# Review of TREC 2016

Jaime Arguello
jarguell@email.unc.edu

January 23, 2017

# Clinical Decision Support

- Goal: developing systems that can link EHRs with scientific literature that are relevant for patient care
- Input:
  - EHR: (1) summary, (2) description, or (3) note
  - Info need: (1) tests,(2) diagnosis, or (3) treatment
- Output: ranked list of scientific articles

# Clinical Decision Support

**Topic 1** – Diagnosis
**Note**:
78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital6 4406**]
(transferred to nursing home for rehab on [**12-8**] after several falls out
of bed.) He was then readmitted to [**Hospital6 1749**] on
[**3120-12-11**] after developing acute pulmonary edema/CHF/unresponsiveness?.
There was a question whether he had a small MI; he reportedly had a
small NQWMI. He improved with diuresis and was not intubated.
.
Yesterday, he was noted to have a melanotic stool earlier this evening
and then approximately 9 loose BM w/ some melena and some frank blood
just prior to transfer, unclear quantity.
**Description**:
78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI. Yesterday, he was noted to have a melanotic stool and then today he had approximately 9 loose BM w/ some melena and some frank blood just prior to transfer, unclear quantity.
**Summary**:
A 78 year old male presents with frequent stools and melena.

# Clinical Decision Support

- Corpus: 1.25 million articles from PubMed Central
- Topics: 30 real de-identified EHRs from Boston ICU's
- Requirements: 2/5 runs had to use notes
- Judgements: produced by physicians (definitely relevant, possibly relevant, not relevant)
- Metrics: P@10, R-precision, infNDCG, infAP

# Contextual Suggestion

- Goal: developing systems that can recommend points of interest (POIs) that are contextually relevant to a user
- Input:
  - Context: city (e.g., Boston), trip type (e.g., business), trip duration (e.g., weekend), group type (e.g., friends), season (e.g. summer)
  - Profile: ratings from POIs for the same user in a different context, possible tags
- Output: ranked list of up to 50 POIs

# Contextual Suggestion

## Context

- City: ID, city, state, latitude, longitude
- Trip type: business, holiday, other
- Trip duration: night out, day trip, weekend, longer
- Group type: alone, friends, family, other
- Season: Winter, summer, fall, spring

# Contextual Suggestion

## Profile


```
{"id":743,
"body": {
    "group": "Friends",
    "season":"Summer",
    "trip_type":"Holiday",
    "duration":"Weekend trip",
    "location":{
        "state":"TX",
        "id":306,
        "name":"Waco",
        "lat":31.54933,
        "lng":-97.14667},
    "person": {
                "gender": "Male",
                "age": 28,
                "id": 15012,
                "preferences":[
                    {
                    "rating":4,
                    "documentId":"TRECCS-00211395-161",
                    "tags":[
                         "Beer",
                         "Culture",
                         "Cocktails",
                         "Restaurants",
                         "Food",
                         "pub-hopping",
                         "cocktails",
                         "bar-hopping"
                         ]},
                    ...
                    ]
                    }},
"candidates":[
    {"documentId":"TRECCS-00267253-306",
    "tags":[
        "Beer",
        "Cocktails",
        "Family Friendly",
        "Restaurants",
        "Food"
        ]},
    {"documentId":"TRECCS-00294259-306",
    "tags":[
    "Tourism",
    "Bar-hopping",
    "Restaurants",
    "Entertainment",
    "Live Music"
    ]},
     ...
    ]
    }
```

# Contextual Suggestion

- Phase 1: return POIs from TREC CS Web Corpus
- Phase 2: return POIs from a set of candidate POIs provided as input

# Live Question Answering

- Goal: developing systems that can respond to questions in real time (within one minute)
- Input: recent and unanswered question posted to Yahoo! Answers (YA) (opinions, advice, polls, etc.)
- Output: a 1000-character response or “no response”

# Live Question Answering
## Questions

- QID: unique identifier
- Title: typically one sentence question
- Body: additional context provided by the user
- Category: arts, beauty, health, home, pets, sports, travel

# Live Question Answering
## Evaluation

- Real-time evaluation: web service accepted requests and responded within one minute
- Responses rated on a 0-3 scale by NIST assessors
- Metrics:
  - AvgScore: average score (null = bad)
  - Succ@i+: % of Q's with score(R) >= i (null = bad)
  - Prec@i+: % of R's with score(R) >= i
- YA answers included in the evaluation

# Real-time Summarization

- Goal: developing systems that can monitor a data stream (e.g., tweets) and push content that is relevant, novel (with respect to previous pushes), and timely
- Input: a query describing the topic or event of interest
- Evaluation period: 8/2/2016 - 8/11/2016 (10-day period)
- Output:
  - Scenario A: up 10 tweets per day (ASAP)
  - Scenario B: up 100 tweets per day (Midnight)

# Real-time Summarization
## Evaluation

- About 200 interest profiles (“predicting the future”)
- Pooling: Scenario A and B results were pooled
- Judging: not relevant, relevant, highly relevant
- Clustering: relevant and highly relevant tweets were clustered manually per query

# Real-time Summarization

## Scenario A: Metrics

- Expected Gain: G(NR) = 0.0, G(R) = 0.5, G(HR) = 1.0

$$\frac{1}{N}\sum \mathrm{G}(t)$$

- Credit given for only first tweet from each cluster
- Separate metrics to model “silent days”

# Real-time Summarization

## Scenario B: Metrics

- NDCG@10
- Separate version to model “silent days”
- On a “silent day”, a system that does not push any results gets a score of 1.0, and a system that does push results gets a score of 0.0

# Tasks Track

- Goal: developing systems that can infer the user's higher-level task(s) and return results that are relevant to the task(s)
- Input: a query describing the information need
- Output:
  - Task Understanding: return a ranked list of 1000 key-phrases describing all possible tasks
  - Task Completion : return a ranked list of 1000 documents relevant to all possible tasks
  - Ad-hoc : return a ranked list of 1000 documents relevant to any possible task

# Tasks Track
## Evaluation

- Possible tasks created by NIST assessors from pooled key phrases returned by participants
- Metrics: favor relevance and novelty

$$\text{METRIC-IA}(\mathcal{R}_Q) = \sum_{I \in \text{INTENTS(Q)}} P(I|Q) \times \text{METRIC}(\mathcal{R}_I)$$

# Open Search Track

- Goal: developing systems that can improve their ranking performance using “real” user implicit feedback
- Live Systems: CiteSeerX, SSOAR, Microsoft Academic

# Open Search Task
## Evaluation

- System: select frequent queries and top-100 results
- System: split into training set and test set
- System: send training set to participant
- Participant: re-rank training set, send back to system
- System: interleave system with participant's results
- System: send interactions (clicks/skips) to participant
- Participant: re-train ranker, re-rank test set, send to system
- System: interleave system with participant's results

# Open Search Task
## metrics

- Wins: number of queries for which more participant results were clicked
- Losses: number of queries for which more system results were clicked
- Ties: number of queries for which the same number of participant and system results were clicked
- Outcome: wins / (wins + losses + ties)

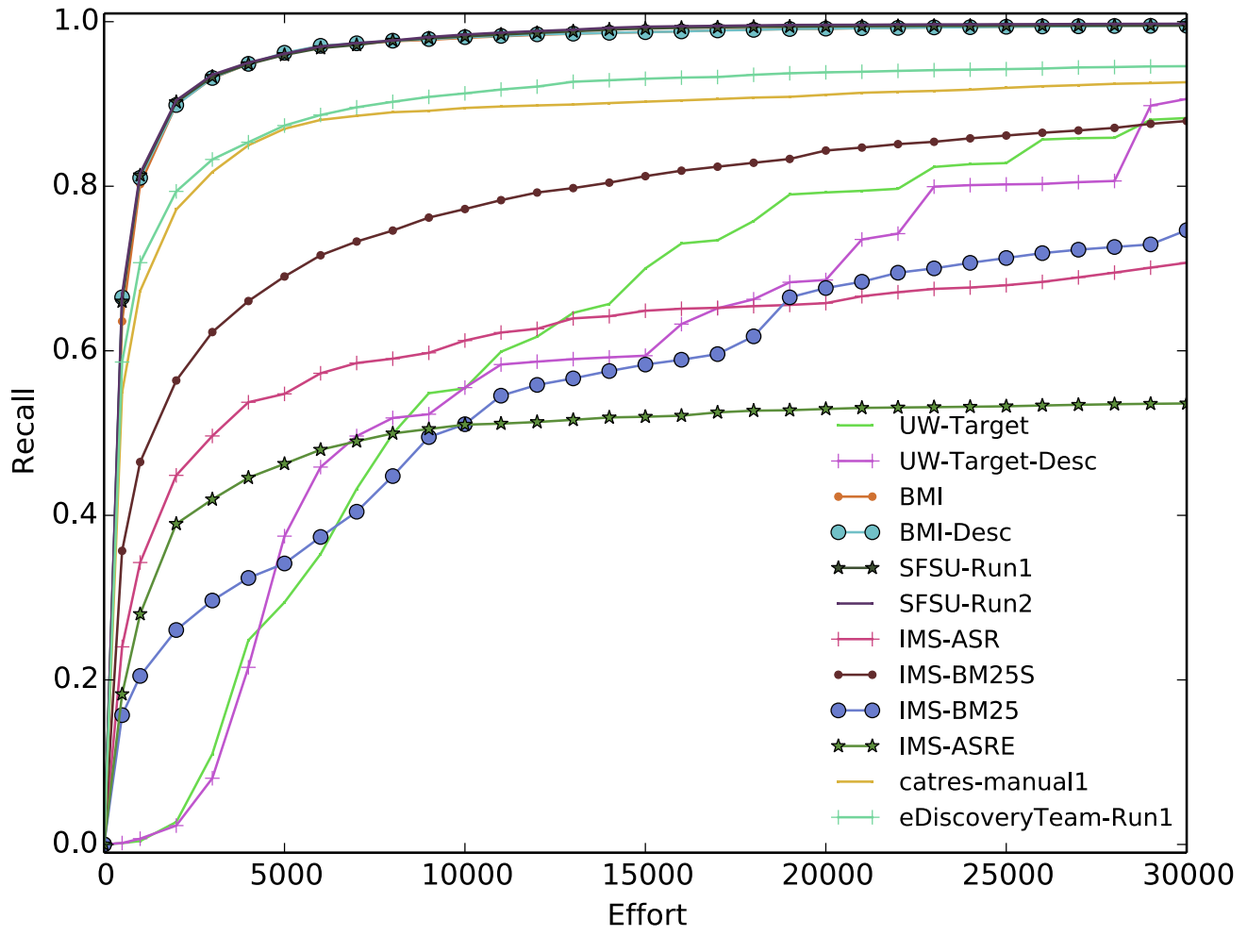
# Total Recall Track

- Goal: developing systems that can produce document-at-a-time results, accept feedback, and achieve high recall and precision (i.e., decide when to stop)
- At home task: downloadable collections, web server to simulate "human in the loop"
- Sandbox task: remote collections, web server provided statistics for ranking and simulated "human in the loop"

# Total Recall

## metrics

- R = number of relevant documents for the query
- Recall@aR+b: recall after showing "human" aR+b results
- Recall vs. Effort



# Dynamic Domain Track

- Goal: developing systems that can produce 5-document-at-a-time results, accept feedback (at the sub-topic and passage level), and decide when to stop

# Dynamic Domain Track

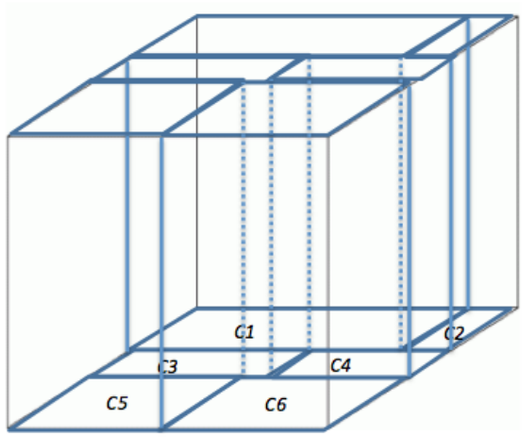



Figure 1: An empty task cube with 6 subtopics.

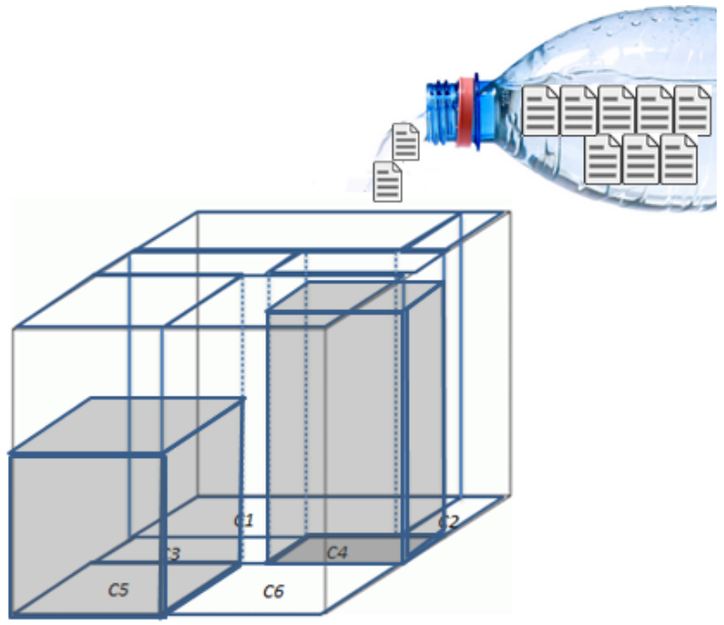



Figure 2: Filling "document water" into the task cube.

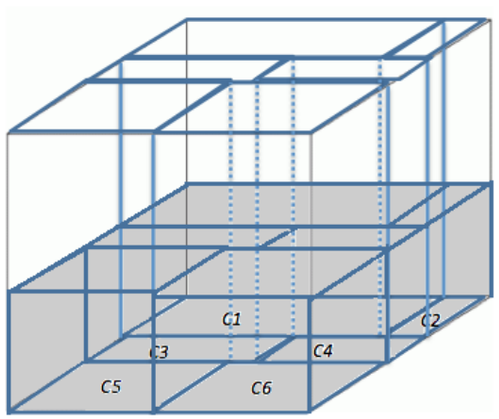



Figure 3: High scoring result.

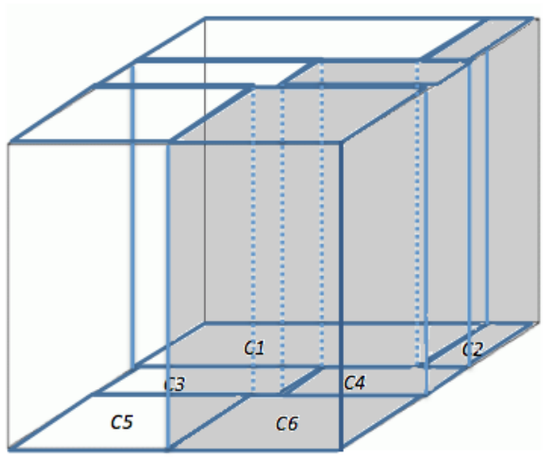



Figure 4: Low scoring result.

# Assignments

- Jeff - Clinical Decision Support (upload Fudan)
- Katherine - Contextual Suggestion
- Tripp - Dynamic Domain
- Pamela - Live QA
- Albert - Task Track
- Bogeum - Total Recall
- Jaime - Real-time summarization
- Jaime - Open Search