Exploiting Structural Patterns for Health Search

Research Track - Project Proposal

Alberto Ueda
CS Dept, UFMG

ABSTRACT

Health related topics have become a common theme within IR. A number of premiere IR publication venues, including SIGIR, have dedicated workshops, tutorials or tracks dedicated to health search. Similarly, other venues such as WSDM, WWW, KDD, and ACL have all hosted health related tutorials or workshops. This shows the rising interest from the research community in contributing to health search, an area with arguably significant social impact. In this project, we propose... The vast literature available for precision medicine can make it difficult to find the most appropriate treatment for the clinician's current patient. The ability to quickly locate relevant information for a current patient using IR can be an important tool for helping clinicians find the most up-to-date evidence-based treatment for their patients.

ACM Reference format:
Alberto Ueda. 2017. Exploiting Structural Patterns for Health Search. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 2 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 THE PROBLEM

Similar to the 2014 track, the focus of the 2015 Clinical Decision Support Track will be the retrieval of biomedical articles relevant for answering generic clinical questions about medical records. [5] A case report typically describes a challenging medical case, and it is often organized as a well-formed narrative summarizing the portions of a patient's medical record that are pertinent to the case. The case narratives describe information such as a 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 finally, the steps taken by a physician to treat the patient.

Participants of the track will be challenged with retrieving for a given case report full-text biomedical articles that answer questions related to several types of clinical information needs. Retrieved articles will be judged relevant if they provide information of the specified type that is pertinent to the given case.

The target document collection for the track is the Open Access Subset of PubMed Central (PMC), an online digital database of freely available full-text biomedical literature. a snapshot of the open access subset on January 21, 2014, which contained a total of 733,138 articles. The topics are annotated according to three common generic clinical question types: *diagnosis, test,* or *treatment.* For example, for a case report labeled "diagnosis" participants should retrieve PMC articles a physician would find useful for determining the diagnosis of the patient described in the report. Similarly, for a case report labeled "treatment," participants should retrieve articles that suggest to a physician the best treatment plan for the condition exhibited by the patient described in the report. Finally, for "test" case reports participants should retrieve articles that suggest relevant interventions that a physician might undertake in diagnosing the patient.

The table below shows examples of the kind of case-based topics we will be using for the track. The document identifiers listed in the last column are relevant for the given cases because they can assist a physician in determining the patient's diagnosis or treatment.

Table 1: Examples of input and expected output for the problem. TREC 2015 Clinical Decision Support Track.

No.	Summary	Relevant Articles
1	Type: Diagnosis. Description: A 26-year-old obese woman with a history of bipolar disorder complains that her recent struggles with her weight and eating have caused her to feel depressed. She states that she has recently had difficulty sleeping and feels excessively anxious and agitated. She also states that she has had thoughts of suicide. She often finds herself fidgety and unable to sit still for extended periods of time. Her family tells her that she is increasingly irritable. Her current medications include lithium carbonate and zolpidem.	1087494 1434505 2031887
2	Type: Treatment. Description: A 21-year-old female is evaluated for progressive arthralgias and malaise. On examination she is found to have alopecia, a rash mainly distributed on the bridge of her nose and her cheeks, a delicate non-palpable purpura on her calves, and swelling and tenderness of her wrists and ankles. Her lab shows normocytic anemia, thrombocytopenia, a 4/4 positive ANA and anti-dsDNA. Her urine is positive for protein and RBC casts.	1065341 1459118 1526641

2 LITERATURE

In previous editions of TREC (2014–2016) [4], similar health search tasks were proposed. Several techniques from IR and NLP were employed by the participant teams, including disease-centered document clustering and semantic word vectors using word embeddings [3], query expansion by identification of clinical intent types, negation-aware ranking models, and synonyms [1], and supervised learning-to-rank approaches based on document similarity [2].

3 OUR APPROACH

In some contexts, finding highly related documents is a desirable feature for IR systems. For instance, given a relevant medicine article with a treatment for a patient's disease, one can be interested in finding as many documents similar to this article as possible.

For such tasks, we can apply document clustering techniques. In particular, using disease-centered document clusters improve clinical document retrieval effectiveness [3]. Our proposal is to still consider the disease but on a treatment-centered network. Given a patient with a specific disease and a reference treatment, we

Conference'17, Washington, DC, USA
2017. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

aim to provide articles with similar treatments or clinical trials. To determine the reference treatment, we can use learning to rank methods based on clinical ground-truths.

4 EVALUATION PROCEDURE

We aim to use both the data and the evaluation procedures provided by previous tracks of the TREC.[1]. The highest ranked articles for each topic will be pooled and judged by medical librarians and physicians from the Department of Medical Informatics of the Oregon Health and Science University. Assessors will be instructed to judge articles as either "definitely relevant" for answering questions of the specified type about the given case report, "definitely not relevant," or "potentially relevant." (the article is not immediately informative on its own, but it may be relevant in the context of a broader literature review). Because we plan to use a graded relevance scale, the performance of the retrieval submissions will be measured using normalized discounted cumulative gain (NDCG). As the previous results of several contestants are also available, we will be able to compare the effectiveness of our approach against the top tier competitors.

[1] http://trec-cds.org/2015.html

REFERENCES

[1] Simon Greuter, Philip Junker, Lorenz Kuhn, Felix Mance, Virgile Mermet, Angela Rellstab, and Carsten Eickhoff. 2016. ETH Zurich at TREC 2016 Clinical Decision Support. In *NIST Special Publication 500-321: The Twenty-Fifth Text REtrieval Conference Proceedings (TREC 2016)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology, S.l.

[2] Harsha Gurulingappa, Luca Toldo, Claudia Schepers, Alexander Bauer, and Gerard Megaro. 2016. Semi-Supervised Information Retrieval System for Clinical Decision Support. In *TREC*.

[3] Seung-Hyeon Jo and Kyung-Soon Lee. 2016. CBNU at TREC 2016 Clinical Decision Support Track. In *TREC*.

[4] Kirk Roberts, Matthew S. Simpson, Ellen M. Voorhees, and William R. Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*. http://trec.nist.gov/pubs/trec24/papers/Overview-CL.pdf

[5] Guido Zuccon and Bevan Koopman. 2017. SIGIR 2017 Tutorial on Health Search (HS2017): A Full-day from Consumers to Clinicians. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 1415–1418. DOI: https://doi.org/10.1145/3077136.3082061