Overview of TREC 2023

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

1 Introduction

TREC 2023 is the thirty-second 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 2023 meeting was held at the National Institute of Standards and Technology (NIST) November 14–17, 2023.

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 2023 contained eight tracks:

Authoring Tools for Multimedia Content (AToMiC): The AToMiC track is about text-to-image and image-to-text search. The user task is that of a Wikipedia author looking for images to illustrate a given page section, or looking for such a section for a given picture.

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

CrisisFACTs: The CrisisFACTs track is about event summarization. During a large emergency event like a wildfire or a hurricane, public safety personnel complete daily status updates that include various facts about the progress of the event. In CrisisFACTs, participants build systems to try and identify those facts automatically in multiple text streams.

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.

Interactive Knowledge Aquisition (IKAT): IKAT, the successor to CAsT, supports creating data to support developing systems that engage users in open-domain, information-centric, conversational dialogues.

Product Search: The Product Search track is about customer-focused e-commerce search, such as would be found on a shopping website.

NeuCLIR: Cross-language search has returned to TREC after a twenty year hiatus. In cross-language search, the information needs are expressed in a different language than the documents being searched. For NeuCLIR, the topics are in English and the documents are in Russian, Chinese, and Farsi. The advancements in neural retrieval architectures are driving TREC to revisit this task.

Tip-of-the-Tongue Search (ToT): Tip-of-the-tongue searches are known-item searches where the user has a fuzzy, vague, incomplete, or only partially correct memory of what they are looking for.

Forty-four groups from fifteen different countries participated in TREC 2023. Table 1 lists the participating organizations.

Table 1: Organizations participating in TREC 2023

Allen Institute for AI	Waseda University
Academia Sinica and National Chengchi University	Siena College Institute for Artificial Intelligence
University of Massachusetts Amherst	University of Amsterdam
University of Chinese Academy of Sciences	University of Milano-Bicocca, Department of Informatics
Carnegie Mellon University Language Technologies Institute	Smucker IR Research Group
CSIRO and Uni of QLD ielab	Toronto Metropolitan University
Commonwealth Scientific and Industrial Research Organisation	Delft University of Technology (TU Delft)
Politecnico di Torino	University of Waterloo (Clarke)
DoSSIER	Webis @ Jena, Leipzig, Weimar
Dalhousie University, University of Manitoba	Endicott College / University of North Carolina
University of Glasgow	University of Waterloo
Humanitarian Informatics Lab, George Mason University	Human Language Technology Center of Excellence, JHU
IDA Center for Computing Sciences	Jeonbuk National University
Indian Institute of Technology Delhi	Nagaoka University of Technology
Indian Institute of Technology (BHU) Varanasi	Melax Tech, now part of IMO
University of Amsterdam	Seoul National University ldilab
Information Sciences Institute	Universidade Federal de Minas Gerais
InfoSense Lab, Georgetown University	University of Maryland, HCIL
Marquette University	University of Glasgow Terrier Team
NeuralMind	University of Tsukuba
Technische Hochschule Nürnberg Georg Simon Ohm	York University
Recherche appliquée en linguistique informatique	INESC-ID: Instituto de Engenharia de Sistemas e Computadores, Lisboa

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.

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 [5, 9], 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 [7]. 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 [10].

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 [8]. 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 [6] and Zobel [11] 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 [3]. 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 [4]. 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.

Recently the emergence of large language models and highly tuned applications based on them has created a lot of interest in generative information seeking. This could be retrieval-augmented generation, where results from a search are summarized (instead of snippets on search results); question answering in situations where answers could be long or detailed (instead of passage retrieval); iterative conversational search (instead of query reformulation and suggestion); and more. Generative methods are harder to evaluate, and those evaluations are typically not reusable like a retrieval test collection. Therefore this is an important area of emerging research.

3 TREC 2023 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 2023 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 Authoring Tools for Multimedia Content (AToMiC)

AToMiC is a new track for 2023 focused on text-to-image and image-to-text search. The user adopted as a model is a Wikipedia editor, who may be looking for images to illustrate a section of a Wikipedia page, or looking for page sections that might be good matches for a given image.

These two activities, "suggestion" and "promotion", are the two tasks in the track. For the suggestion task, the topic is a section of a Wikipedia page, and systems returned a ranked list of images ideally ranked in order of usefulness as an illustration for that section. The promotion task is the mirror image: the topic is an image and systems returned ranked lists of Wikipedia passages.

Participation in the first year was low, and so the organizers produced a set of baselines to include in the pools. For the suggestion task, submissions were pooled to depth 30 and baselines to depth 25. For the promotion task, both groups were pooled to depth 30. These depths were based solely on estimates of how long the pools would take to be assessed. In fact, assessing for this task went very quickly, and we were able to judge pools for 74 suggestion topics and 61 promotion topics.

Four groups participated in the track and submitted a total of 27 experimental and baseline runs. Metrics for this task are still under discussion; NIST furnished full trec_eval scores and reported median scores for success at ranks 1

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	'22	'23
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												
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	10	
Misinfo																												4	9	7	4	
Deep Learning																												16	25	20	14	14
Fair Ranking																												5	6	4	5	
Podcast																													14	11		
Clinical Trials																														26	11	11
NeuCLIR																															12	6
CrisisFACTs																															7	10
AToMiC																																4
IKAT																																8
Product																																4
ToT																																11
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	48	46

and 5, and NDCG at 10.

Table 3: Two sample topics from the 2023 Clinical Trials collection. The template fields are defined in the guidelines.

Glaucoma Template	Patient1	Patient2
diagnosis	POAG	uveitic glaucoma
intraocular pressure	19 mmHg	22 mmHg
visual field	-	advanced damage
visual acuity	20/80	20/200
prior cataract surgery	no	no
prior LASIK surgery	no	no
comorbid ocular diseases	-	uveitis

3.2 Clinical Trials

The Clinical Trials track, now in its third year, is the successor to Precision Medicine. The goal is to identify appropriate clinical trials given a (mock) patient. 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. Instead of using narrative patient histories, this year the track switched to using a questionnaire such as might be used in an intake interview. For each of eight different disorders there is a questionnaire template which may have all or some fields completed.

The document collection for this track is a snapshot of the `clinicaltrials.gov` registry from May 2023, with 451,538 clinical trial descriptions in XML format. The fifty topics each consist of the specific disorder for the patient and the values for some or all of the questionnaire template fields. Figure 3 shows an example topic.

Assessments for this track were done by physicians trained in medical informatics. All runs from each team were pooled to a depth of 50, for a total of 36,791 relevance judgments. Retrieved trials were judged as either "eligible", meaning that the patient met the inclusion criteria and did 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 rank 10, with a gain value of 2 for "eligible" documents and 1 for "excluded" documents. Precision at rank 10 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. 33 runs were submitted to the Clinical Trials track from 11 groups, including 4 manual runs.

3.3 CrisisFACTs

The CrisisFACTs track is the latest edition of a series of tracks going back to TREC 2013 and the Temporal Summarization track, all focused on the problem of providing updates during events using news and/or social media sources. CrisisFACTs uses events and social media content (Twitter) from the Incident Streams track, and adds three additional feeds: news, Reddit, and Facebook.

The challenge of the track is to produce a minimally-redundant set of facts useful to first responders during emergencies. That user model is defined by an official status report form, the ICS-209, which gets completed daily by teams on the ground during a crisis. If any information from the four content feeds corresponds to something that would go in the ICS-209 form, then it should be marked as useful.

Systems are given a definition of the event, including a type and a link to a news article or the Wikipedia page about the event; a user profile that includes general and event-type-specific questions that the user needs to answer (see Figure 1); a set of days of the event to summarize; and the four content streams for that event. Systems take this information and return facts conveying needed information, a set of feed entry IDs as provenance, a timestamp and an importance score.

Facts were aggregated from all runs and clustered using BERTscore similarity. The assessors then reviewed the

```
{
  "queryID": "CrisisFACTS-General-q002",
  "indicativeTerms": "rail closed",
  "query": "Have railways closed",
  "trecisCategoryMapping": "Report-Factoid"
},
{
  "queryID": "CrisisFACTS-General-q003",
  "indicativeTerms": "water supply",
  "query": "Have water supplies been contaminated",
  "trecisCategoryMapping": "Report-EmergingThreats"
},
...,
{
  "queryID": "CrisisFACTS-Wildfire-q001",
  "indicativeTerms": "acres size",
  "query": "What area has the wildfire burned",
  "trecisCategoryMapping": "Report-Factoid"
},
{
  "queryID": "CrisisFACTS-Wildfire-q002",
  "indicativeTerms": "wind speed",
  "query": "Where are wind speeds expected to be high",
  "trecisCategoryMapping": "Report-Weather"
},
```


Figure 1: Two general and two event-specific queries from the CrisisFACTs track.

representatives of each fact "cluster", strung together as if in a temporally-ordered summary. The assessors could label a fact as useful, poor (somewhat useful but flawed), or redundant (the information was covered in a fact above in the summary).

For scoring, a set of gold-standard summaries were taken from real-world ICS-209 reports. Fact lists from runs are scored for comprehensiveness against the gold-standard and redundancy within the list. Additionally, summaries aggregated from system's fact outputs are compared to gold-standard summaries using ROUGE and BERTscore. This scoring approach is a way to measure generative systems automatically, and as such is deserving of study.

Ten groups submitted a total of 26 runs to the CrisisFACTs track.

3.4 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.

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 700 test queries, a sample of which are shown in Figure 4.

This year the track wanted to investigate whether queries could be synthesized to make a reliable, reusable test collection. The 2023 queries break down into three subsets: 200 actual queries submitted to a major search engine, 250 queries generated by a fine-tuned T5 model, and 250 generated using GPT4 with a prompt. The Deep Learning track overview provides more details and an analysis of the results.

Both tasks used version 2 of the MS MARCO dataset.[1] For the Passage Ranking task, the document set was about 138 million passages extracted from web pages using an algorithm to try to identify the most promising passage

[1] `http://www.msmarco.org`, see `https://microsoft.github.io/msmarco/TREC-Deep-Learning` for details. Version 2

Table 4: Examples from the set of questions NIST assessors judged for the Deep Learning track.

2002988	how to check hard disc memory on windows 10
2006929	what is a narwhal tusk made of
2026107	what causes people act for civil rights
2035895	cost of us visa uscis
2041662	what factors determines education
3052503	do lizards eat crickets
3069625	causes of valve prolapses
3100062	How can I use Adobe apps on my Linux Desktop?
3100709	Can I purchase a replacement property from a related party during a 1031 exchange?
3100877	What are the basic financial functions offered by banks and credit unions?

independent of a query. The Document Ranking task set was around 12 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 – 2022 as validation data.

Since MS MARCO v2 is different than the collection used on the MS MARCO leaderboard, a few rules were instituted in 2021 concerning data that participants were permitted to use. The passage-document mapping was permitted. The ORCAS click data[2] 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 in 2021 and remained available.

In 2022, NIST identified near-duplicates in the passage and document collections, using locality-sensitive hashing with word 9-grams and a Jaccard similarity threshold of around 0.85. This process produces a set of equivalence clusters where all members of the cluster are deemed to be near-duplicates. One member of each cluster is chosen to be the cluster representative. This reduced the passage collection by around 150k passages and the document collection to 11.9 million documents. Submitted runs had near-duplicates removed (such that the earliest ranked member of a near-duplicate cluster was replaced with the cluster representative) for pooling only. Relevance judgments were extended to all members of a near-duplicate class. This means that future users of the collection don't need to worry about the near-duplicates, and this seemed more workable than trying to release another version of MARCO.

One goal of the track was to create traditional, reusable ad hoc test sets from this data. This year we simplified the approach from what we did in 2022: runs were pooled to depth 10, and based on those pools, a subset of queries were selected for further assessing. That assessing was done with a NIST reimplementation of Cormack and Grossman's BMI continuous active learning model.

Six teams participated in the Deep Learning track, submitting 35 Passage Ranking runs and 5 Document Ranking runs.

3.5 Interactive Knowledge Acquisition (IKAT)

The Interactive Knowledge Assistance Track (IKAT) is the successor to the Conversational Assistance track (CAsT), which ran from 2019 to 2022. The focus of these tracks is on systems that support conversational information seeking. One inspiration for this is personal assistant "smart speakers", but imagine them able to help you with a complex, multistep search. Such systems need to be able to maintain information about the state of the dialogue ("context") to properly interpret the current information need.

For the first year of IKAT, the biggest change is the addition of the "Personal Text Knowledge Base", or PTKB. The PTKB is a set of natural language statements that describe the user's background, perspective, and context. The PTKB

is not the same as the MS MARCO leaderboard dataset.

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

Table 5: The Personal Text Knowledge Base (PTKB) for a user engaged in a conversational search about dieting.

Finding a diet

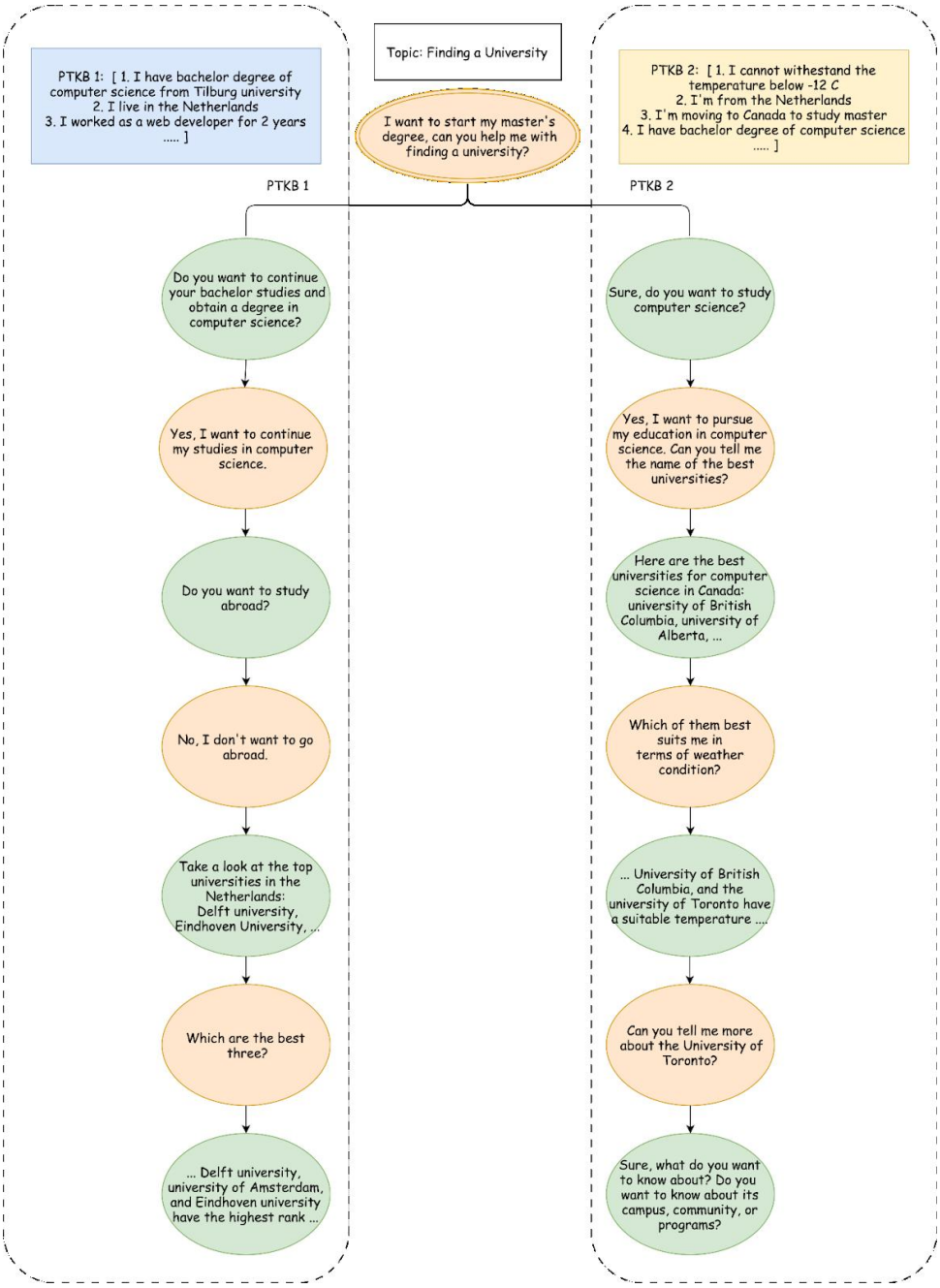

Figure 2: An example conversation "tree" from the Interactive Knowledge Acquisition Track.

is associated with a conversations, and statements in the PTKB may be relevant in a given turn of the conversation. This year, the PTKB is provided, but in future years the track hopes to make a task of generating the PTKB.

Systems treat the user utterances as queries and generate a response for each. That response can be a simple

Table 6: IKAT conversation 9-1, "Finding a diet". The Personal Text Knowledge Base (PTKB) is first and applies to all turns in the conversation.

Title: Finding a diet	
PTKB	
1	I don't like the new spin-off; because I keep comparing the two and it has lower quality.
2	Because of my kidney problem, I have to drink water frequently to stay hydrated.
3	I'm going to change my phone.
4	I can't exercise too much because of the heart problem that I have.
5	I'm vegetarian.
6	I'm lactose intolerant.
7	I'm allergic to soybeans.
8	I just finished watching the Game of Thrones.
9	I didn't like how the series ended, especially the war scenes.
10	I'm an Android user.
Turn	**Conversation Utterances**
1	Can you help me find a diet for myself?
2	Ok, good. Can you tell me what diet is the fastest way to lose some weight?
3	What about the DASH diet? I heard it is a healthy diet.
4	I prefer a natural diet, not a pill-based diet. Which of the aforementioned ones is natural?
5	Can you eat fish in any of them?
7	Thanks, but I also want the diet to be maintainable and not very hard to keep up.

document passage, or a summary (extracted or generated) from one or more passages. Systems also returned a "ptkb_provenance" with scores for PTKB entries, and a "passage_provenance" with a short passage ranking supporting the response.

The document set is a subset of 116,838,987 passages from the new ClueWeb22-B collection. Documents from ClueWeb22 were segmented automatically into passages of at most ten sentences. Systems could return up to 1000 passages in support of a response.

The NIST assessors made judgments for thirteen topics and 176 turns. Submissions were pooled to depth 25.

Each passage was judged on the 5-point scale used since CAsT 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.

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.

NIST was not able to assess response quality.

The track did not specify a primary metric, but used "standard" metrics like precision-at-cutoff, mean average precision, and NDCG. The track received 28 runs from 8 participating teams.

3.6 NeuCLIR

The NeuCLIR track is working on cross-language search. In this second year of the track, the coordinators introduced a new task of multilingual search, meaning that runs would return results in all target languages in a single ranking.

The document collection for this task is a combination of Russian, Chinese, and Farsi news text from Common-Crawl News. The relevance assessors are all bilingual in English and one of the target languages, and many of them are native speakers in the target language.

Last year's topics were developed by a single assessor and then translated to the other languages, and as a result there are a fair number of topics that have relevant documents concentrated in one of the three languages. To attempt to remedy this, we had the assessors work in pairs, each on a different target language, to develop the topics together in two languages simultaneously. NeuCLIR topics follow the traditional TREC topic format with a short title, a one-sentence description, and a paragraph-length narrative.

Documents are assessed for each topic in all three languages. This essentially means judging three times the number of topics, and so the assessment process takes quite a while. It also introduces a level of assessor disagreement noise into the relevance judgments, since those three assessors can't match the thinking of each other exactly.[3] One avenue to investigate is whether the judgments in the different languages rank systems differently, and why.

The scale for relevance was not relevant, topical (being in the ballpark), valuable and very valuable. The distinction between the top two levels was stated in terms of the user task: the user is writing a report on the topic, and very valuable items are the citations of the highest importance and usefulness. They would be cited early in the report, perhaps even on the first page. Documents that were merely "valuable" were valid citations, but they were useful more in a supporting role.

NeuCLIR also had a smaller-than-expected number of participants (six teams), and so the organizers ran a number of baselines in order to beef up the pools. All runs were pooled to depth 20, and a set of runs selected by the organizers were pooled to depth 50.

A third task in the track was cross-language search from English to Chinese over technical scientific documents. Five of the six teams also participated in this task, and their runs were pooled to rank 25. A manual run by the organizers was added to the pools.

220 runs were submitted: 48 Farsi, 48 Russian, and 49 Chinese monolingual runs; 24 multilingual runs; and 51 tech runs.

3.7 Product Search

Product Search is another new track for 2023, and TREC's first track in the e-commerce domain. The primary theme this year is customer-oriented product search. There are three tasks: a reranking task, an end-to-end ranking task, and a multi-modal task where the product information includes structured elements, clicks, and images. The documents and queries this year come from the ESCI Challenge for Improving Product Search [4], a KDD Cup challenge in 2022 administered by Amazon.[5]

All participant runs were pooled to depth 25, reflecting that we were not sure how long or complex the judging task would be. As it turned out, assessing was very quick and we were able to produce judgments for 186 different queries. NIST only selected English-language queries for assessment.

Relevance was judged on a four-point scale:

3 Perfectly relevant: the product exactly matches the query.

2 Highly relevant: the product isn't exactly what the query seemed to be looking for, but would be a reasonable substitute.

1 Related:] the product seems related, but it isn't what the user seemed to be looking for.

0 Irrelevant: the product is not at all relevant.

[3] I do not expect to find assessors who individually are fluent in all four languages.

[4] `https://amazonkddcup.github.io/`

[5] The data is Apache-2.0 licensed and so was available to reuse in TREC.

Some queries were extremely precise, for example, "Ernie Ball Custom Gauge 11 Nickel Guitar String 6 Pack" (200371), and perfectly relevant indicated one specific product, modulo duplicates; essentially a known-item search. Other queries, such as "short black cosplay wig for men" (200218) were less exact and for these there might be several perfectly relevant products.

Four groups participated in the Product Search track, with 62 submitted runs.

3.8 Tip-of-the-Tongue Search (ToT)

Tip-of-the-Tongue search, another new track for 2023, is inspired by sites such as `irememberthismovie.com` and the Reddit board `r/tipofmytongue`. On these sites, people ask about a movie or book or song or place that they don't remember the title of, and don't even completely remember anything about it. Rather, they have an incomplete memory that may even mix up two separate things. People on the site then respond with the book or movie that the poster was trying to remember.

This is a special type of known-item search where the information need is vague, incomplete, and possibly erroneous. Such searches happen in many domains (we've all struggled to find that email...) but as a search query type they are not well studied in the information retrieval community.

The track concept originated from a paper [2] and the accompanying MS-TOT dataset. The original plan was to use queries and ground truth from Reddit, but as the track was being planned, Reddit revised their terms of service to forbid redistribution of data. It is hoped that NIST will develop new topics for this track next year.

Since the MS-TOT dataset already has queries and relevance judgments, there was no pooling or assessment at NIST. Eleven groups submitted 33 runs.

4 Future

TREC will continue in 2024. The AToMiC, IKAT, Product Search, NeuCLIR, and ToT tracks will continue. There will be three new tracks: retrieval-augmented generation, biomedical generative retrieval, and lateral reading (a task similar to fact-checking).

Methods and systems that make use of LLMs to generate text are of great interest at the moment, and this is reflected in the tracks: two are explicitly about generative retrieval, NeuCLIR is planning to pilot a multilingual summarization subtask, and assessing for IKAT will focus more on the generative aspects of the task. Generative tasks pose scaling and reuse problems for dataset creation, and so in TREC 2024 we hope to incubate several ideas geared towards solving that evaluation problem.

In 2001, the Document Understanding Conference (DUC) split off from TREC, with tasks in summarization, question answering, and textual entailment; DUC was renamed TAC in 2008. A TREC 2001 and 2002 track on video retrieval was spun out into it's own venue, TRECVID, in 2003. At that time, IR, NLP, and multimedia research was done in sufficiently different communities that separate venues made sense and allowed more tasks than could fit in a single TREC track. However, in these days where LLMs act as universal vector-space projection, the distinctions don't seem as important as they used to be. Consequently, TAC and TRECVID are merging back into TREC in 2024, and bringing six tracks with them: entity, relation, and event extraction; plain language adaptation of biomedical abstracts; activities in extended video; adhoc video search; video-to-text; and medical video QA. There will also be a track in cooperation with DARPA's Computational Cultural Understanding (CCU) program.

Acknowledgments

TREC could not happen without a program committee, track coordinators, participants, and assessors, and I am grateful to them all for the contributions they make to the community. In particular our assessor team has pushed above and beyond to support this year's tracks in an uncertain calendar.

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] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. Tip of the tongue known-item retrieval: A case study in movie identification. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 5–14, 2021.

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

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

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

[6] 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.

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

[8] 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.

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

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

[11] 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.