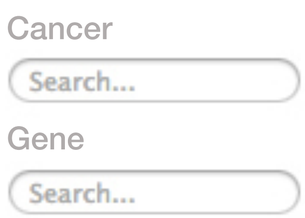
A Clinical Trial Search Engine for Precision Medicine

Course Project for SI 650
Adharsh Murali, Andrew Barber, Jinghui Liu

Introduction

- To personalize treatment and prevention strategies in a scientific rigorous manner has been regarded as the central theme of ***precision medicine (PM)***.
- Given the wide range of treatment and prevention options that need to be considered by clinicians is usually overwhelming, ***how to search and process related information*** and leverage them in clinical practice has become an important issue in promoting and implementing precision medicine.
- Because of the variety and specificity of the relation between cancer and gene, ***genetic mutations of cancer*** has become a specific use case in the precision medicine paradigm.
- We aim to build a search engine that focus on this problem: retrieving useful information to treat cancer patients with specific gene mutations.



- Different from the original search function of *ClinialTrials.gov*, it focuses on the specific gene to help clinicians to implement precision health care.

Data

- Data for clinical trials is a corpus consists of **241,006** past, present, and planning clinical trials derived from *ClinicalTrials.gov*.

NIH ClinicalTrials.gov

- We evaluate our system using **30 patient cases** provided by TREC (Text REtrieval Conference) 2017 Precision Medicine Track, containing a wide range of condition and gene variations. For example:

```
<topic number="2">
    <disease>melanoma</disease>
    <gene>BRAF (V600K)</gene>
   <demographic>54-year-old male</demographic>
   <other>Type II Diabetes</other>
</topic>
```

- Topics are created by precision oncologists at *the University of Texas MD Anderson Cancer Center* and *the Oregon Health & Science University (OHSU) Knight Cancer Institute*. Ground truth for evaluating these topics are provided as well.

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

OHSU KNIGHT CANCER Institute

Methods

- Free-text indexed using common tokenizers and filters to remove stop-words, do stemming and so on.
- Used BM25F as the similarity algorithm.
- Two approaches focused on query modification:

  1) ***Query expansion***, and
  2) ***Query term selection***

- Query expansion uses NCBI (National Center for Biotechnology Information) database and UMLS to find synonyms for gene and disease.

NCBI

- Term selection modifies the component of query based on how many docs the system can retrieve using the query.

Unified Medical Language System

Term Selection	Ex. <disease>thyroid cancer</disease> <gene>BRAF (V600R)</gene>
All terms	thyroid + cancer + BRAF + V600R + other terms
Relaxed variant	thyroid + cancer + BRAF + other terms
Relaxed disease	thyroid + BRAF + V600R + other terms
… …	thyroid cancer BRAF V600R

- Multiple sets of docs are retrieved and merged using these different queries.
- Results post-processed by removing cases not satisfying demographic constraints, including age and gender.

Evaluation

- We used the 30 topics to evaluate our performance. Metrics used for evaluation are Precision@5, Precision@10, and Precision@15, chosen by TREC.

On 30 patient cases	P@5	P@10	P@15
Our Performance	0.36	0.29	0.27
Median	0.28	0.24	0.20
Best run of 2017	0.54	0.44	0.38

Future Work

- Explore re-ranking using learning-to-rank methods;
- Expand the coverage of ontological resource;
- Create richer filter to remove irrelevant documents;
- Implement topic modeling using the ground truth provided for the 30 topics.

(adharsh, acbarber, ljinghui)@umich.edu