News from TREC 2011

Ian Soboroff

National Institute of Standards and Technology
U.S. Department of Commerce

Happy 20th Anniversary, TREC

TREC
2011
20th Anniversary

TREC 2011 Tracks

- **Chemical IR**: patent search and chemical structure search
- **Crowdsourcing**: examine crowd sourced judgments
- **Entity**: semantic search and linked open data
- **Legal**: e-discovery and privilege
- **Medical**: search in unstructured medical records
- **Microblog**: real time search in Twitter
- **Session**: evaluate multi-query sessions
- **Web**: diversity ranking and hard queries

TREC 2011 Tracks

- **Chemical IR**: patent search and chemical structure search
- **Crowdsourcing**: examine crowd sourced judgments
- **Entity**: semantic search and linked open data
- **Legal**: e-discovery and privilege
- **Medical**: search in unstructured medical records
- **Microblog**: real time search in Twitter
- **Session**: evaluate multi-query sessions
- **Web**: diversity ranking and hard queries

Crowdsourcing Track

- New for 2011
- A meta-track:
  - investigate best practices for using crowdsourcing to build IR evaluation resources
- Thanks to CrowdFlower for providing crowdsourcing resources to track participants

Assessment Task

- Investigate how best to gather relevance judgments from a crowd
  - participants collect assessments for sets of topic-doc pairs
  - 5 pairs per set; crowd worker must judge all 5
  - judgments either absolute labels or preference relations (up to participant)
  - topics drawn from previous TRECs; ClueWeb docs
  - evaluate quality of crowdsourcing design by quality of the judgments obtained
    - as computed over either gold standard [previous NIST] or consensus judgments

Consensus Task

- 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
  - entire data set contains 19,033 [topic, doc] pairs judged by 762 workers who produced 89,624 binary judgments; 3275 pairs have NIST judgments
  - evaluate quality of consensus labels as either function of gold standard [NIST] labels or as function of others' consensus labels

Medical Records Track

- New in 2011
- Foster research on providing content-based access to free text fields of electronic health records
  - health IT high priority in US
  - standard IR techniques likely not optimal given nature of language use in records

Medical Records Track Task

- 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

Microblog Track

- New track for 2011
- 58 (!) participants in inaugural year
- Goal
  - examine search tasks and evaluation methodologies for information seeking behaviors in microblogging environments
  - initial task is real-time search task: user wishes to see most recent relevant information to the query

Microblog Track

- Documents
  - Tweets2011 collection
  - sample of Twitter tweets from two-week span Jan24—Feb 8, 2011
  - sample includes spam tweets, retweets, replies, non-English tweets, tweets containing URLs
  - about 16 million tweets
  - corpus released as set of tweet ids plus tools to fetch own copy of those tweets

Microblog Track

Title: BBC World Service staff cuts
querytime: Tue Deb 08 12:30_27 +0000 2011
querytweettime: 34952194402811904

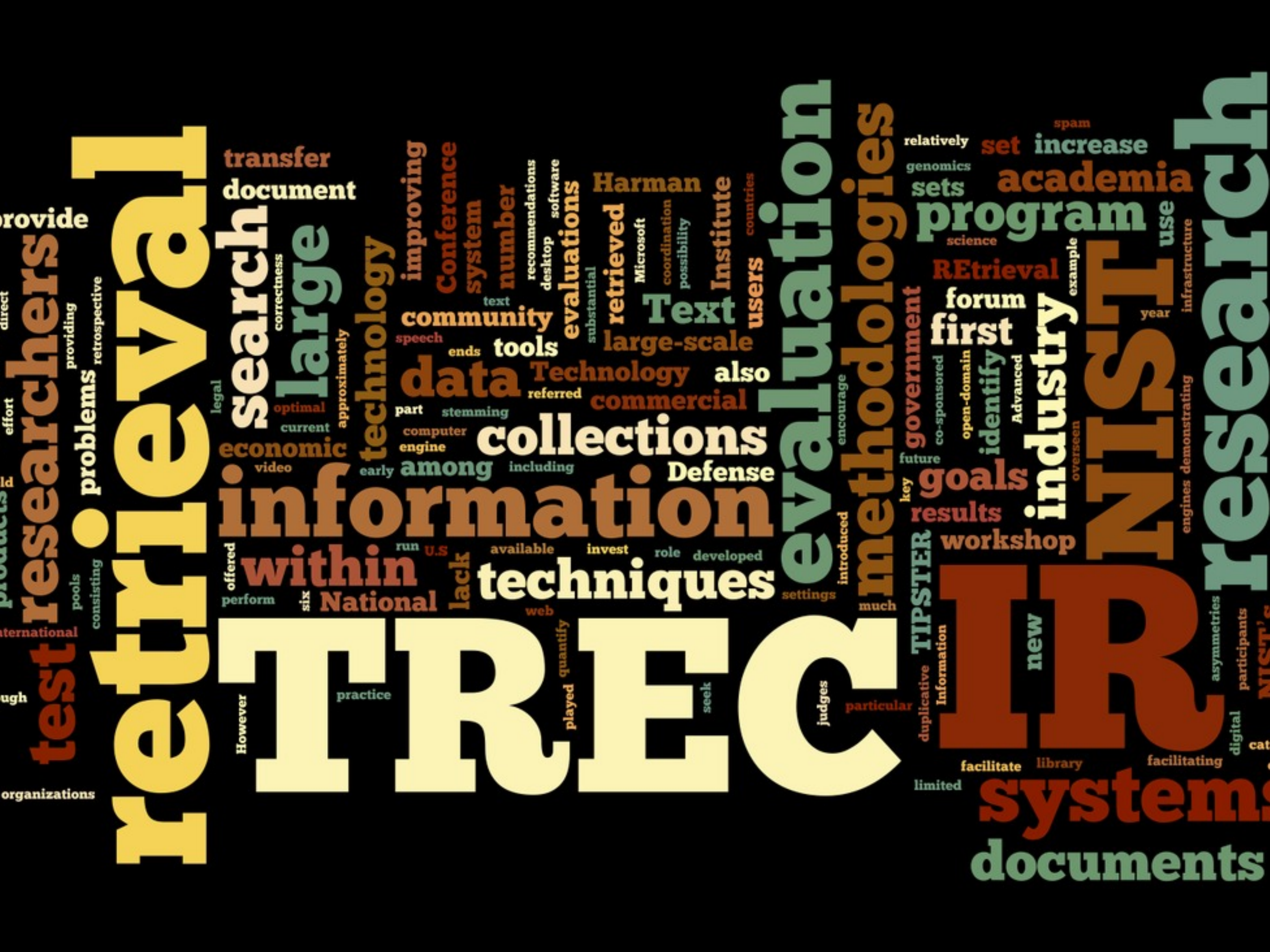
- Topics
  - 50 topics created by NIST assessors
  - [title, triggerTweet] pairs where title is an English statement of the information need and triggerTweet is a pointer to a tweet in the collection
  - triggerTweet defines the "time" of the query
    - triggerTweet may or may not be relevant to query
  - systems return relevant tweets in reverse chronological order starting at time of triggerTweet
  - use of future information (after time of query) or external information (not in Tweets2011) required to be declared

Microblog Track

- Evaluation
  - judgment sets are pools of top 30 tweets from submitted runs (ordered by time)
  - tweet is relevant if it contains relevant information itself or points to relevant information
    - must precede time of triggerTweet
    - non-English declared not relevant by fiat
    - retweets declared not relevant by definition
  - Prec(30) as main evaluation measure

TREC 2012

- TREC to continue in 2012
- Tracks
  - Crowdsourcing, Legal, Microblog, Medical Records, Session, Web continuing
  - New: Knowledge Base Acceleration (KBA)
    - Identify articles to update a Wikipedia page
  - New: Contextual Suggestion
    - "Entertain me" given user and context information


TREC
IR
retrieval
information
research
NIST
evaluation
methodologies
systems
documents
collections
techniques
search
large
researchers
test
within
data
program
academia
industry
Text
first
forum
goals
results
workshop
increase
set
sets
relatively
spam
genomics
science
use
REtrieval
example
infrastructure
year
transfer
document
provide
problems
providing
retrospective
direct
effort
products
pools
consisting
international
organizations
correctness
technology
improving
Conference
system
number
recommendations
desktop
software
evaluations
retrieved
Microsoft
coordination
possibility
countries
Institute
users
Harman
community
text
speech
ends
tools
large-scale
also
Technology
referred
commercial
legal
optimal
approximately
part
stemming
current
computer
engine
economic
video
early
among
including
Defense
run
U.S
available
invest
role
developed
offered
perform
six
lack
National
settings
much
introduced
encourage
TIPSTER
Information
duplicative
new
asymmetries
participants
NIST's
digital
facilitating
library
facilitate
limited
particular
judges
seek
played
quantify
web
However
practice
government
co-sponsored
open-domain
identify
Advanced
future
key
overseen
engines
demonstrating
substantial