The TREC-8 Question Answering Track Report

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899
ellen.voorhees@nist.gov

Abstract

The TREC-8 Question Answering track was the first large-scale evaluation of domain-independent question answering systems. This paper summarizes the results of the track by giving a brief overview of the different approaches taken to solve the problem. The most accurate systems found a correct response for more than 2/3 of the questions. Relatively simple bag-of-words approaches were adequate for finding answers when responses could be as long as a paragraph (250 bytes), but more sophisticated processing was necessary for more direct responses (50 bytes).

The TREC-8 Question Answering track was an initial effort to bring the benefits of large-scale evaluation to bear on a question answering (QA) task. The goal in the QA task is to retrieve small snippets of text that contain the actual answer to a question rather than the document lists traditionally returned by text retrieval systems. The assumption is that users would usually prefer to be given the answer rather than find the answer themselves in a document.

This paper summarizes the retrieval results of the track; a companion paper ("The TREC-8 Question Answering Track Evaluation") gives details about how the evaluation was implemented. By necessity, a track report can give only an overview of the different approaches used in the track. Readers are urged to consult the participants' papers elsewhere in the Proceedings for details regarding a particular approach.

1 The Task

A successful evaluation requires a task that is neither too easy nor too difficult for the current technology. If the task is too simple, all systems do very well and nothing is learned. Similarly, if the task is too difficult, all systems do very poorly and again nothing is learned. Accordingly, we chose a constrained version of the general question answering problem as the focus of the track.

The document collection used in the task was the same as the TREC-8 ad hoc collection, namely the set of documents on TREC disks 4 and 5 minus the *Congressional Record* documents. The documents consist mostly of newspaper articles and thus contain information on a wide variety of subjects. Participants were given 200 fact-based, short-answer questions, such as those given in Figure 1. Each question was guaranteed to have at least one document in the collection that explicitly answered the question.

Participants returned a ranked list of five [*document-id*, *answer-string*] pairs per question such that each answer string was believed to contain an answer to the question. Answer strings were limited to either 50 or 250 bytes, and could either be extracted from the corresponding document or automatically generated from information contained in the document. Human assessors read each string and made a binary decision as to whether the string actually did contain an answer to the question in the context provided by the document. Taking document context into account allowed a system that correctly derived a response from a document that was incorrect to be given full credit for its response.

Given a set of judgments for the strings, the score computed for a submission was mean reciprocal rank, defined as follows. An individual question received a score equal to the reciprocal of the rank at which the first correct response was returned, or 0 if none of the five responses contained a correct answer. The score for a submission was then the mean of the individual questions' reciprocal ranks. The reciprocal rank has several advantages as a scoring metric. It is closely related to the average precision measure used extensively in document retrieval. It is bounded between 0 and 1, inclusive, and averages well. A run is penalized for

- How many calories are there in a Big Mac?
- What two US biochemists won the Nobel Prize in medicine in 1992?
- Who was the first American in space?
- Who is the voice of Miss Piggy?
- Where is the Taj Mahal?
- What costume designer decided that Michael Jackson should only wear one glove?
- In what year did Joe DiMaggio compile his 56-game hitting streak?
- What language is commonly used in Bombay?
- How many Grand Slam titles did Bjorn Borg win?
- Who was the 16th President of the United States?

Figure 1: Example questions used in the question answering track.

AT&T Labs Research	MultText Project	U. of Iowa
CL Research	New Mexico State U.	U. of Maryland, College Park
Cymfony, Inc.	NTT DATA Corp.	U. of Massachusetts
GE/U. of Pennsylvania	National Taiwan U.	U. of Ottawa
IBM Research	Royal Melbourne Inst. Technology	U. of Sheffield
LIMSI-CNRS	Seoul National U.	Xerox Research Centre Europe
MITRE	Southern Methodist U.	

Figure 2: Participants in the Question Answering track.

not retrieving any correct answer for a question, but not unduly so. However, the measure also has some drawbacks. The score for an individual question can take on only six values (0, .2, .25, .33, .5, 1). Question answering systems are given no credit for retrieving multiple (different) correct answers. Also, since the track required at least one response for each question, a system could receive no credit for realizing it did not know the answer.

2 Retrieval Results

Twenty different organizations participated in the Question Answering track. The participants are listed in Figure 2. A total of 45 runs were submitted, 20 runs using the 50-byte limit and 25 runs using the 250-byte limit. Table 1 gives both the mean reciprocal rank and the number of questions for which no answer was found for each run. (Two submissions that contained errors are omitted from the table.) The scores are computed over the 198 questions that comprised the official test set. The table is split between the 50-byte and the 250-byte runs and is sorted by decreasing mean reciprocal rank within run type.

The number of questions for which no answer was found shows that the most accurate systems were able to find an answer for more than 2/3 of the questions. Furthermore, when the answer was found at all it was usually ranked first, as shown by the fact that the mean reciprocal rank is also close to 2/3 for these systems.

While the run with the highest mean reciprocal rank score was a 50-byte run, a direct comparison between 50- and 250-byte submissions from the same participant shows that the 50-byte task is more difficult. For every organization that submitted runs of both lengths, the 250-byte limit run had a higher mean reciprocal rank. This is not a surprising result—a system has a greater chance of including a correct response in a

Table 1: Mean reciprocal rank (MRR) and number of questions for which no correct response was found (# not found) for Question Answering track submissions.

Run Name	Participant	MRR	# not found
textract9908	Cymfony, Inc.	.660	54
SMUNLP1	Southern Methodist U.	.555	63
attqa50e	AT&T Research	.356	109
IBMDR995	IBM	.319	110
xeroxQA8sC	Xerox Research Centre Europe	.317	111
umdqa	U. of Maryland	.298	118
MTR99050	MITRE	.281	118
IBMVS995	IBM	.280	120
nttd8qs1	NTT Data Corp.	.273	121
attqa50p	AT&T Research	.261	121
nttd8qs2	NTT Data	.259	120
CRL50	New Mexico State U.	.220	130
INQ634	U. of Massachusetts	.191	140
CRDBASE050	GE/U. of Pennsylvania	.158	148
INQ638	U. of Massachusetts	.126	158
shefinq50	U. of Sheffield	.081	182
shefatt50	U. of Sheffield	.071	184
UIowaQA3	U. of Iowa	.018	188
UIowaQA4	U. of Iowa	.017	193

a) Runs with a 50-byte limit on the length of the response.

SMUNLP2	Southern Methodist U.	.646	44
attqa250p	AT&T Research	.545	63
GePenn	GE/U. of Pennsylvania	.510	72
attqa250e	AT&T Research	.483	78
uwmt9qa1	MultiText Project	.471	74
mds08q1	Royal Melbourne Inst. Tech	.453	77
xeroxQA8lC	Xerox Research Centre Europe	.453	83
nttd8ql1	NTT Data Corp.	.439	79
MTR99250	MITRE	.434	86
IBMDR992	IBM	.430	89
IBMVS992	IBM	.395	95
INQ635	U. of Massachusetts	.383	95
nttd8ql4	NTT Data Corp.	.371	93
LimsiLC	LIMSI-CNRS	.341	110
INQ639	U. of Massachusetts	.336	104
CRDBASE250	GE/U. of Pennsylvania	.319	111
clr99s	CL Research	.281	115
CRL250	New Mexico State University	.268	122
UIowaQA1	U. of Iowa	.267	117
Scai8QnA	Seoul National U.	.121	154
shefinq250	U. of Sheffield	.111	176
shefatt250	U. of Sheffield	.096	179
NTU99	National Taiwan U.	.087	173
UIowaQA2	U. of Iowa	.060	175

b) Runs with a 250-byte limit on the length of the response.

longer string—but it was not a guaranteed result. That is, longer strings that include a correct response were not always a correct response themselves. Response strings that contained multiple entities of the same semantic type as the answer and did not specifically indicate which of the entities was the answer were marked as incorrect. For example, for the question *What is the capital of Kosovo?* the 50-byte response of

```
0 miles northwest of Pristina, five demonstrators
```

was judged correct, while the 250-byte response of

```
protesters called for military intervention to end "the Albanian uprising."
</P> <P> At Vucitrn, 20 miles northwest of Pristina, five demonstrators were
reported injured, apparently in clashes with police.  </P> <P> Violent clashes
were also repo
```

was judged incorrect since it is unclear from the response whether the capital is Vucitrn or Pristina.

The submissions from AT&T Research Labs demonstrate that existing passage-retrieval techniques can be successful for 250-byte runs, but are not suitable for 50-byte runs [18]. Their question answering system used a traditional vector-based retrieval system to select 50 documents and then scored each sentence within those documents by the number of question words in the surrounding context. For the passage-based runs (`attqa50p` and `attqa250p`), the highest scoring sentences were returned as the response. For their "entity-based" runs (`attqa50e` and `attqa250e`), high scoring sentences were further processed by a linguistic module. The passage-based method was very competitive for the 250-byte limit, but was not nearly as successful when restricted to just 50 bytes. NTT Data Corporation note similar effects in their runs [20]. These results suggest that the relatively simple bag-of-words approaches that are successfully used in text retrieval are not sufficient for extracting specific, fact-based answers.

3 Retrieval Strategies

Many participants used a variant of the following general strategy to the question answering problem. The system first attempted to classify a question according to the type of its answer as suggested by its question word. For example, a question that begins with "who" (*Who is the prime minister of Japan?*) implies a person or an organization is being sought, and a question beginning with "when" (*When did the Jurassic Period end?*) implies a time designation is needed. Next, the system retrieved a small portion of the document collection using standard text retrieval technology and the question as the query. The system performed a shallow parse of the returned documents to detect entities of the same type as the answer. If an entity of the required type was found sufficiently close to the question's words, the system returned that entity as the response. If no appropriate answer type was found, the system fell back to best-matching-passage techniques.

This approach works well provided the query types recognized by the system have broad enough coverage and the system can classify questions sufficiently accurately. Most systems could answer questions that began with "who" very accurately. However, questions that sought a person but did not actually begin with "who" (*Name the first private citizen to fly in space. What Nobel laureate was expelled from the Philippines before the conference on East Timor?*) were much more difficult. More difficult still were questions whose answers were not an entity of a specific type (*What is Head Start? Why did David Koresh ask the FBI for a word processor?*). Of course, pattern matching on expected answer types was not fool-proof even when "good" matches were found. One response to the question *Who was the first American in space?* was Jerry Brown, taken from a document that says

```
As for Wilson himself, he became a senator by defeating Jerry Brown, who has
been called the first American in space.
```

Broadly speaking, each of the following organizations used a variant of the general strategy: AT&T Labs Research [18], CL Research [10], Cymfony, Inc. [19], GE/University of Pennsylvania [13], LIMSI-CNRS [5], MITRE Corporation [2], New Mexico State University [15], NTT Data Corporation [20], Southern Methodist University [12], University of Maryland [14], University of Ottawa/NCR [11], University of Sheffield [8], and Xerox Research Centre Europe [7]. The approach used by IBM Research [16] was very similar in spirit to this approach except they located entities at indexing time and used a bag-of-words scoring metric that incorporated the entities, thus providing efficient retrieval at question-answering time. The University of Iowa [4] classified questions by type and used their filtering system to learn features of answers. Seoul National University [17] performed an initial document retrieval run and then selected phrases from top-ranking documents by extracting the immediate neighborhood of the highest-weighted question word. Finally, the MultiText Project [3], National Taiwan University [9], Royal Melbourne Institute of Technology/CSIRO [6], and University of Massachusetts [1] used traditional passage retrieval techniques alone.

4 Conclusion

The Question Answering track was the first large-scale evaluation of domain-independent question answering systems. The questions used in the track were deliberately constrained to fact-based, short-answer questions to make the task amenable to evaluation. Systems generally classified a question according to the type of its answer, and then performed a shallow parse of likely documents to find objects of the entailed type. The most accurate systems were able to answer more than 2/3 of the questions correctly. Existing passage-retrieval techniques were adequate for finding answers when relatively long responses were permissible, but more sophisticated processing was need to focus on the answer itself.

There will be another Question Answering track in TREC-9, which will be mostly the same as the TREC-8 track. One change in the track will be to have a test set of 500 questions rather than 200 questions, and to have many fewer of the questions be constructed from a target document.

References

[1] James Allan, Jamie Callan, Fang-Fang Feng, and Daniella Malin. INQUERY and TREC-8. In Voorhees and Harman [21].

[2] Eric Breck, John Burger, Lisa Ferro, David House, Marc Light, and Inderjeet Mani. A sys called Qanda. In Voorhees and Harman [21].

[3] G.V. Cormack, C.L.A. Clarke, C.R. Palmer, and D.I.E. Kisman. Fast automatic passage ranking (MultiText experiments for TREC-8). In Voorhees and Harman [21].

[4] David Eichmann and Padmini Srinivasan. Filters, webs and answers: The University of Iowa TREC-8 results. In Voorhees and Harman [21].

[5] Olivier Ferret, Brigitte Grau, Gabriel Illouz, Christian Jacquemin, and Nicolas Masson. QALC—the question-answering program of the Language and Cognition group at LIMSI-CNRS. In Voorhees and Harman [21].

[6] Michael Fuller, Marcin Kaszkiel, Sam Kimberly, Corinna Ng, Ross Wilkinson, Mingfang Wu, and Justin Zobel. The RMIT/CSIRO ad hoc, q&a, web, interactive, and speech experiments at TREC 8. In Voorhees and Harman [21].

[7] David A. Hull. Xerox TREC-8 question answering track report. In Voorhees and Harman [21].

[8] Keven Humphreys, Robert Gaizauskas, Mark Hepple, and Mark Sanderson. University of Sheffield TREC-8 Q & A system. In Voorhees and Harman [21].

[9] Chuan-Jie Lin and Hsin-Hsi Chen. Description of preliminary results to TREC-8 QA task. In Voorhees and Harman [21].

[10] Kenneth C. Litkowski. Question-answering using semantic relation triples. In Voorhees and Harman [21].

[11] Joel Martin and Chris Lankester. Ask Me Tomorrow: The NRC and University of Ottawa question answering system. In Voorhees and Harman [21].

[12] Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana Gîrju, and Vasile Rus. LASSO: A tool for surfing the answer net. In Voorhees and Harman [21].

[13] Thomas S. Morton. Using coreference in question answering. In Voorhees and Harman [21].

[14] Douglas W. Oard, Jianqiang Wang, Dekang Lin, and Ian Soboroff. TREC-8 experiments at Maryland: CLIR, QA and routing. In Voorhees and Harman [21].

[15] Bill Ogden, Jim Cowie, Eugene Ludovik, Hugo Molina-Salgado, Sergei Nirenburg, Nigel Sharples, and Svetlana Sheremtyeva. CRL's TREC-8 systems: Cross-lingual IR and Q&A. In Voorhees and Harman [21].

[16] John Prager, Dragomir Radev, Eric Brown, Anni Coden, and Valerie Samn. The use of predictive annotation for question answering in TREC8. In Voorhees and Harman [21].

[17] Dong-Ho Shin, Yu-Hwan Kim, Sun Kim, Jae-Hong Eom, Hyung-Joo Shin, and Byoung-Tak Zhang. SCAI TREC-8 experiments. In Voorhees and Harman [21].

[18] Amit Singhal, Steve Abney, Michiel Bacciani, Michael Collins, Donald Hindle, and Fernando Pereira. AT&T at TREC-8. In Voorhees and Harman [21].

[19] Rohini Srihari and Wei Li. Information extraction supported question answering. In Voorhees and Harman [21].

[20] Toru Takaki. NTT DATA: Overview of system approach at TREC-8 ad-hoc and question answering. In Voorhees and Harman [21].

[21] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Eighth Text REtrieval Conference (TREC-8).* Electronic version available at http://trec.nist.gov/pubs.html, 2000.