Brown University at TREC Precision Medicine 2019

Abdullah Ahmed, Gil Alon, Bashar Zaidat, Hwai-Liang Tung, Isaac Nathoo, Charles Wang, Carsten Eickhoff
Brown University, USA
carsten@brown.edu

Abstract

This paper describes Brown University's submission to the TREC 2019 Precision Medicine (PM) track. We expand disease and gene name related terms, prune expanded queries and boost the importance of key terms. Our retrieval model is based on BM25F and incorporates heuristic relevance eligibility filters for clinical trials as well as reciprocal rank fusion of constituent runs.

1 Introduction

Precision medicine is a modern field of study that aims to use genomic information in finding more effective treatments for patients. Due to the popularity of the new paradigm, the volume of annually published scholarly precision medicine articles has been growing rapidly in recent years. While this considerable amount of scientific research holds a rich and ever increasing well of knowledge, its sheer scale makes it intractable for manual inspection and mandates the development of dedicated automatic retrieval facilities.

In this paper we present a patient-centric information retrieval system which may be applied in precision medicine. Based on information such as a patient's demographics, known diseases, and genetic configuration involved in the disease, we rank clinical trials according to their relevance to reference patients.

2 Methodology

To build our model we used query expansion, retrieval model fusion, and patient eligibility filtering.

2.1 Indexing

The clinical trials were indexed using Whoosh, a pure Python search engine library [1]. Following suit of Team Cat-Garfield's submission to TREC 2018 [2], we indexed the NCT ID, Title, Brief Title, Brief Summary, Detailed Description, Study Type, Intervention Type, Inclusion Criteria, Exclusion Criteria, Healthy Volunteers, Keywords, Gender, and MeSH Terms as dedicated fields.

Our index for scientific abstracts was corrupted just before the submission deadline, preventing us from submitting official runs to that task. We will report unofficial runs after we resolve the issue.

28th Text Retrieval Conference (TREC) Notebook 2019.

2.2 Query Expansion

The goal of query expansion was to increase recall of our search by extracting synonyms of both the disease and gene. Expanding the query consisted of three steps: 1) finding synonyms for both diseases and genes, 2) reducing the expanded query, and 3) boosting the importance of the original disease term.

Finding Disease and Gene Synonyms. We used Lexigram [3] to extract synonyms of the disease. Lexigram combines medical terminologies of SNOMED CT, MeSH and ICD into a single database [4]. Additionally, since many articles use descriptions to discuss the cancer, if the cancer was not a type of blood cancer, namely aortic aneurysms or leukemia, we appended the term "solid" to the query in addition to the synonyms found in Lexigram. This strategy, used by Agosti [5], was found to be effective in clinical trials where the researchers would use the term "solid tumor" to describe the cancer. Likewise, to expand the query based on gene variants, we used the NCBI Gene database [6] to find synonyms for all genes. We appended all disease synonyms and gene variants found in the database without reducing or manipulating them. There were usually 10-20 synonyms and variants appended to each query.

Query Reduction. To limit the query length, we removed all synonyms from Lexigram that received a relevance score of less than 1.3. Lexigram calculates their relevance score using a "decay function tuned to term relevance and term proximity of the input keywords" [4]. We selected a threshold of 1.3 via manual examination; synonyms scoring less than 1.3 were rarely relevant or connected to the disease, and therefore were removed. Additionally, for the gene field, before searching the NCBI Gene database, if in the TREC field the gene had a specific variant attached to it, such as "BRAF (E586K)", we would remove that specific variant and expand only "BRAF" as the Gene database does not have synonyms for specific variants.

Boosting the Importance of Original Terms. Lastly, to improve accuracy and to ensure that our model was prioritizing the most relevant clinical trials, we boosted the score of each document that contained the original disease term provided in the TREC database. The score of each document was doubled if it contained the original query term.

2.3 Retrieval

We used Whoosh to retrieve clinical trials. Whoosh uses the Okapi BM25F model [7] to retrieve and rank documents. Okapi BM25F is an extension of BM25 that views each document as a set of "streams", or fields. Under this framework, the fields indexed in our clinical trials can be weighted differently according to their importance. We weighted MeSH terms 50% higher as they represent the most important topics of each trial.

2.4 Eligibility Filtering

To retrieve relevant clinical trials after ranking with a BM25F algorithm, the results were passed through several eligibility filters. These post-processing filters included checking the clinical trial for specified age ranges, gender, diseases, and genes to determine the eligibility match between the queried topic and the clinical trials. If the trial was found to be one that the patient can enroll in, it was included in the ranked result list. Conversely, if the trial was not one that matched the topic, it was excluded for re-ranking. For each result in the ranked list, we filtered out documents whose age and gender fields did not include or match the age and gender fields of the query, respectively. Next, if the abstract of the clinical trial includes the strings: "no", "not", "without", or "n't" within 25 characters of the disease name and the disease name was not one of the MeSH terms, the trial was deemed ineligible. Lastly, if the abstract mentioned a gene or mutation associated with the disease, it was checked to see if the topic gene was a match. If the resulting list contained fewer than 1000 trials, the list was expanded using a built-in Whoosh pseudo relevance feedback function based on the first 10 retrieved trials. Whoosh's expansion function extracts keywords from the specified field in the document (in our case, MeSH terms), which are then used as a query to retrieve more documents.

2.5 Re-Ranking

As the final step in our ranking scheme, we performed reciprocal rank fusion (RRF) [8] as implemented by the Polyfuse library[1]. The runs were fused according to the parameters of query expansion (more on this in the next section).

3 Results

We submitted five official TREC PM 2019 runs for evaluation, all of which were for clinical trial retrieval. Table 1 highlights our performance on the following runs:

- **ndng** did not expand the gene or disease.
- **egnd** expanded the gene but not the disease.
- **eged** expanded both the gene and the disease.
- **rrf_1** fused ngnd and egnd together using Polyfuse.
- **rrf_2** fused rrf_1 and eged together using Polyfuse.

Table 1: Results for clinical trial retrieval task. Bold is our best run.

Run Name	infNDCG	P@10	R-Prec
ngnd	0.2314	0.2158	0.1291
egnd	**0.3216**	**0.3211**	**0.2056**
eged	0.1906	0.1632	0.1094
rrf_1	0.2635	0.2211	0.1489
rrf_2	0.2166	0.1763	0.1295

We can see that expanding the gene of a query yielded the best results. Expanding the disease decreased our model's performance, which is indicative of a problem in our use of the Lexigram API rather than our overall methodology. The poor results from expanding the disease decreased our performance on rrf_2 as well.

4 Conclusion

In this paper, we provide an overview of Brown University's contribution to the TREC 2019 Precision Medicine Track. Our method uses API-backed query expansion, retrieval model fusion, and eligibility filtering to complete the task. Our results show significantly increased performance when the gene field of a query is expanded. Next steps include constructing a more stable query expansion tool and using attention mechanisms (*e.g.*, BERT) to augment retrieval and better re-rank documents.

References

[1] Matt Chaput. Whoosh. `https://whoosh.readthedocs.io/en/latest/index.html`.

[2] X. Zhou et al. Team cat-garfield at trec 2018 precision medicine track. *Text REtrieval Conference (TREC) Proceedings*, 27, 2018.

[3] Lexigram. `https://www.lexigram.io/`.

[4] Lexigram api documentation. `https://docs.lexigram.io/v1/lexigraph/getConcept`.

[5] M. Agosti et al. An analysis of query reformulation techniques for precision medicine. *ACM SIGIR*, 42:973–976, 2018.

[1] https://github.com/rmit-ir/polyfuse

[6] NA O'Leary et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(14):6614–24, 2016.

[7] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 2009.

[8] G. Cormack et al. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *ACM SIGIR*, 32:758–759, 2009.