Health Search

From Consumers to Clinicians

Slides available at

https://ielab.io/russir2018-health-search-tutorial/

Guido Zuccon
Queensland University of Technology

@guidozuc

Knowledge based vs data-driven Query Expansion

Knowledge based query expansion

Inference

Concept relationships

Subsumption

Corpus/Data Driven

Multi-evidence

Co-occurences, Latent methods & Word2vec

Combine documents that refer to the same case [Zhu&Carterette, 2012; Limsopatham et al., 2013b]

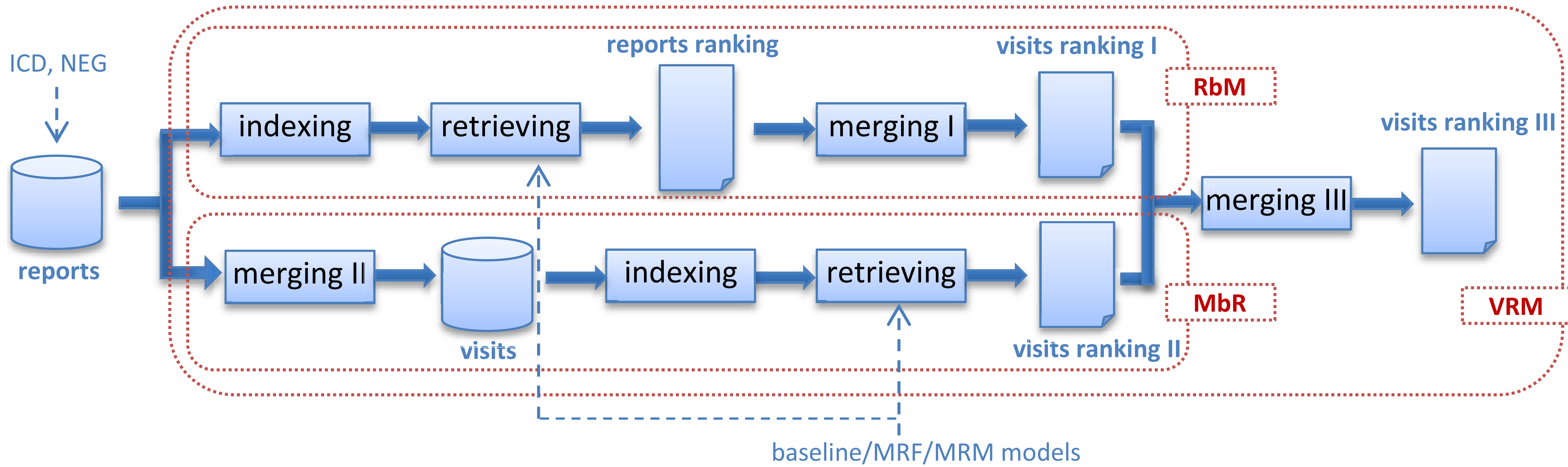
Different, diverse corpora used for query expansion [Zhu&Carterette, 2012 b; Zhu et al., 2014]

Measure the usefulness of different collections [Limsopatham et al., 2015]

…

Combine multiple-evidences in the collection that refer to the same case

[Zhu&Carterette, 2012]



- Ranking generated for each document, individually
- Ranking generated for an aggregated case

} Fused into new ranking

- Online possible in situations where multiple documents are available for one case (e.g. with health records, where case=patient)

Adaptively Combine (or not) Records of a Case

[Limsopatham et al., 2013b]

- Choose between:
  1. Combine records for a patient, then rank patient
  2. Rank records, then identify patients based on relevance of records ranking
- Classifier to **learn to select which ranking approach** to use, depending on query
- Features: query difficulty measures (QPPs), number of medical concepts in query

Different, diverse corpora used for query expansion

[Zhu et al., 2014]

- Mixture of relevance models to **combine evidence from different collections** to derive **query expansions**
  - Collections: Mayo Clinic health records (39M), TREC Genomics (166K), ClueWeb09B (44M), TREC Medical Records (100K)
- Findings:
  - Access to **large clinical corpus** significantly improves query expansion
  - The more difficult the query, the more it benefit expansion benefits from auxiliary collections
  - “use **all available data**" is **sub-optimal**: value in collection curation

Measure the usefulness of different collections

[Limsopatham et al., 2015]

- Automatically decide which collection to use for query expansion evidence
  - 14 different document collections, from domain-specific (e.g. MEDLINE abstracts) to generic (e.g. blogs and webpages)
    - But they are **not all useful**, and **not to the same extent** to generate query expansion terms
- Techniques based on resource selection and learning to rank

Co-occurences, Latent Methods & Word2vec

- (Co-occurence of) concepts as a **graph** -> application of **link analysis** methods [Koopman et al., 2012; Martinez et al., 2014]
- **Explicit** and **latent concepts** [Balaneshin-kordan&Kotov, 2016]
- Word **embeddings** and concept embeddings [Zuccon et al., 2015, b; Nguyen et al., 2017]

Co-occurence Graphs, Semantic Graphs and Page Rank

- [Koopman et al., 2012]:
    1. Build concept **graph** from **document concepts** as they **co-occur** in document
    2. Run Pagerank
    3. Use Pagerank scores as additional weights for retrieval
- [Martinez et al., 2014]:
    1. Build concept **graph** from **query concepts** and **related concepts** in **UMLS**
    2. Run Pagerank
    3. Rank concepts using page rank scores; select top K concepts as query expansion
- Analysis shows expansion terms selected by Pagerank: taxonomic (eg., synonyms) and not taxonomic (eg., disease has associated anatomic site).

Explicit and Latent Concepts

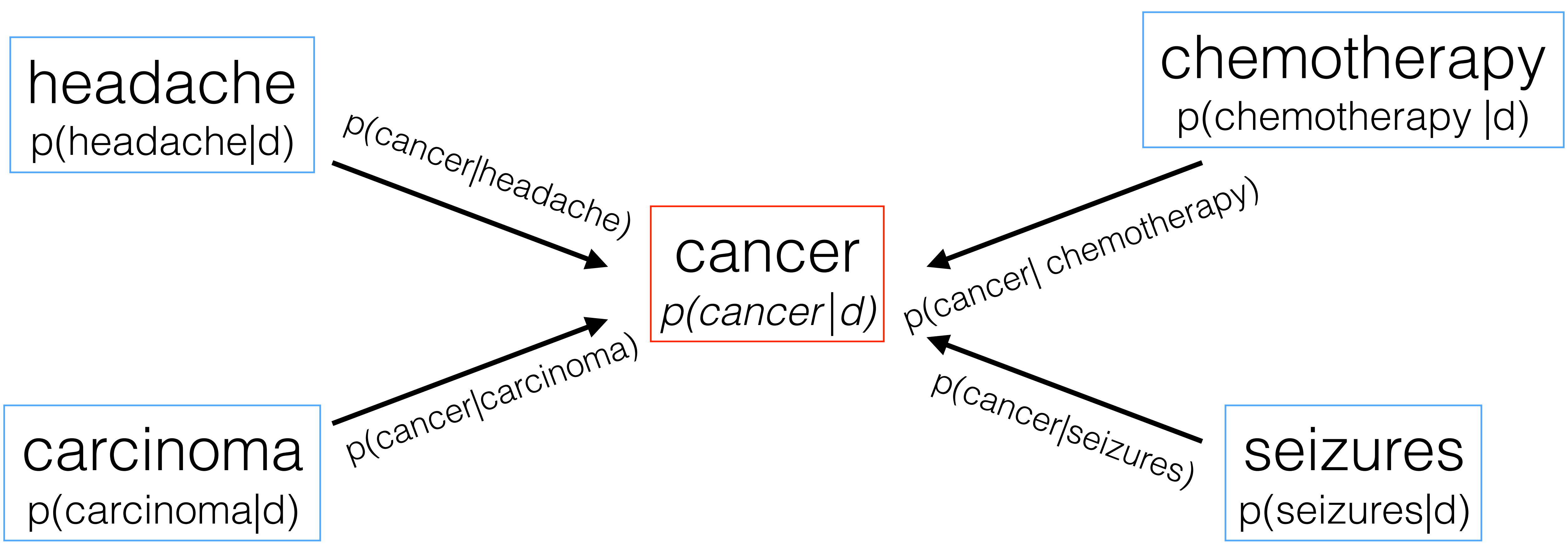
- [Balaneshin-kordan&Kotov, 2016]: different **concept types/ sources** (KBs, PRF) should have **different weights**
- Builds upon Markov Random Field retrieval [Metzler&Croft, 2005]
- Different **features** for different **semantic types** + **topical features** of KB graphs, and **statistics** of concepts in collection
- Learns optimal query concept weight using multivariate optimisation
- Base approach (without optimisation) best system at TREC CDS 2015

Word Embeddings and Concept Embeddings: Neural Translation LM

[Zuccon et al., 2015, b]

$$p_t(w|d) = \sum_{u \in d} \boxed{p_t(w|u)}\, p(u|d)$$

use Word Embeddings for computing this

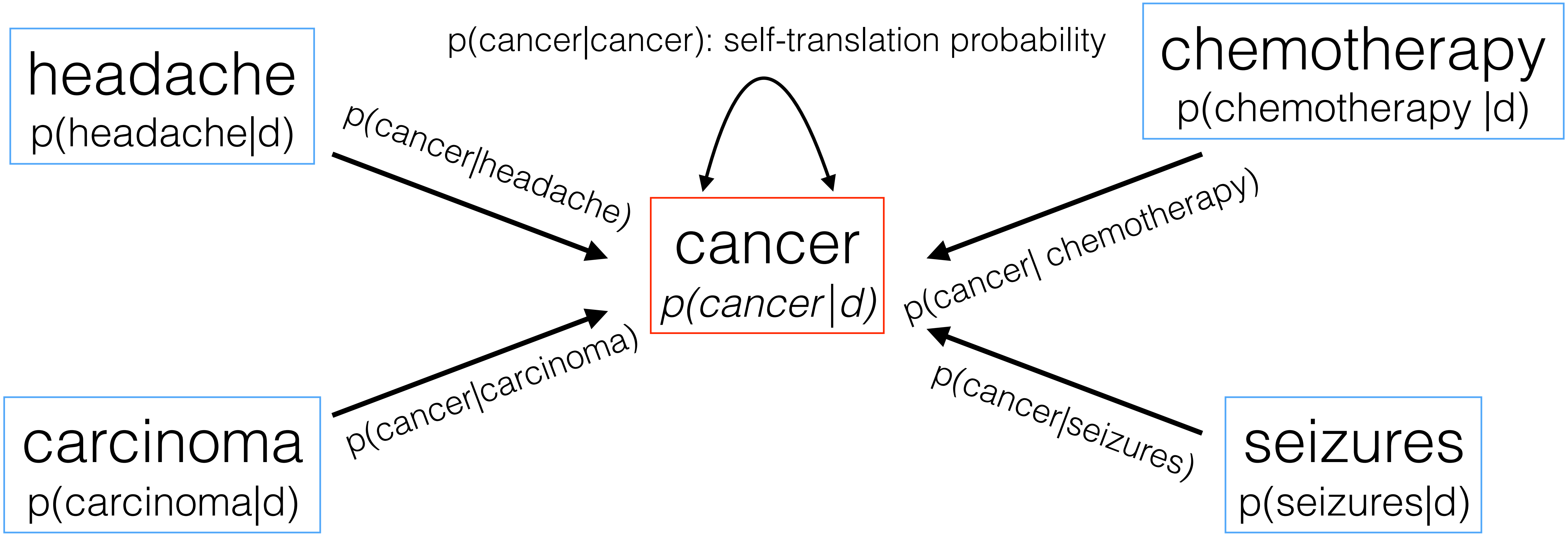


Word Embeddings and Concept Embeddings: Neural Translation LM

[Zuccon et al., 2015, b]

$$p_t(w|d) = \sum_{u \in d} \boxed{p_t(w|u)}\, p(u|d)$$

use Word Embeddings for computing this



Skipped

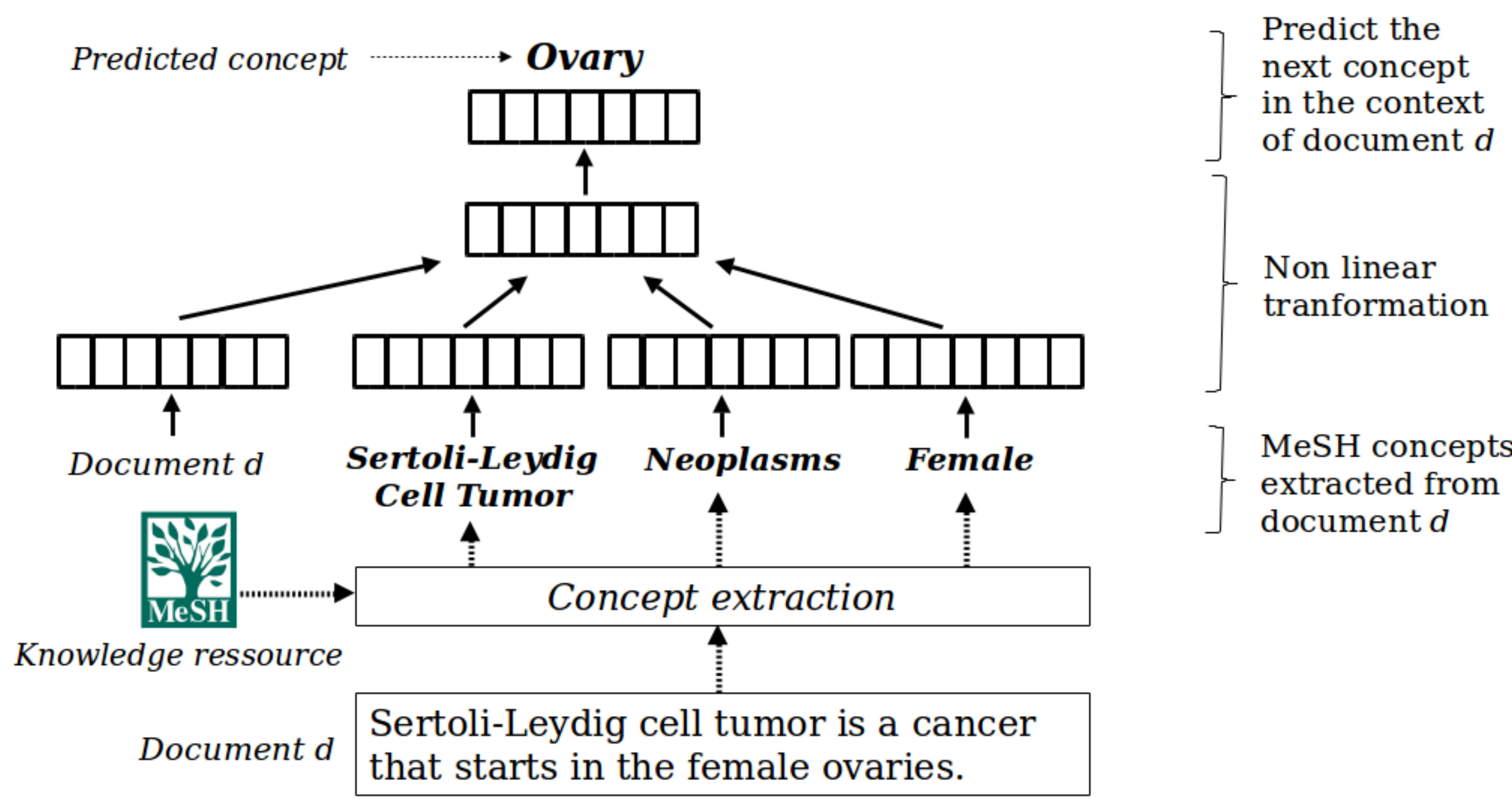
Constraining word embeddings by prior knowledge

- [Liu et al., 2016]: learn concept embeddings **constrained by relations** in KB (UMLS)
- Results in a modified CBOW
- Use word embeddings to **re-rank** results: **interpolate** original relevance score with similarity based on embeddings
- Experiments only limited to synonym relations & single-word concepts

Concept-Driven Medical Document Embeddings

[Nguyen et al., 2017]: optimises document representation for medical content

- Uses neural-based approach (akin to *doc2vec*) to create embedding that captures latent relations from concepts and terms in text.
- Uses embedding to identify top documents
- Extract top words and concepts from top documents to produce expansions



Skipped

Learning to Rank

[Soldaini&Goharian, 2017]: compares 5 LTR in CHS context:

- LTR: logistic regression, random forests, LambdaMART, AdaRank, ListNet
- Features: statistical (36 features), statistical health (9), UMLS (26), latent semantic analysis (2), word embeddings (4).
- **LambdaMART** performed best; **all features** required

Dealing with the nuances of medical language

Negation & Family History

“denies fever”
“no fracture”

“mother had breast cancer”

Negation & Family History

`“denies fever”`
`“no fracture”`

`“mother had breast cancer”`

NegEx/ConText [Harkema et al., 2009]:
Algorithm for extracting negated content

Negation & Family History

“denies fever”
“no fracture”

“mother had breast cancer”

NegEx/ConText [Harkema et al., 2009]:
Algorithm for extracting negated content

- Negated content best handled by:
  - **Not removing** negated content (as is commonly done)
  - **Indexing** positive, negated & family history content **separately** [Limsopatham et al., 2012]
  - **Weighting** content **separately** [Koopman & Zuccon, 2014]

PICO

- PICO: framework for formulating clinical questions
  P: Patient/Problem (P) (e.g., males aged 20-50)
  I: Intervention (e.g., weight loss drug)
  C: Comparison (e.g., controlled exercise regime)
  O: Outcome (e.g., weight loss)

- Exploiting PICO elements in IR:
  - Language modelling based **content weighting** [Boudin et al., 2010]
    - Tagging PICO elements for IR - "I" & "P" elements most beneficial for retrieval
  - **Field retrieval** based on PICO [Scells et al., 2017b]
    - promising, but needs method to predict which keywords require PICO annotations

RobotReviewer [Marshall et al., 2015]: Algorithm for extracting PICO elements from free-text

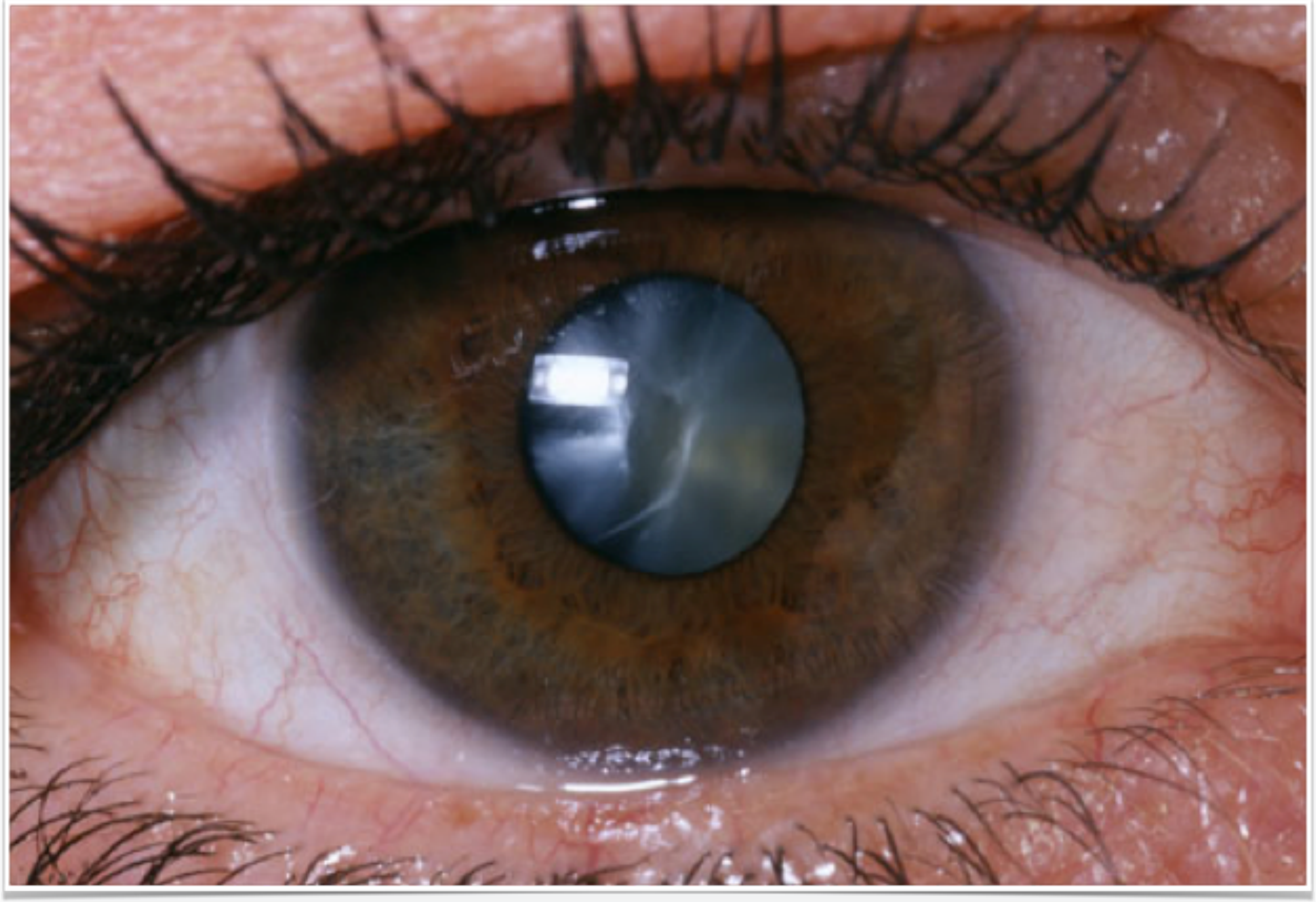
Readability & Understandability

- **Laypeople** do not necessarily **understand medical** documents that clinicians would understand
- Need to retrieve documents that are both understandable and relevant
- [Palotti et al., 2016 b]: LTR with two sets of features:
    - Estimate relevance: standard IR features
    - Estimate understandability: features based on **readability measures** and **medical lexical aspects**

Understanding and aiding query formulation

Skipped

What would **search** for?



Enter your search terms at http://chs.ielab.webfactional.com/

Skipped

"Circumlocutory" queries

Symptom Group	Crowdsourced Circumlocutory Queries
alopecia	baldness in multiple spots, circular bald spots, loss of hair on scalp in an inch width round
angular cheilitis	broken lips, dry cracked lips, lip sores, sores around mouth
edema	fluid in leg, puffy sore calf, swollen legs
exophthalmos	bulging eye, eye balls coming out, swollen eye, swollen eye balls
hematoma	hand turned dark blue, neck hematoma, large purple bruise on arm
jaundice	yellow eyes, eye illness, white part of the eye turned green
psoriasis	red dry skin, dry irritated skin on scalp, silvery-white scalp + inner ear
urticaria	hives all over body, skin rash on chest, extreme red rash on arm

[Stanton et al., 2014]

How effective are Google & Bing at Health Search?

System	ndcg@1		ndcg@5		ndcg@10		P@5		P@10	
	Rel	Hrel	Rel	Hrel	Rel	Hrel	Rel	Hrel	Rel	Hrel
Bing	.3846	.2308	.3812	.2654	.3802	.2764	.4385	.2769	.4308	.2769
Google	.3846	.3077	.4242	.3142	.4252	.3138	.5000	.3154	.4923	.3115

Effectiveness of two widely used commercial search engines when prompted with (circumlocutory) medical queries aimed at self-diagnosis purposes. Results are averaged over 26 queries. P@k stands for precision at rank k; ndcg@k stands for normalised discounted cumulative gain at rank k.

[Zuccon et al., 2015]

How effective are Google & Bing at Health Search?

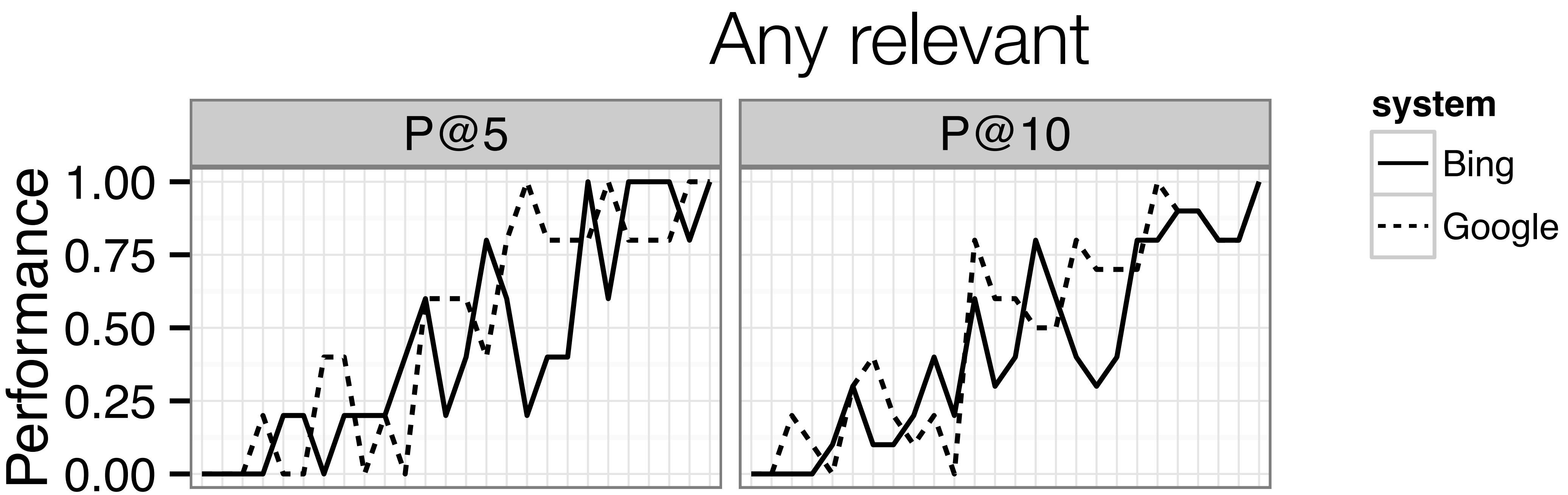
System	ndcg@1		ndcg@5		ndcg@10		P@5		P@10	
	Rel	Hrel	Rel	Hrel	Rel	Hrel	Rel	Hrel	Rel	Hrel
Bing	.3846	.2308	.3812	.2654	.3802	.2764	.4385	.2769	.4308	.2769
Google	.3846	.3077	.4242	.3142	.4252	.3138	.5000	.3154	.4923	.3115

Effectiveness of two widely used commercial search engines when prompted with (circumlocutory) medical queries aimed at self-diagnosis purposes. Results are averaged over 26 queries. P@k stands for precision at rank k; ndcg@k stands for normalised discounted cumulative gain at rank k.

[Zuccon et al., 2015]

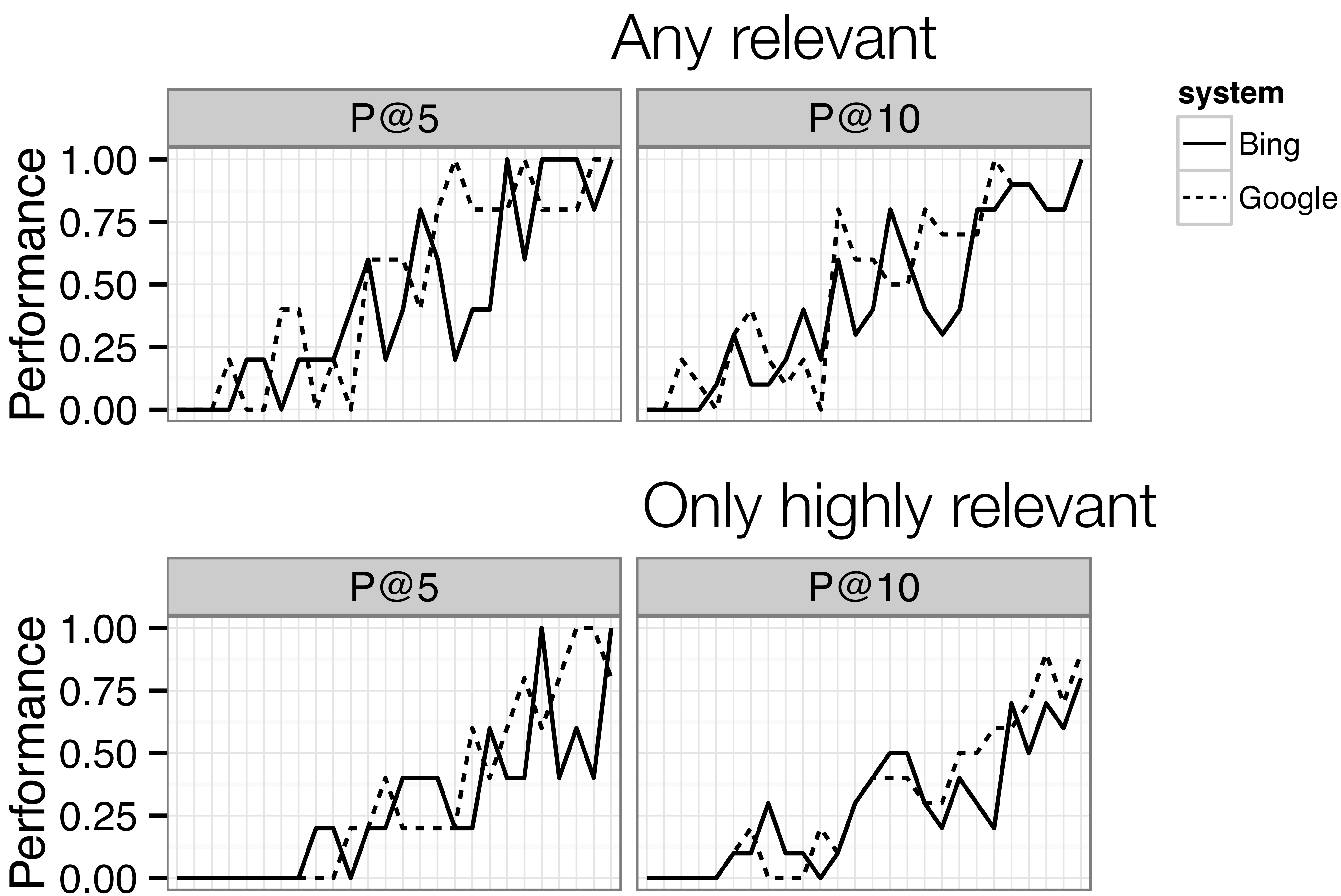
Skipped

Performance per query



[Zuccon et al., 2015]

Performance per query

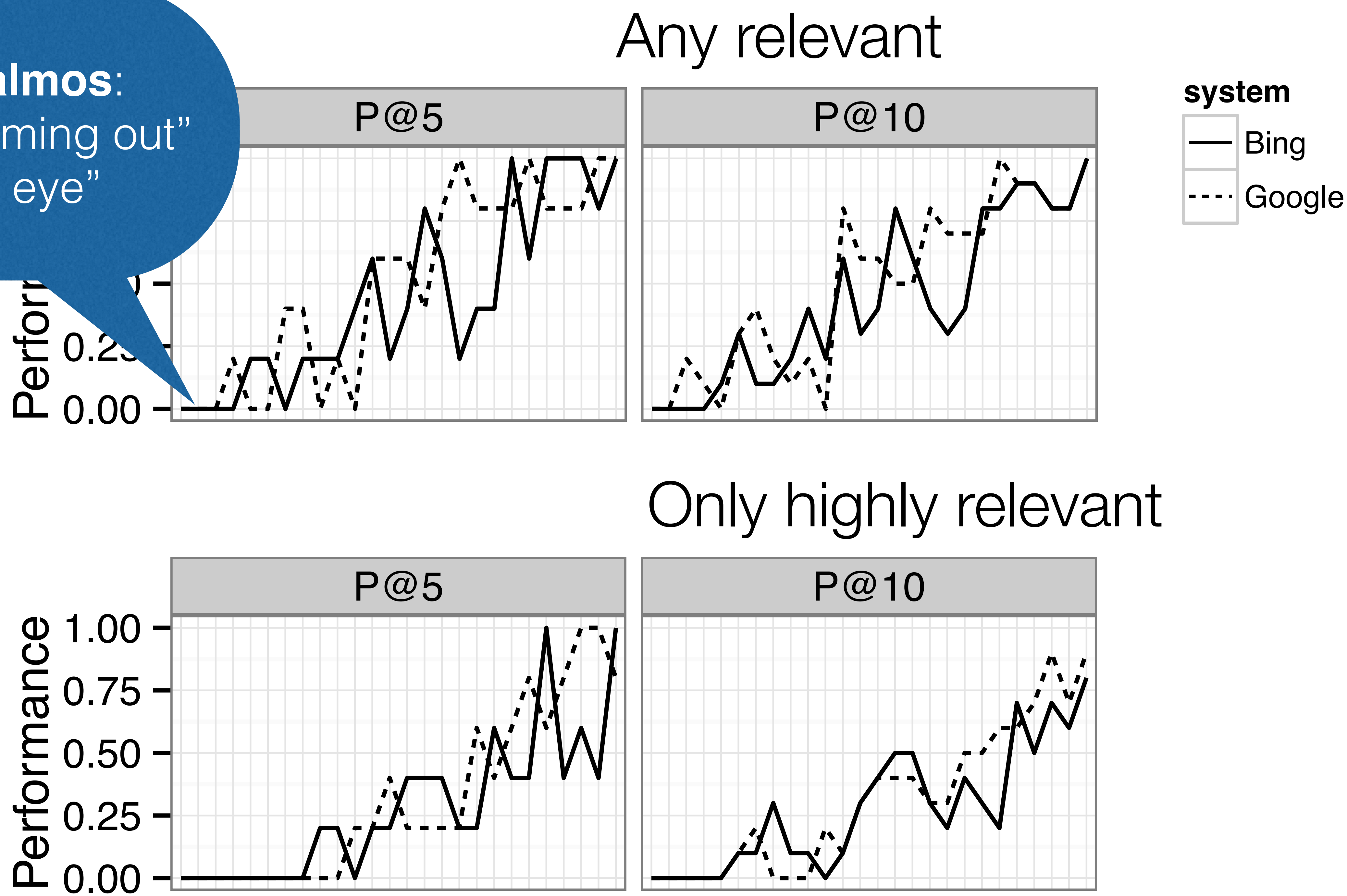


[Zuccon et al., 2015]

Skipped

Performance per query

exophthalmos:
"eye balls coming out"
"swollen eye"



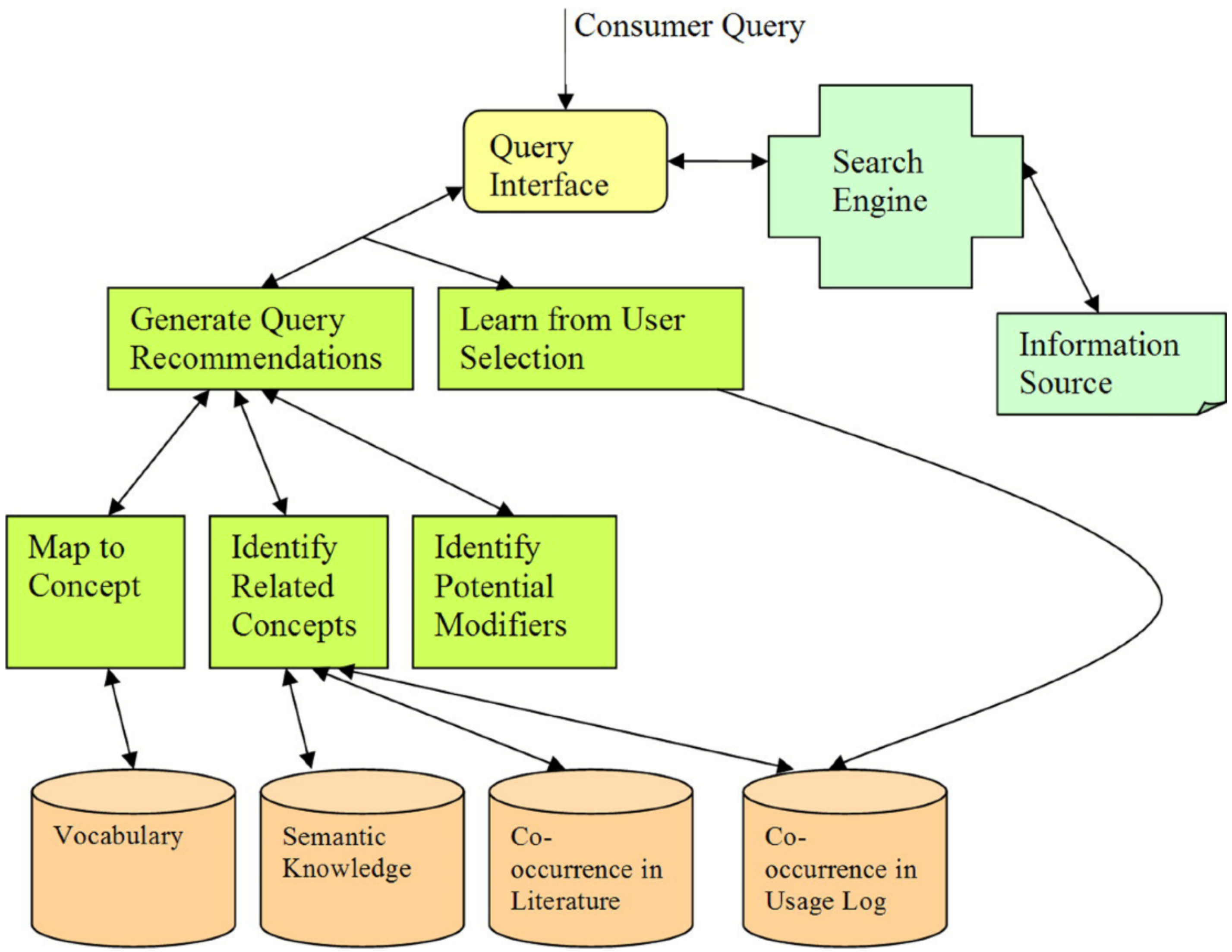
[Zuccon et al., 2015]

Skipped

Query Recommendation

[Zeng et al, 2006]: recommend queries based on UMLS and query log (CHS task)

- Leads to higher user satisfaction and query success rate



Query Reformulation

[Soldaini et al., 2015]: compares the effectiveness of 7 query reformulation techniques (CDS task)

1. **UMLS Concepts Selection** (MMselect): remove all terms with no mapping to any UMLS concepts
2. **Health-related terms selection** (HT): compute ratio of associated Wikipedia page P being health-related over being not-health-related. Retain only query terms with ratio ≥ 2.
3. **Query Quality Predictors** (QQP): use QPPs as features of SVMrank to select query terms.
4. **Faster QQP**: rank sub-queries using MI and retains the top 50. In addition to QQP features, add features: UMLS concepts found, UMLS sem-types found, HT ratio, and MeSH found.

Query Reformulation

[Soldaini et al., 2015]: compares the effectiveness of 7 query reformulation techniques (CDS task)

5. **UMLS Concepts Extraction** (MMexpand): append the preferred terms UMLS query concepts to expand original query
6. **Pseudo Relevance Feedback** (PRF): weight terms in top 10 initial results, rank and add top 20 terms not in original query.
7. **Health Terms PRF** (HT-PRF): as PRF, but candidate expansion terms filtered health term ratio
   - This is empirically identified as the best technique
   - The HT component in general seems effective

Query Reformulation with deep learning

[Soldaini et al., 2017]: considers short clinical notes as queries (CDS task)

1. Generate candidate terms using PRF
2. Train supervised neural network to predict Weight Relevance Ratio (WRR) of candidate terms: importance of term in relevant documents
3. For representations it uses word embeddings, statistical features over multiple collections, syntactical and semantical features

- The neural network approach and HT-PRF perform similarly

Query Clarification

[Soldaini et al., 2016]: **add** the most appropriate **expert expression** to queries submitted by users

- Acquire expert expressions **from 3 KBs**: behavioral (logs), MedSyn, and DBpedia
- Select expression with the highest probabilities of appearing in health-related Wikipedia pages, using logistic regression classifier
- Finding through user study evaluation (CHS task):
    - Expressions from all 3 KBs **improve rate of correct answers** (behavioural KB best)
    - number of **correct** answers significantly increases when users clicked **HON-certified websites**

Skipped

Query Reduction

- [Koopman et al., 2017 c]: reduce verbose clinical queries (health records, CDS task) using generic & domain-specific methods
- Reduce to only UMLS Medical Concepts & Tasked UMLS
- Combined model UMLS + IDF-r (proportion of top-ranked IDF terms retained)
- Comparison vs human-generated queries: human generated queries significantly more effective
    - per-query parameter learning promising
    - automated reduction handicapped in that they only use terms from narrative

Skipped

Query Reduction

[Soldaini et al., 2017 b]: use convolutional neural networks (CNN) to reduce queries (CDS task)

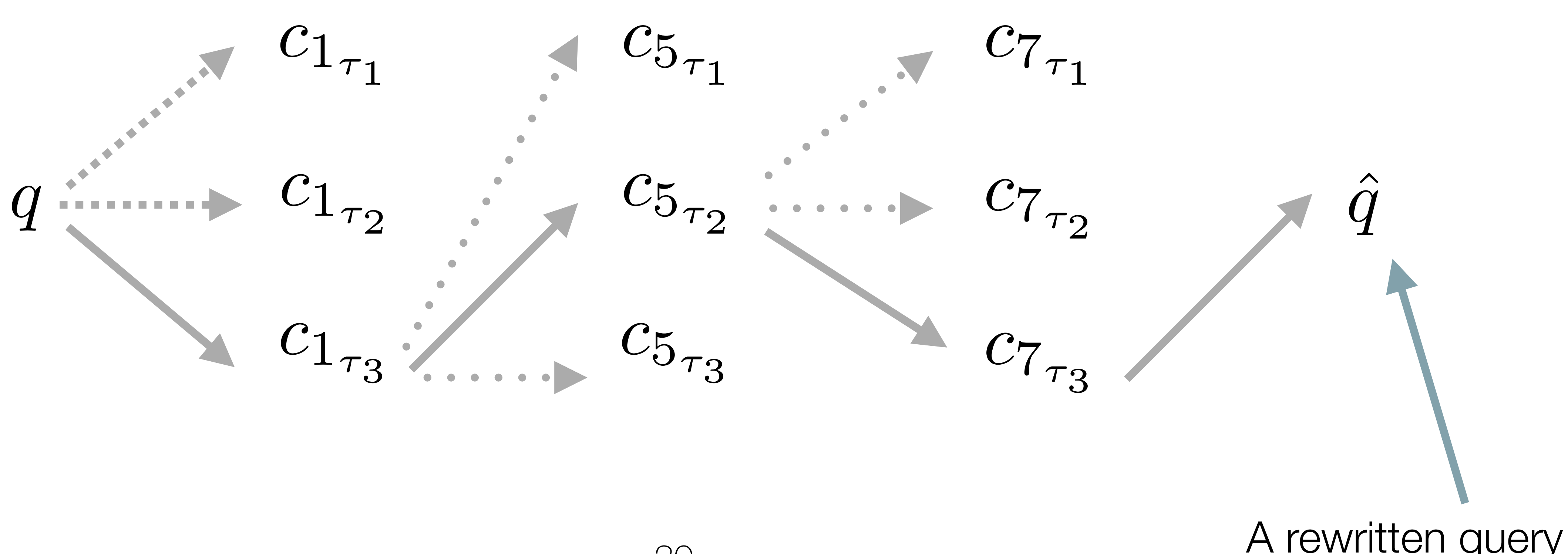
- Queries are short clinical notes
- CNN is used to estimate the importance of each query term
- Given a query, a relevant document and a non-relevant document:

1. Use CNN to determine weights terms in query
2. Use term weights to score relevant and non-relevant documents
3. Back-propagate a loss if non-relevant document is scored higher than relevant document

Query Rewriting

[Scells&Zuccon, 2018]: through a chain of transformation, generates better (Boolean) queries (for systematic reviews compilation)

- Defines set of transformations: mostly syntactic transformations
- Selects transformations based on: heuristics, classifier, learning to rank
- Large gains possible by transforming queries



Skipped

Query Difficulty

- [Boudin et al., 2012]: predictor that **exploits MeSH structure** to ascertain how difficult queries are — estimates query variability and specificity

$$\mathit{MeSH\text{-}QD}(Q,\mathcal{T}) = \sum_{t \in Q} \overbrace{\frac{df(t)}{\sum_{t' \in V(t)} df(t')} \cdot ln\left(1 + \frac{N}{df(t)}\right)}^{\text{term variability}} \cdot \overbrace{\frac{\text{depth}(t)}{\text{length}(t)}}^{\text{term generality}}$$

  - V(t): set of alternative expressions of the concept t; depth/length in MeSH
  - **coverage** of thesaurus & concept **mapping** influence quality
- [Scells et al., 2018]: **standard predictors** for QPP and QVPP (V=variation) in systematic reviews compilation
  - Predictors **not suited** to the domain-specific nature of the task
  - Identifying best performing **variations hard task**

Skipped

Task based retrieval

- Research on how clinicians' query shows a set of standard query types [Ely et al., 2000]
- Can be simplified to three clinical tasks:
  1. searching for **diagnoses** given a list of symptoms;
  2. searching for relevant **tests** given a patient's situation
  3. searching for effective **treatments** given a particular condition.
- These can be exploited in a retrieval scenario…

Skipped

Tasked-based retrieval

- Concept-based approach but “focusing only on medical concepts essential for the information need of a medical search task” [Limsopatham et al., 2013]
- Tasked-oriented filtering, visualisation and retrieval [Koopman et al., 2017 b]

Skipped

Tasked-based retrieval

- Concept-based approach but "focusing only on medical concepts essential for the information need of a medical search task" [Limsopatham et al., 2013]

- Tasked-oriented filtering, vi
[Koopman et al., 2017 b]

MetaMap's Semantic Type	Aspects of the Medical Decision Criteria			
	Symptom	Diagnostic test	Diagnosis	Treatment
Body Location or Region	✔	✔	✔	✔
Body Part, Organ, or Organ Component	✔	✔	✔	✔
Clinical Drug	–	–	–	✔
Diagnostic Procedure	–	✔	–	–
Disease or Syndrome	–	–	✔	–
Finding	✔	–	–	–
Health Care Activity	–	✔	–	✔
Injury or Poisoning	✔	–	–	–
Intellectual Product	–	✔	–	✔
Medical Device	–	✔	–	✔
Mental or Behavioral Dysfunction	✔	–	✔	–
Neoplastic Process	✔	✔	✔	✔
Pathologic Function	✔	–	–	–
Pharmacologic Substance	–	–	–	✔
Sign or Symptom	✔	–	–	–
Therapeutic or Preventive Procedure	–	–	–	✔

Skipped

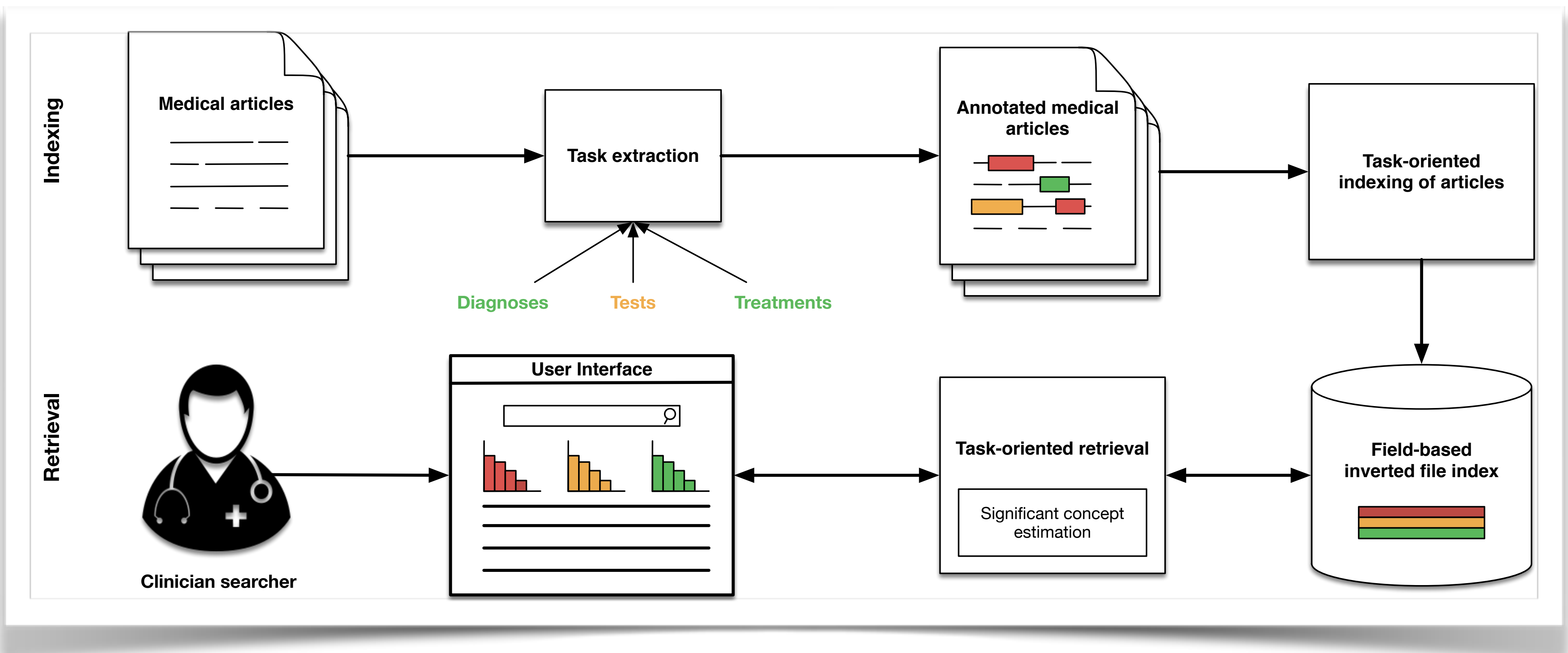
Tasked-based retrieval

- Concept-based approach but “focusing only on medical concepts essential for the information need of a medical search task” [Limsopatham et al., 2013]
- Tasked-oriented filtering, visualisation and retrieval [Koopman et al., 2017 b]

Skipped

Tasked-based retrieval

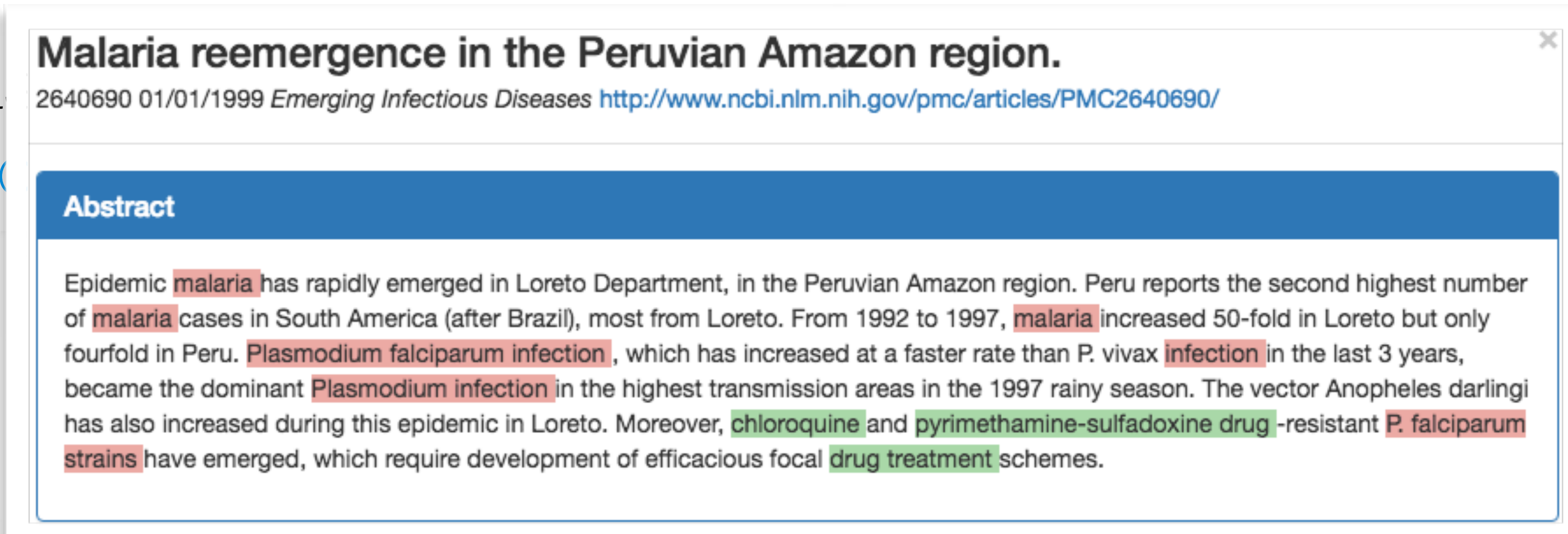
- Concept-based approach but "focusing only on medical concepts essential for the information need of a medical search task" [Limsopatham et al., 2013]
- Tasked-oriented filtering, visualisation and retrieval [Koopman et al., 2017 b]



Skipped

Tasked-based retrieval

- Concept-based approach but "focusing only on medical concepts essential for the information need of a medical search task" [Limsopatham et al., 2013]
- Ta
[K

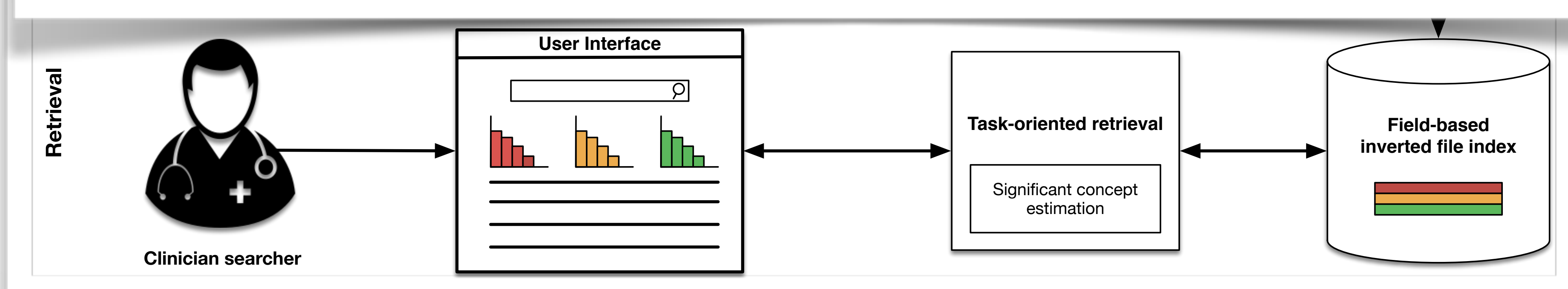


Malaria reemergence in the Peruvian Amazon region.

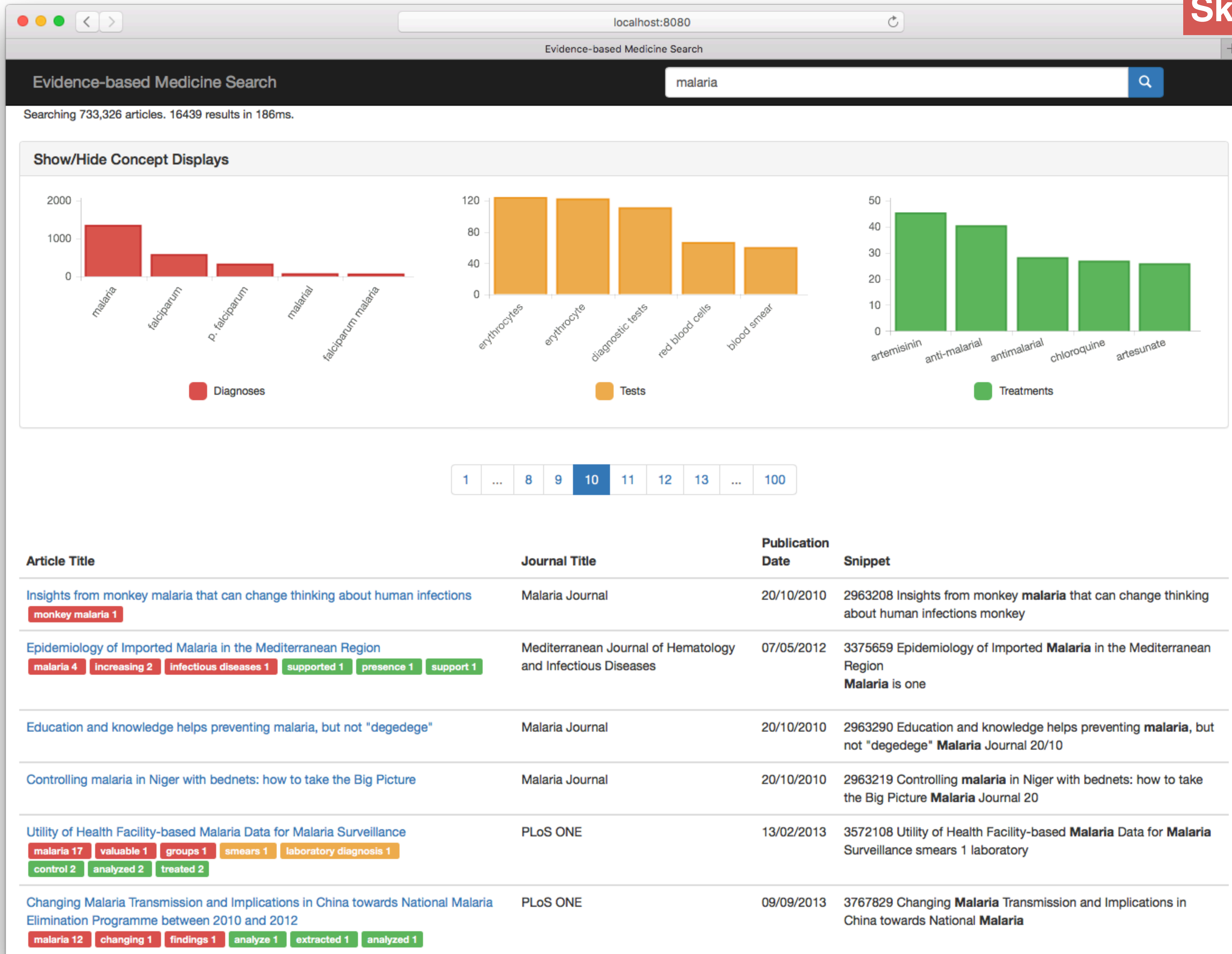
2640690 01/01/1999 *Emerging Infectious Diseases* http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2640690/

Abstract

Epidemic malaria has rapidly emerged in Loreto Department, in the Peruvian Amazon region. Peru reports the second highest number of malaria cases in South America (after Brazil), most from Loreto. From 1992 to 1997, malaria increased 50-fold in Loreto but only fourfold in Peru. Plasmodium falciparum infection , which has increased at a faster rate than P. vivax infection in the last 3 years, became the dominant Plasmodium infection in the highest transmission areas in the 1997 rainy season. The vector Anopheles darlingi has also increased during this epidemic in Loreto. Moreover, chloroquine and pyrimethamine-sulfadoxine drug -resistant P. falciparum strains have emerged, which require development of efficacious focal drug treatment schemes.



Skipped


localhost:8080
Evidence-based Medicine Search
Evidence-based Medicine Search
malaria
Searching 733,326 articles. 16439 results in 186ms.
Show/Hide Concept Displays
2000
1000
0
malaria
falciparum
p. falciparum
malarial
falciparum malaria
Diagnoses
120
80
40
0
erythrocytes
erythrocyte
diagnostic tests
red blood cells
blood smear
Tests
50
40
30
20
10
0
artemisinin
anti-malarial
antimalarial
chloroquine
artesunate
Treatments
1 ... 8 9 10 11 12 13 ... 100
Article Title
Journal Title
Publication Date
Snippet
Insights from monkey malaria that can change thinking about human infections
monkey malaria 1
Malaria Journal
20/10/2010
2963208 Insights from monkey malaria that can change thinking about human infections monkey
Epidemiology of Imported Malaria in the Mediterranean Region
malaria 4
increasing 2
infectious diseases 1
supported 1
presence 1
support 1
Mediterranean Journal of Hematology and Infectious Diseases
07/05/2012
3375659 Epidemiology of Imported Malaria in the Mediterranean Region
Malaria is one
Education and knowledge helps preventing malaria, but not "degedege"
Malaria Journal
20/10/2010
2963290 Education and knowledge helps preventing malaria, but not "degedege" Malaria Journal 20/10
Controlling malaria in Niger with bednets: how to take the Big Picture
Malaria Journal
20/10/2010
2963219 Controlling malaria in Niger with bednets: how to take the Big Picture Malaria Journal 20
Utility of Health Facility-based Malaria Data for Malaria Surveillance
malaria 17
valuable 1
groups 1
smears 1
laboratory diagnosis 1
control 2
analyzed 2
treated 2
PLoS ONE
13/02/2013
3572108 Utility of Health Facility-based Malaria Data for Malaria Surveillance smears 1 laboratory
Changing Malaria Transmission and Implications in China towards National Malaria Elimination Programme between 2010 and 2012
malaria 12
changing 1
findings 1
analyze 1
extracted 1
analyzed 1
PLoS ONE
09/09/2013
3767829 Changing Malaria Transmission and Implications in China towards National Malaria

How does a good health query look like?

- [Tamine&Chouquete, 2017] found that in health search, query quality is influenced by medical expertise
- [Koopman et al., 2017] studied the querying behaviour of 4 clinicians
  - most effective clinicians those who entered **short queries** (but retrieval models optimised for short queries)
  - most effective clinicians those who **inferred novel keywords** most likely to appear in relevant documents
  - most effective clinicians posed queries around **treatments rather than diagnoses** (but influenced by task: searching for clinical trials)

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"*

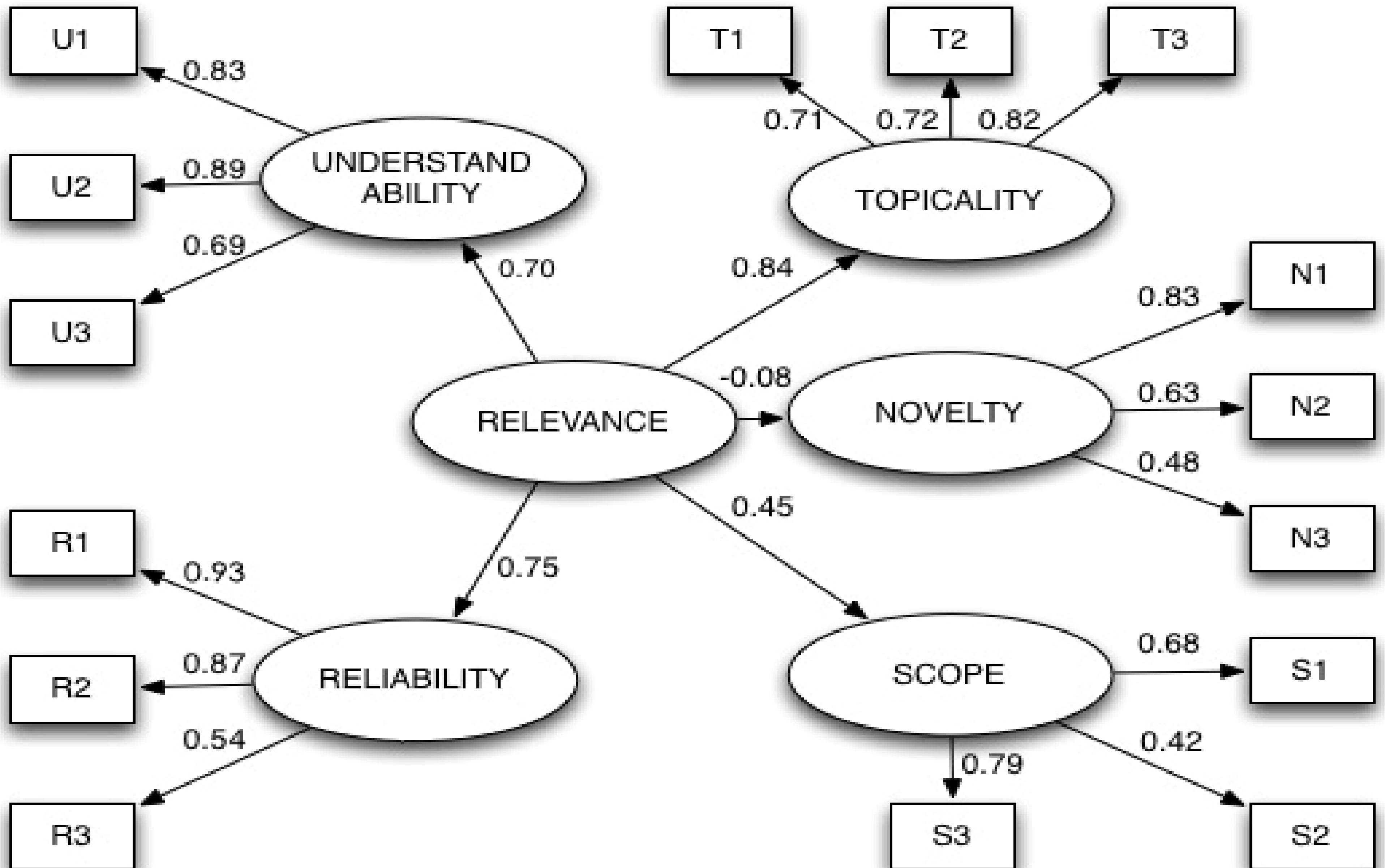
Expertise and Relevance Assessments

[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”`

Skipped

Assessing beyond topical relevance



[Zhang et al., 2014]

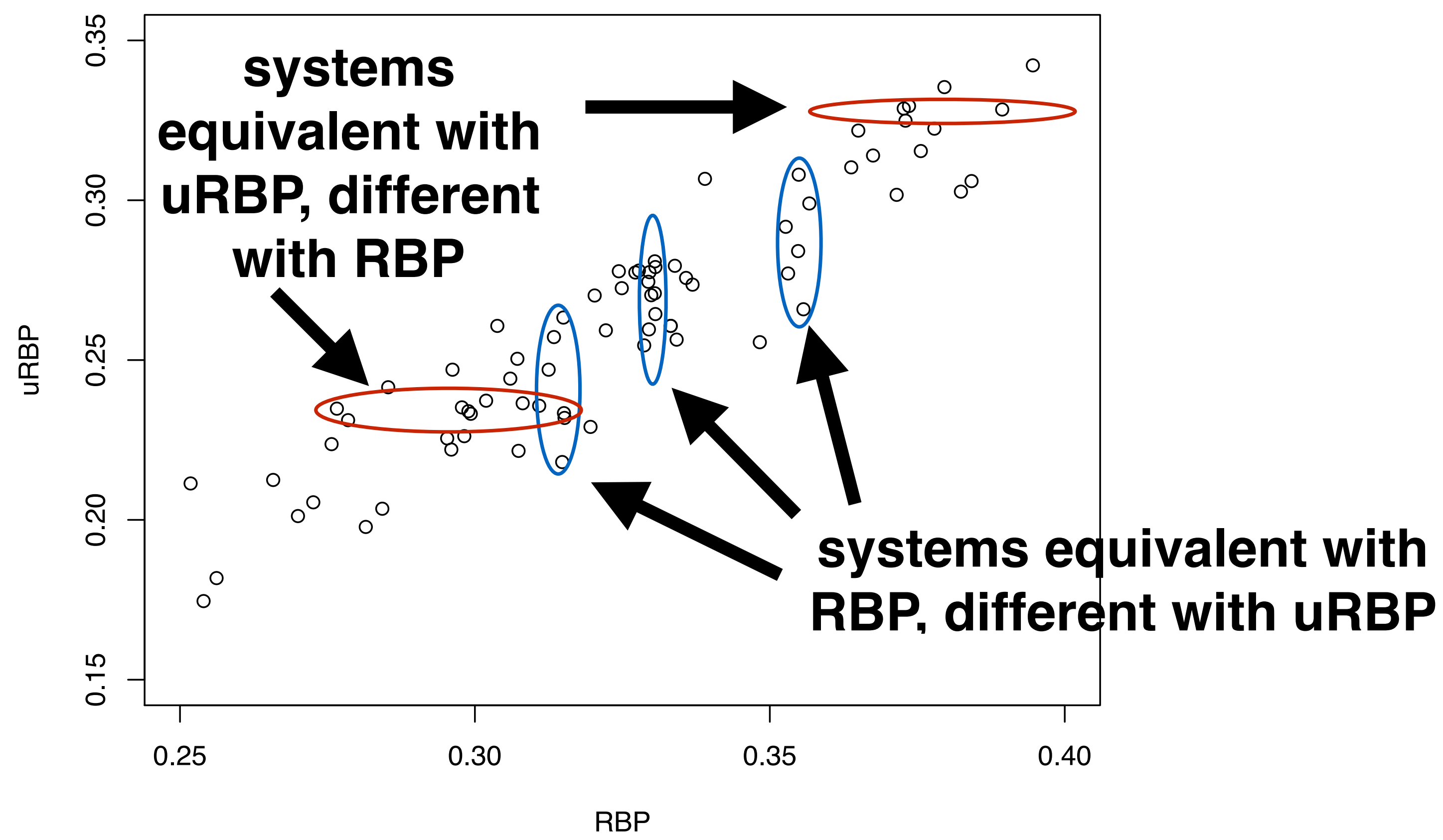
Skipped

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

Skipped

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

Skipped

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]

Skipped

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 heada
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 aspirin, atenolol, Lipitor, Advair,
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

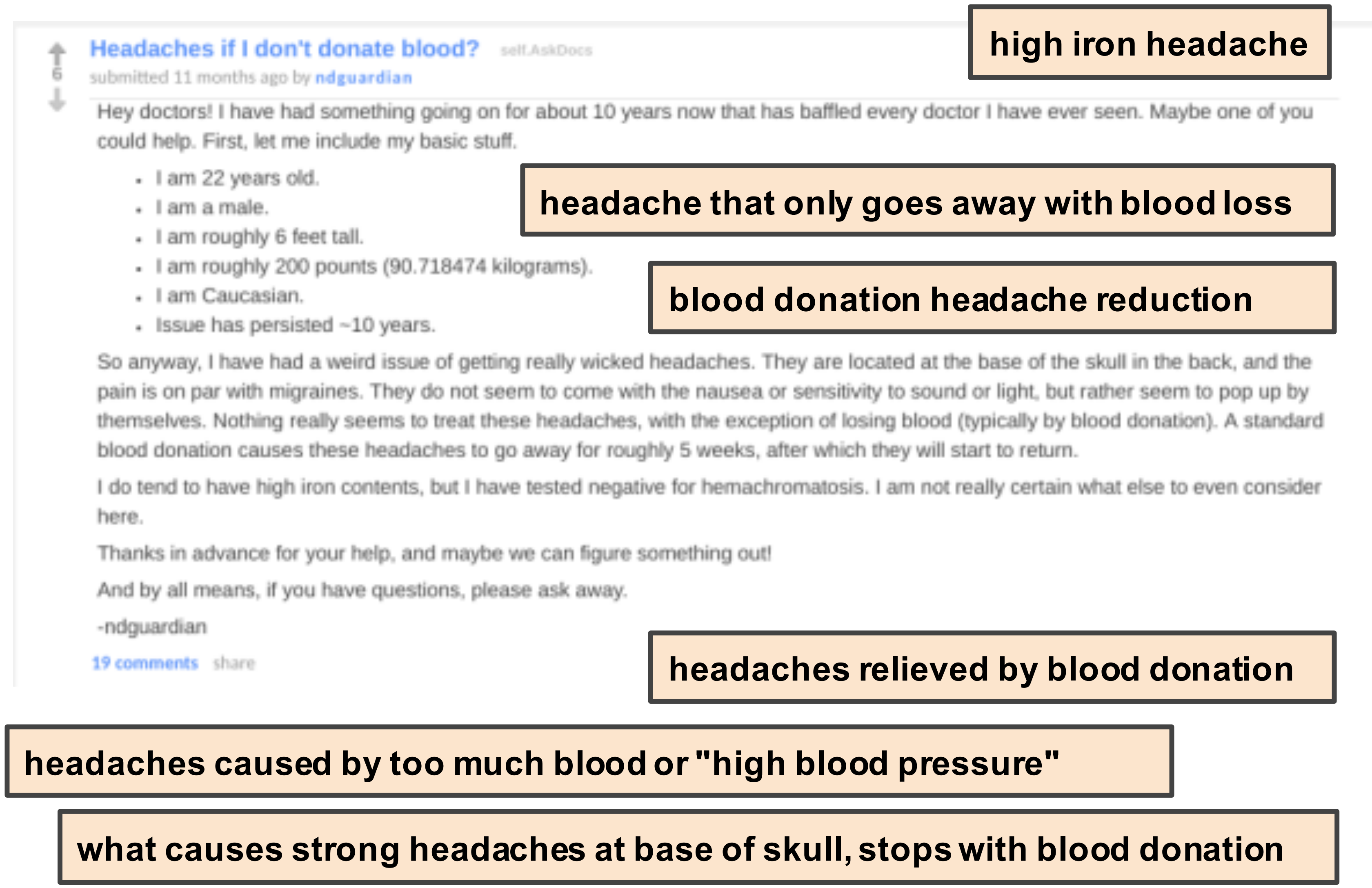
[Zuccon et al., 2016]

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

Skipped

- **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

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.io/russir2018-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"

PhD Projects Available

- We are recruiting PhD students!
- PhD projects available in the areas of interests the ielab:
  - **formal models of IR** (search methods, user models, evaluation)
  - **Health search** and domain-specific search
- Funding:
  - One **full scholarship** available for **CHS** for start in 2019
  - One **full scholarship** available for any topic of interest for start in 2019
  - Other scholarships possibly available through UQ
- **Join the ielab at UQ**:
  - Top-50 University in the World
  - 3 years and half of PhD funding
  - **Great lifestyle** in Brisbane! Avg temp 21 degrees C

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Avg temp in Kazan 4 degrees C

Thanks!

- The material in this lecture series is based on the HS SIGIR 2018 tutorial developed together with **Dr Bevan Koopman** (AEHRC, CSIRO)
- My PhD students **Anton van der Vegt**, **Harrisen Scells** and **Jimmy** have provided comments and support when developing parts of this tutorial
- Thanks to the **RUSSIR organisers** for inviting me and for the **Student Volunteers** for their friendly help and assistance

Thanks for attending!

Guido Zuccon
Queensland University of Technology

@guidozuc

THE END