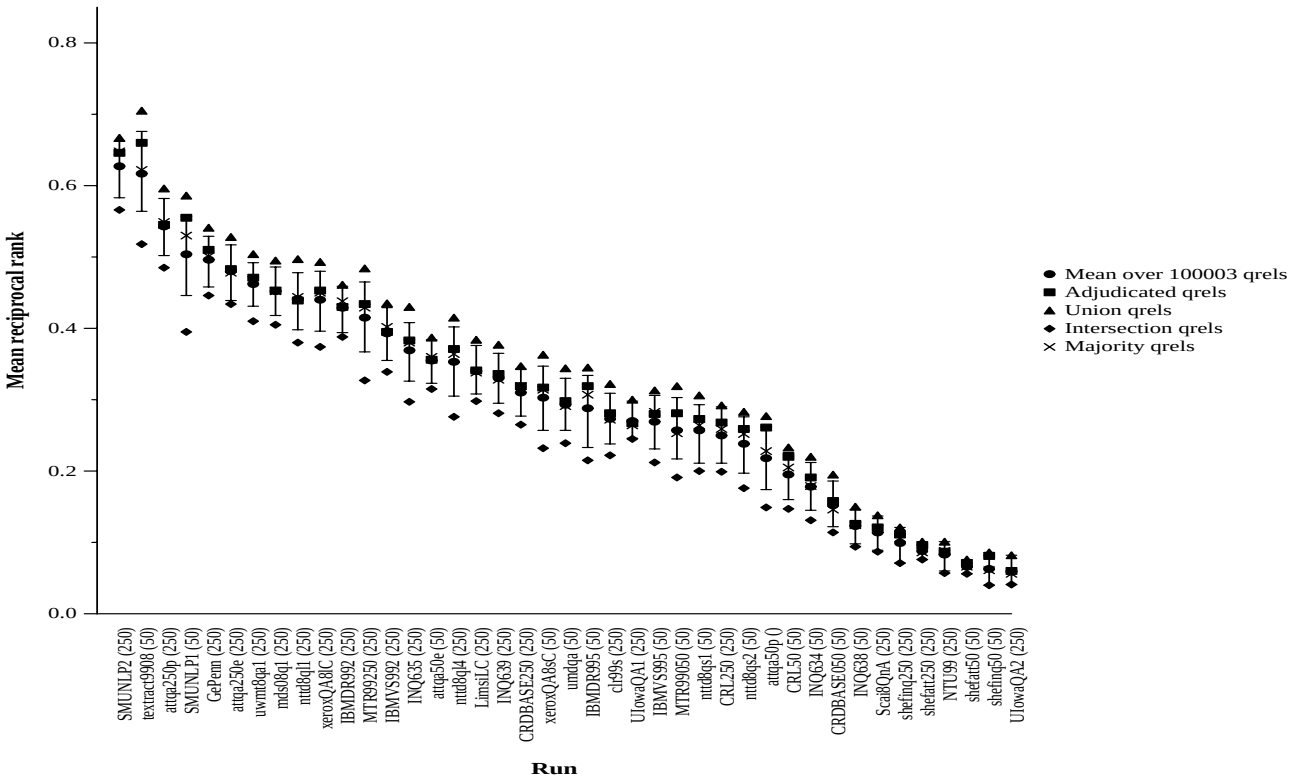

Figure 3: Sample mean, min, and max of the mean reciprocal rank computed for QA runs over a sample of 100,000 single-judge qrels. Also plotted are the mean reciprocal rank for the adjudicated, majority, union, and intersection qrels. Runs are labeled as either 50 byte limit (50) or 250 byte limit (250).

is used as the judgment for each string. The remaining two multiple-judge qrels are the union and intersection qrels. 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 3. 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. 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 single-judge rankings. In the second case, we took the mean of the Kendall's correlation between the ranking produced by the adjudicated qrels and all 100,000 single-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 single-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 single-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. Thus one-judge rankings are essentially equivalent with one another for the purpose of comparative evaluation of QA systems. Furthermore, the second half of Table 2 suggests that single-judge qrels are also equivalent to the expensive adjudicated qrels. As can be seen from Figure 3, the adjudicated score for a run always lies within the boundaries of the minimum and maximum scores obtained on the sample of single-judge qrels. We can conclude, therefore, that using a single human assessor to judge system responses is a viable methodology for comparative evaluation of question answering technology.

There is an important caveat to this conclusion, however, which is 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. From a stability viewpoint, more questions in a test set is always better

	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 single-judge rankings

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

b) correlations with the adjudicated ranking

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.

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 for the number of questions required. Our analysis thus far suggests that 200 (or 198) is sufficient. Since some runs had almost 50 questions that were affected by judgment differences (Voorhees and Tice, 2000), a test set should probably have at least 100 questions.

5. Conclusion

The Question Answering track was the first large-scale evaluation of domain-independent question answering systems. The questions used in the track were deliberately constrained to fact-based, short-answer questions to make the task amenable to evaluation. Systems generally classified a question according to the type of its answer, and then performed a shallow parse of likely documents to find objects of the entailed type. The most accurate systems were able to answer more than 2/3 of the questions correctly. Existing passage-retrieval techniques were adequate for finding answers when relatively long responses were permissible, but more sophisticated processing was need to focus on the answer itself.

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 paper validated the methodology used by showing it was both appropriate and effective. Assessors do have differences of opinion as to whether a particular response answers a question even for these fact-based questions. Having the evaluation accommodate differences of opinion in the answer keys reflects a requirement of the real problem, since if assessors have different opinions then eventual end-users of the technology will have different opinions as well. Comparisons between systems are valid to the extent that they are stable under changes in the judgments that produce the scores.

There will be another Question Answering track in TREC-9, which will be mostly the same as the TREC-8 track. One change in the track will be to have a test set of 500 questions rather than 200 questions, and to have many fewer of the questions be constructed from a target document. A second change is to add a third "exact answer" condition to the 50- and 250-byte-limit conditions. In this condition, answer strings will be judged incorrect if they contain any spurious material.

Acknowledgements

Our thanks to the QA participants who made the track possible, with special thanks to track coordinators Amit Singhal and Tomek Strzalkowski. Donna Harman and Paul Over at NIST assisted with test question selection and other aspects of defining the track evaluation methodology.

6. References

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

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

Singhal, Amit, Steve Abney, Michiel Bacciani, Michael Collins, Donald Hindle, and Fernando Pereira, 2000. AT&T at TREC-8. In (Voorhees and Harman, 2000).

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

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. and Dawn M. Tice, 2000. 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.