Overview of TREC 2021

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, MD 20899

1 Introduction

TREC 2021 is the thirtieth edition of the Text REtrieval Conference (TREC). The main goal of TREC is to create the evaluation infrastructure required for large-scale testing of information retrieval (IR) technology. This includes research on best methods for evaluation as well as development of the evaluation materials themselves. "Retrieval technology" is broadly interpreted to include a variety of techniques that enable and/or facilitate access to information that is not specifically structured for machine use. The TREC 2021 meeting was held at the National Institute of Standards and Technology (NIST) November 15–19, 2021.

Each TREC is organized around a set of focus areas called "tracks". A track has a motivating use case, which is generally an abstraction of a user task. TREC 2021 contained eight tracks:

Clinical Trials: The Clinical Trials track looks to focus research on matching patient health records to suitable clinical trials for that patient.

Conversational Assistance: The Conversation Assistance (CAsT) track looks to build systems that engage users in open-domain, information-centric, conversational dialogues.

Deep Learning: The Deep Learning track focuses on IR tasks where a large training set is available, allowing us to compare a variety of retrieval approaches including deep neural networks and strong non-neural approaches, to see what works best in a large-data regime.

Fair Ranking: The Fair Ranking track focuses on building two-sided systems that offer fair exposure for producers of ranked content while ensuring high results quality for ranking consumers.

Health Misinformation: The Health Misinformation track aims to (1) provide a venue for research on retrieval methods that promote better decision making with search engines, and (2) develop new online and offline evaluation methods to predict the decision making quality induced by search results. Consumer health information is used as the domain of interest in the track.

Incident Streams: The Incident Streams track is designed to develop technologies to automatically process social media streams during emergency situations, with the aim of categorizing information and aid requests made on social media for emergency service operators.

News: The News track is run in partnership with *The Washington Post*, it looks to develop test collections that support the search needs of news readers and news writers in the current news environment.

Podcast: The Podcasts track seeks to develop methods for information retrieval and content understanding from open-domain podcast transcripts and audio.

Seventy-three groups from twenty-seven different countries participated in TREC 2021. Table 1 lists the participating organizations.

This paper serves as an introduction to the research described in detail in the remainder of the proceedings. The next section provides a summary of the retrieval background knowledge that is assumed in the other papers. Section 3 presents a short description of each track—a more complete description of a track can be found in that track's overview paper in the proceedings.

Table 1: Organizations participating in TREC 2021

Alibaba Group	Sabanci University
Assoc para a Investigacao em Valor e Inovacao Tecnologica em Saude	Siena College Institute for Artificial Intelligence
Bosch Center for AI	Spotify
CMU Language Technologies Institute	State University of Campinas
CNR	Swiss Institute of Bioinformatics / HES-SO
Charles University	TDMINER
Chinese Information Processing Lab	TU Wien (2)
Commonwealth Scientific and Industrial Research Organisation	Technical University of Valencia
ElmDoc B.V.	Technische Hochschule Koln
Fernuniversitat in Hagen	The University of Queensland
Gonzaga University	U. of Applied Sciences & Arts of Western Switzerland
Hamad Bin Khalifa University	UniMelb/RMIT/ITTC AI MedTech
IBM Research	Universidad del Pais Vasco
IHS Markit	University College Dublin
Indian Institute of Technology Delhi	University College London
Information Technologies Institute CERTH	University Hospital Essen
Jeonbuk National University	University of Amsterdam (2)
L3S Research Center, Leibniz University	University of Cincinnati College of Medicine
MLIA team - LIP6 - Sorbonne Universite	University of Duisburg-Essen
Max Planck Institute for Informatics	University of Geneva
Middlebury College	University of Glasgow (2)
National Institute of Standards and Technology R&R Group	University of La Rochelle
National Taiwan University NLP Lab	University of Maryland
Naver Labs Europe	University of Milano-Bicocca
New Jersey Institute of Technology	University of North Texas
OpenSource Connections	University of Padova
PingAn Smart Health	University of Santiago de Compostela
Politecnico di Torino	University of Stavanger
Poznan University of Technology	University of Tsukuba
Qatar University	University of Waterloo (4)
RMIT University	Webis@Halle, Leipzig, Weimar
Radboud University	Wilfrid Laurier University
Research Ctr for IT Innovation, Academia Sinica	York University
Ryerson University	

2 Information Retrieval

Information retrieval is concerned with locating information that will help satisfy a user's information need. Traditionally, the emphasis has been on text retrieval: providing access to natural language texts where the set of documents to be searched is large and topically diverse. There is increasing interest, however, in finding appropriate information regardless of the medium that happens to contain that information. Thus "document" can be interpreted as any unit of information such as a tweet, an email message, a medical record, a web page, or an academic paper.

The prototypical retrieval task is a researcher doing a literature search in a library. In this environment the retrieval system knows the set of documents to be searched (the library's holdings), but cannot anticipate the particular topic that will be investigated. We call this an *ad hoc* retrieval task, reflecting the arbitrary subject matter that is the focus of the search and its short duration. A retrieval system's response to an ad hoc search is generally an ordered list of documents sorted such that documents the system believes are more likely to help satisfy the information need are ranked before documents it believes are less likely to satisfy the need.

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [4, 8], and TREC continues this tradition. 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 collections consist of three parts: a set of documents, a set of information needs (called *topics* in TREC), and *relevance judgments*, an indication of which documents should be retrieved in response to which topics. We call the result of a retrieval system executing a task on a test collection a run.

2.1.1 Documents

The document set of a test collection should be a sample of the kinds of texts that will be encountered in the operational setting of interest. It is important that the document set reflect the diversity of subject matter, word choice, literary styles, document formats, etc. of the operational setting for the retrieval results to be representative of the performance in the real task. The initial TREC test collections contained 2 to 3 gigabytes of text and 500,000 to 1,000,000 documents. The document sets used in various tracks throughout the years have been smaller and larger than these initial sets depending on the needs of the track and the availability of data, but the general trend has been toward ever-larger document sets to enhance the realism of the evaluation tasks. Similarly, the initial TREC document sets consisted mostly of newspaper or newswire articles, but later document sets have included a much broader spectrum of document types (such as recordings of speech, web pages, scientific documents, blog posts, email messages, and business documents). Each document is assigned an unique identifier called the DOCNO. For most document sets, high-level structures within a document are tagged using a mark-up language such as SGML or HTML, or broken into fields of a JSON object. In keeping with the spirit of realism, the text is kept as close to the original as possible.

2.1.2 Topics

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of the criteria that make a document relevant. What is now considered the "standard" format of a TREC topic statement—a topic id, a title, a description, and a narrative—was established in TREC-5 (1996). But topic formats vary in support of the task, and few current TREC tasks use topics in this traditional format.

Participants are (usually) free to use any method they wish to create queries from the topic statements. TREC distinguishes among two major categories of query construction techniques, automatic methods and manual methods. An automatic method is a means of deriving a query from the topic statement with no manual intervention whatsoever; a manual method is anything else. The definition of manual query construction methods is very broad, ranging from simple tweaks to an automatically derived query, through manual construction of an initial query, to multiple query reformulations based on the document sets retrieved. Since these methods require radically different amounts of (human) effort, care must be taken when comparing manual results to ensure that the runs are truly comparable.

TREC topics are generally constructed specifically for the task they are to be used in. When outside resources such as search engine logs are used as a source of topics the sample selected for inclusion in the test set is vetted to insure there is a reasonable match with the document set (i.e., neither too many nor too few relevant documents). Topics developed at NIST are created by the NIST *assessors*, the set of people hired to both create topics and make relevance judgments. Most of the NIST assessors are retired intelligence analysts. The assessors receive track-specific training by NIST staff for both topic development and relevance assessment.

2.1.3 Relevance judgments

Relevance judgments turn a set of documents and topics into a test collection. Given a set of relevance judgments, the ad hoc retrieval task is then to retrieve all of the relevant documents and none of the irrelevant documents. Most of the traditional measures of retrieval effectiveness treat relevance judgments as binary indicators—either a document is relevant to the topic or it is not—and the judgments themselves are binary in the original TREC collections. Use

of evaluation measures that incorporate different levels (or grades) of relevance has become much more prevalent in recent TRECs, and today relevance judgments are generally made on a graded scale to support the use of these measures.

Relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [6]. Furthermore, a set of static relevance judgments makes no provision for the fact that a real user's perception of relevance changes as he or she interacts with the retrieved documents. Despite the idiosyncratic nature of relevance, test collections are useful abstractions because the *comparative* effectiveness of different retrieval methods is stable in the face of changes to the relevance judgments [9].

The relevance judgments in the first retrieval test collections were complete. That is, a relevance decision was made for every document in the collection for every topic. The size of the TREC document sets makes complete judgments infeasible, so by necessity TREC collections are created by judging only a subset of the document collection for each topic and then estimating the effectiveness of retrieval results from the judged sample.

"Pooling" is the technique used in early TRECs for selecting the sample of documents for the human assessor to judge [7]. In pooling, the top results from a set of runs are combined to form the pool and only those documents in the pool are judged. Runs are subsequently evaluated assuming that all unpooled (and hence unjudged) documents are not relevant. In more detail, the TREC pooling process proceeds as follows. When participants submit their retrieval runs to NIST, they rank their runs in the order they prefer them to be judged. NIST chooses a number of runs to be merged into the pools, and selects that many runs from each participant respecting the preferred ordering. For each selected run, the top X documents per topic are added to the topics' pools.

The critical factor in pooling is that unjudged documents are assumed to be not relevant when computing traditional evaluation scores such as mean average precision (MAP). This treatment is a direct result of the original premise of pooling: that by taking top-ranked documents from sufficiently many, diverse retrieval runs, the pool will contain the vast majority of the relevant documents in the document set. If this is true, then the resulting relevance judgment sets will be "essentially complete", and the evaluation scores computed using the judgments will be very close to the scores that would have been computed had complete judgments been available.

Various studies have examined the validity of pooling's premise in practice. Harman [5] and Zobel [10] independently showed that early TREC collections in fact had unjudged documents that would have been judged relevant had they been in the pools. But, importantly, the distribution of those "missing" relevant documents was highly skewed by topic (a topic that had lots of known relevant documents had more missing relevant), and uniform across runs. Zobel demonstrated that these "approximately complete" judgments produced by pooling were sufficient to fairly compare retrieval runs. Using the leave-out-uniques (LOU) test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of 0.5 %. The maximum increase for any run was 3.5 %. The average increase for the TREC-3 ad hoc collection was somewhat higher at 2.2 %.

As document sets continue to grow, the proportion of documents contained in standard-sized pools shrinks. At some point, pooling's premise must become invalid. The test collection created in the Robust and HARD tracks in TREC 2005 showed that this point is not at some absolute pool size, but rather when pools are shallow relative to the number of documents in the collection [2]. With shallow pools, the sheer number of documents of a certain type fill up the pools to the exclusion of other types of documents. This produces judgments sets that are biased against runs that retrieve the less popular document type, resulting in an invalid evaluation.

Several TREC tracks have investigated new ways of sampling from very large documents sets to obtain judgment sets that support fair evaluations. The primary goal of the Terabyte track that was part of TRECs 2004–2006 was to investigate new pooling strategies to build reusable, fair collections at a reasonable cost despite collection size. The Million Query track (TRECs 2007–2009) was a successor to the Terabyte track in that it had the same goal, but a different approach. The Common Core track of TREC 2017 and 2018 used multi-arm bandit optimization techniques to select documents to be judged. TRECs since 2019 have experimented with a different approach based on the University of Waterloo's HiCAL [1] system to select the judgment set in the Deep Learning track. Each of these methods reduces the number of relevance judgments made, but can bias the test collection more towards submitted systems, introduce logistical challenges, or both.

2.2 Evaluation

Retrieval runs on a test collection can be evaluated in a number of ways. In TREC, ad hoc tasks that use pooling are evaluated using the `trec_eval` package [3]. This package reports about 85 different numbers for a run, including *recall* and *precision* at various cut-off levels plus single-valued summary measures that are derived from recall and precision. Precision is the proportion of retrieved documents that are relevant (number-retrieved-and-relevant/number-retrieved), while recall is the proportion of relevant documents that are retrieved (number-retrieved-and-relevant/number-relevant). A cut-off level is a rank that defines the retrieved set; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list. The `trec_eval` program reports the scores as averages over the set of topics where each topic is equally weighted. (An alternative is to weight each relevant document equally and thus give more weight to topics with more relevant documents. Evaluation of retrieval effectiveness historically weights topics equally since all users are assumed to be equally important.)

Precision reaches its maximal value of 1.0 when only relevant documents are retrieved, and recall reaches its maximal value (also 1.0) when all the relevant documents are retrieved. Note, however, that these theoretical maximum values are not obtainable as an average over a set of topics at a single cut-off level because different topics have different numbers of relevant documents. For example, a topic that has fewer than ten relevant documents will have a precision score at ten documents retrieved less than 1.0 regardless of how the documents are ranked. Similarly, a topic with more than ten relevant documents must have a recall score at ten documents retrieved less than 1.0. For a single topic, recall and precision at a common cut-off level reflect the same information, namely the number of relevant documents retrieved. At varying cut-off levels, recall and precision tend to be inversely related since retrieving more documents will usually increase recall while degrading precision and vice versa.

3 TREC 2021 Tracks

TREC's track structure began in TREC-3 (1994). The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups. Table 2 lists the different tracks that were in each TREC, the number of groups that submitted runs to that track, and the total number of groups that participated in each TREC.

This section describes the tasks performed in the TREC 2021 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 Clinical Trials

The Clinical Trials track is the successor to Precision Medicine, and is new in 2021. The goal is to identify appropriate clinical trials given a (mock) patient's electronic health record. The vast majority of clinical trials fail to recruit sufficient patients or to recruit them in time for the study. If patients can be matched to appropriate clinical trials efficiently then more trials may be run. There is already sizeable research on the trial matching problem using structured health record data, and so the Clinical Trials track focuses on what can be done with natural language text appearing in the record.

The document collection for this track is a snapshot of the `clinicaltrials.gov` registry from April, 2021, with 375,000 clinical trial descriptions in XML format. The seventy-five topics each consist of a 5-10 sentence patient case description that mimics an admission statement that would be found in an electronic health record. These case descriptions were composed by medical clinicians and not taken from real people's records. Figure 1 shows an example topic.

Assessments for this track were done by the Department of Medical Informatics at Oregon Health and Science University. The four highest-priority runs from each team were pooled to a depth of 10, for a total of 35,832 relevance judgments. Retrieved trials were judged as either "eligible", meaning that the patient met the inclusion criteria and did

Table 2: Number of participants per track and total number of distinct participants in each TREC

Track	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18	'19	'20	'21
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																									
Merging				3	3																									
Filtering				4	7	10	12	14	15	19	21																			
Chinese					9	12																								
NLP					4	2																								
Speech						13	10	10	3																					
Xlingual						13	9	13	16	10	9																			
High Prec						5	4																							
VLC							7	6																						
Query							2	5	6																					
QA								20	28	36	34	33	28	33	31	28								14	16	5				
Web								17	23	30	23	27	18					26	24	16	12	15	10							
Video										12	19																			
Novelty											13	14	14																	
Genomics												29	33	41	30	25														
HARD												14	16	16																
Robust												16	14	17																
Terabyte													17	19	21															
Enterprise														23	25	20	16													
Spam														13	9	12														
Legal															6	14	15	14	17	11										
Blog															16	24	25	11	16											
Mlln Query																11	7	8												
Feedback																	15	20	7											
Chemical																		8	4	9										
Entity																		13	16	10										
Session																			10	13	10	6	11							
Crowd																				11	8	4								
Medical																				29	24									
Microblog																				58	36	20	23	16						
Contextual																					13	19	17	12	13					
KBA																					11	13	11							
Temporal Summ																						7	6	12						
Federated																						11	12							
Clinical																							26	36	26					
Dynamic Domain																								7	6	3				
Tasks																								5	4	2				
Recall																								10	5					
RTS																									19	18	6			
OpenSearch																									9	3				
CAR																										7	9	7		
Core																										14	11			
Precision Medicine																										32	27	15	16	
CENTRE																											1			
Incident Streams																											11	15	6	5
News																											7	13	10	8
CAsT																												21	15	15
Misinfo																												4	9	7
Deep Learning																												16	25	20
Fair Ranking																												5	6	4
Podcast																													14	11
Clinical Trials																														26
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117	107	95	56	67	75	121	83	60	75	87	74	67	57	66	84	74

not meet any exclusion criteria; "excluded", meaning that the patient met the inclusion criteria but was excluded by one or more exclusion criteria; or "not relevant".

The main metric for Clinical Trials is nDCG at ranks 5 and 10, with a gain value of 2 for "eligible" documents and 1 for "excluded" documents. Precision at ranks 5 and 10, R-Precision, and mean reciprocal rank were also reported, counting only "eligible" documents as relevant. This reflects the behavior of real users, who are extremely dissatisfied when shown trials that they are explicitly excluded from. 113 runs were submitted to the Clinical Trials track from 26

Patient is a 45-year-old man with a history of anaplastic astrocytoma of the spine complicated by severe lower extremity weakness and urinary retention s/p Foley catheter, high-dose steroids, hypertension, and chronic pain. The tumor is located in the T-L spine, unresectable anaplastic astrocytoma s/p radiation. Complicated by progressive lower extremity weakness and urinary retention. Patient initially presented with RLE weakness where his right knee gave out with difficulty walking and right anterior thigh numbness. MRI showed a spinal cord conus mass which was biopsied and found to be anaplastic astrocytoma. Therapy included field radiation t10-l1 followed by 11 cycles of temozolomide 7 days on and 7 days off. This was followed by CPT-11 Weekly x4 with Avastin Q2 weeks/ 2 weeks rest and repeat cycle.

Figure 1: The first topic from the 2021 Clinical Trials collection

groups.

3.2 Conversational Assistance

The Conversational Assistance track (CAsT) started in 2019. Its focus is on systems that support conversational information seeking such as automated personal assistants. Such systems need to be able to maintain information about the state of the dialogue ("context") to properly interpret the current information need.

The system task is operationalized as a passage retrieval task where each topic consisted of a series of questions and systems returned a ranked list of passages for each question ("turn") in the series. The topics were developed by the track organizers, who provided a raw version of the series, which was the version required to be used for automatic runs; a manually rewritten version that made the implicit context explicit in each question; and an automatically rewritten version using a T5 transformer. Figure 2 shows three turns with all three question versions for an example topic from the test set. The manual version of the series was used by the assessors during judging and by some manual runs submitted to the track, and can be used as training data for future systems. The test set consisted of 25 topics with a mean of nine turns per topic.

The document set used in the track was the union of MS MARCO,[1] the KILT benchmark version of Wikipedia,[2] and the TREC *Washington Post* collection, version 4.[3] Systems returned a ranked list of up to 1000 passages for each question.

A problem was uncovered at the start of assessing: the track-provided tool to break the collections into passages produced different output depending on the version of a library used. Since canonical passage numbering is essential for valid evaluation and there was no way of knowing which numbering was used in which runs, we were forced to assess documents rather than passages. Assessing documents takes more time, which in turn forced us to use smaller pools than originally planned. The final relevance judgments (qrels) are based on depth-7 pools across all submissions, including baselines provided by the organizers. Not all questions in all topics were able to be judged; please see the Conversational Assistance track overview for details.

NIST assessors made judgments for nineteen topics, with most of those topics judged through question eight.

Each passage was judged on the 5-point scale used in 2020:

4 Fully Meets: The passage is a perfect answer for the turn. It includes all of the information needed to fully answer the turn in the conversation context. It focuses only on the subject and contains little extra information.

3 Highly Meets: The passage answers the question and is focused on the turn. It would be a satisfactory answer if Google Assistant or Alexa returned this passage in response to the query. It may contain limited extraneous information.

[1] http://www.msmarco.org

[2] https://github.com/facebookresearch/KILT/

[3] https://trec.nist.gov/data/wapost/

Turn 1:

Raw utterace : "I just had a breast biopsy for cancer. What are the most common types?"

Manual Rewrite : "I just had a breast biopsy for cancer. What are the most common types of breast cancer?"

Automatic Rewrite : "What are the most common types of cancer in regards to breast biopsy?"

Turn 2:

Raw utterace : "Once it breaks out, how likely is it to spread?"

Manual Rewrite : "Once it breaks out, how likely is lobular carcinoma breast cancer to spread?"

Automatic Rewrite : "Once the cancer breaks out, how likely is it to spread?"

Turn 3:

Raw utterace : "How deadly is it?"

Manual Rewrite : "How deadly is lobular carcinoma in situ?"

Automatic Rewrite : "How deadly is LCIS?"

Figure 2: An example question sequence from the CAsT test set. Automatic systems chose to process either the raw questions or the automatically-rewritten questions. Manual systems could use the manually-rewritten questions. Assessors judged against the manually-rewritten questions.

2 Moderately Meets: The passage answers the turn, but is focused on other information that is unrelated to the question. The passage may contain the answer, but users will need extra effort to pick the correct portion. The passage may be relevant, but it may only partially answer the turn, missing a small aspect of the context.

1 Slightly Meets: The passage includes some information about the turn, but does not directly answer it. Users will find some useful information in the passage that may lead to the correct answer, perhaps after additional rounds of conversation (better than nothing).

0 Fails to Meet: The document is not relevant to the question and is unrelated to the target query.

The rubric was adjusted to refer to documents. The main effect of this was that concerns regarding "extraneous information" were ignored.

Each submission was processed such that the first occurrence of a document in a ranked list was retained and all other occurrences removed. Removing a document moved subsequent documents higher in the ranking and shortened the overall list to fewer than the allowed maximum number retrieved (1000). The de-duplicateds ranked lists were used for pooling and were the lists evaluated.

The main evaluation measures for the track are nDCG@3, nDCG@5, NDCG@500, and AP@500. (500 was used instead of 1000 due to the shortening of ranked lists by de-duping.)

The track received 50 runs from 15 participating teams.

3.3 Deep Learning

The Deep Learning track focuses on a traditional ad hoc retrieval task in an environment where there is a large, labeled training set. The goal is to explore the trade-offs between neural- and dense-retrieval approaches on the one hand and traditional rankers on the other. In 2021, the track refreshed the corpus to unify the document and passage datasets.

168329	does light intensity or concentration of carbon dioxide have a higher rate of photosynthesis
300025	how many whales are caught in fishing nets off california
505390	supartz injections what does it made of
615176	what court has appellate jurisdiction circuits based on population distribution
818583	what is the difference between the range rover and the range rover sport
935353	when and where did the battle of manassas take place
952284	when is the best time to fish with a crawfish color bait
1006728	which cerebral lobe of the brain is involved in our ability to see?
1109840	what law is concerned with the safety and health conditions in the majority of private-sector industries?
1121909	what are the instruments in a woodwind quintet

Figure 3: Example questions from the set of questions NIST assessors judged for the Deep Learning track.

This presented a significant challenge to participants as the new data was made available only a month before the submission deadline.

The track had two tasks, Document Ranking and Passage Ranking. For each task, participants could either do their own retrieval from the full collection or re-rank an initial retrieved set provided by the track organizers. Both tasks used the same set of 477 test questions, a sample of which are shown in Figure 3.

Both tasks used version 2 of the MS MARCO dataset.[4] For the Passage Ranking task, the document set was about 140 million passages extracted from web pages using an algorithm to try to identify the most promising passage independent of a query. The Document Ranking task set was around 11 million documents. There was a set of training queries and relevance judgments, two sets of development queries and judgments, and the queries and judgments from TREC 2019 and 2020 as validation data.

Partly because of the shift from the data used on the official MS MARCO leaderboard to the new version of the collection, there were new rules about what data was permitted to be used by participants. The passage-document mapping was permitted. The ORCAS click data[5] was prohibited, as well as any information that mapped documents and passages in the new collection back to the old collection. Aside from ORCAS, the topics and relevance judgments from previous DL track were permitted. Participants were prohibited from using other MS MARCO resources, such as the QnA or NLGEN data. A segmented document collection and an augmented passage collection were provided by the organizers.

One goal of the track was to create traditional ad hoc test sets from this data. To this end, NIST assessors created much more dense judgments for much smaller sets of questions. NIST originally selected 57 questions from the test set of 477 to be judged for both tasks. The topics were selected by observing the behavior of submitted Document Ranking task runs on the entire test set when using the sparse MS MARCO judgments to evaluate runs. Test questions that had median (across submissions) mean reciprocal rank (MRR) scores of 0.0 or 1.0 were eliminated. Queries for assessment were randomly selected from the remainder such that half were long queries (ten or more words) and half were short.

After question selection, we pooled the top 10 results across all runs in the task. The assessor judged these pool documents first, then another 20-25 documents selected to be judged using the University of Waterloo's CAL [1] system.[6] CAL uses the current set of judgments to build a relevance model using very basic tokenization and logistic regression, and then selects the unjudged document most likely to be relevant as the next document to judge. A topic continued to be assessed using CAL if

- more than half of the documents judged for the topic so far had been judged relevant; or
- if fewer than 150 documents had been judged so far; or

[4] `http://www.msmarco.org`, see `https://microsoft.github.io/msmarco/TREC-Deep-Learning` for details. Version 2 is not the same as the MS MARCO leaderboard dataset.

[5] ORCAS: Open Resource for Click Analysis in Search, `https://microsoft.github.io/msmarco/ORCAS.html`

[6] "CAL" refers to the active learning document selection process; "HiCAL" is a web-based tool that uses CAL under the hood.

- if the topic had not yet had any documents suggested by CAL judged (that is, each topic had at least one CAL iteration); or
- if more than 20% (10% in early iterations) of the documents in the most recent previous iteration had been judged relevant.

Once a topic exited the process, it was never restarted. Assessment for the passage ranking task followed the same procedure and started with the same topics. All 57 topics completed the process for document ranking, but for passage ranking, four were dropped as they had less than five relevant passages.

Documents in the Document Ranking task were judged on a four-point scale of *Irrelevant* (0), *Relevant* (1), *Highly Relevant* (2), and *Perfect* (3) where all but Irrelevant were treated as relevant in CAL and in computing binary-relevance-based measures. For the Passage Ranking task, passages were judged on a four-point scale of *Irrelevant* (0), *Related* (the passage is on-topic but does not answer the question) (1), *Highly Relevant* (2), and *Perfect* (3). In this task, only Highly and Perfectly Relevant were considered to be relevant for binary measures and by CAL, though nDCG scores did use a gain value of 1 for the Related passages.

Most topics have very many relevant (presumably because the document set is much larger than in previous years), and as a result the collection is likely not a reusable collection. A variant of the "Leave-Out-Uniques" test [10] that accommodates the use of CAL shows that some Document Ranking task submissions experience large swings in evaluated effectiveness if their team's uniquely-retrieved relevant documents are removed from the top-10 pools. This suggests future systems may not be evaluated fairly with that collection, and similar concerns are appropriate for the Passage Ranking task collection, too. Evaluation results that might be biased can be detected by the presence of many unjudged passages in a ranking. This does not affect track submissions for high-precision measures such as P@10 or ndcg@10 since all submissions contributed to the depth-10 pools. But so many relevant means these high-precision measures obtain very high scores. While that can be gratifying, it also makes the evaluation less stable because the effective number of topics that are distinguishing the runs from one another in the average is reduced.

Twenty teams participated in the Deep Learning track, submitting 63 Passage Ranking runs and 66 Document Ranking runs.

3.4 Fair Ranking

The Fair Ranking track recognizes that search systems affect both consumers and producers of information, and looks to build systems that provide fair exposure of producers while still returning relevant information. The initial iterations of the track used academic papers from Semantic Scholar, and sought to encourage the development of systems that would rank search results fairly with respect to pre-specified groups of authors, and then in 2020 against arbitrary group definitions.

For 2021, the track initiated a partnership with the Wikimedia Foundation. The track considered two types of users in the Wikipedia environment. The first type of user is a WikiProject coordinator; this is a person or group of people who shepherd a project such as Wikipedia, Wikidata, or any of the other projects hosted at the Foundation. Coordinators keep an eye on the overall direction and quality of the project, and part of that is keeping tabs on pages that need to be worked on. This user was the model for the track's first task. Task two represented a second type of user, a Wikipedia editor who may be looking for pages that would benefit from their effort.

Both tasks shared the same documents, topics, and definition of fairness. The documents were English Wikipedia articles, and the topics focused on topics in individual WikiProjects that would have relevant pages in Wikipedia. Fairness of exposure was with respect to geographic focus and (in the case of biography pages) the gender of the subject of the page. An automated measure of article quality was also provided as a characteristic that should not receive fair exposure.

For the Coordinators task, the output of the search was a single ranking of pages that were both relevant to the project and fair with respect to exposure across the fairness categories. For the Editors task, the output was a sequence of rankings where systems needed to identify relevant pages that needed work, in addition to the fairness criterion. The notion of a set of rankings comes from the idea that a project would have a "standing query" representing information relevant to the project, and different editors would receive different rankings targeted to them, and that whole set of rankings was a candidate for exposure fairness.

id	a topic identifier
title	the WikiProject title
keywords	a collection of search keywords forming the query text
scope	a textual description of the project scope, from its project page
homepage	the URL for the WikiProject. This is provided for attribution only and is not intended for system use.

Figure 4: Structure of Fair Ranking topics.

NIST assessors reviewed pooled Wikipedia pages for relevance. There were two sets of pools: a depth-20 pool of Task 1 rankings, and a depth-5 pool of the first 25 Task 2 rankings from a submission, for 75% of the queries. Quality, gender, and geography categories were sourced automatically.

The metrics for the Fair Ranking track attempt to measure the degree to which the system retrieved relevant documents, and exposed documents in a way that is fair across the fairness categories (geography and gender). One difficulty discussed in the overview paper is that Wikipedia itself has biases in these categories, and as such measuring fairness here is difficult.

Task 1 used attention-weighted rank fairness as a metric, which accumulates exposure across fairness groups. This metric was multiplied by nDCG for relevance to produce the final metric. Task 2 computed expected exposure. Metrics for fairness are an active area of research and we refer the reader to the track overview for more details.

The track received 24 runs from four participating teams.

3.5 Health Misinformation

This is the third year of the Health Misinformation track, a track that investigates how search engines can support sound decision making.[7] The track used search for consumer health information over web documents as its use case since this is a common, important scenario. As operationalized in the track, the search task is similar to a traditional ad hoc search task, with the difference that retrieving on-topic but *incorrect* information is more strongly penalized than simply retrieving non-relevant information.

The document set used in the track was HuggingFace's "noclean" version of Google's C4 dataset[8]. As a typical web collection, it contains documents of widely varying credibility and quality, including documents that actively try to mislead.

Fifty topics were created by track organizers. Two example topics are shown in Figure 5. The topic format is similar to traditional TREC topics except for the addition of three fields, "stance", "evidence", and "disclaimer". The stance indicates a position on whether the treatment is helpful or unhelpful for the health issue. The evidence gives support for that stance. The disclaimer reads, "We do not claim to be providing medical advice, and medical decisions should never be made based on the stance we have chosen. Consult a medical doctor for professional advice." The set of topics was selected to have equal representation of helpful and unhelpful issue-intervention pairs.

Document judging was performed by NIST assessors. Initial judgment pools were created over all runs submitted to the track to depth 35. Documents were judged on three separate dimensions. The first dimension was Usefulness, where documents were judged on a three-point scale of *not-useful* (0), *useful* (1) , and *very-useful* (2) based on whether the document provided information about the health issue, the treatment, or both that would help the user make a decision about the search topic. Provided the usefulness judgment was either useful or very useful, the assessor then judged the Supportiveness and Credibility dimensions. Supportive was judged with respect to *what the document itself claimed*—regardless of any knowledge the assessor might have to the contrary—and was judged as either *dissuades* (0), *neutral* (1), or *supportive* (2). Credibility was judged as either *low* (0), *good* (1), or *excellent* (2) and was based on the assessor's opinion of the document taking into consideration its source and the supporting evidence it contained.[9] The assessors were able to complete judgments for 35 topics.

[7] In 2019 this track started as the "Decision" track, but changed its name in 2020.

[8] `https://huggingface.co/datasets/allenai/c4`

[9] There are some cases where the assessor missed a criterion by mistake; those "judgments" are indicated in the qrels as -2.

Number:	105
Query:	put ice on a burn
Description:	Should I apply ice to a burn?
Narrative:	Many people commonly put on ice on burns in an attempt to stop the burning and pain. A very useful document would discuss the effectiveness of using ice to treat burns. A useful document would help a user decide if putting ice on burns is a recommended treatment by providing information on recommended treatments for burns and may not discuss ice as a treatment, or a useful document may discuss benefits or concerns for application of ice to skin.
Disclaimer:	We do not claim to be providing medical advice, and medical decisions should never be made based on the stance we have chosen. Consult a medical doctor for professional advice.
Stance:	unhelpful
Evidence:	https://www.uwhealth.org/news/the-right-way-to-treat-burns
Number:	106
Query:	vitamin b12 sun exposure vitiligo
Description:	Can vitamin b12 and sun exposure together help treat vitiligo?
Narrative:	Vitiligo is characterized by discolored patches on the skin because the skin loses its pigment cells (melanocytes). A very useful document would discuss the effectiveness of vitamin b12 and sun exposure together for vitiligo. A useful document help a user make a decision about using the combination of vitamin B12 and sunlight to treat vitiligo, and would provide information on recommended treatments for vitiligo, use of vitamin b12 and sun exposure, or both.
Disclaimer:	We do not claim to be providing medical advice, and medical decisions should never be made based on the stance we have chosen. Consult a medical doctor for professional advice.
Stance:	helpful
Evidence:	https://pubmed.ncbi.nlm.nih.gov/9394983/

Figure 5: Example topics from the Health Misinformation track test set.

These judgments were used to create six different derived qrels files, based on different combinations of usefulness, correctness, and credibility. The resulting split is complex, because an answer is "correct" if it supports a helpful treatment or if it dissuades the use of an unhelpful treatment. In addition to reporting nDCG, precision at 10, and average precision using these qrels, the track computed a "compatibility" metric, described in the track overview and at `https://github.com/claclark/Compatibility`.

Seven participating teams submitted 71 runs to the Health Misinformation track.

3.6 Incident Streams

The Incident Streams track is motivated by the rise of social media use during emergency situations. Emergency service operators are now expected to monitor various social media channels and respond to requests found there. However, the emergency services do not have the tools or manpower to categorize, cross-reference, and verify the information spread across multiple platforms. The track is designed to focus research on technologies to automatically process social media streams during emergencies.

The track debuted in TREC 2018, where systems categorized tweets into a set of high-level information types for 15 different events. In subsequent editions, the track has had two evaluation phases, A and B, adding more annotated events in each edition. In 2021, the track distributed a test set with 26 crisis events. 2021-A produced partial assessments for a subset of events, which supported a leaderboard, and 2021-B reviewed deeper pools for more of the

High-level Type	Description
Request-GoodsServices	The user is asking for a particular service or physical good.
Request-SearchAndRescue	The user is requesting a rescue (for themselves or others).
Request-InformationWanted	The user is requesting information.
CallToAction-Volunteer	The user is asking people to volunteer to help the response effort.
CallToAction-Donations	The user is asking people to donate goods/money.
CallToAction-MovePeople	The user is asking people to leave an area or go to another area.
Report-FirstPartyObservation	The user is giving an eye-witness account.
Report-ThirdPartyObservation	The user is reporting information from someone else.
Report-Weather	The user is providing a weather report (current or forecast).
Report-EmergingThreats	Report of a potential problem that may cause future loss of life or damage.
Report-MultimediaShare	The user is sharing images or video.
Report-ServiceAvailable	The user is reporting that someone is providing a service.
Report-Factoid	The user is reporting some facts, typically numerical.
Report-Official	An official report by a government or public safety representative.
Report-CleanUp	A report of the clean up after an event.
Report-Hashtags	Reporting which hashtags correspond to each event.
Report-NewSubEvent	The user is reporting a new occurrence that public safety officers need to respond to.
Report-Location	The post contains information about the user or observer location.
Other-News	The post provides/links to continuous coverage of the event.
Other-Advice	The author is providing some advice to the public.
Other-Sentiment	The post is expressing some sentiment about the event.
Other-Discussion	Users are discussing the event.
Other-Irrelevant	The post is irrelevant, containing no information about the event.
Other-ContextualInformation	The post is generic news, e.g., reporting that the event occurred.
Other-OriginalEvent	The responder already knows this information.

Figure 6: High-level information types used in the Incident Streams 2021 track. Systems categorized tweets in an event's tweet stream by assigning one or more labels to each tweet.

test events.

Systems were given the ontology of high-level information types shown in Figure 6 plus a time-ordered stream of tweets for each event. The systems' task was to label each tweet with all of the information types that applied to it. This categorization was required to be done on-line; that is, each stream needed to be processed in order as if the event were occurring in real time, and the labels for the current tweet had to be assigned before the next tweet was processed.

NIST assessors processed each stream as well, assigning all applicable labels to each tweet. The assessors also indicated categories for images present in a tweet: whether the image was a satellite image, a weather forecast, an infographic, a photo of a weather event, if the image depicted victims, damage, crowds, or recovery efforts, if the image was a screenshot, and whether the image was informative.

Assessors also assigned an information priority level to each tweet. Runs are evaluated by how closely the set of tags the run assigned to each tweet match the set of labels the assessors assigned to that tweet, conditioned on actionable and non-actionable categories. In addition, the information priority level assigned by assessors is used to define a new metric, Accumulated Alert Worth, that aims to capture how well systems perform as an alerting service.

Five groups participated in the Incident Streams track, submitting 14 runs for the 2021-A deadline, and 17 runs for the leaderboard and 2021-B. This is the last year of the Incident Streams track.

3.7 News

The News track is coordinated in partnership with The Washington Post and looks to study the search needs of news readers and writers in the current news environment.

Topic: 936
Title: Nora Ephron dies at 71
Desc: I'm looking for information on the passing of author and screenwriter Nora Ephron.
Narr: Please provide information as to when author and screenwriter Nora Ephron died and of what causes. Biographical details of her life and career are all relevant. Reactions to her passing by friends and the media are relevant as well.
Subtopics:

1. Find details of Nora Ephron's life and accomplishments.
2. What are Nora Ephron's most well-known works?
3. Did Nora Ephron know that Mark Felt was "Deep Throat" of Watergate fame?

Target article: [f831cae6-bfa4-11e1-9ce8-ff26651238d0, https://...]

Nora Ephron, prolific author and screenwriter, dies at age 71

"Take notes," Nora Ephron's mother advised her as a child. "Everything is copy."

Her mother, a Broadway playwright and Hollywood screenwriter, imbued Ms. Ephron with a razor-sharp self-awareness and the ambition to transform workaday absurdities, cultural idiosyncrasies, romantic foibles and even marital calamity into essays, novels and films brimming with invitingly mordant wit. She credited her mother with bestowing "this kind of terrific ability, not to avoid pain but to turn it over and recycle it as soon as possible."

Nora Ephron, who gained a devoted following for her perceptive, deeply personal essays and parlayed that renown into a screenwriting career of wistful romantic comedies such as "When Harry Met Sally" and "You've Got Mail," the marital exposé "Heartburn" and the whistleblower drama "Silkwood," died June 26 at a hospital in New York. She was 71.

Figure 7: Example topic and part of its target document from the News track test set.

The track focuses on two tasks, background linking and wikification.[10] The document collection for both tasks is the *TREC Washington Post* collection.[11] Both tasks also used the same set of 51 topics. For background linking, a topic is an article from the document collection that is designated as the target article. Systems retrieve other articles that provide important context or background information for the target. The goal is to help a reader understand or learn more about the story or the main issues of the target article using the best possible sources. For wikification, the systems were to identify spans of text in the target article and propose links from those spans to either other Washington Post articles or Wikipedia. The links should ideally be those that would help the reader understand the broader context of the article; this is in contrast to other wikification tasks that seek to link all entities in a document.

The Washington Post itself makes recommendations like these, but they are manually curated. Articles frequently have interstitial links to other articles on the same story. In addition, the Post creates story-specific link boxes that appear at the end of some articles, with links to different aspects of the story. To capture this latter idea, the track included title/description/narrative blocks, and a small set of subtopics. The subtopics were not used for diversity ranking; rather, systems submitting results to the subtopics subtask of the background linking task returned a separate ranked list for each subtopic. Figure 7 shows a news track topic statement with a portion of the target article.

For the background linking task (main and subtopics subtasks), systems returned up to 100 documents per topic. The top 20 documents per run were pooled for assessment, with assessors marking each pooled article on the following scale:

0 : The linked document provides little or no useful background information.

1 : The linked document provides some useful background or contextual information that would help the user understand the broader story context of the query article.

2 : The document provides significantly useful background ...

[10] Wikification is automatically linking segments of text to other pages, in the style of Wikipedia.
[11] `https://trec.nist.gov/data/wapost/`

3 : The document provides essential useful background ...

4 : The document *must* be linked otherwise critical context is missing.

The primary metric for the task is nDCG@5 with the gain values as 2^r where r is the value of the relevance label except that non-relevant documents (label 0) contribute no gain.

The track only received wikification task submissions from one group, so NIST staff reviewed top-10 pools. For wikification, a pool entry was composed of an extent in the article and a target link, which could either point to a Wikipedia page or another document in the Washington Post collection. Each anchor-link pair was judged according to the following rubric:

- Is the anchor **sensible**, that is, does it break on word boundaries and seem reasonable lexically?
- Is the anchor **useful**, that is, does it identify an extent which might be usefully linked to another resource to help the reader better understand the article?
- Does the target **match**, that is, is the linked article what a user would reasonably expect to be linked from that anchor?
- Is the target **helpful**, that is, does the linked article actually provide useful background that would help the reader better understand the article?

A retrieved anchor-link pair needed to answer 'yes' to all four questions to count as relevant for the task.

The track received 47 runs from 8 participants: 22 runs for background linking, 20 for the subtopics subtask, and 5 for wikification.

3.8 Podcast

The Podcast track began in 2020 to explore search and summarization tasks in the podcast domain. Podcasts are audio programs distributed as digital files. A podcast comprises a series of episodes, may or may not be on a focused topic, and may have guests or just be a monologue performance by the podcast creator. Podcasts range in quality from professional audio productions to a pair of teenagers with a phone, and in topic from journalism to scientific research to improvised dialogue.

The dataset for the Podcast track was provided by Spotify, and consists of over 100,000 English-language podcast episodes. The episodes are available as audio files or as an ASR transcript generated through Google's Speech-to-Text API. Metadata includes descriptions of the podcast and of the episode, provided by the podcast creator.

The track featured two tasks, retrieval and summarization. The retrieval task topics included "query" and "description" fields. Topics were also of two types: topical, meaning a user was interested in finding podcast episodes about a topic, and known-item, where they were looking for a specific podcast episode. Figure 8 shows three examples.

The episodes were automatically marked with boundaries at each minute, and systems were to retrieve two-minute segments by indicating the start of the segment. The retrieved segments were pooled to depth 20 and judged by NIST assessors for relevance along a scale of perfect-excellent-good-fair-bad. The perfect level was reserved for known-item topics for segments that returned a perfect entry point to the desired podcast episode. For topical topics, the level indicated both the relevance of the result and the quality of the entry point.

A new twist to the retrieval task for 2021 was that systems were required to return four rankings for each topic: a normal "relevance ranking", and three re-orderings of the relevance ranking along the following criteria:

Entertaining: the segment is topically relevant and the topic is presented in a way which the speakers tend to be entertaining to the listener, rather than informative or evaluative.

Subjective: the segment is topically relevant to the topic description and the speaker or speakers explicitly and clearly express a polar opinion about the query topic, so that the approval or disapproval of the speaker is evident in the segment.

Discussion: the segment is topically relevant to the topic description and includes more than one speaker participating with non-trivial topical contribution.

id	97
query	smuggling
type	topical
description	I want to hear stories about smuggling. General discussion about smuggling without reference to actual events are not relevant.
id	98
query	nobel prize laureates
type	topical
description	I want to hear about Nobel prize laureates. Biographies including both personal and professional life is relevant. The segment must name the laureate to be relevant.
id	99
query	samin nosrat
type	known-item
description	I heard that Chef Samin Nosrat makes a surprise appearance on an episode of The Cut and I want to find it.

Figure 8: Three example Podcast track topics. In the test set, 40 are topical and 10 are known-item topics.

All four rankings were included in the depth-20 pool. Assessors judged every pool segment against the four criteria, so as a practical matter the pool depth was more than 20 for both relevance and the re-ranking criteria.

For the summarization task, 1000 podcast episodes were provided with both audio and transcripts, and systems were required to return a textual summary as well as a one-minute audio clip. The intended use of both the summary and the clip were to help the user decide if they would want to listen to the podcast episode. No training summaries were provided, but podcasts and episodes included descriptions which served as baselines in the 2020 track. For 2021, the baseline was the transcript and audio of the first minute of the episode.

The 11 teams participating in the Podcast track submitted 27 retrieval runs and 12 summarization runs.

4 Future

TREC will continue in 2022. The News track is ending, the Podcast track is taking a hiatus, and the Incident Streams track is changing to "Crisis Facts", a turn back to its temporal summarization roots but with some added twists. In addition, TREC 2022 will have a new cross-language retrieval task. Planning sessions for all tracks were included in their presentation block during the virtual program.

Acknowledgments

Much of the text and all of the format of this paper were taken from Ellen Voorhees' 2019 TREC Overview. I am grateful to be allowed to reuse her text. Any errors in this document are my own. The descriptions of the tracks' goals and tasks were adapted from the track guidelines and overview papers authored by the track coordinators. My thanks to the coordinators who by their volunteer efforts make the variety of different tasks addressed in TREC possible.

Disclaimer

Certain companies or commercial products are identified in various papers in the TREC proceedings, including this one, in order to describe the TREC process adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the companies or products identified are necessarily the best available for the purpose.

References

[1] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. A system for efficient high-recall retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1317–1320, 2018.

[2] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10:491–508, 2007.

[3] Chris Buckley et al. trec_eval IR evaluation package. Available from `https://github.com/usnistgov/trec_eval.git`.

[4] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems. Two volumes, Cranfield, England, 1968.

[5] Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, October 1996. NIST Special Publication 500-236.

[6] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

[7] K. Spärck Jones and C. J. 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.

[8] Karen Spärck Jones. *Information Retrieval Experiment*. Butterworths, London, 1981.

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

[10] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.