Overview of the TREC 2003 Question Answering Track

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

Abstract

The TREC 2003 question answering track contained two tasks, the passages task and the main task. In the passages task, systems returned a single text snippet in response to factoid questions; the evaluation metric was the number of snippets that contained a correct answer. The main task contained three separate types of questions, factoid questions, list questions, and definition questions. Each of the questions was tagged as to its type and the different question types were evaluated separately. The final score for a main task run was a combination of the scores for the separate question types.

This paper defines the various tasks included in the track and reports the evaluation results. Since the TREC 2003 track was the first time for significant participation in the definition and list subtasks, the paper also examines the reliability of the evaluation for these tasks.

TREC introduced the first question answering (QA) track in TREC-8 (1999). The goal of the track is to foster research on systems that retrieve answers rather than documents in response to a question, with particular emphasis on systems that can function in unrestricted domains. The tasks in the track have evolved over the years to focus research on particular aspects of the problem deemed important to improving the state-of-the-art.

The task in the original QA tracks required systems to return text snippets drawn from a large corpus of newspaper articles in response to closed-class or factoid questions such as *Who invented the paper clip?*. Each response was judged by a human assessor; a response was marked correct if an answer to the question was contained within the snippet. Unfortunately, the relative effectiveness of different systems was masked by the fact that two different snippets could both contain a correct answer while one was a significantly better response then the other [4]. To force systems to demonstrate their ability to locate the actual answer, the TREC 2002 task required systems to return *exact answers*, text strings consisting of a complete answer and nothing else. Strings that contained a right answer with additional text were judged to be "inexact" and did not contribute to a system's score.

Pinpointing the precise extent of an answer is a more difficult problem than finding a text segment that contains an answer, and there are applications of QA technology that do not require this extra step. To provide a forum for research groups interested in these applications, the TREC 2003 track included a "passages" task that allowed text segments containing answers to be returned. The other task in the track, the main task, required exact responses. While the test set of questions for the passages task contained only factoid questions, the main task contained list and definition questions as well as factoid questions. Each question type was evaluated separately, and the final score for a main task run was a combination of the scores for the three questions types.

This paper provides an overview of the results of the TREC 2003 QA track. The first two sections describe the two tasks in the track and present the evaluation results for the tasks. Since the TREC 2003 track was the first time for significant participation in the definition and list subtasks, Section 4 examines the reliability of the evaluation used for these tasks. This analysis demonstrates that the evaluation results for the definition task must be interpreted with care as using different assessors can cause substantial changes in the relative evaluation scores. Using more questions in the definition test set should increase the stability of the evaluation.

1 The Passages Task

The passages task tested a system's ability to find an answer to a factoid question within a relatively short (250 characters) span of text. In contrast to the first years of the QA track, each text span returned by the system was required to be an extract from a document in the corpus—results files submitted to NIST contained the offset from the beginning of the document of the first character in the span plus the span length rather than actual strings. Requiring

Table 1: Evaluation scores for the best passages task run from each group that submitted a passages run.

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
LCCpass03	Language Computer Corp.	0.685	0.381	0.800
nuslamp03a	National University of Singapore (Lee)	0.419	0.156	0.333
uwmtCQ2	University of Waterloo (MultiText)	0.351	—	—
umassql	University of Massachusetts	0.201	—	—
answfind1	Macquarie University	0.191	—	—
Saarland	Saarland University	0.169	0.097	0.367
IITBQA1	Indian Institute of Technology Bombay	0.133	0.045	0.100
clr03p2	CL Research	0.119	0.109	0.233
UAmsT03P1	University of Amsterdam	0.111	0.128	0.333
pircsqa3	Queens College, CUNY	0.097	0.000	0.000
NSIR	University of Michigan	0.085	0.075	0.100

extracts rather than allowing arbitrary 250-character strings resolves many of the issues surrounding correct answers contained within very poor responses. As in previous years, all processing was required to be completely automatic with no changes to the system permitted once the test questions were released.

The document collection used as the source of answers was the same collection used in the TREC 2002 QA track, the AQUAINT Corpus of English News Text. This collection consists of documents from three different sources: the AP newswire from 1998–2000, the New York Times newswire from 1998–2000, and the (English portion of the) Xinhua News Agency from 1996–2000. There are approximately 1,033,000 documents and 3 gigabytes of text in the collection. The test set of questions contained 413 questions drawn from AOL and MSNSearch logs[1]. Thirty of the questions have no known correct answer in the document collection.

A passages task run consisted of exactly one response for each of the test questions. A response was either a specification of a document extract or the string "NIL" used to indicate the system's belief that there was no correct answer in the collection. NIL was the (only) correct response for the 30 questions with no known answer in the document collection. Each extract was independently judged as correct, unsupported, or incorrect by two human assessors. When the two judgments differed, an adjudicator made the final decision. An extract was judged correct if it contained a right answer to the question, if the document from which it was drawn made it clear that it was a right answer, and if the extract was responsive. Extracts that contain multiple entities of the same semantic category as the correct answer but do not indicate which of those entities is the actual answer (e.g., an extract containing multiple names in response to a who question) are not responsive. Extracts that do not include appropriate units (e.g., "500" instead of "500 meters") are also not responsive. Finally, extracts must refer to the famous entity itself and not imitations, copies, and the like to be responsive to questions about a famous entity (e.g., extracts about the Taj Mahal casino are not responsive to questions regarding "the Taj Mahal") [6]. An extract was judged unsupported if it contained a right answer and was responsive, but the document from which it was drawn does not indicate that it is a right answer. Otherwise, an extract was judged as incorrect.

The main evaluation score for a passages task run 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. Precision of recognizing no answer is the ratio of the number of times NIL was returned and correct to the number of times it was returned; recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct (30).

Twenty-one passages task runs from eleven different participating groups were submitted to the QA track. Table 1 gives evaluation results for the most accurate run from each group. The table includes the run tag, the group that submitted the run, and the accuracy, NIL precision, and NIL recall scores. The NIL precision and recall scores are reported as "—" if a run never returned NIL.

[1] The logs from which questions were drawn were generously donated by Abdur Chowdhury of AOL and Sue Dumais of Microsoft Research.

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

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
LCCmainE03	Language Computer Corp.	0.700	0.381	0.800
lexiclone92	LexiClone	0.622	—	—
nusmml03r1	National University of Singapore (Yang)	0.562	0.160	0.400
isi03a	University of Southern California, ISI	0.337	0.071	0.200
IBM2003a	IBM Research (Prager)	0.298	0.082	0.233
MITCSAIL03b	Massachusetts Institute of Technology	0.295	0.258	0.267
uwbqitekat03	University of Wales, Bangor	0.259	0.092	0.967
Albany03I2	University of Albany	0.240	0.109	0.200
irstqa2003w	ITC-irst	0.235	0.121	0.267
BBN2003B	BBN	0.208	0.068	0.100
FDUT12QA1	Fudan University	0.194	0.077	0.233
ntt2003qam1	NTT Communication Science Labs	0.150	0.090	0.267
MITRE2003A	MITRE Corp.	0.148	0.000	0.000
ICTQA2003C	Chinese Academy of Sciences (CAS-ICT)	0.145	0.105	0.467
UAmsT03M2	University of Amsterdam	0.145	0.273	0.100

2 The Main Task

The main task of the QA track involved three types of questions, factoids, lists, and definitions. Each question was tagged as to its type, and the response formats and evaluation methods differed for each type. The factoid questions for the main task were the same set of 413 questions used in the passages task, and the AQUAINT corpus was used for all subtasks. As in the passages task, completely automatic processing was required.

Fifty-four main task runs from 25 different groups were submitted to the track. Three groups, CL Research, Language Computer Corp., and the University of Amsterdam, submitted both passages and main task runs.

2.1 Factoids

The factoid component of the main task was very similar to the passages task. As noted, the same document and question set were used as in the passages task, and systems returned exactly one response for each factoid question. For the main task, however, systems were required to return an exact answer rather than an extract containing an answer. The answer strings returned by the systems were not required to be an extract from a document; the response for a question was of the form `query-id run-tag doc-id answer-string`. If the system believed there was no correct response in the document collection, `doc-id` was set to NIL and `answer-string` was empty.

Responses were judged as in the passages task except a fourth judgment, not exact, could also be assigned. A judgment was assigned to a response in the following order:

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

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

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

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

The score for the factoid component of the main task was accuracy, the fraction of responses judged correct. Table 2 gives evaluation results for the factoid component corresponding to Table 1 for passages. The table shows the most accurate run for the factoid subtask for each of the top 15 groups, and includes the the accuracy, NIL precision, and NIL recall scores.

1915: List the names of chewing gums.			
Stimorol	Orbit	Winterfresh	Double Bubble
Dirol	Trident	Spearmint	Bazooka
Doublemint	Dentyne	Freedent	Hubba Bubba
Juicy Fruit	Big Red	Chiclets	Nicorette

Figure 1: Answer list for list question 1915 — names of chewing gums found within the AQUAINT corpus.

2.2 Lists

The list task was offered in both the TREC 2001 and the TREC 2002 QA tracks. However, participation in the task has been quite limited as groups concentrated on the main (factoid) task. Encouraging additional participation in other subtasks was the primary motivation for a combined main task in TREC 2003.

The list task requires systems to assemble an answer from information located in multiple documents. In TREC, a list question asks for different instances of a particular kind of information to be retrieved, such as *List the names of chewing gums*. List questions can be thought of as a shorthand for asking the same factoid question multiple times; the set of answers that satisfy the factoid question is the appropriate response for the list question. A system's response to a list question was an unordered set of [*document-id*, *answer-string*] pairs such that each *answer-string* was considered an instance of the requested type.

Unlike the previous two times the list task was run in TREC, this year's list questions did not specify a target number of instances to return. Instead, system were expected to return all of the correct, distinct answers contained in the document collection. Different questions had different numbers of distinct answers, ranging from a low of 3 (*What Chinese provinces have a McDonald's restaurant?*) to a high of 44 (*Which countries were visited by first lady Hillary Clinton?*).

The 37 list questions used in the list subtask were constructed by NIST assessors. The assessors were instructed to construct questions whose answers would be a list of entities (people, places, dates, numbers) such that the list would not likely be found in a reference work such as a gazetteer or almanac. They searched the document collection using the PRISE search engine to find as complete a list of instances as possible. A single document could contain multiple instances, and the same instance might be repeated in multiple documents. After the participants' results were returned to NIST, the assessor who created the question judged the set of responses for that question. If a system found a correct response that the assessor had not found during question development, the system's answer was added to the list of known instances for that question. The final list of known instances was considered the correct answer to the question and was used to score the systems' responses. Figure 1 shows the final list of answers for the chewing gum question.

Judgments of incorrect, not supported, inexact, or correct were made individually for each [*document-id*, *answer-string*] pair as for the factoid subtask. The assessor was given one run's list at a time, and while judging for correctness he also marked a set of responses as distinct. The assessor arbitrarily chose any one of a set of equivalent responses to mark as the distinct one, and marked the remainder as not distinct. Only correct responses could be marked distinct.

The instance precision (IP) and instance recall (IR) for a list question can be computed from the final answer list and the assessor judgments. Let S by the size of the final answer list (i.e., the number of known answers), D be the number of correct, distinct responses returned by the system, and N be the total number of responses returned by the system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined using the F measure with equal weight given to recall and precision:

$$F = \frac{2 * IP * IR}{(IP + IR)}$$

The score for the list component was the average F score over the 37 list questions. Table 3 gives the average F scores for 15 of the main task submissions. The table gives the run with the best list component score for each of the top 15 groups.

Table 3: Average F scores for the list question subtask. Scores are given for the best run from the top 15 groups.

Run Tag	Submitter	F
LCCmainS03	Language Computer Corp.	0.396
nusmml03r2	National University of Singapore (Yang)	0.319
MITCSAIL03c	Massachusetts Institute of Technology	0.134
isi03a	University of Southern California, ISI	0.118
BBN2003B	BBN	0.097
Albany03I4	University of Albany	0.096
ICTQA2003A	Chinese Academy of Sciences (CAS-ICT)	0.091
FDUT12QA1	Fudan University	0.088
IBM2003c	IBM Research (Prager)	0.077
irstqa2003w	ITC-irst	0.076
MITRE2003A	Mitre Corp.	0.069
UAmsT03M1	University of Amsterdam	0.054
CMUJAV2003	Carnegie Mellon University (Javelin)	0.052
lexiclone92	LexiClone	0.048
ntt2003qam1	NTT Communication Science Labs	0.040

2.3 Definitions

Definition questions are questions such as *Who is Colin Powell?* or *What is mold?*. Definition questions occur relatively frequently in logs of web search engines [4] suggesting they are an important type of question. However, evaluating systems that answer definition questions is much more difficult than evaluating systems that answer factoid questions because it is no longer useful to judge a system response as simply right or wrong. Assigning partial credit to a system response requires some mechanism for matching the concepts in the desired response to the concepts present in the system's response. The issues are similar to those that arise in the evaluation of machine translation and automatic summarization.

A small pilot evaluation of definition questions was held as part of the ARDA AQUAINT program in the fall of 2002 [5]. The lessons learned from the pilot were used to help define the definitions component of the main task, which was the first first large scale evaluation of definition questions.

The definition question test set contained 50 questions. The questions were drawn from the same set of search engine logs that the factoid questions were drawn from. Assessors selected a question from the log, and searched the document collection for information about the target. The final set of questions contained 30 questions for which the target was a (perhaps fictional) person (e.g., Vlad the Impaler, Andrea Bocceli, Ben Hur), 10 questions for which the target was an organization (e.g., Bausch & Lomb, Friends of the Earth, Freddie Mac), and 10 questions for which the target was some other thing (e.g., a golden parachute, feng shui, TB). If the question had a qualification in the log, the qualification remained in the test set (e.g., *What is Ph in biology?*).

In evaluations such as TREC, questions are asked in isolation. This is not much of an issue for factoid questions, but becomes much more of an issue for definition questions. Without any idea of who the questioner is and why he or she is asking the question it is essentially impossible for a system to decide what level of detail in a response is appropriate—presumably an elementary-school-aged child and a nuclear physicist should receive different answers for at least some questions. To provide some guidance for the system developers, the following scenario was assumed for definition questions:

> The questioner is an adult, a native speaker of English, and an "average" reader of US newspapers. In reading an article, the user has come across a term that they would like to find out more about. They may have some basic idea of what the term means either from the context of the article (for example, a bandicoot must be a type of animal) or basic background knowledge (Ulysses S. Grant was a US president). They are not experts in the domain of the target, and therefore are not seeking esoteric details (e.g., not a zoologist looking to distinguish the different species in genus Perameles).

As in the list task, a system returned an unordered set of [*document-id*, *answer-string*] pairs as a response for a definition question. Each string was presumed to be a facet in the definition of the target. There were no limits placed on either the length of an individual answer string or on the number of pairs in the list, though systems were penalized for retrieving extraneous information.

Judging the quality of the systems' responses was done in two steps. In the first step, all of the answer-strings from all of responses were presented to the assessor in a single (long) list. Using these responses and his own research done during question development, the assessor first created a list of "information nuggets" about the target. An information nugget was defined as a fact for which the assessor could make a binary decision as to whether a response contained the nugget. At the end of this step, the assessor decided which nuggets were vital—nuggets that must appear in a definition for that definition to be good. The assessor went on to the second step once the nugget list was created. In this step the assessor went through each of the system responses in turn and marked where each nugget appeared in the response. If a system returned a particular nugget more than once, it was marked only once.

Figure 2 shows an example of how one response was judged for the question *What is a golden parachute?*. The top of the figure shows the nugget list developed by the assessor with the vital nuggets indicated. The bottom of the figure shows a system response with the nugget numbers of the nuggets that were assigned to the response to the left of the individual strings that match the nugget. For example, the assessor matched nugget 1 to item b, and nugget 4 to item i.

The judging of the systems' responses was designed in this way to make the evaluation depend only on the *content* of a system response, and not on the particular structure of a system response. Assessors ignored wording differences, making conceptual matches between the system responses and their nuggets, not syntactic matches (that's why human assessors were needed!). A single answer string within the system's response was allowed to match multiple nuggets such as for string 'a' in Figure 2). The design does depend on the assessor nuggets truly being atomic so that a nugget isn't split across system's strings. Occasionally in the judging the assessors found that a nugget wasn't really atomic; in these cases they tended to assign the nugget to a string that had the main part of the concept.

Given the judgments as described above, it is straightforward to compute the nugget recall of a response: it is simply the ratio between the number of correctly retrieved nuggets to the number of nuggets on the assessor's list. But the corresponding measure of nugget precision, the ratio between the number of nuggets correctly retrieved to the total number of nuggets retrieved, is problematic since the correct value for the denominator is unknown. A trial evaluation prior to the pilot showed that assessors found enumerating *all* concepts represented in a response to be so difficult as to be unworkable. For example, how many information units are contained in the string "`Oh, Eaton also has a new golden parachute clause in his contract.''`? Using only nugget recall as a final score is untenable since systems would not be rewarded for being selective. Retrieving the entire document collection is guaranteed to give a perfect recall score for every question.

Borrowing from the evaluation of summarization systems [1], we can use length as a crude approximation to precision. A length-based measure captures the intuition that users would prefer the shorter of two definitions that contain the same concepts. The length-based measure used in both the pilot and the TREC track gives a system an allowance of 100 (non-white-space) characters for each correct nugget it retrieved. The precision score is set to one if the response is no longer than this allowance. If the response is longer than the allowance, the precision score is downgraded using the function precision $= 1 - \frac{\text{length}-\text{allowance}}{\text{length}}$.

Remember that the assessors marked some nuggets as vital and the remainder are not vital. The non-vital nuggets act as a "don't care" condition. That is, systems should be penalized for not retrieving vital nuggets, and penalized for retrieving items that are not on the assessor's nugget list at all, but should be neither penalized nor rewarded for retrieving a non-vital nugget. To implement the don't care condition, nugget recall is computed only over vital nuggets, while the character allowance in the precision computation is based on both vital and non-vital nuggets. The recall for the example in Figure 2 is thus 3/3, and the character allowance is 500.

The final score for a definition response was computed using the F-measure as it was for list questions. For the definition questions, however, the β parameter was set to five indicating that recall was five times as important as precision. The value of five is arbitrary, but reflects both the emphasis given to recall by the assessors in the pilot version of the task, and acknowledges the crudeness of the length approximation to true precision. Figure 3 shows the complete computation of the F($\beta = 5$) score for a single definition question. The score for the definition component of the main task was the average F($\beta = 5$) score over the 50 definition questions.

Table 4 gives the average F($\beta = 5$) score for the best scoring run from a group for the top 15 groups. The table also gives the average length of a definition response for the run.

1	vital	Agreement between companies and top executives
2	vital	Provides remuneration to executives who lose jobs
3	vital	Remuneration is usually very generous
4		Encourages execs not to resist takeover beneficial to shareholders
5		Incentive for execs to join companies
6		Arrangement for which IRS can impose excise tax

a) nugget list

Nuggets		Answer Strings
2, 3	a.	`The arrangement, which includes lucrative stock options, a hefty salary, and a "golden parachute" if Gifford is fired,`
1	b.	`Oh, Eaton also has a new golden parachute clause in his contract.`
	c.	`But some, including many of BofA's top executives, joined the 216 and cashed in their "golden parachute" severance packages.`
	d.	`The big payment that Eyler received in January was intended as a "golden parachute"`
	e.	`Cotsakos' contract included a golden parachute big enough to make a future sale of the company more likely`
	f.	`syndication, the golden parachute for production companies`
6	g.	`But if he quits or is dismissed during the two years after the merger, he will be paid $24.4 million, with DaimlerChrysler paying the "golden parachute" tax for him and the taxes on the compensation paid to cover the tax.`
	h.	`If he left, On leaving, O'Neill could would be able to collect a golden parachute package providing three years of salary and bonuses, stock and other benefits.`
4	i.	`After the takeover, as jobs disappeared and BofA's stock tumbled, many saw him as a bumbler who had sold out his bank, walking away with a golden parachute that gives him $5 million a year for the rest of his life.`
	j.	`And after BofA disclosed that he had a golden parachute agreement giving him some $50 million to $100 million if he left following the merger, he sent a voice mail message to bank employees that he intended to stay.`

b) system response

Figure 2: Assessor annotation of a sample response for definition question 1905 *What is a golden parachute?*

Let r be the number of vital nuggets returned in a response;
a be the number of acceptable (non-vital but on list) nuggets returned in a response;
R be the total number of vital nuggets in the assessor's list;
len be of the number of non-white space characters in an answer string summed over all answer strings in the response;

Then

$$\text{recall} = r/R$$

$$\text{allowance} = 100 * (r + a)$$

$$\text{precision} = \begin{cases} 1 & \text{if len} < \text{allowance} \\ 1 - \frac{\text{len} - \text{allowance}}{\text{len}} & \text{otherwise} \end{cases}$$

$$F(\beta = 5) = \frac{26 * \text{precision} * \text{recall}}{25 * \text{precision} + \text{recall}}$$

Figure 3: Computation of the F($\beta = 5$) score for a definition question.

Table 4: Average F($\beta = 5$) scores for the definition questions subtask. Scores are given for the best run from the top 15 groups. Also given is the average length of a response for the run measured in non-white-space characters.

Run Tag	Submitter	F($\beta = 5$)	Ave Length
BBN2003C	BBN	0.555	2059.20
nusmml03r2	National University of Singapore (Yang)	0.473	1478.74
isi03a	University of Southern California, ISI	0.461	1404.78
LCCmainS03	Language Computer Corp.	0.442	1407.82
cuaqdef2003	Univ. of Colorado/Columbia Univ.	0.338	1685.60
irstqa2003d	ITC-irst	0.318	431.26
UAmsT03M1	University of Amsterdam	0.315	2815.08
MITCSAIL03a	Massachusetts Institute of Technology	0.309	620.28
shef12simple	University of Sheffield	0.236	338.42
UIowaQA0303	University of Iowa	0.231	3039.26
CMUJAV2003	Carnegie Mellon University (Javelin)	0.216	182.34
FDUT12QA3	Fudan University	0.192	203.54
piq002	University of Pisa	0.185	89.52
IBM2003b	IBM Research (Prager)	0.177	223.16
ntt2003qam1	NTT Communication Science Labs	0.169	2219.24

Table 5: Component scores and final combined scores for main task runs. Scores are given for the best run from the top 15 groups.

Run Tag	Submitter	Component Score Factoid	List	Def	Final Score
LCCmainS03	Language Computer Corp.	0.700	0.396	0.442	0.559
nusmml03r2	National University of Singapore (Yang)	0.562	0.319	0.473	0.479
lexiclone92	LexiClone	0.622	0.048	0.159	0.363
isi03a	University of Southern California, ISI	0.337	0.118	0.461	0.313
BBN2003C	BBN	0.206	0.097	0.555	0.266
MITCSAIL03a	Massachusetts Institute of Technology	0.293	0.130	0.309	0.256
irstqa2003w	ITC-irst	0.235	0.076	0.317	0.216
IBM2003c	IBM Research (Prager)	0.298	0.077	0.175	0.212
Albany03I2	University of Albany	0.240	0.085	0.146	0.178
FDUT12QA3	Fudan University	0.191	0.086	0.192	0.165
UAmsT03M1	University of Amsterdam	0.136	0.054	0.315	0.160
shef12simple	University of Sheffield	0.138	0.029	0.236	0.135
CMUJAV2003	Carnegie Mellon University (Javelin)	0.133	0.052	0.216	0.134
ICTQA2003C	Chinese Academy of Sciences (CAS-ICT)	0.145	0.091	0.149	0.133
uwbqitekat03	University of Wales, Bangor	0.259	0.000	0.000	0.130

2.4 Combined results

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

$$\text{FinalScore} = 1/2 * \text{FactoidScore} + 1/4 * \text{ListScore} + 1/4 * \text{DefScore}.$$

Since each of the component scores ranges between 0 and 1, the final score is also in that range. The final score emphasizes the factoid component, which represented the largest number of questions and is the task people are most familiar with. The weight for the other components was made large enough to encourage participation in those subtasks.

Table 5 shows the combined scores for the best run for each of the top 15 groups.

3 Question Answering Techniques

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

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

Answering definition questions generally involved using different techniques than those used for factoid questions. Since the definition task emphasized nugget recall and did not require "exact" answers, most systems first retrieved passages about the target using a recall-oriented search. Subsequent processing reduced the amount of material returned. Many systems used pattern-matching to locate definition-content in text. These patterns, such as looking for copular constructions and appositives, were either hand-constructed or learned from a training corpus. Systems also looked to eliminate redundant information, using either word overlap measures or document summarization techniques. The output from this step was then returned as the definition of the target.

4 Analysis of the Evaluation

Since this was the first year for the definition task, and the first year for significant participation in the list task, it is appropriate to assess how well the definition and list question evaluations measure system effectiveness. Of course, in many respects this is a chicken-or-the-egg proposition since the whole point of the evaluation is to define what constitutes "good system effectiveness". In general, a good evaluation will assign higher scores to responses that are intuitively "better" and lower scores to responses that are intuitively "worse". Evaluation results should also provide guidance to system builders as to how their system can be improved.

There are at least two aspects to the quality of an evaluation, fidelity and reliability. Fidelity is the extent to which the evaluation measures what it is intended to measure. Since an evaluation task is an abstraction of a real user's task, fidelity is the extent to which the abstraction captures (some of) the issues of the real task. Reliability is the extent to which an evaluation result can be trusted. Since TREC evaluations are comparative, reliability amounts to the likelihood that an evaluation ranks a better system ahead of a worse system.

This section analyzes the definition and list question evaluations with respect to fidelity and reliability. The first subsection addresses the fidelity of the definition task by examining the effect varying the relative importance of recall and precision has on the evaluation results. The following two subsections empirically estimate the size of the difference required between scores to confidently conclude that two runs are different.

4.1 Definition task fidelity

The F measure is actually a family of measures based on the value of β that controls the relative importance of recall and precision. The general form of the F measure is

$$F = \frac{\beta^2 PR}{(\beta^2 + 1)P + R}$$

Such parameterization allows the measure to be finely tuned to the expectations of the user, but also means that the expectations of the user must be known to properly evaluate the results.

The pilot evaluation of definition questions contained an additional evaluation of the system responses not included in the TREC track [5]. In this "holistic" evaluation, the assessor assigned one score between 0 and 10 to the content of the entire response. This data provided a target for the more quantitative evaluation adopted by the TREC track. Using a β value of five gave a good correlation between the systems ranked by quantitative score and the systems ranked by the holistic score.

The pilot was small, however, and it is unclear whether the mythical average user would so strongly prefer recall. Such a strong emphasis on recall may not encourage systems to be selective enough in the amount of information they return. The results in Table 4 show a trend—but not a strict ordering—for longer responses to receive higher scores. One way to determine how selective the definition question systems are is to compare the results to a baseline run that returns all the sentences in the corpus that contain the target of the definition question. A slightly smarter baseline that returns a sentence only if its overlap with an already-retrieved sentence is small enough was implemented by Jinxi Xu of BBN and the results given to NIST. The results of this sentence baseline were judged by the assessors as part of the regular definition question judging.

Table 6 gives the F(β) evaluation scores for the set of 15 runs from Table 4, plus the sentence baseline (SENT-BASE), for $\beta = 1 \ldots 5$. The table is sorted by decreasing F($\beta = 5$) scores as in Table 4. The sentence baseline is very effective when recall receives strong emphasis. For F($\beta = 5$), the baseline is the second ranked run in the table and ranked 4 of 55 over all runs (only the three BBN runs have a higher F($\beta = 5$) score). For F($\beta = 1$), the baseline is ranked eleventh in the table and 25 out of 55 overall. Other runs also experience a change in relative effectiveness as the importance of recall varies. (Note that the best run for a group for a β value other than five may not be shown in the table.)

The results in Table 6 do not show which F(β) is a better measure for the definition task, only that the evaluation of runs differs depending on the relative importance given to recall and precision. The pilot evaluation demonstrated that there are at least some people for whom recall plays a very dominant role, and for them the F($\beta = 5$) measure is appropriate. Other users may be better represented by the measures that reward the more selective systems.

Table 6: Average F(β) scores for the definition questions subtask. Scores are given for each of the runs in Table 4 plus a sentence-based baseline run. The β parameter varies from 1 (recall and precision have equal importance) to 5 (recall five times as important as precision). The table is sorted by the F($\beta = 5$) scores.

	F(β) Score				
Run Tag	$\beta = 1$	$\beta = 2$	$\beta = 3$	$\beta = 4$	$\beta = 5$
BBN2003C	0.310	0.423	0.493	0.532	0.555
SENT-BASE	0.205	0.315	0.400	0.456	0.493
nusmml03r2	0.261	0.360	0.421	0.454	0.473
isi03a	0.270	0.353	0.409	0.442	0.461
LCCmainS03	0.332	0.374	0.408	0.429	0.442
cuaqdef2003	0.187	0.256	0.299	0.324	0.338
irstqa2003d	0.310	0.310	0.314	0.316	0.318
UAmsT03M1	0.163	0.215	0.259	0.292	0.315
MITCSAIL03a	0.296	0.298	0.304	0.307	0.309
shef12simple	0.195	0.211	0.224	0.232	0.236
UIowaQA0303	0.156	0.188	0.210	0.223	0.231
CMUJAV2003	0.246	0.223	0.218	0.217	0.216
FDUT12QA3	0.214	0.196	0.193	0.192	0.192
piq002	0.234	0.198	0.189	0.186	0.185
IBM2003b	0.209	0.186	0.180	0.178	0.177
ntt2003qam1	0.145	0.151	0.159	0.165	0.169

4.2 Definition task reliability

In the evaluation methodology used in TREC, one system is considered to be more effective than another if the evaluation score computed for the output of the first system is greater than the evaluation score computed for the output of the second system. However since *all* measurements have some (unknown) error associated with them, there is always a chance that such a comparison can lead to the wrong result. An analysis of the reliability of an evaluation establishes bounds for how likely it is for a comparison to be in error.

The error in a definition question evaluation score arises from a variety of sources of noise. One source is mistakes made during the judgment process: the assessors are humans and humans will make mistakes, especially given the rate at which the NIST assessors are asked to process results. Another source of noise is the fact that humans have differing opinions as to the quality of a response. These differences of opinions are not mistakes, but legitimate differences in what the assessor considers to be acceptable. A third source of noise is the sample of questions used in the evaluation. Since in general system performance depends on the question that is asked, it may be that a better overall system evaluates as worse than another because of the particular sample of questions used in the test set. None of these sources of error is unique to the definition question task. All three sources exist in all of the evaluations in TREC.

The amount of noise contributed by mistakes in judging has been difficult to measure in other TREC tasks because results are pooled. That is, identical responses from different runs are represented only once in what the assessor judges, so there is no opportunity to be inconsistent. The results are not pooled for the definition (and list) task, however, because the assessor must judge the entire response as a unit for this task. Since some main task submissions differed only in the factoid component of the task, some of the 55 responses a definition assessor was asked to judge were exact copies of another response. These pairs of identical, separately judged responses provide the data for estimating an error rate due to assessor mistakes.

The main task runs contained 14 pairs of identical definition components (i.e., exactly the same for all 50 definition questions). The distribution of the number of questions that were judged differently over these 14 pairs was as follows: 0 questions judged differently, 1 pair; 1 question judged differently, 3 pairs; 2 questions, 1 pair; 3 questions, 4 pairs; 4 questions, 2 pairs; 8 questions, 2 pairs; 10 questions, 1 pair. Across the whole set of 14 pairs, 19 questions had some difference in judging, spread relatively uniformly across individual assessors. The differences in the average F($\beta = 5$) scores for the pairs of identical runs ranged from a low of 0.0 to a high of 0.043, and averaged 0.013.

The definition task was a brand new and more difficult task for the assessors. There is reason to believe the number

1	vital	provides remuneration to executives who lose jobs
2	vital	assures officials of rich compensation if lose job due to takeover
3	vital	contract agreement between companies and their top executives
4		aids in hiring and retention
5		encourages officials not to resist a merger
6		IRS can impose taxes

Figure 4: Information nuggets created by second assessor for question 1905 *What is a golden parachute?*.

of judging errors will decrease now that the assessors have experience with the task and the assessment system software is changed to better support this task. However, the number of errors will never be reduced to zero. Since runs that were absolutely identical had differences in average scores of 0.043 in this year's evaluation, distinct runs that differ by less than this amount (such as the `nusmml03r2`, `isi03a`, and `LCCmainS03` runs in Table 4) clearly must be considered equivalently effective.

Another source of noise is the differences in scores caused by differing opinions of assessors. Since assessor opinions as to correctness are known to differ, each question is judged by a single assessor so the judgments are internally consistent (modulo mistakes). However, to quantify the effect different opinions have on definition question scores, each question was independently judged by a second assessor. Since the second assessor did not research the target of the question during question development as the original assessor had, the second assessor was given the initial list of nuggets the original assessor created during question development as a starting point. The second assessor was free to modify that list in any way (add, delete, or alter nuggets) as he proceeded in the two stages of judging the question. The set of information nuggets created by the second assessor for the golden parachute question of Figure 2 is given in Figure 4.

Not surprisingly, there were definite differences across assessors. Some assessors listed many more nuggets—and many more vital nuggets—than other assessors. There was no universal agreement as to what kinds of information should be returned for, say, a "who is" question. Different assessors looked for different information for the same question, and the same assessor looked for different information for different questions. For example, a birth date was usually wanted for an important historical figure, but not for a contemporary sports figure.

To measure the effect of the differences of opinion of different assessors on the evaluation, the set of runs were scored using each of the two sets of judgments and ranked by decreasing average F($\beta = 5$) scores. The distance between the two rankings of runs was computed using a correlation measure based on Kendall's τ [2]. The τ score between the two rankings of runs was 0.848, which corresponds to 113 pairwise differences between the rankings (out of a possible 1485 pairwise difference since there were 55 runs being ranked). While most of the changes in the rankings were between two runs with small differences in F($\beta = 5$) scores as computed using the original assessor's judgments, eight of the differences were between runs with F scores that differed by more than 0.1. The largest difference in original F($\beta = 5$) scores between two runs that evaluated differently when using different judgments was 0.123

It would be nice to have a critical value for τ such that correlations greater than the critical value guarantee a quality evaluation. Unfortunately, no such value can exist since τ values depend on the set of runs being compared. In practice, we have considered correlations greater than 0.9 to indicate an acceptably stable evaluation [3]. The correlation between the rankings of the definition questions is decidedly smaller than 0.9, and the eight swaps with a difference of greater than 0.1 is especially worrisome. These findings suggest that the evaluation results of this first version of the definition evaluation need to be interpreted with great caution, and that steps should be taken in future years to increase the stability of the evaluation.

One way to increase the stability of an evaluation is to increase the number of questions that scores are computed over. Using larger numbers of questions has two effects: first, the sample of questions is larger and thus more likely to include different categories of questions, and second the law of large numbers says the averages will be closer to the true means. We can use the runs submitted to the track to empirically determine the relationship between the number of questions in a test set, the observed difference in scores (δ), and the likelihood that a single comparison of two definition tasks runs leads to the correct conclusion. Once established, the relationship can be used to derive the minimum difference in scores required for a certain level of confidence in the results given there are 50 questions in the test set. In theory, the same relationship can also predict the number of questions to have in the test set given

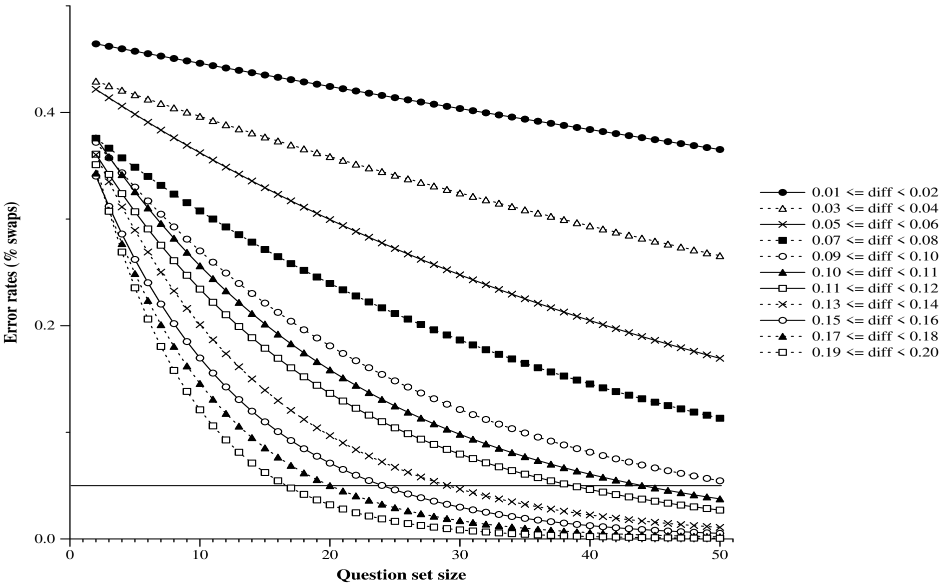

Figure 5: Stability of definition question evaluation as a function of question set size.

that you want a certain level of confidence in the comparison at a given minimum difference in scores. However, the amount of extrapolation involved in that computation makes the prediction of dubious value.

The core of the procedure for establishing the relationship is comparing the effectiveness of a pair runs on two disjoint question sets of equal size to see if the two sets disagree as to which of the runs is better. We define the error rate as the percentage of comparisons that result in a swap. Since the definition component used 50 questions, we can directly compute the error rate for question set sizes up to 25 questions. By fitting curves to the values observed for question set sizes up to 25, we can extrapolate the error rates to question sets up to 50 questions.

When calculating the error rate, the difference between two runs' F($\beta = 5$) scores is categorized into one of 21 bins based on the size of the difference. The first bin contains runs with a difference of less than 0.01 (including no difference at all). The next bin contains runs whose difference is at least 0.01 but less than 0.02. The limits for the remaining bins increase by increments of 0.01, with the last bin containing all runs with a difference of at least 0.2.

Each question set size from 1 to 25 is treated as a separate experiment. Within an experiment, we randomly select two disjoint sets of questions of the required size. We compute the F($\beta = 5$) score over both question sets for all runs (using the original assessor judgments), then count the number of times we see a swap for all pairs of runs using the bins to segregate the counts by size of the difference in scores. The entire procedure is repeated 50 times (i.e., we perform 50 trials), with the counts of the number of swaps kept as running totals over all trials. The ratio of the number of swaps to the total number of cases that land in a bin is the error rate for that bin.

The error rates computed from this procedure are then used to fit curves of the form $ErrorRate = A_1 e^{-A_2 S}$ where A_1 and A_2 are parameters to be estimated and S is the size of the question set. A different curve is fit for each different bin. The input to the curve-fitting procedure used only question set sizes greater than 1 since a single question is known to be very noisy. The largest bin (all differences greater than 0.2) is very flat and therefore difficult to fit, so no curve was fit for it.

The resulting extrapolated error rate curves are plotted in Figure 5. In the figure, the question set size is plotted on the x-axis and the error rate is plotted on the y-axis. The horizontal line in the graph in Figure 5 is drawn at an error rate of 5 %, a level of confidence commonly used in experimental designs. For question set sizes of 50 questions, there needs to be an absolute difference of at least 0.1 in the F($\beta = 5$) scores before the error rate is less than 5 %. Using this standard, swaps caused by different assessor opinions when the difference in F score is less than 0.1 are to be expected since the runs being compared must be considered equally effective. It is the eight swaps among runs with greater differences that are the concern.

Requiring a difference of at least 0.1 (or 0.123) in F($\beta = 5$) scores before considering two evaluation results to be different is necessary to have confidence in the conclusions, but also results in a fairly insensitive test. For example, if we take the run that ranks in the middle of the runs in Table 4, and add ± 0.1 to its score, its rank would range

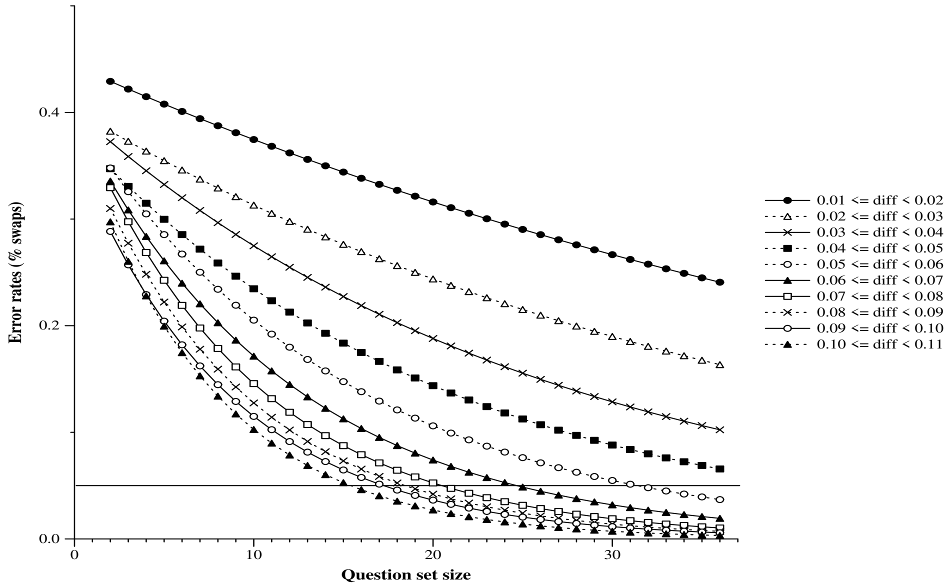

Figure 6: Stability of list question evaluation as a function of question set size.

anywhere from fifth to twelfth. More questions are needed in the test set to increase the sensitivity while remaining equally confident in the result.

4.3 List task reliability

As in the definition question evaluation, list task responses must also be judged as a unit, so an estimate of the error attributable to assessing mistakes can be derived from direct comparison of scores for identical submissions. There were ten pairs of runs that had identical list task components. Over these ten pairs of runs, one pair differed in the score assigned to three questions, four pairs differed in the scores assigned to two questions, four pairs differed in the scores assigned to one question, and one pair was assigned the same score for all questions. The largest difference in the average F score for a pair was 0.004, and the average difference across the ten pairs was 0.002.

The list questions were each judged by only one assessor, so no comparisons across assessors are possible. However, Figure 6 shows the plot of extrapolated error rates by question set size for the list task. Since the list component had only 37 questions, the error rates were directly computed up to question set size 18, and then extrapolated to question set size 36. The plot shows that the error rates decrease quickly as question set size increases: at 36 questions the error rate is less than 5% for a difference in F scores of at least 0.05.

These two tests show that this year's list task results are extremely stable. Remember, however, that the sensitivity analysis is dependent on the runs used to compute it. One reason the list task results appear as stable as they do is because in general the scores for the list task are very low. Only 11 of the 54 main task submissions had an average score for the list component that was greater than 0.1. Only 9 of the 37 questions had a median F score greater than 0.0. Clearly if the scores are similar regardless of which question you choose (in this case, all close to 0), you don't need many questions to converge to the average score.

5 Future of the QA Track

The TREC 2003 QA track offered the first large-scale evaluations of list and definition questions. The results demonstrated that these are challenging tasks for systems, and that they present challenges for evaluation. The proposed task for the TREC 2004 QA track will address these challenges by keeping a combined task, but reorganizing it somewhat so that more definition questions can be accommodated.

There will be approximately 100 targets for a definition. For each target, NIST assessors will define a set of factoid questions and a separate set of list questions. There will also be an implied "tell me more" question, which is to be interpreted as ”Tell me other interesting things about this target I don’t know enough to ask directly”. This last question is roughly equivalent to the definition questions in the TREC 2003 task. As an example, if the target were an author, the factoid questions might ask for things such as birth date, date of death, and nationality. A list question may ask for the names of the author’s works. Finally, the answer to the last question may include such items as one of the author’s books won a Newberry prize, and the author teaches at XYZ University.

Organized this way, the task allows for factoid and list questions as well as more definition questions. The organization also provides context for the factoid and list questions, an important element that the track has not yet successfully incorporated [4].

Acknowledgements

My thanks to Chris Buckley for fitting the error rate curves shown in Figures 5 and 6.

References

[1] Donna Harman and Paul Over. The DUC summarization evaluations. In *Proceedings of the International Conference on Human Language Technology*, 2002.

[2] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[3] Ellen M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, 2001.

[4] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Tenth Text REtreival Conference (TREC 2001)*, pages 42–51, 2002.

[5] Ellen M. Voorhees. Evaluating answers to definition questions. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), Volume 2*, pages 109–111, May 2003.

[6] Ellen M. Voorhees and Dawn M. Tice. The TREC-8 question answering track evaluation. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 83–105, 2000. NIST Special Publication 500-246. Electronic version available at `http://trec.nist.gov/pubs.html`.