Overview of the TREC 2002 Question Answering Track

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

Abstract

The TREC question answering track is an effort to bring the benefits of large-scale evaluation to bear on the question answering problem. The track contained two tasks in TREC 2002, the main task and the list task. Both tasks required that the answer strings returned by the systems consist of nothing more or less than an answer in contrast to the text snippets containing an answer allowed in previous years. A new evaluation measure in the main task, the confidence-weighted score, tested a system's ability to recognize when it has found a correct answer.

The goal of the question answering (QA) track is to foster research on systems that retrieve answers rather than documents in response to a question, with particular emphasis on systems that can function in unrestricted domains. Now in its fourth year, the tasks in the track have evolved over the years to increase the realism of the task and to focus research on particular aspects of the problem deemed important to improving the state-of-the-art. All of the tasks have involved finding answers to closed-class questions within a large corpus of news text.

This paper provides an overview of the TREC 2002 QA track. This year's track contained two tasks, the main task and the list task. Both tasks were also run in TREC 2001, but systems were required to return exact answers this year. That is, the text string returned by the system in response to a question was required to consist of a complete answer and nothing else, in contrast to earlier years where systems could return text strings that simply contained an answer. To make the paper self-contained, the first section recaps the tasks and evaluation procedures used in the first three tracks. The following sections then describe this year's tasks.

1 Evolution of the TREC QA Track

The task in the first two QA tracks (TRECs 8 and 9) was the same. For each question in the question set, systems retrieved a ranked list of up to five text snippets that contained an answer to the question plus a document that supported the answer. The collection of documents from which the support was drawn was a large set of newswire and newspaper articles. The questions were restricted to factoid questions such as *In what year did Joe DiMaggio compile his 56-game hitting streak?* and *Name a film in which Jude Law acted.* Each question was guaranteed to have at least one document in the collection that explicitly answered it. The maximum length of the text snippets was either 50 or 250 bytes, depending on the run type.

Human assessors read each string and decided whether the string actually did contain an answer to the question in the context provided by the document. Given a set of judgments for the strings, the score computed for a submission was the mean reciprocal rank. An individual question received a score equal to the reciprocal of the rank at which the first correct response was returned, or zero 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 TREC-8 track both defined how answer strings were judged, and established that different assessors have different ideas as to what constitutes a correct answer even for the limited type of questions used in the track. A [*document-id*, *answer-string*] pair was judged correct if, in the opinion of the assessor, the answer-string contained an answer to the question, the answer-string was responsive to the question, and the document supported the answer. If the answer-string was responsive and contained a correct answer, but the document did not support that answer, the pair was judged "Not supported". Otherwise, the pair was judged incorrect. Requiring that the answer string be responsive to the question addressed a variety of issues. Answer strings that contained multiple entities of the same semantic category as the correct answer but did not indicate which of those entities was the actual answer (e.g., a list of names in response to a who question) were judged as incorrect. Certain punctuation and units were also required. Thus "5 5 billion'' was not an acceptable substitute for "5.5 billion", nor was "500" acceptable when the correct answer was "$500". Finally, unless the question specifically stated otherwise, correct responses for questions about a famous entity had to refer to the famous entity and not to imitations, copies, etc. For example, two TREC-8 questions asked for the height of the Matterhorn (i.e., the Alp) and the replica of the Matterhorn at Disneyland. Correct responses for one of these questions were incorrect for the other. See [6] for a very detailed discussion of responsiveness.

To test whether assessor opinions vary, each TREC-8 question was independently judged by three different assessors. The separate judgments were combined into a single judgment set through adjudication for the official track evaluation, but the individual judgments were used to measure the effect of differences in judgments on systems' scores. Assessors opinions did vary. For example, assessors differed on how much of a name was required and on the desired granularity of dates and locations. Fortunately, as with document retrieval evaluation, the relative mean reciprocal rank scores between QA systems remain stable despite differences in the judgments used to evaluate them [5].

The TREC 2001 track modified the main task to make it more realistic and introduced two new tasks, the list task and the context task. All runs were restricted to answer strings of maximum length 50 bytes since the results from the earlier tracks clearly demonstrated that allowing 250-byte answer strings was a much simpler problem. In the main task, the guarantee that a question had an answer in the document collection was eliminated. A system returned the string "NIL" to indicate its belief that there was no answer in the document collection. NIL was marked correct if there was no known answer for that question in the collection and incorrect otherwise.

The list task required systems to assemble an answer from information located in multiple documents. Such questions are harder to answer than the questions used in the main task since information duplicated in the documents must be detected and reported only once. Each question in the list task specified a particular kind of information to be retrieved, such as *Who are 6 actors who have played Tevye in "Fiddler on the Roof"?*. Systems returned an unordered list of [*document-id*, *answer-string*] pairs where each pair represented a single instance. Results were scored using mean accuracy, which is the ratio of the number of distinct correct responses retrieved to the target number of responses requested.

The context task was a pilot evaluation for question answering within a particular scenario or context. The task was designed to represent the kind of dialog processing that a system would need to support an interactive user session. Questions were grouped into different series, and the QA system was expected to track the discourse objects across the individual questions of a series. Unfortunately, the results in the pilot were completely dominated by whether or not a system could answer the particular type of question: the ability to correctly answer questions later in a series was uncorrelated with the ability to correctly answer questions earlier in the series. Thus the task was not repeated in TREC 2002.

2 The TREC 2002 QA Track

The TREC 2002 track repeated the main and list tasks from 2001, but with the major difference of requiring systems to return exact answers. The change to exact answers was motivated by the belief that a system's ability to recognize the precise extent of the answer is crucial to improving question answering technology. The problems with using text snippets containing the answer as responses were illustrated in the TREC 2001 track. For example, each of the answer strings shown in Figure 1 was judged correct for the question *What river in the US is known as the Big Muddy?*, yet earlier responses are clearly better than later ones. Judging only exact answers correct forces systems to demonstrate that they know precisely where the answer lies in such strings.

What constitutes an "exact answer"? As with correctness, exactness is essentially a personal opinion. NIST provided guidelines to the assessors so that questions would be judged similarly, but in the end whether or not an answer was exact was up to the assessor. The guidelines given to the assessors are reproduced in Figure 2. Notice that even "good" responses that contain a correct answer and justification for that answer

```
the Mississippi
Known as Big Muddy, the Mississippi is the longest
as Big Muddy , the Mississippi is the longest
messed with .  Known as Big Muddy , the Mississip
Mississippi is the longest river in the US
the Mississippi is the longest river in the US,
the Mississippi is the longest river(Mississippi)
has brought the Mississippi to its lowest
ipes.In Life on the Mississippi,Mark Twain wrote t
Southeast;Mississippi;Mark Twain;officials began
Known; Mississippi; US,; Minnesota; Gulf Mexico
Mud Island,;Mississippi;"The;-- history,;Memphis
```

Figure 1: Correct answer strings for *What river in the US is known as the Big Muddy?*

were considered inexact for the purposes of this evaluation.

A system response consisting of an [*document-id*, *answer-string*] pair was assigned exactly one judgment by a human assessor as follows:

wrong: the answer string does not contain a correct answer or the answer is not responsive;

not supported: the answer string contains a correct answer but the document returned does not support that answer;

not exact: the answer string contains a correct answer and the document supports that answer, but the string contains more than just the answer (or is missing bits of the answer);

right: the answer string consists of exactly a correct answer and that answer is supported by the document returned.

Both QA tasks used the same new document collection as the source of answers. The collection is known as the AQUAINT Corpus of English News Text, which may be obtained from the Linguistic Data Consortium (`www.ldc.upenn.edu`) as catalog number LDC2002T31. The collection is comprised of documents from three different sources: the AP newswire from 1998–2000, the New York Times newswire from 1998–2000, and the (English portion of the) Xinhua News Agency from 1996–2000. There are approximately 1,033,000 documents and 3 gigabytes of text in the collection.

As in previous tracks, NIST provided the ranking of the top 1000 documents retrieved by the PRISE search engine when using the question as a query, and the full text of the top 50 documents per question (as given from that same ranking). This data was provided strictly as a convenience for groups that did not wish to implement their own document retrieval system. There was no guarantee that the ranking would contain the documents that actually answer a question.

All runs submitted to the track were required to be completely automatic; no manual intervention of any kind was permitted. To avoid any confounding effects caused by processing questions in different orders, all questions were required to be processed from the same initial state. That is, the system was not permitted to adapt to test questions that had already been processed.

Thirty-four groups submitted a total of 76 runs to the QA track, 67 main task runs (though two of the runs were mistakenly duplicates of one another) and 9 list task runs. All submissions to the track were judged.

3 The Main Task

In addition to the requirement for exact answers, the TREC 2002 main task had another significant change from earlier QA tasks. Systems were limited to one response per question, not five, and thus the scoring

Consider the question

What is the longest river in the United States?

The following are correct, exact answers

- `Mississippi,`
- `the Mississippi,`
- `the Mississippi River,`
- `Mississippi River`
- `mississippi`

while none of the following are correct, exact answers

- `At 2,348 miles the Mississippi River is the longest river in the US.`
- `2,348 miles; Mississippi`
- `Missipp`
- `Missouri`

You are not required to accept only the most minimal response possible as an exact answer; some redundancy is fine. In the Mississippi River example, we want you to accept "`Mississippi River`" as exact even though "river" is redundant since the correct response must be a river. Similarly, we want you to accept answers of "`<number> X`" for questions that ask "*How many X?*". Ungrammatical responses are probably not exact. A location question can have "`in Pennsylvania`" as an exact answer, but not "`Pennsylvania in`". If the answer string contains several entities of the same type, one of which is correct and the others are not, then the answer string is not exact.

Figure 2: Instructions given to the assessors regarding how to judge exact answers.

metric also changed. The scoring metric used, called the confidence-weighted score, was specifically chosen to test a system's ability to recognize when it has found a correct answer.

This section describes the main task within the TREC 2002 QA track. The first subsection gives the details regarding how the task was implemented. The following subsection provides the results of the evaluation, and the final subsection assesses the quality of the evaluation.

3.1 Task details

As already mentioned, one goal of the main task in this year's QA track was to test a system's ability to recognize when it has found a correct answer. A main task run consisted of exactly one response for each of 500 test questions. A response was either a [*document-id*, *answer-string*] pair or the string "NIL", which was used to indicate the system's belief that there was no correct answer in the collection. Within the submission file, the questions were ordered from most confident response to least confident response. That is, the question for which the system was most confident that it had returned a correct response was ranked first, then the question that the system was next most confident about, etc. so that the last question was the question for which the system was least confident in its response.

Given a question ranking based on confidence of a correct response, an analog of document retrieval's uninterpolated average precision can be computed. This measure rewards a system for a correct answer early in the ranking more than it rewards for a correct answer later in the ranking. More formally, if there are Q questions in the test set, the confidence-weighted score is defined to be

$$\frac{1}{Q}\sum_{i=1}^{Q}\frac{\text{number correct in first } i \text{ ranks}}{i}.$$

The test set of question used in the task were drawn from MSNSearch and AskJeeves logs. These logs are part of the set of logs donated by Microsoft and AskJeeves for use in TREC 2001. As in TREC 2001, NIST assessors searched the document collection for answers to candidate questions. NIST staff selected the final test set from among the candidates that had answers, keeping some questions for which the assessors found no answer. After judging, 46 questions had no known answer in the collection. NIST corrected the spelling, punctuation, and grammar of the questions in the logs, but left the content as it was. Unfortunately, some errors in the test questions remained. For example, question 1445 asked *When is Snoop Dog's birthday?*, when the correct spelling of the name is 'Snoop Dogg'. After discussion on the track mailing list, the track participants decided to evaluate over all 500 questions despite the remaining errors. This decision was based on the difficulty of knowing when to stop calling something an error (is a misplaced apostrophe an error? missing capitalization? alternate spellings?) and on the recognition that deployed systems will have to cope with user errors.

About one quarter of the 2001 test set of questions were definition questions such as *Who is Duke Ellington?* and *What are polymers?*. Having many definition questions is a problem for an evaluation such as TREC where there is no specific target user and thus no way of knowing what kind of response should be produced. Accordingly, NIST did not choose any definition questions for this year's test set, a restriction that most likely made the test set intrinsically easier than last year's set since definition questions are among the more difficult questions to answer. NIST made no other attempt to control the relative number of different types of questions in the test set.

One of the concerns expressed by the track participants regarding the move to exact answers was that submissions containing only exact answers would be too sparse to make good training data for future development. To address this concern and to collect data for a future "answer justification" task, participants were requested to also submit a justification for each of their responses. A justification was defined to be an arbitrary collection of ASCII characters that did not contain newline characters and was no longer than 1024 characters. Justifications were optional in that the justification string could be empty. The vast majority of justifications that were submitted consisted of the piece of text (snippet, sentence, or paragraph) from which the response had been extracted.

3.2 Evaluation results

Table 1 gives evaluation results for a subset of the main task runs. The table includes one run each from the fifteen groups who submitted the top-scoring runs. The run shown in the table is the run with the best confidence-weighted score, and the table is sorted by confidence-weighted score. Also given in the table are the number and percentage of questions answered correctly; the number of questions whose response was judged as inexact; and the precision and recall for recognizing when there is no correct answer in the document collection ("NIL Accuracy"). Precision of recognizing no answer is the ratio of the number of times NIL was returned and correct to the number of times it was returned; recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct (46).

QA systems have become increasingly complex over the four years of the TREC track such that there is now little in common across all systems. Most systems classify an incoming question according to a system-specific ontology of question types as a first step. The ontologies of question types range from small sets of broad categories to highly-detailed hierarchical schemes. Once a question is classified, the system performs type-specific processing. Many TREC 2002 systems used specifc data sources such as name lists, gazetteers, movie databases, and the like that were searched when the system determined the question to be of the appropriate type. The web was used as a data source by most systems, though it was used in different ways. Some systems used the web as the primary source of an answer that the system then mapped to a document in the corpus to return as a response. Other systems did the reverse: used the corpus as the primary source of answers and then verified candidate answers on the web. Still other systems used the web as one of several sources whose combined evidence selected the final response.

TREC 2001 saw an increase in the number of systems using a shallow, data-driven approach to question answering in contrast to systems that attempt a full understanding of the question. Both approaches were well-represented in TREC 2002, as illustrated by the two top-scoring runs *LCCmain2002* from Language Computer Corporation and *exactanswer* from InsightSoft-M. The LCC system, PowerAnswer, transforms questions and possible answer text into a logic representation and then builds a formal proof for valid answers [1]. In contrast, the InsightSoft-M system relies on an extensive set of patterns where an individual

Table 1: Evaluation scores for a subset of the TREC 2002 main task runs. Scores are given for the best run as measured by confidence-weighted score from the top 15 groups. Scores include the confidence-weighted score; the number (#) and percentage (%) of questions that were answered correctly; the number of responses judged inexact; and the precision (Prec) and recall (Recall) for recognizing when there is no correct answer in the collection (NIL Accuracy).

Run Tag	Confidence weighted Score	Correct Answers #	%	Number Inexact	NIL Accuracy Prec	Recall
LCCmain2002	0.856	415	83.0	8	0.578	0.804
exactanswer	0.691	271	54.2	12	0.222	0.848
pris2002	0.610	290	58.0	17	0.241	0.891
IRST02D1	0.589	192	38.4	17	0.167	0.217
IBMPQSQACYC	0.588	179	35.8	9	0.196	0.630
uwmtB3	0.512	184	36.8	20	0.000	0.000
BBN2002C	0.499	142	28.4	18	0.182	0.087
isi02	0.498	149	29.8	15	0.385	0.109
limsiQalir2	0.497	133	26.6	11	0.188	0.196
ali2002b	0.496	181	36.2	15	0.156	0.848
ibmsqa02c	0.455	145	29.0	44	0.224	0.239
FDUT11QA1	0.434	124	24.8	6	0.139	0.957
aranea02a	0.433	152	30.4	36	0.235	0.174
nuslamp2002	0.396	105	21.0	17	0.000	0.000
pqas22	0.358	133	26.6	11	0.145	0.674

pattern is a complex expression built from simpler component structures [2].

The results in Table 1 illustrate that the quality of a system's confidence ranking can have a significant impact on its score. For example, the `FDUT11QA1` and `aranea02a` runs have nearly identical confidence-weighted scores of .434 and .433, but the `aranea02a` run correctly answered 28 more questions than the `FDUT11QA1` run. Figure 3 shows a scatter plot of confidence-weighted score versus number correctly answered for all main task runs. The solid line shows the best possible score a run could achieve for a given number of correctly answered questions; this score corresponds to ranking all correctly answered questions before all incorrectly answered questions. Similarly, the dotted line shows the worst possible score a run could achieve; this score corresponds to ranking all incorrectly answered questions before all correctly answered questions. In general, points are closer to the optimal line than the pessimal line, demonstrating that the systems were at least as good at ranking their responses as random guessing. A dot above and to the left of a second dot represents a system that is better at ranking than the second system since it has a higher confidence-weighted score but correctly answered fewer questions.

The systems used a variety of approaches to creating their question rankings. Almost all systems used question type as a factor since some question types are easier to answer than others. Some systems use a score to rank candidate answers for a question; when that score is comparable across questions, it can also be used to rank questions. A few groups used a training set of previous years' questions and answers to learn a good feature set and corresponding weights to predict confidence. Many systems used NIL as an indicator that the system couldn't find an answer (rather than the system was sure there was no answer), so ranked NIL responses last. With the exception of the *LCCmain2002* run, though, the NIL accuracy scores are low, indicating that systems had trouble recognizing when there was no answer in the document collection.

3.3 Analysis of the evaluation

The TREC-8 track demonstrated that QA evaluation results based on text snippets and mean reciprocal rank scoring is stable despite differences in assessor opinions as to whether an answer is correct [5]. This year's main task included several possible sources of additional instability: a single response per question,

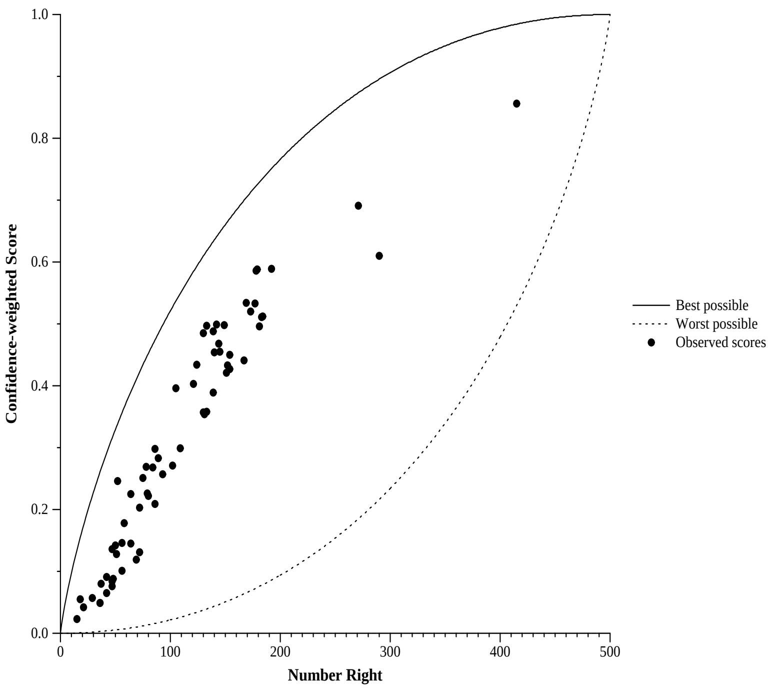

Figure 3: Confidence-weighted score vs. number correctly answered questions for main task runs.

Table 2: Kendall tau correlations for system rankings based on different judgment sets.

	Confidence score			Number correct		
	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
Adjudicated	0.954	0.941	0.944	0.958	0.949	0.960
Set 1		0.920	0.917		0.933	0.944
Set 2			0.906			0.926

confidence-weighting scoring, and exact answers. The methodology used in TREC-8 to test for stability was repeated for this year's main task to assess the effect of these changes. Each question was independently judged by three different assessors. The assessors for a particular question were arbitrarily assigned as assessor 1, assessor 2, or assessor 3. All the assessor 1 judgments for all questions were gathered into judgment set 1, all the assessor 2 judgments into judgment set 2, and all the assessor 3 judgments into judgment set 3. These three judgment sets were combined through adjudication into a final judgment set, which is the judgment set used to produce the official TREC 2002 main task scores.

Each run was scored using each of the four judgment sets. For each judgment set, the runs were ranked in order from most effective to least effective using either the confidence-weighted score or the raw number of correctly answered questions. The distance between two rankings was computed using a correlation measure based on Kendall's tau [3]. 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. Table 2 gives the correlations between all pairs of rankings for both evaluation metrics.

The correlations between rankings are all above 0.9, an acceptable level for the assessor effect. Correlations with the adjudicated judgment set are 0.94 or higher, a better level. The higher correlations with the adjudicated set are probably due to the lower incidence of judgment errors (i.e., just plain mistakes rather than differences of opinion) in an adjudicated set. Correlations are slightly higher for the raw count of the

Table 3: Distribution of disagreements in judgments. Each response pair was independently assigned a judgment of wrong (W), right (R), unsupported (U), or inexact (X) by three assessors. Each entry in the table gives the number (#) and percentage (%) of pairs assigned the given triple of judgments. For example, for 418 pairs or 22.2 % of the total disagreements, two assessors marked the pair right and the third marked it inexact.

	Counts			Counts	
Judgments	#	%	Judgments	#	%
WWR	174	9.2	WXX	86	4.6
WWU	151	8.0	RRU	141	7.5
WWX	141	7.5	RRX	418	22.2
WRR	167	8.9	RUU	87	4.6
WRU	32	1.7	RUX	36	1.9
WRX	93	4.9	RXX	201	10.7
WUU	81	4.3	UUX	23	1.2
WUX	34	1.8	UXX	21	1.1

number of questions correctly answered than for the confidence-weighted score. This likely reflects the fact that the confidence-weighted score is much more sensitive to differences in judgments for questions at small (close to one) ranks.

There was a total of 15,948 [*document-id*, *answer-string*] pairs judged across the 500 questions, an average pool size of 31.9 strings. This is a smaller pool size than in previous tracks because only one response per question per run was allowed, and because the move to exact answers significantly increased the amount of overlap among runs. A total of 1886 pairs (11.8 % of all pairs) had some disagreement among the three assessors as to which judgment should be assigned to the pair. This small percentage of disagreement is misleading, however, since only 3725 pairs had at least one judge assign a judgment that was something other than wrong. In other words, there was some disagreement in assessing for half of all pairs that were not obviously wrong.

Table 3 shows the distribution of the assessors' disagreements. Each response pair is associated with a triple of judgments according to the three judgments assigned by the different assessors. In the table the judgments are denoted by W for wrong, R for right, U for unsupported, and X for inexact. The table shows the number of pairs that are associated with each triple plus the percentage of the total number of disagreements that that triple represents. The largest number of disagreements involves right and inexact judgments: the RRX and RXX combinations account for a third of the total disagreements. Inspection of these disagreements reveals that many of the granularity differences observed in TREC-8 are now reflected in this distinction. For example, question 1395 asks *Who is Tom Cruise married to?*, and a correct response is Nicole Kidman. Some assessors accepted just Kidman, but others marked that as inexact. Some assessors also accepted actress Nicole Kidman, while others marked that as inexact. Similar issues arose with dates and place names. For dates and quantities, there was disagreement whether slightly off responses are wrong or inexact. For example, when the correct response is April 20, 1999, is April 19, 1999 wrong or inexact? This last distinction doesn't matter very much in practice since in either case the response is not right.

Human judgments are not the only source of variability when evaluating QA systems. As is true with document retrieval systems, QA system effectiveness depends on the questions that are asked, so the particular set of questions included in a test set will affect evaluation results. Since the test set of questions is assumed to be a random sample of the universe of possible questions, there is always some chance that a comparison of two systems using any given test set will lead to the wrong conclusion. The probability of an error can be made arbitrarily small by using arbitrarily many questions, but there are practical limits to the number of questions that can be included in an evaluation.

Following our work for document retrieval evaluation [4], we can use the runs submitted to the QA track to empirically determine the relationship between the number of questions in a test set, the observed difference in scores (δ), and the likelihood that a single comparison of two QA runs leads to the correct conclusion. Once established, the relationship can be used to derive the minimum difference in scores required for a certain level of confidence in the results given there are 500 questions in the test set.

The core of the procedure is comparing the effectiveness of a pair runs on two disjoint question sets of equal size to see if the two sets disagree as to which of the runs is better. We define the error rate as the percentage of comparisons that result in a disagreement (a "swap"). Since the QA track used 500 questions, we can directly compute the error rate for question set sizes up to 250 questions. By fitting curves to the values observed for question set sizes up to 250, we can extrapolate the error rates to question sets up to 500 questions.

When calculating the error rate, the difference between two runs' confidence-weighted scores is categorized into one of 21 bins based on the size of the difference. The first bin contains runs with a difference of less than 0.01 (including no difference at all). The next bin contains runs whose difference is at least 0.01 but less than 0.02. The limits for the remaining bins increase by increments of 0.01, with the last bin containing all runs with a difference of at least 0.2.

Each question set size from 1 to 250 is treated as a separate experiment. Within an experiment, we randomly select two disjoint sets of questions of the required size. We compute the confidence-weighted score over both question sets for all runs, then count the number of times we see a swap for all pairs of runs using the bins to segregate the counts by size of the difference in scores. The entire procedure is repeated 10 times (i.e., we perform 10 trials), with the counts of the number of swaps kept as running totals over all trials.

The error rates computed from this procedure are then used to fit curves of the form $ErrorRate = A_1 e^{-A_2 S}$ where A_1 and A_2 are parameters to be estimated and S is the size of the question set. A different curve is fit for each different bin. The input to the curve-fitting procedure used only question set sizes greater than 20 since smaller question set sizes are both uninteresting and very noisy. Curves could not be fit for the first bin (differences less than .01), for the same reason, or for bins where differences were greater than 0.16. Curves could not be fit for large differences because too much of the curve is in the long flat tail.

The resulting extrapolated error rate curves are plotted in Figure 4. In the figure, the question set size is plotted on the x-axis and the error rate is plotted on the y-axis. The error rate for 500 questions when a difference of 0.05 in confidence-weighted scores is observed is approximately 8 %. That is, if we know nothing about two systems except their scores which differ by 0.05, and if we repeat the experiment on 100 different sets of 500 questions, then on average we can expect 8 out of those 100 sets to favor one system while the remaining 92 to favor the other.

The horizontal line in the graph in Figure 4 is drawn at an error rate of 5 %, a level of confidence commonly used in experimental designs. For question set sizes of 500 questions, there needs to be an absolute difference of at least 0.07 in confidence-weighted scores before the error rate is less than 5 %. When using the 5 % error rate standard, many of the runs in Table 1 group into classes that should be considered equally effective. For example, the scores for the *pris2002*, *IRST02D1*, and *IBMPQSQACYC* runs are all within 0.07 of one another. Further, all changes in the system rankings when systems were evaluated using different judgment sets were between pairs of systems whose difference in confidence-weighted scores when evaluated using the adjudicated judgment set was less than 0.07.

4 The List Task

The list task requires systems to assemble an answer from information located in multiple documents. Each question in the list task specified the number of instances of a particular kind of information to be retrieved, such as in the example questions shown in Figure 5. Each instance was guaranteed to obey the same constraints as an individual answer in the main task except that there was known to be at least as many correct instances as requested in the document collection. Systems returned an unordered list of [*document-id*, *answer-string*] pairs where each pair represented a single instance. The list could contain no more than the target number of instances.

The 25 questions used as the list task test set were constructed by NIST assessors. The assessors were instructed to construct questions whose answers would be a list of entities (people, places, dates, numbers) such that the list would not likely be found in a reference work such as a gazetteer or almanac. Each assessor was asked to create one small question (five or fewer expected answers), one large question (between twenty and forty expected answers), and two medium questions (between five and twenty expected answers). They

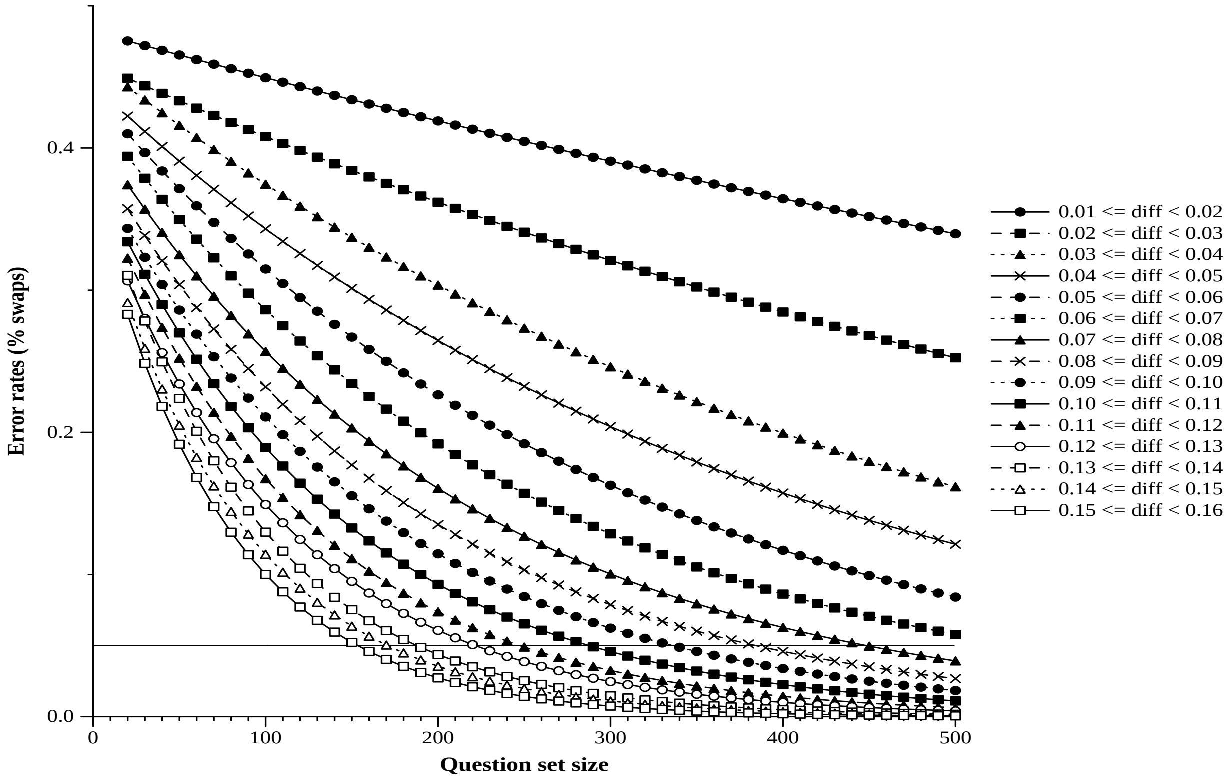

Figure 4: Error rates extrapolated to test sets of 500 questions.

- Name 22 cities that have a subway system.
- What are 5 books written by Mary Higgens Clark?
- List 13 countries that export lobster.
- What are 12 types of clams?
- Name 21 Godzilla movies.

Figure 5: Example list task questions.

Table 4: Average accuracy for list task runs. Accuracy is computed as the number of distinct instances divided by the target number of instances.

Run Tag	Average Accuracy	Run Tag	Average Accuracy
LCClist2002	0.65	shefft11lo	0.06
SUT11IR1LT2	0.15	sheft11log	0.06
SUT11IR1LT	0.11	clr02l1	0.06
UdeMlistNoW	0.07	clr02l2	0.05

searched the document collection using the PRISE search engine to find as complete a list of instances as possible. The target number of instances to retrieve was then selected such that the document collection contained more than the requested number of instances, but more than one document was required to meet the target. A single document could contain multiple instances, and the same instance might be repeated in multiple documents.

Judgments of incorrect, not supported, inexact, or right were made individually for each [*document-id*, *answer-string*] pair as in the main task. The assessor was given one list at a time, and while judging for correctness he also marked a set of responses as distinct. The assessor arbitrarily chose any one of a set of equivalent responses to mark as the distinct one, and marked the remainder as not distinct. Only correct responses could be marked distinct. Each question was judged by only one assessor, though the judgments were reviewed for errors by NIST staff.

List results were evaluated using accuracy, the number of distinct, correct responses divided by the target number of instances. Table 4 gives the average accuracy scores for the eight list task submissions.

In general, the scores for the list task are low. With the change to exact answers, retrieving distinct answers was not an issue: there was only one case across all questions and runs where two correct instances were deemed equivalent. The requirement for exact answers does not appear to be a major problem either. Of the 256 total instances requested across the 25 questions, the eight runs averaged only 7.1 "inexact" judgments. The average for unsupported judgments was similar at 7.9. Instead, it appears that the target answers were just difficult to find.

5 Summary

The TREC 2002 QA track made significant changes to the task definition as compared to earlier TREC tracks. In particular, the 2002 main task required systems to return exact answers, to return only one response per question, and to rank questions by confidence in the response. Major themes for this year's systems were a marked increase in the use of name lists, gazetteers, and the like to answer specific question types, and continued reliance on the web as a system component.

Evaluation of the track results confirmed that system comparisons are sufficiently stable for an effective evaluation. Human assessors do not always agree as to whether an answer is exact, but the differences reflect the well-known differences in opinion as to correctness rather than inherent difficulty in recognizing whether an answer is exact. Empirically-derived error rates based on the sensitivity of the confidence-weighted score to different question sets suggest that scores differing by less than 0.07 are equivalently effective. No pair of systems with a difference in scores of at least 0.07 swapped when evaluated by different judgment sets.

Participation in the list task was quite limited. A majority of main task participants indicated that they did not perform the list task because of time constraints rather than a lack of interest in the task. Accordingly, the current plans for the TREC 2003 QA track are to have one task in which systems are required to answer a variety of question types, including factoid questions, list questions, and definition questions. The test question set will not explicitly distinguish the type of the question. During the evaluation phase, the question set will be partitioned into the three types and each type of question will be scored using the methodology appropriate for that question type.

Acknowledgements

My thanks to John Prager who suggested the plot in Figure 3, and Chris Buckley for fitting the error rate curves shown in Figure 4.

References

[1] Dan Moldovan, Sanda Harabagiu, Roxana Girju, Paul Morarescu, Finley Lacatusu, Adrian Novischi, Adriana Badulescu, and Orest Bolohan. LCC tools for question answering. In Voorhees and Buckland [7].

[2] Martin M. Soubbotin and Sergei M. Soubbotin. Use of patterns for detection of answer strings: A systematic approach. In Voorhees and Buckland [7].

[3] 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.

[4] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, 2002.

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

[6] Ellen M. Voorhees and Dawn M. Tice. The TREC-8 question answering track evaluation. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 83–105, 2000. NIST Special Publication 500-246. Electronic version available at `http://trec.nist.gov/pubs.html`.

[7] E.M. Voorhees and L.P. Buckland, editors. *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2003.