Happy 20th Anniversary, TREC

TREC TRACKS

	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	
																				■	Crowdsourcing
Personal documents															■	■	■	■	■	■	Blog, Microblog
														■	■	■					Spam
Retrieval in a domain																		■	■	■	Chemical IR
											■	■	■	■	■	■				■	Genomics, Medical Records
Answers, not documents											■	■	■								Novelty
								■	■	■	■	■	■	■	■	■	□	■	■	■	QA, Entity
Searching corporate repositories															■	■	■	■	■	■	Legal
														■	■	■	■				Enterprise
Size, efficiency, & web search													■	■	■	■	■	■			Terabyte, Million Query
								■	■	■	■	■	■					■	■	■	Web
						■	■	■	■												VLC
Beyond text										■	■	□	□	□	□	□	□	□	□	□	Video
						■	■	■	■												Speech
				■	■																OCR
Beyond just English						■	■	■	■	■	■										Cross-language
					■	■															Chinese
			■	■	■																Spanish
Human-in-the-loop												■	■	■			■	■	■		HARD, Feedback
			■	■	■	■	■	■	■	■	■								■	■	Interactive, Session
Streamed text				■	■	■	■	■	■	■	■										Filtering
	■	■	■	■	■	■															Routing
Static text	■	■	■	■	■	■	■	■				■	■	■							Ad Hoc, Robust

Text REtrieval Conference (TREC)

Some lessons learned after 20

- People like to do this; an enormous amount of work is done by coordinators & participants!!
- Getting data has been/always will be a problem
  - However having good, freely available test collections is a major contribution of these evaluations
- Two years is about right for a given task!!
  - The first year is an exciting pilot getting the evaluation right; second year has training data for improvements; by the third year its boring (nothing new is being tried or learned)
- The Cranfield paradigm is surprisingly robust
  - The use of pooling for relevance assessment has been shown to result in stable evaluations (at least most of the time)!!
  - The paradigm has been successfully adapted to different media (such as video), different tasks (such as QA)!!

But--

- How do we balance “basic” IR research against moving into new and exciting areas
    - Beyond the early years, we are not seeing real improvements; this is discouraging, especially to today’s students!!
- Is it right to keep adapting Cranfield??
    - Does it mean we are blinding ourselves to “the big picture”; are the tasks we can model the important ones??
    - If not Cranfield, then what??
- How do we work in areas that are known to be important but either lack data (desktop search), or are difficult to cleanly define (such as different relevancy criteria like quality, time-dependency, etc.)

2011 Medical Records Track

- Ad hoc search task
  - set of ~ 100,000 de-identified clinical records assembled by U. of Pittsburgh's BLULab NLP repository
    - assembled into ~17,000 "visits" through mapping table
  - 35 topics developed and judged by physicians enrolled in OHSU bioinformatics program; modeled after inclusion criteria for clinical studies

    patients with complicated GERD who receive endoscopy
  - systems return ranked list of visits
- Evaluation
  - judgment sets produced using deep but sparse stratified sampling
  - bpref as main evaluation metric; inferred measures noisy with type of sampling used

2011 Microblog Track

- Test Collection
  - Tweets2011 collection of about 16 million tweets
  - 50 topics created by NIST assessors consisting of [title, triggerTweet] pairs where title is an English statement of the information need and triggerTweet is a pointer to a tweet in the collection
- Evaluation
  - pools of top 30 tweets from submitted runs
  - tweet is relevant if it contains relevant information itself or points to relevant information
    - Must be in English and NOT a retweet
    - Must precede time of triggerTweet

2011 Session Track

- 76 sessions derived from 62 topics
  - topics taken from previous tracks and faceted like web topics
- A session (created at U. Sheffield) consists of
  - sequence of queries issued to satisfy the information need of the topic; median of two reformations; 38% with more
  - ranked list of (top 10) URLs returned for each query
  - set of URLs clicked on plus dwell times for each query
- Four runs comprise single submission:
  - **R1**: results for last query using no other info
  - **R2**: results for last query, using content of all previous queries in session
  - **R3**: results for last query using content of previous queries plus ranked lists
  - **R4**: results for last query using content of previous queries, ranked lists, and click/dwell time info

2011 Crowdsourcing Track

- Investigate judgments from a crowd
  - participants collect assessments for sets of topic-doc pairs; 5 pairs per set
  - evaluate quality of crowdsourcing design by quality of the judgments (consensus or matching NIST)
- Given a set of labels for same [topic, doc] pair, compute a final label
  - test data built from crowdsourcing judgments collected from TREC 2010 Relevance Feedback track; evaluate quality of consensus labels as either function of gold standard [NIST] labels or as function of others' consensus labels

More TRECs

- TREC 2012 Tracks
  - Crowdsourcing, Microblog, Medical Records, Session, Web continuing (Legal had no new data)
  - Knowledge Base Acceleration (KBA)
    - Update Wikipedia entities based on extraction from streaming data
  - Contextual Suggestion
    - Given a set of profiles, a set of example suggestions, and a set of contexts, for each profile/context pairing, participants should return a ranked list of 50 proposed suggestions
- TREC 2013??