Medical Information Retrieval

Lynda Tamine
Paul Sabatier University
IRIT, Toulouse - France

Lorraine Goeuriot
University of Grenoble Alpes
LIG, Grenoble - France

June 26, 2019

Outline

Introduction

Who are we?

Prof. Lynda Tamine
Department of Computing Science, University Paul Sabatier, Toulouse (France)
Research interests: modelling and evaluation of medical, contextual, collaborative and deep information retrieval.
Works on the characterization of medical queries according to diverse facets such as users expertise, task and difficulty and on semantic search models within medical settings, designing semantic models for health search.

Dr. Lorraine Goeuriot
LIG - Universit Grenoble Alpes
Research interest: modelling and evaluation of medical information retrieval, natural language processing.
Organised the Medical Information Retrieval Workshop in SIGIR. Since 2013, involved in the CLEF eHealth evaluation lab.

Introduction

Who are you?

- Introduce yourself: name, organization, profession
- What is your motivation behind attending the tutorial?
- What do you mainly expect to learn about?

Outline

Information Retrieval in the Medical Domain

Sessions in recent conferences

- Every year papers on medical/health applications in the main IR conferences: ECIR, CIKM, SIGIR
- Every year papers on data and knowledge management in the main Medical informatics conference: AMIA, AIME
- Special tracks on the topic of "Health": the Web Conference in 2018, ECIR'18
- Special tracks on the topic of "Ontologies and knowledge representation and access": AIME'19, AIME'18

Special issues in journals

- 2018: journal of Informatics in Medicine special issue on *"Information Retrieval: A Health Care Perspective"*
- 2017: JASIST special issue on *"Biomedical Information Retrieval"*
- 2016: Journal of Information Retrieval special issue on *"Medical Information Retrieval"*

Workshops (Related to the topic)

- Workshop on Health Search and Discovery at SIGIR 2013
- Workshop Medical Information Retrieval (MedIR) at SIGIR 2014 and 2016
- Workshop Knowledge Representation for Health at AIME 2019
- Workshop Semantic extraction from Medical Texts at AIME 2017

Evaluation challenges

- TREC is running at least one medical task per year: in 2019 TREC Precision Medicine track
- CLEF is running two medical tasks per year: eHealth and imageCLEF medical task
- NTCIR is running the Medical NLP task

Objectives

1. Summarize fundamentals in information retrieval
2. Present a review of tasks, users and resources in the medical domain
3. Present state-of-the art models and techniques in medical information retrieval
4. Present the major medical search evaluation benchmarks and report the key result trends
5. Summarize challenges and research opportunities

Outline

Outline

What is IR?

The IR view

Information Retrieval (Mooers:1951)

"Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his **need for information** into an actual list of citations to **documents** in storage containing information **useful** to him... Information retrieval is crucial to documentation and organization of knowledge"

Information Retrieval [Salton and McGill, 1986]

"Information retrieval systems are designed to help analyze and describe the **items** stored in a file, to organize them and search among them, and finally to retrieve them **in response** to a **user's query**. Designing and using a retrieval system involves four major activities: information analysis, information organization and **search**, **query formulation**, and **information retrieval** and dissemination"

- Main keywords, phrases: naïve definitions
  - *Information item*: a single unit of (textual) information (document, blog, tweet, e-mail, medical visit report, etc.)
  - *Information need*: what the users' seek for
  - *Query*: explicit formulation of the user's information need
  - *Retrieve useful information*: select information items that are relevant to the query

What is IR?

The well known search engines

- When we talk about information retrieval, we think about web search engines...

What is IR?

The well known search engines

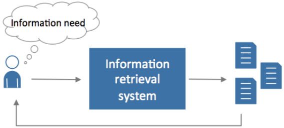
- When we talk about information retrieval, we think about web search engines...

- ...but there also a plenty of other search systems
  - Search in digital libraries
  - Search in entreprise corpora
  - Search for medical patient records
  - Search for legal texts

What is IR?

The Anatomy of an IR process

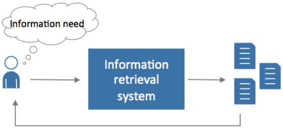
- Look at the IR process from the user side



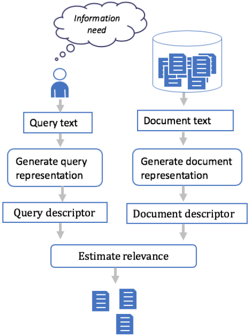
WHAT IS IR?

THE ANATOMY OF AN IR PROCESS

- Look at the IR process from the user side



- Look at the IR process from the system side



What is IR?

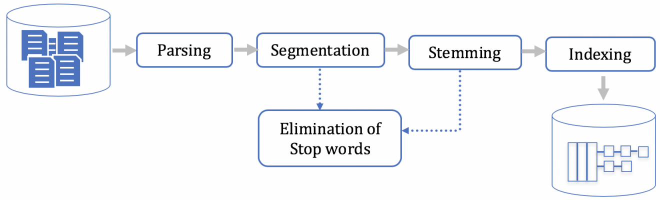
Document representation

- The document
  - Refers to a unit of information searchable by the user
  - Has a syntax and a semantics, specified by the author

What is IR?

Document representation

- The document
  - Refers to a unit of information searchable by the user
  - Has a syntax and a semantics, specified by the author
- Document preprocessing: logical view of a document from full text to a set of index terms
  - Lexical analysis of the text (parsing and segmentation): converting stream of chars into stream of words
  - Elimination of stopwords: words that appear too frequently
  - Stemming of the remaining words: reduce to stems after removing prefixes/suffixes, plural, gerund forms, ...
  - Selection of index terms or keywords: use useful words as index terms (all words, nouns, ...)
  - (Construction of term categorization structures like thesaurus, terminologies, ontologies): list of important words or concepts in a domain, related words, concepts...



What is IR?

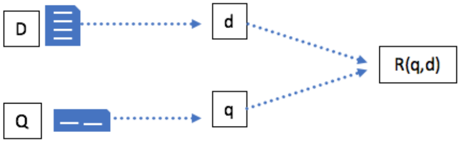
Relevance estimation: how?

- Relevance is a core concept in IR [Borlund, 2003]
  - Relevance is subjective (user's dependent): adequacy of the document to answer the query
  - Multiple dimensions lead to multiple types of relevance: aboutness, novelty, understability, reliability
    - Topical relevance: the major type of relevance addressed in IR based on the "aboutness" criteria
  - Dynamic: user's perception of relevance evolves through the search episodes

What is IR?

Relevance estimation: how?

- Relevance is a core concept in IR [Borlund, 2003]
  - Relevance is subjective (user's dependent): adequacy of the document to answer the query
  - Multiple dimensions lead to multiple types of relevance: aboutness, novelty, understability, reliability
    - Topical relevance: the major type of relevance addressed in IR based on the "aboutness" criteria
  - Dynamic: user's perception of relevance evolves through the search episodes
- IR models: a framework for relevance estimation
  - Algorithmic relevance: assign scores to documents with regard to a given query based on content matching
  - Process in two main stages
    1. Design a logical (formal) framework for representing documents and queries
    2. Define a ranking function that measures the topical similarities between queries and documents and orders the documents according to this measure



What is IR?

The semantic gap issue

She takes **just like a woman**, yeah she does. She makes love **just like a woman**, yeah she does And she aches **just like a woman** But she breaks **just** like a little girl

Just like a woman

One of many Dylan songs with an unclear subject. It's often thought tobe about fellow folk-singer Joan Baez, with whom Dylan had arelationship. Edie Sedgwick, an actress affiliated with Andy Warhol, isalso thought to have inspired the song.

- Representing documents and queries:
  - Understand broad language: what's behind the surface of strings, bag of words?
  - **Semantic representation rather than string/lexical representation**
    - Disambiguation of **entities**, **concepts** and roles
    - Reasoning and inference of relations
- Relevance ranking:
  - Understand broad relevance: what's behind the surface of matching?
  - Semantic matching rather than string matching
  - Relevance matching vs. semantic matching [Guo et al., 2016]

WHAT IS IR?

IR MODELS

- Design retrieval models: a long standing research in IR
- Major IR models: vectorial model [Salton et al., 1975], probabilistic model [Jones et al., 2000], language model [Ponte and Croft, 1998]

Fang and Zhai [2005]: Axiomatic approach
He and Ounis [2005]: TF in BM25 and DFR
Metzler and Croft [2004]: LM and PIN's
Robertson [2005]: Event Spaces
Robertson [2004]: Understanding IDF
Sparck-Jones et al. [2003]: LM and Relevance
Croft and Lafferty [2003], Lafferty and Zhai [2003]: LM book
Zaragoza et al. [2003]: Bayesian extension to LM
Bruza and Song [2003]: probabilistic dependencies in LM
Amati and van Rijsbergen [2002]: DFR
Lavrenko and Croft [2001]: Relevance-based LM
Hiemstra [2000]: TF-IDF and LM
Sparck-Jones et al. [2000]: probabilistic model: status
Ponte and Croft [1998]: LM
Brin and Page [1998a], Kleinberg [1999]: Pagerank and Hits
Robertson et al. [1994], Singhal et al. [1996]:
Pivoted Document Length Normalization
Wong and Yao [1995]: $P(d \rightarrow q)$
Robertson and Walker [1994], Robertson et al. [1995]: 2-Poisson, BM25
Church and Gale [1995], Margulis [1992]: Poisson
Fuhr [1992b]: Probabilistic Models in IR
Turtle and Croft [1991, 1990]: PIN's
Fuhr [1989]: Models for Probabilistic Indexing
Cooper [1991, 1988, 1994]: Beyond Boole, Probability Theory in IR: An Encumbranc
Deerwester et al. [1990], Dumais et al. [1988]: Latent semantic indexing
van Rijsbergen [1986, 1989]: $P(d \rightarrow q)$
Bookstein [1980], Salton et al. [1983]: Fuzzy, extended Boolean
Croft and Harper [1979]: BIR without relevance
Robertson and Sparck-Jones [1976]: BIR
Salton [1971], Salton et al. [1975]: VSM, TF-IDF
Rocchio [1971]: Relevance feedback
Maron and Kuhns [1960]: On Relevance, Probabilistic Indexing, and IR

What is IR?

IR models: the vectorial space model (VSM)

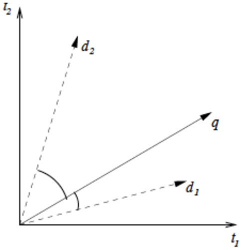
- Still one of the prominent IR frameworks
- A term space is a vector space where each dimension represents one term in the corpora vocabulary
- Each document d_j is represented by a vector of document term weights in the term space
- Each query q is represented by a vector of document term weights in the term space
- Document term weights can be computed, e.g., using tf-idf schema

$$\mathbf{d_j} = \begin{pmatrix} d_{j_1} \\ \boxed{d_{j_2}} \\ \cdots \\ d_{j_n} \end{pmatrix} \qquad \begin{array}{c} \text{Weight of term} \\ t_2 \\ \text{in document} \\ d_j \end{array} \qquad \vec{q} = \begin{pmatrix} q_1 \\ q_2 \\ \cdots \\ q_n \end{pmatrix}$$

What is IR?

IR models: the vectorial space model (VSM)

- The ranking function computes a retrieval status value (RSV) using a vectorial similarity measure
  - Scalar product
  - Jaccard similarity
  - Dice similarity
  - ...
- Documents are ranked according to decreasing RSV



WHAT IS IR?

IR TECHNIQUES: THE WIDELY USED

- Query reformulation
  - Enhance query representation at the retrieval stage by reweighting terms, adding useful terms
  - Use evidence from the user (eg., clicks, search history,...), external resources (eg., thesaurus, corpora, ...), ...
- Document expansion
  - Enhance document representation at the indexing stage by reweighting the terms, adding useful terms
  - Use evidence from external resources (eg., thesaurus, corpora)
- Document re-ranking
  - Update the initial RSV scores of documents
  - Use evidence from the user, external resources, ...

Outline

Why is medical IR important?

IR and society

- Significant societal impact on health care, well being
- Myriad of medical repositories: health records, blogs, guidelines, scientific publications, ...
- Myriad of health search tasks: diagnosis, seek for health advice, patient search (clinical trial), understand conditions and treatments, find health provider,
- Myriad of users' profiles: consumers, clinicians, institutions, ...

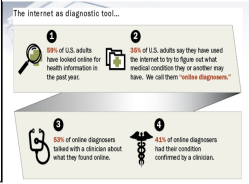
Why is medical IR important?

A largely web-driven activity

- Search engines and social media are popular tools for seeking and sharing information about a range of health conditions PewInternet, October 2013; [De Choudhury et al., 2014, White and Horvitz, 2014]

The % of adult internet users who have looked online in the last 12 months for information about...

55%	Specific disease or medical problem
43	Certain medical treatment or procedure
27	How to lose weight or how to control your weight
25	Health insurance, including private insurance, Medicare or Medicaid
19	Food safety or recalls
16	Drug safety or recalls
16	A drug you saw advertised
15	Medical test results
14	Caring for an aging relative or friend
12	Pregnancy and childbirth
11	How to reduce your health care costs
20	Any other health issue
72	**at least one of the above topics**



Why is medical IR important?

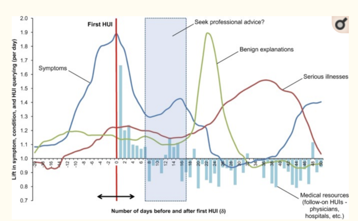
A largely web-driven activity

- Search engines and social media are popular tools for seeking and sharing information about a range of health conditions PewInternet, October 2013; [De Choudhury et al., 2014, White and Horvitz, 2014]

The % of adult internet users who have looked online in the last 12 months for information about...

55%	Specific disease or medical problem
43	Certain medical treatment or procedure
27	How to lose weight or how to control your weight
25	Health insurance, including private insurance, Medicare or Medicaid
19	Food safety or recalls
16	Drug safety or recalls
16	A drug you saw advertised
15	Medical test results
14	Caring for an aging relative or friend
12	Pregnancy and childbirth
11	How to reduce your health care costs
20	Any other health issue
72	**at least one of the above topics**

Existence of behavioral patterns suggesting a strong relationship between search behavior and health care [White and Horvitz, 2010, White and Horvitz, 2014]



What makes medical search challenging?

Overview

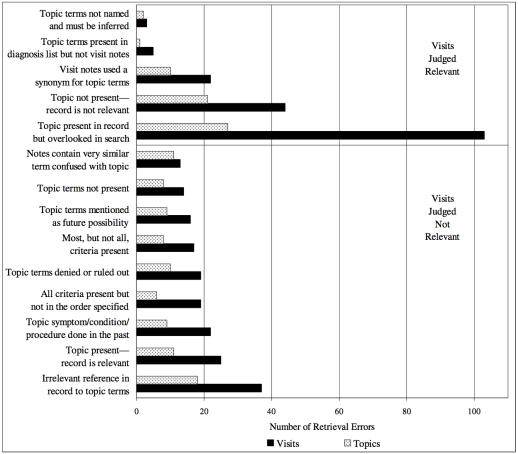
- Lexical representation [Tamine et al., 2015, Stanton et al., 2014, Limsopatham et al., 2013, Edinger et al., 2012, Chapman et al., 2001]
- Lexical matching [Edinger et al., 2012, Dinh and Tamine, 2012]
- Result appraisal [Koopman and Zuccon, 2014, Tamine and Chouquet, 2017, Roberts et al., 2015a, W. White and Horvitz, 2009, White and Horvitz, 2013, Cartright et al., 2011a, Palotti et al., 2016]

		Lexical representation	Lexical Matching	Result appraisal
Text (Queries/Documents)	High occurrence of lexical variants	■	■	■
	Ambiguity of language	■	■	■
	Significant presence of negation	■	■	■
	Time factors	■	■	■
Search task	Domain-specific task	□	■	■
User	Variability in levels of expertise	■	□	■
	Difficulty of understanding medical language	□	□	■
	Cognitive bias	□	□	■

What makes medical search challenging?

Lexical representation and matching issues

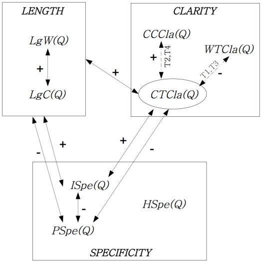
- Failure analysis from the TREC Medical Records Track [Edinger et al., 2012]
  - Same task across queries: retrieve cohorts of patients fitting criteria similar to those specified for participation in clinical studies
  - Same user's profile: domain-expert (physician)
  - Main results: **both precision errors and recall errors were due to bad lexical representations and lexical mismatches**



What makes medical search challenging?

Domain-specific tasks: need of task-dependent matching?

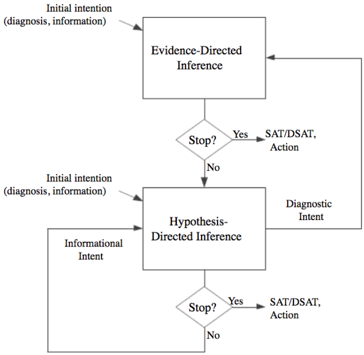
- Study of the peculiarities of queries across medical tasks [Tamine et al., 2015]
  - Study of 4 search tasks:
    - (T1) Retrieve from biomedical scientific literature relevant references that address biology or related protein products (eg., '*Arginin vassopressin*')
    - (T2) Retrieve from medical scientific literature, relevant documents that fit with a patient profile (eg., '*Adult respiratory distress syndroms*')
    - (T3) Identify cohorts in clinical studies for comparative effectiveness research (eg., *Retrieve relevant medical cases including images for differential diagnosis*)
    - (T4) Identify relevant references that deal with typical relations between an entity and a medical process (eg., *What is the role of gene gamma-aminobutyric acid receptors in the process of inhibitory synaptic transmission?*)
  - Same user's profile: domain-expert (physician)
  - Main results:
    - Queries vary significantly across tasks:
      have multiple topical facets expressed using different levels of specifity w.r.t. medical terminology
    - Query performance significantly vary across tasks:
      however, the shorter and less specific the query is,
      the more difficult it is regardless of the task

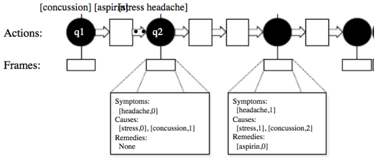


WHAT MAKES MEDICAL SEARCH CHALLENGING?

DOMAIN-SPECIFIC TASKS: NEED OF TASK-DEPENDENT MATCHING?

- Diagnosis, a common medical search setting, is a highly complex task [Cartright et al., 2011a]
- Identifying two medical foci in search sessions corresponding to two iterative and interactive phases:
  - Evidence-directed: findings are merged to build a list of potential explanatory diagnoses ranked by likelihood
  - Hypothesis-directed: list of diagnoses used to guide collection of additional evidence to validate candidate hypotheses

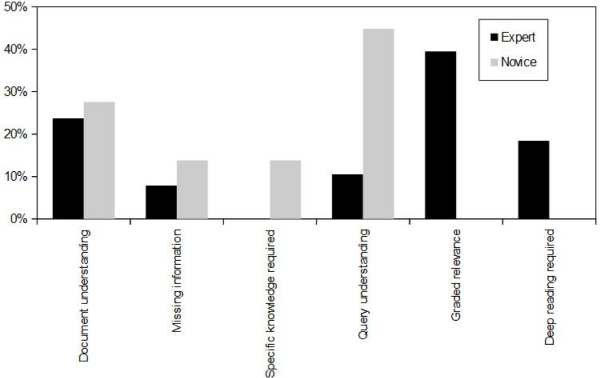




What makes medical search challenging?

Result appraisal: what does make it difficult with respect to user's expertise?

- *Study the factors of the relevance assessment task difficulty* [Tamine and Chouquet, 2017, Koopman and Zuccon, 2014, Roberts et al., 2015a, Voorhees and Hersh, 2012]
  - Qualitative analysis of relevance assessments for equivalent information needs
  - Different user's profiles: domain-expert (physician), domaine-novice (laypeople)
  - Main results:
    - Levels of relevance agreement are low for both experts and novices
    - Better level of relevance agreement among experts than novices
    - More than third of the assessors found the relevance assessment task difficult but the underlying reasons different among experts vs. novices



What makes medical search challenging?

User cognitive bias

- *Study the potential of web search to escalate user's medical concerns* [White and Horvitz, 2013, W. White and Horvitz, 2009]
  - Qualitative and quantitative analysis of users' medical search logs and surveys
  - Observe the escalation within session: severity increase in the search terms of evolving medical search sessions
  - Main results:
    - Using Web search to perform diagnosis is a common user activity (more than 24%) while general-purpose ranking functions are based on the presence of lexical query symptoms, not designed for diagnosis inference
    - Common users consider that the system ranks the potential explanatory diagnoses by likelihood

Table I. Probability of Mention of Cause Given Symptom

Symptom	Cause	Web Crawl	Web Search	Domain Search
headache	caffeine withdrawal	.29	.26	.25
	tension	.68	.48	.75
	brain tumor	.03	.26	.00
muscle twitches	benign fasciculation	.53	.12	.34
	muscle strain	.40	.38	.66
	ALS	.07	.50	.00
chest pain	indigestion	.28	.35	.38
	heartburn	.57	.28	.52
	heart attack	.15	.37	.10

Summary of issues

- Semantic gap
  - Vocabulary mismatch between experts and novices
  - Vocabulary mismatch between the query and documents
  - Interpretation of vocabulary in context: negation, lexical variants, time, task peculiarities
- Result appraisal
  - Understand medical language
  - Ability to interpret the results, make accurate inference, assess the credibility
  - Cognitive bias

Outline

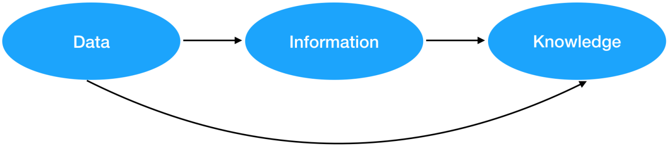
Medical Information

What is information?

Information $\neq$ data $\neq$ knowledge

- Data: observations and measurements made about the world
- Information: data brought together in aggregate to demonstrate facts
- Knowledge: what is learned from the data and information, that can be applied in new situations to understand the world

[Blum, 1984] cited in [Hersh, 2010]



Medical Information

Scientific Information Properties

Properties of scientific texts [Hersh, 2010]:

- **Growth**: The amount of scientific publications is growing exponentially
- **Obsolescence**: scientific advances, constant update of the state-of-the-art and changes in society make information quickly obsolete
- **Fragmentation**: text published often reflects only one part of a problem or situation
- **Links and citations**: strong property of scientific text, links and references allow to generate networks among works and communities
- **Propagation**: simplicity of information flow

Medical Information

A Classification of Textual Health Information

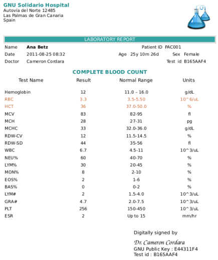
[Hersh, 2010] distinguishes two main categories of textual health documents:

- **Patient-specific information**: applies to individual patients. Tells healthcare providers, administrators and researchers about the health and disease of a patient.
  - Structured: laboratory results, vital signs
  - Narrative: history and physical, progress notes, radiology report
- **Knowledge-based information**: has been derived and organized from observational or experimental research. Usually provided in books, journals or *computerized media*.
  - Primary: original research (in journals, books, reports, etc.)
  - Secondary: summaries of research (in review articles, books, practice guidelines, etc.)

With the emergence of Web2.0, one could also consider **User-generated Content** as another category:

- Collaborative writing: wikipedia, blogs
- Social media: discussion forums, Facebook, Twitter, PatientsLikeMe

Medical Information



GNU Solidario Hospital
Autovía del Norte 12485
Las Palmas de Gran Canaria
Spain

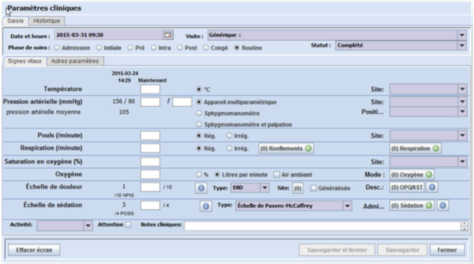
LABORATORY REPORT

Name **Ana Betz** Patient ID PAC001
Date 2011-08-25 08:32 Age 25y 10m 26d Sex Female
Doctor Cameron Cordara Test id B165AAF4

COMPLETE BLOOD COUNT

Test Name	Result	Normal Range	Units
Hemoglobin	12	11.0 – 16.0	g/dL
RBC	3.3	3.5-5.50	10^6/uL
HCT	36	37.0-50.0	%
MCV	83	82-95	fl
MCH	28	27-31	pg
MCHC	33	32.0-36.0	g/dL
RDW-CV	12	11.5-14.5	%
RDW-SD	44	35-56	fl
WBC	6.7	4.5-11	10^3/uL
NEU%	60	40-70	%
LYM%	30	20-45	%
MON%	8	2-10	%
EOS%	2	1-6	%
BAS%	0	0-2	%
LYM#	2	1.5-4.0	10^3/uL
GRA#	4.7	2.0-7.5	10^3/uL
PLT	256	150-450	10^3/uL
ESR	2	Up to 15	mm/hr

Digitally signed by
Dr. Cameron Cordara
GNU Public Key : E44311F4
Test id : B165AAF4



HEMATOLOGIE

NUMERATION ET FORMULE LEUCOCYTAIRE

	Résultats	Normes
HEMATIES	5.320.000 /mm3	[illegible]
HEMOGLOBINE	16,5 g/100 ml	13,0 à 17,0
HEMATOCRITE	48,7 %	40,0 à 54,0
LEUCOCYTES	6.000 /mm3	4.000 à 10.000
PLAQUETTES	226.000 /mm3	150.000 à 500.000

VITESSE DE SEDIMENTATION

1ère heure	8 mm	inf. à 10

CHIMIE DU SANG

	Résultats	Normes
aspect du sérum	normal	
GLYCEMIE	1,08 g/l 5,99 mmol/l	[illegible]
UREE	0,33 g/l 5,48 mmol/l	[illegible]
CREATININEMIE	10 mg/l 88 umol/l	7 à 12 62 à 106

EXPLORATION LIPIDIQUE

CHOLESTEROL TOTAL	2,86 g/l 7,38 mmol/l	inf. à 2,00 inf. à 5,16
H.D.L.	0,53 g/l 1,37 mmol/l	[illegible]
TRIGLYCERIDES	1,54 g/l 1,76 mmol/l	inf. à 2,00 inf. à 2,28
LDL CHOLESTEROL	2,02 g/l 5,23 mmol/l	[illegible]
PROTEINE C-REACTIVE	inf à 3 mg/l	inf. à 5,0

Medical Information

Narrative Patient Specific Information

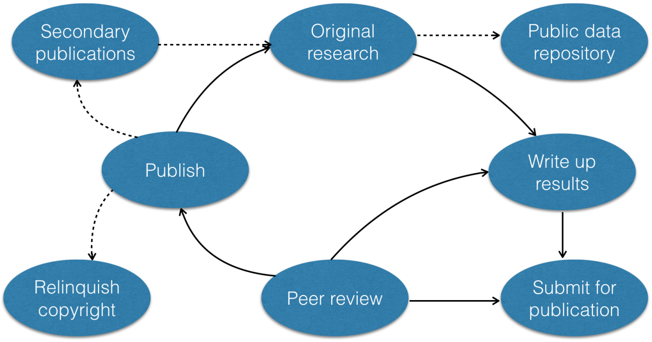
```
Admission Date:  [**2015-03-17**]        Discharge Date:  [**2015-03-24**]
Date of Birth:   [**1974-10-03**]        Sex:  F
Service: Neurosurgery
HISTORY OF PRESENT ILLNESS:  The patient is a 40-year-old female with complaints
of headache and dizziness.  In [**2015-01-14**], the patient had headache with
neck stiffness and was unable to walk for 45 minutes.   [...]
PAST MEDICAL HISTORY:  Hypothyroidism.
ALLERGIES:  Penicillin and Bactrim which causes a rash.
MEDICATIONS:  Levoxyl 1.75 mg.
PHYSICAL EXAMINATION:  On physical examination, her blood pressure was 104/73,
pulse 79.  In general, she was a woman in no acute distress.  HEENT:  Nonicteric.
Pupils are equal, round, and reactive to light.  Extraocular movements are full.
[...]

On postoperative day #1, the patient was taken to arteriogram, where she
underwent a cerebral angiogram to evaluate clipping of the aneurysm.  [          ]
DISCHARGE MEDICATIONS:
1. Hydromorphone 2-6 mg po q4h prn.
2. Synthroid 175 mcg po q day.[...]

CONDITION ON DISCHARGE:  Stable.

FOLLOW-UP INSTRUCTIONS:  She will follow up in 10 days for staple removal with
Dr. [**Last Name (STitle) 570**].

(End of Report)
```

Discharge summary extracted from the MIMIC II dataset
`https://physionet.org/mimic2/`.

Medical Information

Primary Knowledge-based Documents

- Contain reports of research results: discoveries, observations, description of related work and position of the report, conclusions.
- Has never been published before
- Published in books, journals or conference proceedings
- Usually a small number of documents have the highest impact



Medical Information

Primary Knowledge-based Documents

Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search

RYEN W. WHITE and ERIC HORVITZ
Microsoft Research

The World Wide Web provides an abundant source of medical information. This information can assist people who are not healthcare professionals to better understand health and illness, and to provide them with feasible explanations for symptoms. However, the Web has the potential to increase the anxieties of people who have little or no medical training, especially when Web search is employed as a diagnostic procedure. We use the term *cyberchondria* to refer to the unfounded *escalation* of concerns about common symptomatology, based on the review of search results and literature on the Web. We performed a large-scale, longitudinal, log-based study of how people search for medical information online, supported by a survey of 515 individuals' health-related search experiences. We focused on the extent to which common, likely innocuous symptoms can escalate into the review of content on serious, rare conditions that are linked to the common symptoms. Our results show that Web search engines have the potential to escalate medical concerns. We show that escalation is associated with the amount and distribution of medical content viewed by users, the presence of escalatory terminology in pages visited, and a user's predisposition to escalate versus to seek more reasonable explanations for ailments. We also demonstrate the persistence of postsession anxiety following escalations and the effect that such anxieties can have on interrupting user's activities across multiple sessions. Our findings underscore the potential costs and challenges of cyberchondria and suggest actionable design implications that hold opportunity for improving the search and navigation experience for people turning to the Web to interpret common symptoms.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process; query formulation*

General Terms: Human Factors, Experimentation

Additional Key Words and Phrases: Cyberchondria

ACM Reference Format:

White, R. W. and Horvitz, E. 2009. Cyberchondria: Studies of the escalation of medical concerns in Web search. ACM Trans. Inf. Syst. 27, 4, Article 23 (November 2009), 37
DOI = 10.1145/1629096.1629101 http://doi.acm.org/10.1145/1629096.1629101

1. INTRODUCTION

The World Wide Web has the potential to provide valuable medical information to people, where Web sites such as WebMD (http://www.webmd.com) and MSN

Authors' addresses: R. W. White and E. Horvitz, Microsoft Research, One Microsoft Way, Redmond, WA 98052; email: {ryenw, horvitz}@microsoft.com.

DOI 10.1145/1629096.1629101 http://doi.acm.org/10.1145/1629096.1629101

ACM Transactions on Information Systems, Vol. 27, No. 4, Article 23, Publication date: November 2009.

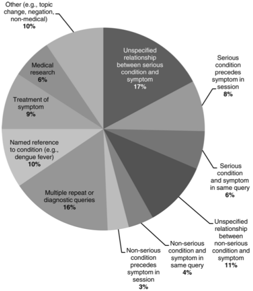


Fig. 1. Distribution of labels assigned to set of hand-labeled no-change sessions.

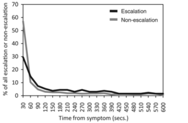


Fig. 2. Temporal distance from initial input of symptom (within session)

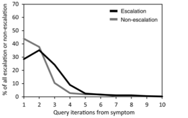


Fig. 3. Query distance from initial input of symptom (within session).

REFERENCES

Agichtein, E., Brill, E., and Dumais, S. 2006. Improving Web search ranking by incorporating user behavior information. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 19–26.
American Psychiatric Association. 1994. *Diagnostic and Statistical Manual of Mental Disorders* 4th Ed. APA, Washington, DC.
Asmundson, J. G., Taylor, S., and Cox, B. J. 2001. *Health Anxiety: Clinical and Research Perspectives on Hypochondriasis and Related Conditions*. Wiley.
Ayers, S. and Kronenfeld, J. 2007. Chronic illness and health-seeking information on the Internet. *Health 11*, 3, 327–347.
Baker, L., Wagner, T. H., Singer, S., and Bundorf, M. K. 2003. Use of the Internet and e-mail for health care information. *J. Amer. Med. Assoc. 289*, 18, 2401–2406.
Barsky, A. J. and Ahern, D. K. 2004. Cognitive behavioral therapy for hypochondriasis. *J. Amer. Med. Assoc. 291*, 12, 1464–1470.
Barsky, A. J. and Klerman, G. L. 1983. Overview: Hypochondriasis, bodily complaints, and somatic styles. *Amer. J. Psychiatry 140*, 273–283.
Benigeri, M. and Pluye, P. 2003. Shortcomings of health-related information on the Internet. *Health Promot. Int. 18*, 4, 381–387.
Berland, G. K., Elliott, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L., Broder, M. S., Kanouse, D. E., Muñoz, J. A., Puyol, J.-A., Marielena, L., Watkins, K. E., Yang, H., and McGlynn, E. A.

Medical Information

Secondary Knowledge-based documents

- All medical professionals are not researchers: primary resources need to be rephrased, summarized, synthetized
- Summary and reviews of primary resources are published in scientific journals
- Quality issue: the editorial process is not the same for secundary than primary resources
- Other category: clinical practice guidelines (many publications, very little control)

Specific case: Systematic Reviews and Meta-Analysis

- Fragmentation of the scientific literature → difficult to identify all the relevant papers on a topic
- In particular with clinical trials, large amount of publications on a similar condition or treatment
- Systematic reviews tackle a precise question, and describe the complete set of related work and factual approaches
- Meta-analysis compare results at the systematic review scale
- Topics: treatment (63%), causality and security (29%), diagnosis (4,4%), prognosis (2.1%) [Montori et al., 2004]
- Cochrane is a non-profit, non-governmental organization formed to organize medical research findings so as to facilitate evidence-based choices about health interventions `http://www.cochranelibrary.com/`

Medical Information

User Generated Content

Collaborative writing websites allow users to edit collaboratively documents. It can have some sort of editorial control. It includes:

- **Wikis** such as wikipedia (collective writing and control of the content)

[Blackman, 2006] showed that information contained on wikipedia wasn't erroneous (comparison on 42 topics with the Britannica Encyclopaedia)

- **Blogs**: discussion or informational website published on the Web consisting of discrete, often informal diary-style text entries ("posts").

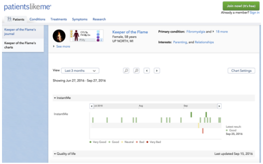
Medical Information

User Generated Content

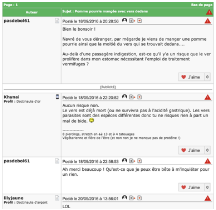
Health topics can be covered on all types of social media:

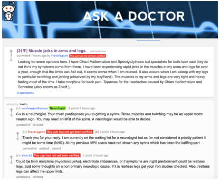
- General social media such as facebook, twitter:

- Medical social media such as PatientsLikeMe:



- Discussion forums: where all kinds of users (patients, doctors, students, nurses...) can discuss health topics





Medical Information

Certification

How can the quality of health information online be guaranteed?

The organization Health On the Net (HON) certifies the quality and validity of medical websites.
HON manually certifies website according to the following principles:

- Principle 1 : Authority - Give qualifications of authors
- Principle 2 : Complementarity - Information to support, not replace
- Principle 3 : Confidentiality - Respect the privacy of site users
- Principle 4 : Attribution - Cite the sources and dates of medical information
- Principle 5 : Justifiability - Justification of claims / balanced and objective claims
- Principle 6 : Transparency - Accessibility, provide valid contact details
- Principle 7 : Financial disclosure - Provide details of funding
- Principle 8 : Advertising - Clearly distinguish advertising from editorial content

`https://www.hon.ch/HONcode/Guidelines/guidelines.html`

Medical Search Tasks

Information need

Information needs [Hersh, 2010]:

- *Retrospective information needs*:
  - The need for help in solving a certain problem or making a decision
  - The need for background information on a topic
- *Current awareness information needs*:
  - The need to keep up with information in a given subject area

Amount of information needed [Lancaster and Warner, 1993]

- A single fact
- One or more documents
- A comprehensive search of the literature

Types of information needs [Wilkinson and Fuller, 1996]

- Fact-finding
- Learning
- Gathering
- Exploring

States of information need [Gorman, 1995]

- Unrecognized need
- Recognized need
- Pursued need
- Satisfied need

Medical Search Queries

Typology

The types of queries that are the most widely studied are:

- Classical keyword-based queries (physician vs patients)
- Boolean queries (systematic reviews)
- Structured queries (PICO)
- Multimodal queries (text + concepts e.g. pubmed search tools)

General classification of search queries from [Broder, 2002]:

- Navigational
- Transactional
- Informational

Classification of search queries for semantic search [Bast et al., 2016]:

- Structured
- Keyword-based
- Natural language-based

Medical Search Queries

Physician Queries

- Study by [Ely et al., 1999] on family doctors questions in their daily practise.
- Observation of 100 doctors from Iowa (US)

Taxonomy of generic questions:

- What is the cause of symptom X?
- What is the dose of drug Y?
- How should I manage disease or finding X?
- How should I treat finding or disease X?
- What is the cause of physical finding X?
- What is the cause of test finding X?
- Could this patient have disease or condition X?
- Is test X indicated in situation Y?
- What is the drug of choice for condition X?
- Is drug X indicated in situation Y?

Medical Search Queries

Physician Queries

- Study by [Ely et al., 1999] on family doctors questions in their daily practise.
- Observation of 100 doctors from Iowa (US)

Taxonomy of generic questions:

- What is the cause of symptom X?
- What is the dose of drug Y?
- How should I manage disease or finding X?
- How should I treat finding or disease X?
- What is the cause of physical finding X?
- What is the cause of test finding X?
- Could this patient have disease or condition X?
- Is test X indicated in situation Y?
- What is the drug of choice for condition X?
- Is drug X indicated in situation Y?

- These are questions and not queries - 64% were not pursued
- In 1999 Internet was not the primary source of information

Medical Search Queries

Clinical Queries

Analysis of search queries in an EHR search utility [Natarajan et al., 2010]

- **Navigational queries (14.5%)**: were mostly aiming at retrieving a specific EHR (e.g. using the record number)
- **Transactional queries (0.4%)**: were representing an action (e.g. adding a new note)
- **Information queries (85.1%)**: the most frequent, especially among clinicians and researchers.

Top 5 semantic types of searches

Semantic type	%	Semantic type	%
Laboratory or test result	29.2	Pharmacologic substance	7.5
Disease or syndrome	21.7	Diagnostic procedure	6.2
Body part, organ or organ component	8.1		

Top 10 most frequent queries

Query	%	Query	%
class	9.8	nephrogenic	1.8
nyha	4.5	hysterectomy	1.5
hodgkins	2.9	cva	1.1
iii	2.4	ef	1.0
iv	2.3	hf	0.9

- Very short queries (1.2 term(s) on average in the corpus)
- Many acronyms (*NYHA*) and abbreviations (*tach* for tachycardia)
- Ambiguous (*class*)

Medical Search Queries
Layperson queries

Particularities and challenges [Zhang et al., 2012]

- *Conceptual level*: layperson have their own understandings and hypotheses about a particular condition.
- *Terminological level*: layperson's vocabulary doesn't match medical terminologies
- *Lexical level*: queries contain mispelling, partial words, etc.
- Short text (on average less than 3 words), ambiguous

[Broder, 2002]

Web search categories:

- Navigational
- Transactional
- **Informational**

[Cartright et al., 2011b]

Topics covered:

- Symptom
- Cause
- Remedy

Types of queries:

- Evidence-directed
- Hypothesis-directed:
  - Diagnosis intent
  - Informational intent

Medical Search Queries

PICO Queries

Designed to answer Evidence-based Medicine problems, PICO stands for:

- **P**atient / **P**roblem / **P**opulation
- **I**ntervention
- **C**omparison / **C**ontrol
- **O**utcome

The formulation of a focused clinical question containing well-articulated PICO elements is widely believed to be **the key to efficiently finding high-quality evidence** and also **the key to evidence-based decisions** [Huang et al., 2006].

Example (from [Boudin et al., 2010]):

```
"children with pain and fever
how does paracetamol compared
with ibuprofen affect levels
of pain and fever?
```

→

Patient/Problem: `children/pain and fever`
Intervention: `paracetamol`
Comparison: `ibuprofen`
Outcome: `levels of pain and fever`

Medical Search Queries

Clinical Search Queries (Genomics)

[Hersh and Voorhees, 2009] categorized clinical queries into seveval Generic Topic Types:

Generic Topic Type	Example Topic
Find articles describing standard methods or protocols for doing some sort of experiment or procedure	Method or protocol: GST fusion protein expression in Sf9 insect cells
Find articles describing the role of a gene involved in a given disease	Gene: DRD4 Disease: alcoholism
Find articles describing the role of a gene in a specific biological process	Gene: Insulin receptor gene Biological process: Signaling tumorigenesis
Find articles describing interactions (e.g. promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease	Genes: HMG and HMGB1 Disease: Hepatitis
Find articles describing one or more mutations of a given gene and its biological impact	Gene with mutation: Ret Biological impact: Thyroid function

Medical Search Queries

Systematic review Queries

- Systematic reviews use boolean queries on specific databases such as the Cochrane library to retrieve all the possible relevant documents on a topic.
- Example (topic extracted from CLEF eHealth Technologically assisted reviews task [Kanoulas et al., 2017]):

```
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
    ...
```

Outline

Semantic Resources

Definitions

- Lexical and semantic resources are used **in many domains**
- Definitions are extracted from [Hersh, 2010] and [Bast et al., 2016]

A concept

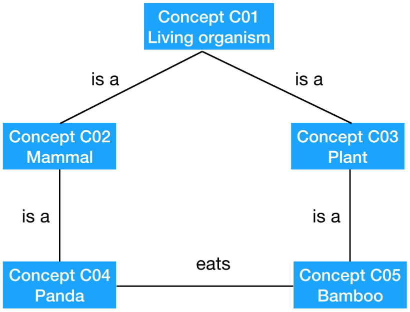
Idea or object that occurs in the world (e.g. *the condition under which human blood pressure is elevated*

A term

String of one or more words that represents a concept (e.g. *hypertension* or *high blood pressure*)

A relationship

Link between 2 concepts (e.g. the *liver* is an *organ*) or terms (e.g. *hypertension* and *high blood pressure* are synonyms)



Semantic Resources
Definitions

A controlled vocabulary or Terminology

Contains a list of terms that are the canonical representation of the concepts.

An ontology

Is a formal description of the concepts in a domain of discourse. Contains a *hierarchy of concepts* with various *relationships* between them. Concepts have *attributes* and *facets*.

A thesaurus

Contains concepts, terms, and relationships between them:

- *Hierarchical* links between broader and narrower concepts
- *Synonymy* and antonymy
- *Related* somehow otherwise related

A Knowledge Base is a collection of records in a database, which typically refers to some kind of knowledge about the world. Records are triples `(subject, predicate, object)`.

- [Bast et al., 2016] calls a knowledge every collection of entities following an ontology.
- A *knowledge-base can be thought as a graph* where entities are the nodes and the relationships are the edges.

Medical Knowledge Resources

Existing Medical thesauri (in English)

1. Medical Subject Headings (MeSH)

- Created by the National Library of Medicine *to index medical documents*
- 28,000 descriptors (concepts) with over 90,000 entry terms
- 3 types of relationships: hierarchical, synonymous, related

Hypertension MeSH Descriptor Data 2018

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading
Hypertension
Tree Number(s)
C14.907.489
Unique ID
D006973
Annotation
not for intracranial or intraocular pressure; relation to BLOOD PRESSURE: Manual 23.27; Goldblatt kidney is HYPERTENSION, GOLDBLATT see HYPERTENSION, RENOVASCULAR; hypertension with kidney disease is probably HYPERTENSION, RENAL, not HYPERTENSION; venous hypertension: index under VENOUS PRESSURE (IM) & do not coordinate with HYPERTENSION; PREHYPERTENSION is also available
Scope Note
Persistently high systemic arterial BLOOD PRESSURE. Based on multiple readings (BLOOD PRESSURE DETERMINATION), hypertension is currently defined as when SYSTOLIC PRESSURE is consistently greater than 140 mm Hg or when DIASTOLIC PRESSURE is consistently 90 mm Hg or more.
Entry Term(s)
Blood Pressure, High
NLM Classification #
WG 340
See Also
Antihypertensive Agents
Vascular Resistance
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2010/06/25

The 16 trees in MeSH

1. Anatomy
2. Organisms
3. Diseases
4. Chemicals and Drugs
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment
6. Psychiatry and Psychology
7. Biological Sciences
8. Natural Sciences
9. Anthropology, Education, Sociology and Social Phenomena
10. Technology, Industry, Agriculture
11. Humanities
12. Information Science
13. Named Groups
14. Health care
15. Publication Characteristics
16. Geographicals

Medical Knowledge Resources

Existing Medical thesauri (in English)

2 International Classification of Medicine (ICD)

- International statistical classification of diseases and health problems
- Coded medical classification including a wide variety of signs, symptoms, trauma, etc.
- Published by the WHO
- Internationaly used to register morbidity and causes and morbidity

ICD-10 Version:2016
- I Certain infectious and parasitic diseases
- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
  - E00-E07 Disorders of thyroid gland
  - E10-E14 Diabetes mellitus
    - E10 Type 1 diabetes mellitus
    - E11 Type 2 diabetes mellitus
    - E12 Malnutrition-related diabetes mellitus
    - E13 Other specified diabetes mellitus
    - E14 Unspecified diabetes mellitus
  - E15-E16 Other disorders of glucose regulation and pancreatic internal secretion
  - E20-E35 Disorders of other endocrine glands
  - E40-E46 Malnutrition
  - E50-E64 Other nutritional deficiencies
  - E65-E68 Obesity and other hyperalimentation
  - E70-E90 Metabolic disorders
- V Mental and behavioural disorders
- VI Diseases of the nervous system
- VII Diseases of the eye and adnexa
- VIII Diseases of the ear and mastoid process
- IX Diseases of the circulatory system
- X Diseases of the respiratory system
- XI Diseases of the digestive system
- XII Diseases of the skin and subcutaneous tissue
- XIII Diseases of the musculoskeletal system and connective

.9 Without complications

E10 Type 1 diabetes mellitus

[See before E10 for subdivisions]

Incl.: diabetes (mellitus):
- brittle
- juvenile-onset
- ketosis-prone

Excl.: diabetes mellitus (in):
- malnutrition-related (E12.-)
- neonatal (P70.2)
- pregnancy, childbirth and the puerperium (O24.-)

glycosuria:
- NOS (R81)
- renal (E74.8)

impaired glucose tolerance (R73.0)
postsurgical hypoinsulinaemia (E89.1)

E11 Type 2 diabetes mellitus

[See before E10 for subdivisions]

Incl.: diabetes (mellitus)(nonobese)(obese):
- adult-onset
- maturity-onset
- nonketotic
- stable

non-insulin-dependent diabetes of the young

Excl.: diabetes mellitus (in):
- malnutrition-related (E12.-)
- neonatal (P70.2)
- pregnancy, childbirth and the puerperium (O24.-)

glycosuria:
- NOS (R81)
- renal (E74.8)

impaired glucose tolerance (R73.0)

ICD Classification

1. Certain infectious and parasitic diseases
2. Neoplasms
3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4. Endocrine, nutritional and metabolic diseases
5. Mental and behavioural disorders
6. Diseases of the nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system
10. Diseases of the respiratory system
11. Diseases of the digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth and the puerperium
16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations and chromosomal abnormalities
18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19. ...

Medical Knowledge Resources

Existing Medical thesauri (in English)

3. Systematized Nomenclature of Medicine (SNOMED): thesaurus designed to process clinical data
4. Cumulative Index to Nursing and Allied Health Literature (CINAHL): classical medical concepts + domain-specific ones
5. EMTREE: European MeSH, used to index EMBASE
6. PsycINFO: psychology and psychiatry thesaurus
7. Gene Ontology: description of biomolecular biology (molecular functions, biological processes, cellular components) - designed to structure the knowledge rather than index content
8. National Cancer Institute (NCI) thesaurus: knowledge model enabling cross-disciplinary communication and collaboration

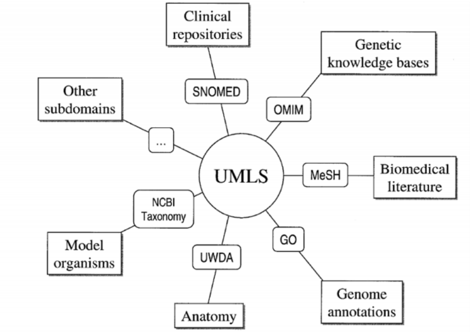
Many thesauri are also available in many well-endowed languages.

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)

- Purpose: provide a mecanism to link existing medical thesaurus and controlled vocabularies
- Initiated in 1986 and maintained by the National Library of Medicine
- Contains: a metathesaurus, a semantic network, NLP tools
- Gathers more than 100 thesauri/vocabulary



Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32, D267-D270.*

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)

- All the entries corresponding to the same thing are registered as a single concept
- To each concept correspond several terms that represent an expression of the concept (not lexical variants)
- For each term are listed the lexical variants, and each string is linked to the atom matching the concept entry
- For each concept, a preferred term and string are given
- Each concept, term, string and atom have a unique identifier: CUI, LUI, SUI, AUI

<table>
<tr><th>Concept (CUI)</th><th>Term (LUI)</th><th>String (SUI)</th><th>Atom (AUI)</th></tr>
<tr><td rowspan="6">C0004238
Atrial fibrillation (preferred)
Atrial fibrillations
Auricular fibrillation
Auricular fibrillations</td><td rowspan="3">L0004238
Atrial fibrillation (preferred)
Atrial fibrillations</td><td rowspan="2">S0016668
Atrial fibrillation (preferred)</td><td>A0027665
Atrial fibrillation (from MSH)</td></tr>
<tr><td>A0027667
Atrial fibrillation (from PSY)</td></tr>
<tr><td>S0016669
Atrial fibrillations (plural variant)</td><td>A0027668
Atrial fibrillations (from MSH)</td></tr>
<tr><td rowspan="2">L0004327
Auricular fibrillation
Auricular fibrillations (synonyms)</td><td>S0016899
Auricular fibrillation (preferred)</td><td>A0027930
Auricular fibrillation (from PSY)</td></tr>
<tr><td>S0016900
Auricular fibrillations (plural variant)</td><td>A0027932
Auricular fibrillations (from MSH)</td></tr>
</table>

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)

Basic View | Report View | Raw View

- Concept: [C0004238] Atrial Fibrillation
- Semantic Type
- Definition
- Synonyms (96)
- Relations (1672) REL | RELA | RSAB| String | CUI

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)

Basic View | Report View | Raw View

⊟ Concept: [C0004238] Atrial Fibrillation

DA Date 1990-09-30 05:00:00.000000000

MR Major Revision Date 2017-09-14 06:00:00.000000000

ST Status R

⊟ Semantic Type

Disease or Syndrome [T047]

⊞ Definition

⊞ Synonyms (96)

⊞ Relations (1672) REL | RELA | RSAB| String | CUI

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)

Basic View | Report View | Raw View

Concept: [C0004238] Atrial Fibrillation

Semantic Type

Definition

CHV/null - rapid tremor and shake of upper chambers of the heart

CSP/null - disorder of cardiac rhythm characterized by rapid, irregular atrial impulses and ineffective atrial contractions.

HPO/null - An atrial arrhythmia characterized by disorganized atrial activity without discrete P waves on the surface EKG, but instead by an undulating baseline or more sharply circumscribed atrial deflections of varying amplitude an frequency ranging from 350 to 600 per minute. [HPO:probinson]

MEDLINEPLUS/null -

An arrhythmia is a problem with the speed or rhythm of the heartbeat. Atrial fibrillation (AF) is the most common type of arrhythmia. The cause is a disorder in the heart's electrical system.

Often, people who have AF may not even feel symptoms. But you may feel

- Palpitations -- an abnormal rapid heartbeat
- Shortness of breath
- Weakness or difficulty exercising
- Chest pain
- Dizziness or fainting
- Fatigue
- Confusion

AF can lead to an increased risk of stroke. In many patients, it can also cause chest pain, heart attack, or heart failure.

Doctors diagnose AF using family and medical history, a physical exam, and a test called an electrocardiogram (EKG), which looks at the electrical waves your heart makes. Treatments include medicines and procedures to restore normal rhythm.

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)

Basic View | Report View | Raw View

- Concept: [C0004238] Atrial Fibrillation
- Semantic Type
- Definition
- Synonyms (96)
  - ACFA (arythmie complète par fibrillation auriculaire)
  - AF
  - AF - Atrial fibrillation
  - AFib
  - ATRIAL FIBRILLATION
  - ATRIJ, FIBRILACIJA
  - AURICULAR FIBRILLATION
  - AURICULAR, FIBRILACION
  - Afib
  - Atrial Fibrillation
  - Atrial Fibrillation [Disease/Finding]
  - Atrial Fibrillations
  - Atrial fibrillation
  - Atrial fibrillation (disorder)
  - Atrieflimmer
  - Atriumfibrillatie
  - Auricular Fibrillation
  - Auricular Fibrillations

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)

Basic View | Report View | Raw View

⊞ Concept: [C0004238] Atrial Fibrillation

⊞ Semantic Type

⊞ Definition

⊞ Synonyms (96)

⊟ Relations (1672) REL | RELA | RSAB| String | CUI

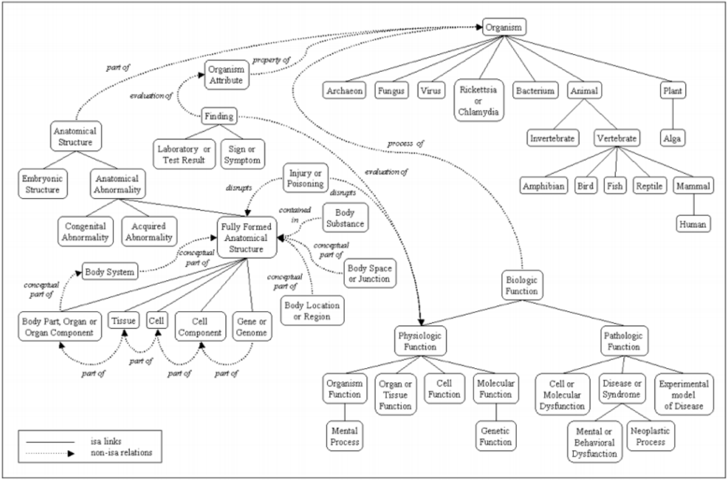
[: 1 - 10 : »]

AQ | | MSH | In Blood | C0005768
AQ | | MSH | In Cerebrospinal Fluid | C0007807
AQ | | MSH | chemically induced | C0007994
AQ | | MSH | Taxonomic | C0008903
AQ | | MSH | Congenital MeSH qualifier | C0009678
AQ | | MSH | nutritional management | C0012160
AQ | | MSH | pharmacotherapeutic | C0013217
AQ | | MSH | Economic | C0013557
AQ | | MSH | embryologic | C0013943
AQ | | MSH | enzymology | C0014445

Medical Knowledge Resources

Existing Medical thesauri (in English)

The Unified Medical Language System (UMLS)



http://genome.tugraz.at/MedicalInformatics/WinterSemester2011Holzinger/3_
LV444_152_STRUCTURED_DATA_HOLZINGER_2011.pdf

Semantic Annotation

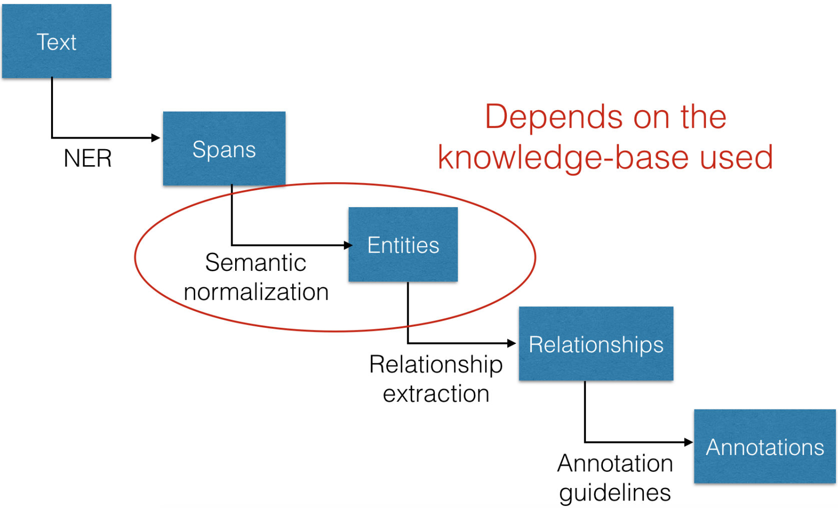
Semantic annotation consists in linking documents to knowledge bases by identifying:

- Entities/concepts (in the document)
- Relationships (implicit or explicit):
  - Between entities: *HK1 involved in glycolytic process*
  - Between an entity and the document:
    - MeSH entities for indexing documents on MEDLINE: *PMID:3207429 is indexed with* `Glucose/metabolism` *and* `Hexokinase/genetics`
    - The ICD code in a medical report

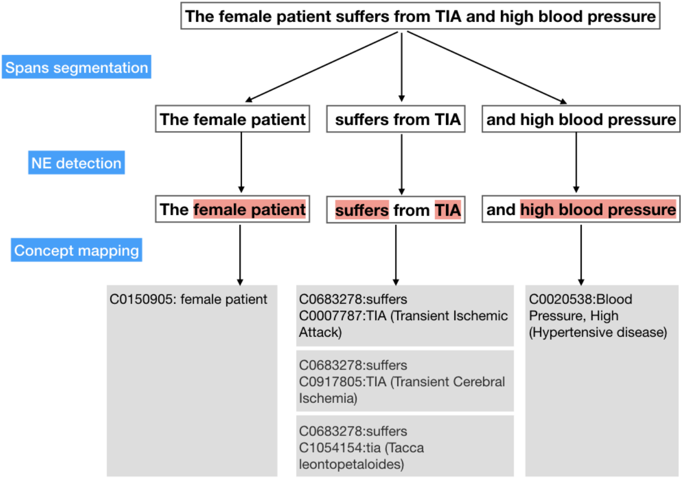
Semantic annotation can be done:

- Automatically
  - Named Entity Recognition and normalization
  - Automatic indexing
- Manually
  - Data curation, manual labelling
  - Manual indexing (majority on MEDLINE)
  - Codes added to patients reports when billing
- Derived from other annotations
  - Using links between concepts and entities

Semantic Annotation



Semantic Annotation



Existing Tools

Metamap [Aronson and Lang, 2010]

- Semantic annotation tool created by the National Library of Medicine
- Designed to annotate medical literature
- Requires a UMLS license
- Interactive, Web API or batch versions available
- `https://metamap.nlm.nih.gov/`

QuickUMLS [Soldaini and Goharian, 2016]

- Semantic annotation tool (written in Python)
- Fast, unsupervised biomedical concept extraction from medical text
- Requires a UMLS license
- `https://github.com/Georgetown-IR-Lab/QuickUMLS`

cTakes [Savova et al., 2010]

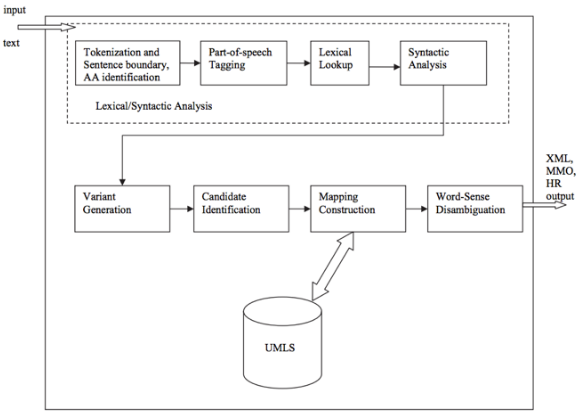
- UIMA-based NLP system for extraction of information from EHR
- Java based library of NLP and information extraction components
- Requires a UMLS license
- `https://ctakes.apache.org/`

SemRep [Rindflesch and Fiszman, 2003]

- Semantic relations extractor from biomedical text
- UMLS-based program extracting three-part propositions (semantic predications)
- Requires a UMLS license
- Batch or interactive mode
- `https://semrep.nlm.nih.gov/`

METAMAP

Figure 1 MetaMap system diagram. HR, human readable; MMO, MetaMap machine output; UMLS, unified medical language system.



Extracted from [Aronson and Lang, 2010]

Semantic Annotation

MetaMap

Input Text:

```
The female patient suffers from TIA and high blood pressure
```

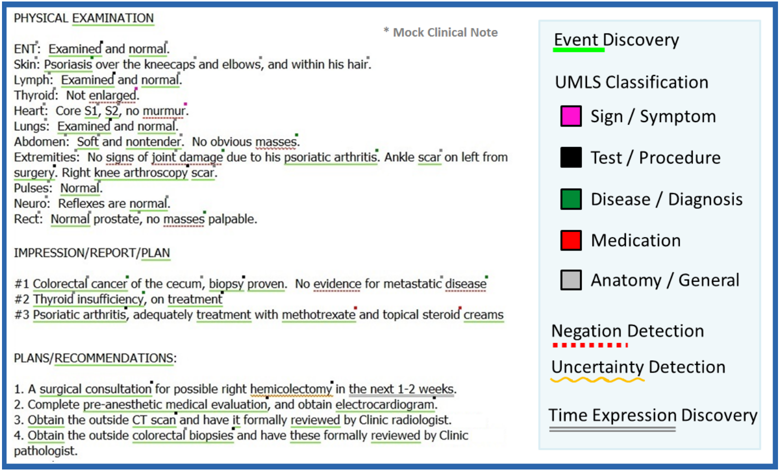
Results:

```
Processing 00000000.tx.1: The female patient suffers from TIA and high blood pressure

Phrase: The female patient
>>>>> Phrase
female patient
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
  1000   C0150905:female patient (patient is female) [Organism Attribute]
<<<<< Mappings

Phrase: suffers from TIA
>>>>> Phrase
suffers from tia
<<<<< Phrase
>>>>> Mappings
Meta Mapping (746):
   790   C0683278:suffers (Mental Suffering) [Mental or Behavioral Dysfunction]
   790   C0007787:TIA (Transient Ischemic Attack) [Disease or Syndrome]
Meta Mapping (746):
   790   C0683278:suffers (Mental Suffering) [Mental or Behavioral Dysfunction]
   790   C0917805:TIA (Transient Cerebral Ischemia) [Disease or Syndrome]
Meta Mapping (746):
   790   C0683278:suffers (Mental Suffering) [Mental or Behavioral Dysfunction]
   790   C1054154:tia (Tacca leontopetaloides) [Plant]
<<<<< Mappings

Phrase: and
>>>>> Phrase
<<<<< Phrase

Phrase: high blood pressure
>>>>> Phrase
high blood pressure
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
  1000   C0020538:Blood Pressure, High (Hypertensive disease) [Disease or Syndrome]
Meta Mapping (1000):
  1000   C2926615:High blood pressure (Ever told by doctor or nurse that you have high blood pressure:Finding:Point in time:^Patient:Ordinal) [Clinical Attribute]
<<<<< Mappings
```

Screenshot from the MetaMap interactive tool

CTAKES

Apache cTakes is a UIMA-based natural language processing system for extraction of information from electronic medical record clinical free-text:


PHYSICAL EXAMINATION

* Mock Clinical Note

ENT: Examined and normal.
Skin: Psoriasis over the kneecaps and elbows, and within his hair.
Lymph: Examined and normal.
Thyroid: Not enlarged.
Heart: Core S1, S2, no murmur.
Lungs: Examined and normal.
Abdomen: Soft and nontender. No obvious masses.
Extremities: No signs of joint damage due to his psoriatic arthritis. Ankle scar on left from surgery. Right knee arthroscopy scar.
Pulses: Normal.
Neuro: Reflexes are normal.
Rect: Normal prostate, no masses palpable.

IMPRESSION/REPORT/PLAN

#1 Colorectal cancer of the cecum, biopsy proven. No evidence for metastatic disease
#2 Thyroid insufficiency, on treatment
#3 Psoriatic arthritis, adequately treatment with methotrexate and topical steroid creams

PLANS/RECOMMENDATIONS:

1. A surgical consultation for possible right hemicolectomy in the next 1-2 weeks.
2. Complete pre-anesthetic medical evaluation, and obtain electrocardiogram.
3. Obtain the outside CT scan and have it formally reviewed by Clinic radiologist.
4. Obtain the outside colorectal biopsies and have these formally reviewed by Clinic pathologist.

Exctracted from `https://ctakes.apache.org/`

QuickUMLS

- A faster option to identify UMLS concepts from unstructured documents
- Gives comparable accuracy on several datasets

Method		*Prec*	*Rec*	*F-1*	*ms/doc*
MetaMap		0.49*	0.48*	0.48*	19,295*
cTAKES		**0.71**	0.53*	0.61	3,852*
QuickUMLS	$\alpha = 0.6$	0.50*	**0.75**	0.60	1,594*
	$\alpha = 0.7$	0.60*	0.66*	**0.63**	680*
	$\alpha = 0.8$	0.63*	0.60*	0.61	332*
	$\alpha = 0.9$	0.64*	0.56*	0.60	193*
	$\alpha = 1.0$	0.67*	0.54*	0.60	**143**

Table 1: Results for the i2b2 dataset. cTAKES outperforms QuickUMLS in precision, but QuickUMLS has better recall. QuickUMLS is 2.5 to 135 times faster than MetaMap or cTAKES. * indicates statistically significant differences from best value (Welch's t-test, $p < 0.05$).

Method		*Prec*	*Rec*	*F-1*	*ms/doc*
MetaMap		0.71*	0.53*	0.61*	15,935
cTAKES		**0.89**	0.55*	0.68*	3,765*
QuickUMLS	$\alpha = 0.6$	0.68*	**0.77**	**0.72**	1,536*
	$\alpha = 0.7$	0.78*	0.67*	**0.72**	646*
	$\alpha = 0.8$	0.83*	0.61*	$0.70^{\dagger}$	340*
	$\alpha = 0.9$	0.85*	0.57*	0.68*	174*
	$\alpha = 1.0$	0.87*	0.55*	0.67*	**141**

Table 2: Results for the THYME corpus. cTAKES achieves the best precision, QuickUMLS the best recall and substantially better throughput than MetaMap or cTAKES. * indicates statistically significant differences from best value (Welch's t-test, $p < 0.05$). † indicates statistical significance from α= 0.6 but not α= 0.7.

SUMMARY

- Many types of data :
  - Patient-specific data
  - Knowledge-based data
  - User-generated data
- Various information needs: search for information, for cohorts, for evidence...
- On the solution side:
  - Some structured knowledge bases: UMLS and all its constituents (and their versions in languages other than English)
  - Several concept annotation tools: MetaMap, cTakes, QuickUMLS...
  - And IR models

Summary of issues

- **Semantic gap**
  - **Vocabulary mismatch between experts and novices**
  - **Vocabulary mismatch between the query and documents**
  - **Interpretation of vocabulary in context: negation, lexical variants, time, task peculiarities**
- Result appraisal
  - Understand medical language
  - Ability to interpret the results, make accurate inference, assess the credibility
  - Cognitive bias

Outline

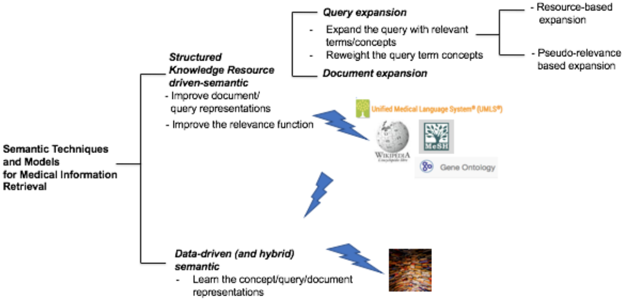
Roadmap

- Traditional IR
  - Q: bag of words
  - D: bag of words
  - RSV(Q,D): Alignment of Q and D
- Semantic (medical) IR
  - Q:
    - Bag of words
    - Bag of words **and concepts/entities**
    - **Embeddings**
  - D:
    - Bag of words
    - Bag of words **and concepts/entities**
    - **Embeddings**
  - RSV(Q,D): **Semantic inference**

ROADMAP

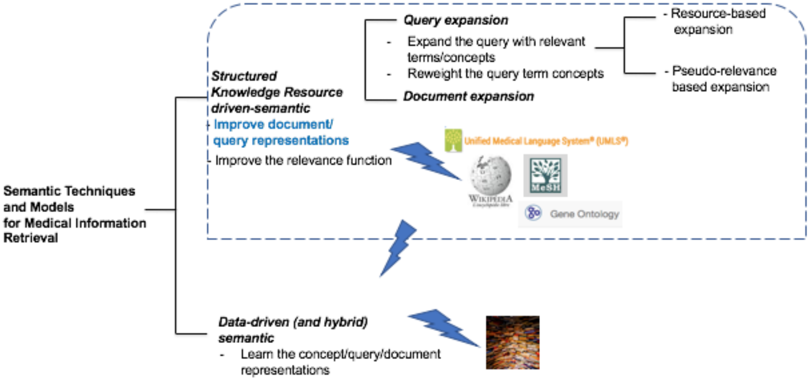
- Traditional IR
  - Q: bag of words
  - D: bag of words
  - RSV(Q,D): Alignment of Q and D

- Semantic (medical) IR
  - Q:
    - Bag of words
    - Bag of words **and concepts/entities**
    - **Embeddings**
  - D:
    - Bag of words
    - Bag of words **and concepts/entities**
    - **Embeddings**
  - RSV(Q,D): **Semantic inference**



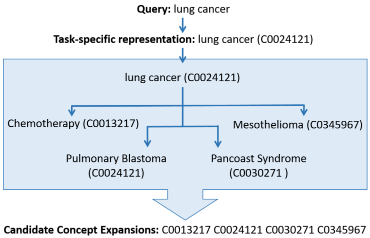
Outline

ROADMAP


Semantic Techniques and Models for Medical Information Retrieval
Structured Knowledge Resource driven-semantic
Improve document/ query representations
Improve the relevance function
Query expansion
- Expand the query with relevant terms/concepts
- Reweight the query term concepts
Document expansion
- Resource-based expansion
- Pseudo-relevance based expansion
Unified Medical Language System® (UMLS®)
WIKIPEDIA
MeSH
Gene Ontology
Data-driven (and hybrid) semantic
- Learn the concept/query/document representations

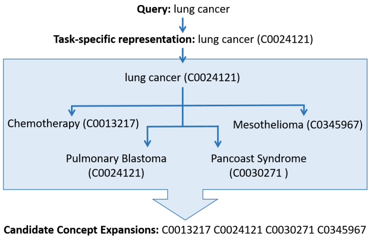
QUERY/DOCUMENT EXPANSION

- Query/document expansion
  - Enhance the Query/Document using:
    - evidence from related words/terms in semantic resources;
    - relevance feedback signals

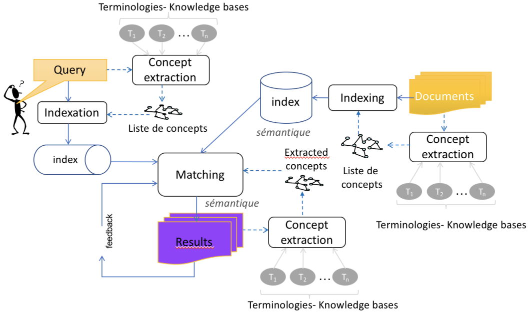


Query/document expansion

- Query/document expansion
  - Enhance the Query/Document using:
    - evidence from related words/terms in semantic resources;
    - relevance feedback signals



  - The full framework



Query/document expansion

- Main impacting factors: [Dinh et al., 2013, Zuccon and Koopman, 2018]
  - Which knowledge-base to use (specialized vs. generic) and how many?
  - Which context to use (global vs. local)?
  - How to select candidate expansion terms and (how to inject them in a retrieval model) ?

Query/document expansion

- Main impacting factors: [Dinh et al., 2013, Zuccon and Koopman, 2018]
  - Which knowledge-base to use (specialized vs. generic) and how many?
  - Which context to use (global vs. local)?
  - How to select candidate expansion terms and (how to inject them in a retrieval model) ?

- Resulting techniques
  - LSMo: Local context, Specialized Mono-Resource [Soldaini et al., 2017, Sondhi et al., 2012]
  - GSMo: Global context, Specialized Mono-Resource [Martinez et al., 2014, Znaidi et al., 2016]
  - LGSMo: Local and Global contexts, Specialized Mono-Resource [Wang and Akella, 2015, Znaidi et al., 2015, Znaidi et al., 2016]
  - GSGMu: Global context, Specialized General Multiple-Resources [Soldaini et al., 2016]
  - LGSM : Local and Global contexts, Specialized Multiple-Resources [Limsopatham et al., 2013, Dinh and Tamine, 2012, Oh and Jung, 2015, Zhu and Carterette, 2012]

		LSMo	GSMo	LGSMo	GSGMu	LGSMu
Context	Local (Pseudo-relevance)	■	□	■	□	■
	Global (Resource)	□	■	■	■	■
Knowledge Base	Specialized	■	■	■	■	■
	General	□	□	□	■	□
	Mono-resource	■	■	■	□	□
	Multiple-resources	□	□	□	■	■

Query/document expansion

Local context, one specialized resource [Sondhi et al., 2012]

- **Context:** Top N retrieved documents
- **Knowledge-Base**: MeSH thesaurus

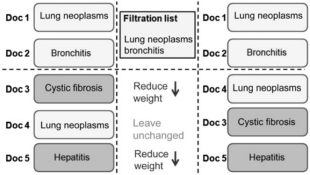
- **Key steps**
  - Map query words to UMLS semantic types and assign weights $c^{'}(w, Q) = c(w, Q)$ if w belong to a relevant type eg., disease, syndrome, body, etc.
  - **Top-N based MeSH feedback**: identify a list of potential diagnoses from N top documents and then rerank the documents w.r.t absence of potential diseases
  - **Distribution-based MeSH feedback**: For each MeSH term, identify all the documents indexed with it, pick the M highest scoring MeSH terms as candidate term expansion
  - Expand the query and then perform a pseudo-relevance feedback based model (PRF) [Zhai and Lafferty, 2001]
    $D(\hat{\theta}_q \parallel \hat{\theta}_d) = -\sum_w p(w \mid \hat{\theta}_q) log\, p(w \mid \hat{\theta_d}) + cons(q)$

QUERY/DOCUMENT EXPANSION

LOCAL CONTEXT, ONE SPECIALIZED RESOURCE [SONDHI ET AL., 2012]

- **Context:** Top N retrieved documents
- **Knowledge-Base**: MeSH thesaurus

- **Key steps**
  - Map query words to UMLS semantic types and assign weights $c'(w, Q) = c(w, Q)$ if w belong to a relevant type eg., disease, syndrome, body, etc.
  - **Top-N based MeSH feedback**: identify a list of potential diagnoses from N top documents and then rerank the documents w.r.t absence of potential diseases
  - **Distribution-based MeSH feedback**: For each MeSH term, identify all the documents indexed with it, pick the M highest scoring MeSH terms as candidate term expansion
  - Expand the query and then perform a pseudo-relevance feedback based model (PRF) [Zhai and Lafferty, 2001]
    $D(\hat{\theta}_q \parallel \hat{\theta}_d) = -\sum_w p(w \mid \hat{\theta}_q) log\, p(w \mid \hat{\theta}_d) + cons(q)$



- **Main results/findings**
  - Slight improvements (more than 6%) over the baseline for the Distribution-based MeSH feedback while the top N based Mesh feedback is worse than the baseline using small datasets (19 queries, 5585 documents)
  - Difficulty in recovering new treatments and rare alternative diseases
  - Confusion between similar conditions/diseases

Query/document expansion

Local context, one general resource [Soldaini et al., 2017]

- **Context:** Top N retrieved documents
- **Knowledge-Base**: Wikipedia
- **Key steps: Health Terms Pseudo Relevance Feedback HTPRF**
  - Retrieve the N Top documents w.r.t query Q
  - For each term from the top N documents, compute a score $s_j = log_{10}(10 + w_j)$ $w_j = \alpha * tf(t_j, Q) + \frac{\beta}{N} \sum_1^N (tf(t_j, D_i) * idf(t_j))$
  - Select the top M terms with the highest score as the candidate expansion terms
  - For each candidate term expansion, compute the likelihood of being health-related. Compute the odds ratio as the proportion of health-related Wikipdia (W_H) documents including term t_i $OR(t_i) = \frac{n(t_i, W_H)}{n(t_i, W)}$
  - Consider the top M ranked terms with score $OR(t_j) > \sigma$
  - Expand the query and perform a pseudo-relevance feedback based model

Query/document expansion

Local context, one general resource [Soldaini et al., 2017]

- **Context:** Top N retrieved documents
- **Knowledge-Base**: Wikipedia

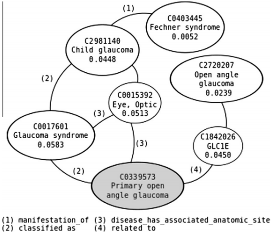
- **Key steps: Health Terms Pseudo Relevance Feedback HTPRF**
  - Retrieve the N Top documents w.r.t query Q
  - For each term from the top N documents, compute a score
    $s_j = log_{10}(10 + w_j)\, w_j = \alpha * tf(t_j, Q) + \frac{\beta}{N} \sum_1^N (tf(t_j, D_i) * idf(t_j))$
  - Select the top M terms with the highest score as the candidate expansion terms
  - For each candidate term expansion, compute the likelihood of being health-related. Compute the odds ratio as the proportion of health-related Wikipdia (W_H) documents including term t_i
    $OR(t_i) = \frac{n(t_i, W_H)}{n(t_i, W)}$
  - Consider the top M ranked terms with score $OR(t_j) > \sigma$
  - Expand the query and perform a pseudo-relevance feedback based model

- **Main results/findings**
  - Mapping the wikipedia terms to UMLS semantic types revealed that 75% are present in the UMLS: 32% are symptoms, 20,3% are treatments, 18% are a diagnosis procedure or test,17,1% are diseases
  - The HTPRF parameters do not significantly impact the results
  - Precision oriented with slight improvement (+3,6%) over state-of the best systems in TREC CDS 2014-TREC CDS 2015

Query/document expansion

Global context, one specialized resource [Martinez et al., 2014]

- **Context:** Concepts and relations between concepts
- **Knowledge-Base**: UMLS thesaurus

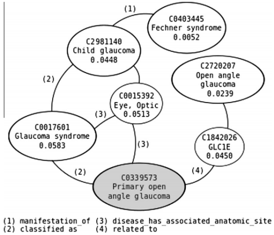


- **Key steps**
  - Map query words to UMLS semantic types
  - Identify the initial sub-graph based concept including query concepts and related UMLS concepts
  - Assign an uniform probability to the concepts in the sub-graph and then run the Page Rank algorithm
  - Rank the concepts using the Page Rank score
  - Expand the query with the N concepts having the highest PageRank Score
  - Perform a basic retrieval model (eg., TF-IDF, BM25)

Query/document expansion

Global context, one specialized resource [Martinez et al., 2014]

- **Context:** Concepts and relations between concepts
- **Knowledge-Base**: UMLS thesaurus



- **Key steps**
  - Map query words to UMLS semantic types
  - Identify the initial sub-graph based concept including query concepts and related UMLS concepts
  - Assign an uniform probability to the concepts in the sub-graph and then run the Page Rank algorithm
  - Rank the concepts using the Page Rank score
  - Expand the query with the N concepts having the highest PageRank Score
  - Perform a basic retrieval model (eg., TF-IDF, BM25)

- **Main results/findings**
  - Experiments on TREC medical records 2011-2012 show significant improvements (+30% in average)
  - Expansion terms are those related to the query with either taxonomic (eg., synonyms) and not taxonomic (eg., disease has associated anatomic site).
  - Useful expansion in the case of a cohort retrieval task.

Queries with highest improvement for PageRank, together with the learnt expansion terms and the Bpref increase.

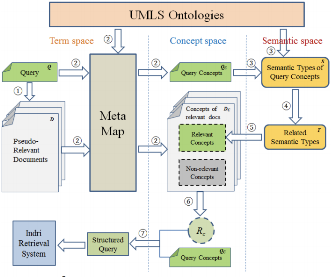
Query	TREC version	Expansion terms	Bpref increase
Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis	2011	MRSA elsewhere/NOS Personal history of poliomyelitis Personal history of other infectious and parasitic disease	0.931
Patients with Primary Open Angle Glaucoma (POAG)	2012	Eye, Eyeball, Globe, Ocular... Glaucoma syndrome Open cleft glaucoma GLC1E	0.742
Patients with adult	2012	Immunology	0.722

Query/document expansion

Global and local contexts, one specialized resource [Wang and Akella, 2015]

- **Key steps**
  - Map query Q and pseudo-relevant document words to UMLS semantic types
  - Build the document model using the distribution of semantic type s in the set of relevant document D_C: $p(c \mid D_C) = \sum_s p(s \mid D_C) P(c \mid s, D_C)$
  - Compute $P(c \mid s, D_C)$ by the maximum likelihood estimation of c in this semantic type, and then smoothed by the collection background model
  - Computed $p(c \mid R_c)$ with R_c the unknown concept relevance model by estimating the joint probability of observing the concept c together with query concepts

    $$p(c \mid R_C) = p(c \mid q_{C1} \ldots q_{Ck}) = \frac{p(c, q_{C1} \cdots q_{Ck})}{(q_{C1} \cdots q_{Ck})}$$
  - Rank the concepts using the $p(c \mid R_c)$ score
  - Expand the query with the N concepts having the highest score
    $p(c \mid Q'_c) = (1 - \lambda) * p(c \mid D_c) + \lambda * p(c \mid R_c)$

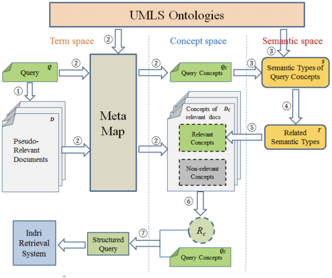


QUERY/DOCUMENT EXPANSION

GLOBAL AND LOCAL CONTEXTS, ONE SPECIALIZED RESOURCE [WANG AND AKELLA, 2015]

- **Key steps**
  - Map query Q and pseudo-relevant document words to UMLS semantic types
  - Build the document model using the distribution of semantic type s in the set of relevant document D_C: $p(c \mid D_C) = \sum_s p(s \mid D_C) P(c \mid s, D_C)$
  - Compute $P(c \mid s, D_C)$ by the maximum likelihood estimation of c in this semantic type, and then smoothed by the collection background model
  - Computed $p(c \mid R_c)$ with R_c the unknown concept relevance model by estimating the joint probability of observing the concept c together with query concepts

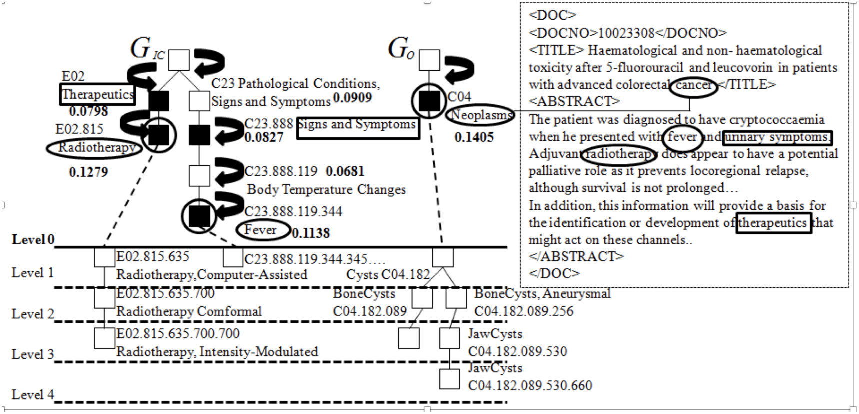
    $$p(c \mid R_C) = p(c \mid q_{C1} \dots q_{Ck}) = \frac{p(c, q_{C1} \cdots q_{Ck})}{(q_{C1} \cdots q_{Ck})}$$
  - Rank the concepts using the $p(c \mid R_c)$ score
  - Expand the query with the N concepts having the highest score $p(c \mid Q'_c) = (1-\lambda) * p(c \mid D_c) + \lambda * p(c \mid R_c)$
- **Main results/findings**
  - Working on concepts leads in a dense space for relevance estimation which is expected to produce better estimates
  - Significant improvement particularly for long documents containing sufficient evidence to infer concept relevance estimates.



Query/document expansion

Global and local contexts, one specialized resource [Znaidi et al., 2016]

- **Context:** Concepts and relations between concepts, Top N retrieved documents
- **Knowledge-Base**: UMLS thesaurus



Query/document expansion

Global and local contexts, one specialized resource [Znaidi et al., 2016]

- **Context:** Concepts and relations between concepts, Top N retrieved documents
- **Knowledge-Base**: UMLS thesaurus

- **Key steps: tailored for PICO queries**
  - Map each PICO facet of query Q to UMLS
  - Identify the UMLS query concepts for each facet
  - Build a concept-based tree for each facet
  - Build the candidate query expansion terms using a recursive propagation algorithm from active concept c through sub-concepts $csub$
    - Compute $score(c) = +level(csub) * score(s)$
  - Expand the query with the N concepts having the highest score
  - Apply a prioritized-scoring based retrieval

$$RSV_{PICO}(Q,d) = \lambda_P * RSV_P(Q_P,d) + \lambda_{IC} * RSV_{IC}(Q_{IC},d) + \lambda_O * RSV_O(Q_O,d) \quad l$$

$\lambda_{IC} = 1 \lambda_P = \lambda_{IC} * RSV(Q^c_{IC},d), \lambda_O = \lambda_P * RSV(Q^c_P,d)$

  - Combine the PICO score with a classic score (eg., BM25)

Algorithm 1: Main

Require: $Q, Q_{PICO}, T, N_d, N_c, MaxLevel$
Ensure: Q^c_P, Q^c_{IC}, Q^c_O
BEGIN
{Initial search}
$D^*_N \leftarrow Top_D(Q, N_d, C)$;
{Query Graph Building}
$Q_P \leftarrow Substr(Q, P)$;
$Q_{IC} \leftarrow Substr(Q, IC)$;
$Q_O \leftarrow Substr(Q, O)$;
$Concepts(Q_P) \leftarrow Extract(Q_P, T)$;
$Concepts(Q_{IC}) \leftarrow Extract(Q_{IC}, T)$;
$Concepts(Q_O) \leftarrow Extract(Q_O, T)$;
$G_P \leftarrow HypG(Concepts(Q_P), T)$;
$G_{IC} \leftarrow HypG(Concepts(Q_{IC}), T)$;
$G_O \leftarrow HypG(Concepts(Q_O), T)$;
$Q^c_P \leftarrow CSelect(G_P)$;
$Q^c_{IC} \leftarrow CSelect(G_{IC})$;
$Q^c_O \leftarrow CSelect(G_O)$;
END

Algorithm 2: Function: *CSelect*

Require: G_x
Ensure: *TopConcepts*
BEGIN
{Process the top ranked documents}
for all $d \in D^*_N$ **do**
  {Extraction of document concepts}
  $TopConcepts \leftarrow Extract(d, G_x)$;
  $level \leftarrow 0$;
  {Score Propagation}
  for all $c \in TopConcepts$ AND $level < Maxlevel$ **do**
    for all $csub \in Hypo(c, G_x)$ **do**
      $Score(csub) \leftarrow (Score(csub) + Lev(csub) * Score(c))$;
      $Score(csub) \leftarrow Normalized(Score(csub))$;
      $level \leftarrow level + 1$;
    end for
  end for
end for
$TopConcepts \leftarrow Top_C(G_x, N_c)$;
return *TopConcepts*;
END

Query/document expansion

Global and local contexts, one specialized resources [Znaidi et al., 2016]

- **Context:** Concepts and relations between concepts, Top N retrieved documents
- **Knowledge-Base**: UMLS thesaurus

Model	Precision			% Change	# RR
	MAP	P@5	P@10		
BM25	0.112	0.1561	0.127	+51.42%**	4574
LM	0.111	0.156	0.130	+52.36%***	4491
PSBM25	0.123	0.151	0.139	+37.94%**	4904
AGM	0.121	0.148	0.135	+40.09%**	4835
PLM	0.163	0.240	–	+4.60%	5770
PSM	**0.170**	**0.254**	**0.198**	–	**5894**

- **Mains findings**
  - Significant improvements (+30%) on a standard PICO collection
  - Slight but non significant improvement in comparison to PICO tailored models
  - The model performs better for relatively short queries (still be long in comparison to classic queries)
  - For long queries, it is more likely that the document matches the query according to different facets but with misleading interpretations of the search intent.

Q	Id	Desc	% Change	#W	#C
R^+	**M6.2**	(**P**) In obese patients diabetes(**/P**)**(IC)** orlistat Placebo(**/IC**) (**O**)Weight loss(**/O**)	+78.08%	7	4
	Q37.2	(**P**) Adults 18 years or more migraine(**/P**)**(IC)**aspirin plus an antiemetic placebo (**/IC**) **O**)pain free(**/O**).	+74.60%	8	5
R^-	**C21.1**	(**P**) Adults 14 years and older GORD (**/P**)**(IC)**Medical management: proton pump inhibitors/histamine receptor antagonists Laparoscopic fundoplication surgery(**/IC**)(**O**) Health-related quality of life (**/O**)	−45.07%	18	7
	Q48.3	(**P**) Adults 18 years or older rheumatoid arthritis (**/P**)**(IC)** methotrexate combined with other non-biologic disease modifying anti-rheumatic drugs (DMARDs) methotrexate alone(**/IC**)(**O**)ACR response of non-MTX DMARDS inadequate response(**/O**).	−83.21%	26	6

Query/document expansion

Global and local contexts, multiple specialized resource [Dinh and Tamine, 2012]

- **Context:** Concepts and relations between concepts, Top N retrieved documents
- **Knowledge-Base**: MeSH, SNOMED, GO, ICD10 thesaurus,

- **Key steps**
  - Document expansion (indexing level)
    - Map each document to N resources and extract N lists of M top ranked candidate concepts
    - Apply a fusion strategy to select the best expansion concepts *C*
    - Expand the document with concepts in *C*

Stage 1 Conceptual Document Indexing

Input: Collection C, Terminologies T
Output: Index I

```
for all document D in C do
  # Mono-terminology extraction
  for all terminology T_i in T do
    R(D, T_i) ← extract(D, T_i);
  end for
  # Concept fusion
  R(D, T) ← ∪_{i=1}^{n} R(D, T_i);
  # Document expansion
  D' ← expand(D, R(D, T));
  # Document indexing
  I ← addIndex(D');
end for
return I;
```

Category	Technique	score(c_j, D)
Rank-based	CombRank	$\sum_{i=1}^{n}(\|R(D,T_i)\| - r_{ji}^{D})$
	CombRCP	$\sum_{i=1}^{n} 1/r_{ji}^{D}$
Score-based	CombSUM	$\sum_{i=1}^{n} w_{ji}^{D}$
	CombMIN	$\min\{w_{ji}^{D}, \forall i = \overline{1..n}\}$
	CombMAX	$\max\{w_{ji}^{D}, \forall i = \overline{1..n}\}$
	CombMED	$median\{w_{ji}^{D}, \forall i = \overline{1..n}\}$
	CombANZ	$\sum_{i=1}^{n} w_{ji}^{D} \div \|\{c_j \in R(D,T)\}\|$
	CombMNZ	$\sum_{i=1}^{n} w_{ji}^{D} \times \|\{c_j \in R(D,T)\}\|$

Query/document expansion

Global and local contexts, multiple specialized resource [Dinh and Tamine, 2012]

- **Context:** Concepts and relations between concepts, Top N retrieved documents
- **Knowledge-Base**: MeSH, SNOMED, GO, ICD10 thesaurus,

- **Key steps**
  - Document expansion (indexing level)
    - Map each document to N resources and extract N lists of M top ranked candidate concepts
    - Apply a fusion strategy to select the best expansion concepts C
    - Expand the document with concepts in C
  - Query expansion
    - Perform an initial search with query
    - Build a document relevance model; use the Bose-Einstein statistics [Amati, 2003] to weight terms in the expanded query q^e derived from the original query Q:

      $$weight(t \in Q^e) = tfq_n + \beta * \frac{\text{Info}_{\text{Bo1}}}{\text{MaxInfo}}$$
  - Perform a post- retrieval with the expanded query

Stage 1 Conceptual Document Indexing

```
Input: Collection C, Terminologies T
Output: Index I
for all document D in C do
   # Mono-terminology extraction
   for all terminology T_i in T do
      R(D,T_i) ← extract(D,T_i);
   end for
   # Concept fusion
   R(D,T) ← ∪_{i=1}^{n} R(D,T_i);
   # Document expansion
   D' ← expand(D,R(D,T));
   # Document indexing
   I ← addIndex(D');
end for
return I;
```

Category	Technique	score(c_j, D)
Rank-based	CombRank	$\sum_{i=1}^{n}(\lVert R(D,T_i)\rVert - r_{ji}^{D})$
	CombRCP	$\sum_{i=1}^{n} 1/r_{ji}^{D}$
Score-based	CombSUM	$\sum_{i=1}^{n} w_{ji}^{D}$
	CombMIN	$\min\{w_{ji}^{D}, \forall i = \overline{1..n}\}$
	CombMAX	$\max\{w_{ji}^{D}, \forall i = \overline{1..n}\}$
	CombMED	$median\{w_{ji}^{D}, \forall i = \overline{1..n}\}$
	CombANZ	$\sum_{i=1}^{n} w_{ji}^{D} \div \lVert\{c_j \in R(D,T)\}\rVert$
	CombMNZ	$\sum_{i=1}^{n} w_{ji}^{D} \times \lVert\{c_j \in R(D,T)\}\rVert$

Query/document expansion

Global and local contexts, multiple specialized resource [Dinh and Tamine, 2012]

- **Context:** Concepts and relations between concepts, Top N retrieved documents
- **Knowledge-Base**: MeSH, SNOMED, GO, ICD10 thesaurus

- **Main results/findings**
  - Multi-terminology based retrieval significantly better than mono-terminology-based retrieval but with varying and moderate improvement rates according to the fusion strategy
  - Need raises toward the weighting of each terminology according to query specificities

Run	*TREC Genomics 2004* MAP	Δ (%)	*TREC Genomics 2005* MAP	Δ (%)
Median	0.2074		0.2173	
Mono-terminology indexing and retrieval				
MeSH	0.4412†††	(+112.73)	0.2639	(+21.45)
SNOMED	0.4222†††	(+103.57)	0.2630	(+21.03)
ICD-10	0.4138†††	(+99.52)	0.2592	(+19.28)
GO	0.4408†††	(+112.54)	0.2536	(+16.71)
Multi-terminology indexing and retrieval				
CombANZ	0.4435†††	(+113.84)	0.2647	(+20.89)
CombMAX	0.4387†††	(+111.52)	0.2684†	(+23.52)
CombMED	0.4459†††	(+115.00)	0.2683†	(+23.47)
CombMIN	0.4440†††	(+114.08)	**0.2685**†	(+23.56)
CombMNZ	**0.4529**†††	(+118.37)	0.2593	(+19.33)
CombRank	0.4407†††	(+112.49)	0.2594	(+19.37)
CombRCP	0.4371†††	(+110.75)	0.2601	(+19.70)
CombSUM	0.4470†††	(+115.53)	0.2601	(+19.70)

Significant changes at $p \leq 0.05, 0.01$ and 0.001 are denoted †, †† and †††.

Query/document expansion

Local and global contexts, multiple specialized and general resources [Zhu and Carterette, 2012]

- **Context:** Medical collections
- **Knowledge-Base**: MeSH thesaurus
- **Key steps**
  - Apply a mixture of relevance model $p(w \mid \theta_Q) = \lambda_Q \frac{c(w,Q)}{|Q|} + (1-\lambda_Q) \sum_C \lambda_C p(w \mid \hat{\theta}_{Q,C})$
    Perform two main steps:
    - Identify relevant candidate expansion terms: MeSH-based terms and collection-based terms
      - Map the query to MeSH thesaurus and select the entry terms of active concepts and descendants e_i;

      - Estimate concept weight p_i as concept popularity using PubMed query log $p_i = \frac{logN_{e_i,G}}{\sum_j logN_{e_i,G}}$

      - For collections C containing full-text articles, compute:

      $p_i = p(e_i \mid \hat{\theta}_{Q,C}) = \sum_j^k exp \frac{tf(e_i,D_j)}{|D_j|} + log \frac{|C|}{df_{e_i,C}} + score(D_j, Q)$
    - Rank the lists E_{c_k} (from MesH thesaurus and each collection) of candidate expansion terms w.r.t. score p_i
    - Compute collection weight: query-adaptive collection weighting strategy by computing

      $\lambda_{c_k} = \frac{J_i^{-1}}{\sum_j J_i^{-1}}$ where $J(\theta_E \parallel \theta_Q) = \frac{1}{2}((K(\theta_E \parallel \theta_M) + K(\theta_Q \parallel \theta_M))$ where E is the expanded query provided by collection c_k.

Query/document expansion

Local and global contexts, multiple specialized and general resources [Zhu and Carterette, 2012]

- **Context:** Medical collections
- **Knowledge-Base**: MeSH thesaurus
- **Key steps**
  - Apply a mixture of relevance model $p(w \mid \theta_Q) = \lambda_Q \frac{c(w,Q)}{|Q|} + (1-\lambda_Q) \sum_C \lambda_C p(w \mid \hat{\theta}_{Q,C})$
    Perform two main steps:
    - Identify relevant candidate expansion terms: MeSH-based terms and collection-based terms
      - Map the query to MeSH thesaurus and select the entry terms of active concepts and descendants e_i;
      - Estimate concept weight p_i as concept popularity using PubMed query log $p_i = \frac{logN_{e_i,G}}{\sum_j logN_{e_i,G}}$
      - For collections C containing full-text articles, compute:
      $p_i = p(e_i \mid \hat{\theta}_{Q,C}) = \sum_j^k exp \frac{tf(e_i,D_j)}{|D_j|} + log \frac{|C|}{df_{e_i,C}} + score(D_j, Q)$
    - Rank the lists E_{c_k} (from MesH thesaurus and each collection) of candidate expansion terms w.r.t. score p_i
    - Compute collection weight: query-adaptive collection weighting strategy by computing
      $\lambda_{c_k} = \frac{J_i^{-1}}{\sum_j J_i^{-1}}$ where $J(\theta_E \parallel \theta_Q) = \frac{1}{2}((K(\theta_E \parallel \theta_M) + K(\theta_Q \parallel \theta_M))$ where E is the expanded query provided by collection c_k.
- **Main results/findings**
  - Significant improvement (+20%) over the baselines using ImageCLEF and Clue Web collections
  - Good balance between expansion terms provided by relations in MeSH thesaurus and term relations provided by the local relevance feedback corpus.
  - Using multiple collections is always better than using just a single collection.

Query/document expansion

Local and global contexts, multiple specialized and general resources [Oh and Jung, 2015]

- **Context:** Top N retrieved documents and collections
- **Knowledge-Base**: Wikipedia
- **Key steps**
  - Build a feedback model from a set of external collections by combining the expand External Expansion Model (EEM) [Weerkamp et al., 2012] and the concept of the cluster-based model (CCBM) [Liu and Croft, 2004]. Compute five components:
    - Prior collection probability $p(\theta_C)$ estimated using collection type (eg., expert-based)
    - Document relevance $p(Q \mid \theta_D)$ estimated using a retrieval model (eg., BM25)
    - Collection relevance $p(Q \mid \theta_C) = \sum_{D \in C} p(Q \mid \theta_D) p(\theta_D \mid C) \propto \frac{1}{|R_C|} \sum_{D \in R_C} p(Q \mid \theta_D)$
    - Document importance $p(D \parallel \theta_C)$: importance of document D in collection C (eg., using PageRank, recency, etc.)
    - Word probability $p(w \mid \theta_D)$: use evidence from a document model smoothed with cluster and collection models
      $p(w \mid \theta_D) = (1 - \lambda_E) \frac{c(w,D) + \mu p(w|C)}{|D| + \mu} + \lambda_E p(w \mid E) p(w \parallel \theta_D) p(Q \mid \theta_C) p(D \mid \theta_C)$
  - Compute the expanded query model based on a feedback model built upon target and external collections
    $p_{EEM}(w \mid Q) \propto \sum_{C \in E} p(Q \mid \theta_C) p(\theta_C) \sum_{D \in C} p(w \mid \theta_D) p(Q \mid \theta_D) p(D \mid \theta_C)$

QUERY/DOCUMENT EXPANSION

LOCAL AND GLOBAL CONTEXTS, MULTIPLE SPECIALIZED AND GENERAL RESOURCES [OH AND JUNG, 2015]

- **Context:** Top N retrieved documents and collections
- **Knowledge-Base**: Wikipedia
- **Key steps**
  - Build a feedback model from a set of external collections by combining the expand External Expansion Model (EEM) [Weerkamp et al., 2012] and the concept of the cluster-based model (CCBM) [Liu and Croft, 2004]. Compute five components:
    - Prior collection probability $p(\theta_C)$ estimated using collection type (eg., expert-based)
    - Document relevance $p(Q \mid \theta_D)$ estimated using a retrieval model (eg., BM25)
    - Collection relevance $p(Q \mid \theta_C) = \sum_{D \in C} p(Q \mid \theta_D) p(\theta_D \mid C) \propto \frac{1}{|R_C|} \sum_{D \in R_C} p(Q \mid \theta_D)$
    - Document importance $p(D \parallel \theta_C)$: importance of document D in collection C (eg., using PageRank, recency, etc.)
    - Word probability $p(w \mid \theta_D)$: use evidence from a document model smoothed with cluster and collection models
      $p(w \mid \theta_D) = (1 - \lambda_E) \frac{c(w,D) + \mu p(w|C)}{|D| + \mu} + \lambda_E p(w \mid E) p(w \parallel \theta_D) p(Q \mid \theta_C) p(D \mid \theta_C)$
  - Compute the expanded query model based on a feedback model built upon target and external collections
    $p_{EEM}(w \mid Q) \propto \sum_{C \in E} p(Q \mid \theta_C) p(\theta_C) \sum_{D \in C} p(w \mid \theta_D) p(Q \mid \theta_D) p(D \mid \theta_C)$
- **Main results/findings**
  - Significant improvement using TREC CDS, Clef E-Health datasets
  - The model is not very sensitive to the size of the initial feedback results
  - Small improvements using a general collection such as Wikipedia.

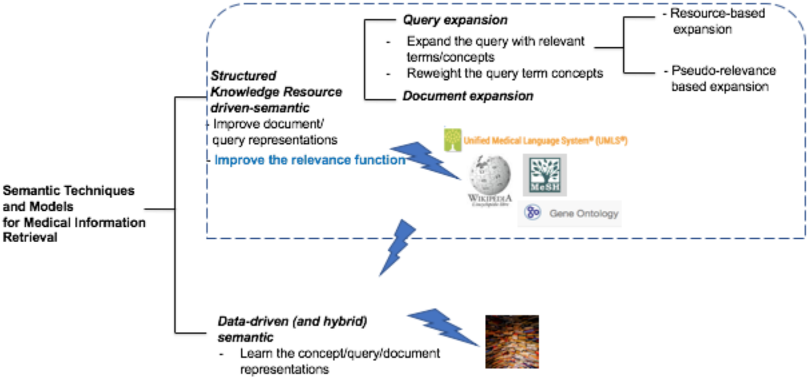
Query/document expansion

Let us discuss

- A large body of state-of-the art work in medical IR rely on query expansion techniques
  - Combine statistical concepts and semantic concepts
  - Exploit generally evidence from multiple resources according to the polyrepresentation view
  - Use generally a local relevance model

- Robustness of query expansion techniques in the medical domain is questionable
  - They do not systematically lead to significant and important improvements over baselines
  - Several impacting factors on performance:
    - The knowledge bases used: general vs. specific
    - The semantic relations exploited: taxonomic vs. non taxonomic
    - The task at hand (collection): searching for cases, scientific literature, etc.
    - Learn more at ECIR'18: *'Jimmy Jimmy, Guido Zuccon and Bevan Koopman. Choices in Knowledge-Base Retrieval for Consumer Health Search'*

ROADMAP


Semantic Techniques and Models for Medical Information Retrieval
Structured Knowledge Resource driven-semantic
Improve document/ query representations
Improve the relevance function
Query expansion
- Expand the query with relevant terms/concepts
- Reweight the query term concepts
Document expansion
- Resource-based expansion
- Pseudo-relevance based expansion
Unified Medical Language System® (UMLS®)
WIKIPEDIA
MeSH
Gene Ontology
Data-driven (and hybrid) semantic
- Learn the concept/query/document representations

RELEVANCE RANKING

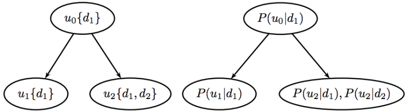
- How to incorporate semantic in the document relevance estimation?
  - Ranking as a semantic inference
    [Goodwin and Harabagiu, 2016, Koopman et al., 2016, Cao et al., 2011]
  - Ranking as learning the discriminant relevant (semantic) features
    [Balaneshin-kordan and Kotov, 2016, Xiong and Callan, 2015, Soldaini and Goharian, 2017]

Relevance ranking

Ranking as a semantic inference: a graph-based approach [Koopman et al., 2016]

- **Key model components**
  - Graph-based representation of the documents
  - Document ranking as an inference process over related concepts in the graph
  - Knowledge resources with directed relationships between concepts
  - Different types of relationships
- **Key inference rationale:** tune the inference mechanism according to semantic gap issues: lexical mismatch, granularity mismatch, conceptual mismatch
  - Lexical mismatch (eg., *hypertension* vs. *high blood pressure*): association and deductive inference
  - Granularity mismatch (eg., *antipsychotic* and *Diazepman*): introduce uncertainty in the taxonomic (hierarchical eg., IS A) relationships
  - Conceptual mismatch (eg., *treatments* $\rightarrow$ *disease*): deductive inference and logical deduction

- The Graph-based corpus representation



(a) Basic node-document representation. (b) Representation with initial probabilities assigned to node.

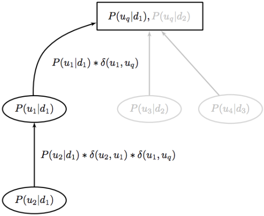
Relevance ranking

Ranking as a semantic inference: a graph-based approach [Koopman et al., 2016]

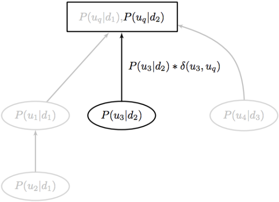
- **The retrieval model**
  - Strength of the association between two information nodes: compute recursively over the graph:
    $\sigma_0(u, u') = \alpha * sim(u, u') + (1 - \alpha) * rel(u, u')$

$$\sigma(u, u') = \begin{cases} 1 \text{ if } u = u' \\ \sigma_0(u, u') \text{ if } uRu' \\ argmax_{u_i \in U : uRu_i} \sigma(u, u_i) \times \sigma(u_i, u'), \text{otherwise} \end{cases} \quad (1)$$

  - Relevance of document-query
    $RSV(d, q) = \prod_{u_q \in q} \prod_{u_d \in d} p(u_d \mid d) \sigma(u_d, u_q)$



(a) Retrieval process for document d_1.



(b) Retrieval process for document d_2.

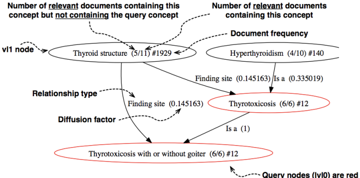
Relevance ranking

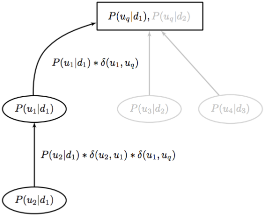
Ranking as a semantic inference: a graph-based approach [Koopman et al., 2016]

- **The retrieval model**
  - Strength of the association between two information nodes: compute recursively over the graph:
    $\sigma_0(u, u') = \alpha * sim(u, u') + (1 - \alpha) * rel(u, u')$

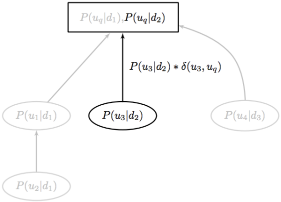
$$\sigma(u, u') = \begin{cases} 1 \ if \ u = u' \\ \sigma_0(u, u') \ if \ uRu' \\ argmax_{u_i \in U: uRu_i} \sigma(u, u_i) \times \sigma(u_i, u'), otherwise \end{cases} \quad (1)$$

  - Relevance of document-query
    $RSV(d, q) = \prod_{u_q \in q} \prod_{u_d \in d} p(u_d \mid d) \sigma(u_d, u_q)$
- **Main results/findings**
  - Effective improvement of queries suffering from the conceptual implication problem
  - Degradation for 'simple' queries do not requiring inference. Inference highlighted general irrelevant concepts





(a) Retrieval process for document d_1.



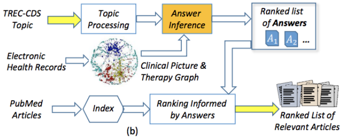
(b) Retrieval process for document d_2.

Relevance ranking

Ranking as a semantic inference: a probabilistic-based approach [Goodwin and Harabagiu, 2016]

- **Key model components**
  - Clinical decision support task through question-answering
  - Document ranking as an inference process over a bayesian graph
  - Medical knowledge-bases for document annotation
  - Full-text documents (EHRs, scientific articles) for background knowledge in the retrieval phase
- **Key inference rationale**: Consider the physician's belief value (eg., hypothetical, present, absent) in the inference process
  - Build the medical knowledge graph as a probabilistic graphical model: compute the probability distribution over all the possible clinical pictures and therapies of patients CPTG
  - Given a question topic t, the set of medical concepts and their assertions derived from the question is viewed as a hidden sketch of the clinical picture and therapy described in the question $Z(t)$ through concepts A with the same type
  - Answering a medical question associated with topic t consists in performing an inference process in the CPTG to determine the medical concepts having the highest likelihood given $Z(t)$

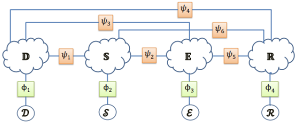
- The architecture of medical question-answering system for clinical decision-support



Relevance ranking

Ranking as a semantic inference: a probabilistic-based approach [Goodwin and Harabagiu, 2016]

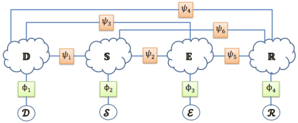
- **The retrieval model**
  - Combine the medical sketch $Z(t)$ with medical concepts from relevant scientific articles $EZ(t,l)$ to provide a complete view of clinical and therapy of a patient:
    $\hat{a} = argmax_{a \in A} p(\{a\} \cup z \mid z) = \frac{p(\{a\} \cup z)}{p(z)}$
    - Ranking answers given concepts in the extended sketch from an article $EZ(t,l)$ and the relevance rank of the article $RRCS(a) = \sum_{r=1}^{|} D| \frac{1}{r} p(\{a\} \cup EZ(t,l_r) \mid EZ(t,l_r)))$
    - Ranking scientific articles that contain at least one of the answers of the medical question $Re(L_i) = p(y_i \mid Z(t)$
    - Generating the clinical picture and therapy graph by modeling six possible types of relations: (1) signs/symptoms S and diagnoses D; (2) signs/symptoms S and tests E; (3) signs/symptoms S and treatments R; (4) diagnoses D and tests E; (5) tests E and treatments R; (6) diagnoses D and treatments R: $p(z) \propto \prod_{i=1}^{6} \omega_i(z) \prod_{j=1}^{4} \phi_j(z)$
    - Optimizing the inference process using Bethe Free-Energy Approximation, pair-wise varitional inference, interpolated smoothing



Relevance ranking

Ranking as a semantic inference: a probabilistic-based approach [Goodwin and Harabagiu, 2016]

- **The retrieval model**
  - Combine the medical sketch $Z(t)$ with medical concepts from relevant scientific articles $EZ(t, l)$ to provide a complete view of clinical and therapy of a patient:
    $\hat{a} = argmax_{a \in A} p(\{a\} \cup z \mid z) = \frac{p(\{a\} \cup z)}{p(z)}$
    - Ranking answers given concepts in the extended sketch from an article $EZ(t, l)$ and the relevance rank of the article $RRCS(a) = \sum_{r=1}^{|} D| \frac{1}{r} p(\{a\} \cup EZ(t, l_r) \mid EZ(t, l_r)))$
    - Ranking scientific articles that contain at least one of the answers of the medical question $Re(L_i) = p(y_i \mid Z(t)$
    - Generating the clinical picture and therapy graph by modeling six possible types of relations: (1) signs/symptoms S and diagnoses D; (2) signs/symptoms S and tests E; (3) signs/symptoms S and treatments R; (4) diagnoses D and tests E; (5) tests E and treatments R; (6) diagnoses D and treatments R: $p(z) \propto \prod_{i=1}^{6} \omega_i(z) \prod_{j=1}^{4} \phi_j(z)$
    - Optimizing the inference process using Bethe Free-Energy Approximation, pair-wise varitional inference, interpolated smoothing



- **Main results/findings**
  - Using the medical knowledge from relevant scientific articles yields significant performance: identifies less common concepts
  - Presence of highly general tests that are not so relevant for the medical expert
  - Issues with negated assertions

Relevance ranking

Ranking as learning the concept-document weighting features [Balaneshin-kordan and Kotov, 2016]

- **Key model components**
  - Document ranking as learning the optimal query concept weighting based on the intuition that different concept types have different importance
  - A set of features characterizing query concept computed on the basis of knowledge-bases (Wikipedia, UMLS) as global context and top retrieved documents as local context of the query
- **The learning process**: A multivariate optimization method to train the weights of all features as contributions to fix the concept weight w.r.t. a document
- **The retrieval model**
  - Relevance status value based on the Markov Random Field retrieval framework [Metzler and Croft, 2005]: weighted linear combination of matching scores of concepts types in a query: $sc(Q, D) = \sum_{c \in C_T} \lambda_T(c) * f_T(c, D)$
  - Importance weight of concept c computed as a linear combination of importance features: $\lambda_T(c) \sum_{n=1}^{N} w_\phi^n \phi^n$
  - Consider global features (eg., popularity concept node in the UMLS graph) and local features (eg., Number of top retrieved documents containing concept c)
  - Learn $\lambda_T(c)$ using a multivariate optimization solving problem.
- **Main results/findings**
  - Reasonable performance improvement (5-9%) over the baselines according to search accuracy
  - Exploiting evidence from multiple resources (Wikipedia and UMLS) positively affects performance

Relevance ranking

Ranking as learning the features that bridge between queries and documents through concepts [Xiong and Callan, 2015]

- **Key model components**
  - A learning to rank model with input features describing connections from queries to objects and from objects to documents
  - Extends ListMLE [Xia et al., 2008] by adding an object layer in the ranking generation process
  - Use features connecting queries to objects (eg., ontology overlap) and features connecting objects to documents (eg., graph connection)
- **The retrieval model**
  - A generative process where the object layer is sampled from objects o belonging to q and the document ranking probability sampled from objects o related to documents d
  - Probability of ranking d given q: $p(d \mid q; w; \theta) = \prod_{i=1}^{n} \sum_{j=1}^{m} p(d_i \mid o_j, S_i) p(o_j \mid q)$
  - Maximize $w*$, $\theta*$ using the EM algorithm
- **Main results/findings**
  - Significant improvements over the baselines for medical and non medical collections
  - The model is sensitive but is still robust to quality of annotation

Relevance ranking

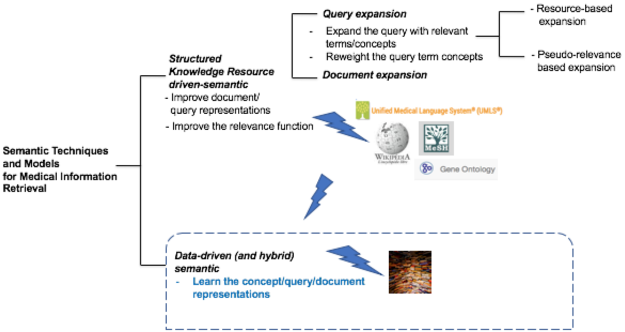
Let us discuss

- **A few work addressed the semantic search at the relevance function level**
  - Identify logical matching between words and concepts
  - Identify relevant semantic features that connect words to concepts, queries to documents

- **Findings: the general trend**
  - High-level inference yiels to high computational complexity
  - The good balance between lexical matching and semantic matching is difficult to tune
  - Robustness to concept annotation quality is important

Outline

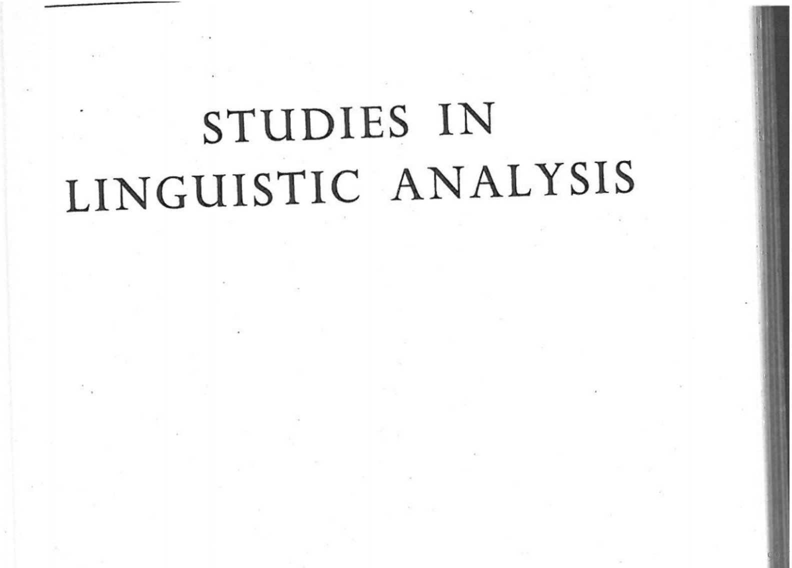
LEARNING


Semantic Techniques and Models for Medical Information Retrieval
Structured Knowledge Resource driven-semantic
- Improve document/ query representations
- Improve the relevance function
Query expansion
- Expand the query with relevant terms/concepts
- Reweight the query term concepts
Document expansion
- Resource-based expansion
- Pseudo-relevance based expansion
Unified Medical Language System® (UMLS®)
WIKIPEDIA
MeSH
Gene Ontology
Data-driven (and hybrid) semantic
- Learn the concept/query/document representations

Fundamentals

Distributional semantics

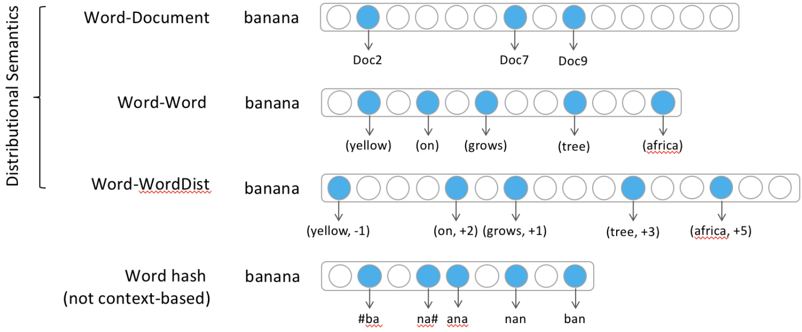
You shall know a word by the company it keeps


STUDIES IN
LINGUISTIC ANALYSIS

FUNDAMENTALS

DISTRIBUTED REPRESENTATIONS OF WORDS

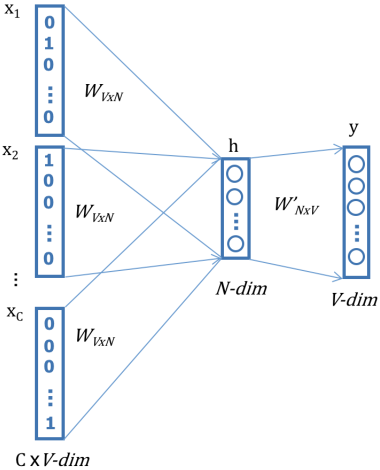
© Tutorial WSDM 2017: Neural Text Embeddings for IR. B. Mitra and N. Craswell



FUNDAMENTALS

DISTRIBUTED REPRESENTATIONS OF WORDS: SEMINAL WORK [MIKOLOV ET AL., 2013]

- **The CBOW model**



$$h = \frac{1}{C} W^{T} \cdot \left(\sum_{i=1}^{C} x_i \right)$$

$$y = W'^{T} \cdot h$$

$$P\left(y_j \mid \{x_1, x_2, \ldots, x_C\}\right) = \frac{exp(y_j)}{\sum_{j'=1}^{V} exp(y_{j'})}$$

Objective function:

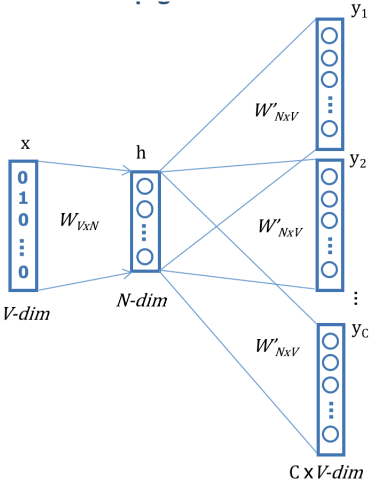
$$E = -\log P(y_j \mid \{x_1, x_2, \ldots, x_C\})$$

$$E = -y_j + \log \sum_{j'=1}^{V} \exp(y_{j'})$$

FUNDAMENTALS

DISTRIBUTED REPRESENTATIONS OF WORDS: SEMINAL WORK [MIKOLOV ET AL., 2013]

- **The Skip-Gram model**



$$h = W^T \cdot x$$

$$y_c = W'^T \cdot h$$

$$P\left(y_{c,j}|x\right) = \frac{exp(y_{c,j})}{\sum_{j'=1}^{V} exp(y_{j'})}$$

Objective function:

$$E = -\log P(y_1, y_2, \ldots, y_c|x)$$

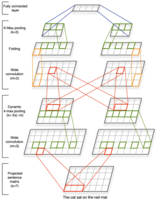
$$= -log \prod_{c=1}^{C} \frac{exp\left(y_{c,j}\right)}{\sum_{j'=1}^{V} exp\left(y_{j'}\right)}$$

$$= -\sum_{j=1}^{C} y_j + C \cdot log \sum_{j'=1}^{V} exp(y_{j'})$$

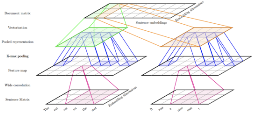
Fundamentals

Deep neural networks for IR

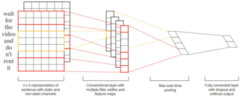
© Tutorial WSDM 2017: Neural Text Embeddings for IR. B. Mitra and N. Craswell



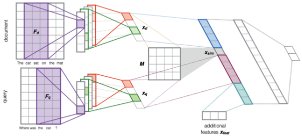
(Kalchbrenner et al., 2014)



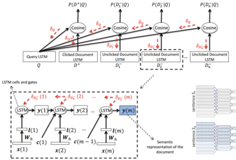
(Denil et al., 2014)



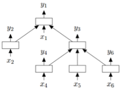
(Kim, 2014)



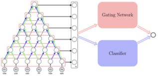
(Severyn and Moschitti, 2015)



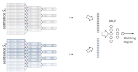
(Palangi et al., 2015)



(Tai et al., 2015)



(Zhao et al., 2015)



(Hu et al., 2014)

Microsoft

Representation learning for medical search

Overview of early research

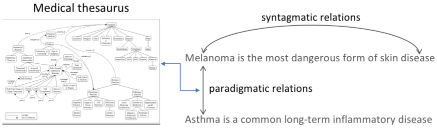
- What do the models learn?
  - *Word, CUI, entity, phrase embeddings*: to learn high-level similarity between information units [De Vine et al., 2014, Limsopatham and Collier, 2016, Liu et al., 2016, Ghosh et al., 2017, Jagannatha and Yu, 2016, Cai et al., 2018, Henry et al., 2018]
  - *Document embeddings*: to improve semantic representations of texts that bridge the gap between data-driven semantic and knowledge resource driven semantic [Minarro-Gimnez et al., 2014, Nguyen et al., 2017, Loza Mencía et al., 2016, Peng et al., 2016, Banerjee et al., 2017, Nguyen et al., 2018a]
  - *Medical objects embeddings*: care events/episodes, disease embeddings [Ghosh et al., 2016, Moen et al., 2015, Choi et al., 2016], patient embeddings [Baytas et al., 2017, Ni et al., 2017, Zhu et al., 2016, Stojanovic et al., 2017, Sushil et al., 2018]
- For which search tasks?
  - Relevance matching RM (eg., document retrieval, care-episode retrieval)
  - Semantic matching (eg., patient similarity)

Representation learning focus	Concept - Word	■	■	□	□
	Document	□	■	□	□
	Care episode	□	□	■	□
	Disease - Disease code	□	□	■	□
	Patient	□	□	□	■
Medical search task	IR task	■	■	□	□
	Concept classification/mapping	■	■	□	□
	Care episode retrieval	□	□	■	□
	Patient search/similarity	□	□	□	■

Representation learning for medical search

Learning Word, CUI, entity, phrase embeddings

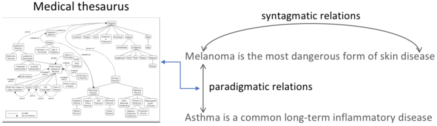
- Different purposes yield to different objective functions
  - Learn readable concept/CUI/entity representations from raw texts: driven by paradigmatic relations provided in knowledge-bases
  - Learn concept and associated poly-senses: learn one vector representation per sense



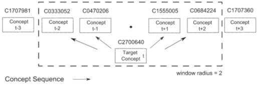
Representation learning for medical search

Learning Word, CUI, entity, phrase embeddings

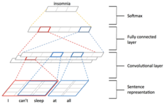
- Different purposes yield to different objective functions
  - Learn readable concept/CUI/entity representations from raw texts: driven by paradigmatic relations provided in knowledge-bases
  - Learn concept and associated poly-senses: learn one vector representation per sense



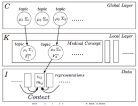
- Different neural architectures
  - Extension of the CBOW and Skip-Gram models
  - Deep architectures (CNN, RNN, ...)



Devine et al. 2014



Limsopatham et al. 2016



Representation learning for medical search

Learning Word, CUI, entity, phrase embeddings [De Vine et al., 2014, Liu et al., 2016]

- Extension of the Skip-Gram model [De Vine et al., 2014]
  - Learn UMLS concept representations from sequences of concepts in annotated texts
  - Maximize the average log probability of the objective function $\frac{1}{2w}\sum_{i=1}^{2w}\sum_{-w\leq j\leq w} log(c_{t+j} \mid c_t)$
  - Valid representations when compared to human-assessments within a concept similarity task (eg., Ped and Cav datasets)
  - Requires hughe amount of annotated data.
  - Sensitivity to concept annotation quality?

Representation learning for medical search

Learning Word, CUI, entity, phrase embeddings [De Vine et al., 2014, Liu et al., 2016]

- Extension of the Skip-Gram model [De Vine et al., 2014]
  - Learn UMLS concept representations from sequences of concepts in annotated texts
  - Maximize the average log probability of the objective function $\frac{1}{2w}\sum_{i=1}^{2w}\sum_{-w\leq j\leq w} log(c_{t+j} \mid c_t)$
  - Valid representations when compared to human-assessments within a concept similarity task (eg., Ped and Cav datasets)
  - Requires hughe amount of annotated data.
  - Sensitivity to concept annotation quality?
- Extension of the CBOW model [Liu et al., 2016]
  - Learn concept representations constrained by relations established in a knowledge base
  - Maximize the log probability of the objective function $L = \sum_{i=1}^{T}(log(p(w_t \mid w_{t+k}) + \alpha \sum_{w_s:(w_t,w_s)\in R} wt(w_s \mid w_t)(logp(w_t \mid w_{t\pm k} - logp(w_s \mid w_{s\pm k})))^2)$ $w_t(w_s \mid w_t) = \frac{f(w_s)}{\sum_{(w_t,w)\in R} f(w)}$
  - Experimental evaluation on IR tasks (query expansion) show: 1) sensitivity to model parameters and collections; 2) ability to select related words in the UMLS thesarus; 3) slight improvement on a medical document search task

The most similar words to « heart »

CBOW		Online	
Cardiac	0.4891	Cardiac	0.5205
Synergist	0.4494	Hearts	0.5030
Hearts	0.4276	Cor	0.4939
Cardiovascular	0.4096	Synergist	0.4690
Acyanotic	0.3987	Cardiovascular	0.4156
Ouvrier	0.3934	Cerebrovascular	0.4149
Multiorgan	0.3931	Acyanotic	0.3985
Ventricular	0.3837	Ventricular	0.3979
Cardiorespiratory	0.3829	Cardiorespiratory	0.3969
Thrive	0.3766	Biventricular	0.3831

Representation learning for medical search

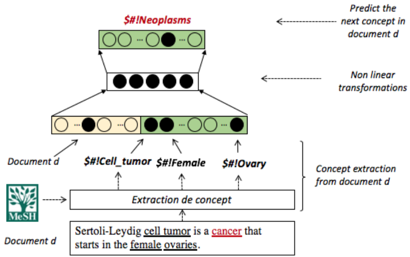
Learning document embeddings [Minarro-Gimnez et al., 2014, Nguyen et al., 2017, Loza Mencía et al., 2016, Peng et al., 2016, Banerjee et al., 2017, Nguyen et al., 2018a]

- Extension of the Doc2Vec model based on an **offline learning** approach [Nguyen et al., 2017]
  - Build the optimal real-valued representation $\hat{d}$ of document d such that the knowledge-based embedding $\hat{d}_i^{(cd2vec)}$ and the corpus-based embedding $\hat{d}^{(PV-DM)}$ are nearby in the latent space. Formally through the minimization problem:

$$\Psi(D) = \sum_{d \in D} \psi(d) = \sum_{d \in D} \left[(1-\beta) \times \|d - \hat{d}^{(cd2vec)}\|^2 + \beta \times \|d - \hat{d}^{(PV-DM)}\|^2 \right]$$

  - Concept-based latent representation of document d is obtained using and extension of the *cd2vec* model. Document vectors $\hat{d}^{(cd2vec)}$ are learned so they allow predicting concepts in their context by maximizing the log-likelihood:

$$\varphi = \sum_{d \in D} logP(d \mid c_1, \ldots, c_m) + \sum_{c_j \in \mathcal{C}_d} logP(c_j \mid c_{j-W} : c_{j+W}, d)$$



Representation learning for medical search

Learning document embeddings [Minarro-Gimnez et al., 2014, Nguyen et al., 2017, Loza Mencía et al., 2016, Peng et al., 2016, Banerjee et al., 2017, Nguyen et al., 2018a]

- Evaluation in a retrieval task [Nguyen et al., 2017]
  - Significant but slight improvement using a query expansion task
  - The model is sensitive to the quality of concept annotation
  - The model allows identifying relevant related concepts

Table: Example of terms/concepts expanded for query 131 in TREC Med

Query text	patients underwent minimally invasive abdominal surgery
Extracted Concepts	Patients; General Surgery;
Added by $Exp_{\hat{d}PV-DM}$	myofascia; ultrasonix; overtube
Added by $Exp_{\hat{d}cd2vec}$	Mesna; Esophageal Sphincter, Upper; Ganglioglioma
Added by Exp_{d}	*umbilical*; *ventral*; *biliary-dilatation*

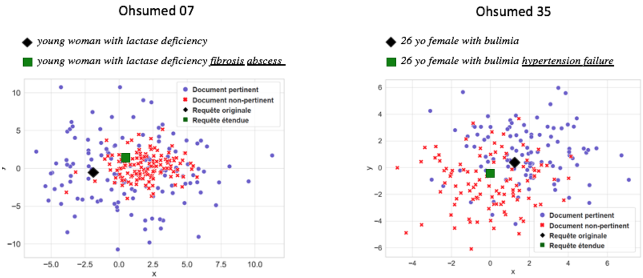
Representation learning for medical search

Learning document embeddings [Minarro-Gimnez et al., 2014, Nguyen et al., 2017, Loza Mencía et al., 2016, Peng et al., 2016, Banerjee et al., 2017, Nguyen et al., 2018a]

- Evaluation in a retrieval task [Nguyen et al., 2017]
  - Significant but slight improvement using a query expansion task
  - The model is sensitive to the quality of concept annotation
  - The model allows identifying relevant related concepts

Table: Example of terms/concepts expanded for query 131 in TREC Med

Query text	patients underwent minimally invasive abdominal surgery
Extracted Concepts	Patients; General Surgery;
Added by $Exp_{\hat{d}PV-DM}$	myofascia; ultrasonix; overtube
Added by $Exp_{\hat{d}cd2vec}$	Mesna; Esophageal Sphincter, Upper; Ganglioglioma
Added by Exp_{d}	*umbilical*; *ventral*; *biliary-dilatation*

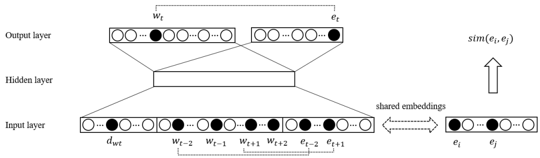


Representation learning for medical search

Learning document embeddings [Minarro-Gimnez et al., 2014, Nguyen et al., 2017, Loza Mencía et al., 2016, Peng et al., 2016, Banerjee et al., 2017, Nguyen et al., 2018a]

- Model improvement by considering knowledge-base constraints using an **online learning** approach [Nguyen et al., 2018b]

$$L_C = \sum_{d \in \mathcal{D}} \sum_{w_t \in \mathcal{W}s_d} [\log p(w_t | w_{t\pm k}, c_{t\pm k}, d) + \log p(c_t | w_{t\pm k}, c_{t\pm k}, d) - \frac{\gamma}{|d|} ||v_d||^2]$$

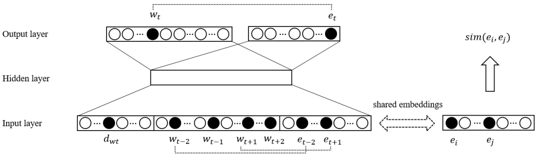


Representation learning for medical search

Learning document embeddings [Minarro-Gimnez et al., 2014, Nguyen et al., 2017, Loza Mencía et al., 2016, Peng et al., 2016, Banerjee et al., 2017, Nguyen et al., 2018a]

- Model improvement by considering knowledge-base constraints using an **online learning** approach [Nguyen et al., 2018b]

$$L_C = \sum_{d \in \mathcal{D}} \sum_{w_t \in \mathcal{W}s_d} [\log p(w_t | w_{t\pm k}, c_{t\pm k}, d) + \log p(c_t | w_{t\pm k}, c_{t\pm k}, d) - \frac{\gamma}{|d|} ||v_d||^2]$$

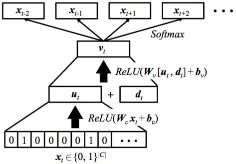


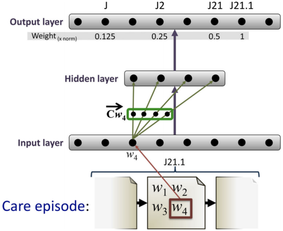
- Evaluation in a retrieval task [Nguyen et al., 2018b]
  - Both offline and online models are more effective in medical IR tasks than in general domain-search tasks. This is reversed in the case of NLP tasks
  - Constraining the learning with relational knowledge is effective in both NLP and IR tasks. The learning leverage from both word-word relations and concept-concept relations

Representation learning for medical search

Learning care episodes, disease codes [Ghosh et al., 2016, Moen et al., 2015, Choi et al., 2016]

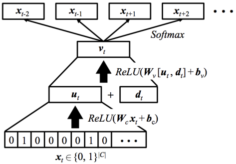
- Learning care episodes [Moen et al., 2015, Choi et al., 2016]
  - Introduce a longitudinal temporal view of medical documents (eg., care episode as successive clinical notes related to patient visits). Two types of relations:
    - Coocurrence of words: basic relation addressed in neural models
    - Sequential order of visits, codes
    - Inputs are high-level data (eg., EHR): computational complexity is questionable
    - Interpretability of the learned representations are highly required



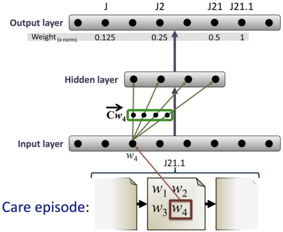


Representation learning for medical search

Learning care episodes, disease codes [Ghosh et al., 2016, Moen et al., 2015, Choi et al., 2016]



- Learning care episodes [Moen et al., 2015, Choi et al., 2016]
  - Introduce a longitudinal temporal view of medical documents (eg., care episode as successive clinical notes related to patient visits). Two types of relations:
    - Coocurrence of words: basic relation addressed in neural models
    - Sequential order of visits, codes
    - Inputs are high-level data (eg., EHR): computational complexity is questionable
    - Interpretability of the learned representations are highly required
  - Evaluation in a retrieval task [Moen et al., 2015]
    - Single care episod vectors: one vector representation per care episode based on average word vectors ; compute vector similarity between care episodes embeddings (eg., cosinus)
    - Average note vector similarity: one vector representation per care episode based on average note vectors; compute average pairwise (note) similarities, optimal pairing, sequence alignment



Representation learning for medical search

Learning care events/episodes, disease representations [Ghosh et al., 2016, Moen et al., 2015, Choi et al., 2016]

- Learning disease: identify taxonomic and interpretable representations of diseases from raw data
- Extension of the Skip-Gram model [Choi et al., 2016]
  - **Input**: ordered sequence of medical codes extracted from patient visits (raw data of the patient visits)
  - **Output**: medical code representation, visit representation
  - **Learning objective**: two levels of learning functions unified in the same framework (yield to representations learned in a shared space)
    - Predict the medical codes given a visit: what happend in the past, what coul happen in the future?

$$min_{W_s, b_s} \frac{1}{T} \sum_{t=1}^{T} \sum_{-w \leq \leq w} -x_{t+1}^{T} log \hat{y}_t - (1 - x_{t+i}^{T} log(1 - \hat{y}_t)$$

    - Predict the medical codes given the code represenions in the same visit

$$min_{W_c'} \frac{1}{T} \sum_{t=1}^{T} \sum_{i:c_i \in V_t} \sum_{j:c_j \in V_t, j \neq i} log(c_j \mid c_i)$$

Representation learning for medical search

Learning care events/episodes, disease representations [Ghosh et al., 2016, Moen et al., 2015, Choi et al., 2016]

- Learning disease: identify taxonomic and interpretable representations of diseases from raw data
- Extension of the Skip-Gram model [Choi et al., 2016]
  - **Input**: ordered sequence of medical codes extracted from patient visits (raw data of the patient visits)
  - **Output**: medical code representation, visit representation
  - **Learning objective**: two levels of learning functions unified in the same framework (yield to representations learned in a shared space)
    - Predict the medical codes given a visit: what happend in the past, what coul happen in the future?

$$min_{W_s,b_s} \frac{1}{T} \sum_{t=1}^{T} \sum_{-w \leq \leq w} -x_{t+1}^{T} log \hat{y}_t - (1 - x_{t+i}^{T} log(1 - \hat{y}_t)$$

    - Predict the medical codes given the code represenions in the same visit

$$min_{W_c'} \frac{1}{T} \sum_{t=1}^{T} \sum_{i:c_i \in V_t} \sum_{j:c_j \in V_t, j \neq i} log(c_j \mid c_i)$$

  - **Qualitative evaluation**
    - Interpretable visit representations validated by medical experts

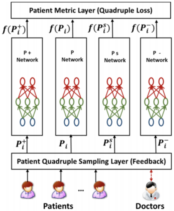
Coordinate 112	Coordinate 152	Coordinate 141
Kidney replaced by transplant (V42.0) Hb-SS disease without crisis (282.61) Heart replaced by transplant (V42.1) RBC antibody screening (P) Complications of transplanted bone marrow (996.85) Sickle-cell disease (282.60) Liver replaced by transplant (V42.7) Hb-SS disease with crisis (282.62) Prograf PO (R) Complications of transplanted heart (996.83)	X-ray, knee (P) X-ray, thoracolumbar (P) Accidents in public building (E849.6) Activities involving gymnastics (E005.2) Struck by objects/persons in sports (E917.0) Encounter for removal of sutures (V58.32) Struck by object in sports (E917.5) Unspecified fracture of ankle (824.8) Accidents occurring in place for recreation and sport (E849.4) Activities involving basketball (E007.6)	Cystic fibrosis (277.02) Intracranial injury (854.00) Persistent mental disorders (294.9) Subdural hemorrhage (432.1) Neurofibromatosis (237.71) Other conditions of brain (348.89) Conductive hearing loss (389.05) Unspecified causes of encephalitis, myelitis, encephalomyelitis (323.9) Sensorineural hearing loss (389.15) Intracerebral hemorrhage (431)
Sickle-cell disease and organ transplant	Sport-related injuries	Brain injuries and hearinh loss

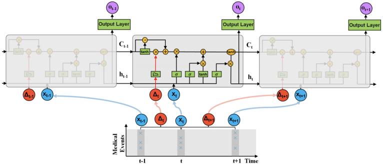
Representation learning for medical search

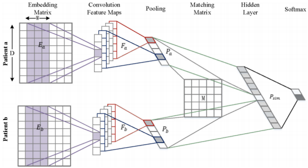
Learning patient profiles, patient similarity

[Baytas et al., 2017, Ni et al., 2017, Zhu et al., 2016, Stojanovic et al., 2017, Sushil et al., 2018]

- Two main objectives
  - Learn patient profile: input (EHR) - output (patient vector) [Baytas et al., 2017, Stojanovic et al., 2017, Sushil et al., 2018]
  - Learn end-to-end patient-patient similarity:
    - Input: EHR patient A, EHR patient B
    - Output: similarity class[Zhu et al., 2016, Ni et al., 2017]
- Input data
  - Heterogeneous patient data: demographic, medication, diagnosis codes etc.
  - Historical data: considering the sequence of medical events with irregular intervals





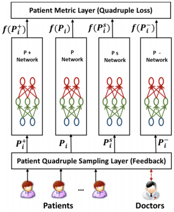


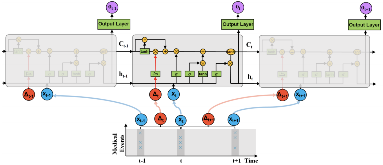
Representation learning for medical search

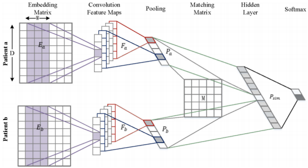
Learning patient profiles, patient similarity

[Baytas et al., 2017, Ni et al., 2017, Zhu et al., 2016, Stojanovic et al., 2017, Sushil et al., 2018]

- Two main objectives
  - Learn patient profile: input (EHR) - output (patient vector) [Baytas et al., 2017, Stojanovic et al., 2017, Sushil et al., 2018]
  - Learn end-to-end patient-patient similarity:
    - Input: EHR patient A, EHR patient B
    - Output: similarity class[Zhu et al., 2016, Ni et al., 2017]
- Input data
  - Heterogeneous patient data: demographic, medication, diagnosis codes etc.
  - Historical data: considering the sequence of medical events with irregular intervals
- Tasks
  - Predict patient mortality, primary diagnosis, length of stay, total incurred charges, ...







Representation learning for medical search

Learning patient profiles, patient similarity

[Baytas et al., 2017, Ni et al., 2017, Zhu et al., 2016, Stojanovic et al., 2017, Sushil et al., 2018]

- Learning patient vector from patient disease and procedure descriptions [Stojanovic et al., 2017]
  - Learn the representations of diseases and procedures (dp2vec): extend the w2vec model by considering patient record as a ”sentence” and diseases/procedures as ”words”
  - Build patient vector: sum the vector representations of in-patient diseases and procedures

Neighbors of *respiratory failure*	Neighbors of *congestive heart failure*
Insertion of endotracheal tube	Insertion of implantable heart assist system
Tracheostomy toilette	Implantation of cardiac resynchronization defibrillator total system (CRT-D)
Other lavage of bronchus and trachea	Implantation of cardiac resynchronization defibrillator pulse generator (CRT-D)
Bronchoscopy through artificial stoma	Insertion of percutaneous external heart assist device
Other oxygen enrichment	Heart transplantation
Other repair and plastic operations on trachea	Excision destruction or exclusion of left atrial appendage (LAA)
Fiber-optic bronchoscopy	Aquapheresis
Infusion of vasopressor agent	Automatic implantable cardioverter-defibrillator (AICD) check
Replacement of tracheostomy tube	Noninvasive programmed electrical stimulation (NIPS)
Replacement of gastrostomy tube	Removal of lead(s) [electrode] without replacement
Complete glossectomy	Endovascular removal of obstruction from head and neck vessel(s)
Other intubation of respiratory tract	Replacement of automatic cardioverter-defibrillator lead(s) only

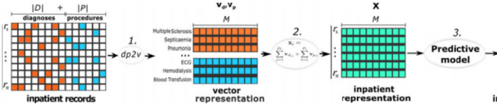
Representation learning for medical search

Learning patient profiles, patient similarity

[Baytas et al., 2017, Ni et al., 2017, Zhu et al., 2016, Stojanovic et al., 2017, Sushil et al., 2018]

- Learning patient vector from patient disease and procedure descriptions [Stojanovic et al., 2017]
    - Learn the representations of diseases and procedures (dp2vec): extend the w2vec model by considering patient record as a "sentence" and diseases/procedures as "words"
    - Build patient vector: sum the vector representations of in-patient diseases and procedures
- Task evaluation
    - Predict length of patient stay and total charges as a regression problem (vector reprsentations used as features)
    - Predict mortality as a classification problem

Neighbors of *respiratory failure*	Neighbors of *congestive heart failure*
Insertion of endotracheal tube	Insertion of implantable heart assist system
Tracheostomy toilette	Implantation of cardiac resynchronization defibrillator total system (CRT-D)
Other lavage of bronchus and trachea	Implantation of cardiac resynchronization defibrillator pulse generator (CRT-D)
Bronchoscopy through artificial stoma	Insertion of percutaneous external heart assist device
Other oxygen enrichment	Heart transplantation
Other repair and plastic operations on trachea	Excision destruction or exclusion of left atrial appendage (LAA)
Fiber-optic bronchoscopy	Aquapheresis
Infusion of vasopressor agent	Automatic implantable cardioverter-defibrillator (AICD) check
Replacement of tracheostomy tube	Noninvasive programmed electrical stimulation (NIPS)
Replacement of gastrostomy tube	Removal of lead(s) [electrode] without replacement
Complete glossectomy	Endovascular removal of obstruction from head and neck vessel(s)
Other intubation of respiratory tract	Replacement of automatic cardioverter-defibrillator lead(s) only



Representation learning for medical search

Let us discuss

- In summary
  - Recent trend toward the use of neural models in medical search: early stage, not yet mature work but seems promising
  - Learned representations reusable in a wide range of search tasks and prediction tasks
  - Background knowledge (eg., Knowledge-base, expert's assessments) driven learning increases the readability of the representations and the explicability of the learning outcomes

- Pending issues
  - What are the impacting factors? What works vs. fails in the black box?
  - Lack of sufficient amount of labeled data to learn accurate representations (eg., patient similarity, IR tasks)
  - Performance sensitivity to a large size of network parameters, hyper-parameters and models parameters
  - Performance variability across medical tasks

Outline

Challenges in Evaluating Medical Information Retrieval

How is Information Retrieval Evaluated

Evaluating an IR system consists in checking how it satisfies an information need

What satisfies a human is different from what satisfies a system!

Two levels of evaluation can be distinguished:

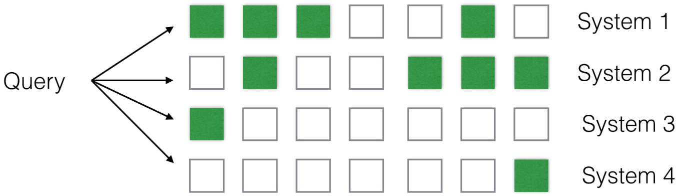
- Evaluation at the document level: does this document satisfy the information need?
- Evaluation at the system level: does this system helped in satisfying the information need? I.e. does it retrieve one or more document(s) satisfying the information need?

Efficient and effective system

- Efficiency: time and space
  - Pre-development specifications
  - Easy to measure
- Effectiveness: good results
  - What is a good result?

Challenges in Evaluating Medical Information Retrieval

Evaluation at the system level



Which system is the best one?

- Looking for the highest number of relevant documents? → Systems 1 and 2
- Looking for a single document giving the answer? → Systems 1, 2, 3 and 4
- Looking at the ranking of the documents? → Systems 1 and 3

Evaluation Paradigms

Laboratory-based: the Cranfield paradigm

- Testing and comparing search systems requires a laboratory environment that doesnt change
- Initiated by Cyril Cleverdon in the Cranfield College of Aeronautics called *the Cranfield Tests*
- Retrieval experiments conducted on test databases in a controlled setting

User studies

- Measuring user satisfaction with feedback from *real users*
- Also allows to measure a system's usability
- Mostly done in Interactive Information Retrieval
- [Kelly et al., 2009]

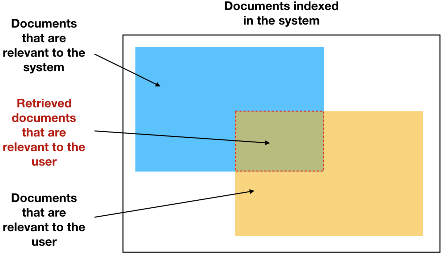
Log-based/click based

- When there is no space/time for laboratory evaluation, e.g. online systems
- A/B testing: redirecting n% of the trafic on a new version of the system
- Click analysis: inference on document relevance/usefulness

Challenges in Evaluating Medical Information Retrieval

Evaluation at the system level

It's all a matter of precision and recall...

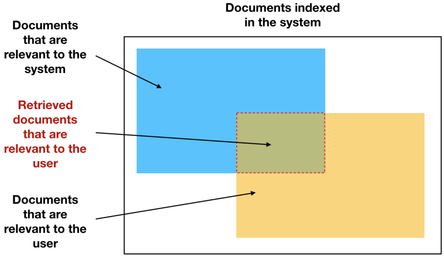


$$Precision = \frac{|P \cap R|}{|R|}, Recall = \frac{|P \cap R|}{|P|}$$

Challenges in Evaluating Medical Information Retrieval

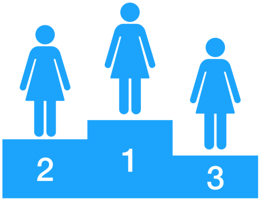
Evaluation at the system level

It's all a matter of precision and recall...



$$Precision = \frac{|P \cap R|}{|R|}, Recall = \frac{|P \cap R|}{|P|}$$

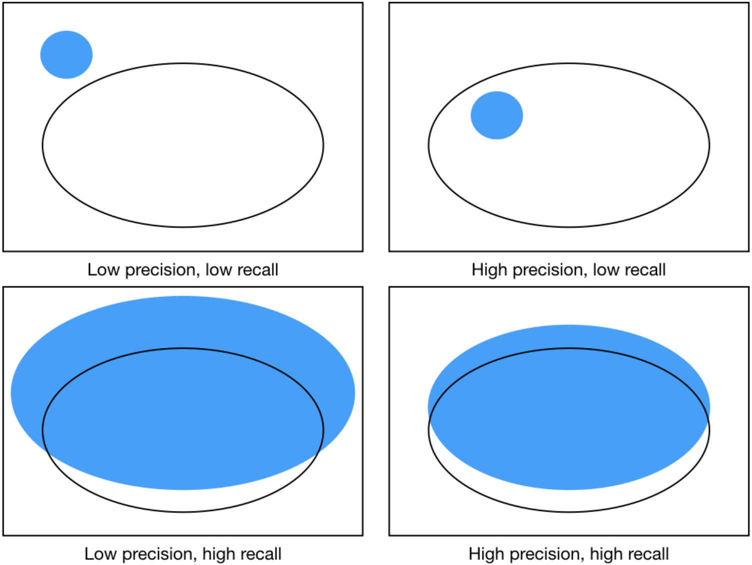
... And of rank!



Unless they are looking for the entire set of documents, nobody goes through the entire set of results.
Ranked metrics:

- P@N
- Mean Average Precision (MAP) [Voorhees, 1998]
- Normative Discounted Cumulation Gain [Jarvelin and Kekalainen, 2000]

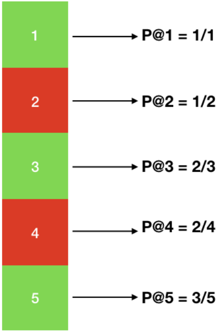
Balancing precision and recall



P@N and Mean Average Precision

- Precision @r: computes the precision at a certain rank
- Mean Average Precision:
  - Average precision over all relevant documents
  - Rewards systems retrieving relevant documents quickly

Results for query *i*



$$\text{Average Precision} = \frac{(1/1 + 2/3 + 3/5)}{3}$$

Mean Average Precision : average on all the queries

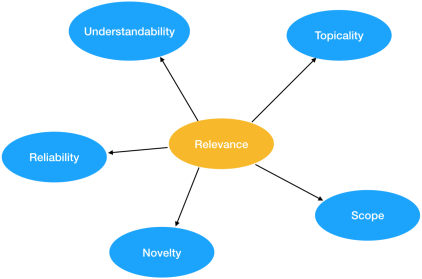
Challenges in Evaluating Medical Information Retrieval

Evaluation at the document level

In classical IR

- From the user point-of-view: a relevant document answer the initial information need
- From the system point-of-view: A relevant document covers the same topic as the query, i.e. contains the query's terms

Relevance has many other dimensions [Zhang et al., 2014]



In the medical domain:

- For patients:
  - Documents must be readable and understandable for a given user
  - The information contained in the documents should be trustworthy
- For medical professionals:
  - Documents must contain up-to-date information
  - Documents must properly cover the topic searched

Challenges in Evaluating Medical Information Retrieval

Each search task has its proper objectives:

- How should the retrieval and the ranking be implemented?
- How should the system be evaluated?

Examples:

- Physician adhoc search: priority given to the rank, P@10, the topicality, scope...
- Patient adhoc search: priority given to the rank, P@10, the topicality, understandability, readability...
- Clinical trials: priority given to the rank, the topicality, the scope, the novelty...
- Systematic reviews: priority given to the recall, the topicality, the scope...

Outline

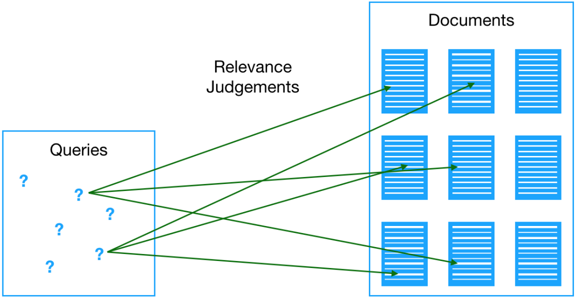
Evaluation Challenges

What is a benchmark?

- Comparing 2 search systems results on a common dataset allows to compare their effectiveness.
- These common datasets are called *benchmarks*.

An IR benchmark contains:

- A document collection that can be indexed
- A set of topics (enriched queries)
- Relevance judgements (linking queries to the relevant documents in the collection)



Evaluation Challenges

The Cranfield Paradigm

Given:

1. A test collection (T, D, R)
2. A retrieval run for the test collection : a doc-list L_t for each topic t in T

For each topic t in T

- Use a measure (e.g. P@10) to compute the quality of L_t

Combine scores:

- Mean average precision

Relevance judgement:

- For a given topic $t \in T$, a given document $d \in D$, $R(d, t)$ is the relevance score of d for topic t.
- $R(d, t)$ can be:
  - a discrete value: e.g. $\in 0, 1$ for binary assessment or $\in 0, 1, 2, 3$ for graded assessment
  - a continuous value: e.g. $\in [0, 1]$
- Assumption: if $R(d, t, u_1)$ is the judgement of assessor u_1 on topic t and document d and $R(d, t, u_2)$ the judgement of assessor u_2 on topic t and document d, $R(d, t, u_1) = R(d, t, u_2)$

Evaluation Challenges

Main Organizers

There are many forums organizing challenges and benchmarking activities, and most of them have some medical tracks:

- **Text REtrieval Conference (TREC)**: organized by the US National Institute of Standards and Technology (NIST). Provides since 1992 numerous evaluation challenges and a forum to discuss the results.
  - `http://trec.nist.gov`
- **Conference and Labs of the Evaluation Forum (CLEF)**: European (smaller) version of TREC, organized by the CLEF initiative. Organizes since 2000 evaluation challenges along with a conference.
  - `http://www.clef-initiative.eu/`
- **NTCIR**: organized by the Japanese National Institute of Informatics (NII). Provides since 1998 numerous evaluation challenges and a forum to discuss the results.
  - `http://ntcir.nii.ac.jp/`
- **Forum for Information Retrieval Evaluation (FIRE)**: South Asian counterpart, organized in India. Provides since 2008 numerous evaluation challenges and a forum to discuss the results.
  - `http://fire.irsi.res.in/`

TREC Medical Evaluation Challenges

TREC Genomics

Ran from 2003-2007 (`https://trec.nist.gov/data/genomics.html`)
Tasks organized:

- **Adhoc retrieval** (2003, 2004, 2005)
  - *Documents*: subset of MEDLINE
  - *Topics*: gene names (2003), information need statement - GTT (2004, 2005)
- **Passage retrieval** (2006):
  - *Documents*: collection of biomedical articles
  - *Topics*: GTTs
- **Entity-based Question-Answering** (2007):
  - *Documents*: collection of biomedical articles
  - *Topics*: questions based on the entities
- **Text annotation** (2003)
  - *Documents*: collection of biomedical articles
  - *Annotation*: title and abstract annotation using geneRIFs (annotation mechanism for genes)
- **Text categorization** (2004-2005)
  - *Documents*: collection of biomedical articles
  - *Categorization* of documents containing data about gene function suitable for "triage" to annotators assigning Gene Ontology (GO) codes for Mouse Genome Informatics database

TREC Medical Evaluation Challenges

TREC Genomics adhoc retrieval Track

[Hersh and Bhupatiraju, 2003, Hersh et al., 2004, Hersh et al., 2005]:

Documents:

- The subset of MEDLINE used for the track consisted of 10 years of completed citations from the database inclusive from 1994 to 2003.
- Records were extracted using the Date Completed (DCOM) field for all references in the range of 19940101 - 20031231.
- This provided a total of 4,591,008 records.

Topics:

- developed from the information needs of real biologists
- 43 interviews with 12 volunteers yielded 74 information needs
- Topics built manually from these information needs
- 50 topics selected

Examples of topic:

```
<TOPIC>
<ID>4</ID>
<TITLE>Gene expression profiles for kidney in mice</TITLE>
<NEED>What mouse genes are specific to the kidney?</NEED>
<CONTEXT> What genes are expressed only in the mouse kidney and not in other tissues?</CONTEXT>
</TOPIC>
<TOPIC>
<ID>5</ID>
<TITLE>Protocols for isolating cell nuclei</TITLE>
<NEED> Articles are relevant if they describe methods for subcellular fractionation of nuclei.</NEED>
<CONTEXT> Laboratory preparations can be enriched for certain kinds of proteins if the cellular
compartment in which they reside is purified away from the rest of the cell contents.</CONTEXT>
</TOPIC>
```

TREC Medical Evaluation Challenges

TREC Genomics adhoc retrieval Track

Pooling:

- Each participating team designated a top-precedence run that would be used for relevance judgement
- They used on average the top 75 documents for each topic from 27 selected runs (removed duplicates)
- average pool size: 976, with a range of 476-1450

Relevance Judgement:

- For each topic, documents are judged as definitely relevant (DR), possibly relevant (PR), or not relevant (NR)
- subset of documents were also judged in duplicate to assess interjudge reliability using the kappa measure: score was 0.51, indicating a fair level of agreement

Evaluation measures:

- The primary evaluation measure for the task was mean average precision (MAP).
- As well as the binary preference (B-Pref), precision at the point of the number of relevant documents retrieved (R-Prec), and precision at varying numbers of documents retrieved (e.g., 5, 10, 30, etc. documents up to 1,000)

TREC Medical Evaluation Challenges

TREC Genomics passage retrieval Track

Details on the passage retrieval task [Hersh et al., 2006]:

Documents:

- new full-text biomedical corpus from Highwire Press
- full text in HTML format, which preserved formatting, structure, table and figure legends, etc.
- full collection contained 162,259 documents

Topics:

- Questions derived from the set of questions based on the GTTs developed in 2005
- 50 topics selected

Examples of topic:

<TOPIC>
<ID>4</ID>
<TITLE>Gene expression profiles for kidney in mice</TITLE>
<NEED>What mouse genes are specific to the kidney?</NEED>
<CONTEXT> What genes are expressed only in the mouse kidney and not in other tissu
</TOPIC>
<TOPIC>
<ID>5</ID>
<TITLE>Protocols for isolating cell nuclei</TITLE>
<NEED> Articles are relevant if they describe methods for subcellular fractionatio
<CONTEXT> Laboratory preparations can be enriched for certain kinds of proteins if
</TOPIC>

TREC Medical Evaluation Challenges

TREC Genomics passage retrieval Track

Results submission:

- Retrieved passages could contain any span of text that did not include any part of an HTML paragraph tag
- Submitted runs could contain up to 1000 passages per topic that were predicted to be relevant to answering the topic question.
- There were 92 submitted runs, with each nominating up to 1000 passages over 28 topics.

Pooling:

- for each topic, pooling was done by taking the top ranked maximal legal span from each submitted run in a round-robin fashion (up to 1000 unique spans)

Relevance Judgement:

- judges were instructed to break down the question into required elements and isolate the minimum contiguous substring that answered the question
- definitely relevant, possibly relevant, not relevant

Evaluation measures:

- Multidimensional evaluation: passage retrieval, aspect retrieval, and document retrieval
- MAP measure for all dimensions

TREC Medical Evaluation Challenges

TREC Genomics

Lessons learned:

- **TREC Genomics adhoc retrieval track**:
  - Symbols and acronym expansion using knowledge bases
  - Domain-specific query expansion
- **TREC Genomics passage retrieval task**:
  - Tricks on improving the results were mostly related to the way passages were extracted and ranked

TREC Medical Evaluation Challenges

TREC Medical Records Track

Organized in 2011 and 2012 [Voorhees and Hersh, 2012].

- **Purpose**: search in a set of EHR to identify patient cohorts for (possible) clinical studies
- **Dataset**:
  - Topics: description of the criteria for inclusion in a study
  - Documents: The document set used in the track is a set of de-identified clinical reports made available to TREC participants through the University of Pittsburgh NLP Repository[1]. EHR are grouped as "visits".
  - Relevance judgement: conducted by physicians
- **Runs**: systems must return a list of visits ordered by the likelihood that the inclusion criteria are satisfied
- `https://trec.nist.gov/data/medical.html`

[1] Use of this dataset is now restricted to people having a license with Pittsburg NLP.

TREC Medical Evaluation Challenges

TREC Medical Records Track

Documents:

- 9 types of documents: Radiology Reports, History and Physicals, Consultation Reports, etc.
- Semi-structured reports with ICD coding, chief complaint
- De-identified and grouped as visits
- 93,551 reports mapped into 17,264 visits

Topics:

- Created by physicians (graduate students in OHSU)
- From a list of research areas the U.S. Institute of Medicine (IOM) has deemed priorities for clinical comparative effectiveness research
- 35 topics in 2011, 50 in 2012

Examples of topics:

136	Children with dental caries
137	Patients with inflammatory disorders receiving TNF-inhibitor treatment
160	Adults under age 60 undergoing alcohol withdrawal
167	Patients with AIDS who develop pancytopenia
169	Elderly patients with subdural hematoma
179	Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

TREC Medical Evaluation Challenges

TREC Medical Records Track

Results submission:

- Several runs were allowed (manual or automatic runs) with ranked list of visits

Pooling:

- Each topic was completely judged (i.e., had all the visits in its judgment set judged)
- All submitted runs contributed to the judgment sets: all visits retrieved in ranks 1-15 by any run in union with a 25% sample of visits not retrieved in the first set that were retrieved in ranks 16-100 by some run
- 25,596 visits were judged (average size of a judgment set: 512 visits)

Relevance Judgement:

- 25 physicians, judged between 19 topics
- rate each visit to determine whether such a patient would be a candidate for a clinical study on the topic
- definitely relevant, possibly relevant / not relevant

Evaluation measures:

- infNDCG, infAP, P@10

TREC Medical Evaluation Challenges

TREC Medical Records Track

Many errors were due to incorrect lexical representations and mismatches

- Non relevant documents contain the query terms
- Relevant documents containing lexical variants
- Document and query processing requires to take into account negation or uncertainty
- Time factors can also generate errors

It is crucial to use proper baselines (very hard to beat BM25) [Leveling et al., 2012].

TREC Medical Evaluation Challenges

TREC Clinical Decision Support / Precision Medicine Track

TREC CDS/PM task evolved over the years, this is a description of the first edition:

- Organized since 2014, started as the Clinical Decision Supports (CDS) track, since 2017 named Precision Medicine (PM) track.
- **Purpose**: the retrieval of biomedical articles relevant for answering generic clinical questions about medical records.
- Given a case report, find full-text biomedical articles that answer questions related to several types of clinical information needs:
  - **Topics**: case reports and one of three generic clinical question types, such as "What is the patient's diagnosis?". Created by expert topic developers at the U.S. National Library of Medicine
  - **Documents**: Full biomedical articles: open access subset1 of PubMed Central (PMC), snapshot of 733,138 articles.
- A case report typically describes a challenging medical case, and it is often organized as a well-formed narrative summarizing the portions of a patient's medical record that are pertinent to the case.

[Simpson et al., 2014, Roberts et al., 2015b, Roberts et al., 2016, Roberts et al., 2017]

TREC Medical Evaluation Challenges

TREC Clinical Decision Support / Precision Medicine Track

Example of topics descriptions and summaries (extracted from [Simpson et al., 2014]):

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.

Topic	Type	Summary
1	Diagnosis	58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.
11	Test	40-year-old woman with severe right arm pain and hypotension. She has no history of trauma and right arm exam reveals no significant findings.
21	Treatment	21-year-old female with progressive arthralgias, fatigue, and butterfly-shaped facial rash. Labs are significant for positive ANA and anti-double-stranded DNA, as well as proteinuria and RBC casts.

TREC Medical Evaluation Challenges

TREC Clinical Decision Support / Precision Medicine Track

Evolution of the dataset over the years:

	Topics	Documents
2014	30 topics (10 per category: diagnosis, treatment, test) based on made-up case reports	733,138 articles from PubMed Central
2015	30 new topics	No change
2016	30 nursing admission notes from the MIMIC-III database	Extension: 1.5M articles from PubMed Central
2017	30 topics describing oncology patient cases: disease, genetic variants, demographic information, other factors	27M MEDLINE abstracts + 250,000 clinical trials

TREC Medical Evaluation Challenges

TREC Clinical Decision Support / Precision Medicine Track

Results submission:

- Up to 5 runs per participants, containing a ranked list of up to 1000 docids.

Pooling:

- all documents retrieved in ranks 120 by any run
- in union with a 20% sample of documents not retrieved in the first set that were retrieved in ranks 21100 by some runs

Relevance Judgement:

- 0: not relevant, 1: possibly relevant, and 2: definitely relevant.
- Assessors: physicians, graduate biomedical students at OHSU
- 34,949 documents were judged across the topics, with a mean of 1265.0 documents judged per topic

Evaluation measures:

- infNDCG and P@10

TREC Medical Evaluation Challenges

TREC Clinical Decision Support / Precision Medicine Track

Lessons learned:

- 2014:
  - topic-type-specific processing in their runs, for example by emphasizing particular MeSH terms related to the topic type when those terms were found in documents
- 2015:
  - Markov Random Fields built from concepts in the description and top retrieved documents
  - combine different information retrieval models along with semantic annotations from DBpedia
  - customized term extraction based on MetaMap (Aronson and Lang, 2010) concept semantic types
- 2016:
  - Query expansion with selected concepts from UMLS
  - Document Distances from Unsupervised Word Embeddings
  - Clinical Causal Relationships Using UMLS and Wikipedia
- 2017:
  - Ontology-based filtering and retrieval
  - Topic analysis and expansion using a medical knowledge graph

CLEF Medical Evaluation Challenges

ImageCLEF

ImageCLEF is one of the oldest CLEF tasks and has been running for 15 years [Mueller et al., 2010]. More information can be found at `http://www.imageclef.org/`

Year	Tasks involving text	Tasks not involving text
2003-2013	Medical retrieval: - subtask adhoc retrieval Topics: images Documents: images with multilingual textual case notes - subtask case-based retrieval (2010-2013) Topics: medical cases description + images Documents: full-text biomedical articles	Medical image annotation (2004-2009) Medical retrieval (2010-2013): - subtask modality detection - subtask compound figure separation (2013) Medical user-oriented image retrieval (2011)
2014-2015		Medical image annotation Medical image clustering Liver CT annotation
2016		Compound figure detection Medical image classification Caption prediction
2017-2018		ImageCLEF tuberculosis

CLEF Medical Evaluation Challenges
ImageCLEF

Here we focus on adhoc and case-based retrieval task [Kalpathy-Cramer et al., 2014].

Results submission:

- Participants could submit results using only the images, only the text, or both

Evaluation measures:

- MAP, BPref, P@5, 10, 30

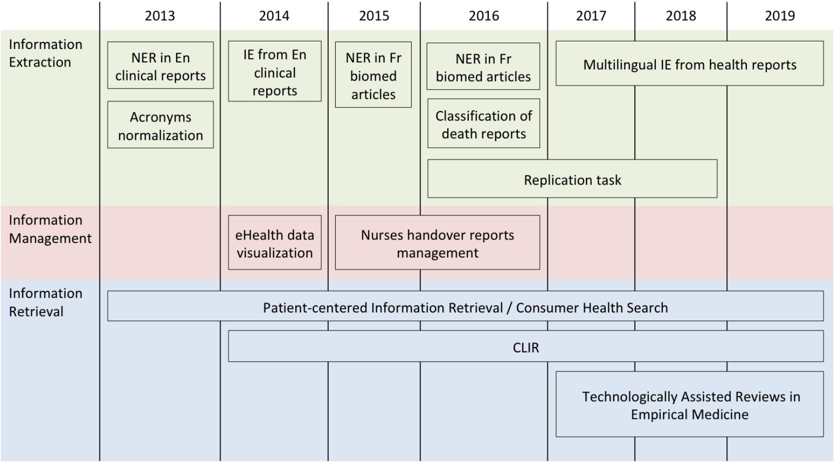
What worked well?

- Using medical knowledge to enrich case reports or expand the queries
- Manually annotated collections have been used to utilize medical knowledge in connection with the original text
- MetaMap used to map biomedical text to the Unified Medical Language System Metathesaurus (UMLS)
- WordNet and EuroWordNet used for term expansion

CLEF Medical Evaluation Challenges

CLEF eHealth

Running since 2013 and proposed a variety of tasks focusing on health-related information extraction and retrieval:



[Kelly et al., 2014, Goeuriot et al., 2015, Kelly et al., 2016, Goeuriot et al., 2017]

CLEF Medical Evaluation Challenges

CLEF eHealth

The patient-centered information retrieval task (2013-2018)

- **Purpose:** retrieval of relevant documents for consumer health search:
- **Datasets**:
  - **Documents**:
    - Set of medical articles and certified documents (2013-2015)
    - Large web crawl (2016-2017)
  - **Topics**: expression of a patient information need

Year	Topics	Example
2013-2014	Disorder extracted from a discharge summary Topic contains: query, description, narrative, patient profile	Asystolic arrest
2015	Description of an image Topic contains: query, image	Weird brown patches on skin
2016-2017	Manually built query from a discussion forum	What causes strong headaches at base of skull, stops with blood donation

CLEF Medical Evaluation Challenges

CLEF eHealth

The patient-centered information retrieval task (2013-2018)

	2013	2014	2015	2016, 2017
Goal	Help laypersons better understand medical reports		Layperson checking their symptoms	
Topics	55 EN topics built from discharge summaries	55 EN topics + translation in CZ, DE, FR	67 EN topics built from images + AR, CZ, DE, FA, FR, IT, PT	300 EN topics built from forum posts + CZ, FR, HU, DE, PO, SW
Docs	Medical document collection provided by Khresmoi project			ClueWeb 12 B13
Relevance judgement	Manual evaluation of relevance of documents	Manual evaluation of relevance and readability of documents		Manual evaluation of relevance, readability and trustworthiness

CLEF Medical Evaluation Challenges

CLEF eHealth

Results submission:

- Participants can submit sevral ranked runs, including a baseline

Pooling:

- Top priority runs are pooled

Relevance Judgement:

- In 2013-2014: only topical relevance was judged
- In 2015: topicality + readability
- In 2016-2017: topicality + readability + reliability

Evaluation measures:

- In 2013-2017: P@10, MAP, NDCG
- In 2015-2017: rank-biased precision, including readability and reliability

What worked well?

- Query expansion methods to close the lexical gap
- Hybrid indexing (terms and concepts)
- Acronym expansion and normalization

CLEF Medical Evaluation Challenges

CLEF eHealth

The Technologically Assisted Reviews in Empirical Medicine Task (2017-2018) [Kanoulas et al., 2017]

- **Purpose**: develop methods to retrieve relevant studies with high precision and high recall
- **Dataset:**
  - **Topics:** 20+30 topics for Diagnostic Test Accuracy (DTA) systematic reviews (training + test sets)
  - **Documents: MEDLINE database documents**

```
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
    ...
```

CLEF Medical Evaluation Challenges

CLEF eHealth

The Technologically Assisted Reviews in Empirical Medicine Task (2017-2018) Results submission:

- Runs with ranked MEDLINE abstracts

Relevance Judgement:

- Assessment was made based on systematic reviews provided by Cochrane Library

Evaluation measures:

- AP, NDCG, recall

What worked well?

- Expansion using implicit (MeSH, UMLS) or explicit semantics (Word Embeddings)

BioASQ Evaluation Challenge

- BioASQ is challenge tackling large-scale biomedical semantic indexing and question answering.
- Running since 2013, in CLEF and then in other venues (BioNLP/ACL, EMNLP)
- In 2019, 3 tasks:
  - Large-Scale Online Biomedical Semantic Indexing (running every year)
  - Biomedical semantic QA (running every year)
  - Medical Semantic Indexing In Spanish

Semantic indexing

Goal: Assign MeSH concepts to biomedical abstracts (to be included in PubMed), i.e. multi-label classification
Data: annotated PubMed abstract (training dataset), non-annotated PubMed abstracts (test dataset)
Metrics: F-measure and variations
What works well:

- Incorporating deep semantic representations into classical approaches works well (doc2vec, word2vec)
- Neural networks approaches facing big data/small data issue (multiclasses problem)

Biomedical QA

Goal: Find the relevant answer to a question
Two steps: find relevant articles (IR), find the answer (QA + summarization)
Data: Questions, document set
Metrics: MAP, F-measure, accuracy, ROUGE
What works well:

- Neural networks are giving best results for answer generation (LSTM, reinforcement learning...)
- Performance strongly relies on the type of question (Yes/no, factoid...)
- Task has to evolve in parallel with the community

Summary of the Benchmarking Activities

Venue	Task	Dataset	Activity
TREC	Genomics adhoc retrieval	Clinical information need Biomedical articles	Terminated
	Genomics passage retrieval	Clinical information need Biomedical articles	Terminated
	Medical records	Patient cohort search	Terminated
	Clinical decision support / Precision medicine	Case reports Biomedical articles	Ongoing
CLEF	ImageCLEF medical retrieval	Image and medical reports Collection of medical images	Terminated
	CLEF eHealth consumer search	Health information need Large web crawl	Ongoing
	CLEF eHealth technological assisted reviews	Boolean queries Biomedical articles	Ongoing
* *	BioASQ	Annotated biomedical abstracts and QA dataset	Ongoing

The majority of these datasets are still available and can be used for research!

CONCLUSION

- *A large and growing body of work on Information Retrieval in the medical domain*
  - Focus on task, user profile, information need elicitation in context (time, task, user's expertise, etc.)
  - Model semantic w.r.t. polyrepresentation view: document collections, knowledge bases, users, etc.
  - Shift from lexical matching to semantic matching by considering domain-specific peculiarities
  - Understand relevance assessment facets according to task, user (laypeople vs.expert)
  - Increasing amount of tasks due to a higher collaboration with the medical community
- *Challenges ahead*
  - IR at the service of the medical community:
    - Domain-driven IR models for medical search: Are IR heuristics similar to medical search heuristics?
    - Complex tasks not solved yet
    - Effort in structuring knowledge: dynamic terminologies, languages other than English...
    - Conceptualization of medical search tasks: eg., model hypothetico-deductive approach of medical experts?
  - Neural networks for all
    - Need for big data
    - Black-box: towards explainability
  - Large-scale evaluation under privacy-constraints

Discussion

References I

Amati, G. (2003).
Probabilistic models for Information Retrieval based on Divergence from Randomness.
PhD thesis, University of Glasgow.

Aronson, A. R. and Lang, F.-M. (2010).
An overview of metamap: historical perspective and recent advances.
Journal of the American Medical Informatics Association, 17(3):229–236.

Balaneshin-kordan, S. and Kotov, A. (2016).
Optimization method for weighting explicit and latent concepts in clinical decision support queries.
In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 241–250, New York, NY, USA. ACM.

Banerjee, I., C. Chen, M., Lungren, M., and Rubin, D. (2017).
Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest ct cohort.
Journal of Biomedical Informatics, 77.

Bast, H., Björn, B., and Haussmann, E. (2016).
Semantic search on text and knowledge bases.
Found. Trends Inf. Retr., 10(2-3):119–271.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., and Zhou, J. (2017).
Patient subtyping via time-aware lstm networks.
In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 65–74.

Blackman, S. (2006).
Nature has wikipedia in its cities.
The scientist, 20(2).

References II

Blum, B. (1984).
Information Systems for Patient Care.
Springer-Verlag.

Borlund, P. (2003).
The concept of relevance in ir.
J. Am. Soc. Inf. Sci. Technol., 54(10):913–925.

Boudin, F., Nie, J.-Y., and Dawes, M. (2010).
Clinical information retrieval using document and pico structure.
In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 822–830, Stroudsburg, PA, USA. Association for Computational Linguistics.

Broder, A. (2002).
A taxonomy of web search.
SIGIR Forum, 36(2):3–10.

Cai, X., Gao, J., Ngiam, K. Y., Ooi, B. C., Zhang, Y., and Yuan, X. (2018).
Medical concept embedding with time-aware attention.
In *IJCAI*.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. (2011).
Askhermes: An online question answering system for complex clinical questions.
Journal of Biomedical Informatics, 44(2):277 – 288.

Cartright, M.-A., White, R. W., and Horvitz, E. (2011a).
Intentions and attention in exploratory health search.
In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 65–74.

References III

Cartright, M.-A., White, R. W., and Horvitz, E. (2011b).
Intentions and attention in exploratory health search.
In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 65–74, New York, NY, USA. ACM.

Chapman, W., Bridewell W, H. P., GF, C., and BG, B. (2001).
Evaluation of negation phrases in narrative clinical reports.
In *Proceedings of the AMIA Symposium*, pages 105–109.

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., and Sun, J. (2016).
Multi-layer representation learning for medical concepts.
In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1495–1504.

De Choudhury, M., Morris, M. R., and White, R. W. (2014).
Seeking and sharing health information online: Comparing search engines and social media.
In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1365–1376.

De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., and Bruza, P. (2014).
Medical semantic similarity with a neural language model.
In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1819–1822, New York, NY, USA. ACM.

Dinh, D. and Tamine, L. (2012).
Towards a context sensitive approach to searching information based on domain specific knowledge sources.
Web Semantics: Science, Services and Agents on the World Wide Web, 12(0).

Dinh, D., Tamine, L., and Boubekeur, F. (2013).
Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies.
Artificial Intelligence in Medicine, 57(2):155 – 167.

References IV

Edinger, T., Cohen, A., Bedrick, S., Ambert, K., and Hersh, W. (2012).
AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2012:180–188.

Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L., and Evans, E. R. (1999).
Analysis of questions asked by family doctors regarding patient care.
BMJ, 319:358–361.

Ghosh, S., Chakraborty, P., Cohn, E., Brownstein, J. S., and Ramakrishnan, N. (2016).
Characterizing diseases from unstructured text: A vocabulary driven word2vec approach.
In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1129–1138.

Ghosh, S., Chakraborty, P., Cohn, E., Brownstein, J. S., and Ramakrishnan, N. (2017).
Joint learning of representations of medical concepts and words from ehr data.
In *Proceedings (IEEE Int Conf Bioinformatics Biomed)*, Biomed '17, pages 764–769.

Goeuriot, L., Kelly, L., Hanlen, L., Suominen, H., Névéol, A., Palotti, J., and Zuccon, G. (2015).
Overview of the CLEF eHealth Evaluation Lab 2015.
In *Proceedings of CLEF 2015*.

Goeuriot, L., Kelly, L., Suominen, H., Nvol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., and Zuccon, G. (2017).
Clef 2017 ehealth evaluation lab overview.
In *Proceedings of CLEF 2017*.

Goodwin, T. R. and Harabagiu, S. M. (2016).
Medical question answering for clinical decision support.
In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 297–306, New York, NY, USA. ACM.

References V

Gorman, P. (1995).
Inforamtion needs of physicians.
Journal of the American Society for Information Science, 46:729–736.

Guo, J., Fan, Y., Ai, Q., and Croft, W. B. (2016).
A deep relevance matching model for ad-hoc retrieval.
In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 55–64, New York, NY, USA. ACM.

Henry, S., Cuffy, C., and McInnes, B. T. (2018).
Vector representations of multi-word terms for semantic relatedness.
Journal of Biomedical Informatics, 77:111 – 119.

Hersh, W. (2010).
Information Retrieval - A Health and Biomedical Perspective.
Springer.

Hersh, W. and Bhupatiraju, R. T. (2003).
TREC genomics track overview.
In *Proceedings of the Text REtrieval Conference*.

Hersh, W., Cohen, A., Yang, J., Bhupatiraju, R. T., Roberts, P., and Hearst, M. (2005).
TREC 2005 genomics track overview.
In *Proceedings of the Text REtrieval Conference*.

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006).
TREC 2006 genomics track overview.
In *Proceedings of the Text REtrieval Conference*.

References VI

Hersh, W. and Voorhees, E. (2009).
Trec genomics special issue overview.
12:1–15.

Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., and Kraemer, D. F. (2004).
TREC 2004 genomics track overview.
In *Proceedings of the Text REtrieval Conference*.

Huang, X., Lin, J., and Demner-Fushman, D. (2006).
Evaluation of pico as a knowledge representation for clinical questions.
In *AMIA Annual Symposium Proceedings*, pages 359–363.

Jagannatha, A. and Yu, H. (2016).
Structured prediction models for RNN based sequence labeling in clinical text.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 856–865.

Jarvelin, K. and Kekalainen, J. (2000).
IR evaluation methods for retrieving highly relevant documents.
In *Proceedings of the 23rd SIGIR conference*, pages 41–48.

Jones, K. S., Walker, S., and Robertson, S. E. (2000).
A probabilistic model of information retrieval: Development and comparative experiments.
Inf. Process. Manage., 36(6):779–808.

Kalpathy-Cramer, J., de Herrera, A. G. S., Demner-Fushman, D., Antani, S., Bedrick, S., and Mller, H. (2014).
Evaluating performance of biomedical image retrieval systems - an overview of the medical image retrieval task at imageclef 2004-2013.
Computerized Medical Imaging and Graphics, 39:55–61.

References VII

Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2017).
Clef 2017 technologically assisted reviews in empirical medicine overview.
In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF) - Online working notes*.

Kelly, D. et al. (2009).
Methods for evaluating interactive information retrieval systems with users.
Foundations and Trends® in Information Retrieval, 3(1–2):1–224.

Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., and Zuccon, G. (2016).
Overview of the CLEF eHealth Evaluation Lab 2016.
In *Proceedings of CLEF 2016*.

Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D. L., Velupillai, S., Chapman, W. W., Zuccon, G., and Palotti, J. (2014).
Overview of the ShARe/CLEF eHealth Evaluation Lab 2014.
In *Proceedings of CLEF 2014*.

Koopman, B. and Zuccon, G. (2014).
Why assessing relevance in medical ir is demanding.
In *Proceedings of the Medical IR workshop, in conjunction with the 2014 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Gold Coast, Australia. ACM.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., and Lawley, M. (2016).
Information retrieval as semantic inference: A graph inference model applied to medical search.
Inf. Retr., 19(1-2):6–37.

Lancaster, F. and Warner, A. (1993).
Information Retrieval Today.

References VIII

Leveling, J., Goeuriot, L., Kelly, L., and Jones, G. J. (2012).
Dcu@ trecmed 2012: Using ad-hoc baselines for domain-specific retrieval.
Technical report, DUBLIN CITY UNIV (IRELAND).

Limsopatham, N. and Collier, N. (2016).
Normalising medical concepts in social media texts by learning semantic representation.
In *Proceedings of the International Conference on the Association for Computational Linguistics*, ACL '16, pages 1014–1023.

Limsopatham, N., Macdonald, C., and Ounis, I. (2013).
Learning to handle negated language in medical records search.
In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1431–1440, New York, NY, USA. ACM.

Liu, X. and Croft, W. B. (2004).
Cluster-based retrieval using language models.
In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 186–193, New York, NY, USA. ACM.

Liu, X., Nie, J.-Y., and Sordoni, A. (2016).
Constraining word embeddings by prior knowledge – application to medical information retrieval.
pages 155–167.

Loza Mencía, E., de Melo, G., and Nam, J. (2016).
Medical concept embeddings via labeled background corpora.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4629–4636.

Martinez, D., Otegi, A., Soroa, A., and Agirre, E. (2014).
Improving search over electronic health records using umls-based query expansion through random walks.
Journal of Biomedical Informatics, 51:100 – 106.

References IX

Metzler, D. and Croft, W. B. (2005).
A markov random field model for term dependencies.
In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 472–479, New York, NY, USA. ACM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
CoRR, abs/1301.3781.

Minarro-Gimnez, J., O, M.-A., and M, S. (2014).
Exploring the application of deep learning techniques on medical text corpora.
In *Studies in health technology and in informatics*, pages 205:584–8.

Moen, H., Ginter, F., Marsi, E., Peltonen, L.-M., T, S., and S., S. (2015).
Care episode retrieval: distributional semantic models for information retrieval in the clinical domain.
BMC Medical Informatics and Decision Making, 15.

Montori, V. M., Wilczynski, N. L., Morgan, D., and Haynes, R. B. (2004).
Systematic reviews: a cross-sectional study of location and citation count.
BMC Medicine, 1(2).

Mueller, H., Clough, P., Deselaers, T., and Caputo, B. (2010).
ImageCLEF Experimental evaluation of visual information retrieval.
Springer.

Natarajan, K., Stein, D., Jain, S., and Elhadad, N. (2010).
An analysis of clinical queries in an electronic health record search utility.
International Journal of Medical Informatics, 79(7):515–522.

REFERENCES X

Nguyen, G.-H., Tamine, L., Soulier, L., and Souf, N. (2017).
Learning Concept-Driven Document Embeddings for Medical Information Search.
In *Conference on Artificial Intelligence in Medicine (AIME 2017)*, Vienna, Austria.

Nguyen, G.-H., Tamine, L., Soulier, L., and Souf, N. (2018a).
A Tri-Partite Neural Document Language Model for Semantic Information Retrieval.
In *ESWC 2018 - 15th European Semantic Web Conference*, volume 10843 of *Lecture Notes in Computer Science*, pages 445–461, Heraklion, Crète, Greece. Springer.

Nguyen, G.-H., Tamine, L., Soulier, L., and Souf, N. (2018b).
A Tri-Partite Neural Document Language Model for Semantic Information Retrieval.
In *Extended Semantic Web Conference (ESWC 2018)*, Vienna, Austria.

Ni, J., Liu, J., Zhang, C., Ye, D., and Ma, Z. (2017).
Fine-grained patient similarity measuring using deep metric learning.
In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1189–1198.

Oh, H.-S. and Jung, Y. (2015).
Cluster-based query expansion using external collections in medical information retrieval.
Journal of Biomedical Informatics, 58:70 – 79.

Palotti, J., Hanbury, A., Müller, H., and Kahn, Jr., C. E. (2016).
How users search and what they search for in the medical domain.
Inf. Retr., 19(1-2):189–224.

Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., and Shanfeng Zhu1, . (2016).
Deepmesh: deep semantic representation for improving large-scale mesh indexing.
Bioinformatics, 32.

References XI

Ponte, J. M. and Croft, W. B. (1998).
A language modeling approach to information retrieval.
In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281.

Rindflesch, T. C. and Fiszman, M. (2003).
The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.
Journal of biomedical informatics, 36(6):462–477.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., and Hersh, W. R. (2016).
Overview of the trec 2016 clinical decision support track.
In *Proceedings of the Text REtrieval Conference*.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A. J., and Pant, S. (2017).
Overview of the trec 2017 precision medicine track.
In *Proceedings of the Text REtrieval Conference*.

Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., and Hersh, W. (2015a).
State-of-the art in biomedical literature rerieval for clinical cases: a survey of the trec 2014 cds track.
Information retrieval.

Roberts, K., Simpson, M. S., Voorhees, E., and Hersh, W. R. (2015b).
Overview of the trec 2015 clinical decision support track.
In *Proceedings of the Text REtrieval Conference*.

Salton, G. and McGill, M. J. (1986).
Introduction to Modern Information Retrieval.
McGraw-Hill, Inc., New York, NY, USA.

References XII

Salton, G., Wong, A., and Yang, C. S. (1975).
A vector space model for automatic indexing.
Commun. ACM, 18(11):613–620.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010).
Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications.
Journal of the American Medical Informatics Association, 17(5):507–513.

Simpson, M. S., Voorhees, E. M., and Hersh, W. (2014).
Overview of the trec 2014 clinical decision support track.
In *Proceedings of the Text REtrieval Conference*.

Soldaini, L. and Goharian, N. (2016).
Quickumls: a fast, unsupervised approach for medical concept extraction.
In *MedIR workshop, sigir*.

Soldaini, L. and Goharian, N. (2017).
Learning to rank for consumer health search: A semantic approach.
In Jose, J. M., Hauff, C., Altıngovde, I. S., Song, D., Albakour, D., Watt, S., and Tait, J., editors, *Advances in Information Retrieval*, pages 640–646, Cham. Springer International Publishing.

Soldaini, L., Yates, A., and Goharian, N. (2017).
Learning to reformulate long queries for clinical decision support.
Journal of the Association for Information Science and Technology, 68(11):2602–2619.

Soldaini, L., Yates, A., Yom-Tov, E., Frieder, O., and Goharian, N. (2016).
Enhancing web search in the medical domain via query clarification.
Inf. Retr., 19(1-2):149–173.

References XIII

Sondhi, P., Sun, J., Zhai, C., Sorrentino, R., and Kohn, M. (2012).
Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries.
Journal of the American Medical Informatics Association: JAMIA, 19(5):851–858.

Stanton, I., Ieong, S., and Mishra, N. (2014).
Circumlocution in diagnostic medical queries.
In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 133–142, New York, NY, USA. ACM.

Stojanovic, J., Gligorijevic, D., Radosavljevic, V., Djuric, N., Grbovic, M., and Obradovic, Z. (2017).
Modeling healthcare quality via compact representations of electronic health records.
IEEE/ACM Trans. Comput. Biol. Bioinformatics, 14(3):545–554.

Sushil, M., uster, S. ., Luyckx, K., and Daelemans, W. (2018).
Patient representation learning and interpretable evaluation using clinical notes.
Journal of Biomedical Informatics, 84:103 – 113.

Tamine, L. and Chouquet, C. (2017).
On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings.
Inf. Process. Manage., 53(2):332–350.

Tamine, L., Chouquet, C., and Palmer, T. (2015).
Analysis of biomedical and health queries: Lessons learned from trec and clef evaluation benchmarks.
JASIST, 66(12):2626–2642.

Voorhees, E. (1998).
Variations in relevance judgments and the measurement of retrieval effectiveness.
In *Proceedings of the 21st SIGIR conference*, pages 315–323.

REFERENCES XIV

Voorhees, E. M. and Hersh, W. (2012).
Overview of the trec 2012 medical records.
In *Proceedings of the Text Retreival Conference TREC*. NIST.

W. White, R. and Horvitz, E. (2009).
Cyberchondria: Studies of the escalation of medical concerns in web search.
ACM Trans. Inf. Syst., 27.

Wang, C. and Akella, R. (2015).
Concept-based relevance models for medical and semantic information retrieval.
In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 173–182, New York, NY, USA. ACM.

Weerkamp, W., Balog, K., and de Rijke, M. (2012).
Exploiting external collections for query expansion.
ACM Trans. Web, 6(4):18:1–18:29.

White, R. and Horvitz, E. (2010).
Web to world: predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search.
In *AMIA Annu Symp Proc*, pages 882–6.

White, R. and Horvitz, E. (2014).
From health search to healthcare: explorations of intention and utilization via query logs and user surveys.
J Am Med Inform Assoc, 21(1):4955.

White, R. W. and Horvitz, E. (2013).
Captions and biases in diagnostic search.
ACM Trans. Web, 7(4):23:1–23:28.

References XV

Wilkinson, R. and Fuller, M. (1996).
Integration of information retrieval and hypertext via structure.
Information Retrieval and Hypertext, pages 257–271.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008).
Listwise approach to learning to rank: Theory and algorithm.
In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1192–1199, New York, NY, USA. ACM.

Xiong, C. and Callan, J. (2015).
Esdrank: Connecting query and documents through external semi-structured data.
In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 951–960, New York, NY, USA. ACM.

Zhai, C. and Lafferty, J. (2001).
Model-based feedback in the language modeling approach to information retrieval.
In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.

Zhang, Y., Wang, P., Heaton, A., and Winkler, H. (2012).
Health information searching behavior in medlineplus and the impact of tasks.
In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI'12, pages 641–650.

Zhang, Y., Zhang, J., Lease, M., and Gwizdka, J. (2014).
Multidimensional relevance modeling via psychometrics and crowdsourcing.
In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 435–444, New York, NY, USA. ACM.

Zhu, D. and Carterette, B. (2012).
Improving health records search using multiple query expansion collections.
In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–7.

References XVI

Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., and Wang, F. (2016).
Measuring patient similarities via a deep architecture with medical concept embedding.
In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 749–758.

Znaidi, E., Tamine, L., and Latiri, C. (2015).
Answering pico clinical questions: A semantic graph-based approach.
In Holmes, J. H., Bellazzi, R., Sacchi, L., and Peek, N., editors, *Artificial Intelligence in Medicine*, pages 232–237, Cham. Springer International Publishing.

Znaidi, E., Tamine, L., and Latiri, C. (2016).
Aggregating semantic information nuggets for answering clinical queries.
In *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016*, pages 1041–1047.

Zuccon, G. and Koopman, B. (2018).
Choices in knowledge-base retrieval for consumer health search.
In *Proceedings of ECIR'2018*. LNCS.