TREC

TEXT RETRIEVAL CONFERENCE

Eric Robinson
Daniel Kunkle
Alan Feuer

What is TREC?

- Not your typical conference!
- NIST (National Institute of Standards and Technology) provides test data and questions
- Participants run their programs on the data
- Judges evaluate the results
- Participants share their experiences in the conference

Why TREC?

- Encourages research in text retrieval in large data sets
- By creating a “practical” conference, increases communication between academia, business, and government
- Provide availability of many different text retrieval techniques

Interesting Facts

- In 2003, included 93 groups from 22 countries
- Makes test collections and submitted retrieval code available to public
- First Large Scale:
  - Non-English retrieval
  - Retrieval of speech recordings
  - Retrieval across multiple languages

Tracks of TREC

- Each conference has many different “tracks”, or challenges
- Past tracks include:
  - Cross-Language Track – Same topic, many languages
  - Filtering Track – Stream of data, choose yes or no
  - Interactive Track – Retrieve while users access data
  - Novelty Track – Find new/original information in data
  - Video Track – Obtain information about video data
  - Web Track – Terabytes of data

Cross-Language Track

- Goal: Find text that pertain to a topic regardless of the language
- Input: Arabic Language Newswire Documents
- Question: English Topic

Video Track

- Goal: Segmentation, indexing, and content-based retrieval of video
- Input: Arabic, Chinese, and English news feeds
- Questions: Short Boundary, Low/High Level Feature, Search

The Seven Tracks in TREC'05

- Enterprise
- Genomics
- HARD
- Question Answering
- Robust Retreival
- SPAM
- Terabyte

Five Tracks in Brief

- The most common application is *document retrieval* in some context (more details later)

1. **Enterprise** - a wide variety of document types from some organization (email, spreadsheets, web, etc.)
2. **Genomics** - documents and data in genomics
3. **HARD** - **H**igh **A**ccuracy **R**etrieval from **D**ocuments, using information about the searcher and context
4. **Robust Retrieval** - focuses on traditionally difficult topics, where retrieval accuracy is consistently low
5. **Terabyte** - retrieval task with data sets of terabyte scale

Question Answering Track

- Short, specific answers to factual questions
- The first problem is usually *problem classification*

Question Classes

- ABBREVIATION
  - abbreviation
  - expression abbreviated
- ENTITY
  - animals
  - organs of body
  - colors
  - books and other creative pieces
  - currency names
  - diseases and medicine
  - events
  - food
  - musical instrument
  - languages
  - letters like a-z
  - other entities
  - plants
  - products
  - religions
  - sports
  - elements and substances
  - symbols and signs
  - techniques and methods
  - equivalent terms
  - vehicles
  - words with a special property
- DESCRIPTION
  - definition
  - description
  - manner of an action
  - reasons
- HUMAN
  - a group or organization of persons
  - an individual
  - title of a person
  - description of a person
- LOCATION
  - cities
  - countries
  - mountains
  - other locations
  - states
- NUMERIC
  - postcodes or other codes
  - number of something
  - dates
  - linear measures (distance)
  - prices
  - ranks
  - other numbers
  - the lasting time of something
  - fractions
  - speed
  - temperature
  - size, area and volume
  - weight

Question Classes

- Abbreviation (2)
  - *What does S.O.S. stand for ?*
- Entity (22)
  - *What fowl grabs the spotlight after the Chinese Year of the Monkey ?*
- Description (4)
  - *Why do heavier objects travel downhill faster ?*
- Human (4)
  - *When Mighty Mouse was conceived , what was his original name ?*
- Location (5)
  - *What country are you in if you woo in the Wu dialect ?*
- Numeric (13)
  - *What is the date of Boxing Day ?*

Answer Evaluation

- An answer contains :
  1. The question number
  2. The id of a document that supports the answer
  3. A rank (1-5) of this response for this question
  4. The text snippet returned as the answer
- The score for a questions is the reciprocal of the rank of the correct answer (0 if no correct answer)
  - e.g. if the system returned five answers and the third one was correct, the score is 1/3
  - If the corresponding document does not support the answer, it may be considered correct or incorrect

SPAM Track

- Classify a chronological sequence of email messages as either SPAM or HAM (not SPAM)
- Supervised learning task of the form:

  `initialize`

  `classify` *emailfile resultfile*

  `(train ham` *emailfile resultfile* OR

  `train spam` *emailfile resultfile)*

  `Finalize`

Evaluating SPAM Filters

- Judged on a human defined *gold standard, w.r.t. the following criteria*
- ham misclassifaction rate (HMR).
  - What fraction of ham messages are misclassified as spam?
- spam misclassifaction rate (SMR).
  - What fraction of spam messages are misclassified as ham?
- ham/spam learning curve.
  - Error rates as a function of number of messages processed
- ham/spam tradeoff curve
  - HMR versus SMR for various SPAM cutoff levels

Information Need

- What is being sought?
  - An information source
  - A particular document
  - An answer
- How can you test if you've found it?
  - Classical IR: Precision & Recall
  - Web search: Success at *n,* Reciprocal Rank

Procedure

1. Index a collection of documents
2. Run queries, generate ranked result list
3. Decide relevance of each result item
4. Compute statistics

Precision

- What fraction of results are relevant?

  relevant_retrieved / total_retrieved

- P@n: What fraction of first n results are relevant?

  relevant_retrieved_of_first_*n* / *n*

Precision

Results List	Relevant?	Rel/Ret	Precision
Doc2505	Yes	1/1	100%
Doc13	No	1/2	50%
Doc1271	No	1/3	33%
Doc16	Yes	2/4	50%
Doc678	Yes	3/5	60%
Doc1003	No	3/6	50%
. . .	. . .	. . .	. . .

Recall

- What fraction of relevant docs are retrieved?

  relevant_retrieved / total_relevant_docs

  ➢ *Need to know total set of relevant docs*

Recall

Results List	Relevant?	Rel/Total	Recall
Doc2505	Yes	1/10	10%
Doc13	No	1/10	10%
Doc1271	No	1/10	10%
Doc16	Yes	2/10	20%
Doc678	Yes	3/10	30%
Doc1003	No	3/10	30%
. . .	. . .	. . .	. . .

Assuming 10 documents are relevant

R Precision

- R Precision
  - Normalizes precision to be independent of the number of relevant documents
  - 11-point average R Precision used in TREC:

    Average of P@*r* for

    r = to recall levels 0.0, .01, …, 1.0

R Precision

Results List	Precision	Recall
Doc2505	100%	10%
Doc13	50%	10%
Doc1271	33%	10%
Doc16	50%	20%
Doc678	60%	30%
Doc1003	50%	30%
. . .	. . .	. . .

11-point R Precision

0% 100%

10% 100%

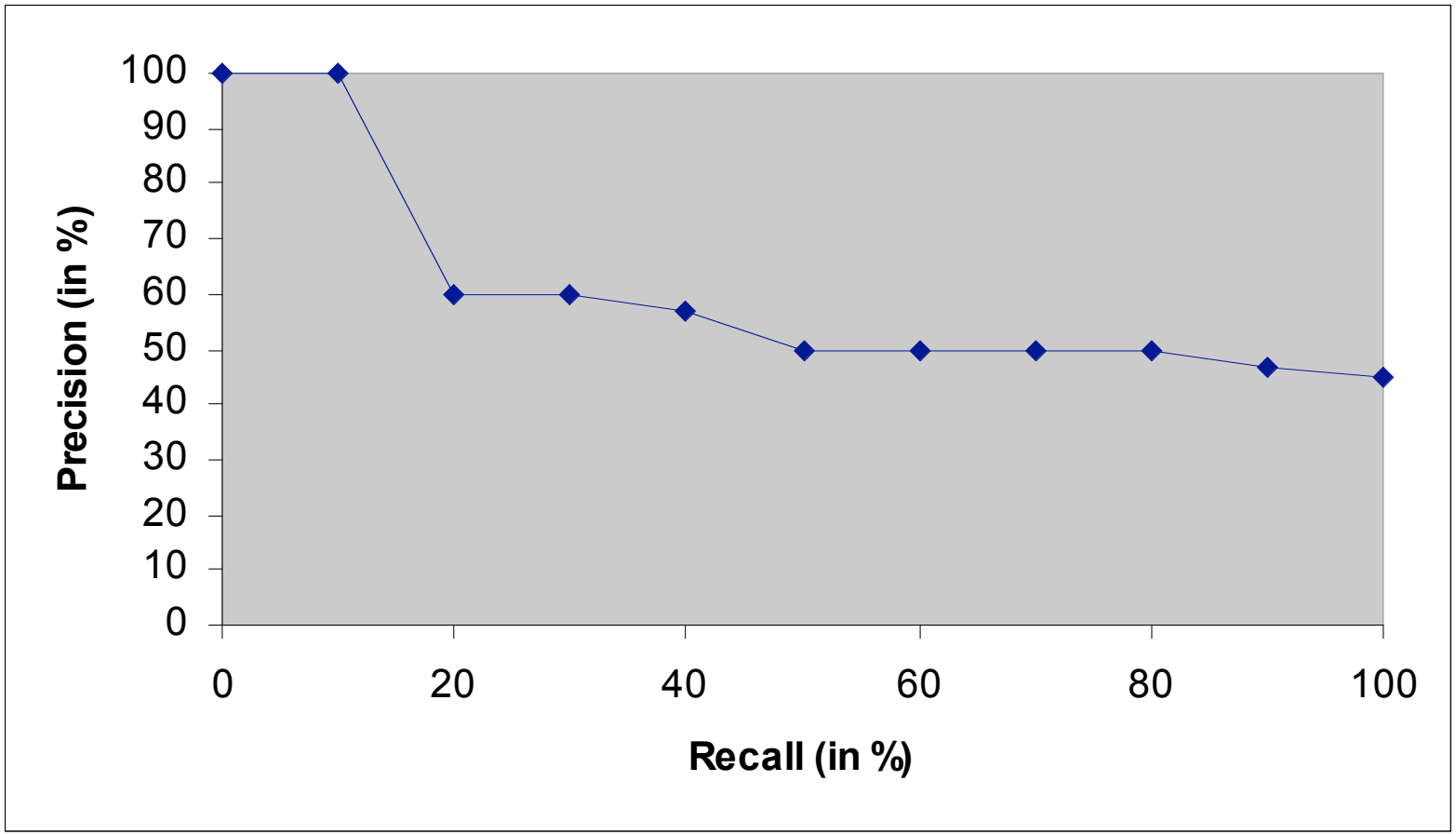
20% 50%

30% 60%

. . .

Mean= (1+1+.5+.6+…) / 11

Recall/Precision Graph



Evaluation of the Evaluation

- How well do TREC metrics predict user satisfaction for general IR?
  - Precision at *n*
    - Suppose fewer than *n* docs are relevant?
  - Doesn't consider presentation of results
    - Document summary
    - List vs. graphic
  - Doesn't consider search conversation
    - Follow-up operations
    - Modified queries

Evaluation of the Evaluation

- How well do TREC metrics predict user satisfaction for Web search?
  - Doesn't consider relative quality of docs
    - Recall often not important
    - Best page may be all that is needed
  - Most people only look at first 10 results
    - Could use P@10, but there may not be 10 relevant docs

Web Search

- Many searches are *navigational*

  Find a known Web site, *S*

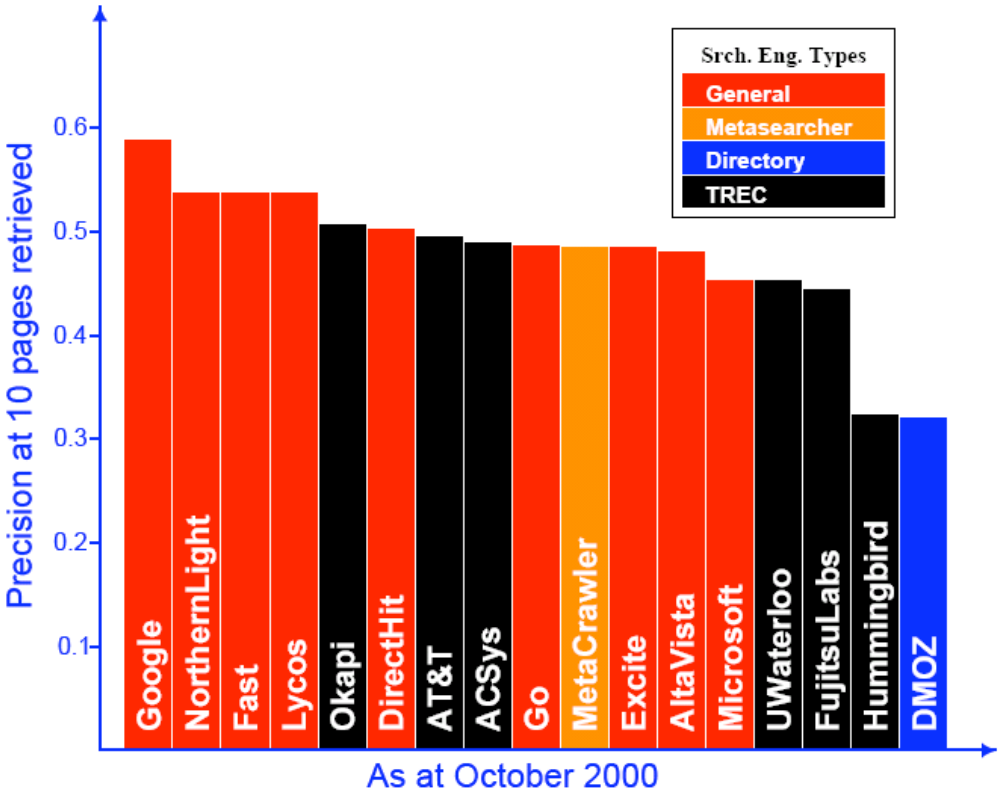
  – **Success at *n***: Is *S* in first *n* hits?

  – **Reciprocal Rank**: RR = 1/*i*

    - Where *i* is the 1-based index of *S* in the result list

➢ *Are precision and recall relevant to the Web?*

TREC vs. Web Search



I guess so!

References

1. David Hawking and Nick Craswell, *Very Large Scale Retrieval and Web Search,* 2004. http://es.csiro.au/pubs/trecbook_for_website.pdf
2. NIST, “Common Evaluation Measures”, in *NIST Special Publication: SP 500-261 The Thirteenth Text Retrieval Conference (TREC 2004).* National Institute of Standards and Technology, 2004.
3. Ellen M. Voorhees. “Overview of TREC 2004” in *NIST Special Publication: SP 500-261 The Thirteenth Text Retrieval Conference (TREC 2004).* National Institute of Standards and Technology, 2004.
4. Ian Witten, Alistair Moffat, and Timothy Bell, *Managing Gigabytes, Second Edition,* Morgan Kaufmann, 1999.
5. Text Retrieval Conference (TREC) Web Site, http://trec.nist.gov/, January 24, 2006.
6. Xin Li, Dan Roth, *Learning Question Classifiers*. COLING'02, Aug., 2002.
7. Experimental Data for Question Classification, http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/, January 24, 2006.
8. Gordon Cormack, Thomas Lynam, *A Study of Supervised Spam Detection Applied to Eight Months of Personal Email,* http://plg.uwaterloo.ca/~gvcormac/spamcormack.html, July 1, 2004