The TREC Conferences: An Introduction

Ellen Voorhees
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Text REtrieval Conference (TREC)
Talk Outline

• General introduction to TREC
  - TREC history
  - TREC impacts

• Cranfield tradition of laboratory tests
  - mechanics of building test collections
  - test collection quality
  - legitimate uses of test collections

• IR evaluation primer
What is TREC?

• A workshop series that provides the infrastructure for large-scale testing of (text) retrieval technology
  - realistic test collections
  - uniform, appropriate scoring procedures
  - a forum for the exchange of research ideas and for the discussion of research methodology
TREC Philosophy

- TREC is a modern example of the Cranfield tradition
  - system evaluation based on test collections
- Emphasis on advancing the state of the art from evaluation results
  - TREC’s primary purpose is not competitive benchmarking
  - experimental workshop: sometimes experiments fail!
Yearly Conference Cycle

- Call for Participation
- Task Definition
- Document Procurement
- Topic Development
- IR Experiments
- Relevance Assessments
- Results Evaluation
- Results Analysis
- TREC Conference
- Proceedings Publication

Text REtrieval Conference (TREC)
TREC 2004 Program Committee

Ellen Voorhees, chair
James Allan
Chris Buckley
Gord Cormack
Sue Dumais
Donna Harman
Dave Hawking
Bill Hersh

David Lewis
John Prager
John Prange
Steve Robertson
Mark Sanderson
Karen Sparck Jones
Ross Wilkinson
TREC 2004 Track Coordinators

Genomics: Bill Hersh
HARD: James Allan
Novelty: Ian Soboroff
Question Answering: Ellen Voorhees
Robust Retrieval: Ellen Voorhees
Terabyte: Charlie Clarke, Ian Soboroff
Web: David Hawking, Nick Craswell, Ian Soboroff
A Brief History of TREC

• 1992: first TREC conference
  - started by Donna Harman and Charles Wayne as 1 of 3 evaluations in DARPA’s TIPSTER program
  - first 3 CDs of documents from this era, hence known as the “TIPSTER” CDs
  - open to IR groups not funded by DARPA
    • 25 groups submitted runs
  - two tasks: ad hoc retrieval, routing
    • 2GB of text, 50 topics
    • primarily an exercise in scaling up systems
A Brief History of TREC

- 1993 (TREC-2)
  - true baseline performance for main tasks
- 1994 (TREC-3)
  - initial exploration of additional tasks in TREC
- 1995 (TREC-4)
  - official beginning of TREC track structure
- 1998 (TREC-7)
  - routing dropped as a main task, though incorporated into filtering track
- 2000 (TREC-9)
  - ad hoc main task dropped; first all-track TREC
TREC Tracks

- Task that focuses on a particular subproblem of text retrieval

- Tracks invigorate TREC & keep TREC ahead of the state-of-the-art
  - specialized collections support research in new areas
  - first large-scale experiments debug what the task really is
  - provide evidence of technology’s robustness
TREC Tracks

- Set of tracks in a particular TREC depends on:
  - interests of participants
  - appropriateness of task to TREC
  - needs of sponsors
  - resource constraints

- Need to submit proposal for new track in writing to NIST
TREC Tracks

- Retrieval in a domain
  - Answers, not docs

- Web searching, size

- Beyond text
- Beyond just English

- Human-in-the-loop

- Streamed text

- Static text

- Genome
- Novelty
- Q&A
- Terabyte Web VLC
- Video Speech OCR
- X→{X,Y,Z}
- Chinese Spanish
- Interactive, HARD
- Filtering Routing
- Ad Hoc, Robust
TRE C Tasks

Number of Experiments


Text REtrieval Conference (TREC)
Participant Growth in TREC

More than 250 distinct groups have participated in at least one TREC.
TREC Impacts

- Test collections
- Incubator for new research areas
- Common evaluation methodology and improved measures for text retrieval
- Open forum for exchange of research
- Technology transfer
TREC Impacts

Cornell University TREC Systems

Mean Average Precision

- TREC-1
- TREC-2
- TREC-3
- TREC-4
- TREC-5
- TREC-6
- TREC-7
- TREC-8
## Ad Hoc Technologies

<table>
<thead>
<tr>
<th></th>
<th>TREC-2</th>
<th>TREC-3</th>
<th>TREC-4</th>
<th>TREC-5</th>
<th>TREC-6</th>
<th>TREC-7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term weights</strong></td>
<td>baseline start of Okapi wts</td>
<td>Okapi perfects “BM25” algorithm</td>
<td>new wts for SMART, INQUERY, PIRCS</td>
<td>Okapi/SMART wts used by others</td>
<td>adaptations of Okapi/SMART algorithm in most systems</td>
<td>new Twente and BBN models</td>
</tr>
<tr>
<td><strong>Passages</strong></td>
<td>use of subdocs by PIRCS</td>
<td>heavy use of passages/subdocs</td>
<td>decline in use of passages</td>
<td></td>
<td>use of passages in relevance feedback</td>
<td>multiple uses of passages</td>
</tr>
<tr>
<td><strong>Auto query expansion</strong></td>
<td>start of expansion using top X documents</td>
<td>heavy use of expansion using top X documents</td>
<td>start of more complex expansion</td>
<td>more sophisticated expansion experiments by many groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Manual query mods</strong></td>
<td>manual expansion using other sources</td>
<td>experiments in manual editing/user-in-the-loop</td>
<td>extensive user-in-the-loop experiments</td>
<td>simpler user-specific strategies tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other new areas</strong></td>
<td>initial use of data fusion</td>
<td></td>
<td>start of concentration on initial topic</td>
<td>more complex use of data fusion</td>
<td>continued focus on initial topic, especially the title</td>
<td></td>
</tr>
</tbody>
</table>
TREC Impacts

• Test collections
  • 28/57 SIGIR 2004 papers used TREC data

• Common evaluation methodology and improved measures for text retrieval
  • documents best practices in IR research methodology for new researchers

• Incubator for new research areas
  • PhD theses resulting from CLIR, SDR, QA participation
TREC Impacts

• Open forum for exchange of research
  • TREC proceedings unreviewed but have CiteSeer impact rating in top 30% of all CS venues (greater than CACM or ACM DL, for example)
  • TREC papers figure prominently in IR syllabi on the web
  • publication of all results prevents unsuccessful research from being duplicated

• Technology transfer
  • impact is far greater than just those who actually participate
Talk Outline

- General introduction to TREC
  - TREC history
  - TREC impacts

- Cranfield tradition of laboratory tests
  - mechanics of building test collections
  - test collection quality
  - legitimate uses of test collections

- IR evaluation primer
Cranfield Tradition

- Laboratory testing of system components
  - fine control over variables
  - abstraction from operational setting
  - comparative testing

- Test collections
  - set of documents
  - set of questions
  - relevance judgments
TREC approach

Assessors create topics at NIST

Topics are sent to participants, who return ranking of best 1000 documents per topic

NIST forms pools of unique documents from all submissions which the assessors judge for relevance

Systems are evaluated using relevance judgments
Creating a test collection for an ad hoc task

- topic statements
  - Automatic: no manual intervention
  - Manual: everything else, including interactive feedback

queries

representative document set

ranked list
Creating Relevance Judgments

RUN A

401

RUN B

401

Top 100

Pools

401

402

403

Alphabetized Docnos

Text REtrieval Conference (TREC)
Documents

• Must be representative of real task of interest
  - genre
  - diversity (subjects, style, vocabulary)
  - amount
  - full text vs. abstract

• TREC
  - generally newswire/newspaper
  - general interest topics
  - fulltext
Topics

- Distinguish between stmt of user need (topic) & system data structure (query)
  - topic gives criteria for relevance
  - allows for different query construction techniques

- TREC topics are NOT all created equal
  - 1-150: very detailed, rich content
  - 151-200: method of topic creation resulted in focused, easy topics
  - 201-250: single sentence only
  - 301-450: title is set of hand-picked keywords
Relevance Judgments

- Main source of criticism of Cranfield tradition
  - In test collections, judgments are usually binary, static, and assumed to be complete.
  - But...
    - “relevance” is highly idiosyncratic
    - relevance does not entail utility
    - documents have different degrees of relevance
    - relevance can change over time for the same user
    - for realistic collections, judgments cannot be complete
Relevance Judgments

• Consistency
  - idiosyncratic nature of relevance judgments does not affect comparative results

• Incompleteness
  - the important issue is that relevant judgments be unbiased
    • complete judgments must be unbiased
  - TREC pooling has been adequate to produce unbiased judgments
Consistency

• Mean Kendall τ between system rankings produced from different qrel sets: .938
• Similar results held for
  • different query sets
  • different evaluation measures
  • different assessor types
  • single opinion vs. group opinion judgments
Average Precision by Qrel

- **Mean**
- **Original**
- **Union**
- **Intersection**
QA Judgments

• Judging correctness, not relevance

• Assessors have differences of opinions as to what constitutes a correct answer
  - granularity of names, dates
  - assumed context

• Comparative evaluation stable despite those differences
Incompleteness

- Study by Zobel [SIGIR-98]:
  - Quality of relevance judgments does depend on pool depth and diversity
  - TREC ad hoc collections not biased against systems that do not contribute to the pools
  - TREC judgments not complete
    - additional relevant documents distributed roughly uniformly across systems but highly skewed across topics
Uniques Effect on Evaluation

![Graph showing the difference in MAP over runs.](Image of graph)
Uniques Effect on Evaluation: Automatic Only

![Graph showing the effect of uniques on evaluation. The x-axis represents the run number, and the y-axis represents the number of unique queries by group. A separate y-axis shows the difference in MAP, with values ranging from 0 to 0.0025.]
Cranfield Tradition

- Test collections are abstractions, but laboratory tests are useful nonetheless
  - evaluation technology is predictive (i.e., results transfer to operational settings)
  - different relevance judgments almost always produce the same comparative results
  - adequate pools allow unbiased evaluation of unjudged runs
Cranfield Tradition

- Note the emphasis on **comparative** !!
  - absolute value of effectiveness measures not meaningful
    - absolute value changes as relevance judgments change
    - theoretical maximum of 1.0 for both recall and precision not obtainable by humans (inter-assessor judgments suggest 65% precision at 65% recall)
  - evaluation results are only comparable when they are from the same collection
    - a subset of a collection is a different collection
    - comparisons between different TREC collections are invalid
Sensitivity Analysis

- With archive of TREC results, have empirically determine relationship between number of topics, $\Delta$ of scores, & error rate [Voorhees & Buckley, 2002]
  - error rates generally larger than accounted for in literature
  - confidence increases with topic set size
  - confidence also increases with larger $\Delta$, but then power of comparison reduced
  - confidence can be increased by repeating experiment on multiple collections
Talk Outline

- General introduction to TREC
  - TREC history
  - TREC impacts

- Cranfield tradition of laboratory tests
  - mechanics of building test collections
  - test collection quality
  - legitimate uses of test collections

IR evaluation primer
Ad hoc results — Cornell University

Summary Statistics

<table>
<thead>
<tr>
<th>Run Number</th>
<th>CertArel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run Description</td>
<td>Category A, Automatic, long</td>
</tr>
<tr>
<td>Number of Topics</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of documents over all topics</td>
<td></td>
</tr>
<tr>
<td>Retrieved</td>
<td>50000</td>
</tr>
<tr>
<td>Relevant</td>
<td>4611</td>
</tr>
<tr>
<td>Rejected</td>
<td>2590</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recall Level Precision Averages</th>
<th>Document Level Averages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>0.00</td>
<td>0.7013</td>
</tr>
<tr>
<td>0.10</td>
<td>0.9550</td>
</tr>
<tr>
<td>0.20</td>
<td>0.9150</td>
</tr>
<tr>
<td>0.30</td>
<td>0.9650</td>
</tr>
<tr>
<td>0.40</td>
<td>0.9287</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1775</td>
</tr>
<tr>
<td>0.60</td>
<td>0.1402</td>
</tr>
<tr>
<td>0.70</td>
<td>0.1015</td>
</tr>
<tr>
<td>0.80</td>
<td>0.0538</td>
</tr>
<tr>
<td>0.90</td>
<td>0.0224</td>
</tr>
<tr>
<td>1.00</td>
<td>0.0091</td>
</tr>
</tbody>
</table>

Average precision over all relevant docs non-interpolated | 0.2639

Text REtrieval Conference (TREC)
Evaluation Measure Criteria

- Related to a user satisfaction
- Interpretable
- Able to average or collect
- Have high discrimination power
- Able to be analyzed
# Evaluation Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Non-Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>r</td>
<td>n-r</td>
</tr>
<tr>
<td>Non-Retrieved</td>
<td>R-r</td>
<td>N-n-R+r</td>
</tr>
</tbody>
</table>

- \( N \) = number docs in collection
- \( n \) = number docs retrieved
- \( R \) = number relevant docs
- \( r \) = number relevant retrieved
Uninterpolated R-P Curve for Single topic

![Graph showing precision versus recall for different runs]

Legend:
- run1
- run2
- run3
Interpolated R-P Curves for Individual Topics
Single Number Summary Scores

- Precision (n): \( r / n \)
- Recall(n): \( r / R \)
- Average precision: \( \text{Avg}_{rd} (\text{Prec(rank of rd)}) \)
- R-Precision: \( \text{Prec}(R) \)
- Recall at .5 precision
  - use \( \text{Prec}(10) \) if precision < .5 in top 10
- Rank of first relevant (expected search length)
Document Level Measures

- **Advantage**
  - immediately interpretable

- **Disadvantages**
  - don’t average well
    - different number of relevant implies topics are in different parts of recall-precision curve
    - theoretical maximum impossible to reach
  - insensitive to ranking: only # rels that cross cut-off affect ranking
    - less useful for tuning a system
Number Relevant

![Bar chart showing number of relevant documents for different document levels.](image)
Average Precision

- Advantages
  - sensitive to entire ranking: changing a single rank will change final score
  - stable: a small change in ranking makes a relatively small change in score
  - has both precision- and recall-oriented factors
    - ranks closest to 1 receive largest weight
    - computed over all relevant documents
- Disadvantages
  - less easily interpreted
## Runs Ranked by Different Measures

<table>
<thead>
<tr>
<th>P(10)</th>
<th>P(30)</th>
<th>R-Prec</th>
<th>Ave Prec</th>
<th>Recall at .5 Prec</th>
<th>Recall (1000)</th>
<th>Total Rel</th>
<th>Rank of 1st Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>INQ502</td>
<td>INQ502</td>
<td>ok7ax</td>
<td>ok7ax</td>
<td>att98atdc</td>
<td>ok7ax</td>
<td>ok7ax</td>
<td>tno7tw4</td>
</tr>
<tr>
<td>ok7ax</td>
<td>ok7ax</td>
<td>INQ502</td>
<td>att98atdc</td>
<td>att98atde</td>
<td>mds98td</td>
<td>ok7ax</td>
<td>bbn1</td>
</tr>
<tr>
<td>att98atdc</td>
<td>att98atdc</td>
<td>ok7am</td>
<td>INQ502</td>
<td>att98atde</td>
<td>ok7am</td>
<td>att98atdc</td>
<td>INQ502</td>
</tr>
<tr>
<td>nect’chall</td>
<td>nect’chall</td>
<td>att98atdc</td>
<td>mds98td</td>
<td>att98atde</td>
<td>mds98td</td>
<td>att98atdc</td>
<td>nect’chall</td>
</tr>
<tr>
<td>att98atde</td>
<td>att98atde</td>
<td>INQ501</td>
<td>att98atdc</td>
<td>INQ502</td>
<td>att98atdc</td>
<td>ok7as</td>
<td>att98atdc</td>
</tr>
<tr>
<td>nect’chall</td>
<td>nect’chall</td>
<td>att98atdc</td>
<td>mds98td</td>
<td>nect’chall</td>
<td>mds98td</td>
<td>bbn1</td>
<td>tno7exp1</td>
</tr>
<tr>
<td>ok7as</td>
<td>ok7as</td>
<td>INQ503</td>
<td>tno7exp1</td>
<td>tno7exp1</td>
<td>tno7exp1</td>
<td>nect’chall</td>
<td>tno7exp1</td>
</tr>
<tr>
<td>nect’chall</td>
<td>nect’chall</td>
<td>bbn1</td>
<td>INQ501</td>
<td>INQ501</td>
<td>INQ501</td>
<td>ok7as</td>
<td>pirc8Aa2</td>
</tr>
<tr>
<td>tno7exp1</td>
<td>INQ502</td>
<td>mds98td</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
</tr>
<tr>
<td>mds98td</td>
<td>ok7as</td>
<td>att98atdc</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
</tr>
<tr>
<td>bbn1</td>
<td>nect’chall</td>
<td>tno7exp1</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Ad</td>
</tr>
<tr>
<td>nect’chall</td>
<td>ok7as</td>
<td>INQ503</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Ad</td>
</tr>
<tr>
<td>ok7as</td>
<td>nect’chall</td>
<td>pirc8Ad</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Ad</td>
</tr>
<tr>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Ad</td>
</tr>
<tr>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Ad</td>
</tr>
<tr>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>pirc8Aa2</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Aa2</td>
<td>nect’chall</td>
<td>pirc8Ad</td>
</tr>
<tr>
<td>Cor7A3rrf</td>
<td>Cor7A3rrf</td>
<td>Cor7A3rrf</td>
<td>Cor7A3rrf</td>
<td>Cor7A3rrf</td>
<td>Cor7A3rrf</td>
<td>Cor7A3rrf</td>
<td>Cor7A3rrf</td>
</tr>
<tr>
<td>tno7tw4</td>
<td>MerAbtnd</td>
<td>acsys7al</td>
<td>iowacuhk1</td>
<td>tno7tw4</td>
<td>MerAbtnd</td>
<td>acsys7al</td>
<td>tno7tw4</td>
</tr>
</tbody>
</table>

Ranked by measure averaged over 50 topics
## Correlations Between Rankings

<table>
<thead>
<tr>
<th></th>
<th>P(30)</th>
<th>R Prec</th>
<th>Ave Prec</th>
<th>Recall at .5 P</th>
<th>Recall (1000)</th>
<th>Total Rels</th>
<th>Rank 1st Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(10)</td>
<td>.8851</td>
<td>.8151</td>
<td>.7899</td>
<td>.7855</td>
<td>.7817</td>
<td>.7718</td>
<td>.6378</td>
</tr>
<tr>
<td>P(30)</td>
<td>.8676</td>
<td>.8446</td>
<td>.8238</td>
<td>.7959</td>
<td>.7915</td>
<td>.8320</td>
<td>.5896</td>
</tr>
<tr>
<td>R Prec</td>
<td></td>
<td></td>
<td>.9245</td>
<td>.8654</td>
<td>.8342</td>
<td>.8320</td>
<td>.5896</td>
</tr>
<tr>
<td>Ave Prec</td>
<td></td>
<td></td>
<td></td>
<td>.8840</td>
<td>.8473</td>
<td>.8495</td>
<td>.5612</td>
</tr>
<tr>
<td>R at .5 P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.7707</td>
<td>.7762</td>
<td>.5349</td>
</tr>
<tr>
<td>Recall(1000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.9212</td>
</tr>
<tr>
<td>Total Rels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.5891</td>
</tr>
</tbody>
</table>

Kendall’s $\tau$ computed between pairs of rankings
Known Item Search Evaluation

- Known item search: find document known to exist in collection
  - named page finding in web track
- Rewarded for retrieving particular target only, not related documents
Known Item Search Evaluation

- Mean reciprocal rank
  - use of reciprocal bounds measure & emphasizes differences that matter
  - equivalent to average precision with 1 rel
  - sensitivity of measure depends on size of ranked list

- Other statistics reported:
  - number of times target in first rank
  - number of times target not retrieved at all
Set-based Evaluation

- Required for some tasks
  - traditional Boolean searches
  - filtering
  - novelty

- 2 main approaches
  - utility functions
  - combinations of recall & precision
    - $F(\beta) = \frac{((\beta^2+1) \times P \times R)}{(\beta^2 P + R)}$
Summary

• TREC emphasizes individual experiments evaluated on a benchmark task
  - leverages modest government investment into substantially more R&D than could be funded directly
  - improves state-of-the-art
  - accelerates technology transfer