The TREC Medical Records Track *

Ellen M. Voorhees
National Institute of Standards and Technology
ellen.voorhees@nist.gov

ABSTRACT

The Text REtrieval Conference (TREC) is a series of annual workshops designed to build the infrastructure for large-scale evaluation of search systems and thus improve the state-of-the-art. Each workshop is organized around a set of "tracks", challenge problems that focus effort in particular research areas. The most recent TRECs have contained a Medical Records track whose goal is to enable semantic access to the free-text fields of electronic health records. Such access will enhance clinical care and support the secondary use of health records.

The specific search task used in the track was a cohort-finding task. A search request described the criteria for inclusion in a (possible, but not actually planned) clinical study and the systems searched a set of de-identified clinical reports to identify candidates who matched the criteria. As anticipated, the search results demonstrate that language use within electronic health records is sufficiently different from general use to warrant domain-specific processing. Top-performing systems each used some sort of vocabulary normalization device specific to the medical domain to accommodate the array of abbreviations, acronyms, and other informal terminology used to designate medical procedures and findings in the records. The use of negative language is also much more prevalent in health records (e.g., *patient denies pain*, *no fever*) and thus requires appropriate handling for good search results.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—Performance Evaluation

General Terms

Measurement, Experimentation

Keywords

electronic health records, TREC

1. INTRODUCTION

Accessing a medical record based on its content is a fundamental usage requirement for electronic health record (EHR) management systems. Today's systems provide access largely based on structured fields—data elements in the record that have been coded to allow effective access. However, the majority of the content of a record is often in the care providers' notes and other free-text fields that are not so structured. Furthermore, standard text processing techniques do not work well for EHR free-text fields because the fields seldom contain well-formed, grammatical sentences; the vocabulary used within the record is highly specialized with many non-word terms (abbreviations, measurements, symbols, etc.); and the notes are often highly elliptical, implicitly referring to various other parts of the record.

Despite the difficulty of automated processing of free-text content, free-text fields within EHRs are nonetheless inevitable and even desirable. Language is the most convenient way for humans to communicate; free-text allows providers to express nuance and exceptional circumstances that are precluded—by definition—from being captured in coded fields. Using data capture methods that are natural and convenient greatly increases the quality of the data so captured. Thus EHR system ease-of-use and record quality concerns both argue for the continuing use of free-text fields within the EHR. What is needed are methods that can provide access to the semantic content of these free-text fields.

This paper describes the findings of the first two years of the Text REtrieval Conference (TREC) Medical Records Track that was established to focus the research community on the problem of providing content-based access to the free-text fields of EHRs. The lack of sharable test corpora has been cited as a major impediment to progress in applying natural language processing techniques to clinical text[4], and the track was established to help fill this void. The first section provides background information on TREC and the

*To adequately describe the findings of the TREC Medical Records track, certain commercial entities, equipment, or materials may be identified in this document. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

BCB '13, September 22 - 25, 2013, Washington, DC, USA

benefits of community evaluations to drive research. Section 3 describes the particular search task used in the track, while the following section summarizes the results and findings of the track. The final section concludes the paper by noting how just two years of the shared task has improved search effectiveness, though further progress is imperiled due to a continuing lack of sharable data.

2. THE TEXT RETRIEVAL CONFERENCE

Information Retrieval (IR) is the academic discipline concerned with providing automated access to content that is not specially structured for machines. Web search engines are probably the best-known examples of IR systems today, but content-based search is decades older and more pervasive than web search. Diverse applications such as intelligence gathering, ecommerce, legal discovery, and scientific research all require IR components.

IR systems have historically been developed through experimentation using *test collections* [5, 10]. A test collection consists of a document set, a set of information need statements (called *topics* in what follows), and a set of *relevance judgments* that define which documents should be retrieved for which topics. An IR system processes an information need statement to produce a list of documents ranked by decreasing likelihood that the document answers the need. The set of lists returned for each topic in a test collection is called a *run*. The quality of an IR system's response for a single topic can be measured as a function of how closely its retrieved set matches the correct response as recorded in the relevance judgments, with the overall score for a run usually computed as the average of the individual topics' scores. There are many possible measures of effectiveness, but most measures in common use are a combination of *precision*, the fraction of retrieved documents that are relevant, and *recall*, the fraction of relevant documents that are retrieved [1].

While test collections can be powerful research tools, a given collection's utility depends on how representative it is of the actual search task being modeled. One of the main considerations is the size of the collection: the document set must be large enough to exhibit the variety of content that would be encountered in the real-world task and the topic set large enough to be representative of the types of questions encountered. Unfortunately, relevance judgments must be created by humans, so the expense and difficulty of creating quality test collections increases with size. The U.S. National Institute of Standards and Technology (NIST) founded the TREC workshop series[1] in 1992 with the goal of building a realistically-large test collection to support IR research. The goal has broadened since then to standardizing and validating IR evaluation methodology in addition to building appropriate test collections for a variety of IR tasks.

Participants in TREC are retrieval research groups drawn from the academic, commercial, and government sectors. TREC organizers release a document and topic set, and participants use their systems to produce runs that they submit to NIST. Since TREC document sets are much too large for each document to be judged for relevance for each topic (the typical TREC collection contains several hundred thousand documents and 50 topics), a subset of documents is created for each topic by sampling from the union of the submitted runs. The documents in the sample are viewed by a human assessor who rates the relevance of a document to the topic. Once all the relevance judgments for all of the topics in the test set are complete, NIST scores the submitted runs on the basis of the relevance judgments and returns the evaluation results to the participants. A TREC cycle ends with the workshop that is a forum for participants to share their experiences. A test collection built during a given TREC cycle is subsequently made available to non-participants (subject to any encumbrances on the document set) to benefit the wider research community.

Each TREC contains a set of tasks, called tracks, that focus on separate subproblems of IR. For example, past tracks have included cross-language retrieval (retrieving documents written in languages that differ from the query language), video retrieval (providing content-based access to digital video), and question answering (retrieving answers themselves rather than documents containing answers). TRECs 2003–2007 contained a Genomics track whose goal was to develop technology that assists biology researchers (i.e., users familiar with the biomedical domain) with keeping current with the biomedical literature [6]. The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often uncovers 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.

When TREC began there was real doubt as to whether the text collection paradigm had out-lived its usefulness for IR research [9], but the 20-plus year history of TREC has demonstrated that it had not. TREC has now built dozens of test collections for a variety of tasks, and in each case the test collections have been integral to progress on the task. For example, retrieval effectiveness doubled in the first six years of TREC for the basic "ad hoc" search task [12]. By defining a common set of tasks, TREC focuses retrieval research on problems that have a significant impact throughout the community. The workshop provides a forum in which researchers can efficiently learn from one another and thus facilitates technology transfer. TREC also provides a forum in which methodological issues can be raised and discussed, resulting in improved retrieval research. The motivation for the TREC Medical Records track is to bring these benefits to the problem of providing content-based access to the free-text fields of electronic health records.

3. MEDICAL RECORDS TRACK TASK

This section describes the set-up of the task in the TREC Medical Records track. Operationalizing an evaluation task requires balancing the competing constraints imposed by the real-world task of interest, data availability, and capabilities of the current technology (i.e., making a task neither too hard nor too easy). Due to the sensitive nature of medical records, data constraints are the overarching constraints for the Medical Records track.

The document set used in the track is a set of de-identified clinical reports made available to TREC participants through the University of Pittsburgh NLP Reposi-

[1] `http://trec.nist.gov`

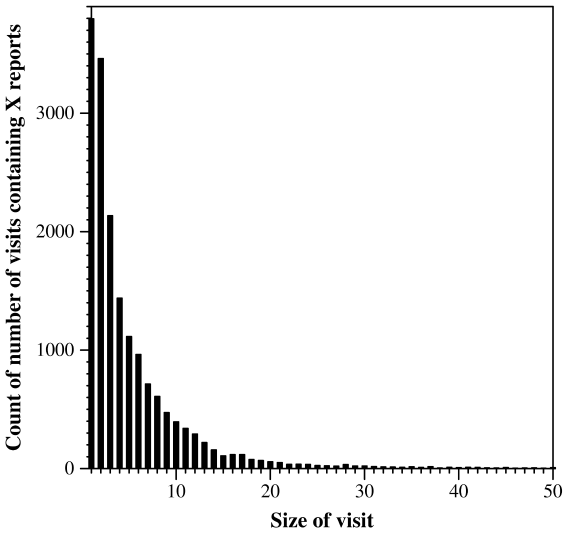

Figure 2: Distribution of visit sizes, truncated at visit size 50. The size of a visit is the number of reports associated with the visit.

tory (called the Pitt record set below)[2] [2]. The Pitt record set contains one month of reports from multiple hospitals, and includes nine types of reports: Radiology Reports, History and Physicals, Consultation Reports, Emergency Department Reports, Progress Notes, Discharge Summaries, Operative Reports, Surgical Pathology Reports, and Cardiology Reports. A report is linked to a "visit" (an individual patient's single stay at a hospital), and contains both the ICD discharge diagnosis codes (primary and secondary) for its visit as well as the free-text "chief complaints" field as captured in the medical record's Discharge Abstract for that visit. Links between the same person's different visits to a hospital are (intentionally) broken as part of the de-identification process, so it is not possible to track a single person through multiple episodes. Nonetheless, a single visit can represent a lengthy hospital stay, and thus a visit may encompass many different reports.

Figure 1 shows the structure of the data set. The many-to-one mapping between reports and visits is codified through a mapping table that gives the corresponding visit-id for each report-id. The report id is an identifier for a file that contains the content of that report. The data set contains 93,551 reports mapped into 17,264 visits. The distribution of visit size—as measured by number of reports—is highly skewed, with a minimum of 1, a maximum of 415, and a median of 3. Figure 2 shows the distribution truncated at visit size 50. The unit of retrieval (a "document") in the track is the visit. That is, for the purposes of the track the content of a document is the union of the content of all the reports associated with a given visit.

The kinds of reports and the number of reports of each type included in the Pitt record set were determined by the corpus builders well before it was known that the record set would be used as the document set in a TREC track. The track organizers therefore needed to choose a retrieval task that was a good fit for the record set in addition to being representative of an interesting real-world problem. The task selected was an ad hoc retrieval task as might be used to identify cohorts for comparative effectiveness research or other types of clinical research. When designing a clinical study, a researcher will usually develop inclusion criteria that describe the kind of patients required for the study. These criteria include attributes such as disease(s) present, treatment(s), age group, gender, and ethnicity. The track's topic statements were modeled after inclusion criteria statements, and systems returned a list of visits ranked by the likelihood that the visit's patient satisfied the inclusion criteria. Several example topics are shown in Figure 3.

In both years of the track, topics were created by physicians who were also students in the Oregon Health & Science University (OHSU) Biomedical Informatics Graduate Program. For 2011, topic developers used a list of research areas the U.S. Institute of Medicine (IOM) has deemed priorities for clinical comparative effectiveness research[3] as a starting point for topic development. Given a topic from the IOM list, the developer searched the Pitt record set using a Boolean retrieval system to develop an estimate of the number of relevant visits in the document set. Topic development for 2012 proceeded similarly, though additional sources for topic ideas were used since the IOM list was exhausted before a sufficient number of new topics was created. These sources included the clinical quality measures for eligible hospitals under the meaningful use incentive program for electronic health record adoption in the US and the OHSUMED medical literature retrieval test collection[4]. At the 2011 TREC meeting, some track participants expressed the opinion that the 2011 topics had been too easy; while selecting more difficult topics was not an explicit design critereon for the 2012 topic creation process—and in any case accurately predicting how difficult a topic will be is itself quite difficult [11]—this may have caused some bias against obviously easy topics in the final selection of 2012 topics. The final test sets contained 35 topics for 2011 and 50 (different) topics for 2012.

The relevance assessing for the track was performed by physicians who were also either students in graduate programs in biomedical informatics (at OHSU and elsewhere) or researchers from the U.S. National Library of Medicine. Assessors judged between 1–9 topics depending on their time availability. A given topic was judged by one primary assessor whose judgments were used in scoring runs. (Some topics were judged by more than one assessor for an eventual study on judgment agreement rates, but only the judgments of the primary assessor were used in scoring.)

Assessors were instructed to rate each visit in the judgment set to determine whether such a patient would be a candidate for a clinical study on the topic. Assessors used an interface that allowed them to expand and contract the individual reports associated with a visit, and judged each visit in the set as either not relevant, partially relevant, or relevant. A definitely relevant judgment meant that the patient was unequivocally a candidate for the study. A possibly relevant judgment meant that the patient might be a candi-

[2]Because of the private nature of medical records—even when de-identified—the University of Pittsburgh distributed the records only to track participants. The records are not available at this time.

[3]http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx

[4]http://ir.ohsu.edu/ohsumed/ohsumed.html

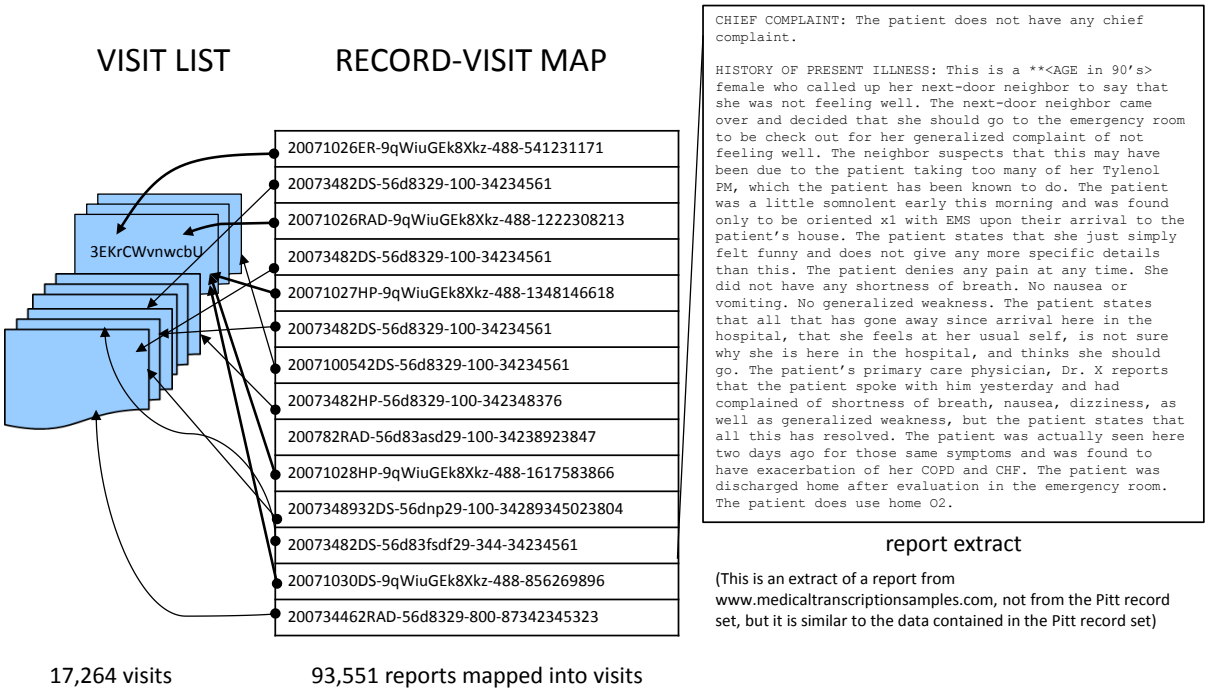

Figure 1: Structure of the Pitt record set. The mapping table gives the visit id associated with each report id. The report id identifies the file containing the content of the report. For the purposes of the track, a document is the visit, the union of all the reports associated with the visit.

136: Children with dental caries

137: Patients with inflammatory disorders receiving TNF-inhibitor treatment

152: Patients with Diabetes exhibiting good Hemoglobin A1c Control (<8.0%)

160: Adults under age 60 undergoing alcohol withdrawal

167: Patients with AIDS who develop pancytopenia

169: Elderly patients with subdural hematoma

179: Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

Figure 3: Example topics from the TREC 2012 Medical Records Track test set.

date for the study but insufficient information was available for a definitive decision. A not relevant judgment meant that the patient was not a candidate for the clinical study.

Despite attempts to make sure that topics had sufficient numbers of relevant visits during topic development, several topics in the test set were omitted from the evaluation because they had fewer than five definitely or possibly relevant visits identified after the judgment process was complete. Evaluating retrieval effectiveness of topics with very few relevant visits is inherently unstable in the sense that small changes in the retrieval results can cause very large changes in effectiveness scores. Relevant visits can be "lost" for a variety of reasons: judging mistakes made in either the topic development or assessing phase; a single assessor changing his or her mind regarding relevance between the two phases; using different assessors whose opinions regarding relevance differ for the two phases; a relevant visit seen during topic development not making it into the judgment set (see below). The evaluation set for TREC 2011 contains 34 topics and the TREC 2012 set contains 47 topics.

The particulars regarding how the judgment sets were constructed differed in the two years, with the end result being that different primary evaluation measures were used in different years. The construction strategy used for TREC 2011 proved to induce unexpectedly noisy evaluation scores, and since results from the first year of any track are often dominated by start-up issues such as lack of training data, the remainder of this report concentrates on the TREC 2012 edition of the track. For 2012, all submitted runs contributed to the judgment sets, which were constructed to be compatible with using extended, inferred Normalized Discounted Cumulative Gain (infNDCG) as the primary evaluation metric [13]. Precision after the first 10 documents retrieved (Prec(10)) was also reported and could be computed exactly since all submitted runs had the top 10 documents retrieved for a topic added to the judgment set for that topic. The union of the judgment sets across the 50 topics in the test set included 25,596 visits to be judged. The average size of a judgment set was 512 visits, with a minimum size of 206 and a maximum size of 919.

Table 1 shows the distribution of the visit sizes in the judged sets (restricted to the 47 topics in the final evaluation set), as well as in the entire set of visits for comparison. The third column gives the absolute number and percentage of the total number of judged-not-relevant visits that contained the given number of records. The fourth column gives the corresponding figures for relevant documents, using both partially relevant and definitely relevant judgments as relevant visits. There was a total of 20,089 visits judged not relevant and 4130 visits judged relevant across the 47 topics in the final evaluation set. The distributions of visit sizes of judged not relevant and relevant visits are equivalent, suggesting that visit size is not a determining factor for relevance. The distribution of visit sizes of the retrieved set (as reflected by the judgment set) does differ from the overall distribution in that it contains many fewer single-record visits. Some single-record visits (visits containing only a single radiology report, for example) contain very little text, making them both less likely to match retrieval systems' queries and less likely to be candidates for the types of studies being modeled in the topics.

Inferred measures are used as a means of getting more accurate estimates of a run's quality than is likely possible with traditionally-defined versions of the measures when judging a comparatively small number of documents. Normalized discounted cumulative gain is an evaluation measure that accommodates different relevance levels by rewarding systems for retrieving documents with a higher relevance rating before documents with a lower relevance rating [7]. The reward (called a gain value) received for retrieving a relevant document with a particular rating is reduced (discounted) the further down in the list the document is retrieved, under the assumption that that document is worth less to a user than a document retrieved higher in the list. All gain values up to a pre-determined rank are summed, and then that value is normalized by the score that a best possible ranking would receive. (An ideal ranking is a ranking that lists all documents from a relevance category with a higher gain value before any document from a relevance category with a lower gain value. The score for an ideal ranking is topic-dependent because it depends on the number of relevant documents in each relevance category.) Normalized discounted cumulative gain has a minimum score of 0 (assuming non-negative gain values) and a maximum score of 1, with higher scores being better. For the TREC 2012 Medical Records track, the gain value for partially relevant was 1 and the gain value for fully relevant documents was 2; the rank cut-off used was 100 (i.e., only the first 100 documents retrieved for a topic are factored into the score).

Table 2: Submissions to the Medical Record Track.

	Number Participants	Total Runs	# Runs by Type	
			automatic	manual
2011	29	127	109	18
2012	24	88	82	6

4. RETRIEVAL RESULTS

Table 2 gives details about the runs received by the track in each of the two years. The table gives the number of participants in the track per year, the total number of runs received, and the split of those runs between automatic and manual runs (discussed below). In TREC 2011, each participant was permitted to submit up to eight runs; for TREC 2012 that limit was reduced to four runs per participant. Some groups participated in both years of the track, though there was a sizeable number of first-time track participants in 2012.

TREC has historically distinguished between runs that are produced from the topic statement with no human intervention of any sort (*automatic* runs) and all other runs (*manual* runs). The definition of manual runs 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. Manual runs allow researchers to gauge how well their systems support users that actively engage in the search, while automatic runs demonstrate what a system can do without any additional direction. Manual results tend to be more variable than automatic results even when putative level of effort is controlled for since some humans are inherently better searchers than others.

Table 1: Distribution of visit sizes for visits, for judged Not Relevant visits and for combined Partially Relevant/Relevant visits. Judged visit counts are computed over 47 topics in the final evaluation set.

Number of Reports	Total Visits		Judged Not Relevant		Relevant	
	Number	%	Number	%	Number	%
1	3846	22	1582	8	235	6
2–5	8315	48	6268	31	1300	31
6–15	4164	24	8159	41	1893	46
16–30	692	4	2382	12	461	11
31–100	226	1	1368	7	208	5
>100	21	0	330	2	33	1

Table 3: Evaluation results for the best runs for the top eight participants in the TREC 2012 track ordered by infNDCG. Run tags that are starred are manual runs.

Run	infNDCG	P(10)
NLMManual*	0.680	0.749
udelSUM	0.578	0.592
sennamed2	0.547	0.557
ohsuManBool*	0.526	0.611
atigeo1	0.524	0.519
UDinfoMed123	0.517	0.528
uogTrMConQRd	0.509	0.553
NICTAUBC4	0.487	0.517

Details regarding the different approaches used by individual participants can be found in the participant reports included in the TREC proceedings (see `http://trec.nist.gov/proceedings/proceedings.html`). In the remainder of this section we summarize the evaluation results and highlight the major themes from across participants, once again focusing the discussion on the TREC 2012 track.

4.1 Retrieval Scores

Table 3 gives the evaluation scores for the best run for the top eight participants in the TREC 2012 track as measured by infNDCG. The table gives the infNDCG and Prec(10) scores averaged over the 47 topics in the final evaluation set. Starred run tags in the table denote manual runs.

The most effective run, `NLMManual`, was a manual run in which physicians modified automatically-generated queries. As measured by Prec(10), this run retrieved about 1.5 more relevant visits in the top 10 visits retrieved on average than the automatic run with the best Prec(10) score, `udelMRF` (7.49 for `NLMManual` vs. 6.04 for `udelMRF`). This difference is highly statistically significant according to a paired t-test ($p<0.0005$), supporting the intuition that actively engaging the human user in the loop is an effective search strategy. The extent to which such a difference would be noticeable in practice—and whether these levels of effectiveness are useful in practice—depends on the application.

As is typical for retrieval performance, individual topic scores varied widely both within and across runs. The run obtaining the best score for a given topic across the 88 submitted runs was a manual run for slightly less than half the topics. The plot in Figure 4 shows results for individual topics using infNDCG as the measure. The line graphs show the median (solid line) and best (dotted line) scores obtained for the given topic as computed over all submissions[5]. The x-axis gives the topic number, with topics sorted by decreasing median infNDCG score. The gray bar chart imposed on the graph shows the number of known relevant (definitely relevant plus partially relevant) visits per topic. The left y-axis plots the infNDCG score value and the right y-axis plots the number of relevant visits.

Also typical for retrieval performance is that the difficulty of a topic, as measured by the evaluation scores obtained for it, is independent of the number of relevant visits it has. The two hardest topics (topics with the worst median and worst best infNDCG scores), topic 167 *Patients with AIDS who develop pancytopenia* and topic 137 *Patients with inflammatory disorders receiving TNF-inhibitor treatments*, each have only six known relevant visits. But there are other topics with similarly small relevant sets that are relatively easy. For example, topic 150 *Patients who have cerebral palsy and depression* has nine relevant visits and is the sixth easiest topic as measured by median infNDCG scores. The topic with the best median infNDCG score is topic 178 *Patients with metastatic breast cancer* with 34 known relevant visits, while the topic with the best best infNDCG score is topic 182 *Patients with Ischemic vascular disease* with 468.

Several topics have large differences between the best and the median scores. Topic 167 mentioned earlier is one example, as is topic 179 *Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression.* One cause of the disparity in this latter case is the failure of the majority of systems to match the generic term 'atypical antipsychotics' to the particular instance of such a drug (e.g., clozapine or risperidone) mentioned in the record.

4.2 Retrieval Approaches

As illustrated by the last example, the use of some sort of vocabulary normalization specific to the medical domain and/or the use of term expansion is necessary for good search performance. The language use in health records is generally informal and a given medical entity (condition, treatment, diagnostic procedure, etc.) is referred to by a wide vari-

[5]An observant reader will notice that the best infNDCG value for topic 182 is slightly greater than 1.0, the theoretical maximum value NDCG. The estimate of 1.012 is caused by sampling errors in the computation of infNDCG. This level of "impossibility" has been observed in the past when the mean estimated values produced by the inferred measures were quite accurate. Hence, we believe the evaluation results reported for TREC 2012 are sound.

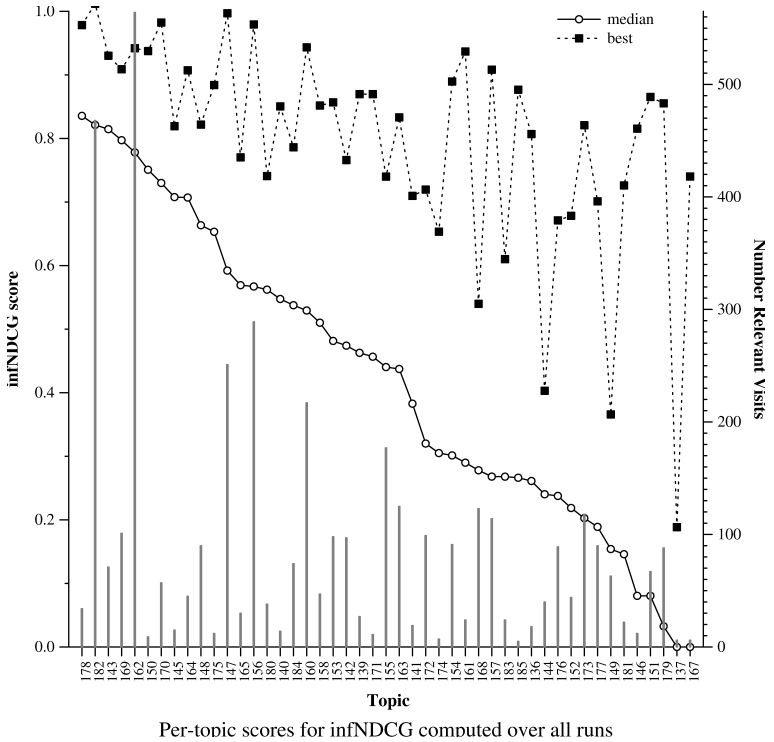

Figure 4: Median and best per-topic scores as measured be inferred NDCG.

ety of acronyms, abbreviations, and informal designations. Many track participants used MetaMap[6] to locate medical terms in text and map those terms to concepts in the UMLS metathesaurus. Once concepts are mapped from UMLS concepts to entries in some medical controlled vocabulary (such as SNOMED-CT or MeSH), terms related to the concept can be added to the query.

Another source of terms for query expansion was ICD-9 codes assigned to the record. The International Classification of Diseases (ICD) codes are designations from a hierarchical classification of human diseases and symptoms maintained by the World Health Organization[7] that are included as part of the structured content of most records, as ICD-9 is required for health care providers to obtain reimbursement from insurance companies for services provided. Most of the participants that used the codes used words from the textual descriptions of codes related to query terms rather than match on the codes themselves. Regardless of the source of query expansion terms, the expansion must be done with care, as some participants reported significant degradation from query expansion for some query types due to query drift.

Health record text is full of negated language constructs documenting the absence of symptoms (*no chest pain or palpitations*), behaviors (*denies use of alcohol*), and abnormal diagnostic results (*temperature not elevated*). Given the prevalence of its use, and the fact that a match with the search criteria often depends on the polarity of an indicator, specific processing for negated language appears necessary for effective retrieval for the cohort-finding task. This is in contrast to search in many domains where such processing generally has little effect. Negated expressions can be found using tools such as NegEx [3], and the object of the negation was frequently just omitted as an index term for the report.

[6] `http://metamap.nlm.nih.gov`

[7] `http://www.who.int/classifications/icd/en/`

5. CONCLUSION

The goal of the TREC Medical Records Track is to bring the benefits of community evaluations—test collections, a focused research community, established research methodology—to the problem of enabling semantic access to the free-text fields of electronic health records. And there is evidence that these benefits have accrued in just the first two years of the track. During the course of its participation in the track, a team from Dublin City University ran a basic BM25 search (a good quality, domain-independent baseline system) first using the TREC 2011 topics and then using the TREC 2012 topics [8]. They found that the relative effectiveness of this baseline run compared to other participants' runs to be greater in 2011 than in 2012. In particular, the BM25 run would have scored among the top five runs in the TREC 2011 track, but was only slightly above the median in the TREC 2012 track. This suggests that, on the whole, the effectiveness of search systems for the cohort finding task improved in the second year of the task despite the absolute value of the effectiveness scores being lower in 2012 (i.e., the 2012 task was inherently harder). While improvement in the second year of a task is to be expected, that is precisely because of the benefits of the paradigm: researchers have more experience with the task and an existing test collection on which to train their systems.

The hypothesis going into the track was that language use within health records is different enough from "nor-

mal" English that existing search systems would have trouble with it. That hypothesis has been largely confirmed. Two main differences are the extensive use of acronyms and abbreviations that need to be resolved to the correct medical concept, and the pervasive use of negative language. Top-performing groups each used some sort of vocabulary normalization device specific to the medical domain. Such devices must be used carefully, however, as multiple groups also demonstrated that aggressive use harms baseline performance. Exploiting human expertise through manual query construction proved most effective.

The future of the track is uncertain since we currently lack a suitable collection of health records to serve as the basis of a test collection. After two years of use, there are few viable topics remaining in the comparatively small (for search test collections) Pitt record set. As noted elsewhere [4], there is a dearth of shared record sets, and we have been unable to find a large record set that is available for the wide-scale use implied by being the basis of a TREC test collection. The track will be on hiatus while we explore creative solutions to resolving the tension between protecting patients' (and record owners') legitimate privacy concerns and the need for realistic data in support of research.

6. ACKNOWLEDGEMENTS

Many thanks to Bill Hersh for his advice and assistance in shaping the track, and to Bill and his colleagues at Oregon Health and Science University for managing the topic development and relevance judgment process for the track. Paul Over at NIST provided significant assistance in running the track. The track's steering committee—consisting of Wendy Chapman, Aaron Cohen, Kevin Cohen, Milton Corn, Bill Hersh, Paul Over, Mark Sanderson, Guergana Savova, Richard Tong, Ozlem Uzuner, and Ellen Voorhees—guided the definition of the track in many ways.

7. REFERENCES

[1] Chris Buckley and Ellen M. Voorhees. Retrieval system evaluation. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–75. MIT Press, 2005.

[2] Wendy Chapman, Melissa Saul, John Houston, Jeannie Irwin, Dannielle Mowery, Henk Harkema, and Michael J. Becich. Creation of a repository of automatically de-identified clinical reports: Processes, people, and permission. In *Proceedings of the American Medical Informatics Association Clinical Reserach Informatics Summit*, March 2011.

[3] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310, 2001.

[4] Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, Leonard W. D'Avolio, Guergana K. Savova, and Ozlem Uzuner. Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Information Association*, 18(5), 2011.

[5] C. W. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 173–192, 1967. (Reprinted in *Readings in Information Retrieval*, K. Spärck-Jones and P. Willett, editors, Morgan Kaufmann, 1997).

[6] William Hersh and Ellen Voorhees. TREC Genomics Track special issue overview. *Information Retrieval*, 12:1–15, 2009.

[7] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[8] Johannes Leveling, Lorraine Goeuriot, Liadh Kelly, and Gareth J. F. Jones. DCU@TRECMed 2012. In *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*, `http://trec.nist.gov/pubs/trec21/papers/DCU.medical.final.pdf`, 2013.

[9] S.E. Robertson and M.M. Hancock-Beaulieu. On the evaluation of IR systems. *Information Processing and Management*, 28(4):457–466, 1992.

[10] Karen Spärck Jones and Peter Willett. Evaluation. In Karen Spärck Jones and Peter Willett, editors, *Readings in Information Retrieval*, chapter 4, pages 167–174. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1997.

[11] Ellen M. Voorhees and Donna Harman. Overview of the sixth Text REtrieval Conference (TREC-6). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 1–24, August 1998. NIST Special Publication 500-240. Electronic version available at http://trec.nist.gov/pubs.html.

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

[13] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the Thirty-First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 603–610, 2008.