Inf Retrieval J
DOI 10.1007/s10791-015-9259-x

MEDICAL INFORMATION RETRIEVAL

State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track

Kirk Roberts[1] · **Matthew Simpson**[1] · **Dina Demner-Fushman**[1] · **Ellen Voorhees**[2] · **William Hersh**[3]

Received: 15 December 2014 / Accepted: 17 June 2015

Abstract Providing access to relevant biomedical literature in a clinical setting has the potential to bridge a critical gap in evidence-based medicine. Here, our goal is specifically to provide relevant articles to clinicians to improve their decision-making in diagnosing, treating, and testing patients. To this end, the TREC 2014 Clinical Decision Support Track evaluated a system's ability to retrieve relevant articles in one of three categories (Diagnosis, Treatment, Test) using an idealized form of a patient medical record. Over 100 submissions from over 25 participants were evaluated on 30 topics, resulting in over 37k relevance judgments. In this article, we provide an overview of the task, a survey of the information retrieval methods employed by the participants, an analysis of the results, and a discussion on the future directions for this challenging yet important task.

Keywords Biomedical information retrieval · Clinical decision support · Information retrieval evaluation

1 Introduction

In the course of caring for patients, physicians often seek out information to improve their clinical decisions. Relevant medical information can help improve this decision-making by providing recent, evidence-based analysis. Common information sought by physicians includes questions about potential diagnoses given a list of symptoms, the most effective treatment given what is known about a patient's condition, and the best test to conduct to

✉ Kirk Roberts
kirk.roberts@nih.gov

[1] Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

[2] National Institute of Standards and Technology, Gaithersburg, MD, USA

[3] Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Published online: 18 July 2015

determine the best care for the patient. In some cases, the best source for this type of information is the published biomedical literature. A significant barrier to the use of the biomedical literature, however, is its large volume and rapid growth. While many tools exist for searching this literature (Lu 2011), few target the clinical environment, and standardized datasets do not exist for evaluating information retrieval (IR) systems designed for this task.

In order to make biomedical information more accessible and to meet the requirements for the meaningful use of electronic health records (EHRs), a goal of modern clinical decision support systems is to anticipate the needs of physicians by linking EHRs with information relevant for patient care. The 2014 TREC Clinical Decision Support (CDS) Track aimed to simulate the requirements of such systems and to encourage the creation of tools and resources necessary for their implementation.

The focus of the 2014 track was the retrieval of biomedical articles relevant for answering generic clinical questions about medical records. Since a reusable, de-identified collection of medical records is not easily available, the track utilized short case reports, such as those published in biomedical articles, as idealized representations of medical records. A case report typically describes a challenging medical case, and it is organized as a well-formed narrative summarizing the portions of a patient's medical record that are pertinent to making decisions about the patient's care. The case reports used in the CDS track were more representative of typical medical cases, though still sufficiently challenging as to merit a literature search. In this article, we demonstrate the difference between an actual clinical record and the cases used in the CDS track.

In this article, we provide an overview of the track and insights into the nature of clinical IR in Sect. 3, including (1) the biomedical literature dataset used in the track, (2) how the topics were developed and the difference between the topics and actual clinical records, (3) how retrieval results were evaluated, (4) data on the participation in the track, and (5) an overview of the relevance assessment process. In Sect. 4, we provide a detailed survey of the methods employed by the many participants in the track. In the initial year of the track, participants combined standard IR techniques (e.g., ranking models, query expansion) with methods designed specifically for clinical queries (e.g., recognition of important patient attributes, utilization of medical ontologies and NLP tools, and categorization of biomedical articles based on their utility to the clinical setting). In Sect. 5, we present the results of the track, including (1) the top-scoring submissions for each participant, (2) statistical significance evaluation for the top participants, and (3) analysis on how results varied between topics. In Sect. 6, we discuss the lessons learned and existing challenges with retrieving literature in the clinical setting, including (1) a comparison of the top-performing systems, (2) our overall impression of which retrieval strategies worked best and how that might change in future systems, (3) an analysis of the types of articles determined as relevant, as opposed to the types of articles in the collection, and (4) future challenges and directions for literature retrieval in the clinical setting. By providing an overview of the TREC 2014 CDS task, surveying the methods used, and analyzing its results, it is our express intention to motivate even further advances in the state-of-the-art for this important task.

2 Background

The Clinical Decision Support Track was designed to complement previous biomedical-inspired TREC tasks, specifically, the Genomics and Medical Records tracks.

The TREC Genomics Track (2003–2007) was designed to retrieve relevant scientific articles to aid biological research. The task in the Genomics track included ad hoc retrieval

(Hersh and Bhupatiraju 2003; Hersh et al. 2004, 2005), summarization (Hersh and Bhupatiraju 2003), categorization of biological articles by function (Hersh et al. 2004, 2005), passage retrieval (Hersh et al. 2006), and question answering (Hersh et al. 2007). Despite its conclusion 8 years ago, the Genomics track continues to serve as a basis for further biomedical IR research for advancing the state-of-the-art (Hu et al. 2011; Zhang et al. 2011; Wang et al. 2012; Kim and Cohen 2013), creating new biomedical IR resources (Ryu and Choi 2013), and the evaluation of IR results (An and Cercone 2014).

Subsequently, the TREC Medical Records Track (2011–2012) was designed to retrieve records from an EHR for patients matching a given condition and demographic description, referred to as cohort retrieval (Voorhees and Tong 2011; Voorhees and Hersh 2012). Example cohort descriptions include "children with dental caries" and "adults under age 60 undergoing alcohol withdrawal". Systems were then tasked with retrieving hospital visit records of matching patients. The track was discontinued after 2012 due to the lack of availability of a sufficiently large, publicly available clinical dataset with unstructured notes. While it is no longer possible to conduct research on the datasets created by the Medical Records track, the track clearly demonstrated that language use in EHRs is sufficiently different from general use to warrent domain-specific processing in clinical IR systems.

The 2014 CDS track was designed to fill in a gap in these two tasks, where literature articles are retrieved (similar to the Genomics track), but the queries are the medical records themselves (whereas the medical records are the retrieval results in the Medical Records track). The task can then be viewed either as an ad hoc retrieval task (where a medical record is created expressly to query the literature) or a passive retrieval task (where relevant literature is pre-retrieved and consulted if necessary). Furthermore, since the query is an idealized medical record, IR systems must deal with the myriad of problems presented by clinical text (Chapman and Cohen 2009; Demner-Fushman et al. 2009). To some degree, these issues are alleviated by utilizing cases instead of the EHR record directly, but most of the language understanding, representation, querying, and ranking issues involved with clinical text must still be addressed by the participants in the track.

The most closely related medical IR tasks outside of TREC have been organized as part of the CLEF eHealth labs. The ShARe/CLEF eHealth Task 3 (Goeuriot et al. 2013, 2014) focused on retrieving medical information from the web for consumers (patients), usually in connection with the patient's discharge summary for context. Also of note is the related ImageCLEFmed task (Kalpathy-Cramer et al. 2014), where cases were also used as a query into the literature, but the focus was on retrieving medical images, and the BioCreative interactive task (Arighi et al. 2011), which focuses on finding scientific articles for genes. Table 1 lists the tasks described above, as well as their reported participation in terms of the number of participants (as opposed to the number of runs). As can be seen in the table, there has been a sustained, high interest in participating in medical IR challenges for over a decade, and the 2014 TREC CDS track had a similar level of participation as the previous tracks.

3 Task overview

3.1 Document collection

The document collection for the CDS track was the open access subset[1] of PubMed Central[2] (PMC). PMC is an online digital database of freely available full-text biomedical

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[2] http://www.ncbi.nlm.nih.gov/pmc/

Table 1 Number of participants in the given medical IR challenges

Task	Year	# Participants
TREC genomics	2003	25
	2004	27
	2005	32
	2006	30
	2007	27
TREC medical records	2011	29
	2012	24
ImageCLEF medical	2012	17
	2013	10
CLEF eHealth	2013	9
	2014	14
TREC CDS	2014	26

For the challenges with non-IR tasks, only the IR-related tasks are reported here

literature. Because documents are constantly being added to PMC, a snapshot of the open access subset (January 21, 2014) was used. This ensured the consistency of the collection for track participants and future researchers. The snapshot contained a total of 733,138 articles. The full text of each article in the open access subset is represented as an NXML file (XML encoded using the US National Library of Medicine's Journal Archiving and Interchange Tag Library[3]). Images and supplemental materials were also available. The total size of the snapshot (without images and supplemental materials) is 9.5 GB.

Each article in the collection is identified by a unique number (PMCID) that was used for run submissions and relevance judgments. The PMCID of an article is specified by the `<article-id >` element within its NXML file. Although each article is represented by multiple identifiers (e.g., PubMed, PMC, Publisher, etc.), only the PMCIDs were used for this task.

3.2 Topics

The topics for the track were medical case reports created by clinical informatics experts (all of whom were physicians) at the US National Library of Medicine. The descriptions of the topics emulated a patient sign-out (information communicated about the patient in the hand-off process—the process of transferring patient care from one clinician to the other[4]) or presenting a patient in a teaching environment, for example, in clinical rounds. By doing this, the goal was to replicate the types of information contained in EHR notes, thus providing as near as possible a realistic evaluation of how such a retrieval system would perform in a clinical environment. The case reports therefore contained some or all of the following elements, depending on the envisioned stage of the hospital encounter: patient's medical history, the patient's current symptoms, tests performed by a physician to diagnose the patient's condition, the patient's eventual diagnosis, and the steps taken by a physician to treat the patient.

Knowing that questions related to diagnoses, treatments, and tests account for a majority (58 %) of the clinical questions posed by primary care physicians (Del Fiol et al. 2014),

[3] http://jats.nlm.nih.gov/archiving/versions.html

[4] http://psnet.ahrq.gov/primer.aspx?primerID=9

the topic creators labeled the case reports they constructed according to these three categories to indicate what question type will most probably arise at this stage. For example, consider the following notes:

1. This is a 62 year-old woman with IDDM, COPD, interstitial lung disease, h/o IVDU on methadone maintenance, hepatitis C, gastric varices, depression, anxiety with recent 40 lb weight loss in setting of dysphagia with thus far unremarkable work up.
2. This 59-year-old African American gentleman initially presented due to dysphagia and weight loss. At that time, he had a barium swallow, which showed a pinpoint narrowing of his distal esophagus. He had endoscopy and underwent dilatation of this stricture. He did not have much improvement with the dilatation.

These real-life examples are partially-redacted extracts from MIMIC-II (Scott et al. 2013), and were transformed into topic 19:

> **TOPIC 19:** A 52-year-old African American man with a history of heavy smoking and drinking, describes progressive dysphagia that began several months ago. He first noticed difficulty swallowing meat. His trouble swallowing then progressed to include other solid foods, soft foods, and then liquids. He is able to locate the point where food is obstructed at the lower end of his sternum. He has lost a total of 25 pounds.

At this time, the clinicians will consider what further tests are needed to diagnose the cause of dysphagia and weight loss. In the cases that served as prototype for the question, the following tests were performed:

1. Work up for weight loss and dysphagia have included recent EGD which demonstrated gastritis and Barrett's esophagus, CT abdomen which demonstrated an isolated pulmonary nodule, gastric varices and a renal cyst. Additionally, barium swallow demonstrated only hiatal hernia.
2. He had a CT scan which showed a 1.5 cm gastrohepatic lymph node. He underwent an upper endoscopy on which they saw distal esophageal narrowing. They also performed multiple biopsies of the area of narrowing. Of note, they saw some ulceration in the GE junction and a thick abnormal fold concerning for esophageal or gastric cardia cancer. The biopsy showed moderate to poorly differentiated adenocarcinoma. After this he underwent endoscopic ultrasound, however, they were unable to pass the ultrasound probe beyond the stricture.

For this Test topic, the participants were expected to retrieve articles that would suggest relevant interventions that a physician might undertake in diagnosing the patient. A Diagnosis topic required participants of the track to retrieve PMC articles a physician would find useful for determining the diagnosis of the patient described in the report. Similarly, for a Treatment topic, participants retrieved articles that would suggest to a physician the best treatment plan for the condition exhibited by the patient described in the report. Figure 1 illustrates the relationships between the findings in a case description and the three topic types. When constructing the case-based topics, the topic creators were careful to omit direct answers to the questions. For example, a Diagnosis topic might have contained information pertaining to a patient's treatments and tests, but not the patient's diagnosis. This was hoped to more accurately mimic real clinical scenarios. The topic creators produced 10 topics for each of the 3 topic types for a total of 30 topics.

In addition to annotating the topics according to the type of clinical information required, participants were also provided two versions of the case reports. The topic "description" contains a reasonably complete account of the patient's visit, including details such as vital statistics, drug dosages, etc., whereas the topic "summary" was a

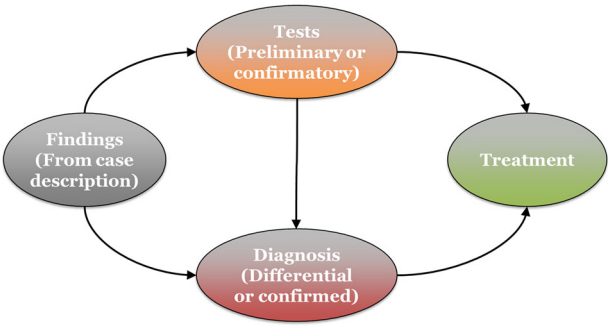

Fig. 1 Relationship between the topic types. The *arrows* represent the clinical decision-making process. Notably, Test topics cover both tests to determine a diagnosis or to find the best treatment

simplified version of the report that contained less irrelevant information. A topic's description and its summary were functionally equivalent: the set of relevant documents was identical for each version. However, the summary versions of the case reports were provided for participants who were not interested in or equipped for processing the detailed descriptions. The process of summarizing the topic descriptions involved using the topic creator's medical knowledge and clinical expertise to remove redundant and non-essential information. For instance, in Topic 19 above, the description notes the patient's progressive dysphagia, then states his difficulty swallowing meats, followed by other solid foods, and finally liquids. This process of having increasing difficulty swallowing foods over time is the chief symptom of progressive dysphagia, and as such is largely redundant. The summary for Topic 19, therefore, is able to condense several sentences down to only the disorder progressive dysphagia:

> **TOPIC 19 (SUMMARY):** 52-year-old man with history of smoking and heavy drinking, now with progressive dysphagia and 25-pound weight loss.

Table 2 shows six topics, two from each topic type, with the description and corresponding summary. All topics are available on the track website.[5]

3.3 Evaluation

The evaluation of the track followed standard TREC evaluation procedures for ad hoc retrieval tasks. Participants were allowed to submit in `trec_eval` format a maximum of five automatic or manual runs per topic, each consisting of a ranked list of up to one thousand PMCIDs.

The primary metric for comparing the retrieval submissions was inferred normalized discounted cumulative gain (infNDCG). Additionally, runs were scored using precision at rank 10 (P@10), inferred average precision (infAP), and R-Precision. Inferred measures are used as a means of getting more accurate estimates of a run's quality than is likely possible with traditional measures when judging a relatively small number of documents (Yilmaz et al. 2008). The submitted runs were sampled following an effective sampling strategy (Voorhees 2014) for computing inferred measures. In particular, judgment sets were created using two strata: all documents retrieved in ranks 1–20 by any run in union

[5] http://www.trec-cds.org

Table 2 Six of the 30 topics from the TREC 2014 CDS track

ID	Type	Topic
2	Diagnosis	**Description:** An 8-year-old male presents in March to the ER with fever up to 39 C, dyspnea and cough for 2 days. He has just returned from a 5 day vacation in Colorado. Parents report that prior to the onset of fever and cough, he had loose stools. He denies upper respiratory tract symptoms. On examination he is in respiratory distress and has bronchial respiratory sounds on the left. A chest X-ray shows bilateral lung infiltrates
		Summary: 8-year-old boy with 2 days of loose stools, fever, and cough after returning from a trip to Colorado. Chest X-ray shows bilateral lung infiltrates
5	Diagnosis	**Description:** A 56-year-old female on 20th day post-left mastectomy presents to the emergency department complaining of shortness of breath and malaise. The patient says that she has remained in bed for the last 2 weeks. The physical examination reveals tenderness on the left upper thoracic wall and right calf. The surgical incision shows no bleeding or signs of infection. Pulmonary auscultation is significant for bilateral decreased breath sounds, especially at the right base. Laboratory tests reveal an elevated D-dimer
		Summary: 56-year-old woman presents with shortness of breath 3 weeks after surgical mastectomy. Physical exam is significant for right calf tenderness and decreased breath sounds at the right base. Her D-dimer level is elevated
12	Test	**Description:** A 25-year-old woman presents to the clinic complaining of prolonged fatigue. She denies difficulty sleeping and sleeps an average of 8 h a night. She also notes hair loss, a change in her voice and weight gain during the previous 6 months. She complains of cold intolerance. On examination she has a prominent, soft, uniform anterior cervical mass at the midline
		Summary: 25-year-old woman with fatigue, hair loss, weight gain, and cold intolerance for 6 months
18	Test	**Description:** A 6-month-old male infant has a urine output of less than 0.2 mL/kg/hr shortly after undergoing major surgery. On examination, he has generalized edema. His blood pressure is 115/80 mm Hg, his pulse is 141/min, and his respirations are 18/min. His blood urea nitrogen is 33 mg/dL, and his serum creatinine is 1.3 mg/dL. Initial urinalysis shows specific gravity of 1.017. Microscopic examination of the urine sample reveals 1 WBC per high-power field (HPF), 18 RBCs per HPF, and 5 granular casts per HPF. His fractional excretion of sodium is 3.3 %.
		Summary: 6-month-old male with decreased urine output and edema several hours after surgery. He is hypertensive and tachycardic, has a high BUN and creatinine, and urine microscopy reveals red blood cells and granular casts
22	Treatment	**Description:** A 15-year-old girl presents to the ER with abdominal pain. The pain appeared gradually and was periumbilical at first, localizing to the right lower quadrant over hours. She has had no appetite since yesterday but denies diarrhea. She has had no sexual partners and her menses are regular. On examination, she has localized rebound tenderness over the right lower quadrant. On an abdominal ultrasound, a markedly edematous appendix is seen
		Summary: 15-year-old girl with right lower quadrant abdominal pain, decreased appetite, and enlarged appendix on abdominal ultrasound
23	Treatment	**Description:** A 63-year-old man presents with cough and shortness of breath. His past medical history is notable for heavy smoking, spinal stenosis, diabetes, hypothyroidism and mild psoriasis. He also has a family history of early onset dementia. His symptoms began about a week prior to his admission, with productive cough, purulent sputum and difficulty breathing, requiring him to use his home oxygen for the past 24 h. He denies fever. On examination he is cyanotic, tachypneic, with a barrel shaped chest and diffuse rales over his lungs. A chest X-ray is notable for hyperinflation with no consolidation
		Summary: 63-year-old heavy smoker with productive cough, shortness of breath, tachypnea, and oxygen requirement. Chest X-ray shows hyperinflation with no consolidation

with a 20 % sample of documents not retrieved in the first set that were retrieved in ranks 21–100 by some run. Documents in the judgment set were judged on a three-point scale of 0 (not relevant), 1 (possibly relevant), and 2 (definitely relevant). For the evaluation reported here, the measures were computed by conflating the possibly relevant and definitely relevant sets into a single relevant set. The exception to this is the infNDCG measure, which makes use of the different relevance grades.

3.4 Participation

The track received a total of 105 runs from 26 different groups. A list of the participating groups, along with the number of runs per group, is shown in Table 3. Not all participants submitted description papers, and those without system descriptions are omitted from the survey of features in Sect. 4. Included in the set of 105 runs was a set of 8 runs from a single group, 3 of which were excluded since the track had a limit of 5 runs. All of the remaining 102 runs contributed to the judgment sets, which were constructed to be compatible with computing the inferred retrieval measures. Of the 102 runs, three were

Table 3 Participating groups and submitted runs

Group	Affiliation	References	# Runs
atigeo	Atigeo	Wei et al. (2014)	5
BigPig	University of Michigan	Joo and Sohn (2014)	3
BiTeM_SIBtex	University of Applied Sciences, Geneva	Gobeill et al. (2014)	5
CSEIITV	Indian Institute of Technology, Varanasi	Singh and Chowdary (2014)	2
cuhk_sls	The Chinese University of Hong Kong	Wan et al. (2014)	4
DA_IICT	Dhirubhai Ambani IICT	Sankhavara et al. (2014)	5
DawitAfshin	York University	Girmay and Deroie (2014)	4
ECNUCS	East China Normal University	Li et al. (2014)	4
Georgetown	Georgetown University	Wing and Yang (2014)	5
georgetown_ir	IR Lab, Georgetown University	Soldaini et al. (2014)	5
HENRI_TUDOR	CRP Henri Tudor	Dinh and Ben Abacha (2014)	4
hltcoe	Johns Hopkins University HLTCOE	Xu et al. (2014)	4
IKMLAB	Institute of Medical Informatics, NCKU		1
ir.cs.sfsu	San Francisco State University	Bhandari and Kulkarni (2014)	1
KISTI	Korea ISTI	Oh and Jung (2014)	8
LIMSI	LIMSI-CNRS	D'hondt et al. (2014)	5
Merck_DA	Merck KGaA		3
NovaSearch	Universidade Nova Lisboa	Mourão et al. (2014)	5
OHSU	Oregon Health & Science University		4
Philips	Philips Research North America	Hasan et al. (2014)	1
SNUMedinfo	Medical Informatics Laboratory	Choi and Choi (2014)	5
super_kxlab	bupt-kxlab	Xue et al. (2014)	1
TUW	Vienna University Of Technology	Palotti et al. (2014)	5
UCLA_MII	Medical Imaging Informatics, UCLA	Garcia-Gathright et al. (2014)	4
udel_fang	InfoLab Group, University of Delaware	Wang and Fang (2014)	5
UTDHLTRI	University of Texas at Dallas	Goodwin and Harabagiu (2014)	4

Only 5 of the KISTI runs were scored

manual runs,[6] while 99 were automatic runs. For the remainder of this article, only the automatic runs are considered. In total, 2,864,988 articles were returned in the 102 runs (691,288 unique topic/article pairs). A total of 37,949 of these, based on the sampling strategy defined above, were considered for assessment.

3.5 Assessment

Relevance assessment was done as in many past TREC biomedical tasks using physicians who were either graduate students in the OHSU biomedical informatics program or otherwise working or training in the informatics field at other institutions. The physician-experts were tasked with providing relevance assessments for the articles retrieved for each of the 30 topics. Based on past experience, potential assessors were told that the judging process takes on average about 1 min per article, or about 21 (range 15–27) hours total work for judging an average-sized pool of 1269 articles. Once assessors signed up, they were sent instructions for the assessment process, including how to log on to the assessment system, screenshots of the system, and instructions for assigning assessments.

Assessors were instructed not to think of the topics as posing questions that require a single "correct" answer. They were told that for an article to be relevant, they must find it useful in addressing the generic question posed by the given case. This means that it had to (a) provide information of importance to the question type (Diagnosis, Test, or Treatment), and (b) provide information that is topical to the patient. It was up to the assessors to determine how well an article fits these criteria.

An article that suggested a particular diagnosis, test, or treatment or a set of potential diagnoses, tests, or treatments that sounded reasonable to the medical expert given the information available in the case should have been judged "definitely relevant". Relevant articles did not necessarily have to address "textbook" presentations, but had to contain enough meaningful information for an expert to find the suggested diagnoses plausible. On the other hand, if the article suggested a diagnosis that obviously was not appropriate for the given case according to the assessors' medical expertise, or if the article did not even describe a medical condition at all, it should have been assessed as "definitely not relevant". Articles that did not fall strictly into the above two categories were to be judged as "potentially relevant". For example, an article could be considered potentially relevant if the expert is unsure whether or not the condition the article describes applies to the given case, but feels there is a reasonable chance that it might if there was only more information available. Relevance assessments were captured by a Web-based system whose interface is shown in Fig. 2.

In order to assess the replicability of judging, eight topics were assigned to two assessors to measure relevance overlap as well as the κ-statistic, a common measure of

[6] Based on the submission form, there were 11 manual runs. One participant submitted 4 manual runs, but no system description paper. Of the 7 verified manual runs, 1 participant (Girmay and Deroie 2014) submitted 4 manual runs and no automatic runs, while the other participants each submitted 1 manual run in addition to their automatic runs. Dinh and Ben Abacha (2014) actually performed worse than their best automatic run by manually expanding and weighting the query; Garcia-Gathright et al. (2014) used an expert to manually tweak the queries, resulting in the second overall run in infNDCG and the first overall run in P@10; Wan et al. (2014) also used an expert to manually edit the queries, but this system would only have resulted in a run at the median score. To our knowledge, none of the manual runs used experts to manually filter the results, and instead only used experts to build queries. In this article, these 3 manual runs from teams who also submitted manual runs are not included, while the 4 manual runs of Girmay and Deroie (2014) are included due to the lack of automatic runs.

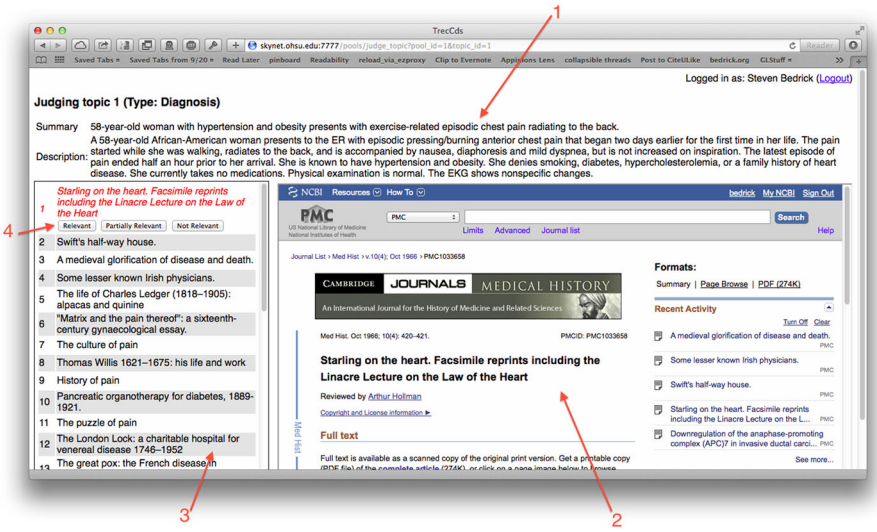

Fig. 2 Features of the assessing interface included (*1*) text of the topic's description and summary, (*2*) article text, direct from PubMed Central, (*3*) list of articles that need to be reviewed, and (*4*) relevance selection buttons

Table 4 Agreement between assessors for the eight double-judged topics

Topic	NN	NR	RR	RN	Overlap	κ
1	1349	32	35	47	0.307	0.442
5	1369	1	14	119	0.104	0.174
12	838	17	114	508	0.178	0.183
17	1040	53	13	6	0.181	0.286
19	977	25	70	134	0.306	0.404
25	1351	70	28	6	0.269	0.404
27	437	17	296	158	0.628	0.614
28	1070	10	35	17	0.565	0.709
Overall					0.332	0.438

inter-observer agreement. The κ-statistic has been used in the past for assessments for a number of ad hoc medical retrieval challenge evaluations, listed below. The results of this evaluation, seen in Table 4, indicate that relevance judgment for this task is quite difficult: assessors were only able to achieve an overlap of 0.332 and a κ of 0.44 on the eight topics. These are, however, consistent with previous medical IR evaluations:

- Medline abstracts: 0.43 (Hersh et al. 1994b)
- Medline abstracts: 0.41 (Hersh et al. 1994a)
- Medical textbook: 0.37 (Hersh and Hickam 1995)
- TREC 2004 Genomics track: 0.51 (Hersh et al. 2004)
- TREC 2005 Genomics track: 0.59 (Hersh et al. 2005)
- ImageCLEFmed images: 0.74 (Hersh and Kim 2006)

4 Retrieval approaches

Participants in the track took a wide variety of approaches for retrieving relevant articles. Here, we break down the retrieval process into coarse-grained modules in order to compare individual elements of these approaches. Note that each participating system likely has implementation details that seem similar on the surface, yet lead to dramatically different results. The aim of this section, therefore, is not to judge the worthiness of individual retrieval features, but to provide a survey of potential methods. While cross-system comparisons based on performance are intentionally avoided, when an individual system experiments with and without a particular feature, we do report whether that feature had a positive, negative, or neutral impact on the system. A high-level interpretation of some of the features is provided in the Discussion (Sect. 6). Tables 5 and 6 note the high-level design decisions of each run based on each participant's system description paper. The numbers that follow only reflect participants that submitted a system description, while some changes were made based on personal communication with the authors.

4.1 Topic representation

For each run, participants could decide to use either the description (long form) or summary (short form) of the case, 12 participants only submitted runs using the description, 5 only submitted runs using the summary, while 6 experimented with both topic representations. In total, 56 runs used the description, while 31 runs used the summary. Of the participants that experimented with both the description and summary, three performed better using the summary (Sankhavara et al. 2014; Singh and Chowdary 2014; Xu et al. 2014), two performed better using the description (Choi and Choi 2014; Wei et al. 2014), and one had mixed results that depended heavily on the evaluation metric (Dinh and Ben Abacha 2014). It is likely that the choice of description or summary is highly linked with the downstream representation and ranking methods: a description provides more keywords which can be difficult to weight properly, while a summary provides fewer and likely requires better query expansion.

4.2 Indexing

For the most part, participants used publicly available indexing software packages to build their retrieval index(es). The most popular were Lucene (Hatcher and Gospodnetic 2004) (9 participants), Terrier (Ounis et al. 2006) (7 participants), and Indri (Strohman et al. 2005) (5 participants). 1 participant used Xapian[7]. 1 participant used Haircut (McNamee et al. 2002), which is specifically designed to perform character n-gram searches (see Sect. 4.5). Finally, 1 participant (D'hondt et al. 2014) used PubMed, then filtered out results that were not in PubMed Central. Presumably, these systems used PubMed's relevance ranking instead of the default chronological ordering. Most indexing software supports multiple ranking models (see Sect. 4.4). While the ranking model is what theoretically governs the quality of results, index implementations often make different underlying assumptions so their effect cannot simply be dismissed. Due to the extensive capabilities of each indexing package, optimal use of an individual package depends

[7] http://xapian.org/

Table 5 Retrieval features used by the participant systems

System	Run	Topic		Index				Field		Model			
		Description	Summary	Indri	Lucene	Terrier	Other	Full Text	Other	TF-IDF	BM25	LM	InExp
Bhandari and Kulkarni	1	x		x				x				x	
Choi and Choi	1		x	x				x				x	
	2		x	x				x				x	
	3		x	x				x				x	
	4	x		x				x				x	
	5	x		x				x				x	
D'hondt et al.	1		x			x		x			x		
	2		x				x		x				
	3		x				x		x				
	4		x			x		x			x		
	5		x			x		x			x		
Dinh and Ben Abacha	1	x				x		x		x	x		x
	2		x			x		x		x	x		x
	3	x				x		x		x	x		x
	4		x			x		x		x	x		x
Garcia-Gathright et al.	1		x		x			x		x			
	2		x		x				x			x	
	3		x		x			x				x	
	4		x		x			x				x	
Girmay and Deroie	1		x			x		x			x		
	2		x			x		x			x		
	3		x			x		x					x
	4		x			x		x					x

Table 5 continued

System	Run	Topic		Index				Field		Model			
		Description	Summary	Indri	Lucene	Terrier	Other	Full Text	Other	TF-IDF	BM25	LM	InExp
Gobeill et al.	1	x				x		x			x		
	2	x				x		x			x		
	3	x				x			x		x		
	4	x				x			x		x		
	5	x				x			x		x		
Goodwin and Harabagiu	1	x			x			x			x		
	2	x			x			x			x		
	3	x						x					
	4	x			x			x			x		
Hasan et al.	1	x							x				
Joo and Sohn	1	x			x			x			x		
	2	x			x			x			x		
	3	x			x			x			x		
Li et al.	1	x		x		x		x		x	x		
	2	x		x		x		x			x		
	3	x		x		x		x		x			
	4	x		x		x		x		x	x		
Mourão et al.	1	x			x			x			x		
	2	x			x			x			x		
	3	x			x			x			x		
	4	x			x			x		x	x	x	
	5	x			x			x			x		

Table 5 continued

System	Run	Topic		Index				Field		Model			
		Description	Summary	Indri	Lucene	Terrier	Other	Full Text	Other	TF-IDF	BM25	LM	InExp
Oh and Jung	1	x			x			x				x	
	2	x			x			x				x	
	3	x			x			x		x		x	
	4	x			x			x		x		x	
	5	x			x			x		x		x	

System	Run	Terms			CE		Negation	AE		QE		Article Pref.
		Words	Concepts	Other	MetaMap	Lexicon		Gender	Age	Lexicon	PRF	
Bhandari and Kulkarni	1		x		x					x		x
Choi and Choi	1	x									x	x
	2	x									x	x
	3	x									x	x
	4	x									x	x
	5	x									x	x
D'hondt et al.	1	x										x
	2		x		x	x		x	x	x		x
	3		x		x	x		x	x	x		x
	4	x			x	x				x		x
	5	x			x	x				x		x
Dinh and Ben Abacha	1	x									x	
	2	x									x	
	3	x									x	
	4	x									x	

Table 5 continued

System	Run	Terms			CE		Negation	AE		QE		Article Pref.
		Words	Concepts	Other	MetaMap	Lexicon		Gender	Age	Lexicon	PRF	
Garcia-Gathright et al.	1	x	x		x		x					x
	2	x	x		x		x					x
	3	x	x		x		x					x
	4	x	x		x		x	x	x			x
Girmay and Deroie	1	x										
	2	x								x		
	3	x										
	4	x								x		
Gobeill et al.	1	x	x			x				x		x
	2	x	x			x				x		x
	3	x	x			x				x		x
	4	x	x			x				x		x
	5	x	x			x				x		x
Goodwin and Harabagiu	1		x			x						
	2		x			x				x		
	3			x								
	4		x			x				x		
Hasan et al.	1		x			x				x		
Joo and Sohn	1		x		x							
	2		x		x							
	3		x		x							
Li et al.	1	x						x	x	x		
	2	x						x	x	x		
	3		x				x	x	x	x		
	4	x	x				x	x	x	x		

Table 5 continued

System	Run	Terms			CE		Negation	AE		QE		Article Pref.
		Words	Concepts	Other	MetaMap	Lexicon		Gender	Age	Lexicon	PRF	
Mourão et al.	1	x									x	
	2	x									x	x
	3	x									x	x
	4	x									x	
	5	x									x	
Oh and Jung	1	x										
	2	x									x	
	3	x	x			x	x				x	
	4	x	x		x		x				x	
	5	x	x		x		x				x	

CE concept extraction, *AE* attribute extraction, *QE* query expansion, *PRF* pseudo-relevance feedback, *Article Pref.* preference/bias for specific article types

Table 6 Retrieval features used by the participant systems

System	Run	Topic		Index				Field		Model			
		Description	Summary	Indri	Lucene	Terrier	Other	Full Text	Other	TF-IDF	BM25	LM	InExp
Palotti et al.	1	x		x				x				x	
	2	x			x			x		x			
	3	x					x	x			x		
	4	x		x	x		x	x		x	x	x	
	5	x						x					
Sankhavara et al.	1	x				x		x					x
	2	x				x		x					x
	3		x			x		x					x
	4	x				x		x					x
	5	x				x		x					x
Singh and Chowdary	1	x			x			x		x			
	2		x		x			x		x			
Soldaini et al.	1	x			x				x				
	2	x			x				x				
	3	x			x				x				
	4	x			x				x				
	5	x			x				x				
Wan et al.	1		x			x		x		x			
	2		x			x		x		x			
	3		x			x		x		x			
	4		x			x		x		x			
Wang and Fang	1		x					x					
	2		x					x					
	3		x					x					
	4		x					x					
	5		x					x					

Table 6 continued

System	Run	Topic		Index				Field		Model			
		Description	Summary	Indri	Lucene	Terrier	Other	Full Text	Other	TF-IDF	BM25	LM	InExp
Wei et al.	1		x	x	x			x		x		x	
	2	x		x	x			x		x		x	
	3		x	x				x				x	
	4	x		x				x				x	
	5		x	x				x				x	
Wing and Yang	1	x		x				x				x	
	2	x		x				x				x	
	3	x		x				x				x	
	4	x		x				x				x	
	5	x		x				x				x	
Xue et al.	1	x						x		x			
Xu et al.	1		x				x	x				x	
	2		x				x	x				x	
	3	x					x	x				x	
	4		x				x	x				x	

System	Run	Terms			CE		Negation	AE		QE		Article Pref.
		Words	Concepts	Other	MetaMap	Lexicon		Gender	Age	Lexicon	PRF	
Palotti et al.	1	x	x		x	x						
	2	x	x		x	x						
	3	x	x		x	x						
	4	x	x		x	x						
	5		x	x	x	x						

Table 6 continued

System	Run	Terms			CE		Negation	AE		QE		Article Pref.
		Words	Concepts	Other	MetaMap	Lexicon		Gender	Age	Lexicon	PRF	
Sankhavara et al.	1	x									x	
	2	x									x	
	3	x									x	
	4	x									x	
	5	x									x	
Singh and Chowdary	1	x										
	2	x										
Soldaini et al.	1	x									x	x
	2	x						x	x		x	x
	3	x									x	x
	4	x	x		x						x	
	5	x	x		x			x	x		x	x
Wan et al.	1	x										x
	2	x	x			x						
	3	x	x			x						
	4	x	x			x				x		
Wang and Fang	1		x		x		x				x	
	2		x		x		x			x	x	
	3	x									x	
	4		x		x		x			x		
	5	x										
Wei et al.	1	x			x		x		x	x		
	2	x			x		x		x	x		
	3	x			x		x		x	x		
	4	x			x		x		x	x		
	5	x			x		x		x	x		

Table 6 continued

System	Run	Terms			CE		Negation	AE		QE		Article Pref.
		Words	Concepts	Other	MetaMap	Lexicon		Gender	Age	Lexicon	PRF	
Wing and Yang	1		x		x		x			x		
	2		x		x		x			x		
	3		x		x		x			x		
	4		x		x		x			x		
	5		x		x		x			x		
Xue et al.	1									x		
Xu et al.	1			x								
	2			x							x	
	3			x							x	
	4	x									x	

See Table 5 footnote for abbreviation definitions

heavily on a participant's experience with that package. As such, performance analysis of participants that used multiple packages is omitted from this article.

4.3 Search fields

The standard information retrieval method involves indexing and searching entire documents, and this is what most participants did. Some participants, however, experimented with searching just the abstracts or sections of the article. Garcia-Gathright et al. (2014) experimented with both full text and abstract searching and concluded that using abstracts degraded performance. This is an important consideration that might outweigh the benefits of searching just the abstract. Similarly, Gobeill et al. (2014) experimented with indexing and searching individual sections, thinking it might reduce noise. However, results from those runs significantly under-performed the full document runs. Finally, D'hondt et al. (2014) experimented with searching entirely with MeSH terms and ignoring the actual article text. Given the fact that many articles do not have MeSH terms, this approach proved detrimental to recall.

4.4 Ranking model

Participants used a wide variety of scoring functions to rank results. 9 participants used the standard TF-IDF vector space with cosine similarity. 8 participants used variants of BM25 [Spärck Jones et al. 2000], including the standard BM25, BM25F [used by Li et al. (2014)], BM25L (used by Mourão et al. (2014)), and BM25+ (used by Mourão et al. (2014)). 8 participants used a unigram language model approach, usually with Dirichlet prior smoothing, though Garcia-Gathright et al. (2014) and Xu et al. (2014) used Jelinek-Mercer smoothing (Jelinek and Mercer 1980). 2 participants used an Inverse Expected Document Frequency (In_exp) model: Dinh and Ben Abacha (2014) used B2 and Sankhavara et al. (2014) used C2. Finally, Dinh and Ben Abacha (2014) used LGD (Clinchant and Gaussier 2010) and Girmay and Deroie (2014) used InL2c1 (Amati and Van Rijsbergen 2002). Additionally, many systems tried ensemble approaches to combine one or more ranking approaches (e.g., Dinh and Ben Abacha (2014), Li et al. (2014), Mourão et al. (2014), Sankhavara et al. (2014), Wei et al. (2014)).

4.5 Term representation

Deciding which terms in the query and document should be included for indexing and retrieval is one of the most critical considerations in developing an IR system. The two most common choices for the 2014 CDS track were to index words or concepts. A word-based representation (often referred to as a bag-of-words) typically removes stopwords and removes word inflection via stemming. A concept representation typically employs a lexicon or model (see Sect. 4.6) to identify concepts. This method typically identifies the most important terms, but often misses key terms that do not fit within the concept model. 8 participants only submitted runs using word representations, 5 only submitted runs using concept representations, while 9 experimented with both (either in separate runs or the same run). In total, 62 runs used a word representation and 43 runs used a concept representation. Additionally, 2 participants (4 runs) used entirely different representations, discussed below.

The systems that experimented with both representations reported mixed results. D'hondt et al. (2014) experimented with MeSH concepts, achieving their best run over a word baseline. However, it is difficult to isolate whether the improved performance is due to the concepts or other features of that run, while both the baseline and concept runs show poor overall performance. Gobeill et al. (2014) used both words and MeSH concepts on every run, while Palotti et al. (2014) and Wei et al. (2014) use both words and UMLS concepts. Presumably, words that are part of a MeSH/UMLS concept were removed and replaced with a normalized concept ID that could be used to identify synonyms. Li et al. (2014) experimented with words and UMLS concepts, achieving their best infNDCG using only UMLS concepts, but the best P@10 by combining. Oh and Jung (2014) used a baseline word representation and then re-ranked using concepts in 3 of their 5 runs. They achieved similar performance to the baseline when the run used MetaMap, but worse when using Wikipedia concepts. Soldaini et al. (2014) submitted 1 run using UMLS concepts and one combined run. The concept-only run was their second best run, while the combined run was weighed down by a separate, low-performing method. Wang and Fang (2014) used a concept representation based on UMLS in 3 of their 5 runs. All 3 runs significantly outperformed the 2 word-based runs.

Additionally, two other types of term representations were used. First, distributional representations were used by two participants: Palotti et al. (2014) submitted a run that used word vectors based on `word2vec`[8], while Goodwin and Harabagiu (2014) submitted a run using Latent Dirichlet Allocation (Blei et al. 2003). Distributional representations have shown great promise in advancing the state-of-the-art in natural language processing. However, both these runs performed significantly lower than both participants' other runs, suggesting further refinements are necessary before it can be used for this IR task. Xu et al. (2014) used character n-grams in 3 of their 4 runs. This method is supported within the HAIRCUT system, and is designed to deal with morphological variations in low-resource languages. Their best run used the character n-grams, outperforming a word-based run that otherwise had the exact same features, and performing as one of the top overall runs in the task despite its simple, knowledge-poor approach.

4.6 Concept extraction

Two main approaches were used to extract concepts for the systems that used concept representations. The first approach is lexicon-based, where a pre-built lexicon identifies the list of valid concepts. The second approach uses natural language information extraction to identify text spans that match a concept from a lexicon, but not necessarily an exact string match (as is required in the lexicon method). Of the 10 systems that utilized the second approach, all used MetaMap (Aronson and Lang 2010) to identify terms within UMLS (Lindberg et al. 1993). Of the lexicons in the first approach, 3 participants used MeSH[9] [(namely Gobeill et al. (2014), Palotti et al. (2014), and Wan et al. (2014)]; Wan et al. (2014) also used ICD-10[10]; D'hondt et al. (2014) used both the Disease Symptom Knowledge Base (Wang et al. 2008) and OrphaNet (Weinreich et al. 2008), Hasan et al. (2014) used LOINC (McDonald et al. 2003), RxNorm (Liu et al. 2005), and SNOMED-CT (Stearns et al. 2001); while Goodwin and Harabagiu (2014) and Oh and Jung (2014) used Wikipedia. In total, 34 runs utilized MetaMap while 19 used lexicons for concept extraction.

[8] https://code.google.com/p/word2vec/

[9] http://www.nlm.nih.gov/mesh/

[10] http://www.who.int/classifications/icd/en/

4.7 Negation detection

Detecting negation is very important in natural language processing, though its importance in IR is somewhat mixed. In this task, 6 participants chose to include a negation component in a total of 22 runs. Only 1 participant experimented both with and without negation.

To detect the negations, Li et al. (2014), Wei et al. (2014), and Wing and Yang (2014) used NegEx (Chapman et al. 2001), while Garcia-Gathright et al. (2014) and Oh and Jung (2014) used MetaMap, and Wang and Fang (2014) didn't specify. In handling negations, both Garcia-Gathright et al. (2014) and Li et al. (2014) simply pruned negated concepts, Oh and Jung (2014) and Wing and Yang (2014) adjusted weights to prefer polarity-matched concepts, and Wang and Fang (2014) and Wei et al. (2014) replaced negated concepts with an altered representation. In this last case, *fever* could be replaced with *nxfever*, so only polarity-matched concepts will be found (an all-or-nothing version of the weighting strategy). While all of Wei et al. (2014)'s runs used negation in some respect, their runs were each ensembles of several component systems, and 3 of their 5 runs utilized negation more heavily than the other 2 runs. It appears that the runs that relied on negation more performed better than those with fewer negation-enabled ensemble components.

4.8 Attribute extraction

Certain patient attributes are often clinically important. For example, young patients may receive entirely different treatments than older patients, while some diagnoses are only relevant in men or women. Several participants tried to capitalize on this to aid their retrieval models. Both age and gender were used by D'hondt et al. (2014), Garcia-Gathright et al. (2014), Li et al. (2014), and Soldaini et al. (2014). Additionally, Wei et al. (2014) captured age only and Soldaini et al. (2014) additionally captured race information. In total, 14 runs utilized age, 9 runs utilized gender, and 2 runs utilized race.

Each system utilized its own rule-based technique to extract attribute information. Both Wei et al. (2014) and Li et al. (2014) replaced attribute-related terms with a normalized form. Both did this for every run so its impact on performance isn't clear. Both D'hondt et al. (2014) and Garcia-Gathright et al. (2014) utilized attribute-related MeSH terms associated with articles by, respectively, hard filtering and weight-boosting articles based on their MeSH terms. (For example, if a given topic is associated with the MeSH term *Colonoscopy*, then a hard filtering approach would remove all articles from the results that did not have the *Colonoscopy* term assigned, while a weight-boosting approach would improve the rank of articles with the *Colonoscopy* term assigned.) For both participants, the scores decreased compared to other runs, though for D'hondt et al. (2014) these runs are not necessarily comparable and for Garcia-Gathright et al. (2014) the scores were actually increased for Treatment and Test topics, but significantly hurt for Diagnosis topics. Soldaini et al. (2014) adjusted article scores based on attribute matches, resulting in their best run.

4.9 Query expansion

Expanding queries by adding potentially relevant terms is a common practice in IR systems, with the goal of adding synonyms, near-synonyms, and other related terms to increase recall of relevant documents. Almost every participant (19 of 23) utilized some form of query expansion. In each case, the expansion methods fall into one of two

categories. The first category involves the use of lexicons and other ontological resources. These approaches combine well with the concept extraction methods described in Sect. 4.6, as the same resources used to identify concepts are also a source of synonyms. The second category utilizes a method known as pseudo-relevance feedback (PRF), which is known to find potentially relevant terms by first querying the index and looking for new terms in high-ranking documents. This method can add very noisy terms, but might also find related terms not available in a static resource. 13 participants utilized lexicons for expansion, while 8 participants utilized PRF. Only 1 participant used both (Wang and Fang 2014).

A wide variety of lexicons were used for query expansion. Bhandari and Kulkarni (2014), Goodwin and Harabagiu (2014), Wang and Fang (2014), Wei et al. (2014), and Wing and Yang (2014) used UMLS. Gobeill et al. (2014) used MeSH. Girmay and Deroie (2014) used SNOMED-CT. Xue et al. (2014) used WordNet (Fellbaum 1998). Wan et al. (2014) used ICD-10. Li et al. (2014) used a (presumably custom) set of abbreviation expansions. Goodwin and Harabagiu (2014) used an automatically built lexicon based on similar terms using `word2vec`. Finally, D'hondt et al. (2014) and Hasan et al. (2014) both attempted to guess the underlying disease for Treatment and Test topics, then utilize the disease(s) as additional query keywords (essentially performing an automatic diagnosis). They utilized knowledge sources to help map from a set of symptoms to a set of hypothetical diseases.

The impact of lexicons for query expansion was generally positive, but demonstrated precautions are necessary. D'hondt et al. (2014) demonstrated a substantial boost using their disease hypothesis method. Girmay and Deroie (2014) reported a large performance increase when using the InL2c1 ranking model, but a small drop in performance when using BM25. Goodwin and Harabagiu (2014) improved on a baseline system using lexicons, but did much worse with the `word2vec`-based lexicon. Wan et al. (2014) incurred a significant performance drop over a comparable method without expansion. Finally, Wang and Fang (2014) reported their best result on a method that utilizes both PRF and UMLS-based expansion, and it appears that UMLS contributed more to the performance.

Most participants that utilized PRF ran an initial query on the indexed articles, extracted terms from the top N articles, gave them a reduced weight, then re-applied the expanded query to get the final document set. Choi and Choi (2014), however, used the top articles to learn relevant MeSH terms, while Wang and Fang (2014) ran their initial queries on the Cases Database.[11] Mourão et al. (2014) largely used the standard approach but experimented with adjusting expanded term weight using the impact factor for the journal each article was published in.

The impact of PRF for query expansion was quite positive, especially for the more traditional forms of PRF. Xu et al. (2014) demonstrated increased performance with PRF, where their best run used PRF and significantly outperformed the most comparable non-PRF run. Oh and Jung (2014) reported a small boost in runs that used PRF. As stated above, Wang and Fang (2014) reported their best run combines PRF with UMLS-based expansion, though it appears the UMLS expansion likely had the greater impact. Finally, Mourão et al. (2014) used PRF in every run, but only used their journal impact factor weighting in one run. It appears this method substantially damaged performance when compared to a run that used more traditional PRF.

[11] http://www.casesdatabase.com/ (Now offline)

4.10 Article preference

An important consideration for this task was in addressing the topic type (Diagnosis, Treatment, Test). Generally, participants that tried to alter their rankings based on the topic type did so by weighting or filtering certain types of articles. 8 participants submitted a total of 27 runs with some form of article preference strategy.

Bhandari and Kulkarni (2014), Garcia-Gathright et al. (2014), Soldaini et al. (2014), and Wan et al. (2014) developed lexicons of diagnoses, treatments, and tests and added these as lower-weight keywords for each corresponding topic. D'hondt et al. (2014) filtered based on MeSH terms, while Gobeill et al. (2014) and Mourão et al. (2014) used a MeSH weighting strategy. Finally, Choi and Choi (2014) and Soldaini et al. (2014) trained machine learning classifiers to categorize articles. Choi and Choi (2014) trained their classifier on the Clinical Hedges Database (Haynes et al. 2005), while Soldaini et al. (2014) manually multi-labeled 1170 articles as a training set.

Three participants experimented with their article preference features in different runs. Mourão et al. (2014) reported that MeSH weighting reduced infNDCG based on a comparable baseline. The lexicon approach used by Soldaini et al. (2014) was their middle-performing run, while their classifier approach was by far their worst performing run. Finally, Wan et al. (2014) also reported that their lexicon approach degraded performance. These results indicate that attempting to bias ranking results based on the topic type can be very detrimental if not done correctly. The overall idea, however, seems valid: some articles are more likely to discuss diagnoses, treatments, and tests. Given the relevance judgments from 2014, appropriate training data will perhaps be available to enable this strategy to be effective.

4.11 Other approaches

A number of features used by the participants do not fit into the above categories, either because they are entirely novel for the task or were only attempted by one or two participants. We highlight two such approaches here.

Gobeill et al. (2014) observed that certain article types are inherently more useful for this task, regardless of the query or topic type. This is indeed the case, as Table 7 shows:

This is similar to the intuition behind the article preference strategies, though we define article preference to relate to the topic type, while the article type largely is independent of the specific information need. One of the main goals of this task is to be able to differentiate between what articles are (narrowly) clinically relevant and what articles are (broadly) biomedically relevant. As such, reviews and cases are over-represented in the relevant document set, while original research articles are under-represented. Gobeill et al.

Table 7 Distribution of article types in both the collection and the articles judged as relevant

Article type	Distribution	
	Collection (%)	Relevant (%)
Research-article	74.3	52.2
Case-report	4.0	20.4
Review-article	6.9	17.9
Other	2.6	3.2
Brief-report	1.1	1.5

Numbers taken from Gobeill et al. (2014)

(2014) gave a 20 % score boost to review articles and case reports, though it isn't clear whether this approach was a helpful way to leverage this article type insight.

Wing and Yang (2014) utilized the insight that query term weights could be altered by looking at the other queries. That is, terms that are common in the 30 topics are less likely to be important and should have their weight reduced. This approach is analogous to using a large EHR corpus to adjust weights (e.g., so frequent symptoms are weighed less than rarer symptoms), so it could be considered an acceptable strategy for this task. The exact approach taken by Wing and Yang (2014), however, did not appear to have a discernible effect on the results.

5 Results

Table 8 shows the scores for the best run for the top 10 systems (by infNDCG). As is common in such evaluations, the top-ranked system depends on the metric used. The top-ranked by infNDCG (Choi and Choi 2014) is the 4th ranked by P@10 and infAP and the 9th ranked by R-prec. The best performing run by P@10 (Mourão et al. 2014) has the 2nd best infNDCG, while the best performing run by infAP and R-prec (Xu et al. 2014) has the 3rd best infNDCG. It should be noted that infNDCG is the only metric that takes the relevance score into account (i.e., definitely relevant vs. possibly relevant). The top 3 participants by infNDCG all used word representations and all utilized PRF for query expansion. None performed concept extraction or utilized any medical resource for query expansion. The only medical resource used by any of the top 3 was that Choi and Choi (2014), who trained a machine learning model on the Clinical Hedges Database to help rank articles more appropriate for clinical review for Test and Treatment topics (runs that used a classifier for Diagnosis topics performed worse). At the other end of the performance spectrum, quite a few systems performed poorly. We attribute this to the fact that it was the first year of this track, so participants had no means of tuning their methods, or of even knowing what a competitive score would be.

Figure 3 shows a box-and-whisker plot of the infNDCG scores across all 30 topics. The figure is organized by topic type, with the left 10 being Diagnosis topics, the middle 10

Table 8 Top 10 ranked results (by infNDCG)

Participant	Run	infNDCG	P@10	infAP	R-prec
Choi and Choi (2014)	5	0.2674	0.3633	0.0659	0.1733
Mourão et al. (2014)	4	0.2631	0.3900	0.0757	0.2165
Xu et al. (2014)	2	0.2587	0.3700	0.0812	0.2193
Sankhavara et al. (2014)	3	0.2562	0.3733	0.0728	0.1990
Li et al. (2014)	3	0.2367	0.2967	0.0563	0.1891
Soldaini et al. (2014)	2	0.2272	0.3367	0.0623	0.1871
Gobeill et al. (2014)	2	0.2150	0.3333	0.0612	0.1874
Goodwin and Harabagiu (2014)	1	0.2047	0.3067	0.0561	0.1699
Wei et al. (2014)	5	0.1996	0.2933	0.0514	0.1626
Oh and Jung (2014)	4	0.1946	0.2933	0.0492	0.1666

The Run number corresponds to the rows in Tables 5 and 6

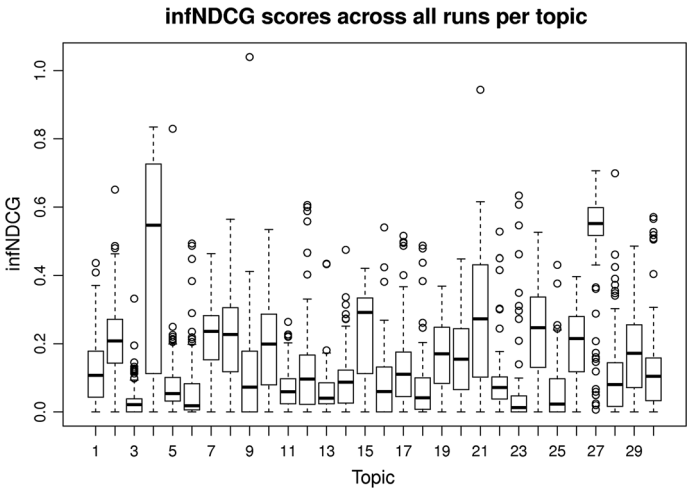

Fig. 3 *Box*-and-*whiskers* plot of the infNDCG scores across the 30 topics. Topics 1–10 are Diagnosis, 11–20 are Test, and 21–30 are Treatment

being Test topics, and the right 10 being Treatment topics. Some topics are clearly quite difficult (e.g., Topics 3, 6, 11, 25), while others are easier (e.g., Topics 4, 27). The Diagnosis and Treatment topics are generally easier (average median score of 0.17283 and 0.17066, respectively) than Test topics (0.11074). However, the Diagnosis and Treatment topics also are more variable (standard deviation of median scores of 0.15450 and 0.15325, respectively) than Test topics (0.07460). Having only 10 topics each is not likely a sufficient sample size to make broad generalizations, so it is difficult to say whether there is something inherent in the notion of a Test topic that is fundamentally more difficult (in terms of understanding, representing, or finding relevant articles). We can speculate, however, that more biomedical articles address the diagnosis and treatment of patients, while fewer are dedicated to the efficacy of the tests themselves. Further, as Fig. 1 shows, Test topics can lead to either a test to make a diagnosis (such as a lab test like hemoglobin A1c) or a test to help determine the best form of treatment (such as a genetic test for cancer therapy). Perhaps reducing the ambiguity of Test topics by splitting them into Diagnosis-Test and Treatment-Test may lead to more relevant information.

5.1 Significance evaluation

Table 9 provides a significance comparison of the runs using pair-wise, two-sided *t* tests. Given only 30 topics in the set and the typical wide variation in effectiveness across topics for different systems (Banks et al. 1999), few run pairs are statistically different from one another. So, for instance, the top-ranked system on topic 1 might be the 14th ranked on topic 2, the 8th ranked on topic 3, etc. The mean infNDCG score of the top run is almost 40 % greater than the mean infNDCG score of the 12th ranked system, and yet the spread in per-topic scores renders the two runs statistically indistinguishable. This not unexpected, and is a issue with many IR evaluations, limiting one's ability to claim one participant's system performs definitively better than another on the task (especially with a small number of topics).

Table 9 Significance (paired *t* test) amongst the systems

Participant		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Choi and Choi (2014)	1	★																									
Mourão et al. (2014)	2	●	★																								
Xu et al. (2014)	3	●	●	★																							
Sankhavara et al. (2014)	4	●	●	●	★																						
Li et al. (2014)	5	●	●	●	●	★																					
Soldaini et al. (2014)	6	○	○	●	●	●	★																				
Gobeill et al. (2014)	7	○	○	○	○	●	●	★																			
Goodwin and Harabagiu (2014)	8	○	○	○	○	●	●	●	★																		
Wei et al. (2014)	9	○	○	○	○	○	●	●	●	★																	
Oh and Jung (2014)	10	○	○	○	○	○	○	●	●	●	★																
Palotti et al. (2014)	11	○	○	○	○	○	○	●	●	●	●	★															
Wang and Fang (2014)	12	○	○	○	○	○	○	●	●	●	●	●	★														
Girmay and Deroie (2014)	13	○	○	○	○	○	○	●	●	●	●	●	●	★													
OHSU	14	○	○	○	○	○	○	○	●	●	●	●	●	●	★												
Wan et al. (2014)	15		○	○	○	○	○	○	○	●	●	●	●	●	●	★											
Garcia-Gathright et al. (2014)	16			○	○	○	○	○	○	○	●	●	●	●	●	●	★										
Dinh and Ben Abacha (2014)	17			○	○	○	○	○	○	○	○	●	○	●	●	●	●	★									
Wing and Yang (2014)	18					○	○	○	○	○	○	○	○	○	●	●	●	●	★								
Joo and Sohn (2014)	19								○	○	○	○	○	○	○	○	○	○	○	★							
Bhandari and Kulkarni (2014)	20											○	○	○	○	○	○	○	○	●	★						
Hasan et al. (2014)	21								○	○	○	○	○	○	○	○	○	○	○	●	●	★					
Merck_DA	22																			○	○	○	★				
Xue et al. (2014)	23																			○	○	○	●	★			
D'hondt et al. (2014)	24																								★		
IKMLAB	25																								●	★	
Singh and Chowdary (2014)	26																								○	○	★

Filled circle indicates runs are nearly the same ($p \geq 0.5$), open circle indicates runs not statistically significantly different ($p \geq 0.05$)

6 Discussion

6.1 Utility of medical resources

In the first instance of this task, it was found that relatively simple, standard IR methods (word representation + PRF) have worked best. We have speculated, though, that this is in large part due to the lack of data available for tuning, as future instances of the track will be able to leverage the 2014 topics and relevance judgments. Such tuning would enable many of the domain-specific features in Sect. 4 to be used more effectively, as we discuss below.

First, there is the issue of how best to use a concept extraction system such as MetaMap (whose default options are tuned for the biomedical literature, and not necessarily clinical text), or which lexicons have a sufficient coverage of the terminology to be useful. Many medical terms have synonyms with no overlapping words (e.g., *hypertension* and *high blood pressure*), so the use of a medical knowledge source to normalize these terms has the potential to retrieve more relevant documents. However, this can come at a cost as well: used incorrectly, MetaMap will generate a significant number of false positive matches (as would a large lexicon such as UMLS). It is likely that an optimal clinical IR system would find a way to normalize concepts and still utilize the words not present in the lexicon, but a priori it is difficult to know how best to combine two different representations.

Second, both negation and attribute extraction (age, gender, etc.) are of profound clinical importance, yet none of the top 4 systems utilized either (the 5th ranked system, Li et al. (2014), utilized both, however). Yet without looking at the topics, participants had to guess the best way of extracting this information (especially for attributes, as there already are well-tested algorithms for negation detection). Further, once extracted, there were three main approaches for handling negations and attributes: simply filtering them out, or normalizing them with either a hard- or soft-matching strategy. Again, a soft-matching strategy seems like it would have the most promise, but the optimal modeling of soft-matching likely requires empirical data.

Finally, the article preference methods should play an important role within a clinical IR system. Informal feedback from the assessors indicated that a significant number of the irrelevant articles were simply not the kind of article that would ever be useful in a clinical setting, or that it had a fundamental mismatch with the clinical important attributes. One of the assessors noted that:

> *I'd say that in around a third of the cases, so far, one wouldn't even need to be a physician to tell there's no possibility that the paper will help choose a test for the given clinical scenario. Not too rarely there were papers on veterinary medicine, very old papers on medical history, papers reporting public health programs in remote locations, and often the papers were about Pediatrics (while the patient in the clinical vignette was the mother, not the child).*

The article preference strategies for trying to identify appropriate articles for Diagnosis, Treatment, and Test topics have the potential to implicitly recognize unworthy articles. As discussed in Sect. 4.10, participants experimented with specialized lexicons, MeSH terms, and machine learning classifiers. None of these approaches can be said to have been definitively successful, though the top-performing system of Choi and Choi (2014) did use a method in all of their runs. Further, Gobeill et al. (2014) tried to utilize the article type descriptor to find articles more likely to prove useful in a clinical setting, though again without success. The availability of 37k judgments as training data, however, should prove useful in future tracks. Additionally, our analysis of MeSH terms in the relevant documents

(see below) should provide some insights to future participants. One of the disadvantages of using PMC, though, is that as an open subset of Medline it lacks many of the meta-analyses and systematic reviews that require paid access.

One of the unique advantages of using PMC as a corpus is the availability of MeSH terms for those PMC articles that are indexed in Medline. MeSH (an abbreviation for Medical Subject Headings) terms provide an authoritative categorization of a biomedical article by a trained expert into an extensive medical ontology. They provide insights on the organism of study (e.g., Humans, Mice), biographical information (Child, Young Adult, Aged, Pregnancy), the type of study (Retrospective Studies, Cross-Sectional Studies, Questionnaires), measurements used (Body Mass Index, Magnetic Resonance Imaging), and much more. MeSH terms are publicly available through a variety of NLM services, such as the Entrez Programming Utilities API (NCBI 2010). Two main issues need to be dealt with in order to effectively utilize MeSH for PMC articles. The first issue is the fact that a large number of PMC articles are not indexed in Medline, and therefore do not have MeSH terms. Of the 733,328 articles in the dataset, only 422,835 (57.7 %) have MeSH terms. Of the 274,339 articles returned by any of the participant systems for any of the topics, only 153,320 (55.9 %) have MeSH terms. Further, of the 3272 articles judged to be relevant for one of the topics, only 1508 (46.1 %) have MeSH terms. This means that any solution that involves MeSH will either have to handle non-indexed articles separately, or apply an automatic technique for indexing the articles (Aronson et al. 2004; Wilbur and Kim 2014). Table 10 demonstrates the advantage of properly leveraging MeSH terms. On the left side of the table are the 50 most common MeSH terms, with their corresponding frequencies, in PMC. On the right are the 50 most common MeSH terms in the relevant document set. While many of the terms overlap, quite a few terms related to cell biology and sequencing (e.g., Molecular Sequence Data; Cell Line; Cells, Cultured; Amino Acid Sequence) are not in the top relevant terms, while non-human MeSH terms (Animals, Mice, Rats) are far less common. This is understandable as these types of publications focus on basic science and have little direct value in the clinical setting. Note that Animals and Mice are still in 177 and 47 relevant documents, respectively. This should motivate re-ranking strategies using MeSH, as opposed to hard filtering, as there are few absolute rules in relevance assessment.

6.2 Limitations and challenges

There are several differences between the track as constituted and how a clinical IR system would have to operate in its intended environment. First, the use of PMC (as opposed to all of Medline) precludes some of the most useful articles from being included in the collection, notably meta-analyses and systematic reviews. A clinical setting such as a hospital would likely have paid access to many of the relevant journals and review services that do not allow their content in PMC. However, PMC is not completely devoid of such articles, they are just likely to be under-represented in an open access collection. This limitation had to be weighed against the ability of teams from around the world, largely at non-medical centers, to participate in the track.

A second limitation is the differences between a topic (in this task) and a EHR note (in a clinical environment). As stated in Sect. 3.2, the topics were inspired by actual medical records in MIMIC-II, though there are some notable differences in style. First, topic creators used fewer abbreviations than are typically found in clinical notes. This is a well-studied NLP problem (Wu et al. 2012), and it was felt that heavy use of abbreviations would burden IR researchers unfamiliar with the task, and unfairly benefit those that were

Table 10 Top 50 MeSH terms in PMC (left) and articles assessed as relevant or possibly relevant (right)

PMC	Relevant
Humans (264,767)	**Humans** (1454)
Animals (152,800)	**Female** (899)
Female (141,798)	**Male** (779)
Male (129,799)	**Middle aged** (627)
Adult (75,098)	**Adult** (522)
Middle Aged (71,229)	**Aged** (478)
Mice (59,900)	**Treatment outcome** (216)
Aged (50,221)	**Adolescent** (195)
Adolescent (29,196)	**Risk factors** (188)
Molecular sequence data (24,250)	**Aged, 80 and over** (186)
Young adult (24,063)	Animals (177)
Rats (22,856)	**Child** (127)
Time factors (21,028)	**Prospective studies** (121)
Child (20,115)	**Young adult** (120)
Aged, 80 and over (18,501)	**Retrospective studies** (118)
Risk factors (17,705)	**Time factors** (112)
Cell line (17,498)	**Child, preschool** (106)
Treatment outcome (16,816)	**Diagnosis, differential** (91)
Cells, cultured (16,586)	**Infant** (90)
Amino acid sequence (15,506)	**Pregnancy** (75)
Base sequence (15,342)	**Sensitivity and specificity** (72)
Cell Line, Tumor (14,413)	**Prognosis** (70)
Child, preschool (13,113)	**Follow-up studies** (66)
Mice, inbred C57BL (13,074)	**Severity of illness index** (62)
Phylogeny (10,751)	**Predictive value of tests** (61)
Reproducibility of results (10,654)	**Prevalence** (59)
Retrospective studies (10,644)	**Tomography, X-ray computed** (58)
Questionnaires (10,272)	**Cohort studies** (54)
Infant (10,242)	**Incidence** (48)
Mutation (10,127)	Case–control studies (47)
Case–control studies (10,092)	Mice (47)
Pregnancy (10,073)	Cross-sectional studies (46)
Protein binding (9584)	**Acute disease** (45)
Genotype (9582)	**Age Factors** (42)
Models, biological (9547)	**Logistic models** (41)
Phenotype (9452)	Questionnaires (39)
Prospective studies (9357)	**Risk assessment** (38)
Cross-sectional studies (9313)	Infant, newborn (38)
Cohort studies (9119)	**Disease progression** (38)
Prognosis (8822)	**Magnetic resonance imaging** (38)
Algorithms (8808)	**Drug therapy, combination** (37)
Immunohistochemistry (8585)	Reproducibility of results (34)
Reverse transcriptase polym… (8324)	**Comorbidity** (31)

Table 10 continued

PMC	Relevant
Follow-up studies (8156)	**Drug combinations** (29)
Disease models, animal (8108)	**Body mass index** (27)
Prevalence (7998)	**Double-blind method** (26)
Sequence analysis, DNA (7755)	**Colonoscopy** (25)
Mice, knockout (7579)	**Biological markers/blood** (25)
Gene expression profiling (7491)	**Randomized controlled trial**… (24)
Kinetics (7379)	Sex factors (24)
Indexed articles: 422,835	Indexed articles: 1507

Article count for the MeSH term is in parentheses. Terms in bold indicate they are significantly more likely to appear in a relevant article than in a random PMC article ($p < 0.001$). Underlined terms indicate they are significantly less likely to appear ($p < 0.001$)

familiar. Second, clinical notes tend to be more telegraphic in style, but the exact manner in which this is done varies from institution to institution. It was thus decided that the topics should be written in as clear a manner as possible. Finally, sometimes there is far more information in a clinical note than what is presented in a topic, oftentimes extraneous information from what is necessary to answer the clinical question. This too varies from institution to institution, though in MIMIC-II it is often not an issue.

Perhaps the most important question to be asked about the track is, qualitatively, how well the systems performed. That is, *is the state-of-the-art where it needs to be to support CDS using the biomedical literature?* The answer to this question is most likely *no*, or at least *not yet*. While there may be some users in a clinical setting who are willing to read through numerous articles, including fundamentally irrelevant articles, to find the answers they are looking for, the (quantitative) scores and the (qualitative) feedback from our assessors indicate that progress is still necessary. (While the assessors did not have access to the ranked results by any particular system, the pooling process should be fairly representative of the top 20–40 results from the systems.) As has been stated numerous times throughout this article, however, this is but the first instance of this track. Next, we discuss the future of the track.

6.3 Future work

The second instance of this track, the TREC 2015 Clinical Decision Support Track, has been accepted as a TREC track and will commence in mid-2015. It is our intention to largely keep the track un-changed to allow participants to leverage the topics and relevance judgments from the 2014 track. Participants will again be provided with 30 topics, from the same three topic types, and be asked to return up to 1,000 ranked articles from the same snapshot of PMC. One additional planned feature of the 2015 track is the release of additional meta-data for the topics, such as the specific diagnosis the topic creator has in mind when writing a treatment or test topic. This meta-data will be released in a second phase, such that the first phase is almost identical to the 2014 track but with different topics, and the second (optional) phase provides additional data systems can use to improve their results. The exact details behind what meta-data will be released will be announced in early 2015.

7 Conclusion

This article has described TREC 2014 Clinical Decision Support Track and the state-of-the-art techniques utilized by the participants. The goal of the track was to develop IR techniques to connect clinicians with the biomedical literature to improve evidence-based decision making. The participants employed a wide variety of techniques, leveraging the standard methods in IR as well as medical resources. The track was very successful in terms of participation, though it is clear further developments are necessary to enable IR systems to be effective in this challenging yet important task.

Acknowledgments Kirk Roberts, Matthew Simpson, and Dina Demner-Fushman were supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. The authors would also like to thank the following participants for providing feedback and clarifications: Raymond Wan, Paul McNamee, Jean Garcia-Gathright, Joao Palotti, Eva D'hondt, Dawit Girmay, Afshin Deroie, Sungbin Choi, Luca Soldaini, Joe McCarthy, and Yi-Shu Wei.

References

Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, *20*(4), 357–389.

An, X., & Cercone, N. (2014). How complementary are different information retrieval techniques? a study in biomedicine domain. In *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing)*, pp. 367–380.

Arighi, C. N., Roberts, P. M., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-aryamontri, A., et al. (2011). BioCreative III interactive task: An overview. *BMC Bioinformatics*, *12*(Suppl 8), S4.

Aronson, A., & Lang, F. M. (2010). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*, 229–236.

Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., & Rogers, W. J. (2004). The NLM indexing initiative's medical text indexer. In *Studies in Health Technology and Informatics (MEDINFO)*, pp. 268–272.

Banks, D., Over, P., & Zhang, N. F. (1999). Blind men and elephants: Six approaches to TREC data. *Information Retrieval*, *1*, 7–34.

Bhandari, A., Kulkarni, A. (2014). San Francisco State University at TREC 2014: Clinical Decision Support Track. In *Proceedings of the 2014 Text Retrieval Conference*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chapman, W. W., & Cohen, K. B. (2009). Current issues in biomedical text mining and natural language processing. *Journal of Biomedical Informatics*, *42*(5), 757–759.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310.

Choi, S., & Choi, J. (2014). SNUMedinfo at TREC CDSS track 2014: Medical case-based retrieval task. In *Proceedings of the 2014 Text Retrieval Conference*.

Clinchant, S., & Gaussier, E. (2010). Information-based Models for Ad Hoc IR. In *Proceedings of the 33rd Annual ACM International Conference on Research and Development in Information Retrieval*.

Del Fiol, G., Workman, T. E., & Gorman, P. N. (2014). Clinical questions raised by clinicians at the point of care: A systematic review. *JAMA Internal Medicine*, *174*(5), 710–718. pMID24663331.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, *42*(5), 760–772.

D'hondt, E., Grau, B., Darmoni, S., Névéol, A., Schuers, M., & Zweigenbaum, P. (2014). LIMSI @ TREC clinical decision support track. In *Proceedings of the 2014 Text Retrieval Conference*.

Dinh, D., & Ben Abacha, A. (2014). CRP Henri Tudor at TREC 2014: Combining Search Results for Clinical Decision Support. In *Proceedings of the 2014 Text Retrieval Conference*.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Garcia-Gathright, J., Meng, F., & Hsu, W. (2014). UCLA at TREC 2014 Clinical Decision Support Track: Exploring Language Models, Query Expansion, and Boosting. In *Proceedings of the 2014 Text Retrieval Conference*.

Girmay, D., & Deroie, A. (2014). Query expansion using SNOMED-CT and weighing schemes. In *Proceedings of the 2014 Text Retrieval Conference*.

Gobeill, J., Gaudinat, A., Pasche, E., & Ruch, P. (2014). Full-texts representation with Medical Subjects Headings, and co-citations network reranking strategies for TREC 2014 Clinical Decision Support Track. In *Proceedings of the 2014 Text Retrieval Conference*

Goeuriot, L., Jones, G. J., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., & Zuccon, G. (2013). ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In *CLEF 2013 Working Notes*.

Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G. J., & Müller, H. (2014). ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In *CLEF 2014 Working Notes*, pp. 43–61

Goodwin, T., & Harabagiu, S. (2014). UTD at TREC 2014: Query expansion for clinical decision support. In *Proceedings of the 2014 Text Retrieval Conference*.

Hasan, S. A., Zhu, X., Dong, Y., Liu, J., & Farri, O. (2014). A hybrid approach to clinical question answering. In *Proceedings of the 2014 Text Retrieval Conference*.

Hatcher, E., & Gospodnetic, O. (2004). *Lucene in action*. Greenwich: Manning Publications.

Haynes, R. B., McKibbon, K. A., Wilczynski, N. L., Walter, S. D., & Werre, S. R., Hedges Team (2005). Optimal search strategies for retrieving scientifically strong studies of treatment from medline: Analytical survey. *BMJ*, *330*, 1179–1185.

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

Hersh, W., & Kim, E. (2006). The impact of relevance judgments and data fusion on results of image retrieval test collections. In *Proceedings of the Second MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, pp. 29–38

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994a). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual ACM International Conference on Research and Development in Information Retrieval*, pp. 192–201

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

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

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

Hersh, W., Cohen, A., Ruslen, L., & Roberts, P. (2007). TREC 2007 genomics track overview. In *Proceedings of the Sixteenth Text Retrieval Conference*.

Hersh, W. R., & Hickam, D. H. (1995). An evaluation of interactive boolean and natural language searching with an online medical textbook. *Journal of the American Society for Information Science*, *46*(7), 478–489.

Hersh, W. R., Hickam, D. H., Haynes, R. B., & McKibbon, K. A. (1994b). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of American Biomedical Informatics*, *1*(1), 51–60.

Hu, Q., Huang, J. X., & Miao, J. (2011). A robust approach to optimizing multi-source information for enhancing genomics retrieval performance. *BMC Bioinformatics*, *12*(Suppl 5), S6.

Jelinek, F., & Mercer, R. (1980). Interpolated estimation of markovsource parameters from sparse data. *Pattern Recognition in Practice* pp. 381–402.

Joo, H., & Sohn, K. (2014). TREC2014 clinical decision support: Concept-based clinical information retrieval using MetaMap. In *Proceedings of the 2014 Text Retrieval Conference*.

Kalpathy-Cramer, J., de Herrera, A. G. S., Demner-Fushman, D., Antani, S., Bedrick, S., & Müller, 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*.

Kim, J. D., & Cohen, K. B. (2013). Natural language query processing for SPARQL generation—A prototype system for SNOMEDCT. In *Proceedings of BioLINK*, pp. 32–36.

Li, M., Song, Y., He, Y., Hu, Q., He, L., & Haacke, E. M. (2014). ECNU at TREC 2014: Clinical decision support track. In *Proceedings of the 2014 Text Retrieval Conference*.

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The unified medical language system. *Methods of Information in Medicine*, *32*(4), 281–291.

Liu, S., Ma, W., Moore, R., Ganesan, V., & Nelson, S. (2005). RxNorm: Prescription for electronic drug information exchange. *IT Professional*, *7*(5), 17–23.

Lu, Z. (2011). *PubMed and beyond: A survey of web tools for searching biomedical literature*. Database 2011

McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., et al. (2003). LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry*, *49*(4), 624–633.

McNamee, P., Mayfield, J., & Piatko, C. (2002). HAIRCUT: A system for multilingual text retrieval in Java. *Journal of Computing Sciences in Colleges*, *17*(3), 8–22.

Mourão, A., Martins, F., & Magalhães, J. (2014). NovaSearch at TREC 2014 clinical decision support track. In *Proceedings of the 2014 Text Retrieval Conference*.

NCBI (2010). *Entrez programming utilities help*. National Center for Biotechnology Information

Oh, H. S., & Jung, Y. (2014). KISTI at TREC 2014 clinical decision support track: Concept-based document re-ranking to biomedical document retrieval. In *Proceedings of the 2014 Text Retrieval Conference*.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: A high performance and scalable information retrieval platform. In *SIGIR Open Source Workshop*.

Palotti, J., Rekabsaz, N., Anderson, L., & Hanbury, A. (2014). TUW @ TREC clinical decision support track. In *Proceedings of the 2014 Text Retrieval Conference*.

Ryu, B., & Choi, J. (2013). Biomedical test collection with multiple query representation. In *Proceedings of the Fifth International Workshop on Evaluating Information Access*, pp. 33–36.

Sankhavara, J., Thakrar, F., Sarkar, S., & Majumder, P. (2014). Fusing manual and machine feedback in biomedical domain. In *Proceedings of the 2014 Text Retrieval Conference*.

Scott, D., Lee, J., Silva, I., Park, S., Moody, G., Celi, L., & Mark, R. (2013). Accessing the public MIMIC-II intensive care relational database for clinical research. *BMC Medical Informatics and Decision Making*, *13*(9). http://www.biomedcentral.com/1472-6947/13/9.

Singh, A., & Chowdary, C. R. (2014). Centrality based document ranking. In *Proceedings of the 2014 Text Retrieval Conference*.

Soldaini, L., Cohan, A., Yates, A., Goharian, N., & Frieder, O. (2014). Query reformulation for clinical decision support search. In *Proceedings of the 2014 Text Retrieval Conference*.

Spärck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing and Management*, *36*(6), 779–840.

Stearns, M. Q., Price, C., Spackman, K. A., & Yang, A. Y. (2001). SNOMED clinical terms: Overview of the development process and project status. In *Proceedings of the AMIA Annual Symposium*, pp. 662–666.

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005). Indri: A language-model based search engine for complex queries. In *International Conference on Intelligence Analysis*

Voorhees, E. M. (2014). The effect of sampling strategy on inferred measures. In *Proceedings of the 37th Annual ACM International Conference on Research and Development in Information Retrieval*, pp. 1119–1122.

Voorhees, E. M., & Hersh, W. (2012). Overview of the TREC 2012 medical records track. In *Proceedings of the 11th Text REtrieval Conference*.

Voorhees, E. M., & Tong, R. M. (2011). Overview of the TREC 2011 medical records track. In *Proceedings of the 10th Text REtrieval Conference*.

Wan, R., Man, J. H. K., & Chan, T. F. (2014). Query modification through external sources to support clinical decisions. In *Proceedings of the 2014 Text Retrieval Conference*.

Wang, X., Chused, A., Elhadad, N., Friedman, C., & Markatou, M. (2008). Automated knowledge acquisition from clinical narrative reports. In *Proceedings of the AMIA Annual Symposium*, pp. 783–787.

Wang, X., Thompson, P., Tsujii, J., & Ananiadou, S. (2012). Biomedical Chinese-English CLIR using an extended CMeSH resource to expand queries. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 1148–1155.

Wang, Y., & Fang, H. (2014). Explore the query expansion methods for concept based representation. In *Proceedings of the 2014 Text Retrieval Conference*.

Wei, Y., Hsu, C., Thomas, A., & McCarthy, J. F. (2014). Atigeo at TREC 2014 clinical decision support task. In *Proceedings of the 2014 Text Retrieval Conference*.

Weinreich, S. S., Mangon, R., Sikkens, J., Teeuw, M., & Cornel, M. (2008). OrphaNet: A european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, *152*(9), 518–519.

Wilbur, W. J., & Kim, W. (2014). Stochastic gradient descent and the prediction of MeSH for PubMed records. In*Proceedings of the AMIA Annual Symposium*, pp. 1198–1207.

Wing, C., & Yang, H. (2014). Query refinement: Negation detection and proximity learning: Georgetown at TREC 2014 clinical decision support track. In *Proceedings of the 2014 Text Retrieval Conference*.

Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A., & Xu, H. (2012). A comparative study on current clinical natural language processing systems on handling abbreviations in discharge summaries. In *Proceedings of the AMIA Annual Symposium*, pp. 997–1003.

Xu, T., McNamee, P., & Oard, D. W. (2014). HLTCOE at TREC 2014: Microblog and clinical decision support. In *Proceedings of the 2014 Text Retrieval Conference*.

Xue, T., Fu, Q., Gu, H., Zhang, S., & Wang, C. (2014). Clinical decision support track of 2014. In *Proceedings of the 2014 Text Retrieval Conference*.

Yilmaz, E., Kanoulas, E., & Aslam, J. A. (2008). A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual ACM International Conference on Research and Development in Information Retrieval*, pp. 603–610.

Zhang, X., Cole, M., & Belkin, N. (2011). Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th Annual ACM International Conference on Research and Development in Information Retrieval*, pp. 1225–1226.