Ranking Clinical Trials Using Elasticsearch

Ajinkya Yogesh Thorve[1] and Haimonti Dutta[2]
[1]Department of Computer Science and Engineering
[2]Department of Management Science and Systems
University at Buffalo, NY 14260
{ajinkyay, haimonti}@buffalo.edu

Abstract

The clinical trials task of the TREC 2017 Precision Medicine Track was designed to represent the potential for connecting patients with experimental treatments if existing treatments were ineffective. Participants were challenged with the task of retrieving appropriate clinical trials from ClinicalTrials.gov for which a patient is eligible. This paper presents an approach to solving the problem by first preparing an index for the clinical trial descriptions based on specific tags in the XML files and querying them using Elasticsearch. Initial results indicate that our approach performed very well for certain kinds of queries – however, more tuning may be required for ensuring generalizable results from the search.

Introduction

The aim of the TREC 2017 Precision Medicine Track [1] is to provide useful information to physicians for treating cancer patients. The goal for participants is to investigate information retrieval techniques that will retrieve the most relevant documents given a case query. There were two document collections that were made available: scientific abstracts (consisting of PubMed abstracts) and clinical trials (consisting of clinical trial descriptions from ClinicalTrials.gov). With PubMed abstracts, the task was to retrieve relevant treatments for the given patients from the given collection. With the clinical trials dataset, the task was to retrieve relevant clinical trials for which the patient is eligible. Considering the computing resources available to us and time constraints, we focused solely on the second task – finding relevant clinical trials.

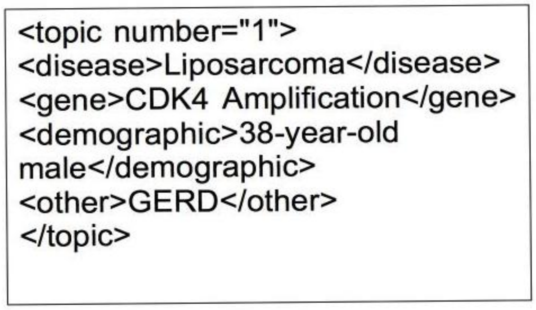
Data

Clinical Trial Descriptions

The clinical trials collection consisted of about 241,006 XML documents . Each of these documents consisted of numerous fields, such as the title, summary, description, eligibility criteria, and some pre-assigned keywords and MeSH terms. Figure 1 shows a sample clinical trial with the brief title and summary of the description.

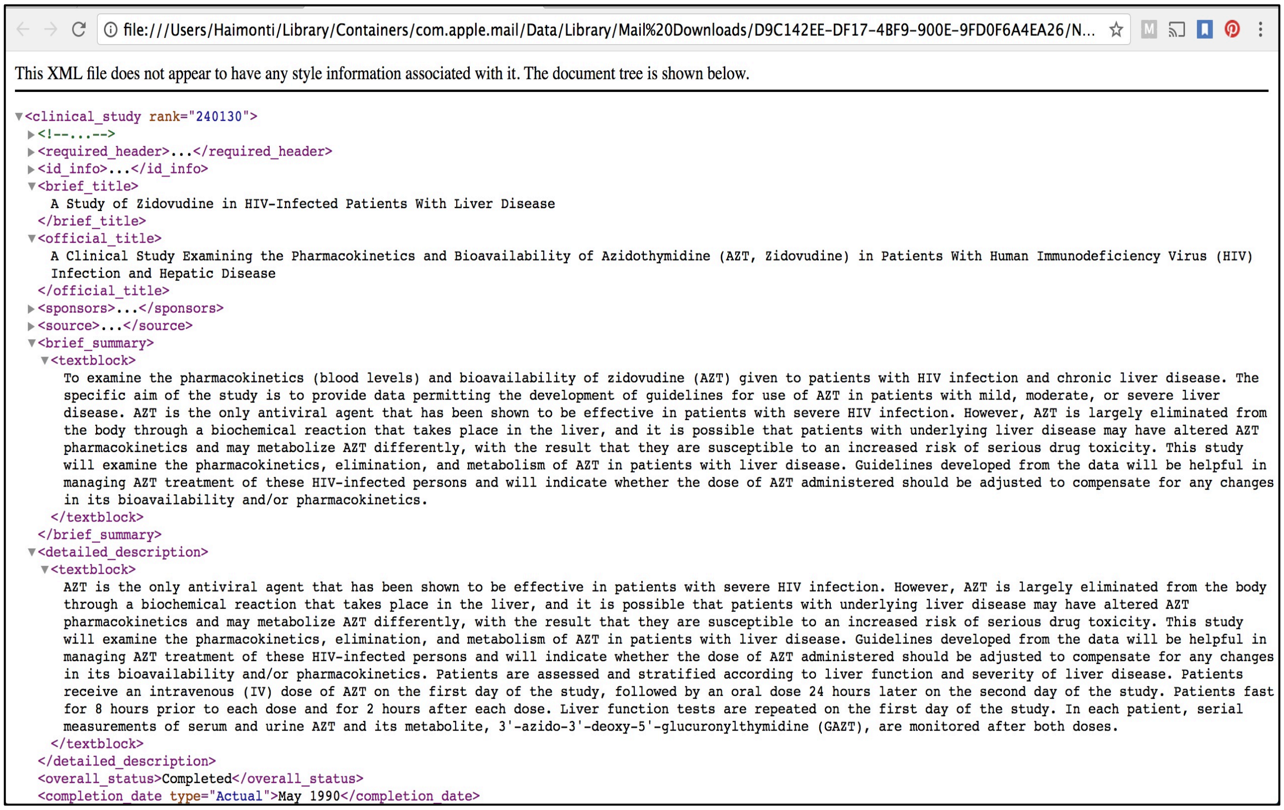
Topics

The topics for the track comprise of synthetic cases created by precision oncologists at the University of Texas, MD Anderson Cancer Center. Each case describes the patient's disease (type of cancer), the relevant genetic variants (which genes), basic demographic information (age, sex), and other potential factors that may be relevant. 30 cases were made available for the retrieval task. Figure 2 shows an illustrative example of a topic. Furthermore, a separate file consisting of 16 extra topics was also provided by TREC. It consisted of some additional query topics accompanied by clinical trial IDs that were partially relevant to the task.



```
<topic number="1">
<disease>Liposarcoma</disease>
<gene>CDK4 Amplification</gene>
<demographic>38-year-old
male</demographic>
<other>GERD</other>
</topic>
```

Figure 2: An example topic.



file:///Users/Haimonti/Library/Containers/com.apple.mail/Data/Library/Mail%20Downloads/D9C142EE-DF17-4BF9-900E-9FD0F6A4EA26/N...

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<clinical_study rank="240130">
  ▶<!--...-->
  ▶<required_header>...</required_header>
  ▶<id_info>...</id_info>
  ▼<brief_title>
     A Study of Zidovudine in HIV-Infected Patients With Liver Disease
   </brief_title>
  ▼<official_title>
     A Clinical Study Examining the Pharmacokinetics and Bioavailability of Azidothymidine (AZT, Zidovudine) in Patients With Human Immunodeficiency Virus (HIV)
     Infection and Hepatic Disease
   </official_title>
  ▶<sponsors>...</sponsors>
  ▶<source>...</source>
  ▼<brief_summary>
    ▼<textblock>
       To examine the pharmacokinetics (blood levels) and bioavailability of zidovudine (AZT) given to patients with HIV infection and chronic liver disease. The
       specific aim of the study is to provide data permitting the development of guidelines for use of AZT in patients with mild, moderate, or severe liver
       disease. AZT is the only antiviral agent that has been shown to be effective in patients with severe HIV infection. However, AZT is largely eliminated from
       the body through a biochemical reaction that takes place in the liver, and it is possible that patients with underlying liver disease may have altered AZT
       pharmacokinetics and may metabolize AZT differently, with the result that they are susceptible to an increased risk of serious drug toxicity. This study
       will examine the pharmacokinetics, elimination, and metabolism of AZT in patients with liver disease. Guidelines developed from the data will be helpful in
       managing AZT treatment of these HIV-infected persons and will indicate whether the dose of AZT administered should be adjusted to compensate for any changes
       in its bioavailability and/or pharmacokinetics.
     </textblock>
   </brief_summary>
  ▼<detailed_description>
    ▼<textblock>
       AZT is the only antiviral agent that has been shown to be effective in patients with severe HIV infection. However, AZT is largely eliminated from the body
       through a biochemical reaction that takes place in the liver, and it is possible that patients with underlying liver disease may have altered AZT
       pharmacokinetics and may metabolize AZT differently, with the result that they are susceptible to an increased risk of serious drug toxicity. This study
       will examine the pharmacokinetics, elimination, and metabolism of AZT in patients with liver disease. Guidelines developed from the data will be helpful in
       managing AZT treatment of these HIV-infected persons and will indicate whether the dose of AZT administered should be adjusted to compensate for any changes
       in its bioavailability and/or pharmacokinetics. Patients are assessed and stratified according to liver function and severity of liver disease. Patients
       receive an intravenous (IV) dose of AZT on the first day of the study, followed by an oral dose 24 hours later on the second day of the study. Patients fast
       for 8 hours prior to each dose and for 2 hours after each dose. Liver function tests are repeated on the first day of the study. In each patient, serial
       measurements of serum and urine AZT and its metabolite, 3'-azido-3'-deoxy-5'-glucuronylthymidine (GAZT), are monitored after both doses.
     </textblock>
   </detailed_description>
   <overall_status>Completed</overall_status>
   <completion_date type="Actual">May 1990</completion_date>
```

Figure 1: A sample clinical trial description from ClinicalTrials.gov illustrating a subset of the tags used for indexing.

Methodology

Our approach makes use of the Elasticsearch [2] search engine to index the available documents in the corpus. The following fields were extracted from each clinical trial description: nct_id, brief_title, brief_summary, detailed_description, overall_status, condition, eligibility, gender, gender_based, minimum_age, maximum_age, keyword, and mesh_term. The entire text in all the fields was used for the purpose of indexing. Following the indexing, queries were run.

Before assigning scores to documents, they are shortlisted by applying a boolean test, so only the documents that match the query are considered. Then scores are assigned based on a BM25 algorithm[3] which is the default similarity algorithm used by Elasticsearch. This algorithm ranks a set of documents based on the query terms appearing in each document. Given a query Q containing keywords $q_1,...,q_n$, the BM25 score of a document is:

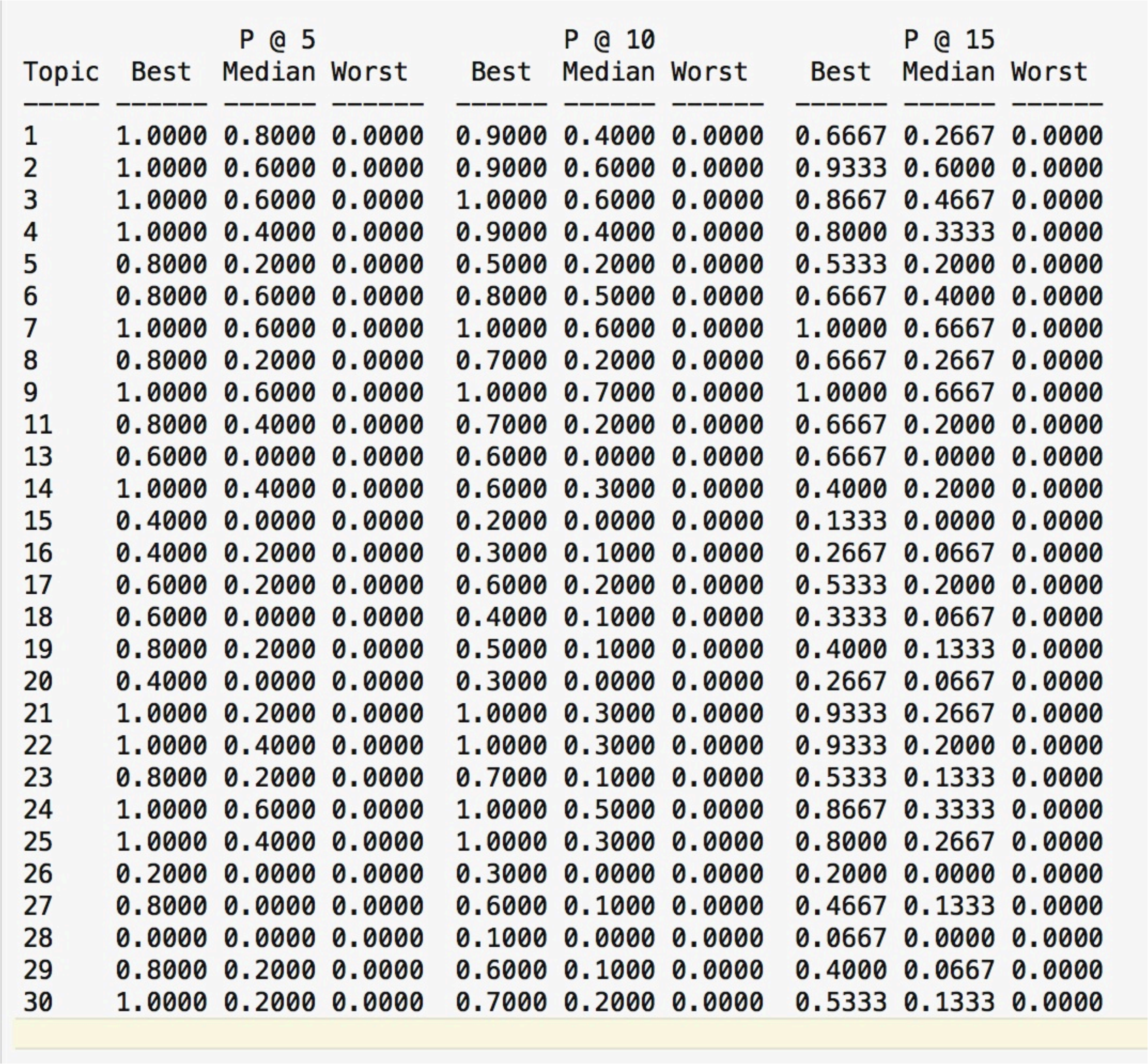
$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,\mathrm{D}).(k_1+1)}{f(q_i,\mathrm{D}) + k_1(1-\mathrm{b}+\mathrm{b}\frac{|D|}{avgdl})}$$

where $f(q_i,\mathrm{D})$ is qi's term frequency in document D, |D| is the length of the document, and avgdl is the average document length in the corpus. k_1 and b are the parameters of the algorithm – (1) k_1 controls non-linear term frequency normalization (saturation); the default value is 1.2 and (2) b controls to what degree document length normalizes tf (term frequency) values. The default value is 0.75. These default values of k1 and b are used in the retrieval process.

For the first run, the disease field in each query was extracted and used to query the corpus. For each query, we limited the results to 1000 each, and these results were ordered in decreasing order based on the relevancy score provided by Elasticsearch. These results were then written to an output file following the standard trec_eval format. For the second run, we used a more complex query. The query was formulated by an AND operation on all the xml tags namely disease, gene, demographic, and other information. Since there was no ground truth available for evaluation, the extra topics provided for the retrieval task was used tuning the Elasticsearch queries for the second run. For example, the query “Colon cancer” is accompanied by NCT IDs NCT02912559 and NCT00898846. While trying out different query configurations, we preferred the ones which gave higher ranking to those two particular NCT IDs. Similarly, other queries too were used in finding out the optimum Elasticsearch query configuration for our second run. After deciding on what seemed to be a reasonable configuration, we followed the similar approach as for our first run for querying Elasticsearch and writing the results to an output file.

Empirical Results.

Table 3 presents the performance for each of the 30 topics used for the retrieval task. Our results indicate that certain queries had very good performance – for example, for the second run, query 9 was known to have 62 relevant documents and our procedure was able to identify 61 of those; query 16 had 5 relevant documents of which 3 were returned by our algorithm. However, these results were not consistent and certain queries were found to not perform very well – for example, none of the 4 relevant documents were identified for query 15 and none of the 5 relevant documents were returned for query 25. Given the wide variance in the performance, we are investigating methods by which the performance on the retrieval task can be enhanced.



```
              P @ 5                       P @ 10                      P @ 15
Topic  Best   Median Worst     Best   Median Worst     Best   Median Worst
-----  ------ ------ ------    ------ ------ ------    ------ ------ ------
1      1.0000 0.8000 0.0000    0.9000 0.4000 0.0000    0.6667 0.2667 0.0000
2      1.0000 0.6000 0.0000    0.9000 0.6000 0.0000    0.9333 0.6000 0.0000
3      1.0000 0.6000 0.0000    1.0000 0.6000 0.0000    0.8667 0.4667 0.0000
4      1.0000 0.4000 0.0000    0.9000 0.4000 0.0000    0.8000 0.3333 0.0000
5      0.8000 0.2000 0.0000    0.5000 0.2000 0.0000    0.5333 0.2000 0.0000
6      0.8000 0.6000 0.0000    0.8000 0.5000 0.0000    0.6667 0.4000 0.0000
7      1.0000 0.6000 0.0000    1.0000 0.6000 0.0000    1.0000 0.6667 0.0000
8      0.8000 0.2000 0.0000    0.7000 0.2000 0.0000    0.6667 0.2667 0.0000
9      1.0000 0.6000 0.0000    1.0000 0.7000 0.0000    1.0000 0.6667 0.0000
11     0.8000 0.4000 0.0000    0.7000 0.2000 0.0000    0.6667 0.2000 0.0000
13     0.6000 0.0000 0.0000    0.6000 0.0000 0.0000    0.6667 0.0000 0.0000
14     1.0000 0.4000 0.0000    0.6000 0.3000 0.0000    0.4000 0.2000 0.0000
15     0.4000 0.0000 0.0000    0.2000 0.0000 0.0000    0.1333 0.0000 0.0000
16     0.4000 0.2000 0.0000    0.3000 0.1000 0.0000    0.2667 0.0667 0.0000
17     0.6000 0.2000 0.0000    0.6000 0.2000 0.0000    0.5333 0.2000 0.0000
18     0.6000 0.0000 0.0000    0.4000 0.1000 0.0000    0.3333 0.0667 0.0000
19     0.8000 0.2000 0.0000    0.5000 0.1000 0.0000    0.4000 0.1333 0.0000
20     0.4000 0.0000 0.0000    0.3000 0.0000 0.0000    0.2667 0.0667 0.0000
21     1.0000 0.2000 0.0000    1.0000 0.3000 0.0000    0.9333 0.2667 0.0000
22     1.0000 0.4000 0.0000    1.0000 0.3000 0.0000    0.9333 0.2000 0.0000
23     0.8000 0.2000 0.0000    0.7000 0.1000 0.0000    0.5333 0.1333 0.0000
24     1.0000 0.6000 0.0000    1.0000 0.5000 0.0000    0.8667 0.3333 0.0000
25     1.0000 0.4000 0.0000    1.0000 0.3000 0.0000    0.8000 0.2667 0.0000
26     0.2000 0.0000 0.0000    0.3000 0.0000 0.0000    0.2000 0.0000 0.0000
27     0.8000 0.0000 0.0000    0.6000 0.1000 0.0000    0.4667 0.1333 0.0000
28     0.0000 0.0000 0.0000    0.1000 0.0000 0.0000    0.0667 0.0000 0.0000
29     0.8000 0.2000 0.0000    0.6000 0.1000 0.0000    0.4000 0.0667 0.0000
30     1.0000 0.2000 0.0000    0.7000 0.2000 0.0000    0.5333 0.1333 0.0000
```

Figure 3: Performance of 30 topic queries on the clinical trials text corpus using Elasticsearch.

Discussions and Future Work

The NLM Unified Medical Language System (UMLS) can be used in our system for extracting the medical terminologies from the clinical trial descriptions. UMLS can be used to identify medical concepts in the topic text and relate them to semantic categories (i.e. disease, symptoms, findings, etc.) and alternative names (i.e. synonyms, preferred names, etc.). Previous studies have found UMLS to be very effective in Named Entity Recognition(NER) tasks on medical documents [4]. MetaMap[5] is a tool that uses knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques to identify UMLS terms in a given text. We are in the process of testing the use of MetaMap to index documents by Elasticsearch with the hope that this will reduce execution time for retrieval of documents. Furthermore, we are in the process of refining the query formulation and evaluating what does not produce good results.

Conclusion

In this paper we described our approach for the TREC 2017 Precision Medicine track. We submitted two automatic runs, which were based on Elasticsearch. In the future we plan to use significantly more computing resources and the MetaMap API as one of the steps in pre-processing to better identify the relevant medical terms in the dataset. Furthermore tuning of Elasticsearch can be possible once we have the ground truth data available.

References

1. http://www.trec-cds.org/2017.html
2. Elasticsearch http://www.elastic.co/products/elasticsearch
3. Stephen Robertson & Hugo Zaragoza (2009). "The Probabilistic Relevance Framework: BM25 and Beyond". **3** (4). Found. Trends Inf. Retr.: 333–389.
4. Park A, Hartzler AL, Huh J, McDonald DW, Pratt W. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. Journal of Medical Internet Research 2015;**17**(8)
5. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA. 2010;17(3):229-236.