Overview of the TREC 2006 Question Answering Track

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

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

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

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

Abstract

The TREC 2006 question answering (QA) track contained two tasks: the main task and the complex, interactive question answering (ciQA) task. As in 2005, the main task consisted of series of factoid, list, and "Other" questions organized around a set of targets; in contrast to previous years, 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). The ciQA task provided a framework for participants to investigate interaction in the context of complex information needs, and was a blend of the TREC 2005 QA relationship task and the TREC 2005 HARD track. Multiple assessors were used to judge the importance of information nuggets used to evaluate the responses to ciQA and "Other" questions, resulting in an evaluation that is more stable and discriminative than one that uses only a single assessor to judge nugget importance.

1 Introduction

The goal of the TREC question answering (QA) track is to foster research on systems that return answers themselves, rather than documents containing answers, in response to a 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.

In TREC 2004 (Voorhees, 2005a), factoid and list questions were grouped into different series, where each series was associated with a target (a person, organization, or thing) and the questions in the series asked for some information about the target. In addition, the final question in each series was an explicit "Other" question, which was to be interpreted as "Tell me other interesting things about this target I don't know enough to ask directly". This last question was roughly equivalent to the definition questions in the TREC 2003 task.

Since the beginning of the QA track, the document returned with an answer had been used to determine the time frame for a question. For example, "Ronald Reagan" was considered a correct answer for the question *Who is the President of the United States?* if that answer was supported by a document from 1987, even if more recent documents supported "George Bush" as the answer. Such guidelines were appropriate because questions were primarily phrased in the present tense without specifying an explicit time frame. However, in the TREC 2005 main task, events were added as a possible target for the question series, and it became clear that the time frame implied by the series could not be ignored when judging the correctness of answers. Event targets and temporally-constrained questions required that questions be interpreted in the temporal context explicit in the question or implicit in the series.

The main task for the TREC 2006 QA track was the same as the main task in 2005, except that the implicit time frame for questions phrased in the present tense was the date of the last document in the document collection, rather than the document returned with the answer. Thus, systems were required to give the most up-to-date answer supported by the document collection. This restriction brought TREC QA more closely in line with question answering in the real world, where users would want the best answer to their question in the document collection, rather than just any answer found in any document. The evaluation of the question series in 2006 also down-weighted factoid questions, which had been tested for many years, by giving equal weight to each of the 3 question types in the final per-series score.

In addition to the main task, the TREC 2006 QA track also contained a complex, interactive QA (ciQA) task. The 2006 ciQA task was a blend of the TREC 2005 relationship task (Voorhees and Dang, 2006) and the TREC 2005 HARD track, which focused on single-iteration clarification dialogues (Allan, 2006). The goals of the ciQA task were to push the frontiers of question answering away from "factoid" questions towards more complex information needs that exist within richer user contexts, and to move away from the one-shot interaction model implicit in previous evaluations towards a model based at least in part on interactions with users. Two metrics were introduced to evaluate answers to complex questions in the ciQA task: modified F-scores based on nugget pyramids and recall plots based on response length.

The remainder of this paper describes each of the two tasks in the TREC 2006 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. The final section looks at the future of the track.

2 Main Task

The scenario for the main task in the TREC 2006 QA track was that an adult, native speaker of English was 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 to 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 XML format used to describe the test set. Example series (minus the XML tags) are shown in Figure 1. The final test set contained 75 series; the targets of these series are given in Table 1. Of the 75 targets, 19 were PERSONs, 19 were ORGANIZATIONs, 19 were EVENTs, and 18 were THINGs. The series contained a total of 403 factoid questions, 89 list questions, and 75 Other questions. Each series contained 6–9 questions (counting the Other question), with most series containing 8 questions.

Participants were required to submit results within one week of receiving the test set. All processing of the questions was required to be strictly automatic. Systems were required to process series independently from one another, and 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). The document collection from which answers were to be drawn was the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31). As a convenience for track participants, NIST made available document rankings of the top 1000 documents per target as produced using the PRISE document retrieval system, with the target as the query. In total, 59 runs from 27 participants were

145	John William King convicted of murder		
	145.1	FACTOID	How many non-white members of the jury were there?
	145.2	FACTOID	Who was the foreman for the jury?
	145.3	FACTOID	Where was the trial held?
	145.4	FACTOID	When was King convicted?
	145.5	FACTOID	Who was the victim of the murder?
	145.6	LIST	What defense and prosecution attorneys participated in the trial?
	145.7	OTHER	
185	Iditarod Race		
	185.1	FACTOID	In what city does the Iditarod start?
	185.2	FACTOID	In what city does the Iditarod end?
	185.3	FACTOID	In what month is it held?
	185.4	FACTOID	Who is the founder of the Iditarod?
	185.5	LIST	Name people who have won the Iditarod.
	185.6	FACTOID	How many miles long is the Iditarod?
	185.7	FACTOID	What is the record time in which the Iditarod was won?
	185.8	LIST	Which companies have sponsored the Iditarod?
	185.9	OTHER	
212	Barry Manilow		
	212.1	FACTOID	What year was he born?
	212.2	FACTOID	How many times has he married?
	212.3	FACTOID	What is the name of the musical that he wrote about the Harmonistas?
	212.4	FACTOID	What music school did he attend?
	212.5	FACTOID	For what female singer was he the musical director and pianist in the 70's?
	212.6	FACTOID	What record label did he sing for in 2000?
	212.7	LIST	List the songs he recorded.
	212.8	OTHER	

Figure 1: Sample question series from the test set. Series 145 has an EVENT as the target, series 185 has a THING as the target, and series 212 has a PERSON as the target.

141	Warren Moon	179	Hedy Lamarr
142	LPGA	180	Lebanese Parliament
143	American Enterprise Institute	181	Manchester United Football Club
144	82nd Airborne Division	182	1998 Edinburgh Fringe
145	John William King convicted of murder	183	Thabo Mbeki elected president of South Africa
146	Pakistani government overthrown in 1999	184	1999 Chicago Marathon
147	Britain's Prince Edward marries	185	Iditarod Race
148	tourists massacred at Luxor in 1997	186	Pyramids of Egypt
149	The Daily Show	187	Amazon River
150	television show Cheers	188	avocados
151	Winston Cup	189	Joanne Kathleen Rowling
152	Wolfgang Amadeus Mozart	190	H. J. Heinz Co.
153	Alfred Hitchcock	191	International Rowing Federation
154	Christopher Reeve	192	Basque ETA
155	Hugo Chavez	193	World Food Program (WFP)
156	NASCAR	194	1996 World Chess Super Tournament
157	United Nations (U.N.)	195	East Timor Independence
158	Tufts University	196	Adoption of the Euro
159	Wal-Mart	197	cloning of mammals (from adult cells)
160	IMF	198	Bushehr Nuclear Facility
161	1999 Baseball All-Star Game	199	Padre Pio
162	Multiple Myeloma	200	Frank Sinatra
163	Hermitage Museum	201	William Shakespeare
164	Judi Dench	202	Cole Porter
165	the Queen Mum's 100th Birthday	203	Nissan Corp.
166	avian flu outbreak in Hong Kong	204	Church of Jesus Christ of Latter-day Saints (Mormons)
167	the Millennium Wheel	205	1991 eruption of Mount Pinatubo
168	Prince Charles' paintings	206	Johnstown flood
169	stone circles	207	Leaning Tower of Pisa
170	John Prine	208	Great Wall of China
171	Stephen Wynn	209	Carolyn Bessette Kennedy
172	Ben & Jerry's	210	Janet Reno
173	World Tourism Organization (WTO)	211	Patsy Cline
174	American Farm Bureau Federation (AFBF)	212	Barry Manilow
175	repatriation of Elian Gonzales	213	Meg Ryan
176	An Officer and a Gentleman	214	2000 Miss America Pageant
177	Deep Blue	215	1999 Sundance Film Festival
178	methamphetamine labs		

Table 1: Targets of the 75 question series.

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 2006 were identical to those used in the TREC 2005 QA track, except that a distinction was made between locally correct answers (supported in the associated document, but contradicted in later documents in the collection) and globally correct answers. 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 a more recent document contradicts the answer;

globally correct: the answer string consists of exactly the correct answer, that answer is supported by the document returned, and there are no later documents that contradict the answer.

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 team. NIL responses were correct only if there was no known answer to the question in the collection. NIL was correct for 17 of the 403 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.

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 (17). 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.

During the evaluation process, the assessor was given an entire system's run at a time. Each instance was evaluated in the same manner as the factoid questions, i.e., assigned one of the following judgments: incorrect, unsupported, not exact, locally correct, and globally correct. In addition to judging for correctness, the assessor also marked the answer instances for distinctness. The assessor arbitrarily chose any one of equivalent responses to be distinct, and the remainder were considered not distinct. Thus, systems were not rewarded (and in fact, penalized) for returning equivalent answer instances multiple times. Only globally correct responses could be marked as distinct.

The final set of globally correct answers for a list question was compiled from the union of distinct globally correct answers across all runs plus instances the assessor found during question development. For the 89 list questions in the test set, the average number of answers per question was 10, with a minimum of 2 and a maximum of 50. 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 instances. Let S be the number of such instances, D be the number of globally correct, distinct responses returned by the system, and N be the total number of responses returned by the system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined 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 89 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 done 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 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.

Given the nugget list and the set of nuggets matched in a system's response, nugget recall was computed as the ratio of the number of matched nuggets to the total number of vital 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 (vital or non-vital) 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 75 Other questions. Table 4 gives the F-score for the best scoring Other question component for each of the top 10 groups.

2.3.1 Nugget Pyramids

The vital/okay distinction has previously been identified as a weakness in the TREC nugget-based evaluation methodology (Hildebrandt et al., 2004). Since only vital nuggets affect nugget recall, it is difficult for systems to achieve non-zero scores on topics with few vital nuggets in the answer key. Thus, scores are easily affected by assessor errors and other random variations in evaluation conditions. One direct consequence is that in previous TREC evaluations, the median score for many questions turned out to be zero (Voorhees, 2005b). A binary distinction on nugget importance is insufficient to discriminate between the quality of runs that return no vital nuggets but different numbers of okay nuggets. To address many of these issues, Lin and Demner-Fushman (2006) proposed an extension called "nugget pyramids", in which multiple assessors provide judgments of whether a nugget is vital or simply okay.

To examine the effectiveness of the pyramid approach, NIST also computed F-scores for Other responses using the pyramid extension. 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. These nugget weights were then incorporated into the nugget recall computation.

The left graph in Figure 2 plots the average F-scores for each run as computed using a single assessor vs. using the nugget pyramid. Even though the nugget pyramid does not represent any single real user, average pyramid F-scores do correlate highly with average single-assessor F-scores; the Pearson's correlation is 0.987, with a 95% confidence interval of [0.980, 1.00].

While the average F-score for a particular run is stable given a large enough number of questions, the F-score for a single Other question does vary depending on the assessor. The right graph in Figure 2 plots the single-assessor and pyramid F-scores for each individual Other question from all submitted runs. The Pearson correlation between single-assessor and pyramid F-scores in this case is 0.870, with a 95% confidence interval of [0.863, 1.00]. For 16.4% of the questions, the nugget pyramid assigned a non-zero F-score where the original single-assessor F-score was zero. Thus, from the perspective of system developers, the F-scores from the nugget pyramids may be more useful since they are more discriminative. For a more detailed analysis of the nugget pyramids extension, please refer to (Dang and Lin, 2007).

2.4 Per-series Combined Weighted 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 2006, the weighted average of the three component scores 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. In contrast to previous years, when factoid questions were weighted more heavily than the other questions,

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lccPA06	Language Computer Corporation (Moldovan)	0.578	0.000	0.000
LCCFerret	Language Computer Corporation (Harabagiu)	0.538	–	0.000
cuhkqaepisto	The Chinese University of Hong Kong	0.390	0.107	0.353
ed06qar1	University of Edinburgh	0.323	0.069	0.294
InsunQA06	Harbin Institute of Technology (HIT)	0.298	0.118	0.353
QACTIS06A	National Security Agency (NSA)	0.266	0.118	0.118
ILQUA1	University of Albany	0.266	0.027	0.059
NUSCHUAQA1	National University of Singapore	0.261	0.000	0.000
asked06c	Tokyo Institute of Technology	0.251	–	0.000
QASCU3	Concordia University (Kosseim)	0.213	0.000	0.000

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

Run Tag	Submitter	F
lccPA06	Language Computer Corporation (Moldovan)	0.433
cuhkqaepisto	The Chinese University of Hong Kong	0.188
NUSCHUAQA1	National University of Singapore	0.171
FDUQAT15A	Fudan University (Wu)	0.165
QACTIS06C	National Security Agency (NSA)	0.156
LCCFerret	Language Computer Corporation (Harabagiu)	0.148
ILQUA1	University of Albany	0.129
Roma2006run3	University of Rome “La Sapienza”	0.127
csail02	Massachusetts Institute of Technology (MIT)	0.125
InsunQA06	Harbin Institute of Technology (HIT)	0.118

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

Run Tag	Submitter	F($\beta = 3$)
ed06qar1	University of Edinburgh	0.250
FDUQAT15A	Fudan University (Wu)	0.223
QASCU3	Concordia University (Kosseim)	0.199
lccPA06	Language Computer Corporation (Moldovan)	0.167
uw574	University of Washington (UW CLMA group)	0.164
Roma2006run3	University of Rome “La Sapienza”	0.164
MITRE2006C	The MITRE Corp.	0.156
QACTIS06C	National Security Agency (NSA)	0.154
NUSCHUAQA3	National University of Singapore	0.150
ISL2	University of Karlsruhe & Carnegie Mellon University	0.150

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

equal weight was given to the three components in 2006. The final per-series score of each run is simply the average of individual per-series scores.

Table 5 shows the final per-series score for the best run from each group. We fit a two-way analysis of variance model with the target type and the best run from each group as factors, and the final per-series score as the dependent variable; we found significant differences between target types ($p = 0.005$) and runs (p essentially equal to 0). To determine which runs were significantly different from each other, we 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; runs sharing a common letter are not significantly different. A similar multiple comparison showed that PERSON targets had significantly higher scores than EVENTs, but no significant differences between any of the other target types were found.

System scores on the main task have declined since TREC 2004 even though the question series format of the main task has been the same. This is not surprising given that the questions have become increasingly more difficult, with "simple" factoid questions requiring higher levels of reasoning to extract the correct answer from the documents. Assessors also have become more strict about disallowing inexact answers as correct answers.

3 The Complex, Interactive QA (ciQA) Task

The goal of the complex, interactive question answering (ciQA) task is 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. (Question series in the main task also exemplify this shift in evaluation focus.)
- A move away from the one-shot interaction model implicit in previous evaluations towards a model based at least in part on interactions with users.

In terms of implementation, the 2006 ciQA task was a blend of the TREC 2005 relationship task (Voorhees and Dang, 2006) and the TREC 2005 HARD track, which focused on single-iteration clarification dialogues (Allan, 2006).

3.1 Complex "Relationship" Questions

The complex information needs explored by ciQA represented an extension and refinement of so-called "relationship" questions piloted in TREC 2005. This choice provided some continuity and training data for participants.

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. 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, which we refer 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 (items in square brackets) 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. Five template types were developed for the ciQA task, enumerated in Figure 3. For the final test set, NIST assessors developed a total of 30 topics, with 6 topics for each of these templates.

Answers to ciQA topics consisted of [*doc-id*, *answer-string*] pairs, and were evaluated using the same nugget-based methodology that was employed for the main task Other questions. However, the total length of system responses was limited to 7,000 non-whitespace characters. Two metrics were employed to quantify answer quality:

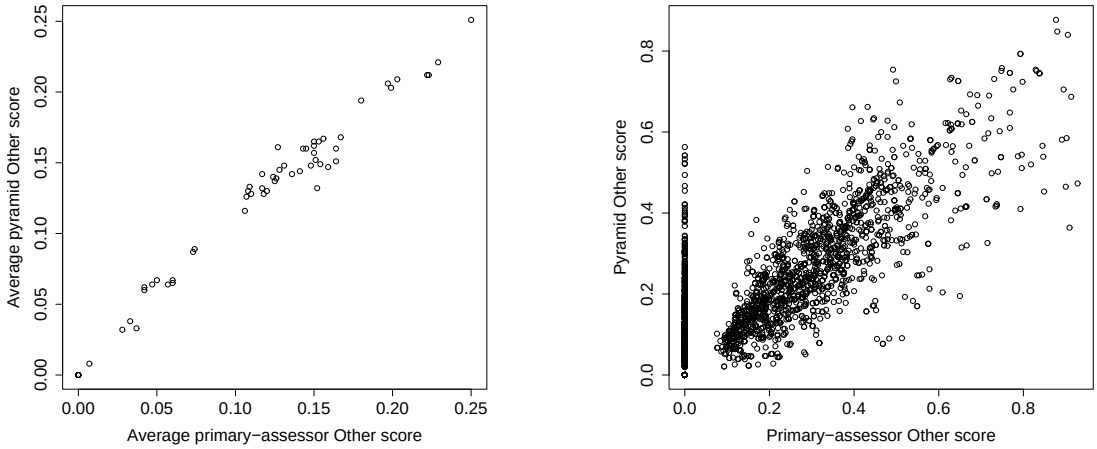

Figure 2: Other F-score computed using a single primary assessor vs. using multiple assessors, by individual question (right), and averaged over all questions for each submitted run (left).

RunID	Per-series score													
lccPA06	0.3938	A												
LCCFerret	0.2644		B											
cuhkqaepisto	0.2310		B	C										
ed06qar1	0.2066		B	C	D									
FDUQAT15A	0.1918			C	D	E								
NUSCHUAQA3	0.1908			C	D	E								
QACTIS06A	0.1853			C	D	E	F							
ILQUA1	0.1713				D	E	F	G						
QASCU1	0.1588				D	E	F	G	H					
Roma2006run3	0.1571				D	E	F	G	H					
InsunQA06	0.1568				D	E	F	G	H					
MITRE2006C	0.1485					E	F	G	H					
ISL2	0.1430					E	F	G	H	I				
csail02	0.1344					E	F	G	H	I				
shef06ss	0.1344					E	F	G	H	I				
lsv2006c	0.1298						F	G	H	I	J			
asked06c	0.1156							G	H	I	J			
uw574	0.1083								H	I	J	K		
DLT06QA02	0.0871									I	J	K	L	
TIQA200601	0.0851									I	J	K	L	
clr06m	0.0763										J	K	L	
TWQA0601	0.0725										J	K	L	M
irstqa06	0.0573											K	L	M
Dal06e	0.0459												L	M
lexiclone06	0.0458												L	M
lf10w10g5	0.0312												L	M
TREC06ST01	0.0167													M

Table 5: Multiple comparison of the best run from each group, based on ANOVA of per-series score.

What evidence is there for transport of [goods] from [entity] to [entity]?
Example: What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?

What [relationship] exist between [entity] and [entity]?
(where [relationship] $\in$ {"financial relationships", "organizational ties", "familial ties", "common interests"})
Example: What [financial relationships] exist between [drug companies] and [universities]?

What influence/effect do(es) [entity] have on/in [entity]?
Example: What effect does [aspirin] have on [coronary heart disease]?

What is the position of [entity] with respect to [issue]?
Example: What is the position of [John McCain] with respect to [the Moral Majority or the Christian Coalition]?

Is there evidence to support the involvement of [entity] in [event/entity]?
Example: Is there evidence to support the involvement of [China] in [human organ transplants from Chinese prisoners]?

Figure 3: The five templates used in the TREC 2006 ciQA task.

- The first and primary metric was the F-score ($\beta = 3$) with the "nugget pyramid" extension.
- The second metric was new for the ciQA task and attempted to graphically capture the tradeoffs between conciseness and completeness (Lin, 2007). The basic idea is to quantify weighted nugget recall (what we call pyramid recall) as a function of answer length (in non-whitespace characters). By the nugget pyramid building process, each nugget is assigned a weight between zero and one. Weighted nugget recall is the sum of weights of all nuggets retrieved divided by the sum of all weights of all nuggets in the assessor's answer key.

Implementing this metric required two important changes to the previous evaluation protocol:

1. Answer strings must be rank ordered, with best first.
2. Assessors must mark the first instance of a nugget in the response set of answer strings.

For the recall plots, the scoring methodology was as follows (character counts do not include whitespaces):

1. For each topic, NIST recorded the cumulative character length and pyramid recall after each answer string had been assessed.
2. Each data point was interpolated to the nearest 100 character increment (longer than the current length). For example, a pyramid recall of 0.25 at 168 characters would be interpolated to (200, 0.25). Plotting these points yielded pyramid recall as a function of answer length for a particular topic.
3. To arrive at the recall plot for a particular system run, the mean of the recall values was taken across all topics at each length increment, i.e., mean pyramid recall over all topics at 100 characters, at 200 characters, at 300 characters, etc.

3.2 Interactive Question Answering

The purpose of the interactive aspect of ciQA was to provide a framework for participants to investigate interaction in the QA context and to provide an opportunity for non-QA researchers to become involved in this area. We consider an interactive system to be a system that gives users control over all or a portion of displayed content. Using this definition, the smallest possible interaction unit consists of the user responding to the system and the system using the

user's response to produce new content. The interactive aspect of ciQA was concerned with this interaction unit and was modeled in part after the HARD track's clarification forms.

The HARD track's clarification forms allowed participants to elicit information from assessors through a single interaction. This interaction consisted of assessors completing forms (i.e., Web pages) that had been created by track participants. The results of these interactions were then returned to the participants so that revised results could be generated—comparison of output before and after the clarification quantified the effects of the interaction.

Although many participants took advantage of the opportunity provided by the HARD track to investigate traditional relevance feedback techniques, this was not a goal of the HARD track nor a condition for participation; there were, in fact, some participants who used clarification forms in novel ways. In the ciQA task, we explicitly encouraged innovative ways of using forms that go beyond traditional relevance feedback. The question answering community has yet to reach common ground on the role of interaction in QA, and the ciQA task was meant to provide a forum for continued dialogue.

The rationale for studying the smallest interaction unit rests on the idea that a good QA system should return relevant information with a minimum amount of interaction. Furthermore, given the potential complexities that are likely to arise with coordinating cross-site interactive evaluations, we believe that using the smallest interaction unit is a reasonable starting point in the exploration of interactive QA. Previous experiences with the TREC interactive track demonstrated that coordinating multi-site interactive IR system evaluation is a challenge and that results are difficult (if not impossible) to compare.

In more detail, interaction forms were HTML pages created by participants that solicited user input via CGI. Although NIST placed no restrictions on the type of content, there were technical restrictions (see below). Each question was associated with a unique form, and each site was limited to two sets of interaction forms (which provided the ability to evaluate two different interaction techniques).

NIST assessors completed the interaction forms on Redhat Enterprise Linux workstations with 20-inch LCD monitors (1600×1200 resolution and millions of colors) using the Firefox Web browser (v1.5.0.2). The machines at NIST were disconnected from all networks and participants were required to provide all necessary information as part of their forms. If a form required multiple files, then it was necessary for such files to be contained within the submitted directory structure. These forms were not allowed to invoke any CGI scripts or write to disk. Javascript was allowed, but Java was not.

Assessors spent no more than three minutes completing each interaction form. This duration included the time needed to load the form, initialize any content, and then render it. At the end of three minutes, if the assessor had not submitted the form, the form timed out and was forcibly submitted. The CGI variable bindings associated with the forms captured the results of the interactions, which NIST returned to the participants.

3.3 Results

The ciQA evaluation proceeded as follows:

1. Participants submitted initial runs and interaction forms.
2. NIST assessors interacted with the forms.
3. NIST returned results of the interaction (i.e., the CGI bindings).
4. Participants submitted final runs based on the results of the interactions.
5. NIST evaluateed both initial and final runs.

As with the main task, the AQUAINT collection of newswire articles served as the official corpus. To support the individual goals of participants, ciQA was entirely independent of the main task; the interactive aspect was also optional, which allowed participants to focus solely on complex QA if they desired. Finally, both automatic and manual runs were allowed. A manual run was defined as any run where human intervention occurred in any part of the process (except assessor interaction with the submitted interaction forms).

The ciQA task drew participation from six groups. NIST received ten initial runs and eleven final runs. A total of ten sets of interaction forms were submitted by the six participants. In addition, a pair of initial/final runs that

used simple sentence retrieval techniques was submitted as a baseline implementation (described below). A set of interaction forms was also associated with this run pair.

We constructed a rotation specifying the order in which interaction forms would be presented to assessors to minimize learning and order effects, and to insure that each form would occupy each position in the rotation (e.g., first, second, third) as equal a number of times as possible. This rotation is shown in Table 6. Row headings show topic numbers, while column headings represent forms. Cell numbers indicate the presentation order of the form; for example, for Topic 26, CLR1 was the fourth form presented and strath3 was the first. This rotation is based on a basic Latin square rotation; the relationship between forms is preserved, but the position of the form is shifted across topics. For example, strath2 always followed CLR1 except when it was the first form in the rotation, and strath2 and CLR1 were each the first, second, third, etc. form in the rotation an equal number of times. To construct the order, forms and topics were randomly assigned to column and row headings, and an order of 1, 2, 3, 4, etc. was assigned to the first row, 2, 3, 4, 5, etc. was assigned to the second row, etc. The table has been sorted according to topic, but one can see that Topics 31, 34 and 39 appeared in either the 1st, 12th, or 23rd row of the randomized table.

In total, there were eleven different initial–final run pairs. The pyramid F-scores of these run pairs are shown in Table 7. Pyramid F-scores were computed using the methodology outlined in (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).

In addition to runs submitted by the participants, the University of Maryland separately prepared a sentence retrieval baseline. 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 initial run (up to the 7,000 character limit). Sentence order within each document and across the ranked list was preserved. The interaction forms associated with this run asked the assessor for relevance judgments on each of the sentences (relevant, not relevant, don't know). The final run was prepared by removing sentences judged not relevant—this had the effect of pulling in more sentences from documents lower in the ranked list. The performance of this sentence retrieval baseline is also shown in Table 7.

Surprisingly, the sentence retrieval baseline performed exceedingly well. Only two initial runs received a higher score, one of which was a manual run. Only two final runs received a higher score, one of which was a manual run. The high baseline performance is consistent with findings from previous TREC results (Voorhees, 2004). Figure 4 shows a scatter plot of the initial and final F-scores for all eleven run pairs. Points below the reference line $y = x$ represent cases in which interaction actually decreased performance—there were two such cases.

Plots of pyramid recall as a function of response length are shown in Figure 5. These graphs attempt to quantify how quickly a user is able to acquire relevant nuggets by reading system responses. Naturally, curves that rise more quickly represent "better" systems. In the top graph, the sentence retrieval baseline is compared against the best automatic run. In the bottom graph, the sentence retrieval baseline is compared against the best manual run. It is interesting to note that for the automatic runs, these recall plots paint a different picture of performance than the pyramid F-scores. Although UWATCIQA4 achieved a higher pyramid F-score than the final submission of the sentence retrieval baseline, the recall plots suggest that the sentence retrieval baseline is able to deliver more information given the same response length. For the manual run, although the recall plots show little difference between the nugget content of the pre- and post-interaction system responses, the pyramid F-scores suggest a difference in answer quality. More work is needed to understand the divergences between pyramid F-scores and these recall plots.

These results appear to suggest that the complex QA task is difficult, but that off–the–shelf IR systems provide a strong baseline. The effective use of linguistic analysis techniques for complex questions remains an open research question. For a more in-depth exploration of these issues and the evaluation methodology, see (Lin, 2007).

4 Future of the QA Track

At the TREC 2006 workshop, participants indicated that they would like 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 would be located at a URL on the participant's machine, and NIST assessors would simply navigate to the URL. The advantage would be that participants would be able to host more complex interaction interfaces. On the other hand, this setup would put additional burden on each partici-

Topic	CLR1	strath2	csaili2	UMDA1	UMAS1	UWAT1	CLR2	csaili1	strath3	UMDM1	Baseline1
26	4	5	6	7	8	9	10	11	1	2	3
27	7	8	9	10	11	1	2	3	4	5	6
28	4	5	6	7	8	9	10	11	1	2	3
29	9	10	11	1	2	3	4	5	6	7	8
30	9	10	11	1	2	3	4	5	6	7	8
31	1	2	3	4	5	6	7	8	9	10	11
32	6	7	8	9	10	11	1	2	3	4	5
33	3	4	5	6	7	8	9	10	11	1	2
34	1	2	3	4	5	6	7	8	9	10	11
35	5	6	7	8	9	10	11	1	2	3	4
36	5	6	7	8	9	10	11	1	2	3	4
37	6	7	8	9	10	11	1	2	3	4	5
38	5	6	7	8	9	10	11	1	2	3	4
39	1	2	3	4	5	6	7	8	9	10	11
40	11	1	2	3	4	5	6	7	8	9	10
41	2	3	4	5	6	7	8	9	10	11	1
42	10	11	1	2	3	4	5	6	7	8	9
43	8	9	10	11	1	2	3	4	5	6	7
44	10	11	1	2	3	4	5	6	7	8	9
45	11	1	2	3	4	5	6	7	8	9	10
46	8	9	10	11	1	2	3	4	5	6	7
47	2	3	4	5	6	7	8	9	10	11	1
48	8	9	10	11	1	2	3	4	5	6	7
49	7	8	9	10	11	1	2	3	4	5	6
50	11	1	2	3	4	5	6	7	8	9	10
51	6	7	8	9	10	11	1	2	3	4	5
52	3	4	5	6	7	8	9	10	11	1	2
53	10	11	1	2	3	4	5	6	7	8	9
54	7	8	9	10	11	1	2	3	4	5	6
55	9	10	11	1	2	3	4	5	6	7	8

Table 6: Form rotation according to topic. As a specific example: for Topic 26, the form CLR1 was presented fourth and the form strath3 was presented first.

Organization	Type	Run tags		Pyramid F-Score	
		Initial	Final	Initial	Final
CL Research	automatic	clr06ci1	clr06ci1r	0.151	0.184
CL Research	automatic	clr06ci2	clr06ci2r	0.175	0.209
MIT	automatic	csail1	csailif1	0.203	0.209
MIT	automatic	csail1	csailif2	0.203	0.203
U. Maryland	automatic	UMDA1pre	UMDA1post	0.224	0.180
U. Maryland	manual	UMDM1pre	UMDM1post	0.316	0.350
U. Mass.	automatic	UMASSauto2	UMASSi2	0.171	0.160
U. Mass.	automatic	UMASSauto1	UMASSi1	0.133	0.150
U. Strathclyde	manual	strath1	strath4	0.227	0.239
U. Waterloo	automatic	UWATCIQA1	UWATCIQA3	0.247	0.247
U. Waterloo	automatic	UWATCIQA1	UWATCIQA4	0.247	0.268
Baseline	automatic	-	-	0.237	0.264

Table 7: Performance of the eleven initial–final pairings for the ciQA task, along with the sentence retrieval baseline.

pant; if the NIST assessor could not reach a site for any reason during the interaction period – even due to problems outside the control of the site – the assessor would simply ignore the site. A straw poll indicated preference for live interactions, and the ciQA task will be repeated in 2007 with live URLs and a longer interaction period. Based on the successful application of the nugget pyramid evaluation method in TREC 2006, the pyramid method will be the official evaluation method for both the ciQa and the Other questions in TREC 2007.

Since the main task had been run largely unchanged for three years, a radical change was proposed to push the state of the art forward. The series format has supported the evaluation of different types of questions (factoid, list and Other) while providing an abstraction of a real user session with a QA system; therefore, rather than changing the series format, it was decided to move the main task forward by changing the genre of the document collection. The main task for the TREC 2007 QA Track will again be series of factoid, list, and Other questions, but the document collection will be a combination of newswire and blogs. Mining blogs for answers will introduce significant new challenges in at least two aspects that are very important for functional 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.

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 Trang 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)*.

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.

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.

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.

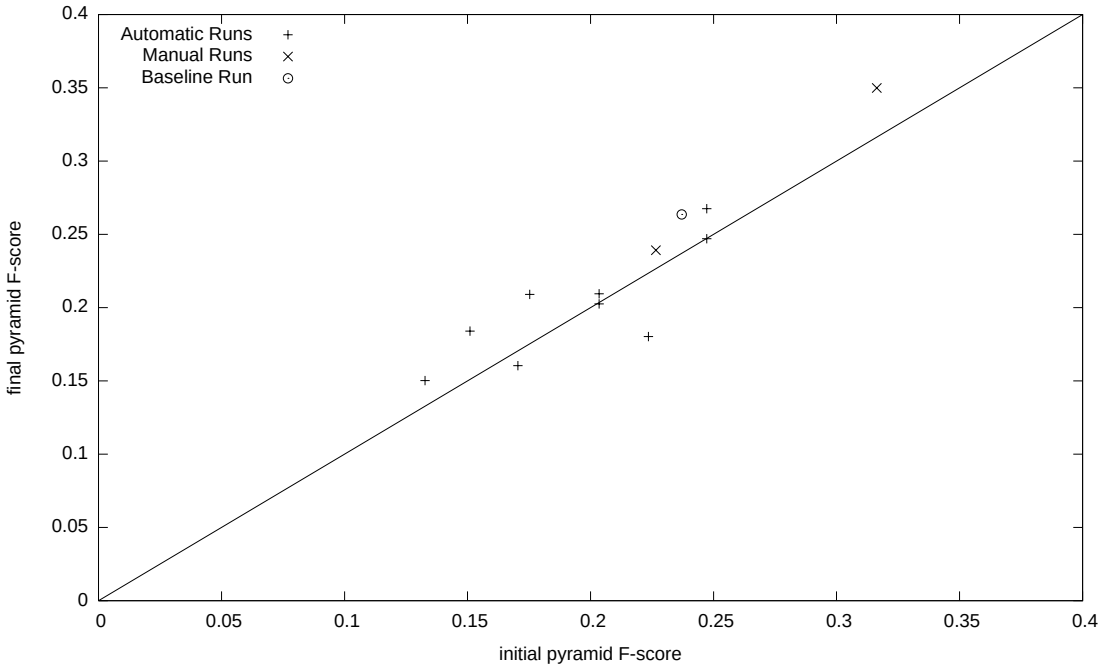

Figure 4: Scatter plot showing initial and final pyramid F-scores for submitted ciQA runs.

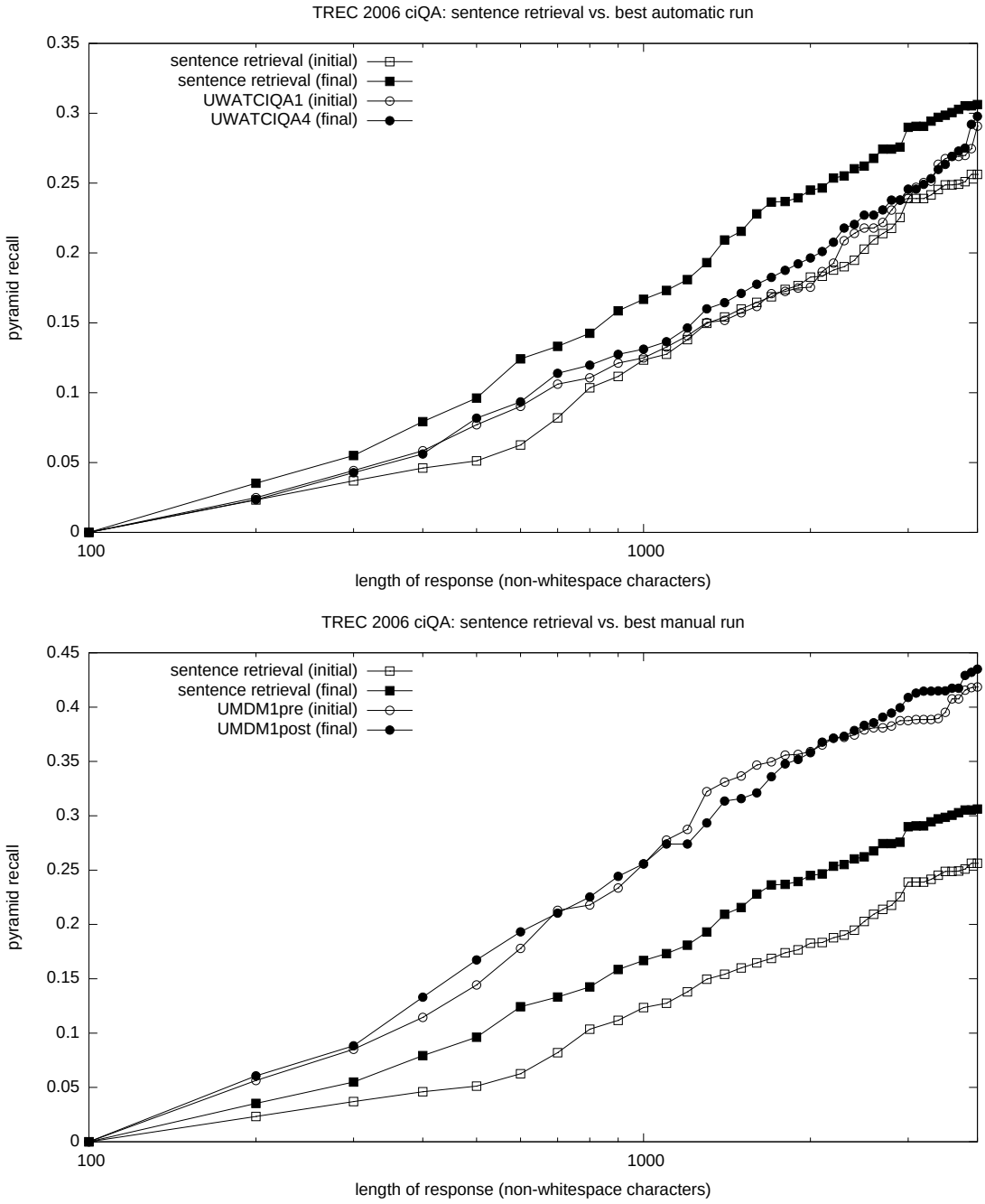

Figure 5: Plots of pyramid recall from ciQA runs as a function of response length: sentence retrieval baseline vs. the best automatic run (top) and vs. the best manual run (bottom)