Overview of the TREC 2007 Question Answering Track

Hoa Trang Dang[1], Diane Kelly[2], and Jimmy Lin[3]

[1]National Institute of Standards and Technology
Gaithersburg, MD 20899
hoa.dang@nist.gov

[2]University of North Carolina
Chapel Hill, NC 27599
dianek@email.unc.edu

[3]University of Maryland
College Park, MD 20742
jimmylin@umd.edu

Abstract

The TREC 2007 question answering (QA) track contained two tasks: the main task consisting of series of factoid, list, and "Other" questions organized around a set of targets, and the complex, interactive question answering (ciQA) task. The main task differed from previous years in that the document collection comprised blogs in addition to newswire documents, requiring systems to process diverse genres of unstructured text. The evaluation of factoid and list responses distinguished between answers that were globally correct (with respect to the document collection), and those that were only locally correct (with respect to the supporting document but not to the overall document collection). The ciQA task provided a framework for participants to investigate interaction in the context of complex information needs. Standing in for surrogate users, assessors interacted with QA systems live over the Web; this setup allowed participants to experiment with more complex interfaces but also revealed limitations in the ciQA design for evaluation of interactive systems.

1 Introduction

The goal of the TREC question answering (QA) track is to foster research on systems that directly return answers, rather than documents containing answers, in response to a natural language question. Since its inception in TREC-8 (1999), the track has steadily expanded both the type and difficulty of the questions asked. 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 contained list and definition questions in addition to factoid questions (Voorhees, 2004). A list question asks for different answer instances that satisfy the information need, such as *List the names of chewing gums*. Answering such questions requires a system to assemble a response 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.

Since TREC 2004 (Voorhees, 2005a), factoid and list questions have been grouped into different series, where each series is associated with a target and the questions in the series ask for some information about the target. In addition, the final question in each series is an explicit "Other" question, which is 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 series format supports the evaluation of different types of questions (factoid, list and Other) while providing an abstraction of a real user session with a QA system.

In TREC 2004, the target for a series could be a person, organization, or thing. Events were added as possible targets in TREC 2005, requiring that answers must be temporally correct with respect to the time-frame defined by the series. In TREC 2006, that requirement for sensitivity to temporal dependencies was made explicit in the distinction between locally and globally correct answers, so that answers for questions phrased in the present tense must not only be supported by the supporting document (locally correct), but must also be the most up-to-date answer in the document collection (globally correct).

The main task in the TREC 2007 QA track repeated the question series format, but with a significant change in the genre of the document collection. Instead of just newswire, the document collection contained both newswire and blogs. Mining blogs for answers introduced significant new challenges in at least two aspects that are very important for real-world QA systems: 1) being able to handle language that is not well-formed, and 2) dealing with discourse structures that are more informal and less reliable than newswire. Based on its successful application in TREC 2006 (Dang and Lin, 2007), the nugget pyramid evaluation method became the official evaluation method for the Other questions in TREC 2007.

In addition to the main task, the TREC 2007 QA track repeated the complex, interactive QA (ciQA) task of TREC 2006. At the TREC 2006 workshop, participants indicated that they wanted to have longer, more complex interactions in the ciQA task rather than short interactions via cached interaction forms. Participants proposed trying "live interactions" for 2007. Under this setup, the interactive QA system was located at a URL (Uniform Resource Locator) on the participant's machine, and NIST assessors simply navigated to the URL. The advantage was that participants were able to explore more complex interactions and interfaces. However, this setup placed the burden on participants to have their systems accessible during the entire interaction period and to record all desired data during the interaction.

The remainder of this paper describes each of the two tasks in the TREC 2007 QA track in more detail. Section 2 describes the questions, evaluation methods, and results for the main task, while Section 3 discusses the ciQA task.

2 Main Task

The scenario for the main task in the TREC 2007 QA track was that an adult, native speaker of English is looking for information about a target of interest. The target could be a person, organization, thing, or event. The user was assumed to be an "average" reader of U.S. newspapers. Serving as surrogate users, NIST assessors developed the questions and judged the system responses.

The main task required systems to provide answers to a series of related questions. A question series, which focused on a target, consisted of several factoid questions, one or two list questions, and exactly one Other question. The order of questions in the series and the type of each question (factoid, list, or Other) were all explicitly encoded in the test set. Example series are shown in Figure 1. The final test set contained 70 series; the targets of these series are given in Table 1. Of the 70 targets, 19 were PERSONs, 17 were

219	Target: Iraqi defector Curveball		
	219.1	FACTOID	What year did Curveball defect?
	219.2	FACTOID	What was Curveball's profession?
	219.3	FACTOID	What is Curveball's real name?
	219.4	FACTOID	Which intelligence service employed Curveball?
	219.5	LIST	Which US government officials accepted his claims regarding Iraqi weapons labs?
	219.6	FACTOID	Where does Curveball now live?
	219.7	OTHER	
254	Target: House of Chanel		
	254.1	FACTOID	Who founded the House of Chanel?
	254.2	FACTOID	In what year was the company founded?
	254.3	FACTOID	Who is the president of the House of Chanel?
	254.4	FACTOID	Who took over the House of Chanel in 1983?
	254.5	LIST	What women have worn Chanel clothing to award ceremonies?
	254.6	LIST	What museums have displayed Chanel clothing?
	254.7	FACTOID	What Chanel creation is the top-selling fragrance in the world?
	254.8	OTHER	
269	Target: Pakistan earthquakes of October 2005		
	269.1	FACTOID	On what date did this earthquake strike?
	269.2	LIST	What countries were affected by this earthquake?
	269.3	FACTOID	What was the final death toll from this earthquake?
	269.4	FACTOID	What was the strength of this earthquake?
	269.5	FACTOID	Where was the epicenter (latitude and longitude)?
	269.6	LIST	What countries supplied aid?
	269.7	OTHER	

Figure 1: Sample question series from the test set. Series 219 has a PERSON as the target, series 254 has an ORGANIZATION as the target, and series 269 has an EVENT as the target.

ORGANIZATIONs, 15 were EVENTs, and 19 were THINGs. The series contained a total of 360 factoid questions, 85 list questions, and 70 Other questions. Each series contained 6–10 questions (counting the Other question), with most series containing 7 questions.

Answers were to be drawn from a document collection comprising the Blog06 corpus (Macdonald and Ounis, 2006) and the AQUAINT-2 Corpus of English News Text. The AQUAINT-2 collection contains approximately 2.5 GB of text (about 907K documents) spanning the time period of October 2004 - March 2006; articles are in English and come from a variety of sources including Agence France Presse, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times-Washington Post News Service, New York Times, and The Associated Press. Blog06 documents were collected by polling 100,649 RSS and Atom feeds over an 11 week period (December 6, 2005 - February 21, 2006). A blog document is defined to be a blog post plus its follow-up comments (a permalink). As a convenience for track participants, NIST made available document rankings of the top 1000 documents per target for each of two corpora, as produced using the PRISE document retrieval system, with the target as the query.

Participants were allowed two weeks to download the test data and submit their results. All processing of the questions was required to be strictly automatic. Systems were required to process series independently

216 Paul Krugman	251 Lyme disease
217 Jay-Z	252 American Girl dolls
218 impressionist Darrell Hammond	253 Kurt Weill
219 Iraqi defector Curveball	254 House of Chanel
220 International Management Group (IMG)	255 British American Tobacco (BAT)
221 U.S. Mint	256 Buffalo Soldiers
222 3M	257 2005 DARPA Grand Challenge
223 Merrill Lynch & Co.	258 2005 presidential election in Egypt
224 WWE	259 2005 World Snooker Championships
225 Sago Mine disaster	260 Teenage Mutant Ninja Turtles (TMNT)
226 Harriet Miers withdraws nomination to Supreme Court	261 marsupials
227 Robert Blake criminal trial	262 kumquat
228 March Madness 2006	263 Ayn Rand
229 first partial face transplant	264 Alan Greenspan
230 AMT	265 Mahmud (or Mahmood, Mahmoud) Ahmadinejad
231 USS Abraham Lincoln	266 Rafik Hariri, former Lebanese Prime Minister
232 Dulles Airport	267 FISA Court
233 comic strip Blondie	268 Israel evacuation of the Gaza Strip
234 Irving Berlin	269 Pakistan earthquakes of October 2005
235 Susan Butcher	270 The Mars rovers, Spirit and Opportunity
236 Boston Pops	271 Jon Bon Jovi
237 Cunard Cruise Lines	272 Barack Obama
238 2004 Baseball World Series	273 Rush Limbaugh
239 game show Jeopardy	274 Exxon Mobile Corp
240 Harry Potter and the Goblet of Fire	275 Dixie Chicks
241 Jasper Fforde	276 B-17 bomber
242 Guinness Brewery	277 Boeing 777 aircraft
243 2005 London terror bombing attacks	278 St. Peter's Basilica
244 Rubik's Cube Competitions	279 Australian wine
245 hybrid cars	280 Angkor Wat temples
246 Michael Brown	281 Joseph Steffen
247 Ella Fitzgerald	282 Orhan Pamuk
248 CSPI	283 Habitat for Humanity
249 Fulbright Program	284 CAFTA approval by U.S. Congress
250 publication of Danish cartoons of Mohammed	285 Yeti

Table 1: Targets of the 70 question series.

from one another, and 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 the same series, but could not "look ahead" and use later questions to help answer earlier questions. Thus, question series can be viewed as an abstraction of an information-seeking dialogue between the user and the system; cf. (Kato et al., 2004). In total, 51 runs from 21 participants were submitted to the main task.

The evaluation of a single run can be decomposed into component evaluations for each of the question types and a final per-series score. Each of the three question types has its own response format and evaluation method. The individual component evaluations in 2007 were identical to those used in the TREC 2006 QA track, except that the official scores for Other questions were computed using multiple assessors' judgments of the importance of information nuggets, and assessors were not restricted in the criteria they could use in distinguishing between locally correct and globally correct answers for factoid and list questions. An aggregate score was computed for each series in a run using a simple average of the component scores of questions in that series, and the final score for the run was computed as the average of its per-series scores.

2.1 Factoid questions

The system response to 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 made the final determination. Each response was assigned exactly one of the following five judgments:

incorrect: 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;

locally correct: the answer string consists of exactly a correct answer that is supported by the document returned, but the document collection contains a contradictory answer that the assessor believes is better;

globally correct: the answer string consists of exactly the correct answer, that answer is supported by the document returned, and the document collection does not contain a contradictory answer that the assessor believes is better.

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 if the question asks about "the Taj Mahal"). Questions also had to be interpreted in the time frame implied by the question series. For example, if the target was the event "France wins World Cup in soccer" and the question was *Who was the coach of the French team?* then the correct answer must be "Aime Jacquet", the name of the coach of the French team in 1998 when France won the World Cup, and not just the name of any past or current coach of the French

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
LymbaPA07	Lymba Corporation	0.706	0.000	0.000
LCCFerret	Language Computer Corporation	0.494	0.000	0.000
lsv2007c	Saarland University	0.289	–	0.000
UofL	University of Lethbridge	0.258	0.052	0.500
QASCU1	Concordia University	0.256	0.000	0.000
FDUQAT16A	Fudan University	0.236	0.053	0.312
pronto07run3	Universita di Roma "La Sapienza"	0.222	0.000	0.000
ILQUA1	State University of New York (SUNY) at Albany	0.222	0.000	0.000
Ephyra3	Carnegie Mellon University and Universitaet Karlsruhe	0.208	0.048	0.062
QUANTA	Tsinghua University (State Key Lab)	0.206	0.091	0.062

Table 2: Evaluation scores for the factoid component. Scores are shown for the best run from the top 10 groups.

team. NIL responses were correct only if there was no known answer to the question in the collection. NIL was correct for 16 of the 360 factoid questions in the test set. For 26 questions, no system returned the correct answer, although those questions did have a correct answer found by the assessors.

It may be the case (especially with the inclusion of blogs) that different documents support contradictory answers as being correct. An exact answer-string that is supported in its associated document is assumed to be globally correct unless there is a *better, contradictory* answer supported elsewhere in the document collection. The assessor was allowed to use any number of criteria in determining that one answer was better than another, including recency of the supporting document, the amount of support provided by each supporting document, the number of distinct sources that support the answer as being correct, and the credibility or authoritativeness of the source. The assessor marked as globally correct one or more of the most credible of the known locally correct answers. "Global" correctness was defined with respect to the document collection, and not necessarily with respect to the real world.

The main evaluation metric for the factoid component was *accuracy*, the fraction of questions judged to be globally correct. Table 2 shows the most accurate run for the factoid component for each of the top 10 groups. 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; NIL recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct in the entire test set (16). If NIL was never returned, NIL precision is undefined and NIL recall is zero.

2.2 List questions

A list question asks for different instances of a particular type. The correct answer for a list question is the set of all such distinct instances in the document collection. A system's response to a list question consists of an unordered set of [*doc-id*, *answer-string*] pairs such that each *answer-string* represents a correct answer instance.

Each instance was evaluated in the same manner as the factoid questions, i.e., assigned one of the following judgments: incorrect, not supported, not exact, locally correct, and globally correct. Instances that were judged to be globally correct were then manually grouped into equivalence classes, where each

Run Tag	Submitter	F
LymbaPA07	Lymba Corporation	0.479
LCCFerret	Language Computer Corporation	0.324
ILQUA1	State University of New York (SUNY) at Albany	0.147
QASCU3	Concordia University	0.145
Ephyra3	Carnegie Mellon University and Universitaet Karlsruhe	0.144
UofL	University of Lethbridge	0.132
FDUQAT16B	Fudan University	0.131
IITDIBM2007T	Indian Institute Of Technology, Delhi	0.125
FDUQAT16A	Fudan University	0.107
pronto07run3	Universita di Roma "La Sapienza"	0.103

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

equivalence class was considered a distinct answer. Thus, systems were not rewarded (and were in fact penalized) for returning equivalent answers multiple times.

The final set of known globally correct answers for a list question was compiled from the union of distinct globally correct answers across all runs plus additional distinct answers the assessor found during question development. For the 85 list questions in the test set, the median number of known distinct globally correct answers per question was 7, with a minimum of 2 and a maximum of 64. A system's response to a list question was scored using instance precision (IP) and instance recall (IR) based on the complete list of known distinct globally correct answers. Let S be the number of such answers, D be the number of distinct globally correct answers returned by the system, and N be the total number of instances returned by the system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined to produce an F-score 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 85 questions. Table 3 gives the average F-score of the run with the best list component score for each of the top 10 groups.

2.3 Other questions

The Other questions were evaluated using the methodology originally developed for the TREC 2003 definition questions. A system's response for an Other question consisted of an unordered set of [*doc-id*, *answer-string*] pairs. The answer strings were presumed to contain interesting "nuggets" about 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 more difficult than answering TREC 2003 definition questions.

Judging the quality of the systems' responses was performed in two steps. In the first step, all of the answer strings from all of the systems were presented to an assessor in a single list. Using all the answer strings and searches performed during question development, the assessor created a list of information nuggets about the target. An information nugget in the context of an Other question is defined as an atomic

Run Tag	Submitter	F($\beta = 3$)
FDUQAT16B	Fudan University	0.329
lsv2007c	Saarland University	0.299
QASCU2	Concordia University	0.281
LymbaPA07	Lymba Corporation	0.281
LCCFerret	Language Computer Corporation	0.261
ILQUA1	State University of New York (SUNY) at Albany	0.242
csail3	Massachusetts Institute of Technology (MIT)	0.235
uams07main	University of Amsterdam	0.209
IITDIBM2007S	Indian Institute Of Technology, Delhi	0.209
QUANTA	Tsinghua University (State Key Lab)	0.194

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

piece of information about the target that is interesting (in the assessor's opinion) and is not part of an earlier question in the series or an answer to an earlier question in the series. An information nugget is considered atomic if the assessor could make a binary decision as to whether the nugget appears in a response. Once the nugget list was created for a target, the assessor decided which were vital, meaning that the information must be returned for a response to be good. Non-vital ("okay") nuggets acted as "don't care" conditions in that the assessor believed the information in the nugget to be interesting enough that returning the information was acceptable in, but not necessary for, a good response.

In the second step of the evaluation process, the assessor went through each system's output in turn and marked which nuggets appeared in the response. An answer string contained a nugget if there was a *conceptual* match between the answer string and the nugget; that is, the match was independent of the particular wording used in either the nugget or the system output. A nugget match was marked at most once per response—if the system output 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.

To address some of the weaknesses of using vital/okay judgments from a single assessor (Hildebrandt et al., 2004), Lin and Demner-Fushman (2006) proposed an extension called "nugget pyramids", in which multiple assessors provide judgments of whether a nugget was vital or simply okay. Dang and Lin (2007) subsequently verified the efficacy of this method, and thus NIST adopted the pyramid extension for computing F-scores for Other responses. Nine different sets of vital/okay judgments were solicited from eight unique assessors (the primary assessor who originally created the nuggets later assigned vital/okay labels again). Each assessor was given all the questions for the series, as well as the nuggets created by the primary assessor. Using the pyramid procedure, a weight was assigned to each nugget based on the number of assessors who marked it as vital.

Given the nugget list and the set of nuggets matched in a system's response, nugget recall was computed as the ratio of the sum of weights of matched nuggets to the sum of weights of all nuggets in the list. Nugget precision was much more difficult to compute since there was no effective way of enumerating all the concepts contained in a particular answer string. Instead, a measure based on length (in non-whitespace characters) was used as an approximation to nugget precision. The length-based measure granted an allowance of 100 characters for each nugget matched. If the total system output was less than this number of

characters, the value of nugget precision was 1.0. Otherwise, the measure's value decreased as the length increased according to the following formula:

$$1 - \frac{length - allowance}{length}.$$

The final score for an Other question was an F-score, with nugget recall weighted more heavily than nugget precision:

$$F(\beta) = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}.$$

The score for the Other questions component was the average F-score (β=3) over the 70 Other questions. Table 4 gives the F-score for the best scoring Other question component for each of the top 10 groups.

2.4 Per-series Combined Scores

The three component scores measure a system's ability to process each type of question, but may not reflect the system's overall usefulness to a user. Since each individual series is an abstraction of a single user's interaction with the system, taking the individual series as the basic unit of evaluation should provide a more accurate representation of the effectiveness of the system from an individual user's perspective. Since each series is a mixture of different question types, we can compute a weighted average of the scores of the three question types on a per-series basis, and take the average of the per-series weighted scores as the final score for the run (Voorhees, 2005b). In 2007, the weighted score for an individual series was computed as:

$$\text{WeightedScore} = \frac{1}{3} \times \text{Factoid} + \frac{1}{3} \times \text{List} + \frac{1}{3} \times \text{Other}.$$

To compute the weighted score for an individual series, only the scores for questions belonging to that series were included in the computation. Since each of the component scores ranges between 0 and 1, the weighted score is also in that range. The final per-series score of each run is simply the average of individual per-series weighted scores.

We fit a two-way analysis of variance (ANOVA) model with the target type and the best run from each group as factors, and the per-series score as the dependent variable; we found significant differences between runs (p essentially equal to 0). To determine which runs were significantly different from each other, we

RunID	Submitter	Score					
LymbaPA07	Lymba Corporation	0.4839	A				
LCCFerret	Language Computer Corporation	0.3575		B			
FDUQAT16B	Fudan University	0.2310			C		
lsv2007c	Saarland University	0.2296			C		
QASCU1	Concordia University	0.2216			C	D	
ILQUA1	State University of New York (SUNY) at Albany	0.2023			C	D	E
Ephyra1	Carnegie Mellon University and Universitaet Karlsruhe	0.1804			C	D	E	F
IITDIBM2007T	Indian Institute Of Technology, Delhi	0.1735				D	E	F
QUANTA	Tsinghua University (State Key Lab)	0.1592					E	F
csail3	Massachusetts Institute of Technology (MIT)	0.1415						F

Table 5: Multiple comparison of the best run from the top 10 groups, based on ANOVA of per-series score.

performed a multiple comparison using Tukey's honestly significant difference criterion and controlling for the experiment-wise Type I error so that the probability of declaring a difference between two runs to be significant, when it is actually not, is at most 5%. Table 5 shows the results of the multiple comparison for the 10 groups with the highest final per-series score; runs sharing a common letter are not significantly different.

2.5 Discussion

Despite the inclusion of the blog corpus, which was expected to make the QA task more difficult, the best component scores in the main task were higher in 2007, after having generally declined each year since TREC 2004.

For each series, an attempt had been made during question development to include at least one question whose answer was found in the Blog06 corpus but not in the AQUAINT-2 corpus. This could be the answer to a factoid question, one of the items answering a list question, or (in rare cases) a nugget for the Other question. NIST assessors varied in their ability to locate blog-specific information that was suitable for the series. In some cases, the assessor could not find an answer in the AQUAINT-2 corpus during topic development, but the answer was later found in AQUAINT-2 during the assessment of system responses. In the end, only 15.0% (54/360) of the factoid questions had an answer that could be found only in the Blog06 corpus; 24.8% (235/946) of the distinct items answering a list question could be found only in the Blog06 corpus; and at most 6.1% (45/735) of the distinct nuggets answering an Other question could be found only in the Blog06 corpus.

The positive contribution of answers from blog documents to the various component scores was likely depressed due to the nature of the questions asked. Because factoid and list questions generally requested factual information, it is not surprising that most of their answers would come from newswire rather than blogs. In addition, assessors tend to place more credibility on newswire documents than blog posts, so if a blog answer contradicted a newswire answer, the newswire answer would be judged as the globally correct one, and the blog answer would at best be judged as locally correct; the effect would be more pronounced for factoid questions (which generally have only one globally correct answer) than for list questions (which allow multiple answers). Finally, assessors were most interested in factual information about their targets, and consequently found very little interesting Other information nuggets in the blog documents.

3 Complex Interactive QA (ciQA) Task

In TREC 2007, the goals of the complex, interactive question answering (ciQA) task remained unchanged from the previous year—to push the frontiers of question answering in two directions:

- A move away from "factoid" questions towards more complex information needs that exist within richer user contexts.
- A move away from the one-shot interaction model implicit in previous evaluations towards a model based on interactions with users.

The ciQA task in TREC 2007 represented the second iteration of the evaluation, which started in 2006. The TREC 2006 ciQA task (Dang et al., 2007), in turn, descended from the TREC 2005 HARD track, which focused on single-iteration clarification dialogues (Allan, 2006). However, there were substantial changes in the evaluation methodology: in TREC 2006, participants "encapsulated" their interactions in HTML forms

that were sent to NIST. This year, the task moved to completely "live" systems where assessors accessed individual QA systems, running at the participants' sites, over the Web.

3.1 Task Definition

3.1.1 Corpus

The ciQA task used the newswire portion of the corpus used by the main QA task (excluding the blog data). Participants were provided with the top 100 documents as retrieved by the PRISE system, using the question template verbatim as the query.

3.2 Complex "Relationship" Questions

The complex information needs explored by ciQA remained unchanged from last year; they represent extensions and refinements of so-called "relationship" questions piloted in TREC 2005 (Voorhees and Dang, 2006).

The concept of a "relationship" is defined as the ability of one entity to influence another, including both the means to influence and the motivation for doing so. Eight "spheres of influence" were noted in a previous pilot study funded by the AQUAINT research program: financial, movement of goods, family ties, communication pathways, organizational ties, co-location, common interests, and temporal. Evidence for both the existence or absence of ties is relevant. The particular relationships of interest naturally depend on the context.

A relationship question in the ciQA task, referred to as a topic, is composed of two parts. Consider an example:

> **Template:** What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?
> **Narrative:** The analyst would like to know of efforts to curtail the transport of drugs from Mexico to the U.S. Specifically, the analyst would like to know of the success of the efforts by local or international authorities.

The question template is a stylized information need that has a fixed structure and free slots whose instantiation varies across different topics. The narrative is free-form natural language text that elaborates on the information need, providing, for example, user context, a more articulated statement of interest, focus on particular topical aspects, etc.

The ciQA task employed the following templates, which were the same as those used in TREC 2006:

- What evidence is there for transport of [goods] from [entity] to [entity]?
- What [relationship] exist between [entity] and [entity]? (where [relationship] is an element of {"financial relationships", "organizational ties", "familial ties", "common interests"})
- What influence/effect do(es) [entity] have on/in [entity]?
- What is the position of [entity] with respect to [issue]?
- Is there evidence to support the involvement of [entity] in [event/entity]?

Thirty topics were developed for this year's task, but they were not distributed evenly across the five templates. In addition, one "throw-away" topic was included for training purposes.

Assessor	Topics
1	57, 69, 83
2	56, 63, 64, 74
3	65, 75, 76, 82
4	61, 68, 80, 85
5	58, 66, 70, 77
6	60, 72, 79, 84
7	62, 73, 81
8	59, 67, 71, 78

Table 6: Mapping between each NIST assessor and the topics they were responsible for.

3.2.1 Interaction Design

The purpose of the interactive aspect of ciQA was to provide a framework for participants to investigate interaction in the QA context. Unlike in TREC 2006, participants were able to deploy full-fledged Web-based QA systems with which the assessors engaged for five minutes per topic. There were no restrictions on the nature of the interaction or the system, except that it had to be accessible from a Web browser. Anything ranging from mixed-initiative dialogues to graphical interfaces was allowed.

To initiate interactions, assessors were directed to URLs provided by the participants. Assessors interacted with each system for five minutes per topic. The interaction length included time spent loading/rendering the page, as well as any delay caused by network traffic. It was the participant's responsibility to ensure that the QA system was Web-accessible during the period of time the assessors were scheduled to interact with submitted systems (a three-day period). If assessors were unable to access the participant's QA system, they skipped that interaction and did not return to it later.

The "throw-away" topic described earlier was used to familiarize assessors with systems before they completed actual test topics. Like other topics, the training period lasted five minutes, and could consist of anything that the participants wanted (e.g., a structured tutorial to introduce system features).

The interactions were completed at NIST using a Redhat Enterprise Linux 4 workstation with a 20-inch LCD monitor with 1600×1200 resolution and true color display (millions of colors), and a Firefox Web browser, v2.0.0.6. In addition, Flash, Acroread, and RealPlayer were enabled.

3.2.2 Experimental Protocol

The basic setup for the task was as follows: Participants first submitted initial runs and URL files to NIST. The URL files provided pointers to the participants' Web-based QA system (one for each topic). Included in the URL files were also pointers to screenshots of the interface, supplied by the participants for archival purposes. NIST assessors interacted with the Web-based QA systems during a three-day period. Results of those interactions were available immediately to participants, since they hosted their own systems. It was each participant's own responsibility to instrument their system to collect whatever data was necessary; NIST did not keep track of the interactions. Eight assessors participated in the task. Most assessors completed four topics; the mapping between assessors and topics is shown in Table 6.

Approximately two weeks following the interaction period, participants submitted final runs based on the results of the interactions to NIST. Assessors evaluated both initial and final runs.

Each participant was allowed to submit a maximum of 2 initial runs, 2 URL files, and 2 final runs. Manual runs were accepted, but had to be marked as such in the run submission interface. The interactive part of ciQA was optional; groups that did not wish to participate in the interactive aspect were asked to simply not submit URL files (however, every team engaged in the interactions). For each final run, participants were asked to supply the run tag of its corresponding initial run—this provided pairs of corresponding initial–final runs that isolated the effects of the interaction.

3.2.3 Evaluation Methodology

System responses were evaluated using the "nugget pyramid" extension of the nugget-based methodology used in previous TREC QA tasks (Lin and Demner-Fushman, 2006). Nine different sets of vital/okay judgments were solicited from eight unique assessors (the assessor who originally created the nuggets later assigned vital/okay labels again). Additional analyses included recall by length plots, as described in (Lin, 2007). A recall plot quantifies pyramid recall as a function of response length, which provides a rough model of how quickly a user can learn about the topic by reading system responses in sequential order. For more information on how this is computed, please refer to (Dang et al., 2007).

In addition to runs submitted by participants, we separately prepared a sentence retrieval baseline, similar to the one prepared last year. This provided a task-wide baseline to serve as a point of comparison. For each topic, the verbatim question template was used as a query to Lucene, which returned the top 20 documents. These documents were then tokenized into individual sentences. Sentences that contained at least one non-stopword from the question were retained and returned as the baseline run (up to a quota of 5,000 characters). Sentence order within each document and across the ranked list was preserved. The interaction associated with this run asked the assessor for relevance judgments on each of the sentences. Three options were given: "relevant", "not relevant", and "no opinion". The final run was prepared by simply removing those sentences judged not relevant—this had the effect of pulling more sentences from documents lower in the ranked list.

After assessors finished their interactions, they completed an online exit questionnaire which asked them to evaluate the various interactions. Assessors evaluated interactions according to several dimensions related to ease of use, usefulness, and effectiveness using 5-point scales. Assessors were also able to provide qualitative feedback about each interaction. Small screenshots of each system were displayed to remind assessors of each interaction. The order of these screenshots (and the order in which assessors evaluated each interaction) was random. A portion of the exit questionnaire, displaying the ciQA baseline interaction (described above), can be seen in Figure 2. At the end of the exit questionnaire, assessors were presented with four open-ended questions that asked them about their overall experiences. These questions were:

1. Of all interactions, which was your favorite and why?
2. What annoyed you about the interactions and why?
3. How different did you find the various interactions from one another and why?
4. Anything else?

3.3 Results

The ciQA task drew participation from seven groups. NIST received twelve initial runs and twelve final runs. A total of fourteen URL files were submitted. For the purposes of evaluation, the sentence retrieval baseline was treated like any other submission. In total, there were twelve initial–final pairs (and the sentence retrieval baseline).

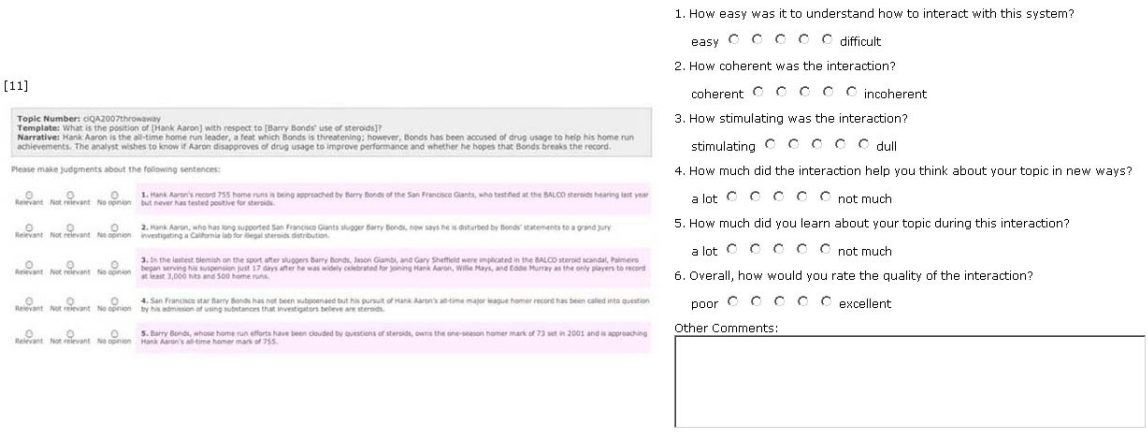

Figure 2: Portion of exit questionnaire for the baseline interaction. On the left the assessor sees a screenshot of the system (not meant to be readable, but simply as a reminder); questions are shown on the right.

3.3.1 System Effectiveness

The pyramid F-scores of the initial–final run pairs are shown in Table 7. By comparing the score of the corresponding runs, we can quantify the effect of the interaction on system performance. The scatter plot in Figure 3 presents a different view of the results—the initial score is plotted on the x axis, and the final score is plotted on the y axis. Points above the reference line $y = x$ represent cases where interaction improved performance.

We note two striking observations: First, unlike last year (Dang et al., 2007), most systems outperformed the baseline.[1] This is encouraging for the development of the field as a whole. Second, many interactions were detrimental, i.e., the pyramid F-score of the final run was higher than that of the initial run. Once again, this was different from last year, where interactions generally yielded small gains. We believe this effect to be caused by a combination of factors: problems with the task setup (more below); technical issues in deploying live Web-based QA systems; and the broadening of the design space that truly allows for effective and non-effective interactions.

3.3.2 Assessors Feedback about Interactions

The majority of interactions submitted by participants involved eliciting some type of relevance feedback from assessors. Items presented to assessors for feedback varied and included terms, sentences, articles from Wikipedia, and entire answer sets. A couple of systems asked assessors to interactively construct answers to their questions using sentences and documents. One interaction technique asked assessors to respond to open-ended questions modeled after a reference exchange, while another technique asked assessors to indicate their preferences for answer types. While most of the interactions went smoothly, at least two sites had network difficulties which impacted the interactions assessors had with their systems.

Figure 4 presents the mean quantitative ratings provided by subjects for three questions:

1. How easy was it to understand how to interact with this system?

[1]There were indexing issues with UNC's initial submission, which readily explains one of the two below-baseline performers. The other run, from the University of Maryland, experimented with *abstractive* techniques for question answering—i.e., the runs contained responses that were not found in any source document.

Organization	Type	Run tags		Pyramid F-Score	
		Initial	Final	Initial	Final
Michigan State U.	automatic	MSUciQAiHeu	MSUciQAfCol	0.359	0.361
Michigan State U.	automatic	MSUciQAiHeu	MSUciQAfInt	0.359	0.370
RMIT	automatic	rmitrun2	rmitrun5	0.361	0.343
RMIT	automatic	rmitrun2	rmitrun6	0.361	0.333
U. Mass	automatic	UMassBaseAut	UMassIntA	0.318	0.347
U. Mass	manual	UMassBaseAut	UMassIntM	0.318	0.503
U. Maryland	automatic	UMD07iMASCa	UMD07iMASCb	0.182	0.156
U. Maryland	automatic	UMD07MMRa	UMD07MMRb	0.333	0.334
U. NC and Yahoo!	automatic	UNCYABL30	UNCYAEX2	0.062	0.374
U. Strathclyde	manual	sicka	sicka2	0.410	0.394
U. Waterloo	manual	UWinitWIKI	UWfinalMAN	0.388	0.386
U. Waterloo	automatic	UWinitWIKI	UWfinalWIKI	0.388	0.380
baseline	automatic	baseA	baseB	0.327	0.327

Table 7: Performance of the twelve initial–final run pairs submitted to the TREC 2007 ciQA task. The sentence retrieval baseline is provided as a reference.

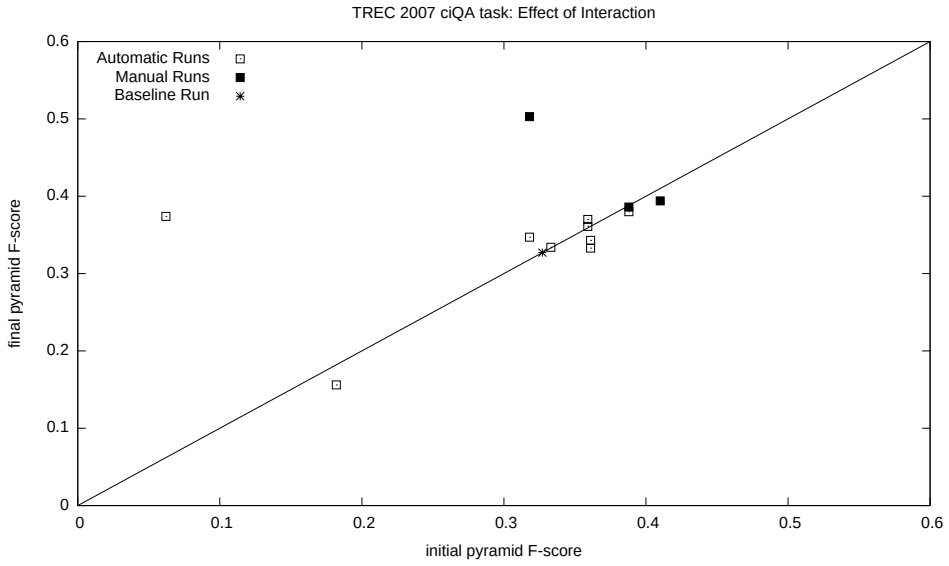

Figure 3: Scatter plot showing initial and final pyramid F-scores for each run pair submitted to the TREC 2007 ciQA task. Points above the line $y = x$ represent interactions that increased answer quality. Note that most systems outperformed the sentence retrieval baseline

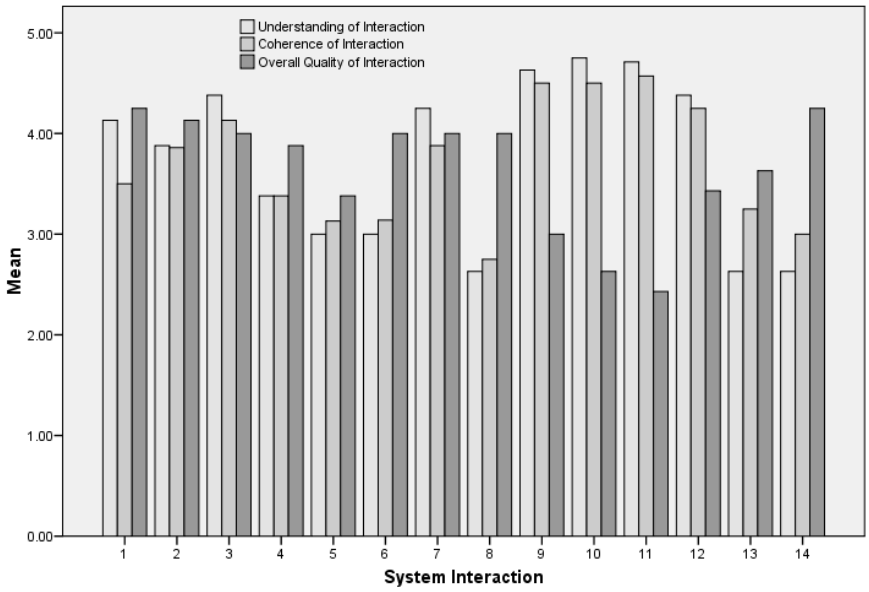

Figure 4: Mean assessors' ratings for each interaction along three dimensions: comprehensibility of interaction, coherence of interaction, and overall quality.

2. How coherent was this interaction?

3. Overall, how would you rate the quality of the interaction?

In all cases, higher scores are more positive. It is important to note that this is a small, unrepresentative, and unusual sample, so these results should be viewed cautiously. These results are by no means definitive and/or generalizable beyond this evaluation context.

Interactions that were rated the lowest with respect to understanding (interactions 8, 13 and 14) and coherence (interactions 8 and 14) had information-dense interfaces and often required multiple steps (one of these interactions was answer construction). Interactions rated most positively for these two attributes were traditional relevance feedback interfaces. Interestingly, understanding and coherence were positively correlated with one another ($r = 0.949$, $p < 0.01$), but were negatively correlated with assessors' overall quality ratings ($r = -0.533$, $p < 0.05$ and $r = -0.674, p < 0.01$, respectively). The interaction that received the lowest quality assessment scored fairly high on understanding and coherence. This interaction was the ciQA baseline interaction, which elicited sentence-level relevance feedback. The interaction that received some of the lowest assessments for understanding and coherence received one of the highest overall quality scores (interaction 14). This interaction consisted of building answers and may have received a higher quality score because of its novelty. It also engaged assessors in the most interaction which may be why scores on these three measures differ.

The qualitative feedback from the final set of questions asking assessors about their entire experiences showed that assessors preferred the traditional relevance feedback interactions, felt considerable time pressure, and did not like the complicated interactions. One assessor indicated a preference for one of the answer construction interactions, while another did not like this interaction. At least two assessors were puzzled about the use of Wikipedia and were displeased with this interaction.

Data from the exit questionnaire should be viewed cautiously for several reasons. Some interactions

were less than perfect because of network problems, so assessors' evaluations, in part, reflect this. Assessors' comments indicated that they felt huge time pressures, which may by why such an overwhelming preference was indicated for simple, easily understood and executed interactions such as those that employed relevance feedback.

One of the most interesting results of this evaluation was that it revealed several limitations of this style of evaluation in the context of TREC. Many of the limitations stem from the fact that assessors already know a great deal about their topics before they engage in interactions. The approach in TREC has traditionally been to have the same person develop the topic and assess its answers, since the assessor is supposed to act as a surrogate user with his/her own particular information needs. However, in developing the topic for ciQA, this "user" researches the topic (to make sure that it is a suitable topic for the particular document collection) and consequently knows more about the topic than a naive user issuing the query.[2] NIST assessors are unusual "users" and it is unrealistic to expect them to assume dual roles as assessors (during topic development and answer evaluation) and naive users (during the interactions).

Helping users learn more about their topics and helping systems learn more about users are central goals of interactive systems. The exit questionnaire reveals that interactive techniques for addressing these goals cannot be evaluated using the ciQA experimental framework. Additionally, not all ciQA participants understood that assessors already knew the answers to the questions they were asking so there may also have been a mismatch between participants' and assessors' expectations of the interactions.

4 Future of the QA Track

TREC 2007 revealed limitations in the ciQA design for evaluating interactive systems. These limitations could not be reconciled within the NIST evaluation framework, and hence it was decided not to attempt another interactive QA task in 2008.

The primary goal of the TREC 2007 main task (and what distinguished it from previous TREC QA tasks) was the introduction of blog text to encourage research in NLP techniques that would handle ill-formed language and discourse structures that are more informal and less reliable than newswire. Questions were asked over a combined newswire (AQUAINT-2) and blog (Blog06) corpus, rather than only a blog corpus, in order to ease participants' transition from newswire. However, because most of the TREC 2007 questions requested factual information, they did not specifically test systems' ability to process blog text, as answers still came predominantly from the AQUAINT-2 corpus.

This mismatch between the corpus and the information need expressed in the questions naturally suggests that in order to move away from traditional newswire towards blogs, the QA task should be changed so that the questions are more targeted towards characteristics that are particular to blogs. Because blogs naturally contain a large amount of opinions, it was decided that the QA task for 2008 should focus on questions that ask about people's *opinions*. Questions would still be grouped into series focused by a particular target (person, organization, etc.), but there would be no factoid questions.[3] Rather, each series would comprise *rigid* list questions (e.g., "What people have good opinions of Sean Hannity?") which would be evaluated in the same manner as TREC 2007 list questions, and *squishy* list questions (e.g., "What reasons do people give for liking Sean Hannity?") which would be evaluated with the nugget pyramid method used for TREC 2007 Other questions.

[2]Results of questions 3, 4, and 5 from the exit questionnaire, which asked assessors to indicate how much they learned about their topics through the interaction (see Figure 2 for specific questions) are not presented because some assessors indicated that these values were low because they already knew about their topics.

[3]It was pointed out that asking factoid type questions about opinions seemed inappropriate, and after nine years of factoid questions (starting in TREC 1999), it was time to retire factoids from the QA track in any case.

References

James Allan. 2006. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

Hoa Dang and Jimmy Lin. 2007. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 768–775, Prague, Czech Republic.

Hoa Trang Dang, Jimmy Lin, and Diane Kelly. 2007. Overview of the TREC 2006 Question Answering Track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, Maryland.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*, pages 49–56.

Tsuneaki Kato, Jun'ichi Fukumoto, Fumito Masui, and Noriko Kando. 2004. 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.

Jimmy Lin and Dina Demner-Fushman. 2006. Will pyramids built of nuggets topple over? In *Proceedings of the 2006 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2006)*, pages 383–390, New York, New York.

Jimmy Lin. 2007. Is question answering better than information retrieval? A task-based evaluation framework for question series. In *Proceedings of the 2007 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2007)*, pages 212–219, Rochester, New York.

Craig Macdonald and Iadh Ounis. 2006. The TREC blog06 collection: Creating and analysing a blog test collection. Technical Report DCS Technical Report TR-2006-224, Department of Computing Science, University of Glasgow.

Ellen M. Voorhees and Hoa T. Dang. 2006. Overview of the TREC 2005 Question Answering Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

Ellen M. Voorhees. 2004. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68.

Ellen M. Voorhees. 2005a. Overview of the TREC 2004 Question Answering track. In *Proceedings of the Thirteenth Text REtreival Conference (TREC 2004)*, pages 52–62.

Ellen M. Voorhees. 2005b. Using question series to evaluate question answering system effectiveness. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 299–306.