Overview of TREC 2002

Sponsored by:
NIST, ARDA, DARPA

Ellen Voorhees

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce
TREC 2002 Program Committee

Ellen Voorhees, chair
James Allan
Nick Belkin
Chris Buckley
Jamie Callan
Gord Cormack
Sue Dumais
Fred Gey

Donna Harman
Dave Hawking
Bill Hersh
Jim Mayfield
John Prange
Steve Robertson
Karen Sparck Jones
Ross Wilkinson
TREC 2002 Track Coordinators

Cross-Language Retrieval: Fred Gey & Doug Oard
Filtering: Steve Robertson, Jamie Callan, Ian Soboroff
Interactive: Bill Hersh & Paul Over
Novelty: Donna Harman
Question Answering: Ellen Voorhees
Video: Alan Smeaton & Paul Over
Web: David Hawking, Nick Craswell, Ian Soboroff
TREC Goals

- To increase research in information retrieval based on large-scale collections
- To provide an open forum for exchange of research ideas to increase communication among academia, industry, and government
- To facilitate technology transfer between research labs and commercial products
- To improve evaluation methodologies and measures for information retrieval
- To create a series of test collections covering different aspects of information retrieval
REtrieval in a domain
Answers, not documents
Web searching
Beyond text
Beyond just English
Human-in-the-loop
Streamed text
Static text

Genome
Novelty
Q&A
Web
Very large corpus
Video
Speech
OCR
X→{X,Y,Z}
Chinese
Spanish
Interactive
Filtering
Routing
Ad Hoc

Common Terminology

• “Document” broadly interpreted
  - page in a web search
  - shot in a video search

• Different types of tasks
  - ad hoc search
  - known-item search
  - filtering
Creating Relevance Judgments

RUN A

RUN B

Pools

401

402

403

Alphabetized Docs

Top 100
TREC 2002 Tracks

• Cross-language
• Filtering
  • adaptive, batch, routing
• Interactive
• Novelty
• Question Answering
  • main, list
• Video
  • shot boundaries, feature extraction, search
• Web
  • topic distillation, named page finding
Cross Language Track

• Task: ad hoc search for documents written in one language using topics in another language

  - Arabic documents:
    • 869 MB news articles from Agence France Presse Arabic newswire
    • May 13, 1994 - December 20, 2000
    • 383,872 articles
    • created & released by LDC
Cross Language Track

- topics:
  - 50 TREC topic statements in English created by bilingual assessors at LDC
  - translated into Arabic by assessor
  - translations vetted by track

- common resources
  - light stemmer
  - bidirectional Arabic-English dictionary
  - tables of translation probabilities
  - web-based bidirectional machine translation system
Cross Language Track

- submitted runs
  - 41 runs from 9 groups
  - 23 cross-language runs, 18 monolingual

- relevance judgments:
  - generally done by assessor who created topic
  - all submitted runs judged to depth 100
  - average pool size was 769, smaller than last year
  - average of 118.2 relevant per topic
    - minimum 3, maximum 523
  - uniques effect comparable to ad hoc collections
English to Arabic Results

![Graph showing precision and recall for different systems]

- UMassX6n
- BBN11XLC
- ibmy02c
- BKYCL2
- AutoClirDoc

Text REtrieval Conference (TREC)
Monolingual Results

![Monolingual Results Graph]

- UniNE3
- BKYM0N
- UMassM
- iit02mp1
- apl11ca1

Text REtrieval Conference (TREC)
Monolingual vs. Crosslingual

![Graph showing precision vs. recall for various models including UniNE3, BKYMON, UMassM, iit02mpl, UMassX6n, BBN11XLC, ibmy02c, and BKYCL2.](image-url)
Filtering Track

- Task: for each document in a document stream, decide whether to retrieve it in response to a standing query
  - 3 subtasks
    - routing
      - rank a new document set, given topic descriptions and a training set of relevant docs
    - batch filtering
      - decide whether to retrieve a document given topic descriptions and a training set of relevant docs
    - adaptive filtering
      - decide whether to retrieve a document given topic descriptions plus judgments for documents retrieved
Filtering Track

- documents: Reuters corpus volume 1
  - each tagged with Reuters category codes

- topics
  - 50 assessor-created topics
    - multiple iterations of judgments during construction to provide necessary relevance data for adaptive filtering
  - 50 topics constructed as intersection of Reuters category pairs
    - chosen to be reasonably meaningful as topic
    - have at least 3 relevant in training set
    - total relevant in same range as assessor-built topics
    - used to investigate categories-as-topics for cheap collection building
Filtering Track

- set evaluation
  - scaled utility
    \[ U = 2R^+ - N^+ \]
    then scaled using Ault's suggestion
  - \( F \) with \( \beta = .5 \)
    \[ T11F = \frac{1.25R^+}{(R^+ + N^+) + 0.25\text{NumRel}} \]

- routing runs produced ranked list so evaluated using mean average precision
Adaptive Filtering Results

Assessor Topics

Intersection Topics

T11SU measure; red line is scaled utility of retrieving no documents
Intersection Topics

- Goal was to test viability of using intersection of categories as topics
  - if documents are already categorized, collection building is very cheap
  - can form collections with many more topics

- Does not appear to be viable alternative
  - successful methods for assessor-built topics aren’t successful for intersection topics
  - true even for routing, where original topic statement is largely immaterial
Interactive Track

- Investigate searching as an interactive task by examining the process as well as the outcome
- Second year of a two-year plan
  - TREC 2001: observational study of subjects using live web to perform search task
  - TREC 2002: controlled laboratory experiment of hypothesis suggested by observations
Interactive Track

• Task: use .GOV collection to accomplish 8 search tasks analogous to those used in TREC 2001
  - four general search activities
    • looking for personal health information
    • seeking guidance on US government laws, regulations, guidelines, or policy
    • making travel plans
    • gathering material for a report on a given subject
  - two templates for searcher tasks
    • find N short answers to a question, where each answer is an instance of the same type
    • find N websites that meet the need
Interactive Studies

CSIRO: Is knowledge of organizational structure helpful for organizing and delivering documents?

Glasgow: Do hierarchical clustering and summarization visualization techniques improve the presentation of long document lists?

OHSU: factors associated with successful search

Rutgers: Does reducing the amount of interaction required of the searcher lead to increased satisfaction?
Does increased query length improve retrieval effectiveness?

UNC, Chapel Hill: Is 3D visualization better than text?

UToronto: What makes a good information exploration interface?
Novelty Track

- New track for TREC 2002
- Goal: investigate systems’ abilities to locate relevant and non-redundant information within a ranked list of docs
- Motivation: reduce user’s workload by eliminating extraneous information from system response
Novelty Track

- **Task**
  - given is ranked list of relevant documents segmented into sentences & topic statement
  - return 1) the set of sentences containing relevant information, and 2) a subset of the relevant sentences such that redundant information is eliminated

- **Collection:**
  - document set used in TREC5 6-8 ad hoc
  - 50 topics taken from TREC5 6-8
Novelty Evaluation

• Reference data created by assessors
  - created the two sentence sets manually
  - each topic independently judged twice
    evaluation based on the sentence sets of the
    judge who selected fewer relevant sentences
  - one topic removed since minimum assessor
    found no relevant sentences

• Measures
  - set recall and precision for both sentence
    sets
  - recall*precision as measure for averaging
Novelty Track Results

![Graph showing average recall vs. precision for various systems. The x-axis represents different systems: CIIRO2, thunv2, pics2N02, ss1, ntcslabvp, U1owa02Nov4, cmu02+300Cr, fdut11n2. The y-axis represents average recall.*precision. The legend includes relevant, new, random-rel, random-new, human-rel, and human-new categories.](image-url)
Question Answering Track

- Goal: encourage research into systems that return answers, rather than document lists
  - 2 subtasks
    - main: for each of 500 questions, return exactly one response and rank questions by confidence in the answer
    - list: assemble a set of instances as the answer to a question
  - for both tasks, response is a [doc, string] pair where string must be an exact answer and doc supports that answer
AQUAINT Document Collection

- New collection created for track
  - LDC catalog number LDC2002T31
- News articles
  - AP newswire, 1998-2000
  - Xinhua News Agency (English), 1996-2000
- 3 gb text, approx. 1,033,000 articles
QA Main Task

- Questions
  - Drawn from MSNSearch and AskJeeves logs
  - no guarantee that question has answer in collection, so a response could be `NIL`
  - else, response was a single [doc, string] pair
  - whole set of questions ranked by confidence in answer

- Evaluated using analog of MAP
  - strict scoring: only `right` answers contributed to score
Exact Answers

- Human assessors judged responses
  - **Wrong**: string does not contain a correct answer or answer is unresponsive
  - **Not Supported**: string contains a correct answer, but doc does not support that answer
  - **Not Exact**: string contains correct answer and doc supports it, but string contains too much (or too little) info
  - **Right**: string is exactly a correct answer that is supported by the doc
Distribution of Judgments

- 15,948 judgments across all questions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12,639</td>
<td>79.3%</td>
</tr>
<tr>
<td>505</td>
<td>3.2%</td>
</tr>
<tr>
<td>442</td>
<td>2.8%</td>
</tr>
<tr>
<td>2,362</td>
<td>14.8%</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong</td>
<td></td>
</tr>
<tr>
<td>Unsupported</td>
<td></td>
</tr>
<tr>
<td>inExact</td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td></td>
</tr>
</tbody>
</table>

- In general, systems can find extent of answer if they can find it at all
  - distribution skewed across systems
  - attempt to get exact answer sometimes caused units to be lost (so marked wrong)
Confidence-weighted Scoring

- Focus on getting systems to know when they have found a good answer
  - questions ranked by confidence in answer
  - compute score based on ranking

\[
\sum_{i=1}^{N} \frac{\text{number right to rank } i}{i}
\]
QA List Task

• Instance-finding task
  - 25 questions that specify a target number of instances to retrieve
    • List 9 types of sweet potatoes.
  - response is an unordered set of the target number of instances
    • an instance is a single [doc, string] pair
    • answer-string required to be exact
  - questions constructed by NIST assessors
    • target chosen such that collection had at least that number of instances but > 1 doc required
    • single document may have > 1 instance
QA List Evaluation

- Each list judged as a unit
  - instances marked right/inexact/unsupported/wrong
  - subset of right instances marked distinct

- Accuracy used as evaluation metric
  \[
  \frac{\# \text{ distinct instances}}{\text{target \# of instances}}
  \]
QA List Results

Average Accuracy

- LCClist2002
- SU11RT2
- UdeMlistNow
- shef11lo
- cl0211

Text REtrieval Conference (TREC)
Video Track

- Track to promote progress in content-based retrieval from digital video
  - three tasks:
    - shot boundary detection
    - feature extraction
    - search
Video Track Collection

- ~73 hours of MPEG-1/VCD recordings
  - participants downloaded directly from Internet Archive and Open Video Project web sites
  - films from 1930’s through 1970’s
    - advertising, educational, industrial films
    - original created by a variety of organizations for a variety of purposes
  - partitioned into different training and test sets for different tasks
Shot Boundary Task

- Task: automatic identification of the shot boundaries in a given clip

- Details:
  - test set: 18 videos
    - 2.88 GB of video
    - 545,068 frames
    - 2,090 shots
  - shots determined manually at NIST
    - cuts(70%), dissolves(24%), fades(3%), other(2%)
    - system’s boundary matched reference boundary if at least one frame overlapped
Shot Boundaries
Cuts vs. Gradual Transitions

![Graph showing precision vs. recall for Cuts and Gradual transitions. The graph indicates that Gradual transitions generally have higher precision at lower recall values, while Cuts tend to have higher precision at higher recall values.](image)
Feature Extraction

- Task: given a set of shot boundaries and a feature definition, find all shots that contain feature

- Details:
  - 23.2 hours (96 videos, 7891 shots) training;
    5.02 hours (23 videos, 1848 shots) test
  - shot has feature iff some frame within shot is characterized by feature
  - each feature tested independently
## Feature Extraction

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Outdoors</td>
<td>6. Landscape</td>
</tr>
<tr>
<td>2. Indoors</td>
<td>7. Text overlay</td>
</tr>
<tr>
<td>3. Face</td>
<td>8. Speech</td>
</tr>
<tr>
<td>4. People</td>
<td>9. Instrumental Sound</td>
</tr>
<tr>
<td>5. Cityscape</td>
<td>10. Monologue</td>
</tr>
</tbody>
</table>

### Average Precision

![Bar chart showing average precision for different categories.](chart.png)

- **x-axis**: Categories (1-10)
- **y-axis**: Average Precision

- **Legend**:
  - Median
  - Maximum
Search

- Task: traditional ad hoc task where “documents” are shots and topics are multimedia information need statements

- Details:
  - 40.12 hours (176 videos; 14,524 shots)
  - 25 topics created by NIST
    - contained textual description plus optional examples in other media
    - defined some topics such that features would be useful; groups shared feature extraction output
  - two types of runs: manual or interactive
Video Search Results

![Graph showing video search results with different lines representing Interactive-Max, Interactive-median, Manual-max, and Manual-median. The x-axis represents topics from 75 to 99, and the y-axis represents average precision ranging from 0 to 1.](Image)
Web Track

- Investigate retrieval behavior on the web
  - two tasks
    - topic distillation: similar to ad hoc task but goal is to find "key" pages, not "relevant" pages
    - named page finding: known-item task to find page specified in topic
  - document set
    - new (Jan. 2002) crawl of .GOV
    - approx. 18 gb
    - 1.25 million documents, including extracted text from PDF, postscript, & word documents
    - images within pages available, not part of 18 gb
Web Topic Distillation Task

- Task definition:
  - assemble a short, but comprehensive, list of pages that are good resources for topic

- Topics similar to ad hoc topics
  - 50 topics created by NIST assessors
  - target content for which .GOV has good resources

- Binary judgments by topic author
  - good key resource/not good key resource

- Evaluation by Prec(10)
  - defined in terms of good resource, not relevance
  - emphasize conciseness
Topic Distillation Results

Top 8 groups by average Precision(10) of best run.
Named Page Finding Task

- Generalization of homepage finding task
- Retrieve ranked list of top 50 pages for 150 requests
  - topic consists of a single phrase
    - US passport renewal
    - Child labor stamp
  - created by NIST assessors for track
- Evaluation: MRR of first correct page
  - small pools judged to find mirrors, aliases
  - 3 correct pages for 2 topics; 2 correct for 16 topics; else 1 known correct page
Web Homepage Finding Results

![Graph showing mean reciprocal rank and percentage of topics with no page found for various systems.](image)

**Text REtrieval Conference (TREC)**