Session 4:

Evaluation & future directions

Outline

- Specific **evaluation challenges**: relevance and beyond
- Evaluation campaigns, **collections** and resources
- **Lessons** learnt from evaluation
- Closing **remarks** and open challenges

Specific evaluation challenges in health search

Relevance Assessments

(and beyond)

- Assessing relevance in health search is **demanding** [Koopman&Zuccon, 2014]
  - no correlation b/w length of document and time to judge document
  - Discharge summaries hard to assess
  - highly relevant documents least demanding to judge; somewhat-relevant documents most demanding

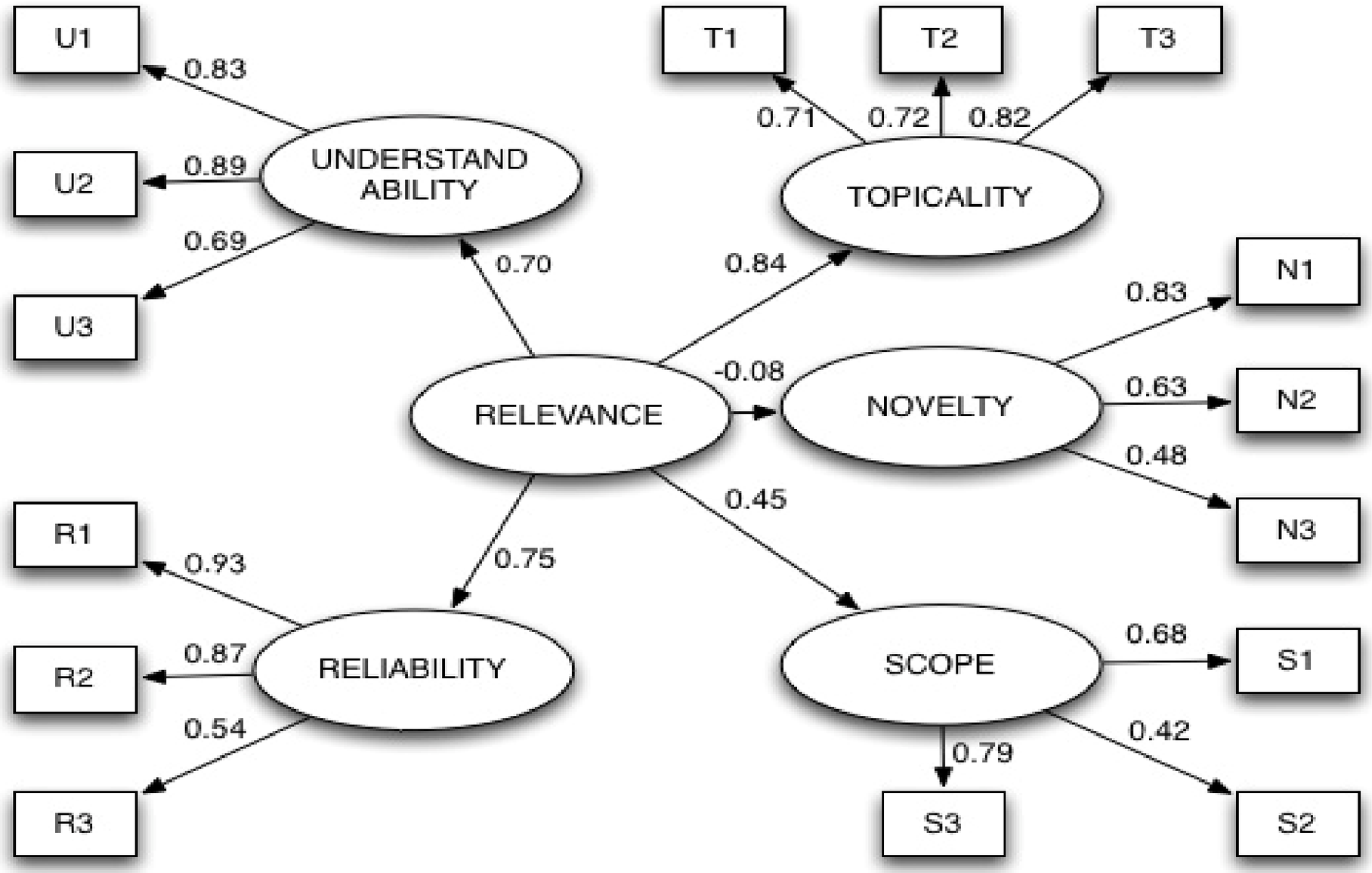
- But **why** is it demanding?
  - vocabulary **mismatch** problem
  - Effect of **temporality** on relevance, *"Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension"*
  - Highly **subjective** *"Patients with hearing loss"*
  - **Dependent aspects** in queries, e.g. *"Patients with complicated GERD who receive endoscopy"*

Expertise and Relevance Assessments

[Palotti et al., 2016 c] + [Tamine&Chouquete, 2017] + [Koopman&Zuccon, 2014]:

- Relevance **agreement low** for both experts and laypeople
- Higher agreement among experts
- medical **expertise** significantly **influences perception** of relevance
- [Tamine&Chouquete, 2017]: `"a single ground truth doesn't exist" -> "variability of system rankings with respect to the level of user's expertise"`

Assessing beyond topical relevance



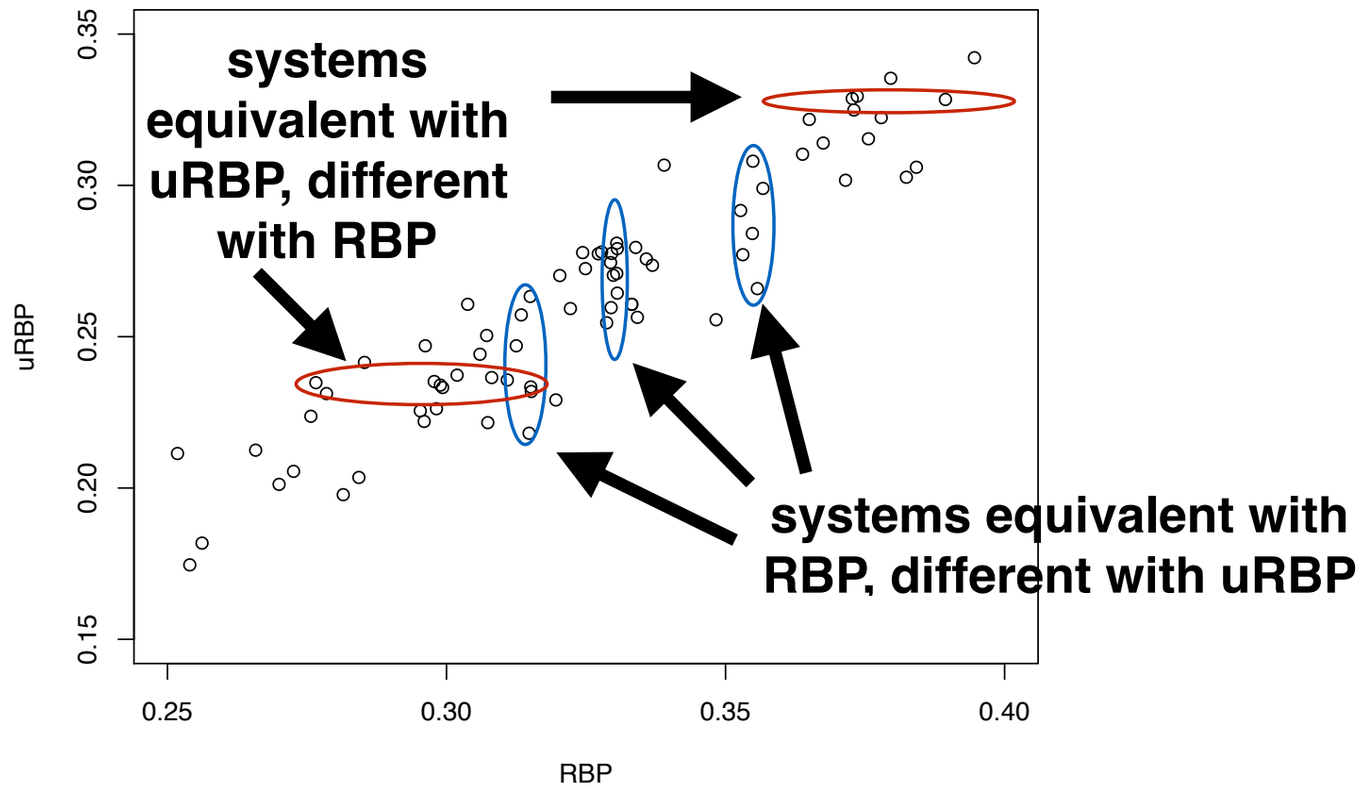
[Zhang et al., 2014]

Integrating Understandability into Gain-Discount Measures

[Zuccon, 2016]

uRB

$$(1-\beta)\sum_{k=1}^{K}\beta^{k-1}P(T|k)P(U|k)$$



- understandability could either be estimated for each document (readability measures as proxy) or computed as a function of understandability label
- framework of evaluation measures that account for dimensions of relevance

Assessing beyond topical relevance

- Integrating **Credibility**: [Lioma et al., 2017]
  - Requires assessments of both relevance and credibility
- **Type I** measures focus on **differences in rank** position of retrieved documents w.r.t. their **ideal rank** (by relevance or credibility).
  - Error based measures
- **Type II** measures operate directly on **document scores**
  - Weighted **cumulative** scores
  - Combination of exisiting evaluation measures (**interpolation**, **harmonic mean**)

Evaluation campaigns, collections and resources

Task	Dataset
Matching patient to clinical trials or trials to patients	1. TREC Medical Records Track [Voorhees&Hersh, 2012] 2. Clinical Trials Test Collection [Koopman&Zuccon, 2016] 3. MIMIC-III: dataset of patient records [Johnson et al., 2016]
Consumer Health Search	1. CLEF eHealth Consumer Health Search Task [Zuccon et al., 2016] 2. FIRE 2016 Consumer Health Information Search
Evidence-based Medicine & Clinical Decision Support (CDS)	1. TREC Genomics Track 2. TREC Clinical Decision Support Track 3. TREC Precision Medicine Track
Compilation of systematic reviews	1. Systematic review test collection [Scells et al., 2017] 2. CLEF eHealth Technology Assisted Review 2017 [Kanoulas et al., 2017]
Image Retrieval	ImageCLEF [Muller et al., 2010]
Identifying concepts from free-text	1. Annotated “problems”, “tests” & “treatments” 2. Annotated SNOMED concept

TREC Genomics

- Run from 2003 to 2007. **Many tasks**, including: ad-hoc, passage retrieval, entity-based QA, text annotation/ categorisation
- Corpus: research articles (e.g. MEDLINE)

Preprocessing&Indexing:
- html -> plain text (tags removal)
- html -> xml (section filtering)
- html -> DB records
- Stemming and stopwords filtering

Query Expansion:
- automated, manual and interactive methods for expansion terms
- Synonyms lookup via UMLS, Entrez Gene, MeSH, HUGO, MetaMAP etc.
- Expansion weighting
- keywords normalisation

Document retrieval:
- tf-idf, BM25, I(n)B2, JelinekMercer smoothing, KLdivergence
- SVM classifiers and an ensemble of standard algorithms

[Hersh&Bhupatiraju, 2003; Hersh, 2005; Hersh et al., 2006]

TREC Genomics

Results are affected by 4 main factors:

1. **Normalization of keywords** in the query into root forms
2. Use of Entrez gene **thesaurus** for **synonymous** look-up

Specific to passage retrieval:

3. **Unit of retrieval** (document, paragraph, subset of paragraphs and a sentence, using these algorithms)
4. Definition of passage

TREC Medical Records

- Run 2011 and 2012.
- Corpus: health records
  - ~93K reports mapped into 17K visits: a patient encounter is made up of one or more reports
  - 9 types of health records
    - ICD coding for each report, plus additional metadata
- Task: identify cohort of patients suitable for specific clinical trials
  - queries: subset of inclusion criteria of trial
  - Some very general, some very specific -> Wide range of number of relevant documents

[Voorhees&Hersh, 2012; Voorhees, 2013]

Example Topics & Documents

Samuel J. Smith

1234567-8

4/5/2006

HISTORY OF PRESENT ILLNESS: Mr. Smith is a 63-year
disease, hypertension, hypercholesterolemia, COPD
well. He did have some more knee pain for a few we
having more trouble with his sinuses. I had starte
He says this has not really helped. Over the past
congestion and thick discharge. No fevers or headа
right-sided teeth pain. He denies any chest pains,
edema or syncope. His breathing is doing fine. No
half-a-pack per day. He plans on trying the patche

CURRENT MEDICATIONS: Updated on CIS. They include
Spiriva, albuterol and will add Singulair today.

ALLERGIES: Sulfa caused a rash.

SOCIAL HISTORY: Smokes as above.

REVIEW OF SYSTEMS: CONSTITUTIONAL: Weight stable. GI: No abdominal pain or change in bowel habits.

PHYSICAL EXAMINATION:

VITAL SIGNS: Weight is 217 lbs, blood pressure 131/61, pulse 63.

HEENT: TMs clear bilaterally, mild maxillary sinus tenderness on the right, nasal mucosa boggy with moderate discharge, teeth in good repair with no erythema or swelling

Topics

136: Children with dental caries

137: Patients with inflammatory disorders receiving TNF-inhibitor treatment

152: Patients with Diabetes exhibiting good Hemoglobin A1c Control (<8.0%)

160: Adults under age 60 undergoing alcohol withdrawal

TREC Clinical Decision Support (CDS)

- Run between 2014 and 2016
  (in 2017 evolved into the Precision Medicine Track)
- Corpus: scientific publications
  - Open Access subset of PubMed Central (PMC); snapshot of ~733K articles in 2014&2015, 1.5M in 2016
- Task: answer clinical questions about health records
  - Queries are very verbose: a summary of the case of a patient
  - 3 types of intents: disease, test, treatment

[Simpson et al, 2014; Roberts et al., 2015]

Example Topics & Documents

Topic	Type	Description
1	Diagnosis	A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease. She currently takes no medications. Physical examination is normal. The EKG shows nonspecific changes.
11	Test	A 40-year-old woman with no past medical history presents to the ER with excruciating pain in her right arm that had started 1 hour prior to her admission. She denies trauma. On examination she is pale and in moderate discomfort, as well as tachypneic and tachycardic. Her body temperature is normal and her blood pressure is 80/60. Her right arm has no discoloration or movement limitation.
21	Treatment	A 21-year-old female is evaluated for progressive arthralgias and malaise. On examination she is found to have alopecia, a rash mainly distributed on the bridge of her nose and her cheeks, a delicate non-palpable purpura on her calves, and swelling and tenderness of her wrists and ankles. Her lab shows normocytic anemia, thrombocytopenia, a 4/4 positive ANA and anti-dsDNA. Her urine is positive for protein and RBC casts.

Sci Rep. 2012;2:685. Epub 2012 Sep 24.

Why large porphyry Cu deposits like high Sr/Y magmas?

Chiaradia M[1], Ulianov A, Kouzmanov K, Beate B.

Author information

Abstract

Porphyry systems supply most copper and significant gold to our economy. Recent studies indicate that they are frequently associated with high Sr/Y magmatic rocks, but the meaning of this association remains elusive. Understanding the association between high Sr/Y magmatic rocks and porphyry-type deposits is essential to develop genetic models that can be used for exploration purposes. Here we present results on a Pleistocene volcano of Ecuador that highlight the behaviour of copper in

TREC Precision Medicine Track

- Run since 2017 (running in 2018)
- Corpus: scientific publications
  - 27M MEDLINE abstracts + 250K clinical trials
- Task: use detailed patient information (genetic information) to identify most effective treatments
  - Focus on oncology
  - Along with the query, comes genetic variants information
  - Primarily needs to identify latest research relevant to patient; otherwise fallback to identify most relevant clinical trials (in case techniques ineffective for patient)

[Roberts et al., 2017]

CLEF eHealth: Consumer Health Search

- Run since 2013 (change name: IR Task, Task 3, Task 2, CHS Task)
- Corpus: web pages
  - 2013-2015: Kreshmoi collection (HON + high quality portals)
  - 2016-2017: Clueweb12b (50M documents)
    - assessments should be used combined for the two years
  - 2018: subset of CommonCrawl: sampled over time via Bing + known reliable&unreliable health websites
- Task: laypeople seeking health advice on the web
  - Many subtasks, including usage of discharge summaries, understandability/ personalisation, query variations, multilingual queries
  - Includes assessments of understandability, trustworthiness

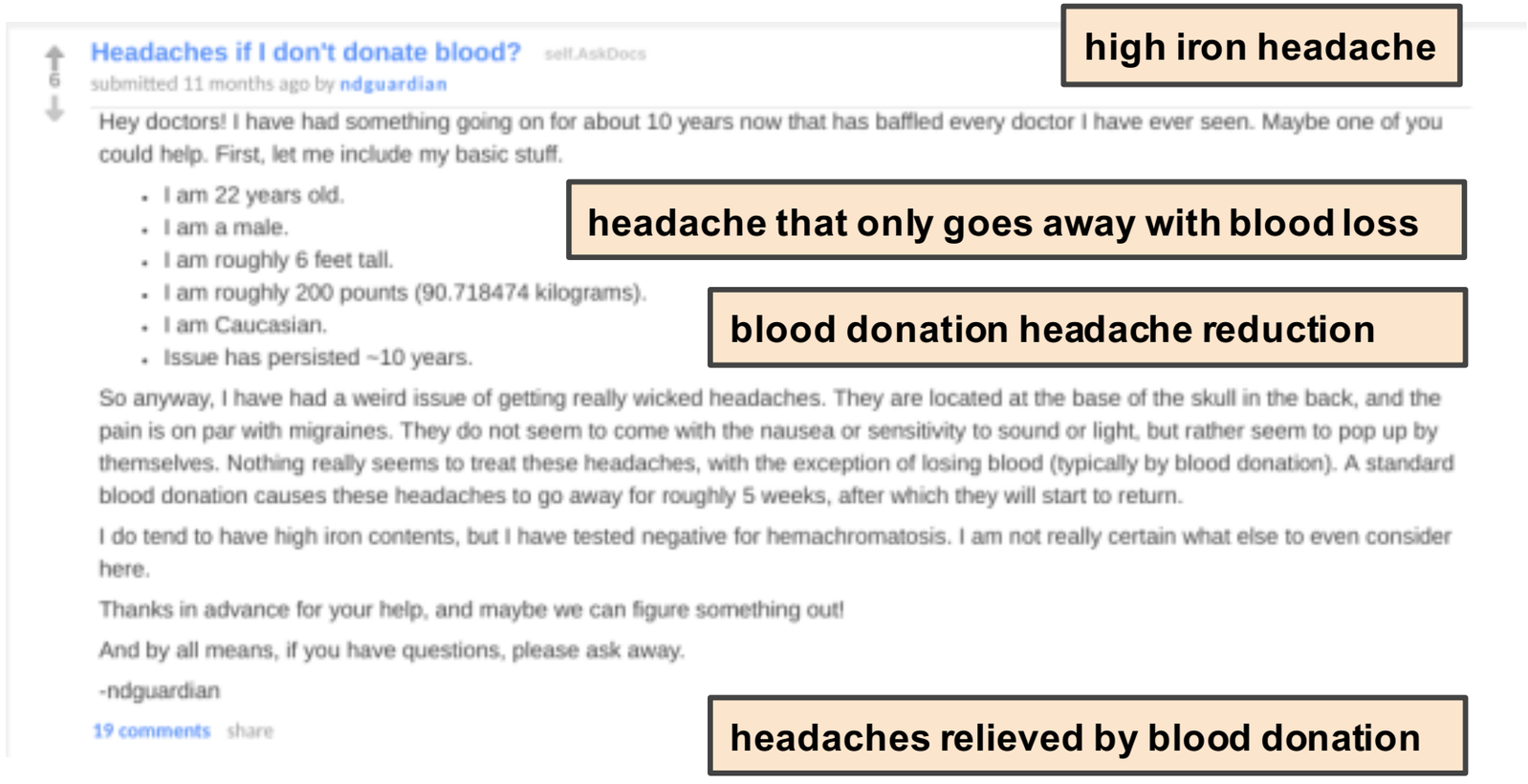
[Zuccon et al., 2016]

The CLEF CHS Queries

- 2013-2014 queries: medical terms extracted from discharge summary (aims to simulate layperson wanting to know more about term)
- 2015: circumlocutory queries sourced via images
- 2016-2017: manually created by external users, via topic description derived from Reddit AskADoctor
- 2018: from HON/TRIP logs

The CLEF CHS Queries: Query Variations

- 2016/2017 (Reddit): 6 variations for each information need (6x50=300)



- Query variations also in 2015 & 2018, but sourced differently

CLEF eHealth: Technology Assisted Review

- Run since 2017
- Corpus: MEDLINE abstracts
- Task: efficient and effective ranking of articles during screening phase (abstract level) of conducting Diagnostic Test Accuracy systematic reviews
    1. ranking: rank all abstracts; goal: retrieve relevant abstracts as early as possible,
    2. thresholding: identify relevant subset of abstracts to be shown, i.e. rank at which to stop in the result list
- Topics: 50 (20 dev + 30 test) reviews
    - Topic, Title, Boolean Query, and PMID (documents to rank)
- Relevance assessments at (a) abstract, (b) document level

[Kanoulas et al., 2017]

CLEF TAR Topic File

```
Topic: CD009551
Title: Polymerase chain reaction blood tests for the diagnosis of
       invasive aspergillosis in immunocompromised people

Query:
exp Aspergillosis/
exp Pulmonary Aspergillosis/
exp Aspergillus/
(aspergillosis or aspergillus or aspergilloma or "A.fumigatus" or
"A. flavus" or "A. clavatus" or "A. terreus" or "A. niger").ti,ab.
or/1-4
exp Nucleic Acid Amplification Techniques/
pcr.ti,ab.
"polymerase chain reaction*".ti,ab.
or/6-8
5 and 9
exp Animals/ not Humans/
10 not 11

Pmid's:
    25815649
    26065322
    ...
```

Title of the Systematic Review

Boolean query in Ovid format

Articles retrieved by the boolean query

Other Health Evaluation Campaigns: ImageCLEF, NTCIR, FIRE

- **NTCIR** medical natural language processing evaluation
  - 2014-2016: information extraction from health records in Japanese
  - 2017: multilingual disease name extraction from tweets and articles (Chinese, English, Japanese)
- **FIRE** 2016 Consumer Health Information Search (CHIS)
  - Task A: classify relevance of sentences in documents
  - Task B: identify whether relevant sentences support or reject claim made in the query
- **ImageCLEF** medical retrieval 2003-2018
  - Many subtasks, both CBIR and TBIR: adhoc retrieval, case-based retrieval, image annotation, modality detection, caption prediction, etc

Other collections, not associated to campaigns

- **Clinical Trial Retrieval** [Koopman&Zuccon, 2016]
  - ~200K clinical trials from ClinicalTrials.gov
  - 60 topics: descriptions of patient cases (from TREC CDS)
  - Relevance assessments w.r.t. referring the patient to the trial + expected number of trials
    - Support for INST evaluation measure
- **Assisting Systematic Reviews** [Scells et al., 2017]
  - ~26M MEDLINE research studies
  - 94 reviews (query topics) extracted from Cochrane + assessments
  - Tasks supported (+specific evaluation measures):
    (1) retrieval for screening; (2) screening prioritisation; (3) stopping point

Good lessons from evaluation campaigns

- **Retrieval of health records for cohort selection** (TREC Medical Records [Edinger et al., 2012])
  - Both **precision** and **recall errors** due to **incorrect lexical representations and lexical mismatches**
    - Non-relevant visits were most often retrieved because they contained a non-relevant reference to the topic terms
    - Relevant visits were most often infrequently retrieved because they used a synonym for a topic term
- Other issues: time factors, negation detection, overlap in terminology between conditions or procedures (hearing loss vs hearing aid)

Good lessons from evaluation campaigns

- **Retrieval of evidence based medicine**
  (TREC CDS [Roberts et al., 2016], analysing 2014 results)

  - How to best to use **concept extraction** system such as MetaMap of key importance: can easily become a red herring

  - **Negation and attribute extraction** (age, gender, etc.) intuitively important, but best systems did not use them
    If negation extraction, soft-matching strategy best

  - **article preference** to identify appropriate articles for Diagnosis, Treatment, and Test (fundamental mismatch b/w irrelevant articles and clinical important attributes)

    - Methods tried did not work: specialised lexicons, MeSH terms, and machine learning classifiers

Good lessons from evaluation campaigns

[Karimi et al., 2018] provides **platform** to facilitate experimentation and hypothesis testing

- Can tease-out which components provide improvements
  - query and document expansion (UMLS), word embeddings, negation detection/removal, LTR
- Main findings on TREC CDS
  - **Articles body** contributes to retrieving over 50% of relevant results
  - adding UMLS concepts does not improve retrieval using titles only
  - concepts in abstracts slightly improved retrieval for queries built using Desc and Sum, but not Note
  - **PRF** works well, also in combination with **word embeddings**; but **LTR** can outperform all these

Closing remarks

Discussion

- What have we learned?
- What open challenges?

Open challenges

- Ethics and sharing of data — privacy concerns vs need for large scale evaluation
- Integration of data driven and symbolic representations
- Inference with knowledge graphs
- Query understanding
- Results presentation

} require personalisation, context understanding, better user understanding

- Translation of IR for impact on health

Where to go for help?

- Content from this tutorial: https://ielab.github.io/health-search-tutorial/
- Bibliography of all literature mentioned here
- Docker image - https://hub.docker.com/r/ielabgroup/health-search-tutorial
- Hersh's book: "Information Retrieval: A Health and Biomedical Perspective"

Thanks for attending!

Guido Zuccon

@guidozuc

Bevan Koopman

@bevan_koopman

THE END