TREC 2007 Genomics Track Overview

William Hersh[1]
Aaron M. Cohen[1]
Phoebe Roberts[2]

[1]Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University

[2]Pfizer Corp.
Cambridge, MA

Acknowledgements

- Data development, management, and scoring – Aaron Cohen, Hari Krishna Rekapalli, Jianji Yang
- Relevance judgment process – Phoebe Roberts
- Data providers
  - Highwire Press
  - National Library of Medicine
- Funder
  - National Science Foundation Grant ITR-0325160
- NIST – Ellen Voorhees, Ian Soboroff, Lori Buckland
- Track steering committee
- All participating groups over 2002-2007

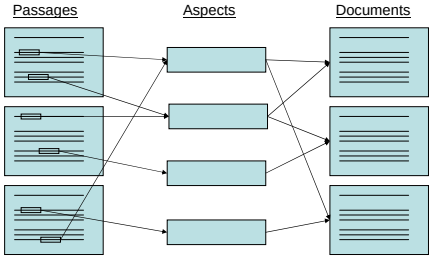
Overview

- Task
- Evaluation measures
- Document collection
- Topics
- Relevance judgments
- Results
- Future Directions

Task (or use case)

- No "central dogma" because it is no longer "dogma!"
- A scientist searching the biomedical literature, wanting to answers to questions but also context
- Aided by a system that provides:
  - Retrieval of passages – portions of text that contain an answer to the question
  - Grouped by aspects – to show which passages provide distinct information; a complete answer may require several different aspects
  - Linked to documents – that the user ultimately wants to retrieve and read
- Evaluated by mean average precision (MAP)-like measures for these three types of retrieval

From passages to aspects to documents



Evaluation measures

- Based on passage spans submitted in runs
  - Passages could not cross <P…> or </P…> boundaries
- Passage-level MAP – two measures:
  - Passage2 (official) – calculated MAP as if each character in each passage were a ranked document
  - Passage (from 2006) – calculated AP based on precision at passage retrieval (derived from HARD Track; Allan, 2004)
    - Found to be problematic after analysis of 2006 results
- Aspect-level MAP
  - Similar to approach used in TREC 6-9 Interactive Track (Hersh, 2001)
- Document-level MAP
  - "Rolled up" from passages to document level

Document collection

- Full text HTML journal articles that preserved formatting, structure, table and figure legends, etc.
- 162,259 documents from 49 journals published by Highwire Press (www.highwire.org)
  - Documents contained 12,641,127 maximum-length legal spans, which consisted of all text delimited by HTML paragraph tags and were listed in file legalspans.txt
- Corresponding MEDLINE record provided by National Library of Medicine (NLM), including Medical Subject Headings (MeSH) terms
  - Full-text file name was PMID provided by Highwire, with link to actual article in file metadata.txt

7

Topics

- Generated from interviews of real biologists
  - Collected ~50, with 36 used as official topics and remainder as sample topics
- Phrased as list entity-based questions
  - More general question format and broader coverage of subject domains
- Expanded to 14 different entity types from 5 generic topic types (GTTs) developed for 2005-2006

8

Example topic

- Collected information need: *What is the genetic component of alcoholism?*
- Transformed into a list question of the form: *What [GENES] are genetically linked to alcoholism?*
- Answers to this question are passages that relate one or more entities of type GENE to alcoholism
  - e.g., a valid and relevant answer to the above question would be: *The DRD4 VNTR polymorphism moderates craving after alcohol consumption.* (from PMID 11950104)
- GENE entity supported by this statement would be *DRD4*

9

Topic entity types and terms

Entity Type	Source of Terms	Topics with Entity Type
ANTIBODIES	MeSH	1
BIOLOGICAL SUBSTANCES	MeSH	3
CELL OR TISSUE TYPES	MeSH	2
DISEASES	MeSH	1
DRUGS	MEDLINEplus	2
GENES	iHoP, Harvester	11
MOLECULAR FUNCTIONS	GO	2
MUTATIONS	MeSH	1
PATHWAYS	BioCarta, KEGG	2
PROTEINS	MeSH	5
STRAINS	Ad hoc	2
SIGNS OR SYMPTOMS	MeSH	1
TOXICITIES	MeSH	2
TUMOR TYPES	MeSH	1

10

Relevance judgments – pooling, judging and processing

- Collected ranked passages in round robin manner from each submission until had 1000 per topic
- Judging led by Phoebe Roberts (PR) and performed by 13 biology experts (mostly with PhD)
- Developed form-based GUI, user documentation, and training session
- Judges instructed to:
  - Select passages (from maximum-length legal spans) that were definitely or possibly relevant
  - Group relevant passages into aspects, designated by one or more judge-assigned terms
- Work reviewed by Phoebe before accepted
- Python programs gathered results and calculated measures

11

Relevance judgment results

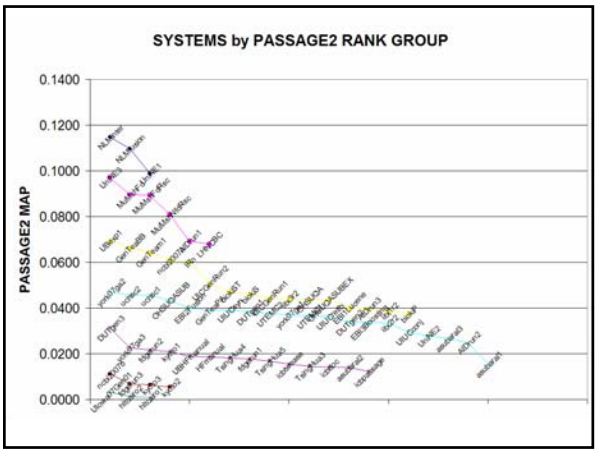
Measure	Average per topic
Relevant passages	124.8 (1-609)
Mean relevant passage length	968.0 (192-2381)
Standard deviation of relevant passage length	1276.2
Aspects	72.3 (1-577)
Mean aspects per relevant passage	1.63 (1.0-3.41)
Relevant documents	69.2 (1-483)

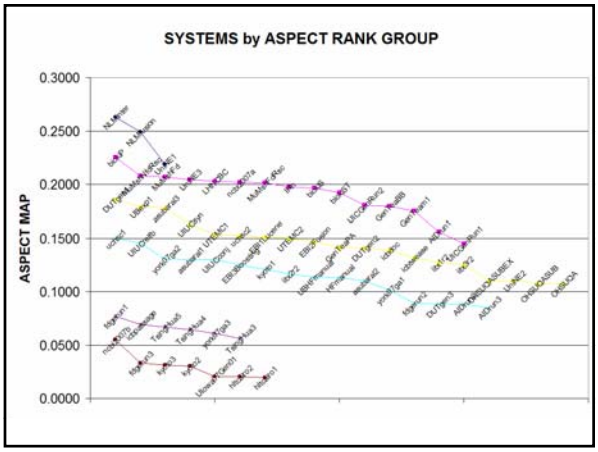
12

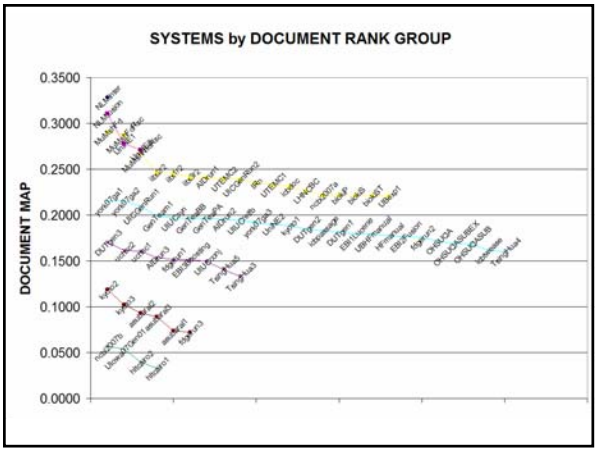
Summary of results – 66 runs from 25 groups

Measure	Passage2 MAP	Passage MAP	Aspect MAP	Document MAP
Minimum	0.0008	0.0029	0.0197	0.0329
Median	0.0377	0.0565	0.1311	0.1897
Mean	0.0398	0.0560	0.1326	0.1862
Maximum	0.1148	0.0976	0.2631	0.3286

- Comparable for automatic and interactive, lower for manual
- Following slides show results using RMEQ tool (Cohen and McWeeney, in preparation) that uses iterative repeated measures analysis to separate groups into statistically distinct rank groups







Preliminary analysis of results

- Level of performance of top systems respectable but somewhat lower than last year
  - List-entity type question more difficult than GTT question?
- Top systems did consistently well on all measures
  - Measures were highly correlated (more so for Passage2 than Passage)
- Unlike last year, aspect MAP was a meaningful measure of system topic coverage this year:
  - While the range of average number of aspects per relevant passage was low (1-3), number of aspects per topic was relatively high (could be over 300)
  - For a system to do well on aspect MAP, a number of passages with complementary aspect information would have to be retrieved and ranked highly, since for most topics, almost no single passages would cover all of the required entities
  - Enabled by allowing the judges to determine the list entities after a passage was judged relevant

Future directions

- This is a high-quality test collection that will hopefully spur further research in genomics/biomedical IR
- With NSF grant ending, this is last year of TREC Genomics Track
  - Test collections will continue to be available
  - Web site and email list will remain for now
  - Call for papers for special issue of *Information Retrieval* (see Web site) – papers due 3/31/08
  - I am looking forward to a real summer vacation in 2008 ☺
- Beyond 2007? Stay tuned…