Overview of the TREC 2005 Question Answering Track

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

Abstract

The TREC 2005 Question Answering (QA) track contained three tasks: the main question answering task, the document ranking task, and the relationship task. In the main task, question series were used to define a set of targets. Each series was about a single target and contained factoid and list questions. The final question in the series was an "Other" question that asked for additional information about the target that was not covered by previous questions in the series. The main task was the same as the single TREC 2004 QA task, except that targets could also be events; the addition of events and dependencies between questions in a series made the task more difficult and resulted in lower evaluation scores than in 2004. The document ranking task was to return a ranked list of documents for each question from a subset of the questions in the main task, where the documents were thought to contain an answer to the question. In the relationship task, systems were given TREC-like topic statements that ended with a question asking for evidence for a particular relationship.

The goal of the TREC question answering (QA) track is to foster research on systems that return answers themselves, rather than documents containing answers, in response to a question. The track started in TREC-8 (1999), with the first several editions of the track focused on *factoid* questions. A factoid question is a fact-based, short answer question such as *How many calories are there in a Big Mac?*. The task in the TREC 2003 QA track contained list and definition questions in addition to factoid questions [1]. A list question asks for different instances of a particular kind of information to be returned, such as *List the names of chewing gums.* Answering such questions requires a system to assemble an answer from information located in multiple documents. A definition question asks for interesting information about a particular person or thing such as *Who is Vlad the Impaler?* or *What is a golden parachute?*. Definition questions also require systems to locate information in multiple documents, but in this case the information of interest is much less crisply delineated.

In TREC 2004 [2], 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.

The TREC 2005 QA track contained three tasks: the main question answering task, the document ranking task, and the relationship task. The document collection from which answers were to be drawn was the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31). The main task was the same as the TREC 2004 task, with one significant change: in addition to persons, organizations, and things, the target could also be an event. Events were added in response to suggestions that the question series include answers that could not be readily found by simply looking up the target in Wikipedia or other pre-compiled Web resources. The runs were evaluated using the same methodology as in TREC 2004, except that the primary measure was the per-series score instead of the combined component score.

The document ranking task was added to build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology. The task was to submit, for a subset of 50 of the questions in the main task, a ranked list of up to 1000 documents for each question. The purpose of the lists was to create document pools both to get a better understanding of the number of instances of correct answers in the collection and to support research on whether some document retrieval techniques are better than others in support of QA. NIST

pooled the document lists for each question, and assessors judged each document in the pool as relevant ("contains an answer") or not relevant ("does not contain an answer"). Document lists were then evaluated using trec_eval measures.

Finally, the relationship task was added. The task was the same as was performed in the AQUAINT 2004 relationship pilot. Systems were given TREC-like topic statements that ended with a question asking for evidence for a particular relationship. The initial part of the topic statement set the context for the question. The question was either a yes/no question, which was understood to be a request for evidence supporting the answer, or an explicit request for the evidence itself. The system response was a set of information nuggets that were evaluated using the same scheme as definition and Other questions.

The remainder of this paper describes each of the three tasks in the TREC 2005 QA track in more detail. Section 1 describes the question series that formed the basis of the main and document ranking tasks; section 2 describes the evaluation method and resulting scores for the runs for the main task, while section 3 describes the evaluation and results of the document ranking task. The questions and results for the relationship task are described in section 4. Section 5 summarizes the technical approaches used by the systems to answer the questions, and the final section looks at the future of the track.

1 Question Series

The main task for the TREC 2005 QA track required providing answers for each question in a set of question series. A question series consists of several factoid questions, one to two list questions, and exactly one Other question. Associated with each series is a definition target. The series that a question belongs to, the order of the question in the series, and the type of each question (factoid, list, or Other) are all explicitly encoded in the XML format used to describe the test set. Example series (minus the XML tags) are shown in figure 1.

95	return of Hong Kong to Chinese sovereignty		
	95.1	FACTOID	What is Hong Kong's population?
	95.2	FACTOID	When was Hong Kong returned to Chinese sovereignty?
	95.3	FACTOID	Who was the Chinese President at the time of the return?
	95.4	FACTOID	Who was the British Foreign Secretary at the time?
	95.5	LIST	What other countries formally congratulated China on the return?
	95.6	OTHER	
111	AMWAY		
	111.1	FACTOID	When was AMWAY founded?
	111.2	FACTOID	Where is it headquartered?
	111.3	FACTOID	Who is the president of the company?
	111.4	LIST	Name the officials of the company.
	111.5	FACTOID	What is the name "AMWAY" short for?
	111.6	OTHER	
136	Shiite		
	136.1	FACTOID	Who was the first Imam of the Shiite sect of Islam?
	136.2	FACTOID	Where is his tomb?
	136.3	FACTOID	What was this person's relationship to the Prophet Mohammad?
	136.4	FACTOID	Who was the third Imam of Shiite Muslims?
	136.5	FACTOID	When did he die?
	136.6	FACTOID	What portion of Muslims are Shiite?
	136.7	LIST	What Shiite leaders were killed in Pakistan?
	136.8	OTHER	

Figure 1: Sample question series from the test set. Series 95 has an EVENT as a target, series 111 has an ORGANIZATION as a target, and series 136 has a THING as a target.

The scenario for the main task was that an adult, native speaker of English was looking for more information about

a target that interested him. The target could be a person, organization, thing, or event. The user was assumed to be an "average" reader of U.S. newspapers. NIST assessors acted as surrogate users and developed the question and judged the system responses.

In TREC 2004, the question series had been written primarily *before* the assessors had searched the AQUAINT document collection; consequently, many of the question series had been unusable because the document collection did not have sufficient information to answer the questions. Therefore, the questions for TREC 2005 were developed by the assessors *after* searching the AQUAINT document collection to make sure that there was sufficient information about the target. The assessors created factoid and list questions whose answers could be found in the document collection; they tried to phrase the questions as something they would have asked if they hadn't seen the documents already. The assessors also recorded other interesting information that was not an answer to a factoid or list question (because the information was not a factoid, or the question would be too obviously a back-formulation of the answer), which could be used to answer the final "Other" question in the series.

Context processing is an important element for question answering systems to possess, so a question in the series could refer to the target or a previous answer using a pronoun, definite noun phrase or other referring expression, as shown in figure 1. Each series is an abstraction of an information dialogue in which the user is trying to define the target, but it is only a limited abstraction. Unlike in a real dialogue, questions could not mention (by name) an answer to a previous question in the series. Because each usable series was *required* to contain a list question whose answers were named entities, assessors sometimes asked list questions that they were not actually interested in. This means that the series may not necessarily be true samples of the assessor's interests in the target.

The final test set contained 75 series; the targets of these series are given in table 1. Of the 75 targets, 19 are PERSONs, 19 are ORGANIZATIONs, 19 are THINGs, and 18 are EVENTs. The series contained a total of 362 factoid questions, 93 list questions, and 75 (one per target) Other questions. Each series contained 6-8 questions (counting the Other question), with most series containing 7 questions.

Participants were required to submit retrieval results within one week of receiving the test set. All processing of the questions was required to be strictly automatic. Systems were required to process series independently from one another, and required to process an individual series in question order. That is, systems were allowed to use questions and answers from earlier questions in a series to answer later questions in that same series, but could not "look ahead" and use later questions to help answer earlier questions. As a convenience for the track, NIST made available document rankings of the top 1000 documents per target as produced using the PRISE document retrieval system and the target as the query. Seventy-one runs from 30 participants were submitted to the main task.

2 Main Task Evaluation

The evaluation of a single run comprises the component evaluations for each of the question types, and a final average per-series score. Each of the three question types has its own response format and evaluation method. The individual component evaluations for 2005 were identical to those used in the TREC 2004 QA track. Next, a per-series score was computed for a run using a weighted 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 for a factoid question was either exactly one [*doc-id*, *answer-string*] pair or the literal string 'NIL'. Since there was no guarantee that a factoid question had an answer in the document collection, NIL was returned by the system when it believed there was no answer. Otherwise, *answer-string* was a string containing precisely an answer to the question, and *doc-id* was the id of a document in the collection that supported *answer-string* as an answer.

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

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

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

Table 1: Targets of the 75 question series.

66	Russian submarine Kursk sinks	104	1999 North American International Auto Show
67	Miss Universe 2000 crowned	105	1980 Mount St. Helens eruption
68	Port Arthur Massacre	106	1998 Baseball World Series
69	France wins World Cup in soccer	107	Chunnel
70	Plane clips cable wires in Italian resort	108	Sony Pictures Entertainment (SPE)
71	F16	109	Telefonica of Spain
72	Bollywood	110	Lions Club International
73	Viagra	111	AMWAY
74	DePauw University	112	McDonald's Corporation
75	Merck and Co.	113	Paul Newman
76	Bing Crosby	114	Jesse Ventura
77	George Foreman	115	Longwood Gardens
78	Akira Kurosawa	116	Camp David
79	Kip Kinkel school shooting	117	kudzu
80	Crash of EgyptAir Flight 990	118	U.S. Medal of Honor
81	Preakness 1998	119	Harley-Davidson
82	Howdy Doody Show	120	Rose Crumb
83	Louvre Museum	121	Rachel Carson
84	meteorites	122	Paul Revere
85	Norwegian Cruise Lines (NCL)	123	Vicente Fox
86	Sani Abacha	124	Rocky Marciano
87	Enrico Fermi	125	Enrico Caruso
88	United Parcel Service (UPS)	126	Pope Pius XII
89	Little League Baseball	127	U.S. Naval Academy
90	Virginia wine	128	OPEC
91	Cliffs Notes	129	NATO
92	Arnold Palmer	130	tsunami
93	first 2000 Bush-Gore presidential debate	131	Hindenburg disaster
94	1998 indictment and trial of Susan McDougal	132	Kim Jong Il
95	return of Hong Kong to Chinese sovereignty	133	Hurricane Mitch
96	1998 Nagano Olympic Games	134	genome
97	Counting Crows	135	Food-for-Oil Agreement
98	American Legion	136	Shiite
99	Woody Guthrie	137	Kinmen Island
100	Sammy Sosa	138	International Bureau of Universal Postal Union (UPU)
101	Michael Weiss	139	Organization of Islamic Conference (OIC)
102	Boston Big Dig	140	PBGC
103	Super Bowl XXXIV		

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

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lcc05	Language Computer Corp.	0.713	0.643	0.529
NUSCHUA1	National Univ. of Singapore	0.666	0.148	0.529
IBM05L3P	IBM T.J. Watson Research	0.326	0.200	0.118
ILQUA2	Univ. of Albany	0.309	0.075	0.235
Insun05QA1	Harbin Inst. of Technology	0.293	0.057	0.176
csail2	MIT	0.273	0.098	0.294
FDUQA14B	Fudan University	0.260	0.082	0.412
QACTIS05v2	National Security Agency (NSA)	0.257	0.045	0.176
mk2005qar2	Saarland University	0.235	0.071	0.353
Edin2005b	Univ. of Edinburgh	0.215	0.068	0.176

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

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

To be responsive, an answer string was required to contain appropriate units and to refer to the correct "famous" entity (e.g., the Taj Mahal casino is not responsive when the question asks about "the Taj Mahal"). 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 are correct only if there is no known answer to the question in the collection and are incorrect otherwise. NIL is correct for 17 of the 362 factoid questions in the test set. (Eighteen questions had no correct response returned by the systems, but did have a correct answer found by the assessors.)

The main evaluation score for the factoid component is *accuracy*, the fraction of questions judged correct. Also reported are the recall and precision of recognizing when no answer exists in the document collection. NIL precision is the ratio of the number of times NIL was returned and correct to the number of times it was returned, whereas NIL recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct (17). If NIL was never returned, NIL precision is undefined and NIL recall is 0.0. Table 2 gives evaluation results for the factoid component. The table shows the most accurate run for the factoid component for each of the top 10 groups. The table gives the accuracy score over the entire set of factoid questions as well as NIL precision and recall scores.

2.2 List questions

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

The final set of correct answers for a list question was compiled from the union of the correct responses across all runs plus the instances the assessor found during question development. For the 93 list questions used in the evaluation, the average number of answers per question is 12.5, with 2 as the smallest number of answers, and 70 as the maximum number of answers. A system's response to a list question was scored using instance precision (IP) and instance recall (IR) based on the list of known instances. Let S be the the number of known instances, D be the number of correct, distinct responses returned by the system, and N be the total number of responses returned by the

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

Run Tag	Submitter	F
lcc05	Language Computer Corp.	0.468
NUSCHUA3	National Univ. of Singapore	0.331
IBM05C3PD	IBM T.J. Watson Research	0.131
ILQUA1	Univ. of Albany	0.120
csail1	MIT	0.110
QACTIS05v1	National Security Agency (NSA)	0.105
Insun05QA1	Harbin Inst. of Technology	0.085
Edin2005a	Univ. of Edinburgh	0.081
MITRE2005B	Mitre Corp.	0.080
shef05lmg	Univ. of Sheffield	0.076

system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined using the F measure with equal weight given to recall and precision:

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

The score for the list component of a run was the average F score over the 93 questions. Table 3 gives the average F scores for 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 same methodology as the TREC 2003 definition questions. A system's response for an Other question was an unordered set of [*doc-id*, *answer-string*] pairs as in the list component. Each string was presumed to be a facet in the definition of the series' target that had not yet been covered by earlier questions in the series. The requirement to not repeat information already covered by earlier questions in the series made answering Other questions somewhat more difficult than answering TREC 2003 definition questions.

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

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

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

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

Run Tag	Submitter	F($\beta = 3$)
QACTIS05v3	National Security Agency (NSA)	0.248
FDUQA14B	Fudan University	0.232
lcc05	Language Computer Corp.	0.228
MITRE2005B	Mitre Corp.	0.217
NUSCHUA3	National Univ. of Singapore	0.211
ILQUA2	Univ. of Albany	0.207
IBM05C3PD	IBM T.J. Watson Research	0.206
uams05be3	Univ. of Amsterdam	0.201
SUNYSB05qa2	SUNY Stony Brook	0.196
UNTQA0501	Univ. of North Texas	0.191

computed as the F measure with nugget recall three times as important as nugget precision:

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

The score for the Other question component was the average F($\beta = 3$) score over 75 Other questions. Table 4 gives the average F($\beta = 3$) score for the best scoring Other question component for each of the top 10 groups.

As a separate experiment, the University of Maryland created a manual "run" for the Other questions, in which a human wrote down what he thought were good nuggets for each of the questions. This manual run was included in the judging of the submitted automatic runs, and received an average F($\beta = 3$) score of 0.299. The low score may indicate the level of variation between humans regarding what information is considered interesting (vital or okay) for a target. However, this score should not be taken as an upper bound on system performance, since the manual run sometimes included information from previous questions in the series (which were explicitly excluded from the desired Other information). The run also had shorter answer strings than the best system responses; this resulted in high average precision (0.482) at the cost of lower recall (0.296), while the scoring method gave greater importance to recall than precision.

2.4 Per-series Combined Weighted Scores

The three component scores measure systems' 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, evaluating over the individual series 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 scores as the final score for the run. The weighted average of the three component scores for a series for a QA run is computed as:

$$\text{WeightedScore} = .5 \times \text{Factoid} + .25 \times \text{List} + .25 \times \text{Other}.$$

To compute the weighted score for an individual series, only the scores for questions belonging to the series were part of the computation. Since each of the component scores ranges between 0 and 1, the weighted score is also in that range. The average per-series weighted score is called the per-series score and gives equal weight to each series. Table 5 shows the per-series score for the best run for each of the top 10 groups.

Each individual series has only a few questions, so the combined weighted score for an individual series will be much less stable than the global score. But the average of 75 per-series scores should be at least as stable as the overall combined weighted average and has some additional advantages. The per-series score is computed at a small enough

Table 5: Per-series scores for QA task runs. Scores are given for the best run from the top 10 groups.

Run Tag	Submitter	Per-series Score
lcc05	Language Computer Corp.	0.534
NUSCHUA3	National Univ. of Singapore	0.464
IBM05C3PD	IBM T.J. Watson Research	0.246
ILQUA2	Univ. of Albany	0.241
QACTIS05v3	National Security Agency (NSA)	0.222
FDUQA14B	Fudan University	0.205
csail2	MIT	0.201
Insun05QA1	Harbin Inst. of Technology	0.187
shef05lmg	Univ. of Sheffield	0.165
mk2005qar2	Saarland University	0.158

granularity to be meaningful at the task-level (i.e., each series representing a single user interaction), and at a large enough granularity for individual scores to be meaningful. As pointed out in [2], many individual questions have zero for a median score over all runs, but only a few series have a zero median per-series score.

We fit a two-way analysis of variance model with the target type and the best run from each group as factors, and the per-series combined score as the dependent variable. Both main effects are significant at a p value essentially equal to 0, which indicates that there are significant differences between runs as well as between target types. 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 6 shows the results of the multiple comparison; runs sharing a common letter are not significantly different.

A similar analysis showed that PERSON and ORGANIZATION type targets having significantly higher per-series scores than EVENT and THING targets. System effectiveness may be higher for persons and organizations because the types of information desired for a person or organization may be more standard than for an event or thing. While it may be possible to come up with templates for events, identifying references to a particular event in a document collection is difficult because events are usually unnamed and the exent of the event is not always well-defined.

3 Document Ranking Task

The goal of the document ranking task was to create pools of documents containing answers to questions in the main series. These pools would provide an estimate of the number of instances of correct answers in the collections for people wanting to use the 2005 evaluated data for post-conference experiments. The task would also support research on whether some document retrieval techniques are better than others in support of QA, since groups were allowed to mix and match different techniques for retrieval and QA.

All TREC 2005 submissions to the main task were required to include a ranked list of documents for each question in the document ranking task; the list represented the set of documents used by the system to create its answer, where the order of the documents in the list was the order in which the system considered the document. There were 77 submissions to the document ranking task. Groups whose primary emphasis was document retrieval rather than QA, were allowed to participate in the document ranking task without submitting actual answers for the main task; three groups participated in the document ranking task without participating in the main task.

The test set for the document ranking task was a list of question numbers for 50 of the questions from the main task. The set of 50 questions comprised all the factoid and list questions from two series and additional factoid questions from other series. Half of these questions contained pronouns or other anaphors that referred to the target or answer to a previous question. For each question, systems returned a ranked list of up to 1000 documents that were thought to contain an answer for the question.

RunID	PMM														
lcc05	0.5343	A													
NUSCHUA3	0.4641		B												
IBM05C3PD	0.2457			C											
ILQUA2	0.2412			C											
QACTIS05v3	0.2219			C	D										
FDUQA14B	0.2050			C	D										
csail2	0.2004			C	D	E									
Insun05QA1	0.1868			C	D	E	F								
shef05lmg	0.1644				D	E	F	G							
mk2005qar2	0.1578				D	E	F	G							
asked05c	0.1568				D	E	F	G							
Edin2005c	0.1552				D	E	F	G	H						
clr05	0.1357					E	F	G	H	I					
UNTQA0503	0.1337					E	F	G	H	I	J				
ASUQA02	0.1332					E	F	G	H	I	J				
MITRE2005B	0.1328					E	F	G	H	I	J				
uams05be3	0.1268						F	G	H	I	J				
talpupc05b	0.1253						F	G	H	I	J	K			
SUNYSB05qa3	0.1232						F	G	H	I	J	K			
DLT05QA01	0.1183						F	G	H	I	J	K	L		
CMUJAV2005	0.1060							G	H	I	J	K	L		
Dal05s	0.0872								H	I	J	K	L	M	
lexicloneB	0.0841									I	J	K	L	M	
TWQA0502	0.0748									I	J	K	L	M	N
Mon05BIMP2	0.0699									I	J	K	L	M	N
thuir051	0.0654										J	K	L	M	N
dggQA05X	0.0568											K	L	M	N
MSRCOMB05	0.0542												L	M	N
UIowaQA0503	0.0271													M	N
afrun1	0.0152														N

Table 6: Multiple comparison of best run from each group, based on ANOVA of per-series score. PMM is the population marginal mean of the per-series score for the run.

Table 7: R-Precision and MAP scores for the document-ranking task runs. Scores are given for the best run from the top 13 groups.

Run Tag	Submitter	R-Prec	MAP
NUSCHUA1	National Univ. of Singapore	0.4570	0.4698
$*$ humQ05xle	Hummingbird	0.4127	0.4468
IBM05C3PD	IBM T.J. Watson Research	0.3978	0.4038
QACTIS05v1	National Security Agency (NSA)	0.3414	0.3498
$*$ apl05aug	Johns Hopkins Univ. Applied Physics Lab	0.3201	0.3417
ASUQA01	Arizona State Univ.	0.2958	0.3321
UNTQA0501	Univ. of North Texas	0.3205	0.3285
$*$ sab05qa1b	Sabir Research	0.3366	0.3197
lcc05	Language Computer Corp.	0.2921	0.3045
afrun1	Macquarie Univ.	0.3038	0.2852
TWQA0501	Peking Univ.	0.2732	0.2832
csail2	MIT	0.2699	0.2808
ILQUA1	Univ. of Albany	0.2445	0.2596

3.1 Evaluation

For each of the 50 questions, the documents in the top 75 ranks for up to two runs per group were pooled and then judged by the human assessor. A document was considered relevant if the document contained a correct, supported answer and not relevant otherwise. Each pool had an average of about 717 documents; the smallest pool had 295 documents, and the largest pool had 1219 documents. The number of relevant documents (containing an answer) in each pool ranged from 1 to 285, with a mean of 31.5 documents and a median of 7 documents. As expected, the number of different documents containing an answer for each question, as judged in the document ranking task, was far higher than the number of different documents containing the right answer as judged in the strict question answering task. Researchers doing post-evaluation analysis should therefore not assume that the set of documents having correct answers in the main series task is complete.

The submitted runs were scored using trec_eval, treating the contains-answer documents as the relevant documents. Unlike other QA evaluations, trec_eval rewards recall, so retrieving more documents with the same answer yields a higher score than retrieving a single document with that answer. Even though a factoid question requires only a single document containing an answer, a recall-based metric for document retrieval may still correlate with performance on the exact factoid QA task because some systems make use of the frequency of candidate answers in determining which candidate to select as the final answer.

Table 7 shows the R-Precision and mean average precision (MAP) scores for the best run for each of the top 13 groups. The runs for the three groups that participated in the document ranking task without participating in the main task are marked with a $*$. R-precision is the precision after retrieving the first R documents, where R is the number of relevant documents in the pool. We found a weak correlation between factoid accuracy and R-precision (Pearson's $\rho = 0.53$, with a 95% confidence interval of [0.38,1.0]).

4 Relationship Task

AQUAINT analysts defined a "relationship" 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 have been noted including financial, movement of goods, family ties, communication pathways, organizational ties, co-location, common interests, and temporal. Recognition of when support for a suspected tie is lacking and determining whether the lack is because the tie doesn't exist or is being hidden/missed is a major concern. The analyst needs sufficient information to establish confidence in any support given. The particular relationships of interest depend on the context.

In the relationship task, 4 military analysts created 25 TREC-like topic statements that set a context. Each topic

Figure 2: Sample relationship topic and nuggets of evidence.

The analyst is concerned with arms trafficking to Colombian insurgents. Specifically, the analyst would like to know of the different routes used for arms entering Colombia and the entities involved.	
Vital?	Nugget
vital	Weapons are flown from Jordan to Peru and air dropped over southern Columbia
okay	Jordan denied that it was involved in smuggling arms to Columbian guerrillas
vital	Jordan contends that a Peruvian general purchased the rifles and arranged to have them shipped to Columbia via the Amazon River.
okay	Peru claims there is no such general
vital	FARC receives arms shipments from various points including Ecuador and the Pacific and Atlantic coasts.
okay	Entry of arms to Columbia comes from different borders, not only Peru

Table 8: Average F($\beta = 3$) scores for the relationship task for each run. Manual runs are marked with a $*$.

Run Tag	Submitter	F($\beta = 3$)
$*$ clr05r1	CL Research	0.276
csail2005a	MIT	0.228
$*$ clr05r2	CL Research	0.216
$*$ lcc05rel1	Language Computer Corp.	0.190
$*$ lcc05rel2	Language Computer Corp.	0.171
uams05s	Univ. of Amsterdam	0.120
uams05l	Univ. of Amsterdam	0.119
$*$ CMUJAVSEMMAN	Carnegie Mellon Univ.	0.096
$*$ UIowa05QAR01	Univ. of Iowa	0.086
CMUJAVSEM	Carnegie Mellon Univ.	0.061

was specific about the type of relationship being sought. The topic ended with a question that was either a yes/no question, which was to be understood as a request for evidence supporting the answer, or a request for the evidence itself. The system response was a set of information nuggets that provided evidence for the answer, in the same format as the Other questions in the main task. Manual processing was allowed.

4.1 Evaluation

The relationship topics were evaluated using the same methodology as the Other questions in the main task. A system's response for a relationship topic was an unordered set of [*doc-id*, *answer-string*] pairs. Each string was presumed to contain evidence for the answer to the question(s) in the topic. The system responses were judged by 5 assessors who were not the same as those who created the topics. An example topic and associated nuggets of evidence are given in Figure 2.

Each nugget created by the assessor was a piece of evidence for the answer, with nuggets marked as either vital or non-vital. Precision, recall, and F measure were calculated for each relationship topic as for the Other questions, and the final score for the relationship task was the average F($\beta = 3$) score over 25 topics. Table 8 gives the average F($\beta = 3$) score for each of the 10 runs submitted for the relationship task. Runs that included manual processing are marked with a $*$.

5 System Approaches

The overall approach taken for answering factoid questions has remained unchanged for the past several years. Systems generally determine the expected answer type of the question, retrieve documents or passages likely to contain answers to the question using important question words and related terms as the query, and then perform a match between the question words and retrieved passages to generate a set of candidate answers. The candidate answers are then ranked to find the most likely answer.

For the document/passage retrieval phase, most systems simply appended the target to the query. This was an effective strategy since in all cases the target was the correct domain for the question, and most of the retrieval methods used treat the query as a simple set of keywords. More and more systems are exploiting the size and redundancy on the Web to help find the answer. Some search the Web to find the answer, and then project the answer back to the AQUAINT corpus to find a supporting document. Others find candidate answers in the AQUAINT corpus and then use the Web to rerank the answers.

Most groups use their factoid-answering system for list questions, returning the top-ranked n candidate answer strings as the final answer list. The number of answer strings returned was a fixed number or was based on some threshold score for the string. Some groups went further and used their initial list items as seeds to find additional items. Systems generally used the same techniques as were used for TREC 2003's definition questions to answer the Other and relationship questions. Most systems first retrieve passages about the target using a recall-oriented retrieval search. Subsequent processing reduces the amount of material returned. Systems also looked to eliminate redundant information, using either word overlap measures or document summarization techniques. The output from the redundancy-reducing step was then returned as the answer for the question.

6 Future of the QA Track

Even though the main task in the TREC 2005 QA task was supposed to be essentially the same as the 2004 task, system performance was noticably lower in 2005 than in 2004. The 2005 task was more difficult because of the introduction of EVENT type targets and the increased dependencies between questions in a series; questions contained a greater number of anaphoric references, many of which referred to answers to previous questions in the series.

The introduction of event targets had additional ramifications for NIST assessors judging the system responses; it became clear that the assessors would not (and should not) ignore the time frame implied by the series when judging the correctness of answers. Before 2005, assessors assumed that the document returned with an answer would be used to set the time frame for the question, because questions were primarily phrased in the present tense without specifying an explicit time frame. Under those guidelines, *Who is the President of the United States?* would be answered correctly by "Ronald Reagan" if the document was from 1987, even if more recent documents supported "George Bush" or "Bill Clinton" as the answer. However, event type targets and temporally-constrained questions require that questions be interpreted in the temporal context that is explicit in the question or implicit in the series.

The main task for the TREC 2006 QA track will be the same as the main task in 2005, except that the implicit time frame for questions phrased in the present tense will be the date of the last document in the document collection, rather than the document returned with the answer. Thus, systems will be required to give the most up-to-date answer supported by the document collection. This brings the TREC QA task closer in line with question-answering in the real world, where users would want the best answer to their question in the document set (rather than just any answer found in any document). The evaluation of the question series in 2006 will also weight each of the 3 question types equally. The document ranking task will not be repeated in 2006, since little was learned from it. However, the relationship task will be repeated and modified to allow clarification forms like the ones used in the 2005 HARD task.

References

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

[2] Ellen M. Voorhees. Overview of the TREC 2004 question answering track. In *Proceedings of the Thirteenth Text REtreival Conference (TREC 2004)*, pages 52–62, 2005.