The TREC Bio/Medical Tracks

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

References


The TREC Bio/Medical Tracks

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

Why is IR important in health and biomedical domain?

- Clinicians cannot keep up – average of 75 clinical trials and 11 systematic reviews published each day (Bastian, 2010)
- Data points per clinical decision increasing (Stead, 2011)
- Search for health information by clinicians, researchers, and patients/consumers is ubiquitous (Fox, 2011; Google/Manhattan Research, 2012)
- Concerns about reproducibility of science (Baker, 2016)
- “Precision medicine” requires retrieval not only of documents but also data (Tenenbaum, 2016)
The TREC Bio/Medical Tracks

- TREC Genomics Track
- ImageCLEFmed
- TREC Medical Records Track
- TREC Clinical Decision Support Track
- TREC Precision Medicine Track

TREC Genomics Track (Hersh, 2009)

- Motivated by completion of Human Genome Project, exploding research in genomics, and inability to biologists to know all that might impact work
- First TREC track devoted to “domain-specific” retrieval, with focus on IR systems for genomics researchers
  - Supported by NSF Information Technology Research (ITR) grant
- History
  - 2004-2005 – ad hoc retrieval and document categorization with bibliographic (MEDLINE) records
  - 2006-2007 – passage retrieval and question-answering (QA) with full-text journal articles
Lessons learned (Hersh, 2009)

• Ad hoc retrieval
  – Modest benefit for techniques known to work well in general IR, e.g., stop word removal, stemming, weighting
  – Query term expansion, especially domain-specific and/or done by humans, helped most

• QA
  – Most consistent benefit from synonym query expansion and paragraph-length or larger passage retrieval

• For all experiments (and papers describing them), major problems were
  – Lack of detailed description of systems
  – Use of low-performing baselines

Image retrieval – ImageCLEF medical image retrieval task

• Biomedical professionals increasingly use images for research, clinical care, and education, yet we know very little about how to best retrieve them

• Developed test collection and exploration of information needs motivating use of image retrieval systems (Hersh, 2006; Hersh, 2009; Müller, 2010)
  – Funded with supplement to NSF ITR grant

• Started with ad hoc retrieval and added tasks
  – Modality detection
  – Case finding

• In general, text retrieval yielded most consistent results with image features providing variable value

• Continues on with highly defined tasks
TREC Medical Records Track
(Voorhees, 2011; Voorhees, 2012)

• Adapting IR techniques to electronic health records (EHRs)
• Use case somewhat different – want to retrieve records and data within them to identify patients who might be candidates for clinical studies
• Motivated by larger desire for “re-use” of clinical data (Safran, 2007)
• Opportunities facilitated by incentives for “meaningful use” of EHRs in the HITECH Act (Blumenthal, 2011; Blumenthal, 2011)

Challenges for informatics research with medical records

• Has always been easier with knowledge-based content than patient-specific data due to a variety of reasons
  – Privacy issues
  – Task issues
• Facilitated with development of large-scale, de-identified data set from University of Pittsburgh Medical Center (UPMC)
• Launched in 2011, repeated in 2012
Test collection

VISIT LIST

REPORT-VISIT MAP

17,265 visits 101,712 reports (93,552 mapped to visits)

TREC Medical Records Track results

- Highly variable across different topics
  - Easiest — consistently best results
    - 105: Patients with dementia
  - Hardest — consistently worst results
    - 108: Patients treated for vascular claudication surgically
  - Large differences between best and worst results
    - 125: Patients co-infected with Hepatitis C and HIV
- Overall results show substantial room for improvement
  - Best results involve manual modification of queries

(Courtesy, Ellen Voorhees, NIST)
Which approaches worked?

- Best results in 2011 and 2012 obtained from NLM group (Demner-Fushman, 2011; Demner-Fushman, 2012)
  - Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007)
- Many approaches known to work in general IR fared less well, e.g., term expansion, document focusing, etc.
  - Other domain-specific approaches also did not show benefit, e.g., creation of PICO frames, negation
- Some success with
  - Results filtered by age, race, gender, admission status; terms expanded by UMLS Metathesaurus (King, 2011)
  - Expansion by concepts and relationships in UMLS Metathesaurus (Martinez, 2014)
  - Pseudorelevance feedback using ICD-9 codes (Amini, 2016)

Failure analysis for 2011 topics (Edinger, 2012)

<table>
<thead>
<tr>
<th>Reasons for Incorrect Retrieval</th>
<th>Number of Visits</th>
<th>Number of Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visits Judged Not Relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic terms mentioned as future possibility</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Topic symptom/condition/procedure done in the past</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>All topic criteria present but not in the time/sequence specified by the topic description</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>Most, but not all, required topic criteria present</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Topic terms denied or ruled out</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>Notes contain very similar term confused with topic term</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>Non-relevant reference in record to topic terms</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>Topic terms not present—unclear why record was ranked highly</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Topic present—record is relevant—disagree with expert judgment</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>Visits Judged Relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic not present—record is not relevant—disagree with expert judgment</td>
<td>44</td>
<td>21</td>
</tr>
<tr>
<td>Topic present in record but overlooked in search</td>
<td>103</td>
<td>27</td>
</tr>
<tr>
<td>Visit notes used a synonym or lexical variant for topic terms</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>Topic terms not named in notes and must be inferred</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Topic terms present in diagnosis, list but not visit notes</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
TREC Clinical Decision Support Track (Roberts, 2016)

- [www.trec-cds.org](http://www.trec-cds.org)
- Ad hoc search of biomedical literature (PubMed Central Open Access Subset – 1.25M articles)
- Topics are patient descriptions in three information need categories
  - Diagnosis
  - Test
  - Treatment
- Currently in third year of operation
- Transitioning to Precision Medicine Track

TREC biomedical tracks have inspired other challenge evaluations

- i2b2
  - [https://www.i2b2.org/NLP](https://www.i2b2.org/NLP)
  - Various NLP-related tasks, including extraction and de-identification
- CLEF eHealth
  - [https://sites.google.com/site/clefehealth/home](https://sites.google.com/site/clefehealth/home)
  - Information extraction and patient-centered IR
- bioCADDIE
  - [https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration](https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration)
  - Data set retrieval
System-oriented retrieval is not enough

- How well do IR systems enable users to perform their tasks?
  - Major task in medicine: answering clinical questions
- Undertook parallel approaches assessing Boolean vs. natural language searching in
  - Medicine – using
    - Electronic textbook – Scientific American Medicine (Hersh, 1995)
    - Bibliographic database – MEDLINE (Hersh, 1996)
  - General news – TREC Interactive Track (Hersh, 2001)
- Largest study employed 45 medical and 21 NP students answering tripartite questions using MEDLINE, with results showing (Hersh, 2002)
  - Medical students improved from chance (33%) to 50% correct, whereas NP students showed little benefit from using IR system
  - Recall and precision of searches showed no association with correct answers

Conclusions

- TREC challenges have provided insights and enduring test collections for continued research
- Importance of IR in biomedicine will continue, especially as volume, variety, and velocity of science continue to expand
  - Varying benefits for different use cases, but in general, medical vocabulary resources offer most value via query expansion
- While ad hoc IR for general information needs relatively understood, still challenges with
  - Novel types of data, e.g., EHRs and other structured data
  - High-recall tasks, e.g., systematic reviews
- Research confounded by larger issues, e.g.,
  - Private data
  - Proprietary data