The Development and Evolution of TREC and DUC

Donna Harman Retrieval Group, Information Access Division Information Technology Laboratory National Institute of Standards and Technology

Abstract

The Text REtrieval Conference (TREC) has been running for 11 years now, with 93 participants in the last round of evaluation. This paper chronicles the changes in TREC over that time, emphasizing the evolution in the tasks that were evaluated rather than discussing the results of the specific evaluations. The development of the new Document Understanding Conference (DUC) is also discussed, including the evaluation issues that have surfaced during its first two years.

Introduction

In 1992 a new test collection was built at the National Institute of Standards and Technology (NIST) as part of the TIPSTER project [8], a large evaluation project for text retrieval and extraction sponsored by the U.S. Government. This test collection was larger than the older test collections by a factor of 1000, i.e., instead of 2 megabytes of abstracts, this collection contained 2 gigabytes of full-text documents. One of the reasons that NIST undertook the building of the test collection was to enable the entire information retrieval research community to have access to a test collection more suitable to the 1990s.

In addition to making this collection available to researchers, NIST also created an evaluation effort [5] to use this collection. This evaluation, called the Text REtrieval Conference (TREC), has been running for 11 years now, with 93 participants in the last round of evaluation. Complete details of TREC, including ten years of proceedings and information on how to obtain the test collections can be found at http: //trec.nist.gov.

This series of evaluations has added a new dimension to information retrieval experiments. Participating groups not only use the same test collection, but work on the same tasks, use the same metrics, and then compare results at the November meetings. Note that the importance of this cross-system comparison is not to determine the best system but to give researchers some basis for understanding the strengths and weaknesses of the various techniques developed by others. This encourages the transfer of good ideas, and the identification of appropriate performance benchmarks for new approaches to match.

This paper chronicles the changes in TREC over the past 11 years, emphasizing the evolution in the tasks that were evaluated rather than discussing the results of the specific evaluations. The goal of the paper is to present the issues that have caused those evolutions, both as a record of what happened and as a guidepost to other evaluation efforts such as NTCIR who confront some of the same evaluation problems. The development of the new Document Understanding Conference (DUC) is also discussed, including the evaluation issues that have surfaced during its first two years.

Evolution in TREC

The initial tasks for the first three TRECs were "ad hoc" retrieval from static collections and "routing" retrieval using static information needs against streaming (new) documents. The first year NIST worked with the sponsoring agencies to create 50 complex descriptions of information needs (called topics in TREC), and asked participants to retrieve a ranked list of 200 documents from the 2 gigabytes of data as an answer to each topic. Additionally some routing topics were devised, along with a small amount of training material, and groups used these to develop standing queries that were then tested against new data.

Most groups in TREC-1 concentrated on the necessary rebuilding of their basic systems to handle the huge scale-up in collection size. But an important part of the November 1992 meeting was to examine the various evaluation issues *as a group* and to decide collectively what improvements were needed for the next year. One of the major discussion points that first year was the large size of the topics; it was strongly felt that users would not submit such long topics (107 words on average) and therefore a much shorter topic would be more appropriate. However, in the interests of keeping a similar task for the second year, this change was postponed. A second issue was the surprisingly large number of relevant documents that were found in the collection (an average of 277 relevant documents per topic). This caused evaluation at a cutoff of 200 documents to be inaccurate after a 40% recall, and therefore it was determined to ask systems to return a ranked list of the top 1000 documents in TREC-2.

Evolution from TREC-2 to TREC-3 was more drastic. The topic length problem was revisited, with the result that topics for TREC-3 had no "concepts" field, a field that had contained useful manually-assigned keywords. This topic change made for more realistic topics, and also for much more challenging research. The ad hoc topics got even smaller in TREC-4 by losing the narrative field, but this caused problems in making relevance judgments. The basic TREC-3 topic format was used for the ad hoc track in TRECs 5, 6, 7, and 8. The initial ad hoc task was suspended after TREC-8 due to lack of general progress in the track.

It should be noted that the evolution of the ad hoc task reflects a general pattern seen in all evolutions in TREC; the participants and/or NIST identify an issue, bring it for discussion at the November meeting, and then decide as a group what changes are necessary. These changes could be a simple correction to solve some evaluation or procedural problem, as in TREC-1, but more often the desired change comes because the task needs to be made more challenging. The final part of the evolution is the decision to suspend or terminate a TREC task when it is no longer challenging or when it is no longer clear where the task is headed.

Evolution in the TREC tracks

TREC-3 also added new tasks, called tracks in TREC. In general there have been two types of tracks used throughout the years in TREC, either tracks that use tasks that are variations of the initial ad hoc or routing tasks, or tracks that evaluate tasks that are considered important by the information retrieval community but are not addressed by the two initial main tasks.

For TREC-3 two new tracks were added, one of each type. The ad hoc task was also run in Spanish, i.e., Spanish topics against Spanish documents, and a new interactive track was added, reflecting the importance of user interaction in the IR community.

Table 1 lists the different tracks that have been held in each TREC, the number of groups that submitted runs to that track, and the total number of groups that participated in each TREC. It should be noted that the tasks within the tracks have changed; in fact the evolution of the tasks in TREC has been a critical factor in keeping TREC relevant over the 11-year period.

Each track has a set of guidelines developed under the direction of the track coordinator and participants are free to choose which of the tracks they will join. The selection of the task for a given year, and the detailed guidelines that define that task must follow several conflicting constraints. In particular the tasks must be focused enough to allow commonality in experimentation, but interesting/challenging enough to attract participants. The evaluation methodology must be cleanly defined and within NIST's capabilities.

The set of tracks, their primary goals and discussion of their evolution are presented in this section. See the track reports in the online TREC proceedings for a more complete description of each track and its results (the track reports are folded into the main TREC overview up through TREC-4; afterwards the track reports are separate papers in the proceedings). Also see the individual papers for a given TREC to get experimental details for each participant in the tracks.

Evolution of the interactive track

The interactive track, one of the first tracks to be started in TREC, has studied text retrieval systems in interaction with users and has been interested in the search process as well as the results.

The interactive track was in pilot status for TREC-3, using the TREC-3 routing topics (which were available earlier than the ad hoc ones), but processing them in an interactive way. This unfortunately led to a comparison of the human performance to that of the automatic machine learning being used in the routing task. Therefore in TREC-4 the interactive track used the TREC-4 ad hoc topics. There were 11 groups who worked in the interactive track that year, many performing interesting comparisons of interactive trive retrieval versus automatic retrieval.

The major interactive evaluation issue from TREC-4 was the desire to be able to compare interactive systems across sites in spite of site dependencies (searchers, hardware, ...). The track attempted to do this in TRECs 5 and 6 by installing a common system at all participating sites and comparing the systems across sites indirectly by looking at how they measured up to the control system[7]. In the end the use of a common control system could not be justified since the experimental cost was high and efforts to prove its effectiveness were inconclusive.

More generally, beginning with TRECs 5 and 6, the interactive track disinguished itself from other TREC tasks and tracks by accepting the use of small topic sets (less than 10) and emphasizing the importance of good experimental design/analysis in detecting the often small system effects among the other main effects (e.g., searcher, topic) and their interactions. The track also moved away from the standard ad hoc document retrieval task into question answering and away from static document collections to web data.

The track identified a primary need to increase experimental power by reducing variability and addressed this need with a replicated Latin square design that allowed comparison systems within topic and within searcher. It recognized the desirability of large numbers of searchers but also the practical impossi-

	TREC										
Track	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
	TREC-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8	T-9	T-2001	T-2002
Ad Hoc	18	24	26	23	28	31	42	41	—	—	
Routing	16	25	25	15	16	21	—	—	—	—	
Interactive			3	11	2	9	8	7	6	6	6
Spanish	_		4	10	7	—	—	_	_	—	
Confusion			—	4	5	—	—	—		—	
Database Merging			—	3	3	—	—	—		—	
Filtering	—		—	4	7	10	12	14	15	19	21
Chinese			—	—	9	12	—	—		—	
NLP	—		—	—	4	2	—	—		—	
Speech	_		—	—	—	13	10	10	3	—	—
Cross-Language			—	—	—	13	9	13	16	10	9
High Precision			—	—	—	5	4	—	—	—	
Very Large Corpus	—		—	—	—	—	7	6	_	—	—
Query			—	—	—	—	2	5	6	—	
Question Answering			—	—	—	—	—	20	28	36	34
Web	—		—	—		—	—	17	23	30	23
Video	—		—	—	—	—	—	—	—	12	19
Novelty	_										13
Total participants	22	31	33	36	38	51	56	66	69	87	93

Table 1. Number of participants per track and total number of distinct participants in each TREC

bility of finding these for most participating research groups. The experimental design set a minimum of 4 searchers, which grew quickly to 16 and was often exceeded by several fold in some groups. Lack of searchers remains a major weakness of the interactive track work.

More than other tracks, the interactive track explored the question of how much freedom the track guidelines should allow. What is the correct balance between freedom that may allow more groups to test a wider set of approaches and the potential cost of narrowing the scope of system comparison (e.g., no cross-site comparisons), failing to support repeatability (outside TREC), and eroding the common basis for understanding and discussion at the workshop?

Evolution in non-English tracks

The second longest running track has been the use of the ad hoc task in languages other than English. Whereas the task has been held constant, there has been a steady escalation in language difficulty as systems demonstrated rapid progress in each language.

The first language tackled was Spanish, a language of interest to many of the participants and a language similar to English in structure. Over a three year period (TRECs 3, 4, and 5), 75 topics were constructed in Spanish at NIST, using a document collection consisting of about 500 megabytes of the *El Norte* newspaper from Monterey, Mexico and the 1994 newswire from *Agence France Presse*.

After 3 years of Spanish the participants wanted a new challenge, in particular a language with very different characteristics than English. Chinese was chosen to allow investigation of retrieval performance for a language whose orthographics are not wordoriented. The document set was a collection of 168,811 articles (170 megabytes) selected from the *Peoples Daily* newspaper and the *Xinhua* newswire. Twenty-eight topics were created for the track in TREC-5, with an additional 26 topics for TREC-6.

The evolution of the non-English retrieval task took an additional twist in 1997. Inspired by a new test collection[9] created at the Swiss Institute of Technology (ETH) Zurich, the new cross-language (CLIR) track in TREC-6 focused on retrieving documents that are written in different languages (French, German and English) using topics that are in one language only. The collection was composed of 250 megabytes of French and 330 megabytes of German articles from the Swiss news agency Schweizerische Depeschen Agentur (SDA) plus 750 megabytes of English newswires from NIST's AP collection. All of the document sets contained news stories from approximately the same time period. The 25 topic descriptions used in TREC-6 were provided by NIST in English, French and German, using translations of topics originally written mostly in English.

Unlike the monolingual ad hoc task, the CLIR track presented some difficult evaluation challenges. The first was the large range of language pairs that required testing: a potential permutation of all pairs of languages being used. This led to the identification of specific language pairs for bilingual testing and a new multilingual task where the results were a merged ranked list including documents from all languages.

The second issue was the use of non-native speakers for the evaluation. The TREC-6 topics were created at NIST by two persons who were native English speakers but who had strong skills in French and German. There were some problems with the topic translations produced at NIST, but also a serious lack of speed in making relevance assessments for non-native languages. This problem led to forming collaborative partnerships for the evaluation effort in TREC-7.

For TRECs 7 and 8, the track was run in cooperation with four European institutions: University of Zurich, Switzerland (working on the French portion); Social Science Information Centre, Bonn and the University of Koblenz (working on the German portion); and CNR, Pisa, Italy (doing a new Italian portion). There were 25 new topics built for TREC-7 with an additional 28 topics built for TREC-8. Note that these topics were created in each of the cooperating institutions in their native language and then translated to the other languages. Additionally assessments were done in a distributed mode according to the language of the retrieved documents.

The success of the collaborative partnerships inspired a new European evaluation effort, and in 2000 the European CLIR effort moved to the CLEF workshop (http://www.iei.pi.cnr.it/DELOS/CLEF).

But the CLIR task remained at TREC for 3 more years. TREC-9 had a CLIR task that used 25 English topics against 126 megabytes of Chinese documents from Hong Kong news-papers. Since Chinese CLIR was also being done in Asia in the new NTCIR workshop (http://research.nii.ac.jp/ntcir/workshop), the TREC CLIR task switched to English/Arabic. TRECs 2001 and 2002 used a total of 75 English topics against 896 megabytes of Arabic from the Agence France Press Arabic newspaper.

TREC 2002 was the final run of the CLIR task in TREC itself. The results of 9 years of monolingual and cross-lingual evaluation has shown that information retrieval methods used in English port surprisingly well for other languages. Equally important, monolingual and cross-lingual efforts are being evaluated elsewhere and there is no need for duplicate effort.

Evolution of the Routing/Filtering track

The routing task has been run since TREC-1, but with major evolutions in the data, in the metrics, and in the tasks. Since the routing task involves the use of training data to "learn" profiles/standing queries, it has always been a problem to find appropriate test data that is similar to the training data. The plan for first three TRECs was to use the topics and relevance judgments from the ad hoc task in one year as training data for the routing task the following year, and introduce new data for testing of the routing queries. This plan worked well only for TREC-2; for TREC-3 no new data could be found to match the training data and old data had to be reused. For TREC-4 it was decided to find new data *first* and then select appropriate old topics for training. Similar methods were used in TRECs 5 and 6.

There was also disagreement on how to evaluate the routing task. Many real routing applications require a system to make a binary decision whether or not to retrieve the current document, not to form a ranking of a document set. Since these results are an unordered set of documents, the rank-based measures in routing task are not appropriate. Starting in TREC-4, a filtering track has worked on the binary version of the routing task, using utility functions as measures of the quality of the retrieved set. Most TRECs since TREC-4 have tried different types of utility functions in search of the elusive "ideal" measure.

The filtering track was merged with the routing task starting in TREC-7, and a third task was added that models an interactive or *adaptive* filtering application. In this task, a filtering system starts with just the query derived from the topic statement, and processes documents one at a time in date order. If the system decides to retrieve a document, it obtains the relevance judgment for it, and can modify its query as desired. In TREC-7 all three tasks used the routing topics from TREC-1 (topics 1-50), with the AP newswire from 1988-1990 as the test documents, and in TREC-8 used the TREC-7 ad hoc topics and the Financial Times data (no new data for testing).

Having exhausted all the TREC data, the group decided to try the OHSUMED collection [6] of 350,000 Medline documents in TREC-9. Some of the OHSUMED questions were used, but additionally there was some pseudo-classification work done using the MeSH headings. No new relevance assessments were made.

Good luck favored the filtering track in 1999 when Reuters made available a new collection containing 800,000 news stories from 1996-1997. Since no relevance assessments were available for training, and NIST was unable to build topics, the TREC 2001 filtering track used the category codes from Reuters to approximate topics. In TREC 2002 NIST was able to build filtering topics and create training data specifically for the filtering track.

The filtering track has had a long and varied run at TREC. In 2002 it was decided to suspend this track for one year to allow time to rethink the task.

Evolution of non-text retrieval

Retrieval of non-textual material has also been done at TREC. Evolution here has been into increasingly difficult retrieval areas, with three different tracks run to date (confusion/OCR, speech and video). Note that each of these tracks have involved new communities for TREC since they merge the information retrieval interests with those of OCR and speech recognition and later the video retrieval/video processing communities.

The confusion track in TRECs 4 and 5 used input from OCR as the documents and known-item searching as the task. This task proved to be less challenging than expected, with groups generally able to cope with OCR error rates on the order of 20%. Then starting in TREC-6, retrieval of speech documents (broadcast news) was tried, using known item searching at first but then moving on to the building of topics in a similar manner to the ad hoc task. The speech track ran for 4 years, each year working with more speech (87 hours in TREC-7; 500 hours in TREC-8). This was a major challenge to the speech recognition community in terms of scaling up automatic speech transcription, and a good chance for the speech community and the information retrieval community to work together on retrieval of "corrupted" text. After TREC-9 the participants felt that the speech retrieval task could be called a success and the track was terminated.

In TREC-2001 a new track was started to search 11 hours of MPEG-1 video for specific shot sequences of items such as "waterskiers behind a speedboat". The track included the task of shot-boundary detection as well as 74 specific searches to be performed either automatically or manually. For TREC-2002, the number of hours of video was increased to 73.3 hours from the Internet Archive (http://www.archive.org), with similar search and shot-boundary detection tasks. An additional semantic feature finding task was added in 2002 at the request of the participants, with 10 different features defined such as cityscape, face, and instrumental sound.

The video track will become an independent evaluation in 2003.

Evolution of the Very Large Corpus (VLC) and Web tracks

The VLC track was started in TREC-6 by the Australian National University (ANU) to explore how well retrieval algorithms scale to larger document collections. In contrast to the ad hoc task that used a 2 gigabyte document collection, the first running of the VLC track used a 20 gigabyte collection. Precision at 20 was the major effectiveness measure, with assessments done at ANU for the top 20 documents retrieved for each of the TREC-6 ad hoc topics. Also reported were query response time; data structure (e.g., inverted index) building time; and a cost measure of number of queries processed per minute per hardware dollar.

The TREC-7 VLC collection consisted of World Wide Web data that was collected by the Internet Archive. A 100 gigabyte sample of this data was used, along with TREC-7 ad hoc topics (and relevance judgments by ANU as before).

The use of web data suggested the possibility of a track on web searching, and the VLC track migrated to this task in TREC-8. Two subtasks, the small web and the large web tasks, were performed, using a 2 gigabyte subset of the collection, and the entire collection respectively.

The large web task focused on the ad hoc task only, using queries extracted from logs from the Alta Vista and Electric Monk search engines. Fifty of the 10,000 queries were selected for judgment, with the top 20 documents for each run evaluated.

The small web task was exactly the same as the TREC-8 ad hoc task except that the web documents were searched instead of the documents on Disks 4 and 5. The NIST relevance assessors who judged the ad hoc pools also judged the corresponding small web pools. The focus of the small web task was on answering two questions: do the best methods used in the TREC ad hoc task also work best on web data and can link information in web data be used to obtain more effective search rankings than can be obtained using page content alone?

There were some serious questions about the effects of the size and structure of the 2-gigabyte web data on the results and this led to a much larger (10 gigabyte) web track in TREC-9 that had a more controlled structure design. In addition to the more realistic web structure, the TREC-9 topics were specifically generated from real web logs, with the title line of the topic containing the actual web log terms (including misspellings). Three-level judgments (highly relevant, relevant and non-relevant) were used (for the first time in TREC).

TREC 2001 used the same web collection, but added a homepage finding task to the search task used in TREC-9. In TREC 2002 the web track moved to a larger collection (18 gigabytes), formed naturally by a partial crawl of the .GOV domain. The homepage task migrated to a named-page finding task, and a second task of topic distillation was added.

The web track will continue in TREC 2003, with a group of interested researchers currently working on the guidelines. Possibly an even larger crawl of the web will be made and the large web subtask revived.

Evolution of the question-answering tracks

TREC-8 was the first time the Question Answering track was run. The purpose of the track was to encour-

age research into systems that return actual answers, as opposed to ranked lists of documents, in response to a question.

The track used the TREC-8 ad hoc document collection and 198 fact-based, short-answer questions such as "How many calories are there in a Big Mac?". Each question was guaranteed to have at least one document in the collection that answered the question. Participants were to return a ranked list of five strings per question such that each string was believed to contain an answer to the question. Depending on the run type, answer strings were limited to either 50 or 250 bytes. Human assessors read each string and made a binary decision as to whether or not the string actually did contain an answer to the question.

This track was a surprisingly huge success, attracting groups in the natural language processing area as well as groups from the traditional information retrieval area. The track was expanded for TREC-9, using 3 gigabytes of documents for searching and 682 questions. Whereas the questions were similar in type to those used in TREC-8, they were taken from real search logs (Encarta and Excite) and turned out to be more difficult.

TREC 2001 continued the same task, this time using search logs from Microsoft and AskJeeves to generate 500 questions. Strings were limited to 50 bytes (250 bytes being considered too easy by this time), and questions were allowed *not* to have an answer. Additionally a new list type question was used requiring that participants return lists of requested items gleaned from multiple documents ("Name 4 U.S. cities that have a "Shubert" theatre").

TREC 2002 saw further tightening of the task to make it more challenging. Participants were only allowed one answer (not a ranked list of 5 possible answers), and had to return the exact answer rather than a 50-byte string. The list task was continued.

TREC 2002 also had a new version of the questionanswering task: the novelty track. The novelty track was targetted at the opposite end of the questionanswering spectrum, i.e., given a TREC topic (not a fact-based short-answer question) and an ordered list of relevant documents, find the "novel" (and relevant) sentences that should be returned to the user from this set. The novelty track used old TREC ad hoc topics and evaluated the automatic selection of novel sentences against a manual selection of these sentences.

Both the novelty track and the QA track will continue in TREC 2003, with new topics being generated for the novelty track and some new types of questions being created and evaluated in the QA track. Note that the ARDA AQUAINT program is focused on the QA task and this will drive the TREC QA task into much more complex types of questions (and evaluations).

Summarization at DUC

Research in summarization was one of the first efforts to use computers to understand language. Along with the research came efforts to evaluate automatic summarization performance. Two major types of evaluation have been used: *intrinsic* evaluation where the emphasis is on measuring the quality of the created summary directly, and *extrinsic* evaluation where the emphasis is on measuring how well the summary performs within a given task. For examples of extrinsic evaluations, see the TIPSTER SUMMAC evaluation [3] and the Japanese NTCIR evaluation [4].

The DARPA TIDES program offered the opportunity to tackle summarization evaluation once again and a long-term roadmap [1] to guide this evaluation was created. This roadmap provided guidance for the new Document Understanding Conference (DUC), with a pilot run in 2000, and the first major evaluation in the fall of 2001.

For further information on DUC, including online proceedings from the conferences, see http: //duc.nist.gov

DUC 2001

The roadmap called for evaluation of generic summaries of both single documents and sets of multiple documents, at specified levels of text compression. For DUC 2001, fluent, generic extracts or abstracts were to be produced from largely domain-independent English newswire and newspaper articles. The implementation of this in DUC 2001 was as follows:

- Sixty sets of approximately 10 newspaper documents each were provided as system input for this task. Each document in the set was to be mainly about a single topic and each set of documents was to be centered around a specific event, a set of similar events, a specific subject, a specific person, or a set of opinions about a specific concept.
- 2. Given such a set of documents, the systems were to automatically create 100-word generic summaries for each document. Additionally they were to create generic summaries of the entire set, one summary at each of four target lengths (approximately 400, 200, 100, and 50 words).
- 3. The sets of documents were created at NIST by 10 TREC assessors. Each person created six document sets, and then created 100-word manual abstracts for each document, and for the entire document set at the 400, 200, 100 and 50 word lengths. Thirty of the sets (documents and manual abstracts) were distributed as training data; the rest were used in the testing.

The evaluation plan as specified in the roadmap was for NIST to concentrate on manual comparison of the system results with the humanly-constructed abstracts. Manual evaluation was considered critical because there was a general concern that automatic evaluation would not be adequate to deal properly with linguistic devices such as paraphrasing, or with abstracting methods that produce results differing greatly from simple extracts of the initial text. Additionally there was a desire to measure the coherence and organization of generated summaries.

Central to the manual comparison was a new tool developed by Chin-Yew Lin at the Information Sciences Institute, University of Southern California (http://www.isi.edu/~cyl/SEE/). This tool allows pairwise comparison of two summaries, and was used at NIST for the DUC 2001 evaluation. Human evaluation was done using the same personnel who created the manual abstracts (called model summaries in DUC). These people did pairwise comparisons between their "model" summaries and the "peer" summaries. Peer summaries include system-generated summaries, additional manual abstracts generated by others, and baseline summaries.

Two specific areas of evaluation were examined in DUC 2001 using the SEE interface. The first area involved the quality of the summary. Each systemgenerated summary, baseline summary, and "duplicate" manual summary was judged for grammaticality, cohesion and organization using a five-point scale.

The second area of evaluation dealt with coverage, i.e., how well did the peer summaries cover the content of the documents (as expressed by the model summary). The pairwise summary comparison was used in this part of the evaluation and judges were asked to do detailed coverage comparisons. SEE allowed the judges to step through predefined units of the model summary and for each unit of that summary, mark the sentences in the peer summary that expressed [all, most, some, hardly any or none] of the content in the current model summary unit.

The choice of units for the model and peer summaries involved practical as well as theoretical considerations. Researchers wanted units smaller than a sentence for better diagnostic information. Automatically determined units called elementary discourse units (EDUs) [2] based on rhetorical structure theory were proposed. Because these required some human post-editing, it was only practical to use them for the model summaries, which are many fewer in number than the summaries produced by the research systems. The latter were automatically divided into sentences.

Evolution to DUC 2002

Part of the September DUC 2001 meeting was spent discussing changes that needed to be made for DUC

2002. There was general agreement as to the success of using the SEE interface for judgments. It was felt that the judges were able to be as consistent as possible and that the interface allowed the evaluation to focus on comparisons that are useful to the summarization research community. The only suggested change was to replace the five-point intervals [All, most, some, hardly any, or none] with percentages [0, 20, 40, 60, 80, 100] to permit better score averaging and more consistent judgment by the evaluators.

However, poor discrimination in the quality part of the evaluation (grammar, cohesion and organization) made these scores mostly useless. The NIST assessors doing the evaluation had difficulty in separating the concepts: grammar was almost always judged as excellent, whereas the cohesion and organization scores tended to be similar for a given summary. For DUC 2002 it was resolved to evaluate the grammaticality, coherence and organization items by a series of questions about effective use of pronouns, dangling connectives, subject/verb agreement, etc. These more concrete questions would provide a consistent basis for quality assessment and allow the participants to more readily pinpoint text generation problems.

No standard metrics had been used in DUC 2001 and all participants were free to explore with their own set of metrics. This exploration led to convergence and the decision was that NIST would develop a new length-adjusted coverage metric for DUC 2002. In addition to measuring how well a peer summary covers the points in the model summary (the coverage metric), brevity would be rewarded.

In terms of input documents, NIST was asked to generate more document sets (60) to improve statistical analysis of the results. Additionally it was decided that the document sets would be broken into more tightly specified categories: documents about a single natural disaster event with all documents created at most in a seven day window, documents about a single event in any domain but with the same time constraint, documents about multiple distinct events of a single type (no time constraint), and documents that present biographical information mainly about a single individual.

In terms of output, the target lengths of the summaries were shortened to eliminate the 400-word summary and to include a headline length (10-word) summary. It was felt that these were more useful lengths to evaluate; the headline length has many important applications and the 400-word length might be better addressed using extracts.

DUC 2002 and evolution to DUC 2003

DUC 2002, the following July, represented the stabilization of the evaluation framework. The modified SEE interface worked well, and the new lengthadjusted coverage metric proved effective. This metric provides a way of assigning a single score for coverage regardless of the length of the input summary. It essentially assigns an adjustable bonus for writing shorter summaries that still cover the material in the ideal summary.

The coverage evaluation was shown to be stable when averaged over the 60 sets of documents. A Tukey test showed significant differences between three main groups of summaries for the multidocument summarization task: the manual(ideal) summaries in one group, some of the automatic summaries plus one baseline summary in the second group, and the rest of the automatic summaries plus the second baseline in the third group.

The automatic summaries were analyzed for readability by scoring them against 12 specific questions, such as "About how many times are unrelated fragments joined into one sentence?". Results from this part of the evaluation provided a combined readability score and input for system failure analysis and were a major improvement over the DUC 2001 quality evaluation.

One of the interesting outcomes of the DUC 2002 evaluation was that methods/systems performed about the same for all levels of text compression; that is, systems that did well at the 50 word summaries also did well on the 200 word summaries. For this reason it was decided to eliminate the different levels of text compression in DUC 2003 and to concentrate on summaries at the 100-word length.

With the stable evaluation framework and the new metrics completed, the evaluation discussion at DUC 2002 turned to the difficult problem of getting better agreement between humans on what constitutes a good summary. Note that the roadmap effort had called for evaluation of "generic" summaries, i.e., the creation of abstracts that summarize a document or groups of documents without regard to the use of the summary. This lack of focus tends to widen the natural human differences as to the importance of different concepts: the coverage metric comparing two humangenerated summaries averaged 50% for single document summaries and less than 40% for multidocument summaries. Whereas some of this disagreement is unavoidable, plans were made to hopefully increase the agreement by doing more focused summaries in DUC 2003.

Four different summarization tasks are currently ongoing for DUC 2003, each task carefully modeling a real and focused application of summarization. First, single documents will be summarized at the very short (headline) compression level, modeling the very short summaries seen as output of search engines. Second, there will be multi-document summarization of 10 documents drawn from multiple sources in the same timeframe about a given event. This task is now done for evolving news summary applications, such as that implemented by Columbia University's News-Blaster. The third task will be similar to the DUC 2002 multi-document summarization task except the human summarizers will be asked to state the viewpoint from which they are writing the summary. The fourth task will be to generate a multi-document summary with respect to a given question (TREC novelty questions will be used here). A pilot extrinsic evaluation will also be done for the headline-length summaries and for the question-answering summaries.

References

- [1] B. Baldwin, R. Donaway, E. Hovy, E. Liddy, I. Mani, D. Marcu, K. McKeown, V. Mittal, M. Moens, D. Radev, K. S. Jones, B. Sundheim, S. Teufel, R. Weischedel, and M. White. An evaluation road map for summarization research. http://duc.nist.gov/roadmapping.html, 2000.
- [2] L. Carlson, J. M. Conroy, D. P. O'Leary, D. Marcu, M. E. Okurowski, A. Taylor, and W. Wong. An empirical study of the relation between abstracts, extracts, and the discourse structure of texts. In *Proceedings of the 2001 Document Understanding Conference* (DUC2001), pages 11–18, 2001.
- [3] T. Firmin and M. Chrzanowski. An evaluation of automatic text summarization systems. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 325–340. MIT Press., Cambridge, Massachusetts, 1998.
- [4] T. Fukusima and M. Okumura. Text summarization challenge: Text summarization evaluation at ntcir workshop2. In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, pages 9–14, 2001.
- [5] D. Harman. Overview of the first Text REtrieval Conference (TREC-1). In *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 1–20, 1993. NIST Special Publication 500-207.
- [6] W. Hersh. Ohsumed: An interactive evaluation and new large test collection for research. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 192–201, 1994.
- [7] E. Lagergren and P. Over. Comparing interactive information retrieval systems across sites: The trec-6 interactive track matrix experiment. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 164–172, 1998.
- [8] R. Merchant. TIPSTER phase I: Program overview. In The Proceedings of the TIPSTER Text Program - Phase I, pages 1–2, 1994.
- [9] P. Sheridan, J. Ballerini, and P. Schäuble. Building a large multilingual test collection from comparable news documents. In G. Grefenstette, editor, *Cross-Language Information Retrieval*, chapter 11. Kluwer Academic Publishers, Boston, Massachusetts, 1998.