Overview of the TREC 2001 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. In its third year, the track continued to focus on retrieving small snippets of text that contain an answer to a question. However, several new conditions were added to increase the realism, and the difficulty, of the task. In the main task, questions were no longer guaranteed to have an answer in the collection; systems returned a response of 'NIL' to indicate their belief that no answer was present. In the new list task, systems assembled a set of instances as the response for a question, requiring the ability to distinguish among instances found in multiple documents. Another new task, the context task, required systems to track discourse objects through a series of questions.

The TREC 2001 question answering (QA) track was the third running of a QA track in TREC. The goal of the track has remained the same each year: to foster research on systems that retrieve answers rather than documents in response to a question, with a particular emphasis on systems that can function in unrestricted domains. Systems are given a large corpus of newspaper and newswire articles and a set of closed-class questions such as *Who invented the paper clip?*. They return a short ($\leq$ 50 bytes) text snippet and a document as a response to a question, where the snippet contains an answer to the question and the document supports that answer.

While the overall goal has remained the same in each running of the track, this year's track introduced new conditions to increase the realism of the task. The track included three separate tasks this year, the main task, the list task, and the context task. The main task was essentially the same as the task in previous years except questions were no longer guaranteed to have an answer in the collection. In the list task, systems assembled a set of instances as the response for a question, requiring the ability to distinguish among instances found in multiple documents. The context task required systems to track discourse objects through a series of questions.

This paper gives an overview of the TREC 2001 track. The next section provides the background that is common to all of this year's tasks. The following three sections then describe each of this year's tasks in turn. The final section discusses the future of the QA track.

1 Background

1.1 History of the TREC QA track

The QA track was started in 1999 (TREC-8) to support research on methods that would allow systems to move away from *document* retrieval toward *information* retrieval. An additional goal of the track was to define a task that would appeal to both the document retrieval and information extraction communities. Information extraction (IE) systems, such as those used in the Message Understanding Conferences (MUCs, see `http://www.itl.nist.gov/iad/894.02/related_projects/muc`), recognize particular kinds of entities and relationships among those entities in running text. Since answers to closed-class questions are generally entities of the types recognized by IE systems, the QA task was limited to answering closed-class questions. There were no restrictions on the domain the questions could be drawn from, however, and the data source was a large collection of free-text documents.

TREC QA track participants were given the document collection and a test set of questions. The questions were generally fact-based, short-answer 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. 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 bytes 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 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.

Not surprisingly, allowing 250 bytes in a response was an easier task than limiting responses to 50 bytes: for every organization that submitted runs of both lengths, the 250 byte limit run had a higher mean reciprocal rank. In the 50 byte limit case, the best performing systems were able to answer about 70 % of the questions in TREC-8 and about 65 % of the questions in TREC-9. While the 65 % score was a slightly worse result than the TREC-8 scores in absolute terms, it represented a very significant improvement in question answering systems. The TREC-9 task was considerably harder than the TREC-8 task because TREC-9 used actual users' questions while TREC-8 used questions constructed specifically for the track.

Most participants used a version of the following general approach to the question answering problem. The system first attempted 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" implies a person or an organization is being sought, and a question beginning with "when" implies a time designation is needed. Next, the system retrieved a small portion of the document collection using standard text retrieval technology and the question as the query. The system performed a shallow parse of the returned documents to detect entities of the same type as the answer. If an entity of the required type was found sufficiently close to the question's words, the system returned that entity as the response. If no appropriate answer type was found, the system fell back to best-matching-passage techniques. Improvements in TREC-9 systems generally resulted from doing a better job of classifying questions as to the expected answer type, and using a wider variety of methods for finding the entailed answer types in retrieved passages.

In October 2000, the DARPA TIDES project released a white paper that included a roadmap for question answering research [3]. The paper described an ambitious program to increase the complexity of the types of questions that can be answered, the diversity of sources from which the answers can be drawn, and the means by which answers are displayed. It also included a five year plan for introducing aspects of these research areas into the TREC QA track, with the TREC 2001 track as the first year of the plan. The two new requirements suggested by the roadmap for TREC 2001 were including questions whose answers were scattered across multiple documents, and no longer guaranteeing an answer is present in the document collection. These new requirements were the motivation for the list task and removing the guarantee in the main task. The context task was added as a pilot study for TREC 2002 since the roadmap's new requirement for next year is question answering within a context.

1.2 Answer assessment

The TREC QA evaluations have been based on the assumption that different people will have different ideas as to what constitutes a correct answer. This assumption was demonstrated to be true during the TREC-8 evaluation. For TREC-8, 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. Assessors had legitimate differences of opinion as to what constituted an acceptable answer even for the deliberately constrained questions used in the track. Two prime examples of where such differences arise are the completeness of names and the granularity of dates and locations.

Fortunately, as with document retrieval evaluation, the relative scores between QA systems remain stable despite differences in the judgments used to evaluate them [4]. The lack of a definitive answer key does mean that evaluation scores are only meaningful in relation to other scores on the same data set. Absolute scores *do* change if you use a different set of judges, or a different set of questions. However, this is an unavoidable characteristic of QA evaluation. Given that assessors' opinions of correctness differ, the eventual end users of the QA systems will have similar differences of opinion, and thus any evaluation of the technology must accommodate these differences.

A [*document-id*, *answer-string*] pair was judged correct if, in the opinion of the NIST 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" (except in TREC-8 where it was marked correct). 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 [5] for a very detailed discussion of responsiveness.

The basic unit of response for each of the tasks in this year's QA track was once again the [*document-id*, *answer-string*] pair, though all strings were limited to no more than 50 bytes. Response pairs were judged as described above. For the main task, each question was independently judged by two assessors despite the TREC-8 results that showed using multiple assessors per question is not necessary to get stable evaluation results. The TREC-9 judgments, which used only one assessor per question, contain a larger number of blunders (out-and-out mistakes, not differences of opinions) than anticipated. While comparative evaluation results are stable despite such errors, the judgments are also used as training data and the effect of the errors for training is not clear. To reduce the number of errors for the TREC 2001 main task, each question was judged by two assessors and differences between judges were flagged. NIST staff reviewed each flagged response to determine if the difference was a matter of opinion or a mistake. If the reviewer found the difference to be a matter of opinion, the opinion of the first judge prevailed. Otherwise, the reviewer corrected the mistake. On average, the two assessors disagreed on 5 % of the responses, and the initial judgment was changed in 30 % of the cases where there was a disagreement. As a check of the earlier TREC-8 result, we computed the correlations among the system rankings produced by evaluating the main task runs on the different judgment sets. Once again the correlations were very high (greater than 0.96), indicating that the results are stable regardless of which judgment set is used.

The QA roadmap called for another change for the TREC 2001 track that was not implemented: requiring systems to return an actual answer rather than a string that contains an answer. This change was not implemented because it is not clear how to operationalize "actual answer". Is a string wrong if it contains an answer and justification of that answer? Are titles required parts of names or extraneous information? Nonetheless, some move toward "actual answer" will be necessary for future tracks since allowing assessors to pick out the answer from a returned string is masking large differences between systems. For example, Figure 1 shows some of the answer strings that were judged correct for question 916, *What river in the US is known as the Big Muddy?*. Each of these strings should be marked correct according to the current assessment procedure, but some are much better responses than others.

1.3 The TREC 2001 track

The document set for all tasks was the set of news articles on the combined set of TIPSTER/TREC disks. In particular, this includes the AP newswire from disks 1–3, the *Wall Street Journal* from disks 1–2, the *San Jose Mercury News* from disk 3, the *Financial Times* from disk 4, the *Los Angeles Times* from disk 5, and the Foreign Broadcast Information Service (FBIS) from disk 5. This set contains approximately 979,000 articles in 3,033 megabytes of text, and covers a broad spectrum of subjects.

As a service to the track, 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

```
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?*

Table 1: Number of runs per task (Main, List, Context) submitted by TREC 2001 QA track participants.

Organization	M	L	C	Organization	M	L	C
Alicante University	2	–	–	National Taiwan University	2	2	1
Chinese Academy of Sciences	3	–	–	NTT Communication Sci. Labs	1	–	–
CL Research	2	2	1	Oracle	1	–	–
Conexor Oy	1	–	–	Pohang U. of Sci. & Tech.	1	–	–
EC Wise, Inc.	2	–	–	Queens College, CUNY	3	2	2
Fudan University	1	–	–	Sun Microsystems Labs	2	–	–
Harbin Institute of Technology	1	–	–	Syracuse University	2	2	–
IBM (Ittycheriah)	3	–	–	Tilburg University	2	–	–
IBM (Prager)	3	–	–	Universite de Montreal	3	2	–
InsightSoft-M	1	–	–	University of Alberta	1	–	–
ITC-irst	1	–	–	University of Amsterdam	3	2	–
KAIST	2	2	1	U. Illinois, Urbana/Champaign	1	–	–
KCSL	1	–	–	University of Iowa	2	–	–
Korea University	2	–	–	University of Pennsylvania	1	–	–
Language Computer Corp.	1	1	1	University of Pisa	3	–	–
LIMSI	3	–	–	U. of Southern California, ISI	2	1	–
Microsoft Research	2	–	–	University of Waterloo	3	2	1
MITRE	1	–	–	University of York	2	–	–

given from that same ranking). For the context task, the rankings were produced for the first question in a series only. 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-six groups submitted a total of 92 runs to the QA track. Table 1 lists each participating group and the number of runs that group submitted for each task.

2 The Main Task

The main QA task was very similar to previous years' tasks, providing continuity with previous years and giving newcomers to the track a stable task with which to begin. Participants received a set of closed-class questions and searched a large document set to extract (or construct) an answer to each 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 and the document supported that answer. Answer strings were limited to no more than 50 bytes.

Questions were not guaranteed to have an answer in the document collection. Recognizing that there is no answer is a challenging task, but it is an important ability for operational systems to possess since returning an incorrect answer is usually worse than not returning an answer at all. Systems indicated their belief that there was no answer in the document collection by returning 'NIL' rather than a [*document-id*, *answer-string*] pair as one of the responses to a question. The NIL response was scored the same as other responses; NIL was correct when no correct answer was known to exist in the collection for that question. A correct answer was known to exist if the assessor found a correct answer during question development, or if some system returned a correct, *supported* response to the question. Forty-nine questions had no known correct response in the document collection.

Recognizing that no answer exists in the document collection is a different task from having the system recognize that it does not know what the answer is. This latter task is also important, but is more difficult to evaluate because systems must then be scored using a combination of questions attempted and attempted questions correctly answered. As an initial foray into evaluating whether systems can determine if they know the answer, systems were required to report a single final answer for each question in addition to the ranked list of 5 responses. The final answer was either an integer from one to five that referred to a position in the ranked list of responses for that question, or the string 'UNSURE' that indicated the system did not know what the answer was. While the vast majority of systems returned the answer at rank one if they were confident of the answer, a few systems did return an alternate rank.

2.1 Test questions

The test set of questions continued a progression of using more realistic questions in each of the three runnings of the track. In TREC-8, the majority of the questions were created expressly for the track, and thus tended to be back-formulations of a statement in a document. In TREC-9, the questions were selected from an Encarta log that contained actual questions, and a raw Excite log. Since the raw Excite log did not contain many grammatically well-formed questions, NIST staff used the Excite log as a source of ideas for actual questions. All the questions were created without looking at any documents. The resulting test set of questions was much more difficult than the TREC-8 set, mainly because the TREC-9 set contained many more high-level questions such as *Who is Colin Powell?*. For this year's main task, the source of questions was a set of filtered MSNSearch logs and AskJeeves logs. Raw logs were automatically filtered (at Microsoft and AskJeeves) to select queries that contained a question word (e.g., what, when, where, which, etc.) anywhere in the query; that began with modals or the verb to be (e.g., are, can, could, define, describe, does, do, etc.); or that ended with a question mark. NIST did additional human filtering on these logs, removing queries that were not in fact questions; questions that asked for a list of items; procedural questions; questions that asked for the location of something on the web (e.g., pictures of someone); yes/no questions; and questions that were obviously too current for the document collection (e.g., questions about Britney Spears, etc.). The assessors then searched the collection looking for answers for the queries that remained.

The final question set consisted of 500 questions. NIST fixed the spelling, punctuation, and sometimes the grammar of the queries selected to be in the final question set, but except for a very few (less than 10) questions, the content of the question was precisely what was in the log. The few changes that were made were simple changes such as substituting one Greek god for another so that the question would have an answer in the collection.

NIST has made no attempt to control the relative number of different types of questions in the test set from year to year. Instead, the distribution of question types in the final test set has reflected the distribution

Table 2: Evaluation scores for a subset of the TREC 2001 main task runs. Scores are given for the best run from the top 15 groups. Scores include the mean reciprocal rank (MRR), and number (# qs) and percentage (%) of questions for which no correct response was returned for both strict (unsupported responses counted as wrong) and lenient (unsupported responses counted as correct) evaluation. Also included are the number of questions for which the run returned 'NIL' as a response (NIL Returned), the number of questions for which 'NIL' was correctly returned as a response (NIL Correct), the percentage of questions for which the system was sure of its final answer (Final Sure), and the percentage of questions for which the final answer was correct when the system was sure (Sure Correct).

	Strict Evaluation			Lenient Evaluation			# qs	# qs		
		No Correct			No Correct		NIL	NIL	Final	Sure
Run Tag	MRR	# qs	%	MRR	# qs	%	Returned	Correct	Sure	Correct
insight	0.68	152	30.9	0.69	147	29.9	120	38	75 %	77 %
LCC1	0.57	171	34.8	0.59	159	32.3	41	31	100 %	51 %
orcl1	0.48	193	39.2	0.49	184	37.4	82	35	100 %	40 %
isi1a50	0.43	205	41.7	0.45	196	39.8	407	33	80 %	38 %
uwmta1	0.43	212	43.1	0.46	200	40.7	492	49	100 %	35 %
mtsuna0	0.41	220	44.7	0.42	213	43.3	492	49	100 %	32 %
ibmsqa01a	0.39	218	44.3	0.40	212	43.1	192	28	100 %	30 %
IBMKS1M3	0.36	220	44.7	0.36	211	42.9	206	27	100 %	24 %
askmsr	0.35	242	49.2	0.43	197	40.0	491	49	100 %	27 %
pir1Qqa3	0.33	264	53.7	0.33	260	52.8	5	0	100 %	24 %
posqa10a	0.32	276	56.1	0.34	260	52.8	13	3	100 %	24 %
ALIC01M2	0.30	297	60.4	0.31	293	59.6	4	0	100 %	23 %
gazoo	0.30	304	61.8	0.31	300	61.0	11	0	100 %	24 %
kuqa1	0.29	298	60.6	0.30	295	60.0	6	0	100 %	23 %
prun001	0.27	333	67.7	0.27	332	67.5	201	38	100 %	24 %

in the source of questions. This year, the number of questions that asked for a definition was dramatically greater than in previous years. (Ken Litkowski of CL Research puts the count at 135/500 definition questions for TREC 2001 compared to 31/500 for TREC-9.) While a large fraction of definition questions is "real" in that the filtered MSNSearch and AskJeeves logs contain many definition questions, there are easier ways to find the definitions of terms than searching for a concise definition in a corpus of news articles. NIST will need to exert more control over the distribution of question types in future tracks.

Eight questions were removed from the evaluation, mostly due to spelling mistakes in the question. A ninth question, question 1070, also contains a typo, spelling 'Louvre' as 'Lourve'. However, that mistake was not noted until all results had been evaluated, so it remains in the test set.

2.2 Retrieval results

Table 2 gives evaluation results for the top fifteen groups. Only one run per group is included in the table. The table gives the mean reciprocal rank (MRR) scores, and the number and percentage of questions for which no correct response was returned for both strict (unsupported responses counted as wrong) and lenient (unsupported responses counted as correct) evaluation. The table also gives statistics regarding NIL and final answer processing.

Detecting whether or not an answer exists in the collection is feasible—the LCC1 run had an accuracy of 31/41 or 0.76—but apparently difficult—only five runs had an accuracy greater than 0.25 (see Table 3). (Accuracy is computed as the number of questions for which NIL was correctly returned divided by the total number of questions for which NIL was returned.) Since systems could return a ranked list of up to five responses per question, some systems returned NIL as one of the responses for every question. This resulted in an accuracy of only 0.1 (49/492), but tended to increase the overall MRR score of those systems somewhat since it is relatively rare to get the first correct response at large ranks when there is an answer

Table 3: Main task runs that had an accuracy greater than 0.25 in detecting when no answer was present in the collection. Returning NIL for all questions produced an accuracy of 0.1.

Run Tag	Returned	Correct	Accuracy
LCC1	41	31	0.76
orcl1	82	35	0.43
insight	120	38	0.37
ICTQA10a	35	10	0.29
ICTQA10b	55	15	0.27

in the collection. See the MultiText project's paper in this proceedings for an analysis of this effect [1].

In final answer processing the system was to indicate whether it was confident in the answer it produced or it recognized that it did not know what the answer was. The purpose of final answer processing in TREC 2001 was to gather data for investigating evaluation strategies for systems that can return "I don't know" as a response. Unfortunately, not enough data was collected to analyze new strategies since more than half of the runs were always confident in their response (see the last two columns of Table 2).

Almost all systems used variants of the strategy seen in earlier TRECs to perform the main task: determine the answer type from the form of the question; retrieve a small portion of the document set; and find the correct answer type in a document piece. Most systems used a lexicon, usually WordNet [2], to verify that a candidate response was of the correct type. Some systems also used the lexicon as a source of definitions.

While most systems used the same basic strategy, there was much less agreement on the best approaches to realizing that strategy. Many groups continued to build systems that attempt a full understanding of the question, but increasingly many groups took a more shallow, data-driven approach. The data-driven approaches rely on simpler pattern matching methods using very large corpora (frequently the web) rather than sophisticated language processing. The idea exploited in the massive data approach is the fact that in a large enough data source a correct answer will usually be repeated often enough to distinguish it from the noise that happens to occasionally match simple patterns.

A second area in which there is no consensus as to the best approach is classification schemes for answer types. Some systems use a few very broad classes of answer types, while others use many specialized classes. The difference is a standard trade-off between coverage and accuracy. With many specialized answer types, finding the actual answer once an answer type is correctly classified is much easier because of the specificity of the class. However, deciding which class is correct, and ensuring there is a class for all questions, is much more difficult with many specialized classes. Some systems use a hierarchical answer typology to exploit this trade-off.

3 The List Task

As mentioned above, one of the goals for the TREC 2001 QA track was to require 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.

The list task accomplished this goal. 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 2. Each instance was guaranteed to obey the same constraints as an individual answer in the main task and was judged in the same manner as a response in the main task: each true instance was no more than 50 characters long; some document was explicit that it was an instance of the desired type; each answer string had to have an associated document that supported the answer; answer strings had to be responsive; etc. The document collection was guaranteed to contain at least the target number of instances. 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 and NIST staff since

- Name 4 U.S. cities that have a "Shubert" theater.
- Name 30 individuals who served as a cabinet officer under Ronald Reagan.
- Who are 6 actors who have played Tevye in "Fiddler on the Roof"?
- Name 4 countries that can produce synthetic diamonds.
- What are 9 novels written by John Updike?

Figure 2: Example list task questions.

Table 4: Average accuracy of the TREC 2001 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
LCC2	0.76	UdeMlistP	0.15
isi1l50	0.45	qntual2	0.14
pir1Qli1	0.34	UAmsT10qaL2	0.13
SUT10PARLT	0.33	clr01l1	0.13
SUT10DOCLT	0.25	UAmsT10qaL1	0.12
uwmtal1	0.25	clr01l2	0.12
uwmtal0	0.23	KAISTQALIST1	0.08
pir1Qli2	0.20	KAISTQALIST2	0.07
qntual1	0.18	UdeMlistB	0.07

there were not enough appropriate questions in the logs. 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 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 correct, incorrect, or not supported were made individually for each [*document-id*, *answer-string*] pair. 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. Incorrect responses were always marked as not distinct (the assessment software enforced this), but unsupported responses could be marked distinct.

Since answer strings could be up to fifty bytes long, a single string might contain more than one instance. The track guidelines specified that the left-most instance in a string was always counted as the instance of record for that string. For example for the question *Name 9 countries that import Cuban sugar.*, the string `China and Russia imported Cuban sugar` was counted as an instance of China only. If another answer string in the list was `China imports`, one of the two responses would be marked as distinct for China, and Russia still would not be counted as a retrieved instance.

List results were evaluated using accuracy, the number of distinct responses divided by the target number of instances. Note that since unsupported responses could be marked distinct, the reported accuracy is a lenient evaluation. Table 4 gives the average accuracy scores for all of the list task submissions.

Given the way the questions were constructed for the list task, the list task questions were intrinsically easier than the questions in the main task. Most systems found at at least one instance for most questions. Each system returned some duplicate responses, but duplication was not a major source of error for any of the runs. (Each run contained many more wrong responses than duplicate responses.) With just 18 runs, there is not enough data to know if the lack of duplication is because the systems are good at recognizing and eliminating duplicate responses, or if there simply wasn't all that much duplication in the document set.

4 The Context Task

The context task was intended to test the systems' ability to track discourse objects (context) through a series of questions. Eventual users of QA systems will likely interact with the system on a regular basis, and the user will expect the system to have some basic understanding of previous interactions. The TREC 2001 context task was designed to represent the kind of dialog processing that a system would require to support an interactive user session.

The type of questions that were used in the context task was the same as in the main task. However, the questions were grouped into different series, and the QA system was expected to track the discourse objects across the individual questions of a series. That is, the interpretation of a question later in the series could depend on the meaning or answer of an earlier question in the series. Correct interpretation of a question often involved resolving referential links within and across questions. Figure 3 gives three examples of series used in the context task.

NIST staff created ten question series for the context task. Most series contained three or four questions, though one series contained nine questions. There were 42 questions across all series, and each question was guaranteed to have an answer in the document collection. A context task run consisted of a ranked list of up to five [*document-id*, *answer-string*] pairs per question as in the main task. The context task questions were judged and evaluated as in the main task as well, except that only lenient evaluation scores (unsupported as correct) were computed. All questions were judged by the same assessor.

Seven runs were submitted to the context task, with unexpected results. The ability to correctly answer questions later in a series was uncorrelated with the ability to correctly answer questions earlier in the series. The first question in a series defined a small enough subset of documents that results were dominated by whether the system could answer the particular type of the current question, rather than by the systems' ability to track context. Thus, this task is not a suitable methodology for evaluating context-sensitive processing for the current state-of-the-art in question answering.

5 Future

The TREC QA track will continue, with the selection of tasks included in future tracks influenced by both the QA roadmap and the ARDA AQUAINT program (see `http://www.ic-arda.org/InfoExploit/aquaint/index.html`). The goal of future tracks is to increase the kinds and difficulty of the questions that systems can answer.

The main task in TREC 2002 will focus on having systems retrieve the *exact* answer as opposed to text snippets that contain the answer. While this will entail marking as incorrect "good" responses such as an answer plus justification, we believe that forcing systems to be precise will ultimately produce better QA technology. Systems will be allowed to return only one response per question, another change aimed at forcing systems to be more precise. NIST will exert more control over the relative proportions of different kinds of questions in the test set. In particular, definitional questions will be a very small percentage of the total question set.

The list task will be repeated in essentially the same form as TREC 2001. NIST will attempt to find naturally occurring list questions in logs, but appropriate questions are rare, so some constructed questions may also be used. We hope also to have a new context task, though the exact nature of that task is still undefined.

The main focus of the ARDA AQUAINT program is to move beyond the simple factoid questions that have been the focus of the TREC tracks. Of particular concern for evaluation is how to score responses that cannot be marked simply correct/incorrect, but instead need to incorporate a fine-grained measure of the quality of the response. We expect that NIST and the AQUAINT contractors will run pilot studies to

CTX1a Which museum in Florence was damaged by a major bomb explosion in 1993?

CTX1b On what day did this happen?

CTX1c Which galleries were involved?

CTX1d How many people were killed?

CTX1e Where were these people located?

CTX1f How much explosive was used?

CTX3a What grape variety is used in Chateau Petrus Bordeaux?

CTX3b How much did the futures cost for the 1989 vintage?

CTX3c Where did the winery's owner go to college?

CTX3d What California winery does he own?

CTX10a How many species of spiders are there?

CTX10b How many are poisonous to humans?

CTX10c What percentage of spider bites in the U.S. are fatal?

Figure 3: Example question series for the context task.

experiment with different measures in the first year of AQUAINT (2002). Promising measures will then be put to a broader test by being incorporated into later TREC tracks.

References

[1] C.L.A. Clarke, G.V. Cormack, T.R. Lynam, C.M. Li, and G.L. McLearn. Web reinforced question answering (MultText experiments for TREC 2001). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2002.

[2] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[3] Sanda Harabagiu, John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. Issues, tasks, and program structures to roadmap research in question & answering (Q&A), October 2000. `http://www-nlpir.nist.gov/projects/duc/roadmapping.html`.

[4] 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.

[5] 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`.