Clinical Trial Search Using Lucene and UMLS

Yanqing Ji[1], Yun Tian[2], Hao Ying[3], John Tran[4]

[1] Department of Electrical and Computer Engineering, Gonzaga University, Spokane, Washington, USA
[2] Department of Computer Sci. and Electrical Eng., Eastern Washington University, Spokane WA, USA
[3] Department of Electrical and Computer Engineering, Wayne State University, Detroit, Michigan, USA
[4] Department of Psychiatry, University of California San Francisco at Fresno, Fresno, California, USA

Abstract

We approached the clinical trial search task of the 2021 TREC Clinical Trials Track as a query problem. A query (also known as a topic in 2021 TREC) is the free text description of a patient record, while the corpus is a large set of clinical trials descriptions. A commercial search engine, Lucene, was utilized for this clinical trial matching process. Namely, given a query, the system searches in the corpus and returns a subset of clinical trials with specific requirements. In this study, Unified Medical Language System (UMLS) was employed to convert the free text of both topics and clinical trials to more meaningful biomedical concepts, each of which is represented as a Concept Unique Identifier (CUI). An expansion technique based on Medical Subject Headings (MeSH) was used to expand all the condition terms for each clinical trial to their child terms.

To assess whether UMLS can improve the search accuracy, we designed two groups of tests: one group uses free text, while the other uses CUIs for both queries and clinical trial corpuses. As the inclusion/exclusion criteria represent the core aspect of each trial description, we also examined whether the exclusion criteria played an important role in the process of selecting a clinical trial. We extracted the inclusion criteria and exclusion criteria for each clinical study and saved them into two separate corpuses. For each group of tests, we searched against the corpus with inclusion criteria only and also against both corpuses. When searching in both corpuses, for each query (i.e. a patient profile), the search results were sorted using the difference of the two Lucene scores that were returned when searching against the two corpuses, respectively. For the free text group, we also tested whether the expansion technique could improve the performance.

Our experiment results demonstrated that the CUI-based search clearly outperformed the text-based search, which indicated the effectiveness of UMLS in text preprocessing and biomedical concept extraction. Our methods of using the exclusion criteria information and the MeSH-based term expansion technique were not as effective as we expected.

1. Introduction

Recruitment of patients for clinical trials has been and continues to be a challenge. About 80% trials fail to meet their initial recruitment target and timeline, and the delays can cause lost revenue of $8 million per day for drug companies (Johnson, 2015). Besides economic lose, inefficient recruitment may also have scientific and ethical consequences as insufficient sample size may lead to invalid or inconclusive results (Gul & Ali, 2010).

Traditionally, physician referrals play a critical role in recruiting participants for clinical studies. To increase recruitment rate and reduce cost, online recruitment methods (e.g., through social media, Google search engine advertisements, and other website campaigns) have been proposed (Brøgger-Mikkelsen, Ali, Zibert, Andersen, & Thomsen, 2020) (Akers & Gordon, 2018) (Jones, Lacroix, & Porcher, 2017). An alternative solution is to use the huge amount of patient data in electronic health records (Hersh, 2007). The 2021 TREC Clinical Trials Track employs a patient-to-trials paradigm to enable the evaluation of different patient matching systems. That is, a query or topic is the description of a patient and the corpus is a large set of clinical trial descriptions retrieved from clinicaltrials.gov.

Previous studies have shown that Lucene (Apache Lucene, 2021) is effective in medical records retrieval (Demner-fushman, et al., 2011) (Demner-Fushman, et al., 2012). It was selected as the key search tool in this study. In addition, we assume that UMLS (Bodenreider, 2004) is able to improve the system performance as it can extract biomedical terms from free text and determine whether a term is negated in the context of a sentence. Each term identified by UMLS has a Concept Unique Identifier (CUI). UMLS has been utilized by many biomedical applications (Amos, Anderson, Brody, Ripple, & Humphreys, 2020) (Jing, 2021). We also developed an expansion technique based on MeSH (Medical Subject Headings, 2021) which is a NLM (National Library of Medicine) controlled vocabulary thesaurus used for indexing articles for PubMed.

Clinical trial descriptions are generally very long, the original clinical trial files were preprocessed and only important pieces of information were extracted. The extracted data were put in two separate corpuses which include the inclusion criteria and exclusion criteria, respectively. For each query, the difference of the two Lucene scores searched against the two corpuses was used to sort the results.

Using Lucene as the search engine, we examined the effectiveness of UMLS and whether the exclusion criteria can make a difference when searching clinical trials for a patient profile. We submitted five runs that are summarized in Table 1 in Section 5.

We briefly describe our data preprocessing, including relevant information extraction and free-text conversion into CUIs using UMLS in Section 2. Our term expansion technique is presented in Section 3. We illustrate the matching process in Section 4 and present our experiments and results in Section 5. We wrap up with conclusions in Section 6.

2. Data Preprocessing

The clinical trials data from clinicaltrials.gov are stored in xml files, one for each clinical trial. The descriptions of each trial are divided into many sections, many (e.g., sponsors, location, etc.) of which are not relevant to patient-to-trial matching. The useful pieces of information were extracted and stored in separate files which were then converted to standard biomedical concepts using UMLS.

2.1 Relevant Information Extraction

We believe that the eligibility section in a clinical trial file represents the most important and useful information to determine whether a trial is eligible for a given patient. Figure 1 gives an example of the eligibility section of a typical clinical trial file. The textblock subsection (a HTML element) provides free-text descriptions of the eligibility criteria which normally contain both the inclusion criteria and exclusion criteria of a clinical trial. This subsection was extracted, separated and stored into two files which contain the inclusion criteria and exclusion criteria, respectively. Each file has the same name as the original clinical trial file. All the files containing the inclusion criteria form a corpus while those containing exclusion criteria form a different corpus. Please note that the inclusion and exclusion criteria are not provided in a unique format in different clinical trial files. Some files do not even contain explicit exclusion criteria. Different formats were analyzed when extracting and separating the information. The gender and age information were extracted from the gender, minimum_age and maximum_age subtions and put into corresponding variables.

```
<eligibility>
        <criteria>
             <textblock> Inclusion Criteria: - You must be 18-75 years of age and be diagnosed with
             schizophrenia or schizoaffective disorder - You must be able to visit the doctor's office thirteen
             (13) times over a twenty-six (26) week period Exclusion Criteria: - You are a woman and are
             pregnant or breastfeeding - You have an acute or unstable medical illness, such as heart, liver,
             or kidney disease, or you have a seizure disorder. (Note: If you are uncertain about a particular
             condition, please discuss it with your physician.) - You have a history of allergic reaction or
             intolerance to olanzapine or quetiapine
             </textblock>
        </criteria>
        <gender>All</gender>
        <minimum_age>18 Years</minimum_age>
        <maximum_age>75 Years</maximum_age>
        <healthy_volunteers>No</healthy_volunteers>
</eligibility>
```

Figure 1. Eligibility Section of a Sample Clinical Trial File

Each clinical trial file often contains one or more condition sections which consists of the medical conditions of an eligible patient. In addition, some files contain a study_pop section that provides free-text descriptions of the study population for a clinical trial. We believe that these data are also useful in matching a proper clinical trial for a patient. They were also extracted and added to the file that contains the inclusion criteria information.

We also extracted the gender and age information from each topic/patient profile. These pieces of information must meet the gender and age requirements extracted from the clinical trials. Our general approach is that, given a topic (i.e., text description of a patient), we first search the inclusion criteria corpus and/or the exclusion criteria corpus using Lucene to get a list of clinical trials. Then, those trials for which the gender and age information are not satisfied are excluded from the list.

2.2 Free-Text Conversion Using UMLS

The UMLS Metathesaurus integrates about 900,000 biomedical concepts from more than 60 families of biomedical vocabularies such as International Classification of Diseases (ICD), Current Procedural Terminology (CPT), etc. (Bodenreider, 2004). One application of UMLS is to extract meaningful biomedical concepts from free text. Each concept has a Concept Unique Identifier (CUI) as well as specific attributes defining its meaning.

The extracted clinical trials data were converted to CUIs via MetaMap (MetaMap - A Tool For Recognizing UMLS Concepts in Text, 2021) – a tool developed by the NLM) to discover Metathesaurus concepts referred to in text. The MetaMap's strict data model was used so as to remove similar and/or redundant concepts. In addition, the "--negex" parameter was enabled in order to get the negated concepts. The CUIs converted from the extracted data relevant to one clinical trial were saved in a separate file using the same file name as the original clinical trial file name. All these CUI files form a corpus. Each topic was also converted to CUIs which were used as a query to search the CUI corpus.

3. Term Expansion

The language and granularity of patients' medical condition terms in topics can be different from those found in clinical trials. The required condition terms in clinical trials are generally more coarse than those in topics. For example, the condition of a recent clinical study (i.e., NCT00000105) is *cancer* while a patient might have a specific type of cancer such as *lung cancer*. Therefore, we expanded the condition terms in clinical trials with their child terms in MeSH. This was implemented by first finding a condition term in the MeSH tree and then retrieving all its child terms in the subtrees. It is worth mentioning that parent terms were excluded because the addition of them might change the study scope of a clinical trial. For instance, if a clinical trial studies *lung cancer,* it does not make sense to recruit patients with *cancer* in their profiles as the patients may have other types of cancers.

Please also note that the expanded condition terms were not added to the extracted data corpus as some medical conditions may have hundreds of child terms. With a search engine like Lucene, extensive expansion of the corpus might not help improve the search results. Instead, we only used the expanded terms as a pruning technique to postprocess the returned results. That is, if none of the expanded conditions of a returned clinical trial was included in the search topic, the trial would be removed from the returned result list.

4. Patient-to-Trial Matching Process

Figure 2 shows the big picture of the CUI-based patient-to-trial matching process. The first step was to extract relevant data from the original clinical trial files (.xml). For each clinical trial, the required conditions, descriptions of study population and inclusion criteria were stored in one file while the exclusion criteria were stored in a separate file with the same file name but in a different corpus/folder. The text files in both corpuses were then matched with the UMLS metathesaurus and converted to corresponding CUI files which formed two CUI corpuses. The CUI files were indexed using a standard Lucene analyzer and became ready for topic search.

Each topic (i.e., patient profile) was also converted to CUIs which were the input to the Lucene search engine. Lucene's scoring function is based on the traditional TF(term frequency)-IDF(inverse document frequency) model with various extensions such as query boost, normalization, etc. For each topic, if only the inclusion criteria files (containing condition, study population and inclusion criteria data) were searched, the returned clinical trials were sorted using each trial's score. If the exclusion criteria files were also searched, a second score was obtained for each trial. The clinical trials were then sorted using the difference of the two scores associated with the same trial. After the initial results were retrieved, a pruning technique was used to remove those clinical trials whose gender and age requirements were not met by the patient's gender and age.

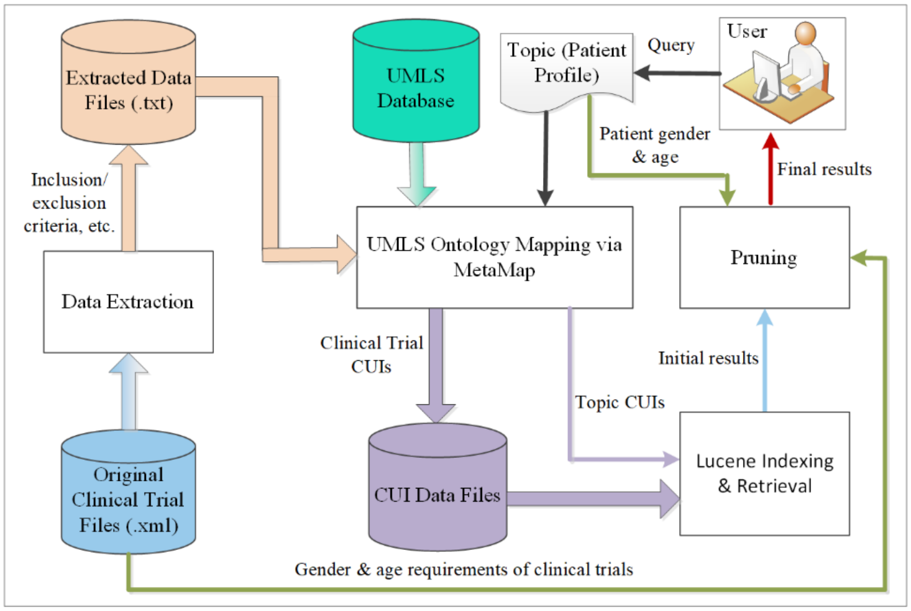
Extracted Data Files (.txt)
UMLS Database
Topic (Patient Profile)
Query
User
Inclusion/ exclusion criteria, etc.
Patient gender & age
Final results
UMLS Ontology Mapping via MetaMap
Pruning
Data Extraction
Clinical Trial CUIs
Initial results
Topic CUIs
Lucene Indexing & Retrieval
Original Clinical Trial Files (.xml)
CUI Data Files
Gender & age requirements of clinical trials

Figure 2. CUI-Based Patient-to-Trial Matching Process

For text-based search, the extracted free-text data files were directly indexed and the free-text topics were the inputs to Lucene. In addition, the expanded medical condition terms were used in the pruning technique only for the text-based search in our implementation.

5. Experiments and Discussions

In our experiments, we examined whether UMLS can improve the performance of our Lucene-based patient-to-trial matching system. We also tested whether our method of using the exclusion criteria in the eligibility section of a clinical trial is valuable. Lastly, we investigated the effectiveness of our expansion technique. Table 1 shows the five runs we submitted. The TxtIncExcExp run represents a strict model where one of the expanded conditions of the final returned clinical trials must be contained by the patient profile.

Table 1. Five Runs Submitted to Clinical Trials Track

Run	Description
TxtInc	Text-based search using inclusion criteria files
TxtIncExc	Text-based search using both inclusion criteria and exclusion criteria files
TxtIncExcExp	Text-based search using both inclusion criteria and exclusion criteria files with term expansion technique
CUIInc	CUI-based search using inclusion criteria files
CUIIncExc	CUI-based search using both inclusion criteria and exclusion criteria files

Table 2 shows the evaluation results of the five runs. The results clearly indicate that the CUI-based search performs better than the text-based search, which verifies UMLS's effectiveness in improving biomedical information retrieval accuracy. The underlying reason is that UMLS not only preprocesses free text (e.g., normalization, acronym extension, stop word removal, etc.), but also converts free text to more meaningful biomedical concepts.

Our approach of using the exclusion criteria data seems not effective. For text-based search, the performance becomes worse for all the measures except *bpref* when the exclusion criteria data are used. This is mainly because Lucene is a keyword-based search engine. Given a query, the matching score of a document depends on the query term frequency in the document and the inverse document frequency in the whole corpus. For a clinical trial, the terms used to describe its exclusion criteria generally belong to the same category of terminology as the inclusion criteria. For instance, both the inclusion and exclusion criteria of a clinical trial designed for a psychiatric disease might include similar terms in the discipline of psychiatry. Therefore, it might be not a good strategy to use the difference of two matching scores searched against the inclusion criteria and exclusion criteria respectively to sort the returned clinical trials. For CUI-based search, the use of exclusion criteria achieved better performance for P@10. For the other three measures, worse performance was obtained, but their numeric values for CUIInc and CUIIncExc are close.

Table 2. Evaluation Results

Run	P@10	bpref	Rprec	ndcg@10
TxtInc	0.1747	0.1275	0.0918	0.2907
TxtIncExc	0.1693	0.1289	0.0680	0.2589
TxtIncExcExp	0.1267	0.0963	0.0450	0.1964
CUIInc	0.2280	0.1293	0.1053	0.3548
CUIIncExc	0.2427	0.1267	0.0930	0.3417

We consistently observed worse performance in the TxtIncExcExp run for the four measures compared with those runs without using the expansion technique. For a topic, the requirement of containing at lease one of the expanded conditions seems so strict that some positive results were pruned. This might be due to various reasons such as complexity of natural language, various ways and formats of expressing a concept, use of acronyms, etc.

6. Conclusion

Our patient-to-trial matching system was based on Lucene along with concept extraction using UMLS, use of exclusion criteria and an expansion technique. Our experiments indicated that using the extracted concepts obtained through UMLS could improve the system's overall performance compared with free-text-based matching. Our approaches that involve using the exclusion criteria and MeSH term expansion did not result in better performance in this study. More effective strategies will be further studied in the future.

References

(2021, October). (The Apache Software Foundation) Retrieved from Apache Lucene: https://lucene.apache.org/

Akers, L., & Gordon, J. S. (2018, November). Using Facebook for Large-Scale Online Randomized Clinical Trial Recruitment: Effective Advertising Strategies. *Journal of Medical Internet Research, 20*(11), e290. doi:10.2196/jmir.9372

Brøgger-Mikkelsen, M., Ali, Z., Zibert, J. R., Andersen, A. D., & Thomsen, S. F. (2020, November). Online Patient Recruitment in Clinical Trials: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research, 22*(11), e22179. doi:10.2196/22179

Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Lang, F.-M., Mork, J. G., . . . Aronson, A. (2012). NLM at TREC 2012 Medical Records Track. Gaithersburg, Maryland.

Demner-fushman, D., Abhyankar, S., Jimeno-yepes, A., Loane, R., Rance, B., Lang, F., . . . Aronson, A. R. (2011). A knowledge-based approach to medical records retrieval. *The Twentieth Text REtrieval Conference (TREC 2011) Proceedings.* Gaithersburg, Maryland.

Gul, R. B., & Ali, P. A. (2010, January). Clinical trials: the challenge of recruitment and retention of participants. *Journal of Clinical Nursing, 19*(1-2), 227-233. doi:10.1111/j.1365-2702.2009.03041.x.

Hersh, W. R. (2007, June). Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *The American Journal of Managed Care, 13*(6 Part 1), 277-8.

Johnson, O. (2015, May). An evidence-based approach to conducting clinical trial feasibility assessments. *Clinical Investigation, 5*(5), 491-499. doi:10.4155/CLI.14.139

Jones, R., Lacroix, L. J., & Porcher, E. (2017, November). Facebook Advertising to Recruit Young, Urban Women into an HIV Prevention Clinical Trial. *AIDS and Behavior, 21*(11), 3141-3153. doi:10.1007/s10461-017-1797-3