Report on TREC-9

Ellen M. Voorhees

National Institute of Standards and Technology

ellen.voorhees@nist.gov

1   Introduction

The ninth Text REtrieval Conference (TREC-9) was held at the National Institute of Standards and Technology (NIST) on November 13–16, 2000. The conference was co-sponsored by NIST, the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA/ITO), and the Advanced Research and Development Activity (ARDA) office of the Department of Defense.

TREC-9 is the latest in a series of workshops designed to foster research in text retrieval. The workshop series has four goals:

- to encourage research in text retrieval based on large test collections;

- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;

- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and

- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

The previous eight TRECs each had an "ad hoc" main task through which eight large test collections were built [8]. In recognition that sufficient infrastructure exists to support researchers interested in this traditional retrieval task, the ad hoc main task was discontinued in TREC-9 so that more TREC resources could be focused on building evaluation infrastructure for other retrieval tasks (called "tracks"). The seven tracks included in TREC-9 were web retrieval, cross-language retrieval, spoken document retrieval, query analysis, question answering, interactive retrieval, and filtering. Sixty-nine groups including organizations from 17 different countries participated in TREC-9.

This paper provides a short overview of the TREC-9 workshop. More details about individual tracks and the results of individual participants' investigations can be found in the TREC-9 proceedings. An electronic version of the proceedings is available on the TREC web site at `http://trec.nist.gov/pubs.html`. This report is extracted from the Overview paper in the proceedings.

2   Some Terminology

TREC classifies tasks into several broad categories. An *ad hoc* task is the traditional IR task. In this task the retrieval system knows the set of documents to be searched, but cannot anticipate the particular topic to be investigated. The result of an ad hoc search is assumed to be a list of documents ranked by decreasing similarity to the question. In a document routing or *filtering* task, the topic of interest is known and stable, but the document collection is constantly changing. The filtering task generally requires a retrieval system to make a binary decision whether to retrieve each document in the document stream as the system sees it. The result of a filtering search is therefore an unordered set of documents. A *question answering* task is one in which the system response is the actual answer rather than a document containing the answer.

A test collection is an abstraction of an operational retrieval environment that provides a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. Test

```
<num> Number:  453
<title> hunger

<desc> Description:
Find documents that discuss organizations/groups that are aiding in the eradication of the
worldwide hunger problem.
<narr> Narrative:
Relevant documents contain the name of any organization or group that is attempting to
relieve the hunger problem in the world.  Documents that address the problem only without
providing names of organizations/groups that are working on hunger are irrelevant.
```

Figure 1: A sample topic.

collections consist of three parts: a set of documents, a set of information needs, and *relevance judgments*, an indication of which documents should be retrieved in response to which information need. TREC makes a distinction between the expression of an information need (called the *topic*) and the data structure that is actually given to a retrieval system (called the *query*). TREC provides relatively lengthy, structured topics to allow a wide range of query construction methods to be tested and also to include a clear statement of what criteria make a document relevant. A topic statement generally consists of four sections: an identifier, a title, a description, and a narrative, as shown in the example topic in Figure 1.

The TREC *assessors* both create the topic statements and make the relevance judgments for the topics. Given the size of the document sets used in TREC, it is infeasible to judge every document for every topic. Instead, TREC uses a technique called pooling [6] to create a subset of the documents (the *pool*) to judge for a topic. Each document in the pool for a topic is judged for relevance by the topic author. Documents that are not in the pool are assumed to be irrelevant to that topic. The pool for a topic is created by merging the top X (usually, $X = 100$) documents retrieved by each system for that topic.

Retrieval runs on a test collection can be evaluated in a number of ways. The most common measures used in TREC are recall-precision graphs and average precision, where the scores are reported as averages over a set of topics. *Precision* is the proportion of retrieved documents that are relevant, while *recall* is the proportion of relevant documents that are retrieved. To make a recall-precision graph, precision is plotted as a function of recall. Since the actual recall values obtained for a topic depend on the number of relevant documents, the average recall-precision curve for a set of topics is interpolated to a set of standard recall values. The *average precision* of a single topic is the mean of the precision obtained after each relevant document is retrieved (using zero as the precision for relevant documents that are not retrieved). The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the individual topics in the run.

3   The TREC-9 Tracks

Each track has a set of guidelines developed under the direction of the track coordinator. Participants are free to choose which track they will join. This section briefly describes the tasks and results for each TREC-9 track.

3.1   The Web track

The purpose of the web track was to build a test collection that more closely mimics the retrieval environment of the World Wide Web. In creating such a collection, a variety of web retrieval strategies were also investigated. The task in the track was a traditional ad hoc retrieval task where the documents were a collection of web pages.

The web track was coordinated by the WAR project of ACSys CRC. They obtained a snapshot of the web from 1997 from the Internet Archive, and produced several subsets of that spidering. A 10 gigabyte subset known as WT10g was used for the main task in the web track. There was also a separate large task

in the web track that used the 100 gigabyte VLC2/WT100g collection and a set of 10,000 queries selected from Electronic Monk and AltaVista query logs. See the web track report in the TREC-9 proceedings for more details about the large web task.

The topics used in the main web task were TREC topics 451–500, which were especially created for the track using a process that differed from standard TREC topic creation methods. There was a desire among the track participants to use "real web queries", but also a desire to retain the statement of how the query would be judged. As a compromise, standard TREC topic statements were retrofitted around queries drawn from a log of the Excite search engine.[1] A sample of the queries in the log was given to the assessors. Each assessor selected a query from the sample and developed a description and narrative for that query. The assessors were instructed that the original query might well be ambiguous (e.g., "cats"), and they were to develop a description and narrative that were consistent with any one interpretation of the original (e.g., "Where is the musical Cats playing?"). They then searched the web document collection to estimate the likely number of relevant documents for that topic. The "title" field of topics 451–500 contains the literal query that was the seed of the topic, including any spelling or grammatical errors of the original.

Relevance assessing for the main web task was also done differently from the standard ad hoc assessing. For the first time in TREC, three-way judgments (not relevant, relevant, and highly relevant) were used. In addition, assessors were asked to select the best document from among all the documents in the pool for each topic. While the official results of the task were scored by conflating the relevant and highly relevant categories, the additional information collected during assessing can be used to develop other evaluation schemes for web retrieval.

Twenty-three groups submitted 105 runs to the main task of the web track. Twelve of the runs used manual query construction techniques, 40 of the runs were automatic runs that used only the original Excite query, and the remaining 53 runs were automatic runs that used some other part of the topic in addition to the Excite query. Runs using the description field of the topic were always more effective than the corresponding run using only the original Excite query, in large part because the spelling errors in the original query were corrected in the description. Seven topics had some sort of error in the original Excite query, five of which were serious (e.g., one-word queries "nativityscenes" and "angioplast7").

The order of systems ranked by average effectiveness differed depending on whether the evaluation used both relevant and highly relevant documents or highly relevant documents only. This finding implies that retrieving highly relevant documents is a different task from retrieving generally relevant documents, at least to the extent that different retrieval techniques should be used. More research is needed to determine precisely which techniques work better for which task and why.

The motivation for distinguishing between highly relevant and generally relevant documents is a widespread belief that web users will be better served by systems that retrieve highly relevant documents. Taking this reasoning a step further, some have argued that web search engines should actually be evaluated on their ability to retrieve the very best page. However, the results of the web track demonstrate that using the best page as the single relevant document is too unstable to be a reliable evaluation strategy.

Once the web track assessing was complete, NIST gave the set of relevant documents (both highly relevant and generally relevant) to two additional assessors and asked them to select the best page. In all cases the definition of best was left up to the assessor. There was significant disagreement among the assessors as to the best page for a topic: all three assessors disagreed with one another for 17 of the 50 topics, and of the 13 topics for which all three assessors picked the same page, 5 topics had three or fewer pages in its relevant set. Furthermore, unlike traditional retrieval system evaluation [7], best document evaluation is not robust against changes caused by using different assessors to select best documents. The Kendall tau correlation between system rankings produced by using different assessors averaged only about .75 when using best document evaluation as compared to over .9 for traditional evaluation.

3.2   The Cross-Language (CLIR) track

The CLIR task is an ad hoc retrieval task in which the documents are in one language and the topics are in a different language. The goal of the track is to facilitate research on systems that are able to retrieve relevant documents regardless of the language a document happens to be written in. The TREC-9 cross-language

---

[1] Jack Xu of Excite released this log on his ftp site at `ftp.excite.com/pub/jack`.

track used Chinese documents and English topics. A Chinese version of the topics was also developed so that cross-language retrieval performance could be compared with the equivalent monolingual performance.

The document set was approximately 250 megabytes of news articles taken from the Hong Kong Commercial Daily, the Hong Kong Daily News, and Takungpao. The documents were made available for use in TREC by Wisers, Ltd. Twenty-five topics were developed by NIST assessors. The assessment pools were created using each group's first choice cross-language run and first-choice monolingual run (if any), using the top 50 documents from each run.

Fifty-two runs from 15 different groups were submitted to the track. Thirteen of the runs were monolingual runs. Only one run (a cross-language run) was a manual run.

The effectiveness of cross-language runs is frequently reported as a percentage of monolingual effectiveness. In the CLIR track, the cross-language run submitted by the BBN group was not only better than their monolingual run (as measured by mean average precision), it was better than all the submitted monolingual runs. While BBN has since produced a monolingual run that is better than the best of their cross-language runs [10], the effectiveness of English to Chinese cross-language retrieval remains high.

3.3   The Spoken Document Retrieval (SDR) track

The SDR track fosters research on retrieval methodologies for spoken documents (i.e., recordings of speech). The task in the track is an ad hoc task in which the documents are transcriptions of audio signals.

The SDR track has run in several TRECs and the track has had the same general structure each year. Participants worked with different versions of transcripts of news broadcasts to judge the effects of errors in the transcripts on retrieval performance. The *reference* transcripts were manually produced and assumed to be perfect. (For TREC-9 the reference transcripts were a combination of human reference transcripts, closed captioning transcripts, and automatically-combined (using NIST's ROVER algorithm) automatic transcripts.) The *baseline* transcripts were produced by one automatic speech recognizer and made available to all participants. There was one TREC-9 baseline transcript, which was the "B2" transcript used in the TREC-8 track. This transcript was produced using NIST's installation of the BBN Rough 'N Ready BYBLOS speech recognizer. The *recognizer* transcripts were produced by the participants' own recognizer systems. The recognizer transcripts of the different participants were made available to one another so that participants could perform retrieval runs against their own recognizer transcripts as well as others' recognizer transcripts (*cross-recognizer* runs). The different versions of the transcripts allowed participants to observe the effect of recognizer errors on their retrieval strategy. The different recognizer runs provide a comparison of how different recognition strategies affect retrieval.

The document collection used in TREC-9 was the audio portion of the TDT-2 News Corpus as collected by the Linguistic Data Consortium (LDC). This corpus contains 557 hours of audio representing 1,064 news shows, which were segmented into approximately 21,500 documents. Two different versions of 50 new topics (numbers 124–173) were created for the track. The "standard" version of the topics was a one sentence description, while the "terse" version of the topics was just a few words. As in the TREC-8 track, the TREC-9 track focused on the unknown boundary condition, that is, retrieving documents from the audio stream when the system was not given the story boundaries.

Three groups participated in the track, submitting a total of 64 runs. Overall, the retrieval results were excellent: the systems could find relevant passages produced by a variety of recognizers on the full unsegmented news broadcasts, using either the terse or longer standard queries. Indeed, for each of the participants, retrieval from the transcripts created by their own recognizer was comparable to the retrieval from the human reference transcripts.

3.4   The Query track

The task in the query track was an ad hoc task using old TREC document and topic sets. The focus in the track was not on absolute retrieval effectiveness, but on the variability of topic performance. A variety of research (for example, see [1]) has shown that the difference in retrieval effectiveness for a given retrieval system on different topics is much greater on average than the difference in retrieval effectiveness between systems for the same topic. The development of query-specific processing strategies has been hampered as a result because the available topic sets (of size 50 for most TREC collections) are too small to isolate

Table 1: Average, minimum, and maximum Mean Average Precision scores for each query track system computed over the 43 different querysets submitted to the track. Each queryset contains queries created for TREC topics 51–100.

	Average	Minimum	Maximum		Average	Minimum	Maximum
IN7a	0.1799	0.1017	0.2534	UoMl	0.1539	0.0812	0.2099
IN7e	0.2288	0.1456	0.3168	hum4	0.1713	0.0907	0.2580
IN7p	0.1848	0.1095	0.2593	humA	0.1741	0.0912	0.2482
SUN	0.0572	0.0062	0.1247	humB	0.1732	0.0908	0.2420
Sunl	0.0677	0.0268	0.1152	humD	0.1771	0.0897	0.2508
Saba	0.1924	0.1089	0.2665	humI	0.1736	0.0910	0.2418
Sabe	0.2516	0.1481	0.3202	humK	0.1713	0.0863	0.2425
Sabm	0.2321	0.1347	0.3057	humV	0.1648	0.0788	0.2453
UoMd	0.1612	0.1010	0.2091	ok9u	0.1917	0.0963	0.2608

the effects caused by different topics. The query track was designed as a means for creating a large set of different queries for an existing TREC topic set as a first step toward query-specific processing.

Six groups participated in the TREC-9 query track, with each group running each of 43 different querysets using one or more variants of their retrieval system. A *queryset* consists of one query for each of 50 topics (TREC topics 51-100) where each query is from the same category of queries. Three different query categories were used.

1. Short: 2-4 words selected by reading the topic statement.

2. Sentence: a sentence—normally less than one line—developed after reading the topic statement and possibly some relevant documents.

3. Sentence-Rel: a sentence developed after reading a handful of relevant documents. The topic statement was *not* used for this category of query.

Relevant documents from TREC disk 2 were used to construct the queries. Twenty-one of the querysets were developed for the TREC-8 query track, and the remaining 22 querysets were developed for the TREC-9 track.

A *runset* is the result of running one version of a retrieval system on all 43 querysets against the documents on TREC disk 1. Eighteen runsets were submitted, for a total of 774 (18×43) runs submitted to the track. Table 1 gives the average, minimum, and maximum mean average precision score obtained for each of the 18 systems computed over the 43 different querysets. As can be seen from wide range of scores for each system, the individual query formulations again had a pronounced effect on retrieval performance.

3.5   The Question Answering (QA) track

The purpose of the question answering track was to encourage research into systems that return actual answers, as opposed to ranked lists of documents, in response to a question. Participants received a set of fact-based, short-answer questions and searched a large document set to extract (or construct) an answer to each 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 and the document supported that answer. Answer strings were limited to either 50 or 250 bytes depending on the run type. Each question was guaranteed to have an answer in the collection. 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. Question answering systems were given no credit for retrieving multiple (different) correct answers, or for recognizing that they did not know the answer.

There were several changes between the TREC-9 track and the initial running of the track in TREC-8. In TREC-9, the document set was larger, consisting of all the news articles on TREC disks 1–5. The test set of questions was also much larger, consisting of 693 questions rather than 200. A more substantial difference was the way in which the questions were created. Instead of using questions created especially for the track, which tended to be back-formulations of a sentence in some document in the collection, questions were selected from query logs. Some questions were taken from a log of questions that had been submitted to Encarta and were made available to NIST by Microsoft. Other questions were created by NIST staff using the Excite log from which the web track queries were selected for suggestions. In all cases, the questions were created without reference to the document set. Once the questions were created, NIST assessors searched the document set to find which questions had answers in the test document set. Five hundred questions were selected from among the candidate questions that had an answer in the document set. In a separate pass, NIST assessors were given a subset of the questions (but not their answers) and were asked to create equivalent, re-worded questions. For example, the question "How tall is the Empire State building?" might be re-worded as "How high is the Empire State building?", "What is the height of the Empire State building?", "The Empire State Building is how tall?", etc. A total of 193 question variants were added to the set of 500 to make the final question set of 693 questions.

Human assessors read each string and decided whether the answer string contained an answer to the question. If not, the response was judged as incorrect. If so, the assessor decided whether the answer was supported by the document returned with the string. If the answer was not supported by that document, the response was judged as "Not Supported". If it was supported, the response was judged as correct. The official scoring for the track treated Not Supported answers as incorrect.

Twenty-eight groups submitted 78 runs to the track, with 34 runs using the 50-byte-limit and 44 runs using the 250-byte-limit. The best performing system, from Southern Methodist University, was able to extract a correct answer about 65% of the time by integrating multiple natural language processing techniques with abductive reasoning [3]. While the 65% score is a slightly worse result than the TREC-8 scores in absolute terms, it represents a very significant improvement in question answering systems. The TREC-9 task was considerably harder than the TREC-8 task because of the switch to "real" questions (which tend to be far more ambiguous than the questions constructed for the TREC-8 task). The SMU system found an answer about a third again as often as the next best system (66% of the questions vs. 42% of the questions).

3.6 The Interactive track

The interactive track was one of the first tracks to be introduced into TREC. Since its inception, the high-level goal of the track has been the investigation of searching as an interactive task by examining the process as well as the outcome. One of the main problems with studying interactive behavior of retrieval systems is that both searchers and topics generally have a much larger effect on search results than does the retrieval system used.

The task in the TREC-9 track was a question answering task. Two different types of questions were used:

- find any n Xs (for example, "Name 3 US Senators on committees regulating the nuclear industry.")

- compare two specific Xs (for example, "Do more people graduate with an MBA from Harvard Business School or MIT Sloan?")

Human searchers were given a maximum of 5 minutes to find the answer to a question and support that answer with a set of documents. A total of 8 questions was used in the track.

The document set was the set of documents used in the question answering track (the news articles from TREC disks 1–5). The track defined an experimental framework that specified a minimum number of searchers, the order in which searchers were assigned questions, and the set of data to be collected. This framework did not provide for a comparison of systems across sites, but did allow groups to estimate the effect of their own experimental manipulation free and clear of the main (additive) effects of searcher and topic.

Six groups participated in the interactive track, with some groups performing more than the minimum number of searches. Of the 829 responses submitted to TREC across all topics and groups, 309 (37%) found

no correct answer, suggesting that the 5-minute time limit made the task challenging. The percentage of no answer found was roughly the same for the two types of questions (36% for the find any n questions and 39% for the comparison questions).

3.7 The Filtering track

The filtering task is to retrieve just those documents in a document stream that match the user's interest as represented by the query. The main focus of the track was an *adaptive* filtering task. In this task, a filtering system starts with just a query derived from the topic statement (for TREC-9, the system also received a few (< 5) relevant documents), 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 the query based on the judgment if desired. Two other, simpler tasks were also part of the track. In the *batch* filtering task, the system is given a topic and a (relatively large) set of known relevant documents. The system creates a query from the topic and known relevant documents, and must then decide whether or not to retrieve each document in the test portion of the collection. In the *routing* task, the system again builds a query from a topic statement and a set of relevant documents, but then uses the query to rank the test portion of the collection. Ranking the collection by similarity to the query (routing) is an easier problem than making a binary decision as to whether a document should be retrieved (batch filtering) because the latter requires a threshold that is difficult to set appropriately.

The document set for the TREC-9 filtering task was the OHSUMED test collection [5]. The test documents were the documents from 1988–1991, while a set of documents from 1987 were available for training (if the particular task allowed training). There were three topic sets: the queries from the OHSUMED test collection, a set of almost 5000 MeSH headings that were treated as topic statements, and a subset of 500 MeSH headings. Since the track used an existing test collection, no relevance judgments were made at NIST for the track.

Research into appropriate evaluation methods for filtering runs (which do not produced a ranked list and therefore cannot be evaluated by the usual IR evaluation measures) has been a major thrust of the filtering track. The earliest filtering tracks used linear utility functions as the evaluation metric. With a linear utility function, a system is rewarded some number of points for retrieving a relevant document and penalized a different number of points for retrieving an irrelevant document. Utility functions are attractive because they directly reflect the experience of a user of the filtering system. Unfortunately, there are drawbacks to the functions as evaluation measures. Utilities do not average well because the best possible score for each topic is a function of the number of relevant documents for that topic, and the worst possible score is essentially unbounded. Thus topics that have many relevant documents will dominate an average, and a single poorly performing topic can eclipse all other topics. Furthermore, it is difficult to know how to set the relative worth of relevant and irrelevant documents. For example, one of the utility functions used in the TREC-8 track rewarded systems with three points for retrieving a relevant document and penalized systems two points for retrieving an irrelevant document. This actually defines a very difficult filtering task: using this utility function, the average behavior of each of the TREC-8 filtering systems was worse than the baseline of retrieving no documents at all. The TREC-9 track used a bounded utility function as one measure and introduced a new "precision-oriented" measure, T9P. The bounded utility function rewarded systems two points for a relevant document and penalized systems one point for an irrelevant document, and in addition set a limit on the worst possible score a topic could receive. The idea of the precision-oriented measure was to penalize systems for not retrieving a sufficient number of documents, which for TREC-9 was 50 documents. Thus,

$$T9P = \frac{\text{number of relevant retrieved}}{\max(\text{number retrieved}, 50)}$$

Seventy-five runs from 14 different groups were submitted to the filtering track. Forty-one of the runs were adaptive filtering runs, 19 runs were batch filtering runs, and 15 runs were routing runs. Unlike TREC-8, several TREC-9 adaptive filtering systems obtained good average utilities while retrieving an adequate number of documents. In addition, the adaptive filtering scores were relatively close to the scores for the much easier routing task, as demonstrated by comparing the scores from the new T9P measure to the average of precision at 50 documents retrieved for the routing runs.

4 TREC 2001

The next TREC, TREC 2001, will see a few changes in the tracks that are offered. The spoken document track has met its goals and will therefore be discontinued. A new track that will focus on content-based access to digital video will continue TREC's interest in multimedia retrieval. The query track will cease to be a track and evolve into a "station". That is, the TREC web site will serve as a repository for both query statements and retrieval results produced from those queries. This will allow the research of the track to continue without the time pressures of the TREC deadlines.

The remaining tracks will continue, though the specific task involved in the track will likely change. The web track will include so-called navigational topics [2] in addition to the informational topics TREC has traditionally used. The cross-language track will focus on retrieving Arabic documents using English (and probably French) topics. In addition to the fact-based, short-answer questions used in the first two years of the question answering track, the 2001 track will feature a pilot study using questions that require information from multiple documents to be combined to form a correct answer. The task in the interactive track will most likely involve observing subjects using the live web to accomplish a specific task. The observations made in the track will be used to inform the definition of a task with a metrics-based evaluation plan for the following year's track, as suggested by the SIGIR Workshop on Interactive Retrieval at TREC and Beyond [4] held after SIGIR 2000. The filtering track will continue to focus on adaptive filtering.

The call to participate in TREC is issued each December. Applications to participate are requested by mid-February, but are accepted whenever the Call is posted on the main page of the TREC web site (http://trec.nist.gov). To join the "TREC friends" mailing list, and thus be sent the next call for participation directly from NIST, send a request to Ellen Voorhees, ellen.voorhees@nist.gov .

References

[1] David Banks, Paul Over, and Nien-Fan Zhang. Blind men and elephants: Six approaches to TREC data. *Information Retrieval*, 1:7–34, 1999.

[2] Andrei Broder. The need behind the query: Web search vs classic information retrieval. http://www.infonortics.com/searchengines/sh01/sh01pro.html.

[3] 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].

[4] William Hersh and Paul Over. SIGIR workshop on interactive retrieval at TREC and beyond. *SIGIR Forum*, 34(1):24–27, Spring 2000.

[5] W.R. Hersh, C. Buckley, T.J. Leone, and D.H. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–201, 1994.

[6] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

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

[8] Ellen M. Voorhees and Donna Harman. Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1):3–35, January 2000.

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

[10] J. Xu and R. Weischedel. TREC-9 cross-lingual retrieval at BBN. In Voorhees and Harman [9].