Challenge Evaluations in Biomedical Information Retrieval

William Hersh, MD
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


Challenge Evaluations in Biomedical Information Retrieval

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

Two talks today

• Primer on information retrieval and challenge evaluations

• TREC challenge evaluations – practice talk for TREC 25th anniversary event
Learning objectives

• Define the evaluation measures used in information retrieval system evaluation and how they are used in challenge evaluations
• Describe the biomedical “tracks” in the Text Retrieval Conference (TREC) challenge evaluations
• Discuss the major results and findings of the TREC biomedical tracks

Information retrieval (IR, aka search)

• Focus on indexing and retrieval of (predominantly) knowledge-based information
• Historically centered on text in knowledge-based documents, but increasingly associated with many types of content
• [www.irbook.info](http://www.irbook.info)
Use cases for IR

- Historically, retrieval of knowledge
  - Documents, especially journal articles (originally abstracts)
  - Multimedia – images, sounds, video, etc.
  - Hypermedia – Web-based content
- Newer foci
  - Clinical data – e.g., cohort discovery from electronic health records
  - Data – e.g., finding data sets
Evaluation of IR systems has always been important

- System-oriented – how well system performs
  - Historically focused on relevance-based measures
    - Recall and precision – proportions of relevant documents retrieved
  - When documents ranked, can combine both in a single measure
    - Mean average precision (MAP)
    - Normal discounted cumulative gain (NDCG)
    - Binary preference (Bpref)
- User-oriented – how well user performs with system
  - e.g., performing task, user satisfaction, etc.

System-oriented IR evaluation

- Historically assessed with test collections, which consist of
  - Content – fixed yet realistic collections of documents, images, etc.
  - Topics – statements of information need that can be fashioned into queries entered into retrieval systems
  - Relevance judgments – by expert humans for which content items should be retrieved for which topics
- Evaluation consists of runs using a specific IR approach with output for each topic measured and averaged across topics
Recall and precision

• Recall
  \[ R = \frac{\text{#retrieved and relevant documents}}{\text{#relevant documents in collection}} \]
  – Usually use relative recall when not all relevant documents known, where denominator is number of known relevant documents in collection

• Precision
  \[ P = \frac{\text{#retrieved and relevant documents}}{\text{#retrieved documents}} \]

Some measures can be combined into a single aggregated measure

• Mean average precision (MAP) is mean of average precision for each topic (Harman, 2005)
  – Average precision is average of precision at each point of recall (relevant document retrieved)
  – Despite name, emphasizes recall

• Bpref accounts for when relevance information is significantly incomplete (Buckley, 2004)

• Normal discounted cumulative gain (NDCG) allows for graded relevance judgments (Jarvelin, 2002)

• MAP and NCDG can be “inferred” when there are incomplete judgments (Yilmaz, 2008)
Challenge evaluations

- A common approach in computer science, not limited to IR
- Develop a common task, data set, evaluation metrics, etc., ideally aiming for real-world size and representation for data, tasks, etc.
- In case of IR, this usually means
  - Test collection of content items
  - Topics of items to be retrieved – usually want 25-30 for “stability” (Buckley, 2000)
  - Runs from participating groups with retrieval for each topic
  - Relevance judgments of which content items are relevant to which topics – judged items derived from submitted runs

Challenge evaluations (cont.)

- Typical flow of events in an IR challenge evaluation
  - Release of document collection to participating groups
  - Experimental runs and submission of results
  - Relevance judgments
  - Analysis of results

- In IR, challenge evaluation results usually show wide variation between topics and between systems
  - Should be viewed as relative, not absolute performance
  - Averages can obscure variations
Some well-known challenge evaluations in IR

  - Many “tracks” of interest, such as routing/filtering, Web searching, question-answering, etc.
  - Mostly non-biomedical, but some tracks focused on genomics, EHRs, etc.
  - Started as track in TREC in 1996, spun off in 2000 to Cross-Language Evaluation Forum
  - Focus on retrieval across languages, European-based
  - Additional focus on image retrieval, which includes medical image retrieval tasks – www.imageclef.org (Hersh, 2009; Müller, 2010)
- TREC has inspired other challenge evaluations, e.g.,
  - i2b2 NLP Shared Task, https://www.i2b2.org/NLP/
  - bioCADDIE Dataset Retrieval Challenge – https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration

IR and text mining in context of biomedical knowledge management

All literature

↓

Possibly relevant literature (abstracts)

↓

Definitely relevant literature (full text)

↓

Actionable knowledge

(Hersh, 2009)

Information retrieval

Information extraction, text mining
The TREC Bio/Medical Tracks

- Why is information retrieval (IR) important in biomedicine
- TREC Genomics Track
- ImageCLEFmed
- TREC Medical Records Track
- TREC Clinical Decision Support Track
- TREC Precision Medicine Track
- Beyond system-oriented evaluation
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, 2012)
- “Precision medicine” will increase quantity and complexity of data (Tenenbaum, 2016)

TREC Genomics Track (Hersh, 2009)

- Motivated by 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 – focus on ad hoc retrieval and document categorization
  - 2006-2007 – focus on passage retrieval and question-answering as means to improve document retrieval
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 query expansion and paragraph-length 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)
• Started with ad hoc retrieval and added tasks
  – Modality detection
  – Case finding
• Overall conclusions: text yielded most consistent results with image features providing variable value
• Continues on with highly defined tasks
TREC Medical Records Track (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

RESULTS SUMMARY
1. Malignant Tumors
2. Neurologies
3. Gastroenterology
4. Gynecology
5. Endocrinology
6. Ophthalmology
7. History of smoking

Report Extract
20071020ER-9yWuGIE5Xz4-488-54123111
20073482DS-5668322-100-34234561
20071020RAD-9yWuGIE5Xz4-488-1222308213
20073482DS-5668322-100-34234561
20071027HP-9yWuGIE5Xz4-488-1348148618
20073482DS-5668322-100-34234561
2007100542DS-5668322-100-34234561
20073482HP-5668322-100-342348370
200729AC-5668322-100-34234832847
20071023HP-9yWuGIE5Xz4-488-1617508066
2007348932DS-5668322-100-34234561
2007348932DS-5668322-100-34234561
20071030DS-9yWuGIE5Xz4-488-865668696
2007348922RAD-5668322-800-8734245323

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

Results for 2012

<table>
<thead>
<tr>
<th>Run</th>
<th>mINDCG</th>
<th>mAP</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLMMenal*</td>
<td>0.600</td>
<td>0.306</td>
<td>0.759</td>
</tr>
<tr>
<td>uselSUM</td>
<td>0.578</td>
<td>0.286</td>
<td>0.592</td>
</tr>
<tr>
<td>unnaaan2</td>
<td>0.547</td>
<td>0.275</td>
<td>0.557</td>
</tr>
<tr>
<td>obsuMallod*</td>
<td>0.526</td>
<td>0.250</td>
<td>0.611</td>
</tr>
<tr>
<td>notgeo1</td>
<td>0.524</td>
<td>0.224</td>
<td>0.519</td>
</tr>
<tr>
<td>UDarboMed123</td>
<td>0.517</td>
<td>0.236</td>
<td>0.528</td>
</tr>
<tr>
<td>wogTrMcRbR8</td>
<td>0.509</td>
<td>0.231</td>
<td>0.553</td>
</tr>
<tr>
<td>NICTAUBC4</td>
<td>0.487</td>
<td>0.216</td>
<td>0.517</td>
</tr>
</tbody>
</table>

Per-topic scores for mINDCG computed over all runs

[Graph showing per-topic scores for mINDCG computed over all runs]
Which approaches did (and did not) work?

- 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
- 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 has inspired and guided other challenge evaluations in biomedicine

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

- My initial focused on concept-based searching (Hersh, 1990)
  - Did not impart value over word indexing and searching (Hersh, JAMIA, 1994)
- Experience of several evaluations led to concern with evaluation focus on recall/precision (Hersh, JASIS, 1994)
  - How much difference is meaningful?
  - How valid is batch evaluation for understand how well user will search?

Led to “task-oriented” evaluation approaches

- Motivated by Egan (1989) and Mynatt (1992)
- Major task in medicine: answering questions
- How can we evaluate systems in interactive use for answering questions?
- Undertook parallel approaches in
  - Medicine – using
    - Electronic textbook – Scientific American Medicine (Hersh, 1995)
    - Bibliographic database – MEDLINE (Hersh, 1996)
  - General news – TREC Interactive Track (Hersh, 2001)
Factors associated with successful searching (Hersh, 2002)

- Medical and nurse practitioner (NP) students success of using a retrieval system to answer clinical questions
  - Had to provide not only answer but level of evidence supporting it
    - Yes with good evidence
    - Indeterminate evidence
    - No with good evidence
- Look at factors associated with success
  - Based on model of factors associated with successful use of retrieval systems (Fidel, 1983) adapted to this setting
    - Including recall and precision
  - Dependent variable was correctness of answer
- Major results
  - Before searching, correct rate due to chance (~32%)
  - Medical students (~50%) but not NP students (~33%) improved with searching
  - Spatial visualization associated with higher rate of success
  - Recall and precision had no association with success

Conclusions

- Importance of IR in biomedicine will not diminish 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 solved, 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