Overview of the TREC 2012 Medical Records Track

Ellen M. Voorhees
National Institute of Standards and Technology

William Hersh
Oregon Health & Science University

Abstract

The TREC Medical Records track fosters research that allows electronic health records to be retrieved based on the semantic content of free-text fields. The ability to find records by matching semantic content will enhance clinical care and support the secondary use of medical records in clinical trials and epidemiological studies. TREC 2012 is the sophomore year of the track, which attracted 24 participating research groups.

The track repeated the cohort-finding task from its initial year. This task is an ad hoc search task in which systems search a set of de-identified clinical reports to identify cohorts for (possible) clinical studies. A topic statement for the task describes the criteria for inclusion in a study, and a system returns a list of "visits" ordered by the likelihood that the inclusion criteria are satisfied. Physicians created fifty topics and performed relevance judgments for the track.

Top-performing groups each used some sort of vocabulary normalization device specific to the medical domain, supporting the hypothesis that language use within electronic health records is sufficiently different from general use to warrant domain-specific processing. 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.

Today's electronic health record (EHR) systems generally provide access to records based on structured fields, data elements in the record that have been coded to allow effective access. Yet the majority of the content of a record is often in the provider's notes and other free-text fields that are not so structured. Free-text allows providers to express nuance and exceptional circumstances that are precluded—by definition—from being captured in coded fields. Thus EHR system ease-of-use and record quality concerns argue for the continuing use of free-text, provided that that content can be effectively searched. The TREC Medical Records track was established to focus a research community on the problem of enabling content-based access to the free-text fields of EHRs and to build the infrastructure necessary for such research.

1 The Medical Records Track Task

The lack of sharable test corpora has been cited as a major impediment to progress in applying natural language processing techniques to clinical text[1]. The TREC Medical Records track looks to help fill this void in the face of pragmatic concerns that constrain what can be done. Due to the sensitive nature of medical records, data constraints are the overarching factor for the Medical Records track. This section first describes the data set used in the track and then motivates the retrieval task.

1.1 Documents

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 Repository (called the Pitt record set below). This is the same document set that was used in the TREC 2011 track. Because of the private nature of medical records (even when de-identified), the University of Pittsburgh distributes the records only to track participants.

The repository 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 International Classification of Diseases (ICD)

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

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.

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 data set used in TREC 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 (see Table 1 for more details). 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 the given visit.

1.2 Retrieval task

The retrieval task used in the track is 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 1.

Topics were created by physicians who were also students in the Oregon Health & Science University (OHSU) Biomedical Informatics Graduate Program. Their goal was to develop fifty topics that each matched a reasonable number of visits (more than a few but less than several hundred) in the Pitt record set.

OHSU physician-students had also developed the topics for the TREC 2011 track. For 2011, they used a list of research areas the U.S. Institute of Medicine (IOM) has deemed priorities for clinical comparative effectiveness research (`http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx`) 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. The final test set of 35 topics for 2011 was drawn from exploring 54 candidates from the IOM list. The main reasons for excluding a candidate was either that the topic was not a good fit for a collection of hospital-based medical records (e.g., *preventing dental caries in children* or *testing new biomarkers for cancer diagnosis and treatment*) or that there were too few or too many relevant visits.

Topic development for 2012 proceeded similarly using three physicians from the OHSU Biomedical Informatics program as developers. They started with the remaining 46 topics in the IOM collection as topic candidates. Sixteen of those 46 IOM topics were deemed to be usable for the TREC track, necessitating additional sources of topic ideas. The first additional source used was the clinical quality measures for eligible hospitals under the "meaningful use"

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	Total Visits		Judged Not Relevant		Relevant	
Reports	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

incentive program for electronic health record adoption in the US. Some of these quality measures are very close conceptually (e.g., measurements of test ordering or patient outcome for the same disease), so these criteria yielded 12 additional topics. The second additional source of topics was the OHSUMED literature retrieval test collection, with the topic statements modified if necessary to reflect querying of a medical records system and evaluated for appropriate numbers of relevant visits. This source provided the remaining 22 topics to round out the 2012 test set of 50 topics.

1.3 Relevance judgments

The relevance assessing for the track was performed over judgment sets of retrieved visits constructed as described below. OHSU initially recruited judges who were physicians and students in the OHSU Biomedical Informatics Graduate Program. This provided an insufficient number of judges, however, so physician researchers from the US National Library of Medicine (NLM) as well as physicians who are students in graduate programs in biomedical informatics funded by training grants from the NLM at other universities were also recruited. All told, 25 physicians judged between 1–9 topics depending on their time availability.

Judges were instructed to rate each visit to determine whether such a patient would be a candidate for a clinical study on the topic. 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 candidate 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.

Each topic was completely judged (i.e., had all the visits in its judgment set judged) by at least one single judge. In addition, six topics were partially or fully judged by a second judge. Three topics had fewer than five definitely or possibly relevant visits. These topics—138, 159, and 166—were omitted from the evaluation, leaving 47 topics contained in the evaluation set.

All submitted runs contributed to the judgment sets, which were constructed to be compatible with computing extended inferred measures (see below). In particular, the judgment sets were created using two strata: all visits retrieved in ranks 1-15 by any run in union with a 25% sample of visits not retrieved in the first set that were retrieved in ranks 16-100 by some run. 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 (topic 169) and a maximum size of 919 (topic 137).

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 ‘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.

Table 2: Groups participating in the Medical Records track.

Atigeo LLC	Australian e-Health Research Center
Dublin City University	Institute of Medical Informatics, NCKU
LSIS - Aix-Marseille University	Mayo Clinic
NEC Laboratories America	NICTA - National ICT Australia
Oregon Health & Science University	Pattern Recognition and Intelligence System Lab
Queensland University of Technology	RMIT University
The Siena College Institute for Artificial Intelligence	Seoul National University
University College Dublin	Universidad Complutense de Madrid
University of Delaware (Carterette)	University of Delaware (Fang)
University of Glasgow (Terrier)	University of South Florida
University of Texas at Dallas	University of Utah
US National Library of Medicine	York University

Table 3: Evaluation results for the best runs for the top eight groups ordered by infNDCG. Run tags that are starred are manual runs.

Run	infNDCG	infAP	P(10)
NLMManual*	0.680	0.366	0.749
udelSUM	0.578	0.286	0.592
sennamed2	0.547	0.275	0.557
ohsuManBool*	0.526	0.250	0.611
atigeo1	0.524	0.224	0.519
UDinfoMed123	0.517	0.236	0.528
uogTrMConQRd	0.509	0.231	0.553
NICTAUBC4	0.487	0.216	0.517

2 Retrieval Results

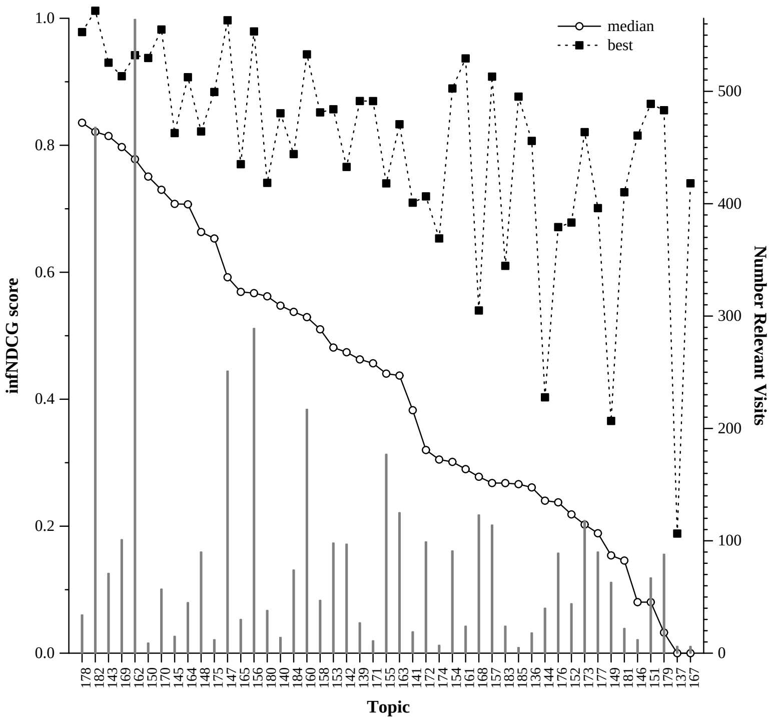
The Medical Records track received a total of 88 runs from the 24 groups listed in Table 2. Six of the runs were manual runs and the remainder were automatic runs, meaning there was no human intervention of any sort in producing the ranked list of visits given the topic statement.

As noted above, the judgment sets were created to support the computation of extended inferred evaluation measures [3]. Inferred measures are used as a means of getting more accurate estimates of a run's quality than is likely possible with traditional measures when judging a relatively small number of documents. Since a run had all of its top 10 documents judged, Precision(10) could be computed exactly. For measures other than inferred NDCG (infNDCG), the partially relevant and fully relevant sets were conflated into a single relevant set. For infNDCG, the gain value for partially relevant was 1 and the gain value for fully relevant documents was 2. InfNDCG was computed using a cut-off of 100.

Table 3 gives the evaluation scores for the best run for the top eight groups as measured by infNDCG. The table gives the infNDCG, inferred average precision (infAP), and precision at rank 10 (P(10)) scores averaged over the 47 topics in the final evaluation set. Starred run tags in the table denote manual runs.

The plot in Figure 2 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 the entire set of 88 runs[1]. The x-axis gives the topic number, with topics sorted by decreasing median infNDCG score. The gray bar chart

[1] 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. While values much greater than 1.0 were the tip-off to excessive instability in the estimates for inferred measures in the TREC 2011 track, this current 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 results being reported here are sound.



Per-topic scores for infNDCG computed over all runs

Figure 2: Median and best per-topic scores as measured be inferred NDCG.

imposed on the graph shows the number of relevant (both highly relevant and 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.

The most effective run, `NLMManual`, was a manual run in which physicians modified automatically-generated queries. As measured by P(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 P(10) score, `udelMRF` (7.49 for `NLMManual` vs. 6.04 for `udelMRF`). 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.

During the course of its participation in the track, the Dublin City University team ran a basic BM25 search first using the TREC 2011 topics and then using the TREC 2012 topics [2]. They found that the relative effectiveness of this baseline run compared to other participants' runs to be greater in 2011 than in 2012. This suggests that, on the whole, the effectiveness of search systems for the cohort finding task improved in this 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). Improvement in the second year of a task is to be expected: researchers have more experience with the task and an existing test collection on which to train their systems. And, indeed, TREC 2012 Medical Records track participants did use the TREC 2011 collection as training material. Nonetheless, the improvement is a reminder of the power of the test collection paradigm in advancing the state-of-the-art.

2.1 Participant results

Details regarding the different approaches used by individual participants can be found in the participant reports included elsewhere in the proceedings. Here, we highlight the major themes from across participants.

A large majority of participants, including top-scoring participants, used some sort of vocabulary normalization specific to the medical domain and/or term expansion. 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 variety of acronyms, abbreviations, and informal designations. Frequently, term normalization was done using MetaMap[2] to locate medical terms in text and map those terms to concepts in the UMLS metathesaurus. There are also terminology granularity considerations when matching queries and records. For example, topic 179 asks for patients taking "atypical antipsychotics"; relevant records indicate the use of a particular instance of such a drug (e.g., clozapine or risperidone) without mentioning the category of drug at all. 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[3] 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 many other ad hoc search tasks where such processing generally has little effect.

3 Conclusion

The TREC Medical Records track has concluded its second year of examining the problem of providing content-based access to free-text fields within health records. Top-performing groups each used some sort of vocabulary normalization device specific to the medical domain, supporting the hypothesis that language use within electronic health records is sufficiently different from general use to warrant domain-specific processing. 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. The track will be on hiatus in TREC 2013 as we try to resolve the data issues.

Acknowledgments

Many thanks to the Medical Records Track Steering Committee for their advice and assistance in shaping the track. Members of the committee include Wendy Chapman, Aaron Cohen, Kevin Cohen, Milton Corn, Paul Over, Mark Sanderson, Guergana Savova and Ozlem Uzuner, in addition to the authors.

References

[1] 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.

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

[3] 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.

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

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