Overview of the TREC-9 Question Answering Track

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

Abstract

The TREC question answering track is an effort to bring the benefits of large-scale evaluation to bear on the question answering problem. The track has run twice so far, where the goal both times was 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 best performing system in TREC-9, the Falcon system from Southern Methodist University, was able to answer about 65% of the questions (compared to approximately 42% of the questions for the next best systems) by combining abductive reasoning with various natural language processing techniques. The 65% score is slightly less than the best scores for TREC-8 in absolute terms, but it represents a very significant improvement in question answering systems. The TREC-9 task was considerably harder than the TREC-8 task because the TREC-9 track used actual users' questions rather than questions constructed specifically for the track.

The TREC-9 question answering (QA) track was the second running of a QA track in TREC. The goal of the track is to foster research on systems that retrieve answers rather than documents in response to a question, with an emphasis on systems that can function in unrestricted domains. While the subject matter of the questions is not restricted, the type of questions the systems are expected to process is limited. Questions are restricted to fact-based, short-answer questions such as *How many calories are there in a Big Mac?*. The answers to such questions are usually entities familiar to information extraction systems. In this way, the track provides an opportunity for the information retrieval and information extraction communities to work on a common problem.

The first QA track in TREC-8 established two facts. First, since the best systems in TREC-8 were able to answer about 70% of the questions, the task is at the right level of difficulty for the current state of the art—challenging, but not impossible. Second, the evaluation methodology used in the track is valid. Although different people's opinions as to whether a retrieved string constitutes a correct answer differ even for this simple type of question, the relative quality of different systems remains stable regardless of whose opinions are used[7]. The TREC-9 track therefore used the same task as in TREC-8 (with some minor differences described below) to provide stability and allow confirmation of the retrieval results observed in TREC-8.

This paper gives an overview of the TREC-9 track. The next section defines the task in detail. Section 2 presents the retrieval results and summarizes the approaches taken by the participating systems. Section 3 examines the effect a particular way a question is phrased had on system processing, and section 4 looks at the prospects for building a reusable QA test collection. The final section discusses the future of the QA track.

1 The TREC-9 QA Task

As mentioned above, the TREC-9 QA task was essentially the same as the TREC-8 task. Participants were given a large corpus of newspaper/newswire documents and a test set of questions. The questions were closed-class questions of the type shown in Figure 1. Each question was guaranteed to have at least one document in the collection that explicitly answered it. The answer was guaranteed to be no more than 50 characters long.

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. All processing was required to be strictly

- How much folic acid should an expectant mother get daily?
- Who invented the paper clip?
- What university was Woodrow Wilson president of?
- Where is Rider College located?
- Name a film in which Jude Law acted.
- Where do lobsters like to live?

Figure 1: Example questions from the TREC-9 question answering track.

automatic (i.e., there could be no human intervention of any kind in producing an answer), and participants were not permitted to change their systems once they received the test questions. Answer strings were limited to either 50 or 250 bytes depending on the run type. The strings could either be extracted from the corresponding document or automatically generated from information contained in the document. Human assessors read each string and decided whether the string actually did contain an answer to the question in the context provided by the document.

Given a set of judgments for the strings, the score computed for a submission was the mean reciprocal rank (MRR). 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. Note that with this scoring metric systems are given no credit for retrieving multiple (different) correct answers. Also, since the track required at least one response for each question, systems could receive no credit for realizing they did not know the answer.

While the same basic task was performed in the TREC-8 and TREC-9 tracks, there were some differences between the tracks. Both the document set and the test set of questions was larger for TREC-9, as shown in Table 1. A more substantive difference was the source of the questions used in each TREC. Some of the questions used in TREC-8 were drawn from a log of questions submitted to the FAQ Finder system, but most of the questions in the log did not have answers in the document collection. As a result, the majority of questions used in TREC-8 were developed by either the participants or NIST assessors specifically for the track. Questions created for the track were often back-formulations of statements in the documents, which

Table 1: Data used in the two TREC question answering tracks.

	TREC-8	TREC-9
number of documents	528,000	979,000
megabytes of document text	1904	3033
document sources	TREC disks 4–5: LA Times, Financial Times, FBIS, Federal Register	news from TREC disks 1–5: AP newswire, Wall Street Journal, San Jose Mercury News, Financial Times, LA Times, FBIS
number of questions released	200	693
number of questions evaluated	198	682
question sources	FAQ Finder log, assessors, participants	Encarta log, Excite log

made the questions somewhat unnatural and also made the task easier since the target document contained most of the question words. For the TREC-9 track, NIST obtained two query logs and used those as a source of questions. An Encarta log, made available to NIST by Microsoft, contained grammatical questions. The other log was a log of queries submitted to the Excite search engine on December 20, 1999. Since the Excite log contains relatively few grammatically well-formed questions, the log was used as a source of ideas for NIST staff who created well-formed questions from query words without referring to the document collection. NIST assessors then checked whether each candidate question had an answer in the document collection, and a candidate question was discarded if no answer was found.

The TREC-9 question set contained 500 questions drawn from the logs, plus an additional 193 questions that were syntactic variants of an original question. The purpose of the syntactic variants was to investigate whether QA systems are robust to the variety of different ways a question can be phrased. Once the first 500 questions were selected, NIST assessors were given a subset of the questions and asked to create "natural" variants of the question. The intent was that the variant should have the same semantic meaning of the original, as well as be phrased in a way that a native English speaker might ask the question. For example, the test set contained four variants for the question *What is the tallest mountain?*: *What is the world's highest peak?*, *What is the highest mountain in the world?*, *Name the highest mountain.*, and *What is the name of the tallest mountain in the world?*. The 193 variants included variants for 54 different original questions, with a range of one to seven new questions per original.

Another difference between the TREC-8 and TREC-9 tracks was the addition of an "unsupported" judgment in TREC-9. There were a number of instances during the TREC-8 judging when an answer string contained the correct answer, but that answer could not possibly have been determined from the document returned. For example, the correct answer for *Who is the 16th President of the United States?* is Abraham Lincoln. One of the answer strings returned contained Abraham Lincoln, but the associated document discussed Lincoln's Gettysburg Address. The document does not even mention that Lincoln was president, let alone that he was the sixteenth president. Since the TREC-8 task did not specifically require that the document returned with the answer string support the string as the answer, these cases were judged as correct in TREC-8, even though the assessors were uncomfortable doing so. In TREC-9, the track guidelines required that the document returned with the answer string actually support the answer contained in the string. If the answer string did not contain a correct answer, the response was judged incorrect. If the string did contain a correct answer, but the document did not support that answer, the response was judged unsupported. Otherwise, the response was judged correct. Two scores were computed for each TREC-9 run, a *strict* score in which unsupported answers were considered incorrect, and a *lenient* score in which unsupported answers were considered correct.

Since the TREC-8 track showed that single-opinion judgments are comparable to adjudicated, multiple-opinion judgments for the purpose of comparing question answering system effectiveness [7], each TREC-9 question was judged by only one assessor. The savings resulting from using only one judge per question permitted the increase from 200 to 693 questions in the test set. Unfortunately, subsequent analysis has shown that the judgment set for the TREC-9 QA track contains a somewhat higher error rate (i.e., judgments that are just plain wrong rather than differences of opinion) than expected. The cause of the increased error rate is not clear. The TREC assessors were asked to do a lot of judging of different types for TREC-9, and it could be that some assessors confused tasks. It is also likely that real users' questions, being more ambiguous than the constructed questions used in TREC-8, are more difficult for humans to judge.

Strings were judged in the same way as they had been in TREC-8: the string did not have to provide justification of the answer, but was required to be responsive to the question. Being responsive includes such things as not containing distracting information, containing appropriate units, and pertaining to the entity asked about rather than replicas or imitations of the entity. Voorhees and Tice give a detailed description of answer string judging in [8]. Each individual variant in a variant question set was judged as a separate question, though the entire set was judged by the same assessor and variants were judged consecutively.

2 Track Results

Both the TREC-8 and TREC-9 QA tracks offered two experimental conditions: answer strings limited to 250 bytes, and answer strings limited to 50 bytes. Participants were permitted to submit up to two runs for each condition (four runs total), where a run consisted of a ranked list of up to five [*document-id*, *answer-string*] pairs for each question in the test set.

Twenty-eight organizations participated in the TREC-9 question answering track. A total of 78 runs was submitted, 34 runs using the 50-byte limit and 44 runs using the 250-byte limit. Table 2 gives the mean reciprocal rank (MRR) and number of questions for which no correct answer was found (# not found) using strict evaluation for 20 runs for each run type. The table is split between the 50-byte and 250-byte runs and is sorted by decreasing mean reciprocal rank within run type. Only one run for each organization within run type is included in the table.

Two main conclusions can be drawn from Table 2: scores are generally lower than the best scores from TREC-8, and the best performing system (from Southern Methodist University) did substantially better than the other systems. Despite the drop in the absolute value of the evaluation scores, the performance of the TREC-9 systems represents a significant improvement in question answering technology. The switch to "real" questions, rather than questions created especially for the track, made the TREC-9 task much more difficult than the TREC-8 task. The motivation for using actual user questions was the belief that constructed questions are easier for QA systems because the question and answer document share the same vocabulary. However, the difference between the TREC-8 and TREC-9 question sets was larger than just vocabulary issues. TREC-8 questions had been restricted to those with an "obvious" answer. While the subsequent differences in opinion during judging demonstrated that there is no such thing as an obvious answer, the questions were still far less ambiguous than the questions mined from logs. Real users ask vague questions such as *Who is Colin Powell?* and *Where do lobsters like to live?*. These questions are substantially harder for both the systems to answer and the assessors to judge.

The differences between the TREC-8 and TREC-9 questions also meant that that the TREC-8 questions were not a representative training set for the TREC-9 task. The problem can be best illustrated by "who" questions. In TREC-8, all of the questions that began with "who" asked for the name of a person or organization such as *Who is the prime minister of Japan?*, and most systems could answer this type of question very accurately. In TREC-9, a sizable fraction of the "who" questions were requests for information about a person such as *Who is Colin Powell?* and *Who was Jane Goodall?*. Information requests are a much more difficult type of question to answer.

The improvement in TREC-9 QA systems came from refinements to the individual steps of the general strategy used by TREC-8 systems rather than entirely new approaches. The general strategy used in TREC-8 was as follows. 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 beginning with "when" implies a time designation is needed. Next, the system retrieved a small portion of the document collection using standard document 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 a response. If no appropriate answer type was found, the system fell back to best-matching-passage techniques. TREC-9 systems were better at classifying questions as to the expected answer type, and used a wider variety of methods for finding the entailed answer types in retrieved passages.

Many TREC-9 systems used WordNet [1] as a source of related words for the initial query and as a means of determining whether an entity extracted from a passage matched the required answer type. Results from Queens College, CUNY demonstrated that high-quality document retrieval in the initial step is helpful [5]. This group used comparatively simple answer extraction techniques, yet performed relatively well, especially in the 250-byte condition.

The Southern Methodist University system, called Falcon [2, 4], classifies questions by expected answer type, but also includes successive feedback loops that try progressively larger modifications to the original question until it finds an answer that can be justified as an abductive proof. The system first parses the question and recognizes entities contained in it to create a question semantic form. The semantic form of the question is used to determine the expected answer type by finding the phrase that is most connected to other concepts in the question. The system uses an answer taxonomy that contains WordNet subhierarchies and thus has broad coverage. Falcon next retrieves paragraphs from the corpus using Boolean queries and terms drawn from the original question, related concepts from WordNet, and an indication of the expected answer type. The paragraph retrieval is repeated using different term combinations until the query returns

Table 2: Mean reciprocal rank (MRR) and number of questions for which no correct response was found (# not found) using strict evaluation for top TREC-9 QA track submissions.

Run Name	Participant	MRR	# not found	
LCCSMU2	Southern Methodist U.	0.58	229	(34%)
ISI0A50	ISI, U. of So. California	0.32	385	(57%)
uwmt9qas0	MultiText, U. of Waterloo	0.32	395	(58%)
IBMKR50	IBM (Prager)	0.32	402	(59%)
ibmhlt0050	IBM (Ittycheriah)	0.29	394	(58%)
pir0qas2	Queens College, CUNY	0.28	401	(59%)
SUT9bn3c050	Syracuse U.	0.25	436	(64%)
NTTD9QAa2S	NTT Data Corp.	0.23	439	(64%)
ALI9C50	U. de Alicante	0.23	451	(66%)
xeroxQA9s	Xeroc Research Centre Europe	0.23	453	(66%)
ICrjc99a	Imperial College	0.23	454	(67%)
KAIST9qa1	KAIST	0.21	468	(69%)
shef50	U. of Sheffield	0.21	470	(69%)
msq9L50	Microsoft Research, Ltd.	0.20	475	(70%)
FDUT9QS2	Fudan U.	0.20	495	(73%)
UdeMshrt	U. de Montreal	0.18	486	(71%)
ualberta	U. of Alberta	0.18	497	(73%)
lcat050	LIMSI	0.18	499	(73%)
clr00b2	CL Research	0.14	550	(81%)
Scai9QnA2	Seoul National U.	0.10	577	(85%)

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

LCCSMU1	Southern Methodist U.	0.76	95	(14%)
ibmhlt00250	IBM (Ittycheriah)	0.46	263	(39%)
pir0qal2	Queens College, CUNY	0.46	264	(39%)
uwmt9qal1	MultiText, U. of Waterloo	0.46	265	(39%)
IBMKA250	IBM (Prager)	0.42	294	(43%)
lcat250	LIMSI-CNRS	0.41	307	(45%)
NTTD9QAa1L	NTT Data Corp.	0.39	299	(44%)
SUT9p2c3c250	Syracuse U.	0.39	319	(47%)
ICrjc99b	Imperial College	0.39	348	(51%)
UdeMlng2	U. de Montreal	0.37	325	(48%)
KUQA250a	Korea U.	0.37	338	(50%)
ALI9C250	U. de Alicante	0.36	321	(47%)
xeroxQA9l	Xerox Research Centre Europe	0.35	349	(51%)
shef250p	U. of Sheffield	0.34	335	(49%)
INQ9AND	U. of Massachusetts	0.34	344	(50%)
SunToo	Sun Microsystems	0.34	362	(53%)
FDUT9QL1	Fudan University	0.34	369	(54%)
KAIST9qa2	KAIST	0.33	362	(53%)
qntua02	National Taiwan U.	0.32	376	(55%)
clr00s2	CL Research	0.30	386	(57%)

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

a number of paragraphs in a pre-determined range. The retrieved paragraphs are parsed into their semantic forms, and a unification procedure is run between the question semantic form and each paragraph semantic form. If the unification fails for all paragraphs, a new set of paragraphs is retrieved using synonyms and morphological derivations of the previous query. When the unification procedure succeeds, the semantic forms are translated into logical forms, and a logical proof in the form of an abductive backchaining from the answer to the question is attempted. If the proof succeeds, the answer from the proof is returned as the answer string. Otherwise, terms that are semantically related to important question concepts are drawn from WordNet and a new set of paragraphs is retrieved.

3 Question Variants

Question variants were introduced into the test set to explore whether systems' question processing—especially the determination of the expected answer type—could handle different formulations of the same basic question. While the intent had been for each variant to have identical semantics, assessment demonstrated that this was not always the case. Sometimes rewording a question caused the focus of the question to change slightly, so that some answer strings were acceptable for some variants but not others. For example, the assessor accepted "November 29" as a correct response for *What is Dick Clark's birthday?*, but required the year as well for *When was Dick Clark born?*. Similarly, the question *Where is the location of the Orange Bowl?* had many more acceptable responses than did *What city is the Orange Bowl in?*.

The change in the focus of a question sometimes resulted in a variant for which the guarantee of an answer in the document collection could no longer be made. For example, *Who invented silly putty?* had an answer of "a General Electric engineer". That answer was not acceptable for the variant *What is the name of the inventor of silly putty?*, so the variant was removed from the test set during evaluation. Eleven questions in all were removed from the test set either because of the change in focus or because the assessor who did the judging disagreed with the answer that was accepted during the candidate question verification phase.

Systems that parsed questions into a common representation generally had fewer differences in their responses to question variants than did systems that relied on templates to classify questions by answer types. The Falcon system cached responses to questions and returned exactly the same response for a question that was sufficiently similar to an earlier question.

Figure 2 shows a plot of the average score for each question in a variant set. The average score for a question is the mean of the reciprocal rank scores averaged over the 33 runs that used the 50-byte limit on responses and using strict evaluation. The y-axis in the plot is the average score and the x-axis represents the different variant sets. The variant sets are identified by the question number of the original question that was used to generate the variants.

Many variant sets show little variability in the average score. Generally, the average score for each of these variants is low, indicating that the underlying information being sought was difficult to obtain no matter how the question was phrased. A few variant sets did have a wide range of average scores. Frequently the difference was caused by different word choices in the variants. For example, the variant set generated from question 413 asked for the location of the U.S. headquarters of Proctor & Gamble. The variant with the lowest average score was question 725 which used "corporate offices" instead of "headquarters". For the variant set generated from question 440, the original question was *Where was Poe born?*, which had a much higher score than any of the variants that all asked for Poe's birthplace. The unintentional change in focus of some variants also made differences in average scores. "New Jersey" was an acceptable (and common) answer to Question 448, *Where is Rider College located?*, but it was not acceptable for the variant *Rider College is located in what city?*.

4 Question Answering Test Collections

The primary way TREC has been successful in improving document retrieval performance is by creating appropriate test collections for researchers to use when developing their systems. A document retrieval test collection consists of a set of documents, a set of information needs, and a set of relevance judgments that list the documents that are relevant to (i.e. should be retrieved for) each information need. Obtaining an

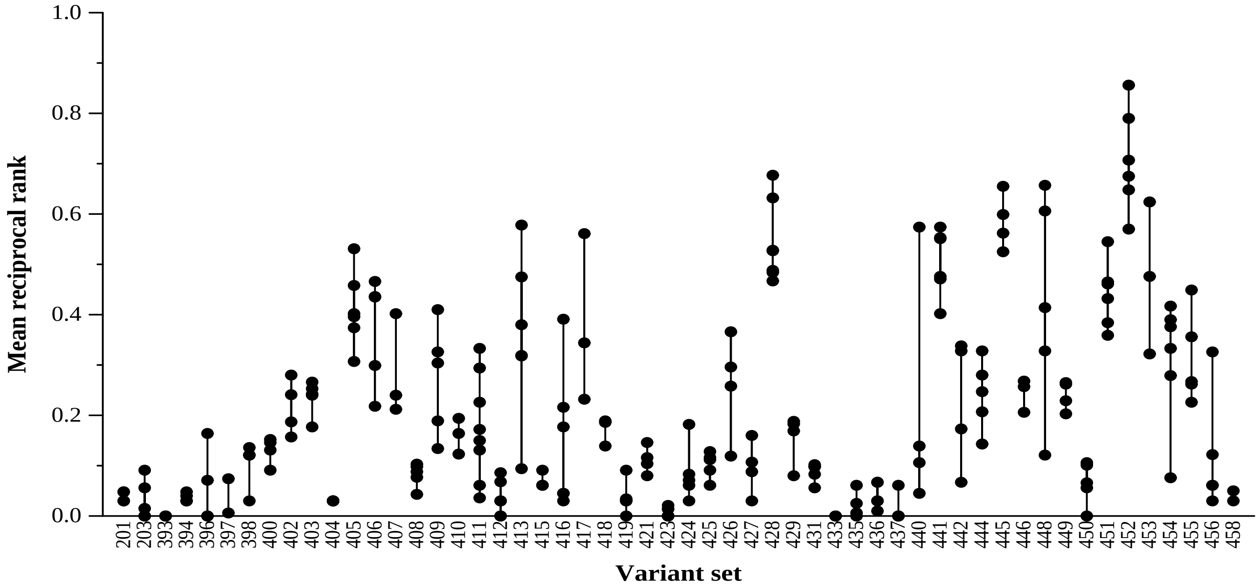

Figure 2: Average reciprocal rank for each question variant computed over 33 50-byte runs using strict evaluation. The x-axis represents different variant sets, identified by the question number of the original question from which the variants were generated.

adequate set of relevance judgments for a large collection can be time-consuming and expensive, but once a test collection is created researchers can automatically evaluate the effectiveness of a retrieval run. One of the key goals of the QA track was to build a reusable QA test collection—that is, to devise a means to evaluate a QA run that uses the same document and question sets but was not among the runs judged by the assessors.

Unfortunately, the judgment sets produced by the assessors for the TREC QA track do not constitute a reusable test collection because the unit that is judged is the entire answer string. Different QA runs very seldom return exactly the same answer strings, and it is quite difficult to determine automatically whether the difference between a new string and a judged string is significant with respect to the correctness of the answer. Document retrieval test collections do not have this problem because the unique identifiers assigned to the documents makes it trivial to decide whether or not a document retrieved in a new run has been judged.

As an approximate solution to this problem, NIST created a set of perl string-matching patterns from the set of strings that the assessors judged correct [7]. An answer string that matches any pattern for its question is marked correct, and is marked incorrect otherwise. The patterns have been created such that almost all strings that were judged correct would be marked correct, sometimes at the expense of marking as correct strings that were judged incorrect. Patterns are constrained to match at word boundaries and case is ignored.

The pattern sets for three questions are given in Figure 3. An average of 1.7 patterns per question was created for the TREC-8 test set, with 65% of the questions having a single pattern. The TREC-9 set averaged 3.5 patterns per question with only 45% of the questions having a single pattern. The increase in the number of patterns per question for the TREC-9 set is another indication that the TREC-9 test set was more difficult. For example, there are a variety of acceptable answers to the question *Who was Jane Goodall?*. The complexity of a pattern set is also affected by the particular answer strings that were judged. The question *Where is the location of the Orange Bowl?* has a complicated pattern set because many of the incorrect answer strings talk about football teams, including the University of Miami, that played in the college bowl game known as the Orange Bowl.

Using the patterns to evaluate the TREC-8 runs produced differences in the relative scores of different systems that were comparable to the differences caused by using different human assessors. The differences in relative scores for TREC-9 were larger, especially for the 50-byte runs. Using Kendall's τ as the measure

Who invented Silly Putty?

```
General\s+Electric
```

Where is the location of the Orange Bowl?

```
^\s*Miami\s*$                     ^\s*in\s+Miami\s*\.?\s*$
to\s+Miami                        at\s+Miami
Miami\s*'\s*s\s+downtown          Orange.*\s+in\s+.*Miami
Orange\s+Bowl\s*,\s*Miami         Miami\s*'?\s*s\s+Orange
Dade County
```

Who was Jane Goodall?

```
naturalist                        expert\s+on\s+chimps
chimpanzee\s+specialist           chimpanzee\s+researcher
chimpanzee\s*-?\s*observer        ethologists?
pioneered.*study\s+of\s+primates  anthropologist
ethnologist                       primatologist
animal\s+behaviorist              wife.*van\s*Lawick
scientist\s+of\s+unquestionable\s+reputation
most\s+recognizable\s+living\s+scientist
```

Figure 3: The pattern sets for three TREC-9 questions. Each pattern is a perl string-matching pattern. A response is considered correct if any pattern matches the answer string and incorrect otherwise.

of association between system rankings [6, 7], the τ over all runs for TREC-8 was .96, while for TREC-9 the τ was .94 for 250-byte runs and only .89 for 50-byte runs. This smaller correlation is probably the result of several factors. The TREC-8 human judgment scores were produced using an adjudicated judgment set that was a combination of three different assessors' judgments, and therefore was of particularly high quality. As mentioned above, the TREC-9 judgment set is known to contain more errors than the TREC-8 judgment set, and these errors could reduce the correlation with the pattern judgments. The more ambiguous questions in the TREC-9 test set were also harder to create patterns for.

Answer patterns are not a true solution to the problem of building a reusable test collection for question answering. Unlike differences of opinion between human assessors, patterns misjudge broad classes of responses—classes that are the cases that are difficult for the original systems being evaluated. For example, patterns cannot recognize when a correct answer is given in the wrong context and do not penalize systems that engage in answer stuffing. Nonetheless, the patterns can be useful for providing quick feedback as to the relative quality of question answering techniques provided their limitations are understood.

5 The Future

A roadmap for question answering research was recently developed under the auspices of the DARPA TIDES project [3]. The roadmap describes a highly ambitious program to increase the complexity of the types of questions that can be answered, the diversity of sources from which the answers can be drawn, and the means by which answers are displayed. The roadmap also includes a five year plan for introducing aspects of these research areas as subtasks of the TREC QA track.

The QA track in TREC 2001 (TREC-10) will include the first steps of the roadmap. The main task in the track will be similar to the task used in TRECs 8 and 9, but there will be no guarantee that an answer is actually contained in the corpus. Recognizing that the answer is not available is challenging, but it is an important ability for operational systems to possess since returning an incorrect answer is usually worse than not returning an answer at all. The track will also contain a subtask in which each question will require information from more than one document to be assembled to produce the answer. For example, a list question such as *Name the countries the Pope visited in 1994.* will require finding multiple documents that describe the Pope's visits and extracting the country from each. The system will also need to detect duplicate reports of the same visit so that countries are listed only once per visit.

Acknowledgements

My thanks to QA track coordinators Amit Singhal and Tomek Strzalkowski.

References

[1] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.

[2] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Gîrju, V. Rus, and P. Morarescu. FALCON: Boosting knowledge for answer engines. In Voorhees and Harman [9].

[3] Sanda Harabagiu, John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. Issues, tasks, and program structures to roadmap research in question & answering (Q&A), October 2000. `http://www-nlpir.nist.gov/projects/duc/roadmapping.html`.

[4] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the Association for Computational Linguistics*, pages 274–281, July 2001.

[5] K.L. Kwok, L. Grunfeld, N. Dinstl, and M. Chan. TREC-9 cross language, web and question-answering track experiments using PIRCS. In Voorhees and Harman [9].

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

[7] Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, July 2000.

[8] 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`.

[9] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Ninth Text REtreival Conference (TREC-9)*, 2001.