Overview of the TREC 2004 Question Answering Track

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

Abstract

The TREC 2004 Question Answering track contained a single task in which question series were used to define a set of targets. Each series contained factoid and list questions and related to a single target. The final question in the series was an "Other" question that asked for additional information about the target that was not covered by previous questions in the series. Each question type was evaluated separately with the final score a weighted average of the different component scores. Applying the combined measure on a per-series basis produces a QA task evaluation that more closely mimics classic document retrieval evaluation.

The goal of the TREC question answering (QA) track is to foster research on systems that return answers themselves, rather than documents containing answers, in response to a question. The track started in TREC-8 (1999), with the first several editions of the track focused on *factoid* questions. A factoid question is a fact-based, short answer question such as *How many calories are there in a Big Mac?*. The task in the TREC 2003 QA track was a combined task that contained list and definition questions in addition to factoid questions [3]. A list question asks for different instances of a particular kind of information to be returned, such as *List the names of chewing gums*. Answering such questions requires a system to assemble an answer from information located in multiple documents. A definition question asks for interesting information about a particular person or thing such as *Who is Vlad the Impaler?* or *What is a golden parachute?*. Definition questions also require systems to locate information in multiple documents, but in this case the information of interest is much less crisply delineated.

The TREC 2003 track was the first large-scale evaluation of list and definition questions, and the results of the track demonstrated that not only are list and definition questions challenging tasks for systems, but they present evaluation challenges as well. Definition task scores contained a relatively large error term in comparison to the size of the difference between scores of different systems. For example, the analysis of the TREC 2003 definition evaluation performed as part of TREC 2003 showed that an absolute difference in scores of 0.1 was needed to have 95% confidence that the comparison represented a true difference in scores when the test set contained 50 questions. Yet relatively few of the runs submitted to TREC 2003 differed by this amount. Reducing the error term requires more definition questions in the test set. The task for the TREC 2004 QA track was designed to accommodate more definition questions while keeping a mix of different question types.

The TREC 2004 test set contained factoid and list questions grouped into different series, where each series had the target of a definition associated with it. Each question in a series asked for some information about the target. In addition, the final question in each series was an explicit "other" question, which was to be interpreted as "Tell me other interesting things about this target I don't know enough to ask directly". This last question is roughly equivalent to the definition questions in the TREC 2003 task.

The reorganization of the combined task into question series has an important additional benefit. Each series is a (limited) abstraction of an information dialog in which the user is trying to define the target. The target and earlier questions in a series provide the context for the current question. Context processing is an important element for question answering systems to possess, but its use has not yet been successfully incorporated into the TREC QA track [2].

The remainder of this paper describes the TREC 2004 QA track in more detail. The next section describes the question series that formed the basis of the evaluation. The following section describes the way the individual question types were evaluated and gives the scores for the runs for that component. Section 3 summarizes the technical

3	Hale Bopp comet		
	3.1	FACTOID	When was the comet discovered?
	3.2	FACTOID	How often does it approach the earth?
	3.3	LIST	In what countries was the comet visible on its last return?
	3.4	OTHER	
21	Club Med		
	21.1	FACTOID	How many Club Med vacation spots are there worldwide?
	21.2	LIST	List the spots in the United States.
	21.3	FACTOID	Where is an adults-only Club Med?
	21.4	OTHER	
22	Franz Kafka		
	22.1	FACTOID	Where was Franz Kafka born?
	22.2	FACTOID	When was he born?
	22.3	FACTOID	What is his ethnic background?
	22.4	LIST	What books did he author?
	22.5	OTHER	

Figure 1: Sample question series from the test set. Series 3 has a THING as a target, series 21 has an ORGANIZATION as a target, and series 22 has a PERSON as a target.

approaches used by the systems to answer the questions. Section 4 looks at the advantages of evaluating runs using a per-series combined score rather than an overall combined score. The final section looks at the future of the track.

1 Question Series

The TREC 2004 QA track consisted of a single task, providing answers for each question in a set of question series. A question series consisted of several factoid questions, zero to two list questions, and exactly one Other question. Associated with each series was a definition target. The series a question belonged to, the order of the question in the series, and the type of each question (factoid, list, or Other) were all explicitly encoded in the XML format used to describe the test set. Example series (minus the XML tags) are shown in figure 1.

The question series were developed as follows. NIST staff searched search engines logs[1] for definition targets. A target was a person, an organization, or thing that was a plausible match for the scenario assumed for the task. The task scenario was the same as in the 2003 track: the questioner was an adult, a native speaker of English, and an "average" reader of US newspapers who was looking for more information about a term encountered while reading the paper.

The set of candidate targets were then given to the assessors, the humans who act as surrogate users and judge the system responses. An assessor selected a target and wrote down questions regarding things he or shee would want to know about the target. The assessor then searched the document collection looking for answers to those questions, plus recording other information about the target that had not asked about but they found interesting. For the most part, the assessors created the questions before doing any searching. However, if the assessor did not know anything about the target (and therefore could create no questions), they first did a Google search to learn about the target, then created questions, and finally searched the document collection. The document collection was the same document set used by the participants as the source of answers, the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31).

NIST staff reviewed the information found by the assessors and constructed the final question series. Because most questions in the final test set had to contain answers in the document collection, and there needed to be sufficient "other" information for the final question in the series, the final series were heavily edited versions of the assessors' original series. This process proved to be more time-consuming than expected, so a few of the question series were constructed directly from searches of the document collection (i.e., the target was not selected from the logs and the questions were developed only after the search).

[1]The search engine logs were donated by Abdur Chowdhury of AOL and Susan Dumais of Microsoft Research for the TREC 2003 track.

Table 1: Targets of the 65 question series.

S1	Crips	S2	Fred Durst	S3	Hale Bopp comet
S4	James Dean	S5	AARP	S6	Rhodes scholars
S7	agouti	S8	Black Panthers	S9	Insane Clown Posse
S10	prions	S11	the band Nirvana	S12	Rohm and Haas
S13	Jar Jar Binks	S14	Horus	S15	Rat Pack
S16	cataract	S17	International Criminal Court	S18	boxer Floyd Patterson
S19	Kibbutz	S20	Concorde	S21	Club Med
S22	Franz Kafka	S23	Gordon Gekko	S24	architect Frank Gehry
S25	Harlem Globe Trotters	S26	Ice-T	S27	Jennifer Capriati
S28	Abercrombie and Fitch	S29	'Tale of Genji'	S30	minstrel Al Jolson
S31	Jean Harlow	S32	Wicca	S33	Florence Nightingale
S34	Amtrak	S35	Jack Welch	S36	Khmer Rouge
S37	Wiggles	S38	quarks	S39	The Clash
S40	Chester Nimitz	S41	Teapot Dome scandal	S42	USS Constitution
S43	Nobel prize	S44	Sacajawea	S45	International Finance Corporation
S46	Heaven's Gate	S47	Bashar Assad	S48	Abu Nidal
S49	Carlos the Jackal	S50	Cassini space probe	S51	Kurds
S52	Burger King	S53	Conde Nast	S54	Eileen Marie Collins
S55	Walter Mosley	S56	Good Friday Agreement	S57	Liberty Bell 7
S58	philanthropist Alberto Vilar	S59	Public Citizen	S60	senator Jim Inhofe
S61	Muslim Brotherhood	S62	Berkman Center for Internet and Society	S63	boll weevil
S64	Johnny Appleseed	S65	space shuttles		

The final test set contained 65 series; the targets of these series are given in table 1. Of the 65 targets, 23 are PERSONs, 25 are ORGANIZATIONs, and 17 are THINGs. The series contain a total of 230 factoid questions, 56 list questions, and 65 (one per target) Other questions. Each series contains at least 4 questions (counting the Other question), with most series containing 5 or 6 questions. The maximum number of questions in a series is 10.

The question series used in the TREC 2004 track are similar to the QACIAD challenge (Question Answering Challenge for Information Access Dialogue) of NTCIR4 [1]. However, there are some important differences. The heavy editing of the assessors' original questions required to make a usable evaluation test set means the TREC series are not true samples of the assessors' original interests in the target. There were many questions that were eliminated because they did not have answers in the document collection or because they did not meet some other evaluation criterion (for example, the answers for many of the original list questions were not named entities). The TREC series are also not true samples of naturally occurring user-system dialog. In a true dialog, the user would most likely mention answers of previous questions in later questions, but the TREC test set specifically did not do this. This appears as a stilted conversational style when viewed from the perspective of true dialog.

Participants were required to submit retrieval results within one week of receiving the test set. All processing of the questions was required to be strictly automatic. Systems were required to process series independently from one another, and required to process an individual series in question order. That is, systems were allowed to use questions and answers from earlier questions in a series to answer later questions in that same series, but could not "look ahead" and use later questions to help answer earlier questions. As a convenience for the track, NIST made available document rankings of the top 1000 documents per target as produced using the PRISE document retrieval system and the target as the query. Sixty-three runs from 28 participants were submitted to the track.

2 Component Evaluations

The questions in the series were tagged as to which type of question they were because each question type had its own response format and evaluation method. The final score for a run was computed as a weighted average of the component scores. The individual component evaluations for 2004 were identical to those used in the TREC 2003 QA track, and are briefly summarized in this section.

2.1 Factoid questions

The system response for a factoid question was either exactly one [*doc-id*, *answer-string*] pair or the literal string 'NIL'. Since there was no guarantee that a factoid question had an answer in the document collection, NIL was returned by the system when it believed there was no answer. Otherwise, *answer-string* was a string containing precisely an answer to the question, and *doc-id* was the id of a document in the collection that supported *answer-string* as an answer.

Each response was independently judged by two human assessors. When the two assessors disagreed in their judgments, a third adjudicator (a NIST staff member) made the final determination. Each response was assigned exactly one of the following four judgments:

incorrect: the answer string does not contain a right answer or the answer is not responsive;

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

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

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

To be responsive, an answer string was required to contain appropriate units and to refer to the correct "famous" entity (e.g., the Taj Mahal casino is not responsive when the question asks about "the Taj Mahal"). NIL responses are correct only if there is no known answer to the question in the collection and are incorrect otherwise. NIL is correct for 22 of the 230 factoid questions in the test set.

The main evaluation score for the factoid component is *accuracy*, the fraction of questions judged correct. Also reported are the recall and precision of recognizing when no answer exists in the document collection. NIL precision is the ratio of the number of times NIL was returned and correct to the number of times it was returned, whereas NIL recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct (22). If NIL was never returned, NIL precision is undefined and NIL recall is 0.0.

Table 2 gives evaluation results for the factoid component. The table shows the most accurate run for the factoid component for each of the top 10 groups. The table gives the accuracy score over the entire set of factoid questions as well as NIL precision and recall scores. In addition, the table reports accuracy for two subsets of the factoid questions: those factoid questions that were the first question in their series (Initial), and those factoid questions that were not the first questions in their series (Non-Initial). As suggested by QACIAD [1], these last two accuracy scores may indicate whether systems had difficulty with context processing in that the first question in a series is usually more fully specified than later questions in a series. (But note there are only 62 initial factoid questions and 168 non-initial factoid questions.)

2.2 List questions

A list question can be thought of as a shorthand for asking the same factoid question multiple times. The set of all correct, distinct answers in the document collection that satisfy the factoid question is the correct answer for the list question.

A system's response for a list question was an unordered set of [*doc-id*, *answer-string*] pairs such that each *answer-string* was considered an instance of the requested type. Judgments of incorrect, unsupported, not exact, and correct were made for individual response pairs as in the factoid judging. The assessor was given one run's entire list at a time, and while judging for correctness also marked a set of responses as distinct. The assessor arbitrarily chose any one of equivalent responses to be distinct, and the remainder were not distinct. Only correct responses could be marked as distinct.

The final set of correct answers for a list question was compiled from the union of the correct responses across all runs plus the instances the assessor found during question development. For the 55 list questions used in the evaluation (one list question was dropped because the assessor decided there were no correct answers during judging), the average

Table 2: Evaluation scores for runs with the best factoid component.

Run Tag	Submitter	Accuracy			NIL Prec	NIL Recall
		All	Initial	Non-Initial		
lcc1	Language Computer Corp.	0.770	0.839	0.744	0.857	0.545
uwbqitekat04	Univ. of Wales, Bangor	0.643	0.694	0.625	0.247	0.864
NUSCHUA1	National Univ. of Singapore	0.626	0.710	0.595	0.333	0.273
mk2004qar1	Saarland University	0.343	0.419	0.315	0.177	0.500
IBM1	IBM Research	0.313	0.435	0.268	—	0.000
mit1	MIT	0.313	0.468	0.256	0.083	0.045
irst04higher	ITC-irst	0.291	0.355	0.268	0.167	0.091
FDUQA13a	Fudan University (Wu)	0.257	0.355	0.220	0.167	0.091
KUQA1	Korea University	0.222	0.226	0.220	0.042	0.045
shef04afv	University of Sheffield	0.213	0.177	0.226	0.071	0.136

Table 3: Average F scores for the list question component. Scores are given for the best run from the top 10 groups.

Run Tag	Submitter	F
lcc1	Language Computer Corp.	0.622
NUSCHUA2	National Univ. of Singapore	0.486
uwbqitekat04	Univ. of Wales, Bangor	0.258
IBM1	IBM Research	0.200
KUQA1	Korea University	0.159
FDUQA13a	Fudan University (Wu)	0.143
MITRE2004B	Mitre Corp.	0.143
UNTQA04M1	University of North Texas	0.128
mk2004qar3	Saarland University	0.125
shef04afv	University of Sheffield	0.125

number of answers per question is 8.8, with 2 as the smallest number of answers, and 41 as the maximum number of answers. A system's response to a list question was scored using instance precision (IP) and instance recall (IR) based on the list of known instances. Let S be the the number of known instances, D be the number of correct, distinct responses returned by the system, and N be the total number of responses returned by the system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined using the F measure with equal weight given to recall and precision:

$$F = \frac{2 \times IP \times IR}{IP + IR}$$

The score for the list component of a run was the average F score over the 55 questions. Table 3 gives the average F scores for the run with the best list component score for each of the top 10 groups.

As happened last year, some submitted runs contained identical list question components as another run submitted by the same group. Since assessors see the lists for each run separately, it can happen that identical components receive different scores. NIST tries to minimize judging differences by making sure the same assessor judges all runs and completes judging one question before moving on to another, but differences remain. These differences are one measure of the error inherent in the evaluation. NIST does not adjust the judgments to make identical runs match because then we wouldn't know what the naturally occurring error rate was, and doing so would bias the scores of systems that submitted identical component runs.

There were 15 pairs of runs with identical list components. Seven pairs had identical average F scores, though some of those seven did have individual questions judged differently. The largest difference in average F scores for identical list components was 0.006, and the largest number of individual questions judged differently for a single run pair was 7.

2.3 Other questions

The Other questions were evaluated using the same methodology as the TREC 2003 definition questions. A system's response for an Other question was an unordered set of [*doc-id*, *answer-string*] pairs as in the list component. Each string was presumed to be a facet in the definition of the series' target that had not yet been covered by earlier questions in the series. The requirement to not repeat information already covered by earlier questions in the series made answering Other questions somewhat more difficult than answering TREC 2003 definition questions.

Judging the quality of the systems' responses was done in two steps. In the first step, all of the answer strings from all of the systems' responses were presented to the assessor in a single list. Using these responses and the searches done during question development, the assessor created a list of information nuggets about the target. An information nugget is an atomic piece of information about the target that is interesting (in the assessor's opinion) and was not part of an earlier question in the series or an answer to an earlier question in the series. An information nugget is atomic if the assessor can make a binary decision as to whether the nugget appears in a response. Once the nugget list was created for a target, the assessor marked some nuggets as vital, meaning that this information must be returned for a response to be good. Non-vital nuggets act as don't care conditions in that the assessor believes the information in the nugget to be interesting enough that returning the information is acceptable in, but not necessary for, a good response.

In the second step of judging the responses, the assessor went through each system's response in turn and marked which nuggets appeared in the response. A response contained a nugget if there was a *conceptual* match between the response and the nugget; that is, the match was independent of the particular wording used in either the nugget or the response. A nugget match was marked at most once per response—if the response contained more than one match for a nugget, an arbitrary match was marked and the remainder were left unmarked. A single [*doc-id*, *answer-string*] pair in a system response could match 0, 1, or multiple nuggets.

Given the nugget list and the set of nuggets matched in a system's response, the nugget recall of the response is the ratio of the number of matched nuggets to the total number of vital nuggets in the list. Nugget precision is much more difficult to compute since there is no effective way of enumerating all the concepts in a response. Instead, a measure based on length (in non-white space characters) is used as an approximation to nugget precision. The length-based measure starts with an initial allowance of 100 characters for each (vital or non-vital) nugget matched. If the total system response is less than this number of characters, the value of the measure is 1.0. Otherwise, the measure's value decreases as the length increases using the function $1 - \frac{length - allowance}{length}$. The final score for an Other question was computed as the F measure with nugget recall three times as important as nugget precision:

$$F(\beta = 3) = \frac{10 \times \text{precision} \times \text{recall}}{9 \times \text{precision} + \text{recall}}.$$

The score for the Other question component was the average F($\beta = 3$) score over 64 Other questions. The Other question for series S7 was mistakenly left unjudged, so it was removed from the evaluation. Table 4 gives the average F($\beta = 3$) score for the best scoring Other question component for each of the top 10 groups.

As with list questions, a system's response for an Other question must be judged as a unit, so identical responses may receive different scores. There were 13 pairs of runs with identical Other question components. The differences between the run pairs' average F($\beta = 3$) scores were {0.012, 0.0, 0.0, 0.0, 0.0, 0.007, 0.0, 0.007, .003, 0.007, 0.0, 0.012, 0.003}, and the number of Other questions that received a different score between the run pairs was {12, 0, 0, 0, 0, 5, 5, 4, 3, 3, 10, 4, 1} respectively.

2.4 Combined weighted average

The final score for a QA run was computed as a weighted average of the three component scores:

$$\text{FinalScore} = .5 \times \text{FactoidAccuracy} + .25 \times \text{ListAveF} + .25 \times \text{OtherAveF}.$$

Since each of the component scores ranges between 0 and 1, the final score is also in that range. Table 5 shows the combined scores for the best run for each of the top 10 groups. Also given in the table are the weighted component scores that make up the final sum.

Table 4: Average F($\beta = 3$) scores for the Other questions component. Scores are given for the best run from the top 10 groups.

Run Tag	Submitter	F($\beta = 3$)
NUSCHUA2	National Univ. of Singapore	0.460
FDUQA13a	Fudan University (Wu)	0.404
NSAQACTIS1	National Security Agency	0.376
ShefMadCow20	University of Sheffield	0.321
UNTQA04M3	University of North Texas	0.307
IBM1	IBM Research	0.285
KUQA3	Korea University	0.247
lcc1	Language Computer Corp.	0.240
clr04r1	CL Research	0.239
mk2004qar3	Saarland University	0.211

Table 5: Weighted component scores and final combined scores for QA task runs. Scores are given for the best run from the top 10 groups.

		Weighted Component Score			Final
Run Tag	Submitter	Factoid	List	Other	Score
lcc1	Language Computer Corp.	0.385	0.155	0.060	0.601
NUSCHUA1	National Univ. of Singapore	0.313	0.120	0.112	0.545
uwbqitekat04	Univ. of Wales, Bangor	0.322	0.065	0.000	0.386
IBM1	IBM Research	0.157	0.050	0.071	0.278
FDUQA13a	Fudan University (Wu)	0.129	0.036	0.101	0.265
mk2004qar3	Saarland University	0.172	0.031	0.053	0.256
mit1	MIT	0.157	0.030	0.046	0.232
irst04higher	ITC-irst	0.145	0.026	0.052	0.223
shef04afv	University of Sheffield	0.106	0.031	0.078	0.216
KUQA1	Korea University	0.111	0.040	0.061	0.212

3 System Approaches

The overall approach taken for answering factoid questions has remained unchanged for the past several years. Systems generally determine the expected answer type of the question, retrieve documents or passages likely to contain answers to the question using important question words and related terms as the query, and then perform a match between the question words and retrieved passages to extract the answer. While the overall approach has remained the same, individual groups continue to refine their techniques for these three steps, increasing the coverage and accuracy of their systems.

Most groups use their factoid-answering system for list questions, changing only the number of responses returned as the answer. The main issue is determining the number of responses to return. Systems whose matching phase creates a question-independent score for each passage return all answers whose score is above an empirically determined threshold. Other systems return all answers whose scores were within an empirically determined fraction of the top result's score.

The fact that target and list questions did not necessarily explicitly include the target of the question was a new difficulty in this year's track. For the document/passage retrieval phase, most systems simply appended the target to the query. This was an effective strategy since in all cases the target was the correct domain for the question, and most of the retrieval methods used treat the query as a simple set of keywords. There were a variety of approaches taken by different systems to address this difficulty in phases that require more detailed processing of the question. While a few systems made no attempt to include the target in the question, a much more common approach was to append the target to the question. Another common approach was to replace all pronouns in the questions with the target. While many (but not all) pronouns in the questions did in fact refer to the target, this approach suffered when the question used a definite noun phrase rather than a pronoun to refer to the target (e.g., using "the band" when the target was Nirvana). Finally, other systems tried varying degrees of true anaphora resolution to appropriately resolve references in the questions. It is difficult to judge how much benefit these systems received from this more extensive processing since the majority of pronoun references were to the target.

Systems generally used the same techniques as were used for TREC 2003's definition questions to answer the Other questions. Most systems first retrieve passages about the target using a recall-oriented retrieval search. Subsequent processing reduces the amount of material returned. Some systems used pattern-matching to locate definition-content in text. These patterns, such as looking for copular constructions and appositives, were either hand-constructed or learned from a training corpus. Systems also looked to eliminate redundant information, using either word overlap measures or document summarization techniques. Unlike last year, answers to previous questions in the series had to be incorporated as part of the redundant information for this year's task. The output from the redundancy-reducing step was then returned as the answer for the Other question.

4 Per-series Combined Weighted Scores

The series play no role in computing the combined average score as above. That is, questions are treated independently without regard to the series they appear in for scoring purposes. This is unfortunate since each individual series is an abstraction of a single user's interaction with the system. Evaluating over the individual series should provide a more accurate representation of the effectiveness of the system from an individual user's perspective. This section examines the effectiveness of a per-series evaluation.

Since each series is a mixture of different question types, we can compute the weighted average score on a per-series basis, and take the average of the per-series scores as the final score for the run. Note that the average per-series weighted score (call this the per-series score) will not in general be equal to the final score computed as the weighted average of the three component scores (the global score) since the two averages emphasize different things. The global score gives equal weight to individual questions within a component. The per-series score gives equal weight to each series. (This is the same difference between micro- and macro-averaging of document retrieval scores.) To compute the combined score for an individual series that contained all three question types, the same weighted average of the different question types was used, but only the scores for questions belonging to the series were part of the computation. For those series that did not contain any list questions, the weighted score was computed as $.67 \times \text{FactoidAccuracy} + .33 \times \text{OtherF}$. All of series S7 was eliminated from the evaluation since that was the series

Table 6: Per-series scores for QA task runs. Scores are given for the best run from the top 10 groups. Also given is the global score (as given in Table 5) for comparison.

Run Tag	Submitter	Per-series	Global
lcc1	Language Computer Corp.	0.609	0.601
NUSCHUA1	National Univ. of Singapore	0.557	0.545
uwbqitekat04	Univ. of Wales, Bangor	0.401	0.386
IBM1	IBM Research	0.289	0.278
FDUQA13a	Fudan University (Wu)	0.289	0.265
mk2004qar3	Saarland University	0.271	0.256
mit1	MIT	0.253	0.232
irst04higher	ITC-irst	0.239	0.223
shef04afv	University of Sheffield	0.230	0.216
NSAQACTIS1	National Security Agency	0.226	0.211

whose Other question was not evaluated.

Table 6 shows the per-series score for the best run for each of the top 10 groups. The global score is repeated in the table for comparison. For the particular set of runs shown in the table, all of the runs rank in the same order by the two scoring methods, except that the tenth run is different for the two schemes (the NSAQACTIS1 run edges out the KUQA1 run when using the per-series score). The absolute value of the per-series score is somewhat greater than the global score for these runs, though it is possible for the global score to be the greater of the two.

Each individual series has only a few questions, so the combined weighted score for an individual series will be much less stable than the global score. But the average of 64 per-series scores should be at least as stable as the overall combined weighted average and has some additional advantages. The per-series score is computed at a small enough granularity to be meaningful at the task-level (i.e., each series representing a single user interaction), and at a large enough granularity for individual scores to be meaningful. Figure 2 shows a box-and-whiskers plot of the per-series scores across all runs for each series. A box in the plot shows the extent of the middle half of the scores for that series, with the median score indicated by the line through the box. The dotted lines (the "whiskers") extend to a point that is 1.5 times the interquartile distance, or the most extreme score, whichever is less. Extreme scores that are greater than the 1.5 times the interquartile distance are plotted as circles. The plot shows that only a few series (S21, S25, S37, S39) have median scores of 0.0. This is in sharp contrast to the median scores of individual questions. For factoid questions, 212 of the 230 questions (92.2%) have a zero median; for list questions 39 of 55 questions (70.9%) have a zero median; and for Other questions 41 of 64 questions (64.1%) have a zero median.

One of the hypotheses during question development was that system effectiveness would depend on the type of target. For example, PERSON targets may be easier for systems to define since the set of information desired for a person may be more standard then the set of information desired for a THING. This hypothesis has little support in the overall results of the track (there may be individual systems that show stronger dependencies). The average of the average per-series score across all runs and all series is 0.187. The averages for series restricted to particular target types are 0.184 for PERSON targets, 0.179 for ORGANIZATION targets, and 0.206 for THING targets.

5 Future of the QA Track

Several concerns regarding the TREC 2005 QA track were raised during the TREC 2004 QA breakout session. Since the TREC 2004 task was rather different from previous years' tasks, there was the desire to repeat the task largely unchanged. There was also the desire to build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology. As a result of this discussion, the main task for the 2005 QA track was decided to be essentially the same as the 2004 task in that the test set will consist of a set of question series where each series asks for information regarding a particular target. As in TREC 2004, the targets will include people, organizations, and other entities; unlike TREC 2004 the target can also be an event. Events were added since the document set from which the answers are to be drawn are newswire articles. Each question series will consist of some

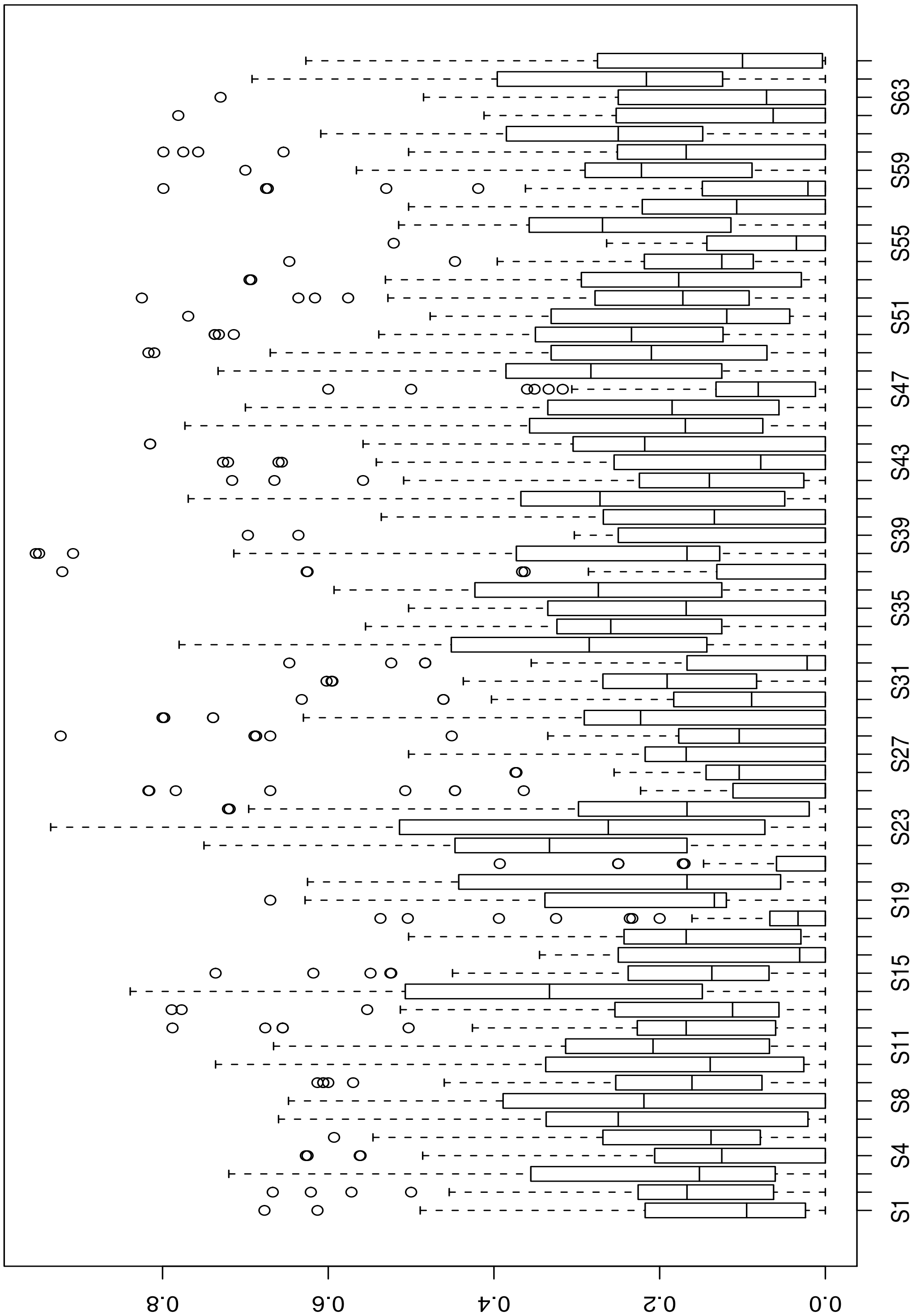

Figure 2: Box and whiskers plot of per-series combined weighted scores across all runs. The x-axis shows the series number (recall that series S7 was omitted), and the y-axis the combined weighted score for that series.

factoid and some list questions and will end with exactly one "Other" question. The answer to the "Other" question is to be interesting information about the target that is not covered by the preceding questions in the series. The runs will be evaluated using the same methodology as in TREC 2004, though the primary measure will be the per-series score.

To address the concern regarding document retrieval and QA, TREC 2005 submissions will be required to include an ordered list of documents for each question. This list will represent the the set of documents used by the system to create its answer, where the order of the documents in the list is the order in which the system considered the document. The purpose of the lists is to create document pools both to get a better understanding of the number of instances of correct answers in the collection and to support research on whether some document retrieval techniques are better than others in support of QA. For some subset of approximately 50 questions, NIST will pool the document lists, and assessors will judge each document in the pool as relevant ("contains an answer") or not relevant ("does not contain an answer"). Document lists will then be evaluated using using trec_eval measures.

References

[1] Tsuneaki Kato, Jun'ichi Fukumoto, Fumito Masui, and Noriko Kando. Handling information access dialogue through QA technologies—A novel challenge for open-domain question answering. In *Proceedings of the HLT-NAACL 2004 Workshop on Pragmatics of Question Answering*, pages 70–77, May 2004.

[2] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text REtreival Conference (TREC 2001)*, pages 42–51, 2002.

[3] Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, 2004.