A Clinical Trial Search Engine for Precision Medicine

Adharsh Murali
School of Information
University of Michigan
Ann Arbor MI USA
adharsh@umich.edu

Jinghui Liu
School of Information
University of Michigan
Ann Arbor MI USA
ljinghui@email.com

Andrew Barber
School of Information
University of Michigan
Ann Arbor MI USA
acbarber@umich.edu

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 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.

CCS CONCEPTS

• Information Retrieval • Query Expansion • Iterative Retrieval

KEYWORDS

Precision Medicine, Cancer, Search Engine

ACM Reference format:

Andrew Barber, Jinghui Liu and Adharsh Murali. 2018. A Clinical Trial Search Engine for Precision Medicine. University of Michigan School of Information SI 650, *Ann Arbor, MI, USA, 3 pages.*

1 Introduction

There is no "one size fits all" solution when treating patients diagnosed with complex diseases. This is especially true when treating cancer - a lifesaving treatment for one person could prove ineffective or even deadly for another. Precision medicine can result in better patient outcomes when compared to using the same strategy for everyone; it assigns treatment and prevention strategies based on individual characteristics rather than relying on cookie-cutter treatment plans.[1] Clinical research in the field of oncology has shown the implication of individualized care when contending with genetic variants of cancer. For example, patients with significant genetic variation in BRCA2 and CHEK2 genotypes have an increased risk of being affected by Lung Cancer.[2] Identifying genetic variants helps in predicting a patient's risk of cancer and is therefore 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, clinicians are overwhelmed with information. This in turn can inhibit them from choosing the appropriate treatment for their patients. Since most of the information is buried in the scientific literature it becomes difficult for clinicians to find the most relevant information. Information retrieval engines provide an effective and efficient way to retrieve relevant and updated information from a very large corpus in minimal time. This can facilitate clinicians in making well-informed decisions.

In this paper, we explain how our retrieval engine can be used to retrieve the most relevant clinical trials based on specific patient cases – we will allow clinicians to search for clinical trial documents based on disease-type as well as the genetic variation of the disease. We will explain how this tool was built, optimized and deployed to return the most pertinent information to our users.

2 Data

Our index was created using a total of 241,006 documents curated from www.clinicaltrials.gov. The documents are available in text format as well as semi-structured XML format. The semi-structured format was used in building our retrieval engine. The data provided information regarding the title of the clinical trial, a detailed description of the trial's findings, inclusion and exclusion criteria for the study and a brief summary of the trial. In order to evaluate the performance of the retrieval engine, the TREC precision medicine Track used a panel of precision oncologists to design 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 patient which includes the age and 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: 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.

3 Methods

We built our retrieval system using a single-step approach. Initially, the documents were indexed using a customized analyzer based on a regular expression tokenizer that filters stop words and special characters. The analyzer also stemmed the words contained in the clinical trial documents, creating a solid inverted index without multiple pointers to the same words. The query terms were expanded by appending synonyms for disease and gene variant information, making use of the well-established medical ontologies: National Center for Biotechnology Information (NCBI) and 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 any clinical trials that did not match the demographic constraints specified by the user.

3.1 Creating the Index

We used the Whoosh Python library to index the clinical trial document collection. The documents were indexed using a customized tokenizer based on regular expressions, which also filtered the stop words and stemmed the words to avoid superlatives and different forms of the same words in the index. For every document, the brief and official title of the trial, brief summary, detailed description and the minimum and maximum age of the patients in the trial were the only query-able fields. Both brief and official titles are merged into a TITLE field for title query, and a brief summary of the trial is indexed together with description in another field named CONTENT. A multi-field query parser is then used to process query that retrieves from both fields, with TITLE field having a higher weight than the CONTENT field.

3.2 Query Expansion

Query expansion was one of the major steps in optimizing the performance of our retrieval engine. Synonyms based on gene name and disease type were appended to the query in order to retrieve more relevant documents containing different descriptions of certain topics. It is known that query expansion is useful particularly when searching for literature in specific domains. The two sources for this expansion task, NCBI and UMLS, are national knowledge-bases which set the standards for research language in the field of medicine and biomedical research. It is believed that these two databases would be sufficient to provide information related closely to the precision medicine paradigm. The NCBI gene database can be downloaded directly from NCBI while UMLS provides an API for individual researchers.[3,4]

3.3 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 tokens 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 exclude terms from the topic itself, such as gene variation. This is because that sometimes a relevant trial is implemented on a wide range of genes and are not explicit about the variation information, which would lead to some blind spots for retrieval. By having this set of different queries, it is believed that the system could have better performance in retrieving the most relevant documents for common topics, while being capable of handling rare cases given the tiered modified queries.

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

Table 1: Query modification based on specificity and selection of terms. The example is a patient case that has thyroid cancer and a mutation on the V600R locus on BRAF gene.

3.4 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 containing the 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 document sets for each topic. Such iteration would continue to retrieve documents until either all of the queries are used for the retrieval or there are in total more than one hundred unique documents retrieved. These sets are subsequently ranked based on the specificity of the queries that are used to retrieve them. All these results will be appended together after dropping duplicates. At the final step, demographic information including gender and age in the index is used to filter out trials that exclude the current patient case from their enrolled population. The search engine will then return the ranking of the rest of the documents as its final output for the task.

4 Evaluation

The retrieval engine was evaluated using the 30 queries on 30 different cancer patients released by the TREC 2017 PM Task B. We evaluated the system using the measure of precision at 5, precision at 10 and precision at 15 (See *Table 1 below*). These metrics are chosen by TREC to evaluate the effectiveness of a system as the real-world application for oncology practitioners, which is expected to return to the top those results that are as relevant as possible. From the analysis of the result, our system was able to beat the median precision of all retrieval engines that participated in the challenge. The performance under the three metrics are consistently higher than the median score, though they have a clear disadvantage compared with the best run in that year. With the score of 0.29 for precision at 10, the system could be considered usable for clinical scenarios to provide relevant information to clinicians for their further evaluation.

	Precision@ 5	Precision@ 10	Precision@ 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

Table 2: Performance of our search engine compared to the median and maximum performance of retrieval system submissions to TREC 2017 Precision Medicine Track.

5 Discussion

It can be seen that the strategies proposed in this project could produce a 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. The expansion strategy makes use of two classic medical ontologies that are considered as standard in life research. The rich information helps the search engine to better recognize the potential information need from the clinician users. The modification strategy focuses on how to rank higher documents that are most relevant, at the same time adopts an iteration of retrieval based on the query modification that ensure the system to find enough documents for extremely rare cases. 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 which synonyms should be appropriate. For example, NCBI would sometimes offer a synonym of a certain gene that actually does not exist in the human genome. For query modification, the problem lies in whether to evaluate the specificity of the queries using the number of retrieved documents, which consequentially decides their ranking. Future work for this project will involve evaluating these issues and exploring the development of more relevant queries that can more accurately represent specific patient cases.

6 Conclusion

This report introduces a search engine that focuses on retrieving relevant clinical trial document for clinician to treat specific patient cases. With domain experts as target users, the performance of the engine is also evaluated using the ground truth provided by clinical practitioners. Specifically, 30 patient cases provided by TREC 2017 Precision Medicine Track are used for the evaluation and the system achieve 0.36 on the main metric precision at 5. It is believed that the search engine has potential to contribute to the real-world patient care and treatment of rare cancer diseases in the future development of precision medicine.

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)

[3] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.

[4] Betsy L. Humphreys, Donald A. B. Lindberg, Harold M. Schoolman, G. Octo Barnett; The Unified Medical Language System: An Informatics Research Collaboration, Journal of the American Medical Informatics Association, Volume 5, Issue 1, 1 January 1998, Pages 1–11