Leveraging both Structured and Unstructured Data for Precision Information Retrieval

Yanshan Wang, Ravikumar Komandur-Elayavilli, Majid Rastegar-Mojarad,
and Hongfang Liu

Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
{Wang.Yanshan, KomandurElayavilli.Ravikumar, Mojarad.Majid,
Liu.Hongfang}@mayo.edu

Abstract. This paper describes the participation of the Mayo Clinic NLP team in the Text REtreival Conference (TREC) 2017 Precision Medicine track. The novelty of our systems is four-fold. First, compared to our submissions in the previous year, our systems utilized an enhanced named entity recognition (NER) method to extract genes, variants, proteins, and diseases from PubMed articles. This NER method combined several state-of-the-art NER tools including TaggerOne, beCAS, Reach and tmVAR. The extracted entities were indexed in different fields and treated as structured data for retrieval. Second, we used multi-field querying in a Pseudo Relevance Feedback (PRF) model. We first query the unstructured fields (i.e., the fields of title and abstract) and utilize information in structured fields from top-ranked documents as feedback for query expansion. Third, we explored the use of MeSH on Demand, a web service identifying MeSH terms in free-text and recommending similar PubMed articles which are relevant to the text, to boost the performance of our retrieval systems. The reason we used MeSH on Demand is due to its effectiveness for recommending relevant PubMed articles based on our manual judgments. Fourth, we utilized the demographic information (i.e., age and sex) as structured data to filter out the clinical trials that did not meet the criteria in each topic.

Keywords: information retrieval, named entity recognition, pseudo relevance feedback, mesh on demand

1 Introduction

Precision medicine can result in better treatment outcomes than using the same strategy for everyone since it assigns preventive measures or treatment interventions on the basis of individual characteristics [23]. Advances in next-generation sequencing technology have led to the development of genetic testing for the molecular diagnosis of diseases, particularly cancers. Genetic variants have been shown to be factors implicated in various cancers, such as breast cancer (e.g., mutations in the BRCA1 and BRCA2 genes account for a small proportion of breast cancer cases in the general population [14]), lung cancer (e.g., variants

of large effect in BRCA2 and CHEK2 affect risk of lung cancer [26]), colorectal cancer (e.g., variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk [8]), prostate cancer (e.g., variants in three independent regions at 8q24 and in one region at 17q12 and another at 17q24.3 are significantly associated with a risk of prostate cancer [29]), etc. Since the genetic variant indicates the risk level of a cancer, knowing the specific genetic variant of a cancer should be informative for prevention and treatment [5]. For example, ovarian cancer patients carrying a rare homozygous genotype of rs1425486 in PDGFC have poorer overall survival and worse treatment response than patients carraying common homozygous and heterozygous genotypes [10]. This knowledge is commonly buried in the biomedical literature [6], however, finding the most relevant and recent research can be quite challenging due to the high volume of scientific literature [22].

Information retrieval (IR) provides an efficient and effective way to retrieve relevant documents from a large corpus [25]. In order to evaluate IR methods that retrieve useful precision medicine-related literature for clinicians treating cancer patients, the Text REtrieval Conference (TREC) Precision Medicine (PM) track was organized to encourage IR researchers to study the use of IR in such a clinical setting to help clinicians make better decisions. In this challenge, 30 queries were created by precision oncologists using synthetic cases at the University of Texas MD Anderson Cancer Center. Each query represents a cancer patient, which includes four fields, namely the patient's disease (type of cancer), the relevant genetic variants (which genes and/or which variant), basic demographic information (age and sex), and other potential factors that may be relevant. Two corpus were provided for participants, biomedical articles and clinical trials. Biomedical articles were in the form of article abstracts from MEDLINE/PubMed, addressing relevant treatments for the given patient. Clinical trials were curated from www.ClinicalTrials.gov, addressing relevant clinical trials for which the patient was eligible. The first set of results represents the retrieval of existing knowledge in the scientific literature, while the second represents the potential for connecting patients with experimental treatments if existing treatments have been ineffective. Participants may submit a maximum of five automatic or manual runs for each corpus, each run consisting of a ranked list of up to one thousand documents.

In this paper, we describe our participation in the TREC PM track. The novelty of our systems is four-fold. First, compared to our submissions in the previous year, this year's systems utilized an enhanced named entity recognition (NER) method to extract genes, variants, proteins, and diseases from PubMed articles. This NER method combined several state-of-the-art NER tools including TaggerOne, beCAS, Reach and tmVAR. The extracted entities were indexed in different fields and treated as structured data for retrieval. Second, we used multi-field querying in a Pseudo Relevance Feedback (PRF) model. We first query the unstructured fields (i.e., the fields of title and abstract) and utilize information in structured fields from top-ranked documents as feedback for query expansion. Third, we explored the use of MeSH on Demand, a web service identi-

fying MeSH terms in free-text and recommending similar PubMed articles, which are relevant to the text, to boost the performance of our retrieval systems. The reason we used MeSH on Demand is due to its effectiveness for recommending relevant PubMed articles based on our manual judgments. Fourth, we utilized the demographic information (i.e., age and sex) as structured data to filter out the clinical trials that did not meet the criteria in each topic.

In the rest of this paper, we describe the methods utilized in our systems, including the NER method, the PRF model, the Markov Random Field (MRF) model, and the MeSH on Demand, in Section 2. We describe the details of each submitted run in Section 3, and show the experimental results in Section 4. Finally, we conclude the study in Section 5.

2 Methods

2.1 Named Entity Recognition

A state-of-the-art NER tool developed in our previous studies [21,20] was utilized to extract gene and disease entities from the given texts. The tool used an ensemble of the state-of-the-art named entity normalization tools, PubTator [27] and beCAS [19], supplemented by a dictionary-based lookup for identifying the entities and normalizing them to standard identifiers.

First we used the REST-API services provided by PubTator and beCAS to detect entities from texts. Subsequently, we built a dictionary by compiling different dictionaries from multiple knowledge sources such as Entrez [12], UniProtKB [1], Gene ontology [3], CTD [13], and MeSH [11], and looked up gene and disease names in the composite dictionary. This dictionary lookup resolved three problems where PubTator and beCAS failed: 1) noun phrases lacking morphological features were detected (e.g., PubTator and beCAS failed to detect "bone morphogenetic protein-2" while the tokenization component in the dictionary lookup translated the phrase to "bone morphogenetic protein 2" that could be exactly matched in the dictionary Entrez); 2) acronyms were detected; 3) strings with high surface similarity were detected (e.g., both "Gialpha(1)" and "Gi alpha(2)" were detected by the dictionary lookup while PubTator failed to detect "Gi alpha(2)").

We had certain priority rules to resolve conflicts between the entity recognition systems. Specifically, we utilized the annotations of PubTator for genes, proteins, variants, and diseases when conflicts existed between PubTator and other systems. When PubTator failed to detect those entities, we considered beCAS and the dictionary lookup. Moreover, when a phrase was matched more than once in the dictionary lookup, we chose the dictionary that exactly matched the phrase instead of those with partial matches.

2.2 Retrieval Models

Markov Random Field Model The MRF model is a general framework for IR, which alleviates the bag-of-word assumption by modeling term dependencies

[15]. It takes into account the relationships between query terms by leveraging the Markov property. The MRF model can be represented as an undirected graphical model. Each node in the graph represents a random variable and each edge between two random variables stands for the independence semantics between the random variables. Since the Markov property states that every node in the graph is independent of its non-neighbors given the observed neighbor nodes, different edge configurations indicate distinct term independence assumptions.

Since IR applications mostly consider three independence types, namely *full independence* (FI), *sequential dependence* (SD), and *full dependence* (FD), the corresponding three generalized graph structures are considered in the MRF model. Figure 1 depicts an example of three clique sets in this MRF model given three query terms. As one can see, in the FI structure each query term is assumed to be independent of each other given the document D. The SD structure assumes dependence between adjacent query terms and the FD structure makes the assumption that all query terms are dependent on each other. An ensemble of the three structures can be utilized to rank the retrieved documents. Since there is no existing training data for this task, we intuitively set the weights for FI, SD, and FD to 0.8, 0.1, and 0.1, respectively [25].
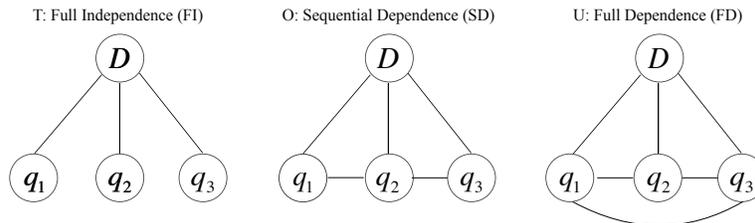

Fig. 1: Alternative dependence structures for an MRF model with three query terms.

Multi-field Pseudo Relevance Feedback The PRF model is a query expansion and refinement technique [4]. It simply assumes that the top-ranked documents retrieved by an IR system are relevant, and words occuring in these documents may be used to expand the initial query. The PRF model has been extensively investigated in the literature, and has been shown to be an effective technique for improving ranking [28,24]. The expansion terms generated by the conventional PRF model depend on the term frequencies in the top-ranked documents. However, this strategy may not be appropriate to use in the PM challenge. The most frequent terms from the top-ranked documents may not be relevant to the query and thus, the expansion terms may worsen the final document ranking.

Since we can extract gene and disease entities from each document, we hypothesize that these genes and diseases are important for retrieving relevant documents. Thus, we expanded the queries by adding these entities from the top-ranked documents (top 20 in our systems). As genes, diseases, and full-text articles were indexed in multiple fields, we called this approach multi-field PRF.

2.3 MeSH on Demand

MeSH on Demand[1] is a tool developed by the National Library of Medicine (NLM) and can be used to identify Medical Subject Heading (MeSH) terms from free text and retrieve similar PubMed or MEDLINE articles according to the MeSH terms. This tool uses the NLM Medical Text Indexer (MTI) [2,18,17] to find MeSH terms. Figure 2 shows a screenshot of MeSH on Demand on the example query "Liposarcoma CDK4 Amplification 38-year-old male GERD" into the web application. Four MeSH terms, namely "Liposarcoma", "Cyclin-Dependent Kinase 4", "Male", and "Gastroesophageal Reflux", can be identified and ten most similar PubMed articles are suggested by MeSH on Demand using both keywords and these MeSH terms. Though the MTI was developed in 2002, it is still relevant and useful for automated indexing recommendations based on recent findings [16]. In order to evaluate the performance of MeSH on Demand for recommending relevant PubMed articles, we manually judged the recommended PubMed articles for the first three topics. We found that these articles were very relevant to the topics, which is consistent with the findings in Mork et al's study [16]. Therefore, the ten MeSH on Demand recommended articles were placed on the top of the final relevant document list with the ranking order computed by our IR systems.

2.4 Query Preprocessing

The query topic in this challenge consists of disease, genetic variants, demographic, and potentially other information about the patients. An example of the provided topics is shown below:
<disease>Melanoma</disease>
<gene>NRAS (Q61K)</gene>
<demographic>55-year-old male</demographic>
<other>Hypertension</other>.
First, we split the "gene" field into two fields, i.e., "gene" and "variant", and put the variant into a "variant" field if a variant is found for a specific gene. Second, we split the information in the "demographic" field into three fields, i.e., "age", "agedes", and "sex". The "age" field contains the age value, the "agedes" field contains the MeSH description for that age (i.e., "aged" for the age older than 65, "middle aged" for the age between 45 and 65, "adult" for the age less than 45), and the "sex" field contains "male" or "female". After the processing, the above query becomes:

---

[1] https://meshb.nlm.nih.gov/MeSHonDemand

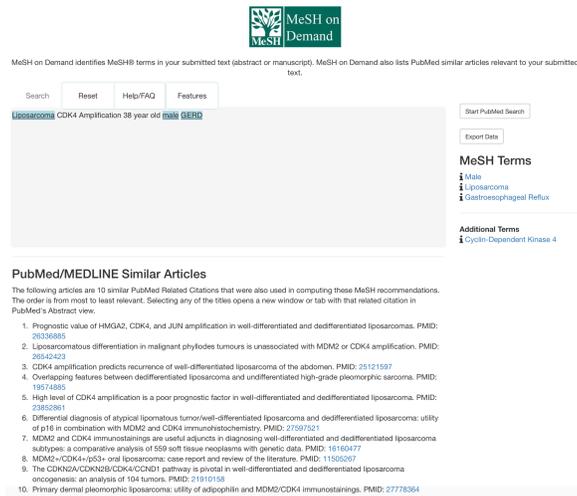Fig. 2: Screenshot of MeSH on Demand on example query "Liposarcoma CDK4 Amplification 38-year-old male GERD".

<disease>Melanoma</disease>
<gene>NRAS </gene>
<variant>Q61K</variant>
<age>55</age>
<agedes>middle aged</agedes>
<sex>male</sex>
<other>Hypertension</other>.

2.5 Indexing

We utilized ElasticSearch[2], an open-source Lucene-based text search engine, for indexing.

For PubMed articles, the title and abstract of each article were merged and indexed into a *abstract* field. The extracted gene and disease entities were indexed in *gene* and *disease* fields, respectively. In addition, we extracted MeSH terms using MetaMap for each PubMed article and indexed them in a *mesh* field. Fields such as the *abstract* field which contained unstructured free-texts were regarded as unstructured fields, and fields such as *gene* and *disease* fields which contained named entities were regarded as structured fields. Various IR models could be applied to query unstructured fields while exact matching was sufficient for the structured fields. For clinical trials, we indexed *brief_summary*, *condition*, *detailed_description*, *keyword*, *mesh_term*, *gender*, *minimum_age*, and *maximum_age* fields. Since many clinical trials contained intervention drugs, we

---

[2] https://www.elastic.co/

utilized semantic MEDLINE [9] to find the ten most related genes to a drug, and indexed the genes in a *gene_semantic_medline* field. Figure 3 shows the indexing fields for both data sources.

PubMed Index

unstructured
abstract

structured
disease
gene
mesh

Clinical Trial Index

unstructured
brief_summary
condition
detailed_description
keyword

structured
mesh_term
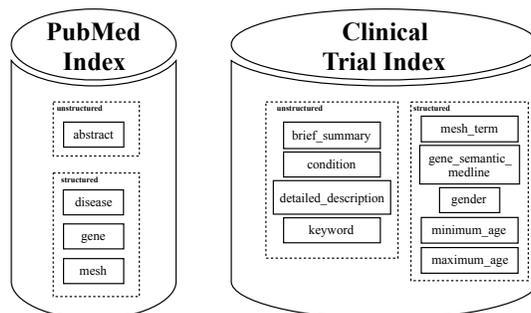gene_semantic_medline
gender
minimum_age
maximum_age

Fig. 3: Indexing fields of PubMed articles and clinical trials.

2.6 Re-ranking

Since both queries and documents contained structured fields, we used exact matching to re-rank the retrieved documents with structured fields. Specifically, a document was ranked higher then other documents if more terms in structured fields were exact matched.

3 Runs

In this section, we describe our submitted runs. Three boolean operators are defined as follows. MUST($q(\cdot)$, $d(\cdot)$) denotes that all the terms in the query field $q(\cdot)$ must be matched in the document field $d(\cdot)$. SHOULD($q(\cdot)$, $d(\cdot)$) represents that at least one of the terms in the query field $q(\cdot)$ should be matched in the document field $d(\cdot)$. MATCH($q(\cdot)$, $d(\cdot)$) represents exact matching when $q(\cdot)$ and $d(\cdot)$ are both structured fields, which is similar to the *WHERE* clause in SQL for relational database. The boolean operators in Elasticsearch takes a more-matches-is-better approach, so the score will be added together to provide the final ranking score for each document.

3.1 PubMed article retrieval task

Figures 4-7 illustrate the pipelines of submitted Runs 1-4. In Run 5, we utilized the pipeline in Run 4 and additionally conducted the re-ranking step before applying the MeSH on Demand. The re-ranking step is the same as used in Run 3 (note that we used the original query terms without query expansion).
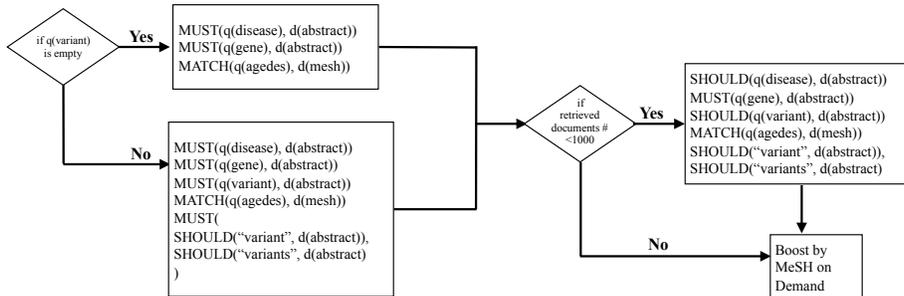
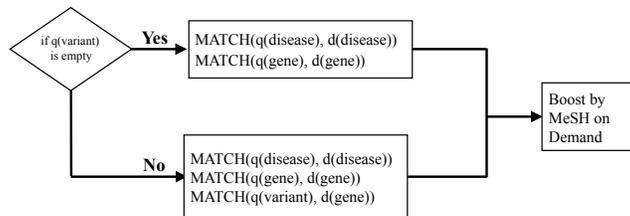Fig. 4: Pipeline of Run 1 for the PubMed article retrieval task.

Fig. 5: Pipeline of Run 2 for the PubMed article retrieval task.
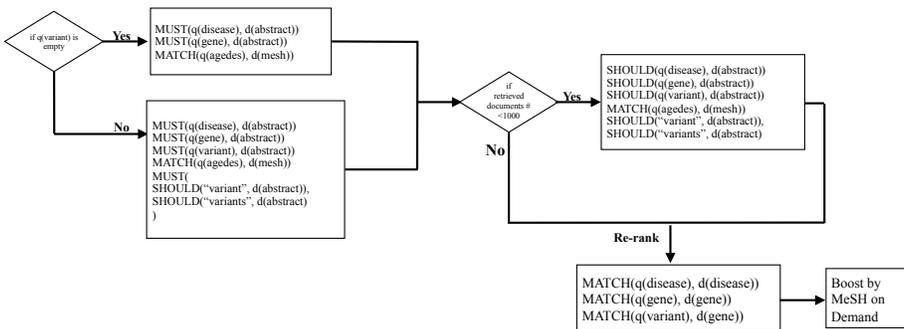
Fig. 6: Pipeline of Run 3 for the PubMed article retrieval task.

3.2 Clinical trial retrieval task

Figures 8-12 illustrate the pipelines of Runs 1-5 for the clinical trial retrieval task. In Runs 3 and 5, we first expanded queries by searching PubMed articles and then used the expanded queries to search clinical trials. Note that a filtering step was applied for each run where we checked whether d(gender) in each retrieved document matched q(sex) and whether q(age) was between d(minimum_age) and d(maximum_age). Documents that did not qualify these two criteria were removed from the final list.

Fig. 7 flowchart:

- if q(variant) is empty
  - **Yes** →
    ```
    MUST(q(disease), d(abstract))
    MUST(q(gene), d(abstract))
    MATCH(q(agedes), d(mesh))
    ```
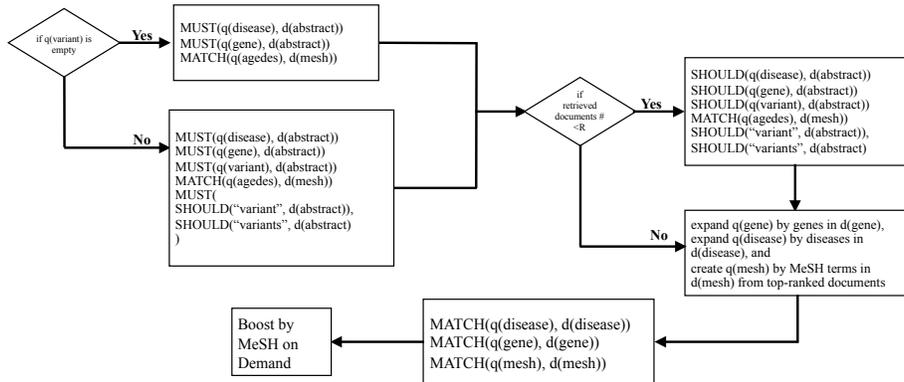 - **No** →
    ```
    MUST(q(disease), d(abstract))
    MUST(q(gene), d(abstract))
    MUST(q(variant), d(abstract))
    MATCH(q(agedes), d(mesh))
    MUST(
    SHOULD("variant", d(abstract)),
    SHOULD("variants", d(abstract)
    )
    ```
- if retrieved documents # <R
  - **Yes** →
    ```
    SHOULD(q(disease), d(abstract))
    SHOULD(q(gene), d(abstract))
    SHOULD(q(variant), d(abstract))
    MATCH(q(agedes), d(mesh))
    SHOULD("variant", d(abstract)),
    SHOULD("variants", d(abstract)
    ```
  - **No** →
    ```
    expand q(gene) by genes in d(gene),
    expand q(disease) by diseases in
    d(disease), and
    create q(mesh) by MeSH terms in
    d(mesh) from top-ranked documents
    ```
- ```
  MATCH(q(disease), d(disease))
  MATCH(q(gene), d(gene))
  MATCH(q(mesh), d(mesh))
  ```
- Boost by MeSH on Demand

Fig. 7: Pipeline of Run 4 for the PubMed article retrieval task.

Fig. 8 box:

```
SHOULD(q(disease), d(brief_summary))
SHOULD(q(disease), d(detailed_description))
SHOULD(q(disease), d(condition))
SHOULD(q(gene), d(detailed_description))
```
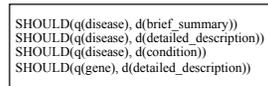
Fig. 8: Pipeline of Run 1 for the clinical trial retrieval task.

Fig. 9 box:

```
SHOULD(q(disease), d(brief_summary))
SHOULD(q(disease), d(detailed_description))
MUST(q(disease), d(condition))
MATCH(q(gene), d(gene_semantic_medline))
SHOULD(q(gene), d(detailed_description))
```
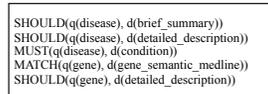
Fig. 9: Pipeline of Run 2 for the clinical trial retrieval task.

Fig. 10 flowchart:

- if q(variant) is empty
  - **Yes** →
    ```
    MUST(q(disease), d(abstract))
    MUST(q(gene), d(abstract))
    MATCH(q(agedes), d(mesh))
    ```
  - **No** →
    ```
    MUST(q(disease), d(abstract))
    MUST(q(gene), d(abstract))
    MUST(q(variant), d(abstract))
    MATCH(q(agedes), d(mesh))
    MUST(
    SHOULD("variant", d(abstract)),
    SHOULD("variants", d(abstract)
    )
    ```
- if retrieved documents # <R
  - **Yes** →
    ```
    SHOULD(q(disease), d(abstract))
    SHOULD(q(gene), d(abstract))
    SHOULD(q(variant), d(abstract))
    MATCH(q(agedes), d(mesh))
    SHOULD("variant", d(abstract)),
    SHOULD("variants", d(abstract)
    ```
  - **No** →
    ```
    expand q(gene) by genes in d(gene),
    expand q(disease) by diseases in
    d(disease), and
    create q(mesh) by MeSH terms in
    d(mesh) from top-ranked documents
    ```
- ```
  SHOULD(q(disease), d(brief_summary))
  SHOULD(q(disease), d(detailed_description))
  SHOULD(q(disease), d(condition))
  SHOULD(q(gene), d(detailed_description))
  MATCH(q(mesh), d(mesh_term))
  ```

Fig. 10: Pipeline of Run 3 for the clinical trial retrieval task.

```
┌────────────────────┐        ┌────────────────────────────────────────────┐
│ Get MeSH terms from │        │ SHOULD(q(disease), d(brief_summary))         │
│ MeSH on Demond for  │───────▶│ SHOULD(q(disease), d(detailed_description))  │
│ each query and      │        │ SHOULD(q(disease), d(condition))             │
│ create q(mesh)      │        │ SHOULD(q(gene), d(detailed_description))     │
└────────────────────┘        │ MATCH(q(mesh), d(mesh_term))                 │
                              └────────────────────────────────────────────┘
```

Fig. 11: Pipeline of Run 4 for the clinical trial retrieval task.

Fig. 12: Pipeline of Run 5 for the clinical trial retrieval task.

4 Results

Tables 1 and 2 list the overall results of the official runs for the PubMed article retrieval task and clinical trial retrieval task, respectively.

As shown in Table 1, the performance of Runs 3 and 5 is identical and the best among the submitted runs. Since Run 5 utilizes the same strategy as Run 4 with an additional re-ranking component as used in Run 3 and since Run 4 underperforms Run 3, we can conclude that the re-ranking by structured data contributes to ranking relevant articles high in the final list. Comparing the performance of Run 4 with other systems in terms of P@10, we find that query expansion decreases the P@10 dramatically. This may be because the query expansion adds too many noisy terms to the original query. It is also interesting that the query expansion fails to improve recall since Run 5 performs as well as Run 3. Thus, query expansion using multi-field PRF fails to improve the IR performance for this task.

Run 2 is inferior to other systems significantly. Since Run 2 only uses the structured field, we can view it as a relational database system. This result implies that adding an IR component to a conventional relational database system will significantly improve the system performance.

Table 2 lists overall results of our submitted runs for the clinical trial retrieval task. As shown in the table, the simplest method, Run 1, has the best performance among the 5 runs. This result indicates that the unstructured free-text

Table 1: Overall results of official submitted runs for the PubMed article retrieval task.

Run Name	infNDCG	P@10	R-prec
Run 1	0.2383	0.3793	0.1352
Run 2	0.2199	0.3724	0.1271
Run 3	**0.2864**	**0.3931**	**0.1698**
Run 4	0.2359	0.0975	0.1370
Run 5	**0.2864**	**0.3931**	**0.1698**

description in each clinical trial is crucial for searching relevant clinical trials. The fact that Run 5 outperforms Run 3 and Run 4 outperforms Run 2 shows that the MeSH terms provided by the MeSH on Demand contribute to the retrieval. Another observation is that incorporating genes from the gene-drug mapping using semantic MEDLINE decreases the performance significantly (Run 2 versus Run 1). The performance of Runs 3 and 5 is worse than Runs 1 and 4, which indicates that query expansion by using feedbacks from relevant biomedical literature fails to improve retrieving relevant clinical trials. This result is consistent with a study that discovers information discrepancies between clinical trial data from ClinicalTrials.gov and biomedical publications from PubMed [7].

Table 2: Overall results of official submitted runs for the clinical trial retrieval task.

Run Name	P@5	P@10	P@15
Run 1	**0.2857**	**0.2393**	**0.2095**
Run 2	0.1357	0.1250	0.1262
Run 3	0.2143	0.1821	0.1690
Run 4	0.2571	0.2250	0.1905
Run 5	0.2357	0.1929	0.1690

Figures 13, 14 and 15 depict results across topics for the PubMed article retrieval task in terms of infNDCG, P@10, and R-Prec, respectively. Our best runs (Runs 3 and 5) outperform the median infNDCG on 17 out of 30 topics. A very interesting observation is that Run 2 performs promisingly on the topics where the best runs perform bad (e.g., topics 3, 11, 13, and 14). Take topic 14 below as an example:
<disease>Cholangiocarcinoma</disease>
<gene>IDH1 (R132H)</gene>
<demographic¿64-year-old male</demographic>
<other>Neuropathy</other>
Runs 1, 3 and 5 fail to retrieve relevant articles. We also find that the P@10 is 0 for this topic for all submitted runs. This result means that the ten articles suggested by MeSH on Demand are judged nonrelevant. The MeSH terms extracted

by MeSH on Demand for this topic are "Cholangiocarcinoma", "Isocitrate Dehydrogenase", "Bile Duct Neoplasms" and "Bile Ducts, Intrahepatic". The suggested articles contain "IDH1" and "IDH1 (R132H)" in the title or abstract, however, most of them are related to "glioma" instead of "cholangiocarcinoma". Our IR systems also failed to find relevant articles for this topic. The reason might be that the term frequency of "IDH1" or "IDH1 (R132H)" is much higher than "cholangiocarcinoma" so the score of matching the term "cholangiocarcinoma" is underweighted in the IR model. The same reason accounts for topic 26 where most retrieved articles are about "NRAS mutations" and "melanoma" instead of "breast cancer".

Figure 14 shows the performance of MeSH on Demand. We observe that it performs well on some topics, such as topics 2, 16, and 23 where P@10 is 1.0 but performs poorly on other topics, such as topics 14, 26, 27, and 30.
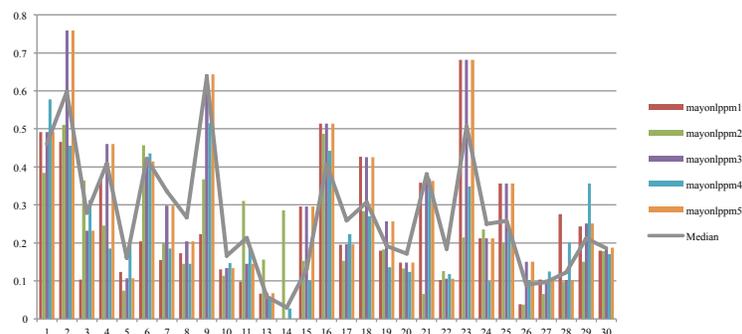
Fig. 13: Distribution of infNDCG over topics compared to median infNDCG for the PubMed article retrieval task.
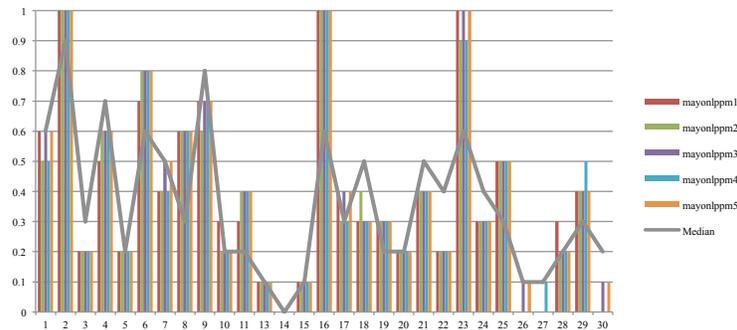
Fig. 14: Distribution of P@10 over topics compared to median P@10 for the PubMed article retrieval task.
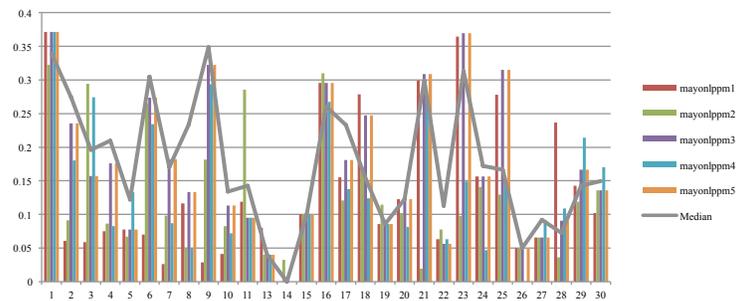

Fig. 15: Distribution of R-Prec over topics compared to median R-Prec for the PubMed article retrieval task.

Since the judgment sets for this task were created using simple depth-15 pools and there existed very few relevant trials, only the precision at cut-off levels of 5, 10, and 15 were used as the official metrics for the clinical trial retrieval task. Moreover, topic 10 was dropped from the evaluation for this task since there were no known relevant trials for it and topic 12 was dropped since the judgments were not completed. Figures 16, 17 and 18 provide results across topics for the clinical trial retrieval task in terms of P@5, P@10, and P@15, respectively. The performance of our best system, Run 1, is better or equal to the median P@5 for 21 out of 28 topics, which implies that the simplest method performs well for the clinical trial retrieval task.

5   Conclusion

This paper describes the participation of Mayo Clinic NLP team in the Text REtreival Conference (TREC) 2017 Precision Medicine track. We explored dif-
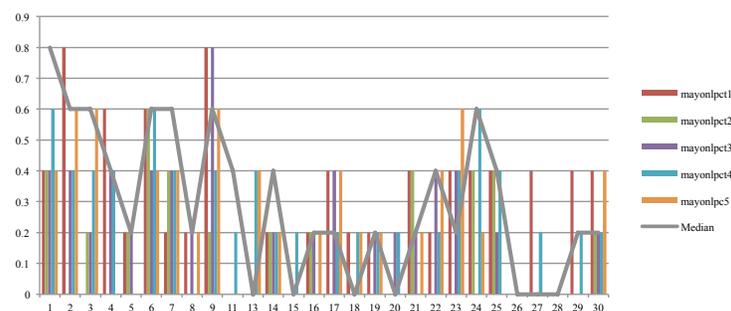
Fig. 16: Distribution of P@5 over topics compared to median P@5 for the clinical trial retrieval task.
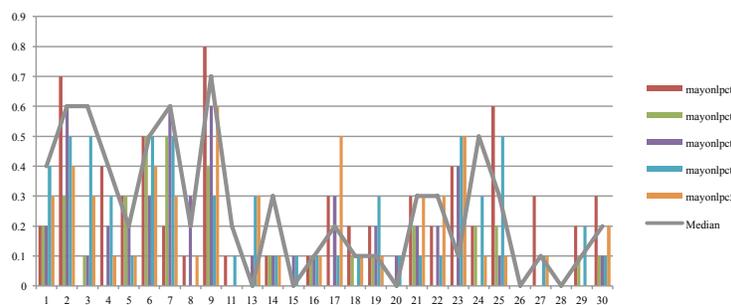

Fig. 17: Distribution of P@10 over topics compared to median P@10 for the clinical trial retrieval task.

ferent approaches in the submitted runs, including the novel NER tool, the PRF model, MeSH on Demand, re-ranking using structured data, and filtering using demographic information. From the experimental results, we conclude that 1) query expansion using PRF in our method failed to improve the IR performance for the PubMed article retrieval and query expansion by using feedbacks from searching biomedical literature failed to improve the IR performance for the clinical trial retrieval; 2) MeSH on Demand performs well on some topics and poorly on others; 3) re-ranking by structured data contributes to ranking relevant articles high in the final list; and 4) unstructured free-text descriptions in clinical trials are crucial for retrieving relevant clinical trials.
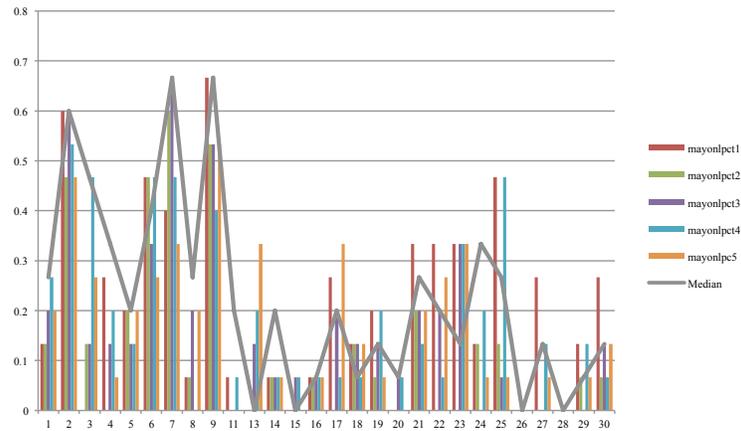
Acknowledgments

Fig. 18: Distribution of P@15 over topics compared to median P@15 for the clinical trial retrieval task.

References

1. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al.: Uniprot: the universal protein knowledgebase. Nucleic acids research 32(suppl_1), D115–D119 (2004)

2. Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J.: The nlm indexing initiative's medical text indexer. Medinfo 89 (2004)

3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics 25(1), 25 (2000)

4. Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice, vol. 283. Addison-Wesley Reading (2010)

5. Cuzick, J., Sestak, I., Bonanni, B., Costantino, J.P., Cummings, S., DeCensi, A., Dowsett, M., Forbes, J.F., Ford, L., LaCroix, A.Z., et al.: Selective oestrogen receptor modulators in prevention of breast cancer: an updated meta-analysis of individual participant data. The Lancet 381(9880), 1827–1834 (2013)

6. Doughty, E., Kertesz-Farkas, A., Bodenreider, O., Thompson, G., Adadey, A., Peterson, T., Kann, M.G.: Toward an automatic method for extracting cancer-and other disease-related point mutations from the biomedical literature. Bioinformatics 27(3), 408–415 (2010)

7. Hartung, D.M., Zarin, D.A., Guise, J.M., McDonagh, M., Paynter, R., Helfand, M.: Reporting discrepancies between the clinicaltrials. gov results database and peer-reviewed publicationsdiscrepancies between clinicaltrials. gov and peer-reviewed publications. Annals of internal medicine 160(7), 477–483 (2014)

8. Jaeger, E., Webb, E., Howarth, K., Carvajal-Carmona, L., Rowan, A., Broderick, P., Walther, A., Spain, S., Pittman, A., Kemp, Z., et al.: Common genetic variants at the crac1 (hmps) locus on chromosome 15q13. 3 influence colorectal cancer risk. Nature genetics 40(1), 26–28 (2008)

9. Kilicoglu, H., Fiszman, M., Rodriguez, A., Shin, D., Ripple, A., Rindflesch, T.C.: Semantic medline: a web application for managing the results of pubmed searches.

In: Proceedings of the third international symposium for semantic mining in biomedicine. vol. 2008, pp. 69–76 (2008)

10. Liang, D., Meyer, L., Chang, D.W., Lin, J., Pu, X., Ye, Y., Gu, J., Wu, X., Lu, K.: Genetic variants in microrna biosynthesis pathways and binding sites modify ovarian cancer risk, survival, and treatment response. Cancer research 70(23), 9765–9776 (2010)

11. Lipscomb, C.E.: Medical subject headings (mesh). Bulletin of the Medical Library Association 88(3), 265 (2000)

12. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez gene: gene-centered information at ncbi. Nucleic acids research 33(suppl_1), D54–D58 (2005)

13. Mattingly, C.J., Colby, G.T., Forrest, J.N., Boyer, J.L.: The comparative toxicogenomics database (ctd). Environmental health perspectives 111(6), 793 (2003)

14. Mavaddat, N., Pharoah, P.D., Michailidou, K., Tyrer, J., Brook, M.N., Bolla, M.K., Wang, Q., Dennis, J., Dunning, A.M., Shah, M., et al.: Prediction of breast cancer risk based on profiling with common genetic variants. JNCI: Journal of the National Cancer Institute 107(5) (2015)

15. Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 472–479. ACM (2005)

16. Mork, J., Aronson, A., Demner-Fushman, D.: 12 years on–is the nlm medical text indexer still useful and relevant? Journal of biomedical semantics 8(1), 8 (2017)

17. Mork, J.G., Demner-Fushman, D., Schmidt, S., Aronson, A.R.: Recent enhancements to the nlm medical text indexer. In: CLEF (Working Notes). pp. 1328–1336 (2014)

18. Mork, J.G., Jimeno-Yepes, A., Aronson, A.R.: The nlm medical text indexer system for indexing biomedical literature. In: BioASQ@ CLEF (2013)

19. Nunes, T., Campos, D., Matos, S., Oliveira, J.L.: Becas: biomedical concept recognition services and visualization. Bioinformatics 29(15), 1915–1916 (2013)

20. Ravikumar, K., Rastegar-Mojarad, M., Liu, H.: Belminer ? information extraction system to extract bel relationships. In: Proceedings of the Biocreative Challenge Workshop (2017)

21. Ravikumar, K., Rastegar-Mojarad, M., Liu, H.: Belminer: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. Database 2017(1), baw156 (2017)

22. Roberts, K., Simpson, M.S., Voorhees, E.M., Hersh, W.R.: Overview of the trec 2015 clinical decision support track. In: TREC (2015)

23. 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)

24. Wang, Y., Rastegar-Mojarad, M., Elayavilli, R.K., Liu, S., Liu, H.: An ensemble model of clinical information extraction and information retrieval for clinical decision support. In: TREC (2016)

25. Wang, Y., Wu, S., Li, D., Mehrabi, S., Liu, H.: A part-of-speech term weighting scheme for biomedical information retrieval. Journal of biomedical informatics 63, 379–389 (2016)

26. 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 in brca2 and chek2 affect risk of lung cancer. Nature genetics 46(7), 736–741 (2014)

27. Wei, C.H., Kao, H.Y., Lu, Z.: Pubtator: a web-based text mining tool for assisting biocuration. Nucleic acids research 41(W1), W518–W522 (2013)

28. Yu, S., Cai, D., Wen, J.R., Ma, W.Y.: Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In: Proceedings of the 12th international conference on World Wide Web. pp. 11–18. ACM (2003)
29. Zheng, S.L., Sun, J., Wiklund, F., Smith, S., Stattin, P., Li, G., Adami, H.O., Hsu, F.C., Zhu, Y., Bälter, K., et al.: Cumulative association of five genetic variants with prostate cancer. New England Journal of Medicine 358(9), 910–919 (2008)