Title: A Clinical Trial Search Engine for Precision Medicine

Abstract:

In the pursuit of value-based care, the “precision medicine” paradigm has become increasingly prevalent in the modern clinical discourse. Personalized evidence-based treatment strategies are considered the pinnacle of clinical care. This is especially true in the area of oncology, where properly tailored treatment can make a lifesaving difference. However, much of the evidence used to develop these strategies is still not reasonably accessible for many practitioners. It is therefore the goal of our team to create a search engine used by clinicians to examine clinical trials based on genetically unique varieties of cancer. We created a search engine for 241,006 past, present and future clinical trial documents derived from clinicaltrials.gov. Our model was evaluated using 30 patient cases established by the Text REtrieval Conference (TREC) 2017 Precision Medicine (PM) Track. To optimize the performance of our model, we introduced query expansion and iterative retrieval techniques leveraging the National Center for Biotechnology Information (NCBI) and Unified Medical Language System (UMLS) databases. Our search engine outperformed the median precision@5, 10 and 15 for the established TREC 2017 PM Track.

CSS Concepts: Information Retrieval, Query Expansion, Iterative Retrieval

Keywords: Cancer, Precision Medicine, Search Engine

Introduction:

There is no “one size fits all” solution in treating patients diagnosed with complex diseases. Especially in treating cancer a lifesaving treatment for a person could prove ineffective and also can be deadly for some other patient. Precision medicine can result in a better patient outcome than using the same strategy for everyone since it assigns the treatment and prevention strategy based on individual characteristics.[1]The clinical research in the field of oncology has shown the implication of genetic variants in cancer. For example, the people with Gene variants of large effect in BRCA2 and CHEK2 have an increased risk of being affected by Lung Cancer.[2] The Gene Variant helps in identifying the risk of cancer hence it is essential in identifying the most effective treatment and preventive measures for cancer patients. However, with increased production of knowledge in the field of Precision Medicine and Oncology, the clinicians are overwhelmed with information which in turn can inhibit them from choosing the appropriate treatment for the patients. Since most of the information is buried in the scientific literature it becomes difficult for clinicians in finding the most relevant article. Information Retrieval engine provides an effective and efficient way to retrieve relevant and updated information from a very large corpus in minimal time which could facilitate the clinicians in making a well-informed decision.

In this paper, we explain the retrieval engine we built to retrieve most relevant clinical trial for which the patient is eligible based on his/her gene variant information. The different

sections in the paper explain the different methods that were used in building the retrieval engine.

Data:

Our index was created using a total of 241,006 documents curated from www.clinicaltrials.gov. The documents were available in text format and also in semi-structured XML format. The semi-structured format was used in building the retrieval engine. The data provided information on the title of the literature, a detailed description of the findings in the literature,Inclusion and the Exclusion Criteria in the study and also a brief summary of the article. In order to evaluate the performance of the retrieval engine, the Text Retrieval Engine Conference Precision Medicine track using the brains of the precision oncologist created 30 queries based on synthetic cases at the University of Texas MD Anderson Cancer Center and the Oregon Health & Science University (OHSU) Knight Cancer Institute. Each query represents a cancer patient, which includes four additional fields 1) Patient disease (i.e.) Type of cancer 2)The Genetic Variant information of the patient 3) The demographic information about the user which includes the age, gender 4)Other Potential factors that are relevant to the disease.

```
topics task="2017 TREC Precision Medicine">
<topic number="1">
 <disease>Acute lymphoblastic leukemia</disease>
 <gene>ABL1, PTPN11</gene>
 <demographic>12-year-old male</demographic>
 <other>No relevant factors</other>
</topic>
...
/topics>
```

Figure 1: (Much better quality in Word) A sample topic for the TREC 2017 Precision Medicine track. Reprinted from *Trec Precision Medicine/Clinical Decision Support Track*, 2017, Retrieved from http://www.trec-cds.org/2017.html.

Methods:

We built the retrieval system in a single step approach. Initially, the documents were indexed using a customized analyzer based on the Regular expression tokenizer that filters the stop words and special characters the analyzer also stemmed the words to have a solid inverted index without multiple pointers to the same words. The query term was expanded by expanding the disease and gene information making use of the well established medical ontologies 1) National Center for Biotechnology Information (NCBI) 2) Unified Medical Language System (UMLS). The improved results were retrieved using the BM25F similarity algorithm. The retrieved results were scored based on their relevance and were presented in the sorted order.

The results were post-processed to remove the results of the document that do not match the demographic constraints.

Creating the Index:

We used whoosh in python to index the clinical trial document collection. The documents were indexed using a customized tokenizer based on regular expression tokenizer, which also filtered the stop words, stemmed the words to avoid superlatives and different forms of the same words in the index. For every document, the title of the trial, detailed description about the trial, the minimum and maximum age of the patients in the trial were only indexed.

Query Expansion:

The Query expansion was one of the major steps in developing the retrieval engine. Synonyms based on gene name and disease type were appended to the query in order to retrieve more relevant documents that contain different description of certain topics. It is known that query expansion is useful particularly in the searching of literature in specific domain. The two sources for this expansion task, NCBI and UMLS, are national knowledge base that sets standard of the research language in the field of medicine and biomedical research. It is believed that these two database would be sufficient to provide information related closely to the precision medicine paradigm. To access them, gene database can be downloaded directly from NCBI while UMLS provides an API for individual researchers.

Query Modification:

With the expansion terms, each topic is parsed into a set of queries that are different from each other. Each query is modified based on its specificity in terms of how many meaningful token it contains. For example, a specific query would include all the disease terms and gene terms, with their respective synonyms. In comparison, a less specific query would include fewer expansion terms or even get rid of gene variation. By having this set of different queries, it is believed that the system could have better performance in retrieving most relevant documents for common topics, while being capable to handle rare cases given the tiered modified queries.

Post Processing:

Documents are retrieved for each topic using the set of queries that are developed with the two strategies described above. Among these queries, the strictest one that contains most query terms is used first, and documents retrieved using it are treated as the top-ranked documents. Then a less strict query will be used to do the same. This creates an iterative pattern to retrieve a number of sets of documents for each topic, and these sets are ranked based on the specificity of the queries used to retrieve them. Finally, all these results will be appended together after dropping duplicates, and they are considered as the output for a topic.

Results or Evaluation:

The retrieval engine was evaluated using the 30 queries on 30 different cancer patients released by the TREC 2017 PM task B. We evaluate the system using the measure of precision at 5, precision at 10 and precision at 15.

Patient cases	Precision@5	Precision@ 10	Precision@ 15
Patient Cases	0.36	0.29	0.27
Median	0.28	0.24	0.20
Best Run	0.54	0.44	0.38

Table 1: Performance of the retrieval engine in comparison to the Median performances of the retrieval engine in the challenge and the best-run retrieval engine.

From the analysis of the result, the built system was able to beat the median precision of the retrieval engine participated in the challenge. The improved performance of the system can be attributed to the inverted index of the retrieval system. The index of the system was built using the entire corpus by filtering the stop words and stemming the words which in turn has contributed to the overall performance of the engine.

Discussion:
It can be seen that the strategies proposed in this project could have reasonable performance in the task of retrieving clinical trials for precision cancer cases. It is believed that such performance comes from the expansion and modification of queries. However, these two strategies are by no means perfect. For query expansion, the two ontological resources often offer too much information and it is sometimes hard to select what synonyms should be appropriate. For example, NCBI would sometimes offer a synonym of certain gene that actually does not belong to human. For query modification, the problem lays in whether the approach of using the number of retrieved documents to evaluate the specificity of queries, and consequently decide their ranking. The future work of the project will try to evaluate these issues and explore building more accurate queries that can represent patient cases well.

References:

1) Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H.S., Diver, W.R., Thun, M.J., Cox, D.G., Hankinson, S.E., Kraft, P., et al.: Performance of common genetic variants in breast-cancer risk models. New England Journal of Medicine 362(11), 986–993 (2010)

2) Wang, Y., McKay, J.D., Rafnar, T., Wang, Z., Timofeeva, M.N., Broderick, P., Zong, X., Laplana, M., Wei, Y., Han, Y., et al.: Rare variants of large effect inbrca2 and chek2 affect risk of lung cancer. Nature Genetics 46(7), 736–741 (2014)